Author:
Chrysos, Alexandros

Title:
Economic Analyses of a Novel Diagnostic Device in Endocrine Disease

Author:
Chrysos, Alexandros

Title:
Economic Analyses of a Novel Diagnostic Device in Endocrine Disease

ECONOMIC ANALYSES OF A NOVEL DIAGNOSTIC DEVICE IN ENDOCRINE DISEASE

By

Alexandros Epameinondas Chrysos

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Health Sciences

Bristol Medical School

October 2020

Word Count: 77,485

Abstract

Background

Diagnosing endocrine disorders with conventional methods is challenging and expensive. *U-Rhythm* is a novel device that collects 24-hour hormone profiles while users continue with daily activities. An early case-control study (*ULTRADIAN*) showed that *U-Rhythm* can play a role in differentiating endocrine conditions. The aim of this thesis was to assess *U-Rhythm*'s potential health economic value and explore the challenges of conducting an early economic evaluation.

Methods

A systematic review of economic evaluations of diagnostic tests for six selected endocrine disorders was performed. *ULTRADIAN* data on patients' and healthy controls' device satisfaction, recent healthcare usage and health-related quality of life were collected prospectively at four European centres. The burden of primary aldosteronism for the UK healthcare system was measured in a retrospective analysis of routinely collected healthcare data (CPRD-HES). A decision tree and Markov model was developed to conduct a cost-effectiveness, price threshold and value of information analysis to explore the potential cost-effectiveness of *U-Rhythm* in the diagnosis of primary aldosteronism and identify areas for further research.

Results

Only seven economic studies for three endocrine disorders were identified. *ULTRADIAN* data indicated patients' general satisfaction with *U-Rhythm*, higher healthcare usage and lower health-related quality of life compared to healthy volunteers. The twenty-year healthcare costs associated with primary aldosteronism were double those of patients without the disease. *U-Rhythm* was found to be less cost-effective than existing tests for the confirmatory diagnosis of primary aldosteronism. Further investment in research and development was not worthwhile unless *U-Rhythm*'s costs are reduced and/or diagnostic accuracy increases as the technology evolves.

Conclusions

Early economic evaluation of diagnostics is hard due to limited and low-quality data, multiple potential uses in clinical practice, and efficiency depending on subsequent technologies. Decisions on further investment should be taken after their key parameters have been established and compared with currently available products.

Dedication and Acknowledgements

The author would like to dedicate this PhD research to his loving parents Olga and George who have instilled in him the value of education, and the drive and ambition to succeed in all his personal, educational and career pursuits. He would also like to dedicate this work to his sister Tina and to the memory of his grandmother Dina who have morally supported him throughout his entire life and made him the person he is today.

The researcher would like to express his sincere gratitude to his two PhD supervisors, William Hollingworth (Professor of Health Economics, University of Bristol) and Stafford Louis Lightman (Professor of Medicine, University of Bristol), for entrusting him to undertake this project; providing him with valuable knowledge in the amazing areas of health economics and endocrinology; and supporting him mentally and financially throughout these years. Thank you for your endless time and help. Without you this hard task would have been impossible.

The author would also like to express his deepest appreciation to several people that have assisted him with completing this project. Firstly, he would like to thank Dr Thomas Upton and Dr Georgina Russell (University of Bristol) for all the clinical knowledge and help that they provided throughout the study period. Secondly, he would like to express his gratefulness to Dr Marianne Aardal Grytaas, Dr Katerina Simunkova and Dr Marianne Øksnes (University of Bergen) for all their time and help with the development of the decision-analytic model and 'device satisfaction' questionnaire as well as the collection of the *ULTRADIAN* data. Thirdly, the author is thankful to Dr Diane Fraser (University of Exeter) and Dr Eder Zavala (University of Birmingham) for providing him the data regarding the diagnostic accuracy of the *U-Rhythm* device; Dr Timothy Jones, Dr Maria Theresa Redaniel (University of Bristol) and Dr Manjula Nugawela (University College London) for providing the author the CPRD, HES, IMD and ONS data, and assisting with their cleaning; Ms Catherine Borwick (University of Bristol) for her advice on the search strategy of the systematic review; Dr Padraig Dixon (University of Oxford) for advising on regression and econometric modelling; Dr Georgina Hazell (University of Bristol) and Mr Robin Crossley (Designworks) for the information given on the further development and studies, CE marking, and commercialisation of the *U-Rhythm* device; and

the rest of the *ULTRADIAN* partners for sharing together the experience of seeing the development of *U-Rhythm* over these four years. Furthermore, the researcher would like to thank all the PhD students with whom he shared the same offices and nice moments as well as all the members of the Health Economics team in Bristol for their help and support whenever they were asked for. Lastly, a special thank you should be said to the author's PhD examiners, Dr Paul McCrone (University of Greenwich) and Dr Sabina Sanghera (University of Bristol), for the time they spent to read this thesis, and for their comments on aspects that have helped the author become a more responsible and independent researcher.

Given that this thesis used CPRD, HES, IMD and ONS data, the following acknowledgement statements should also be included:

"*This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data are provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author alone*".

Author’s Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: [redacted] **DATE:** 4 October 2020

Table of Contents

List of Tables

List of Figures

CHAPTER I

INTRODUCTION

INTRODUCTION TO DIAGNOSIS IN RARE ENDOCRINE DISEASE AND EARLY ECONOMIC EVALUATION OF DIAGNOSTIC TESTS AND MEDICAL DEVICES

CHAPTER I OVERVIEW

Diagnosis has always been a key feature of clinical practice since it is the medical process that seeks to interpret the nature and causes of a patient's condition before determining optimal management. The evaluation of diagnostic technologies in terms of their quality, accuracy, safety, effectiveness and cost-effectiveness has recently received increasing attention since many national and international regulatory bodies require this before a product receives marketing and reimbursement approval. The first four elements are tested in the laboratory and during clinical trials, while cost-effectiveness is examined by conducting an economic evaluation (i.e. systematic process that compares alternative interventions on both their costs and health benefits). Although economic evaluation has been widely used for therapies (e.g. drugs), its use in diagnostics has been more limited and can be challenging. Furthermore, economic evaluation is very useful when performed at an early stage of product development since it can help manufacturers to reduce development and production costs, and policy makers to plan on how to allocate resources more efficiently. Nevertheless, early economic evaluation is difficult to conduct due to the limited availability, low quality and uncertainty of the data, while it is even harder for diagnostics (e.g. multiple differential diagnoses with different prognoses and treatment pathways).

This PhD thesis focuses on the early economic evaluation of diagnostic medical devices. To do so, it uses a case-control study (*ULTRADIAN*) of a novel test (*U-Rhythm*) that might play a role in the diagnosis and monitoring of a range of endocrine disorders to explore some analytical methods of early economic evaluation of diagnostic tests and provide preliminary evidence on the device's cost-effectiveness in one of the diseases (primary aldosteronism). **Chapter I** starts by providing a general background of this PhD project and information on its rationale, scope and objectives. It also describes briefly the *ULTRADIAN* study and the endocrine diseases that it examined as well as the current regulations and health technology assessment guidelines that control the marketing and reimbursement of medical devices and diagnostic tests in the European Union and the United Kingdom. **Chapter I** ends by giving a brief overview of the analyses performed in the **following Chapters (II-VI)**.

Introduction to Diagnosis in Rare Endocrine Disease and Early Economic Evaluation of Diagnostic Tests and Medical Devices

1.1 Background Information

"Diagnosis is not the end, but the beginning of practice". These words were said by *Martin Henry Fischer*, a 19th century American physician-scientist (1), to show the importance and place of *diagnosis* in clinical practice (**Figure 1**). Diagnosis is the clinical process that explains the nature of a patient's clinical signs and symptoms by determining the causes (i.e. disease) that are underlying and distinguishing them from other possible reasons. The term *'diagnosis'* has its origins from the Greek word *'gnosis'* which means *'knowledge'*, and is used to describe the knowledge that is required to predict the development of a condition (*'prognosis'*) and make a decision on the appropriate therapy. The diagnostic process can either be simple, i.e. collecting the patient's medical history and/or performing a physical examination, or more complicated, e.g. using laboratory and/or imaging tests to define signs and symptoms that might overlap between different diseases (*'differential diagnosis'*). Diagnosis is usually more challenging when made at an earlier disease stage. This is because, at that point, most signs and symptoms are non-disease-specific and difficult to differentiate. However, it is important that diagnosis is accurate and made in a timely manner to increase the chances for successful treatment, and subsequently a patient to have a positive health outcome (2-4).

Steps	STEP 1: Patient having a medical problem	STEP 2: Contacting a physician	STEP 3: Data gathering and clinical reasoning	STEP 4: Diagnosis and problem definition	STEP 5: Treatment options and management decisions	STEP 6: Monitoring
Aim		Patient seeks medical advice and care	Physician begins to understand the problem	Physician defines the problem to address	Physician formulates a solution strategy	Physician evaluates new data and modifies plan
Procedure	Patient presents signs and symptoms that affect personal health and general quality of life	Patient contacts a physician to inform about the problem (e.g. family doctor, specialist)	Data gathering: • Patient's concerns, complaints and/or personal characteristics • Medical history • Physical examination • Laboratory and/or imaging testing	Differential diagnosis and/or problem list creation: • Translating data into a defined diagnostic construct • Analysing probabilities and managing uncertainty • Broad definition of problems (e.g. physical, mental, social)	Treatment/solution options: • Evidence-based solutions • Patient context factors (e.g. goals, values, expectations, willingness) • Assessing the risks and benefits of treatment • Ethical dilemma	Monitoring response to management plan: • Use of time • Emergence of new data
		Visit takes place in one or multiple clinical settings (e.g. office, home)	Clinical reasoning and hypothesis generation: • Analysing, interpreting and grouping data • Examining test and/or sign characteristics	Prioritising and selecting diagnosis and problems to address first: • Assessing prognosis • Ordering problems by concern and/or urgency	Management plan decision: • Selecting appropriate treatment/solution • Patient preferences	
Challenges	• Undifferentiated, multidimentional or vague problems	• Patient's inability to contact the physician due to disease severity • Doctor's inavailability • Timing of contact	• Inaccurate and/or incomplete data (bias) • Time pressures • Premature closure • Knowledge gaps	• Wrong diagnosis • Ignoring/not identifying important problems and/or other risk factors • Selecting easy problems	• Ignoring patient context factors and compliance • Failure to seek patient input and preferences	• Failure to use time to reveal more data

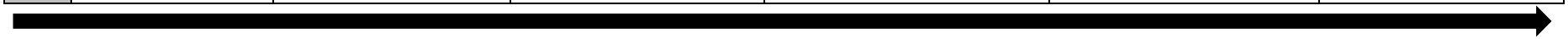


Figure 1: Clinical pathway [**Source:** Image developed based on a figure presented in *Rakel et al. (2011) (2)*]

1.1.1 History of Diagnosis

Thousands of years ago in Ancient Egypt, Mesopotamia and Greece early physicians predicted the progression and outcome of an illness through observation, palpation and auscultation, or later by examining human specimens (**Table 1**). Urine was the first body fluid to be used for diagnostic purposes (before 400 BC), while uroscopy became an essential part of clinical practice in the Middle Ages (5^{th}-15^{th} century), making the urine flask the emblem of medieval medicine. After the discovery of blood circulation and the invention of the compound microscope (16^{th}-17^{th} century), physicians started to understand better a variety of body functions and processes (e.g. metabolism, respiration) as well as the structure of its cells and tissues. Additionally, they began explaining how microorganisms cause complications, dysfunctions and disorders (4-10).

In the 1800s, more advanced diagnostic tools and techniques started being used. For example, the thermometer for measuring body temperature and the stethoscope for measuring heart rate, which together with X-rays, chemical and bacteriological tests were mainly established in clinical routine at the end of the century. At the beginning of the 20^{th} century, the clinical laboratory became essential in medicine with many of them including new diagnostic tools, such as radioactive isotopes (i.e. variants of chemical elements) and electroencephalogram. Before the 2000s, more sophisticated diagnostic methods started becoming widely used (e.g. magnetic resonance imaging), and today, there is a huge variety of diagnostic tests and procedures that are commonly used in clinical practice. Most of these methods can generally be divided into three categories based on the technology that they use: medical imaging, in vitro diagnostics and laboratory tests, and medical devices (4-10). The latter group is examined in the **next Chapters** of this thesis.

Table 1: Important dates in the history of diagnosis

Date*	Historical Event
c. 300 BC	Hippocrates uses urine samples, lung sounds, skin colour, bodily signs and outward appearances for diagnostic purposes
AD 130-201	Galen's work makes him the first experimental physiologist in history
1267	R. Bacon uses optics in his experiments and potentially produces the first microscope
1300-1500	Uroscopy and other forms of urinalysis are routinely used in Europe as diagnostic techniques
1543	A. Vesalius' work on the dissection of human bodies makes him the founder of modern anatomy
1578-1657	W. Harvey discovers the circulation of blood and marks the beginning of a period of explanations for several body functions
1590	H. Janssen and his son invent the compound microscope
1592	G. Galilei invents the first thermometer; later, he constructs a microscope and a telescope
1707	J. Floyer introduces the concept of measuring pulse rate by timing pulse beats with a watch
1714	D.G. Fahrenheit develops the mercury thermometer and creates the Fahrenheit temperature scale
1742	A. Celsius develops the Celsius temperature scale
1754-1761	L. Auenbrugger starts using chest percussion for diagnostic purposes; the method is popularised in 1808 by J.N. Corvisart
1770	J. Hill begins collecting specimens and using them in microscopic study
1816	R.T.H. Laennec invents the stethoscope; G.B. Amici develops the dioptric/achromatic microscope
1828	The first clinical wards appear in British hospitals
1830	G.J. Mulder conducts the first chemical analysis of proteins; J.J. Lister uses an achromatic microscope (dark-field microscopy)
1850	H. von Helmholtz invents the ophthalmoscope

Date*	Historical Event
1865	G. Mendel's work on explaining hereditary traits using peas marks the beginning of molecular genetics
1866	T.C. Allbutt introduces the modern clinical thermometer; the method is popularised later (1871) by C.R.A. Wunderlich
1871	The dry-plate photographic process is invented
1875	W. Erb and C. Westphal are the first to use a hammer to test reflexes
1881	S.S.K von Basch invents the sphygmomanometer; the tool becomes popular in 1901 by H. Cushing
1874	R.H. Chittenden establishes the first laboratory of physiological chemistry in the United States
1875	The first public health laboratory is established in England
1895	W.C. Roentgen discovers the medical use of X-rays
1897	The first commercial clinical laboratory is established in England
1903	W. Einthoven creates the first electrocardiograph machine
1904-1922	O. Follin develops a quantitative method for urine analytes
1916	K.M.G. Siegbahn develops the X-ray spectroscopy
1919	F.W. Aston develops the mass spectroscopy
1945	Hospital services started being charged depending on the cost of the items that were used
1947	E. Land develops the polaroid camera
1950	J. Watson and F. Crick propose a structure for DNA based on the crystallography work of R. Franklin and M. Wilkins
1959	I. Donald uses ultrasound for diagnosis in obstetrics and gynaecology
1972	G. Hounsfield and A. Cormack develop the first computerised axial tomography scanner

Abbreviations:* *AD: After Death; BC: Before Christ; c: century; DNA: deoxyribonucleic acid* [Sources:** (4-7, 10, 11)]

1.2 Research Rationale

Over the last two decades, the evaluation of diagnostics has received increasing attention with many studies suggesting that most of the technologies that have been or are being introduced into medical practice have not been adequately assessed in terms of their costs and health benefits. Additionally, although the methodology when examining the accuracy of diagnostic tests (i.e. ability to confirm/exclude the disease) is well-established, it is currently difficult to relate them to patient outcomes since by definition they should be combined with an effective therapy to improve patient health (12, 13). Therefore, two main questions that should be answered when assessing a new diagnostic technology and before introducing it into clinical practice are:

a) Does this improve patient care and outcomes?
b) Is the cost justified by the benefits that it provides? In other words, is the new test *cost-effective*?

1.2.1 Clinical Evaluation of Diagnostics

In 1986, *Guyatt et al. (12)* tried to answer the question on effectiveness by publishing a comprehensive framework for the clinical evaluation of diagnostics. The authors described six criteria that should be considered before the final market dissemination of a test:

1. **Technological capability**, i.e. the ability to meet laboratory specifications;
2. **Range of possible uses**, i.e. the spectrum of all the conditions in which it can be used;
3. **Diagnostic accuracy**, i.e. the ability to distinguish patients with and without the disease and inform on its severity;
4. **Impact on healthcare providers**, i.e. the confidence that it gives to health professionals that its results are a reliable basis for diagnosis;
5. **Therapeutic impact**, i.e. how it affects the decision regarding treatment; and
6. **Impact on patient outcome**, i.e. what the benefit for the patient is when this is used.

1.2.2 Economic Evaluation of Diagnostics

In addition to the six abovementioned criteria, a seventh criterion related to both the cost and health benefits of a test is equally important and should be examined in parallel. Cost-effectiveness is an essential element for any intervention to ensure that it is worthwhile to invest in this compared to other alternatives (14-16). This is because, no matter how accurate a test is, the decision for using it is restricted by the inability to adopt all interventions due to resource scarcity (e.g. time, human and economic resources). In economics, this is the concept of *'opportunity cost'*, i.e. the value of benefits obtained by one intervention that will be forgone if an alternative is chosen to be reimbursed (17-19). For instance, if a test is very expensive and is used frequently, other tests (and treatments), which may yield more benefits for patients, will be crowded out of the healthcare system. To overcome this problem, an *economic evaluation (EE)* should be conducted (17).

EE is a systematic process in which different healthcare programmes are compared in terms of both their costs and health consequences to identify the most cost-effective option (17). In recent years, an increasing number of countries, especially those whose healthcare system is predominantly publicly funded (e.g. Australia, Canada, the United Kingdom), have made cost-effectiveness a requirement for some medical products (e.g. drugs) before they receive reimbursement (20-25). Nevertheless, the role of EE in non-pharmaceutical technologies, especially diagnostics, has often been overlooked despite their growing use in clinical practice (22-24, 26, 27). Today, several health technology assessment authorities around the world have made the step forward and developed EE guidelines for diagnostic tests. However, these are often provided as a non-mandatory guidance for product developers and commissioners rather than being a requirement before coverage and/or reimbursement (28).

EE is usually conducted at a late stage of development (e.g. phases III-IV for pharmaceuticals) (20, 21, 29), after demonstration of quality, safety and effectiveness (**Figure 2**). For this reason, cost-effectiveness is often described in the literature as the *'fourth hurdle'* in product development (20, 21). According to *Vallejo-Torres et al. (2008) (25)*, the innovation pathway of a medical product can be divided into three stages: *early*, *mid* and *late*. In the early stage, no clinical or economic data are available, and EE relies on assumptions and expert beliefs

about the likely costs (e.g. test cost) and benefits (e.g. impact on therapy). In the mid phase, some evidence has become available from preclinical studies (i.e. before clinical trials), but uncertainty still surrounds the data. The aim here is to identify the parameters with the biggest impact on the product's cost-effectiveness and the key areas where further research should be conducted. In the late stage, data from clinical trials are available, and EE can be performed to inform market approval and reimbursement decisions (25, 29, 30). Sometimes, EE is even conducted after an innovation receives regulatory approval and enters the market. This is because it is only then that developers need information on cost-effectiveness to convince regulators, payers, and clinicians that their technology offers value for money. Additionally, at this point, enough clinical and economic data have become available to more accurately inform coverage decisions (14, 23-25, 30-33).

Lately, many manufacturers are turning their attention to determining a product's potential clinical and economic value at earlier stages of development (R&D) to understand its potential role in clinical practice, and evaluate the costs and outcomes that are important for patients and society (14, 23-25, 32-34). For diagnostics, these early phases are the *'concept'* and *'early development'* stages (equivalent to preclinical and phases I-II for drugs). The concept stage is when the test is first discovered or formed as an idea and includes the period where no data are available. The early development stage follows and is when the product passes some preclinical and clinical tests before being used in bigger trials. Here, a few experimental data are available, and can be used in clinical and economic assessments (23, 24).

By using early EE, developers can examine the market, clinical area and type of technology in which they should invest; the product's role and cost-effectiveness in different settings or scenarios (e.g. disease groups); the product's competing interventions; and the outcomes that should be measured. Early EE can lead to lower production costs since developers can reduce uncertainty when making go/no-go decisions, and avoid investing in projects or trials for diagnostic tests that might be unsuccessful or not cost-effective (14, 22, 24, 25, 29-34). This in combination with the knowledge of the product's highest possible price in a given market can then lead to increased revenues, providing potential to re-invest in new innovations (30, 32-34). Policy makers can also benefit from earlier and better information

about emerging technologies, and make faster reimbursement decisions, which means faster accessibility to new products and health improvement for patients (14, 23, 34, 35).

<table>
<tr><td>Innovation Phase</td><td colspan="3">Early Stage</td><td colspan="3">Mid Stage</td><td>Late Stage</td></tr>
<tr><td>Drugs</td><td>Idea</td><td>Invention prototype</td><td>Preclinical</td><td>Phase I</td><td>Phase II</td><td>Phase III</td><td>Phase IV</td></tr>
<tr><td rowspan="2">Diagnostics</td><td colspan="3">Concept</td><td colspan="3">Test development</td><td>Post-market</td></tr>
<tr><td colspan="2">Test performance</td><td colspan="2">Test accuracy</td><td colspan="3">Clinical effectiveness</td></tr>
<tr><td>Type of Economic Evaluation</td><td colspan="5">Early economic evaluation</td><td colspan="3">Late economic evaluation</td></tr>
<tr><td rowspan="2">Applications</td><td rowspan="2">Strategic R&D decision making</td><td colspan="2" rowspan="2">Preclinical preliminary market assessments</td><td colspan="2">Go/no-go decisions</td><td colspan="3">Assessment of future reimbursement and pricing scenarios</td></tr>
<tr><td colspan="2">Development of future study design</td><td colspan="3">Price determination</td></tr>
<tr><td>Decisions</td><td>Invest in fundamental research?</td><td colspan="2">Continue developing?</td><td colspan="2">Continue to patient-oriented research?</td><td colspan="3">Bring into the market and include into insurance packages?</td></tr>
</table>

[**Sources:** Image developed based on figures presented in *Frempong et al. (2018) (23)* and *Redekop et al. (2013) (30)*]

Figure 2: Health technology innovation pathway

Nevertheless, early EE is difficult mainly because of the limited availability, low quality and uncertainty that surrounds the data. For diagnostics, key parameters to consider are: a) disease prevalence; b) diagnostic accuracy; c) disease state transition probabilities; d) test and treatment costs; and e) treatment effectiveness. In addition, the place of a test in the clinical pathway is not always known; sources of evidence are inaccurate (e.g. expert opinion; estimates/assumptions based on unpublished data; early-phase study results); and data differ from real-world practice. Therefore, any conclusions should be drawn with caution to guide the product development (14, 22-25, 32-34, 36). Due to these difficulties, manufacturers may be reluctant to prioritise resources (time and money) to conduct early EEs. They have more interest in quickly launching their product ahead of competitors and conducting EEs, if required, at the time when their technology is marketed (24, 33). Consequently, if early EE is to be practical, it needs to be simple and informative.

1.3 Scope of PhD Project

The aim of this thesis is to first identify and appraise the different frameworks that have been proposed for the early EE of diagnostic devices, and then make suggestions on the tools that could be used in the future. To examine their suitability and applicability, this thesis uses as an example an innovative device (*U-Rhythm*) for the diagnosis and treatment monitoring of rare endocrine disorders that is currently under development. The clinical and economic data for the device were collected from a European Union (EU) Commission funded study, called *Dynamic Hormone Diagnostics (ULTRADIAN, http://www.uib.no/en/ultradian)*, which was designed to further develop and evaluate the medical device. This thesis mainly focuses on examining the cost-effectiveness of the new device when this is used in the confirmatory diagnosis of primary aldosteronism. Additionally, since this thesis was conducted in the UK, it uses the National Health Service (NHS) as the setting for the analyses.

1.4 Dynamic Hormone Diagnostics (ULTRADIAN)

1.4.1 ULTRADIAN Background

The *endocrine system* is a series of glands that secrete chemical molecules, called *hormones*, into the bloodstream, which are used to send signals and control a variety of bodily functions, such as growth, sexual function, blood sugar and calcium control, and homeostatic regulation. Most hormones are normally released in pulses whose amplitude and frequency are used by the body to fine-tune their biological effect. Variations in hormone levels can appear in several cyclical patterns of different time scales, such as <24 hours (*ultradian* rhythm), the light-dark 24-hour cycle (*diurnal* or *circadian* rhythm), or >24 hours (e.g. monthly menstrual cycle). Any abnormal changes in hormone release patterns can cause diseases, called *endocrine disorders*, which are often difficult to diagnose and treat (37-40).

Diagnostic guidelines provided by the *Endocrine Society (www.endocrine.org)* (41-46) emphasise the difficulty in differentiating endocrine disorders using existing diagnostic tools. This is mainly due to the wide variation in symptoms between patients, which often are non-disease-specific, and the natural fluctuation in hormone levels that makes the use and interpretation of conventional tests challenging. Many current diagnostic and monitoring tests are *'static'*[1], with samples taken at one or small number of points in time (47). This means that they cannot adequately take into account natural fluctuations in hormone levels occurring over a period of time, resulting in broad reference ranges and uncertainty about test interpretation (38-40). As a result, diagnosis can sometimes be protracted, inaccurate and expensive, requiring several primary care consultations and visits to hospital specialists as well as several inpatient, imaging and/or laboratory testing before the diagnosis is finally determined. Furthermore, the treatment of endocrine diseases is not always straightforward and may involve different types of surgery, long-term medication and regular monitoring, meaning that the total cost of care can be relatively high (48-52).

[1] Tests are divided into *'static'* and *'dynamic'*. Static tests are single assays (e.g. of hormones) where samples are collected at specific timepoints and results are compared to acceptable baseline values to examine whether an organ system is functioning normally. Dynamic testing involves the use of exogenous factors (e.g. drug, physical stress) to evaluate the response of the patient's organ system at different times (47).

1.4.2 ULTRADIAN Objectives and Design

There are many endocrine diseases some of which are more common (e.g. diabetes mellitus) and some are very rare (e.g. Addison's disease) (37). In most of them, there is limited evidence on the use and costs of primary and secondary care that is provided before and after their diagnosis (48-52). Between 2016-2020, the *ULTRADIAN* study aimed to improve and simplify the diagnosis, management and treatment monitoring of rare endocrine disease in Europe. This was a multi-centre, observational, prospective study, involving four European clinical centres (University of Bergen, Norway; Karolinska Institute, Sweden; University of Bristol, UK; the Evangelismos General Hospital, Greece).

The study evaluated an innovative, portable, minimally invasive diagnostic device (*U-Rhythm*) to investigate and compare 24-hour non-protein-bound (free) hormone profiles in the subcutaneous tissue of healthy individuals and patients with acromegaly, Addison's disease, congenital adrenal hyperplasia, pituitary or adrenal Cushing's syndrome, growth hormone deficiency, and/or unilateral or bilateral primary aldosteronism. In other words, the study aimed to characterise the physiological rhythm of hormones throughout a 24-hour period, compare it with pathophysiological rhythms caused by the six diseases, and define the limits of normality allowing the early and accurate diagnosis of endocrine disease.

1.4.3 Endocrine Disease Description

Below, a brief description of the main causes; clinical characteristics; and ways of diagnosis, management/treatment and monitoring of primary aldosteronism, Cushing's syndrome and Addison's disease is provided (**Table 2**). The main analysis of this thesis focuses only on these three conditions. ***Appendix 1*** gives similar information for the other three endocrine diseases.

1.4.3.1 Primary Aldosteronism

Primary (hyper)aldosteronism (PA) or *Conn's syndrome* (first described by *Jerome Conn* in 1955) (53) affects 5-15% of the resistant hypertensive population annually (46, 54, 55) and is the result of autonomous overproduction of *aldosterone* which cannot be suppressed by sodium loading. Aldosterone is responsible for keeping a balance between sodium-potassium

levels in the blood. Its secretion is normally regulated by the *renin-angiotensin system*, which controls blood pressure and fluid balance. When aldosterone is overproduced, the mineralocorticoid receptors in the kidneys are activated, potassium is excreted from the body and sodium is retained. If prolonged and severe, the increased potassium excretion may lead to hypokalaemia. The excess sodium levels can result in water retention, which subsequently increases the blood volume causing high blood pressure (46, 53, 56-58).

PA most commonly presents in adults. It has many symptoms, including muscular weakness, and can result in several complications (e.g. cardiovascular, renal). Among the causes of PA are aldosterone-producing adrenal adenomas or bilateral adrenal hyperplasia. In rare cases, PA can be caused by adrenal carcinomas or familial hyperaldosteronism (46, 53, 56-58). To identify patients with PA and differentiate them from other causes of hypertension, blood test screening measuring the plasma aldosterone-renin ratio is commonly performed first. If the screening is positive, additional confirmatory testing (e.g. saline infusion test) is recommended to confirm/exclude the disease. If PA is confirmed, computerised tomography (CT) is used to image the adrenals and identify the disease subtype. If a unilateral adenoma is suspected, adrenal venous sampling should be used for diagnostic confirmation. Treatment with unilateral laparoscopic adrenalectomy or medication (mineralocorticoid receptor antagonists and/or other antihypertensive drugs) is provided depending on PA subtype. Antihypertensive drugs are often still needed in lower doses after a successful surgery since blood pressure is not always completely normalised (46, 58).

1.4.3.2 Cushing's Syndrome

Hypercortisolaemia or *Cushing's syndrome (CS)* (first described by *Harvey Cushing* in 1932) (59) is a multisystem endocrine disorder with an incidence of 0.7-2.4 per million patients per year (60, 61). Cushing's disease is the most common cause of endogenous CS (80-85%) and is defined by a prolonged hypersecretion of *cortisol* by the adrenal glands due to overproduction of *adrenocorticotropic hormone (ACTH)* from the anterior pituitary. CS can also result either from autonomous steroid secretion from an adrenal adenoma/carcinoma or from ectopic ACTH secretion from another source, or from treatment with exogenous steroids (exogenous/iatrogenic CS) (44, 62, 63).

The hypercortisolaemia that is caused by CS can dysregulate cardiovascular, immunological, metabolic, cognitive and bone functions. Clinical symptoms include weight gain/obesity, high blood pressure, impaired immunological function, poor memory/concentration, irritability or reddish face, excess hair growth and/or menstrual irregularity (in women), extra fat on the face (*'moon/rounded face'*) and/or around the neck (*'buffalo hump'*), and psychiatric disorders (e.g. anxiety, psychosis). Less common symptoms, like acne, muscle and/or bone weakness, feet/legs swelling, and diabetes mellitus, are also observed (44, 62-64).

For the diagnosis of suspected pituitary and adrenal CS, a thorough drug history should first be taken to exclude the possibility of iatrogenic disease. Diagnosis is normally initiated with the use of more than one of the following biochemical screening tests: a) 24-hour urinary free cortisol test; b) late-night (midnight) salivary cortisol test; c) overnight or two-day low-dose dexamethasone suppression test; and/or d) midnight serum cortisol test. If test results are equivocal and the patient has a high pre-test disease probability, confirmatory testing with one or more of the other tests and/or the dexamethasone/corticotropin-releasing hormone test is performed. If test results are positive, then a plasma ACTH test is provided to distinguish between patients with adrenal CS (low blood ACTH concentration) and patients with pituitary causes (normal/high ACTH levels). Inferior petrosal sinus sampling is also provided to differentiate pituitary from ectopic ACTH secretion (44, 62, 63).

If CS is found to be caused by an adrenal tumour, CT is performed to scan the tumour before laparoscopic adrenalectomy surgery. If CS is found to be caused by a pituitary tumour, magnetic resonance imaging is performed to visualise the gland/tumour. If the test is positive and the tumour is not extending into areas outside the pituitary, minimally invasive pituitary transsphenoidal surgery is conducted. Radiation therapy/radiosurgery is often administered if surgery is unsuccessful and hypercortisolaemia is persistent. Radiotherapy may also be used after bilateral adrenalectomy to reduce the patient's risk of developing *Nelson's syndrome* (i.e. pituitary tumour that releases high *corticotropin* levels resulting in pigmentation). Medical treatment is often required pre-operatively to render a patient fit for surgery. Glucocorticoid receptor antagonists and compounds that reduce ACTH release or inhibit the production of steroids (*'steroidogenesis'*) are also used in some cases (e.g. metyrapone, ketoconazole, mitotane) (44, 62, 63).

1.4.3.3 Addison's Disease

Primary adrenal insufficiency or *Addison's disease (AD)* (first described by *Thomas Addison*, 1849) is a chronic endocrine disorder with an annual prevalence of 50-100 cases per million people (65-67) that is characterised by a deficient production of glucocorticoids (cortisol) and mineralocorticoids (aldosterone) from the outer layer (cortex) of the adrenal glands. AD is a severe disease that can be potentially lethal if left untreated since steroid hormones play an essential role in the energy, salt and fluid stability/homeostasis of the body. AD is normally associated with non-disease-specific symptoms, such as weight loss, weakness, fatigue and abdominal pain/discomfort, while signs of the adrenal/Addisonian crisis (i.e. a life-threatening situation characterised by clinical manifestations, such as orthostatic hypotension, loss of consciousness and back pain) might also be present. Due to cortisol deficiency, AD is usually followed by ACTH hypersecretion from the pituitary gland, which often results in a characteristic hyperpigmentation (darkening) of the skin and mucosal surfaces (e.g. skin folds; lining of the cheek) (42, 68-70).

AD can affect people of any age and sex. Its causes are divided into three categories based on the mechanism that damages the adrenal glands: a) autoimmunity (i.e. the immune system attacks-destroys the body's cells/organs); b) infectious diseases (e.g. tuberculosis, fungal infections); and c) other causes (e.g. cancer, genetic defects, medication-related causes). Autoimmunity is currently the most common cause and occurs in the form of autoimmune adrenalitis (i.e. destruction of the adrenal cortex) or types 1/2 autoimmune polyglandular-polyendocrinopathy syndromes. These are associated with other autoimmune disorders, e.g. type 1 diabetes, primary gonadal failure and autoimmune thyroid disease (42, 68-70).

The diagnosis of AD relies on serum-plasma cortisol measurements and is highly likely to be present when low stimulated cortisol combined with elevated ACTH levels are observed. Diagnosis is confirmed by using the corticotropin (or ACTH, cosyntropin or short Synacthen) stimulation test. Additionally, simultaneous measurement of plasma renin-aldosterone is recommended to confirm mineralocorticoid deficiency. Diagnosis is recommended for all acutely ill patients, and patients with clinical signs and symptoms suggesting AD, while further investigation is needed after the diagnosis of the disease to identify the underlying cause. If

the patient presents with severe AD or adrenal crisis, immediate therapy with intravenous hydrocortisone and saline is recommended (42, 68-70).

AD is mainly treated with glucocorticoid therapy to replace the missing hormones. Cortisol (hydrocortisone) or cortisone are orally administered in several doses, while prednisolone and very rarely dexamethasone can also be given to patients with reduced compliance. Clinical assessment and monitoring should follow the glucocorticoid replacement regimen to adjust treatment based on patient needs. In patients with confirmed aldosterone deficiency, mineralocorticoid replacement therapy with fludrocortisone is provided. Again, therapy monitoring should be conducted to adjust treatment dose based on clinical response. Other therapies, such as dehydroepiandrosterone replacement, can also be used in specific cases (e.g. AD women with low libido; depression; low energy levels). In all cases, patients should be screened for associated autoimmune disorders (42, 68, 70).

Table 2: Summary information on endocrine disease

Endocrine Disease	Prevalence/Incidence	Key Signs and Symptoms	Diagnosis	Treatment
Primary aldosteronism (unilateral or bilateral)	5-15% of resistant hypertensive patients per year	• Aldosterone overproduction (not suppressed by sodium loading) • Excess potassium excretion (causes hypokalaemia) • Sodium retention (causes high blood pressure) • Muscular weakness	<u>Screening:</u> • Plasma aldosterone-renin ratio <u>Confirmatory testing:</u> • Saline infusion test • Oral sodium loading test • Fludrocortisone suppression test • Captopril challenge test • Furosemide upright test (Japan mainly) <u>Subtype classification:</u> • Adrenal CT • Adrenal venous sampling	• Unilateral laparoscopic adrenalectomy • MRAs for bilateral adrenal disease or when the patient is unwilling/unable to undergo surgery • Antihypertensive drugs (with some MRAs) after successful surgery if blood pressure is not completely normalised • Antihypertensive drugs (including MRAs) after unsuccessful surgery

Endocrine Disease	Prevalence/Incidence	Key Signs and Symptoms	Diagnosis	Treatment
Cushing's syndrome (pituitary or adrenal)	0.7-2.4 cases per million people per year	• Weight gain/obesity • Hypertension • Reddish face • 'Moon face' • 'Buffalo hump' • Muscle/bone weakness • Feet/legs swelling • Diabetes	<u>Diagnosis:</u> • 24-hour urinary free cortisol test • Late-night/midnight salivary cortisol test • Overnight or two-day low-dose dexamethasone suppression test • Midnight serum cortisol test <u>Confirmatory testing:</u> • One/more of the above tests • Dexamethasone/corticotropin-releasing hormone test • ACTH blood and/or stimulation test <u>Subtype classification:</u> • Inferior petrosal sinus sampling • Adrenal CT • Head magnetic resonance imaging	• Unilateral/bilateral laparoscopic adrenalectomy • Pituitary transsphenoidal surgery • Radiosurgery after unsuccessful pituitary surgery or persistent hypercortisolaemia • Medical treatment with glucocorticoid receptor antagonists and compounds that reduce ACTH release or inhibit steroid production before and/or after surgery or when the patient is unwilling/unable to receive surgery

Endocrine Disease	Prevalence/Incidence	Key Signs and Symptoms	Diagnosis	Treatment
Addison's disease (primary adrenal insufficiency)	50-100 cases per million people per year	• Weight loss • Weakness and fatigue • Abdominal discomfort • Adrenal crisis • Skin and mucus darkening	Diagnosis: • Serum-plasma cortisol immunoassay Confirmatory testing: • Plasma ACTH or short Synacthen stimulation test	• Orally administrated glucocorticoid therapy (hydrocortisone, cortisone, prednisolone, dexamethasone) • Mineralocorticoid replacement therapy for patients with aldosterone deficiency • Dehydroepiandrosterone replacement • Therapy monitoring and screening for associated autoimmune disorders

Abbreviations: ACTH: adrenocorticotropic hormone; CT: computerised tomography; MRAs: mineralocorticoid receptor antagonists* [Sources:** (42, 44, 46, 53-70)]

1.5 Medical Devices

1.5.1 General Information

Over the last few decades, a rapid introduction of many medical devices (MDs) in the market and clinical pathway of many diseases has been recorded. According to the EU guidelines (71-73), a *'medical device'* is defined as *"any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes:*

- *diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease,*
- *diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability,*
- *investigation, replacement or modification of the anatomy or of a physiological or pathological process or state,*
- *providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations,*

and which does not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means".

MDs are an integral part of today's clinical practice and patient care (74, 75). MDs include a wide range of heterogenous products and although, like any other health technology (e.g. drugs), they aim to restore or improve health, they are substantially different in their mode of action since they make use of a great variety of actions and reactions (e.g. mechanical, electrical), while they can be used for many clinical purposes. These two characteristics in combination with those presented in **Table 3** make MDs a special healthcare tool which is sometimes difficult to be appropriately evaluated in terms of its clinical and cost effectiveness (76-79).

Table 3: Special characteristics of medical devices

Characteristics	Explanation
Multiple applications	Difficult to estimate the overall value of a device that has multiple uses (in multiple diseases) since these cannot easily be divided. Therefore, some weighted average is needed, which is normally hard to calculate.
Use in diagnosis	Difficult to understand whether the impact on patient outcomes is caused by an improved and/or more accurate diagnosis, or the value of the subsequent treatment.
Device-operator interaction (*'learning curve'*)	The value of MDs depends on the interaction between the device (actual efficacy), the clinical procedure (intended use and/or condition), and the clinician/patient (operator). This means that improvements in the performance and outcomes of a device might be observed over time due to the end-user's acquired skills and experience (potential confounder in studies). The latter is difficult to be measured in experimental studies.
Incremental innovation	Over time, technological innovation takes place, leading to modifications/improvements in the performance, safety and outcomes of a device, and/or the development of new MDs, making older devices obsolete (short product life cycles). Therefore, it is unlikely to find a 'steady-state' period during which the device can be assessed in experimental studies. Consequently, the evaluation needs to be viewed as an iterative process.
Primary outcomes	There are a lot of variations in the primary outcome measures that are used when assessing/comparing devices.
Licensing process & market access approval	A pre-market approval application with sufficient evidence of the device's safety and effectiveness for the intended use is required. Normally, no long-term efficacy data are required.

Characteristics	Explanation
Manufacturer incentives & availability of evidence	There is little available evidence regarding the efficacy and performance of MDs. This is because manufacturers are more interested in placing their product first on the market and less interested in investing in research, especially since regulations are more relaxed compared to other health technologies (e.g. drugs).
Robustness of evidence	Evidence usually comes from small clinical trials and/or observational studies in specific groups of patients and not randomised controlled trials (blinding might be difficult, non-feasible and/or unethical due to the high risk, complexity and/or invasiveness that are related to a device), leading to selection bias and confounding factors.
Evidence comparisons	Lack of equivalent clinical evidence for all devices makes comparisons between them difficult, unless there is available evidence that differentiates the products. Nevertheless, different MDs might have different physical properties, even modes or actions. Additionally, there is a lack of head-to-head comparisons between different devices or between devices and other technologies.
Implementation & running costs	The use of a device can result in wider organisational and economic implications (e.g. need for user training; changes in the local infrastructure). Running costs also include maintenance and consumable costs.
Dynamic pricing	Prices are much more likely to change over time as more new products enter the market or because of how product procurement is conducted in different clinical settings (e.g. buying larger quantities in lower prices).
Statistical analysis	Due to the abovementioned characteristics, statistical analyses might be complicated.

Abbreviations:* MDs: medical devices* [Sources:** (74, 76-88)]

1.5.2 Medical Device Classification

The European guidelines (see *Section 1.6.1* below) categorise general MDs into four different classes (I, IIa, IIb, III) based on some general criteria: potential risks/vulnerability of the human body associated with their design, manufacture and use; intended purpose; invasiveness; duration of continuous contact with the body; tissue/location of administration; and active (i.e. rely on a source of energy) or non-active. Irrespective of the class, all devices must meet the essential requirements and be *CE marked*[2] before being marketed. Classes I-IIa contain the lower-risk devices and those that do not make changes to the human body, whereas classes IIb-III include the higher-risk devices. The device classification is not always straightforward, and must be considered and determined individually (71, 89).

Although the abovementioned classification is suitable for regulatory purposes, it cannot be used when assessing MDs in terms of their costs and effects. To do so, researchers from an EU-funded project (*Advance-HTA*) – which aimed to advance and strengthen the way that economic assessment is applied and implemented in all health technologies in Europe, North and Latin America – tried to create a taxonomic model that categorises MDs from the health economic viewpoint by combining different available classification schemes used for different regulatory and reporting purposes (**Figure 3**). The goal was to assist researchers and policy makers with understanding how to assess different device categories and which tools should be used given the device-specific elements (i.e. safety, efficacy/functionality) that are present each time (78).

Regarding the *U-Rhythm* device, although so far it has followed the IIa development pathway, considering that it is an active device intended to remove body liquids in a non-hazardous way (in terms of the samples' nature, part of the body concerned and mode of application), in 2020, UK authorities classified the device as a *'research use only'* (i.e. no intended medical purpose) in vitro diagnostic device since it is not itself invasive (indirect body contact) and is only intended for sample collection (90). As a research tool the device does not require a CE mark for its use, but this will be needed when *U-Rhythm* is used in future clinical practice.

[2] 'CE' stands for the French phrase *'Conformité Européenne'* (i.e. European Conformity) and means that the product conforms with the provisions of the relevant European legislation.

European device classification according to risk aspects		Classification according to the relevance of product and service, reimbursement characteristics, and assessment logic					
		Diagnostic			Therapeutic		
		A1 Assistive technology devices (directly used by patients)	**B1** Artificial body parts (implanted by medical procedure)	**C1** Medical devices for the assistance of medical professional	**A2** Assistive technology devices (directly used by patients)	**B2** Artificial body parts (implanted by medical procedure)	**C2** Medical devices for the assistance of medical professional
General	**I**	Manual blood pressure meter		Stethoscope	Conventional wheelchair		Spatula
	IIa	Pulse oximeter		Clinical thermometer	Contact lenses	Dental crown	Infusion cannula
	IIb			X-ray machine	Insulin pen	Bone prosthesis	Radio-therapy unit
	III			Intracardiac catheter	Condom with spermicide	Breast implant	Stent delivery catheter/system
In vitro diagnostic		Pregnancy test		Ebola virus antigen			

[**Source:** Image developed based on a figure presented in *Henschke et al. (2015) (78)*]

Figure 3: Taxonomy of medical devices in the context of economic assessment

**Description: The model indicates the importance of assessing medical devices based on their intended purpose, and risk (safety) and benefit for the patient; Red, Low importance; Yellow, Context-sensitive importance; Green, High importance; Grey, No medical devices identified.*

1.6 Approval Regulations for Medical Devices

To understand more about the data that should be collected from *ULTRADIAN*, the types of analysis that should be conducted during the clinical and (early) EE of diagnostic MDs, and the outcomes that are needed to convince health care policy makers about the cost-effectiveness of *U-Rhythm* when used in PA, it would be useful to provide a brief overview of the existing regulations for the approval of MDs and how their health technology assessment (see *Section 1.7* below) is currently performed in the EU and the UK.

1.6.1 European Regulations

EU law states several common regulations (*'Council Directives'*) that operate within each Member State but also gives the right to each country to have its own harmonised national measures based on its constitutional/legal requirements and system (71, 89). Although there are several Directives that regulate the safety and marketing approval of the different types of MDs (71, 89-91), this thesis focuses only on the main *Medical Devices Directive (MDD) (93/42/EEC) (71)*, which applied to *U-Rhythm* during its development, and includes the major processes and requirements for the marketing authorisation of all types of MDs.

1.6.1.1 Essential Requirements

The MDD (71, 89) establishes the rules on all the different aspects of a device which must be followed to allow its use within the internal EU market. In general, during its design, the potential risk of use error due to its ergonomic features and intended users (e.g. technical knowledge/experience) must be considered. During its construction, the right materials must be chosen (e.g. human body compatible), while adequate protection measures, labelling and instructions for the intended users must be included. During packaging, it must be ensured that the characteristics and performance of the device will not be affected and that the risk of contamination is limited. Lastly, during use, the device must provide sufficient accuracy and stability, and must have limited infection/contamination risk factors and undesirable side-effects for the patient and/or third parties.

1.6.1.2 Conformity Assessment

There are two ways that can be used, independently or in combination, to prove that a device complies with the essential requirements and secure an indication of conformity (CE marking). Firstly, a critical evaluation of the relevant scientific literature on the clinical, technical and performance characteristics, and the intended purpose of the device can be used. This presupposes that there are adequate data and/or equivalence between the new device and the device(s) to which the data relate (e.g. product used in clinical practice). Otherwise, a clinical investigation under normal conditions of use must be conducted after receiving approval from the national regulators (*'Competent Authorities'*) and sometimes agreement from a research ethics committee. A clinical investigation must demonstrate that the new device satisfies the essential criteria, give information on the performance and any undesirable side-effects or serious adverse events that are expected, and ensure that there are more benefits than harms when using it. In addition, the manufacturer must always prove that an adequate and efficacious product quality management system is kept and that good manufacturing practices are followed, while a technical documentation with all product information must be stored for future reference (71, 89, 92-94).

A Member State should allow a device to be placed on its market if this has been approved by another EU country and CE marked but can also not approve or temporarily withdraw any product that is suspected to be dangerous for the health and/or safety of patients. The country must then inform the Commission of any measures taken to confront these incidents and the latter must investigate whether the decision is justified. Each country can designate private independent non-governmental organisations (*'Notified Bodies'*) to be involved in the conformity review process and certification and examine/ensure that any device in the market meets the essential criteria. These parties should have the appropriate scientific and technical staff; not be involved in the design, manufacture or marketing of the devices; and have no interests, especially financial, when making their decisions, while they must be closely monitored by the State (71, 89).

1.6.1.3 New Regulations

In 2017, the EU Commission published a new regulation *(2017/745) (73)* for the approval and post-marketing surveillance of MDs. This regulation will fully come into force in spring 2021. Several important amendments made include stricter rules for the clinical evidence that is required; tighter controls on high-risk devices; reinforcement of the criteria when designating and overseeing processes conducted by Notified Bodies; improved transparency of information for consumers through a wider use of *EUDAMED* (central secure European database that contains all the important for the device data/files); strengthening of post-market surveillance requirements for manufacturers; and improved and closer coordination between States in the fields of vigilance and market surveillance. The reasons for these changes are to improve the quality, safety and reliability of MDs, and reduce the problems with diverging interpretation of the current Directive.

1.6.2 UK Regulations

1.6.2.1 Related Historical Events

Before describing the UK regulations for the approval of MDs, it would be useful to provide a summary of several important historical events related to them (**Table 4**).

Table 4: Historical events related to the UK regulations for medical devices

Date	Event
Late 1960s	The *Scientific and Technical Branch* is created. This together with the *Department of Health and Social Care* – then *Department of Health and Social Security* – are responsible for improving and ensuring the quality and safety of medical equipment
1980s	The Scientific and Technical Branch becomes part of the *NHS Procurement Directorate*; later divides into the *NHS Supplies Authority* and the *Medical Devices Directorate*

Date	Event
1989	The *Medicines Control Agency* is established for the control and licensing of medicines based on the Medicines Act of 1968
Mid 1990s	The formal European regulation of medical devices starts
1994	The Medical Devices Directorate becomes the *Medical Devices Agency*
2002	The *Medical Devices Regulations (Statutory Instrument 2002/618)* with guidelines for the approval of medical devices comes into effect
2003	The Medicines Control Agency is merged with the Medical Devices Agency to form the current *Medicines and Healthcare products Regulatory Agency*

Abbreviations:* NHS: National Health Service* [Source:** (95)]

1.6.2.2 MHRA

The Medicines and Healthcare products Regulatory Agency (MHRA) is the UK Competent Authority responsible for regulating, monitoring and ensuring the quality, safety and efficacy of all medicines, MDs and medical equipment, and for providing marketing authorisation and licensing when the products meet the essential requirements. If the MHRA considers that a product may put patients' health at risk, it is allowed to take the necessary measures, such as withdraw the product from the market, suspend its production, issue warnings or even prosecute its manufacturer/authorised representative if the regulations have been breached (95). The MHRA is a governmental body, and therefore it is mainly funded by the Department of Health and Social Care (DHSC) but also from the fees that manufacturers pay to get approval for their products. It is recognised as an independent body of expertise for the evaluation of medical products and collaborates with the *European Medicines Agency*, influencing the EU regulation, but also with international agencies, e.g. the US *Food and Drug Administration* (95).

1.6.2.3 Medical Device Approval Process

The Medical Devices Regulations (MDR) of 2002 (Statutory Instrument 2002 No. 618) constitute the UK national device-specific legislation, which, respecting the UK Consumer Protection Act 1987, implements the provisions of the MDD. The MDR inform on the classification and essential requirements for MDs, the obligations of the manufacturer or authorised representative, the functions of the MHRA and UK Notified Bodies, and any fees that are charged by the Secretary of State for Health and Social Care (SSHSC) (93, 94, 96). As mentioned above, every MD that is traded within the EU market must bear a CE marking. Since most device categories need a Notified Body before the CE marking can be affixed, the MHRA is responsible for certifying these bodies in the UK and ensuring that they work properly and impartially. For the CE marking to be affixed, the device must conform to all the Directives that apply to it and the manufacturer must provide the necessary clinical data to the MHRA. If a clinical investigation is planned to be conducted within the UK, an approval from the MHRA on behalf of the SSHSC, the Research Ethics Committee and the management of the NHS organisations that will be involved are required (93-96).

1.7 Health Technology Assessment of Medical Devices

1.7.1 Definition and Challenges

Although the EU and UK guidelines state the essential criteria regarding the quality, safety and efficacy of MDs, they make no reference to the need of measuring their cost-effectiveness before and/or after they are launched (78). In most European countries with a publicly funded healthcare system, the cost-effectiveness of many existing and emerging healthcare programmes (e.g. drugs) is evaluated using a procedure called *'health technology assessment (HTA)'*. HTA is a systematic and multidisciplinary process that identifies, reviews, assesses and appraises the medical, economic, organisational, political, social and ethical evidence related to the use of a healthcare product/service to assess the added clinical and economic value compared to existing care. HTA is conducted by national authorities and/or independent agencies (*'HTA bodies'*) to fully inform policy makers and help decisions on which health technologies should be reimbursed (76, 84, 85, 97-99). HTA is a much broader process

than EE. However, the two terms are often used interchangeably, as in this thesis, since cost-effectiveness plays a key role in coverage and reimbursement decisions (99).

Many HTA bodies consider HTA as a one-off process at a late stage of a product's development or sometimes even after it is marketed (29). This can lead to a situation summarised by *Buxton's Law (1987) (100)* which states that *"It's always too early until, unfortunately, it's suddenly too late!"*. The dilemma outlined by Buxton is that once a technology is in regular use, clinicians, patients and ethics committees may consider it unethical to randomise patients (to get the technology or not) – thereby, severely limiting the opportunities to collect unbiased evidence on the costs and benefits. Although Buxton's view was expressed about surgical procedures, it can be applied to every health technology. Since more clinical and economic data become available as a technology develops and/or is used in clinical practice, it is essential to understand that HTA should be an iterative process. This will give more reassurance to policy makers that they have made the right decision about an intervention or inform them whether they need to change it to use resources more efficiently (29, 101).

In HTA, evidence synthesis is challenging given the difficulty to identify relevant, high-quality, unbiased clinical and economic evidence. Data may be available from various imperfect study designs, studies with narrow research questions, inadequate or inconsistent clinical registries, and/or expert opinions. When data are not available from head-to-head comparison studies, indirect comparisons are usually made, which do not always check whether the patient characteristics are similar. Data might also not be relevant to the eventual real-world use, particularly in early HTAs where eventual use is still to be established. For example, evidence might come from highly selected patients and healthy controls or different clinical or organisational settings, which means that they do not always represent clinical practice, leading to incorrect decisions (102).

Moreover, it is not always clear which costs (e.g. direct costs) and/or units of effectiveness (e.g. life-years) should be used. In addition, decisions driven by only considering the product's cost-effectiveness only partially consider the benefits to the patients, healthcare system and society (102). These problems in combination with the special characteristics of MDs have resulted in an existing difficulty in conducting HTAs for these products in Europe, and

consequently the need for a consensus on how this should be done. Therefore, until now, HTA bodies have been assessing devices using similar HTA approaches to those of other health technologies, especially pharmaceuticals, since their evaluation is more straightforward, ignoring the device-specific issues and the complex way that they affect patient outcomes (76, 80, 81, 85).

1.7.1.1 MedTecHTA

To improve the existing HTA methodological framework for the assessment of MDs as well as the decisions concerning their cost-effectiveness, their appropriate use and patients' access, an EU-funded project (*MedTecHTA, http://www.medtechta.eu/*) took place between 2013-2016. This project aimed to analyse current cross-country HTA guidelines; the geographical variation in the access/use of MDs; the methods used for the evaluation of their comparative effectiveness; the methods used and the uncertainty in their EE; and their organisational impact (82). The results indicated that different regulators have different requirements that a device should meet before getting pre- and/or post-market approval/surveillance. Despite this, in all countries, device regulations are less demanding than those of other health technologies. Additionally, in all guidelines, the clinical evidence that is needed depends on the risk-category of the device, meaning that there are cases (e.g. lower-risk devices) where evidence is limited. The lack of clinical evidence causes difficulties when conducting HTAs since it leads to delays in funding and making technologies accessible to patients. Furthermore, regional bodies have developed specific organisational procedures for the evaluation of MDs, making it difficult for manufacturers to simultaneously prepare and submit their reports in different countries. For this reason, it is important to choose the right study design when evaluating the safety and efficacy of different devices and have an international harmonisation of regulatory requirements. Pre- and post-marketing surveillance is needed to ensure that devices continue being effective even when they are used in clinical practice. Collection and dissemination of real-word data (e.g. observational studies, registries) should also be encouraged to inform healthcare workers and patients regarding any adverse events and improve the learning curve and wider organisational impact of MDs. Lastly, there is a need for a better linkage between device approval regulations and HTA processes since the latter can secure that the device will be reimbursed (84, 85).

1.7.2 European Guidelines

The absence of a harmonised European guideline for HTAs means that each Member State can apply its own rules (77). To address this, the EU Commission introduced a European collaboration body, called *HTA Network*, through the Directive 2011/24/EU (article 15) (103), in which EU national authorities and/or regional HTA bodies can voluntarily participate. This is implemented by the scientific and technical component of the Network, called the *European Network for Health Technology Assessment (EUnetHTA)*. EUnetHTA aims to facilitate the efficient and optimised use of the available resources; create sustainable systems of HTA transparent knowledge and experience sharing; avoid duplication in HTA activities; and promote good practices in the methods and processes used for HTAs (*HTA Core Model®*) (97, 98, 104-106). In 2015, EUnetHTA produced a guideline of the most appropriate methods for conducting EE. This handbook contains information on how health economic evaluations are conducted by EUnetHTA partners and how they should be conducted by making recommendations on the type of analysis that should be performed, the comparators and time horizon that should be included, the way that costs and effects should be measured, and the different analytical methods that can be used (104).

1.7.3 UK Guidelines

1.7.3.1 UK HTA Bodies

In England, the main HTA agency is the *National Institute for Health and Care Excellence (NICE, https://www.nice.org.uk/)*, which was established in 1999 by the DHSC to evaluate existing and emerging health technologies, and recommend whether they should be funded by the NHS and how they should be used in clinical practice. NICE works independently from the NHS and its remit has expanded over time to provide national guidance, advice, quality standards and information services, mainly in the form of health technology appraisals and clinical guidelines, that aim to improve health and standardise patient access to health, public health and social care (107, 108). Other parts of the UK have different HTA bodies. More precisely, in Scotland and Wales, similar bodies for medicines are the *Scottish Medicines Consortium (https://www.scottishmedicines.org.uk/)* and the *All Wales Medicines Strategy Group*

(https://awmsg.nhs.wales/), respectively. Since this PhD thesis was produced in England, analysis was conducted in accordance with NICE's guidelines and recommendations.

1.7.3.2 NICE HTA Programmes for Diagnostic Devices

1.7.3.2.1 Technology Appraisal Programme

The *Technology Appraisal Programme (TAP)* is one of the earliest HTA programmes that was established by NICE for the evaluation of all types of health technologies, including medicines and treatments, medical devices, and diagnostics. TAP currently contains three forms of processes. First, the *single technology appraisal*, which is used for single technologies with a single indication. Second, the *fast track appraisal*, which is a faster and less resource-intensive process for single technologies with single indications that are also very cost-effective. Third, the *multiple technology appraisal*, which is used when more than one technology is being assessed or for a single technology with more than one indication. These processes typically cover new products or new licensed indications. Technologies must have been granted or be soon to receive marketing authorisation. NICE produces HTA guidance after reviewing the available clinical and economic evidence. Product developers or representatives first submit the principal evidence (published and unpublished), which is reviewed by an independent external academic centre (*Evidence Review Group, ERG*). NICE also invites other consultees (e.g. healthcare professional bodies; clinical commissioning groups; patients and carers) to participate. Consultees comment on the scope of the appraisal (i.e. disease; targeted patients; technologies that will be covered), can submit additional evidence and can consult on the generated guidance. NICE's standing advisory committee (*Technology Appraisal Committee*), which includes NHS workers, lay members, relevant academics, and representatives of the pharmaceutical or MD industries, reviews all evidence, consultees' comments and the ERG's assessment report to decide whether the technology should be adopted and if yes, for which groups of patients. The Committee provides its recommendations initially in an appraisal consultation document (if the technology is not recommended or is recommended with a limited use; or further clarification/evidence is needed), and then a final appraisal document (the basis of NICE's guidance). Here, it should be noted that NHS commissioners are legally obliged to fund and promptly provide any medicines and treatments recommended by TAP (109).

1.7.3.2.2 Medical Technologies Evaluation Programme

The UK is one of the few countries worldwide (e.g. Brazil, France) in which an HTA handbook for devices has been developed (81). In 2009, NICE established the *Medical Technology Evaluation Programme (MTEP)* aiming to produce national guidance for the assessment of novel medical technologies (110), including MDs and diagnostics (87, 111, 112). The MTEP process normally starts when a manufacturer/supplier submits a notification form regarding an approved or soon to be approved (e.g. CE marked) device (new or innovative modification). This submission must contain published and unpublished clinical and economic evidence, state the decision problem, and indicate how the product can be used in current practice and how it performs compared to existing alternatives (111-113). The MTEP team rates the extent to which the technology meets the programme's selection criteria: a) benefit to patients; b) benefit to the health and social care system; c) targeted patient population; d) disease impact on quality of life and life expectancy; e) cost considerations; and f) technology sustainability (87, 112). Based on these criteria, the MTEP team decides whether a *Medical Technologies Guidance*, a guideline from another NICE programme or a Medtech innovation briefing[3] should be developed. If an MTEP guidance is selected, the submitted evidence is reviewed by an external assessment centre, which prepares an assessment report. The MTEP team may also ask expert advisers, and patient and carer organisations for contributions. Afterwards, an independent committee (*Medical Technologies Advisory Committee, MTAC*) examines all the evidence and prepares a consultation document with its provisional recommendations. Registered stakeholders, health professionals and members of the public can comment, and after considering their comments, the MTAC makes its final recommendations. Here, it should be highlighted that any MTEP recommendation is currently in the form of guidance and is not mandatory (87, 110, 112). This is because the use and cost-effectiveness of such technologies may vary depending on the targeted patients and current practice. Therefore, adoption decisions are better to be taken by clinicians at a local level (114).

[3] Information on MDs and diagnostics produced after quickly reviewing the relevant published evidence and the likely costs of using these technologies. The aim is to accelerate innovation and help local decision-making (87).

1.7.3.2.3 Diagnostics Assessment Programme

In 2011, NICE created the *Diagnostics Assessment Programme (DAP)* to assess and promote the rapid and consistent adoption of clinically and cost-effective diagnostics. DAP focuses on physiological measurements; laboratory, pathological and imaging tests; endoscopy; decision rules and algorithms; questionnaires, structured interviews and surveys; or test combinations that involve the use of instruments/devices and can be used for any diagnosis-related purpose (i.e. diagnosis, monitoring, screening, prognosis). The technologies under evaluation should have previously received marketing approval and are selected by the MTAC to be routed to DAP due to the complexity of their evaluation (e.g. consideration of multiple technologies or indications). Companion diagnostic technologies, whose main purpose is to identify patients who would benefit from a new medicine, are normally appraised in conjunction with the drug under TAP. In DAP, recommendations are formed by an independent advisory committee (*Diagnostics Advisory Committee, DAC*) that consists of professional and lay specialists in the area. External expert input (*External Assessment Group*) is also required for examining the evidence and preparing the diagnostics assessment report. The latter together with any other available information are then reviewed by the DAC. The DAC prepares the diagnostics consultation document with its draft recommendations and asks registered stakeholders, health professionals and members of the public to comment. Based on their comments, the DAC publishes the final diagnostics guidance documents with its final recommendations. As with MTEP, DAP recommendations are currently used as guidance and are not mandatory (115, 116).

1.7.3.3 NICE Decision-Making Criteria for Diagnostic Devices

MDs and diagnostic technologies can be assessed under either TAP, MTEP or DAP, depending on the nature of the technology and its value proposition. These technologies are normally routed to TAP or DAP when they are compared to a reference case and are associated with higher costs and health benefits. MTEP does not require a reference case and it assesses MDs that are cost saving with the same clinical benefit or cost neutral with a greater clinical benefit than current practice. In all programmes, the scope of the evaluation (i.e. patient population; comparators; care pathway; costs and outcomes; time horizon; other special considerations) is first defined. Costs are measured using official national sources (e.g. NHS reference costs) and based on them, the NHS budget impact of the technology is determined. In TAP and DAP, health benefits are expressed in *quality-adjusted life-years (QALYs)* (see *Section 2.3.4.4.1* below), preferably measured using the EuroQol EQ-5D health questionnaire, while in MTEP, several outcome measures may be considered (e.g. clinical benefit, resource consequences). To demonstrate cost-effectiveness in TAP and DAP, NICE uses a £20,000-£30,000 per QALY willingness-to-pay threshold, which means that if the evaluated technology costs £20,000-£30,000 more than the existing alternative, it should provide at least one more QALY to be adopted. The innovation of the technology; the nature and quality of evidence; uncertainty; differential benefits or side/adverse outcomes; the technology's position in the care pathway; and scientific and social value judgements are additional factors that are taken into account. In DAP, test accuracy also plays a key role in decision-making. In MTEP, a willingness-to-pay threshold does not exist, and decision-making is based on whether the new technology provides substantial benefit to the patient, and the health and social care system when compared to current practice. Lastly, in all programmes, maximising the population's health and equality in the distribution of the available resources have a strong impact on the decisions made (87, 115, 117).

1.8 PhD Project Objectives

As shown in **this Chapter**, diagnosis can be challenging, especially for endocrine disorders where dynamic testing is likely to be more accurate than standard static testing. Evaluation of diagnostic tests requires assessment of the whole pathway (including diagnostic impact, changes in therapy, costs and patient outcomes), making it a complex procedure. The current regulations governing the use/marketing of MDs and the HTA process for assessing the cost-effectiveness of diagnostic devices, in particular, varies between EU counties and is less well-established than equivalent processes for other health technologies (e.g. drugs). Early EE might have a role for device manufacturers (in stop/go R&D decisions) and for healthcare regulators (in early adoption decisions), but this role is ill-defined. The overall aim of this PhD thesis is to apply health economic methods to summarise current knowledge on the diagnosis and monitoring of the six abovementioned endocrine disorders; measure satisfaction with the *U-Rhythm* device, resource usage costs and quality of life in patients with PA, CS and AD; and examine the potential clinical and economic value of *U-Rhythm* in PA. This will lead to a wider discussion of the role of economic methods in early diagnostic MD innovation, adding further evidence in this under-researched area.

To do so, several analyses are performed:

1. An exploratory literature review to give a brief overview of the methods that are currently used for the clinical and early economic assessment of diagnostic technologies, especially of MDs (**Chapter II**).

2. A systematic literature review (**Chapter III**) to retrieve and appraise the quality of the most recent economic evidence on optimal diagnosis and monitoring in the six endocrine disorders to inform the evaluation of a decision-analytic model structure for *U-Rhythm* for PA later (**Chapter VI**).

3. A descriptive analysis to compare the participant data collected from *ULTRADIAN*. Data were based on three questionnaires asking about the person's satisfaction after using *U-Rhythm*, their healthcare resource usage and impact of their condition on their work

productivity in the three months before sampling, and their health-related quality of life (**Chapter IV**).

4. A detailed costing study using UK healthcare data to inform on the patients' demographic, socioeconomic and clinical characteristics, and the healthcare resources that are used for the diagnosis and treatment of PA. These data are then used to measure the total cost of care that is associated with the disease and define the parameters that drive this cost for the NHS (**Chapter V**).

5. A brief review of the available literature, including evidence derived in **Chapters IV-V** to obtain data on the parameters for a decision model for the diagnosis and treatment of PA. A probabilistic decision analysis using modelling techniques to estimate the long-term costs and health benefits of the current diagnostic and treatment pathways compared to a new course of action that uses *U-Rhythm* for PA. A value of information analysis is conducted to address uncertainty issues in the decision that is taken (**Chapter VI**).

6. A detailed discussion on the applicability and suitability of early EE methods when used for diagnostic devices (**Chapter VII**).

CHAPTER II

HEALTH TECHNOLOGY ASSESSMENT

CLINICAL AND ECONOMIC EVALUATION OF HEALTH TECHNOLOGIES

CHAPTER II OVERVIEW

Chapter I gave an overview of this PhD thesis by describing the rationale and objectives of the research that is conducted; the *ULTRADIAN* study and the *U-Rhythm* device; the importance of clinical and early economic evaluation of diagnostic technologies and medical devices for manufacturers, policy makers, healthcare providers and patients; the regulatory framework under which the economic evaluation of (diagnostic) medical devices must be conducted; and the objectives of the reviews and analyses that are presented in the **next Chapters (III-VI)**.

Chapter II describes the elements that should be considered in a clinical and early economic evaluation of diagnostic medical devices. It starts by giving some general information on the different types of medical tests that exist based on their clinical purpose and method that they use; the way that the clinical assessment of diagnostics should be conducted; the most commonly used diagnostic accuracy measures for characterising test performance; and the study designs that are preferred in diagnostic evaluation studies. Afterwards, the concept of economic evaluation is defined by presenting the decision perspectives on which it can be based, the types of analysis that can be performed and the ways that it can be conducted. A greater focus on the economic evaluation of diagnostic technologies is put by presenting its methodological challenges. Later, the analytical methods that have been proposed or used in the early economic evaluation of health technologies, especially of diagnostic tests, are described to inform on how and at which stage of product development they are most often used. At the end, the applicability and suitability of early economic evaluation methods for *U-Rhythm* are examined, and the decision on the methods that are used in this thesis is justified.

Clinical and Economic Evaluation of Health Technologies

2.1 Assessment Rationale

Conducting a comprehensive clinical and economic evaluation of diagnostic technologies before and after they are marketed is extremely significant for the manufacturer, the policy maker, the payer, the doctor and the patient. During these processes, it should be determined whether the new product is safe and acceptable for patients; is accurate, and leads to prompt and effective treatment; and is affordable for the healthcare system or payer (12, 118-120). Establishing cost-effectiveness often requires study designs in which participants receive the new test, the existing comparator test(s) or all tests, with different designs having different advantages and disadvantages (12). A variety of diagnostic accuracy measures also exists (121-124), while economic evaluation (EE) can be conducted in various ways depending on how effectiveness is measured (17). Therefore, the choice of the most appropriate study design, diagnostic accuracy and effectiveness measure, and EE type are crucial when measuring cost-effectiveness in diagnostics and should reflect the research question.

2.2 Clinical Evaluation of Diagnostic Technologies

The first step in the evaluation of any health technology, including diagnostics, is to examine whether it is clinically effective. In diagnostic technologies, clinical effectiveness is determined based on the test's diagnostic accuracy, impact on the treatment that follows, and effect on the patient's health status (12, 16, 26). To describe these three characteristics, some general information on the different types of diagnostic tests, diagnostic accuracy measures and study designs that can be used is provided.

2.2.1 Medical Tests

A diagnostic test, or better, a *medical test* is a clinical procedure that is performed to screen/detect, diagnose, classify/stage and/or monitor one or more diseases (16, 121, 124, 125). Tests can be categorised based on their clinical purpose or the method that they use.

2.2.1.1 Types of Medical Tests by Purpose

Table 5 presents the different types of tests based on their role in clinical practice. Knowing the test's clinical role is important when evaluating it. Obviously, one test can have one or multiple clinical uses in different diseases (125).

Table 5: Types of medical tests based on clinical purpose

Role	Definition
Screening	To detect or predict the presence of a disease early on in asymptomatic patients
Diagnosis	To confirm/exclude the presence of a disease in an individual suspected of having the disease (e.g. after signs and symptoms)
Triage	To determine whether there is a need for further testing
Monitoring	Periodic testing in patients who have chronic diseases or have received treatment to monitor the progress of the disease or the patient's response to treatment
Prognosis	To predict the likelihood of developing a disease or disease progression

2.2.1.2 Types of Medical Tests by Method

In addition to the division of tests into categories based on their clinical role, they can also be classified into one of the following method groups (**Figure 4**) (125):

- Reported or recorded patient observations (i.e. signs and symptoms);
- Questions regarding the patient's or patient's family medical history;
- Tests performed during physical examination (e.g. palpation, auscultation);
- Tests where a sample of tissue or bodily fluid is analysed in the laboratory;
- Radiologic or imaging tests used to visualise a body target (e.g. organ, tissue);
- Tests where an agent or equipment is inserted in the body (i.e. invasively);
- Tests that focus on the physiological function of a body target (e.g. organ, biomarker).

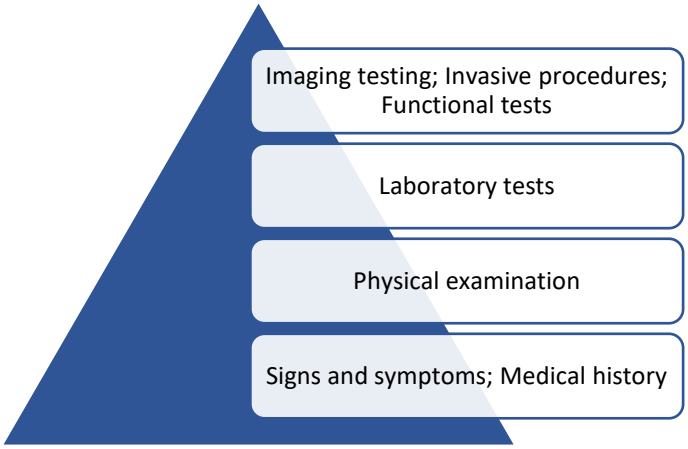


Figure 4: Hierarchy of medical tests based on the method used

2.2.1.3 Comparing Medical Tests

In today's clinical practice, although there are many diagnostic tests and procedures used for most diseases, there is a constant need for new ones since (119, 125):

- New types/subtypes of diseases are discovered or appear;
- A test that works for one patient might not work for another (*'personalised medicine'*);
- There is a need to reduce the risk to patient from existing tests;
- Need for less invasive, painful and/or uncomfortable tests for the patient;
- Need for faster diagnosis (i.e. quicker tests, easier to perform and/or interpret);
- Need for less expensive tests or tests that involve less healthcare resources.

Nevertheless, due to the rapid pace of new entrants in the market, there is also some concern about *over-testing* and *overdiagnosis* (i.e. identification of signs and symptoms that are not harmful for the individual) associated with new and more testing, which can cause substantial costs for the healthcare system or payer if not addressed appropriately (126). Therefore, it is important to understand whether the examined (*'index'*) test is really needed (i.e. its potential place in current practice compared to existing alternatives) (**Figure 5**). An index test can be provided before the existing test(s) to decide whether further investigation is needed (*'triage test'*). Triage tests are normally non-invasive, simple to perform and cheap, and are used to avoid the additional costs and/or discomfort of further testing. Nevertheless, they are less accurate than existing tests, meaning that they cannot adequately replace them. An index test can be used instead of existing alternatives (*'replacement test'*) if it provides additional benefits (e.g. more accurate and faster diagnosis). Lastly, an index test can be used as an *'add-on'* if this is provided after the testing pathway. The benefit of these tests is that they can work as additional confirmation that the disease is present/absent and target therapy (16, 119, 127).

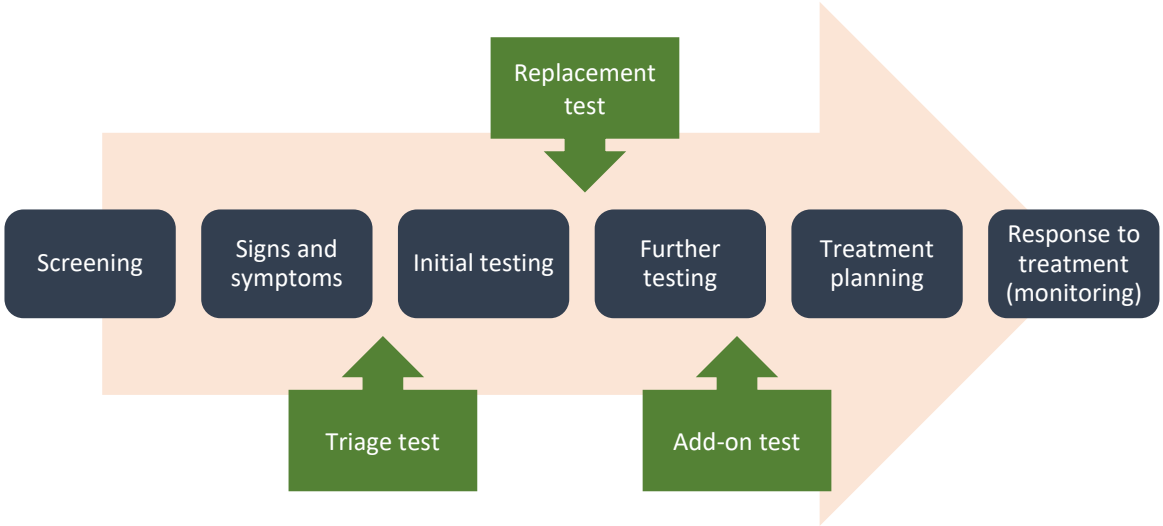


Figure 5: Roles of tests and their position in clinical pathway

2.2.2 Clinical Assessment Procedure

A properly conducted clinical evaluation is essential for helping clinicians select the most appropriate medical test for each condition, correctly interpret its results and choose the right management plan that will improve patient outcomes (15, 119, 121). For this reason, at the beginning of a test's development, its *technical performance* (i.e. technological requirements; quality; safety; provision of usable information) should be checked in the laboratory, without the use of human participants (12, 15). Moreover, the clinical area(s) in which the technology could potentially be used should be decided (128), while its *analytical sensitivity/specificity* (i.e. ability to indicate the presence/absence of disease) and *reproducibility* (i.e. ability to produce same results every time it is used) should be examined (16).

After a new technology passes all initial checks, it should be evaluated in the clinical setting, under a variety of conditions/circumstances, using human participants. Here, all the different clinical conditions in which the test could potentially be used as well as its result reliability (i.e. stability/consistency), validity (i.e. accuracy) and reproducibility should be examined. This is essential for manufacturers to decide whether they should invest in the technology (12, 16, 121). Additionally, all information on patient characteristics and their underlying conditions; contraindications to the test's use (*'harm'*); technical failures (*'feasibility'*); and difficulties in interpretation (*'interpretability'*) should be collected. This would help to examine where in clinical practice the technology can be more beneficial, and whether it could potentially increase/decrease the time to diagnosis and the need for further investigations, affecting the *impact on diagnostic thinking* (i.e. confidence on test results), *patient management* and *clinical outcome* when using the test (12, 16, 128).

Once the range of possible uses has been demonstrated, it is important to estimate the *diagnostic performance/accuracy* of the new intervention for the disease(s) in question (12, 15, 16, 120). To do so, representative groups of individuals (i.e. non-diseased; diseased – with different levels of severity and/or conditions that have overlapping signs and symptoms) should be used. Diagnostic accuracy is a joint function of the test's performance and the observer's interpretation (15, 120). Diagnosis is rarely 100% certain based on test results, so a test is only capable of providing supportive evidence to confirm/exclude a disease (120,

121). In some cases, measuring whether the benefits from treating patients outweigh the harms of not treating them and comparing the probability that the disease is present to a well-established cut-off point is the only aim of a test (121). Since a *reference standard* (i.e. best available diagnostic test) is normally required to examine the diagnostic accuracy of an index test, investigators should objectively check the data from all comparators while being unaware of the results of the other test(s) to avoid biased decisions. In cases where a reference standard is not available, establishing diagnostic accuracy is difficult since investigators should compare the results of the index test to a consensus diagnosis made based on the combined results of a number of other tests and often long-term clinical follow-up, and then examine the degree to which the test measures what it claims to measure (i.e. *construct validity*) (12, 120, 121).

Another parameter that should be examined is the impact of the test on healthcare providers. In clinical practice, physicians would prefer a test that reassures them that the disease is present/absent without the need for further investigation. This means that sometimes procedures are used without evidence of being totally accurate and/or effective (12, 15, 16, 128). The doctor's emotional, cognitive, social and/or behavioural perspectives can also modify the test-treatment pathway (15, 128). These are parameters that are generally difficult to measure but should be considered when launching a new technology and furtherly explored in the future (12, 128). Another complexity with diagnostics is that their clinical effectiveness mainly depends on the efficacy of the subsequent treatment (26, 128). In practice, a test might have high diagnostic accuracy but not be used because: a) clinicians cannot correctly interpret its results; b) physicians are unaware/unfamiliar with the available treatment; c) the patient is unwilling to receive treatment (e.g. surgery); d) there is no available therapy; or e) the patient is already receiving the best treatment (12, 15, 16, 120). Therefore, it is wrong to consider a test's diagnostic accuracy as the sole measure of its clinical effectiveness. On the contrary, the whole test-treatment pathway (**Table 6**) should be identified, described and assessed (125, 128, 129).

Although the impact of a test on the treatment is essential, one more clinical element that should be demonstrated before a new test is widely disseminated is its impact on patient outcomes (*'clinical utility'*). This is because a new intervention might lead more patients to

treatment and in a shorter period of time, but if the treatment remains the same, there is no difference in the patient outcome from before (12, 15). Health outcomes can also be affected by the patient's perspective, an unpredictable factor that can enhance/deteriorate the results of a test-treatment pathway. The way that the patient perceives the testing procedure, the previous experience with the testing process and/or the understanding of test results can all influence the patient's health status directly or indirectly (e.g. psychologically). In a similar way, diagnostic placebo effects can occur when the patient considers a test-treatment as clinically effective, improving patient outcomes (128).

Table 6: List of test-treatment pathway attributes that affect patient health

Pathway Component & Mechanism	**Definition**
Step 1: Diagnostic test provision	
Timing	How fast/soon the test is performed/provided within the management strategy.
Feasibility	Ability of the diagnostic procedure to be completed. (Reasons for non-completion: patient refusal, test contraindications and/or technical failure).
Process	Interaction between the patient and the test. (The test can cause physical or psychological harm or benefit).
Step 2: Test result	
Interpretability	Degree to which test results can inform diagnosis and disease classification.
Accuracy	Ability to distinguish between patients with and without the disease, or patients with a mild or severe form of the disease.
Timing	How fast/soon test results are made available.
Step 3: Diagnosis	
Timing	How fast/soon the diagnostic decision is made.

Pathway Component & Mechanism	Definition
Diagnostic yield	Degree to which the test contributes to diagnosis (any form), i.e. definitive diagnosis, confirmatory diagnosis, exclusion of a working diagnosis, and/or distinction between alternative diagnoses with different treatment implications. Diagnostic yield is different from accuracy because it includes all information and factors used by the doctor to make a diagnosis (e.g. previous test results).
Diagnostic confidence	Degree of confidence/reassurance that the test gives to the doctors and patients regarding its validity and/or applicability.
Step 4: Management decision	
Therapeutic yield	Degree to which the decision on diagnosis affects the therapy that is chosen to be followed.
Therapeutic confidence	Confidence with which the doctors and patients follow a specific treatment plan.
Step 5: Treatment administration	
Timing	How fast/soon the treatment is provided to patients.
Efficacy	Ability of the treatment to improve patient outcomes and health status.
Adherence	Extent to which patients comply/adhere to the treatment/management plan that is suggested by their doctor to achieve therapeutic goals.

[**Source:** Table developed based on a table presented in *Ferrante di Ruffano et al. (2012) (128)*]

2.2.3 Diagnostic Accuracy Measures

There are several measures of diagnostic accuracy that can be used (**Table 7**) and the selection of the most appropriate depends on the aspects of the diagnostic procedure that are assessed each time (e.g. disease discrimination, disease prediction). It should be highlighted that all measures are sensitive to the characteristics of the population that is examined (i.e. disease prevalence, spectrum of the disease), meaning that it is not only crucial to know how to interpret them but also under which circumstances to implement them (122). Diagnostic accuracy is often measured using the test's *sensitivity* (i.e. proportion of individuals with a disease who are correctly identified as sick) and *specificity* (i.e. proportion of individuals without the disease who are correctly identified as such) (12, 16, 121-124). Neither of these measures is affected by the disease prevalence, meaning that their results are transferable between different settings. However, both can vary depending on the patient characteristics and the disease spectrum (122-124). Since most tests are 'imperfect' (i.e. sensitivity and/or specificity are <100%), there is often a trade-off between the two measures. For this reason, the *receiver-operating characteristic (ROC)* curve is constructed (12, 26) by plotting paired values of '1-specificity' (x-axis) and 'sensitivity' (y-axis). The area under the curve (AUC_{ROC}) is then calculated to estimate the probability that a randomly selected individual with the disease has a higher test score compared to one without the disease. The larger the AUC_{ROC} is, the better the test's discriminatory ability (122, 123).

Instead of a test's sensitivity/specificity, its *positive (PPV)* and *negative (NPV) predictive values* are sometimes used to measure the proportion of patients with positive and negative test results that have (*'true positive'*) or do not have (*'true negative'*) the disease, respectively. Unlike sensitivity/specificity, PPVs/NPVs are affected by the pre-test disease probability. The *likelihood ratio (LR)* (i.e. the probability of an expected test result in individuals with the disease of interest over the probability in individuals without the disease) is another diagnostic measure that is often used (121-124). LRs give the probability that a disease exists at different levels of a test's results by combining sensitivity and specificity in a simple formula. This means that LRs are only affected by the disease spectrum (12, 121-123). Lastly, diagnostic accuracy is less commonly measured using the *diagnostic odds ratio*, the *diagnostic effectiveness (accuracy)* and/or *Youden's index* (122, 123, 130).

Table 7: Diagnostic accuracy measures

Total Population (N)	**Disease (D)**	**No Disease (ND)**	Prevalence = D/N	Accuracy = (TP + TN)/N	
Positive Test (T+)	True positive (TP) (Power)	False positive (FP) (Type I error)	Positive predictive value (PPV) (Precision) TP/(TP + FP)	False discovery rate (FDR) FP/(TP + FP)	
Negative Test (T-)	False negative (FN) (Type II error)	True negative (TN)	False omission rate (FOR) FN/(FN + TN)	Negative predictive value (NPV) TN/(FN + TN)	
	True positive rate (TPR) (Sensitivity) TP/(TP + FN)	False positive rate (FPR) (Fall-out) FP/(FP + TN)	Positive likelihood ratio (LR+) TPR/FPR	Diagnostic odds ratio (DOR) LR+/LR-	Youden's index (TPR + TNR) - 1
	False negative rate (FNR) (Miss rate) FN/(TP + FN)	True negative rate (TNR) (Specificity) TN/(FP + TN)	Negative likelihood ratio (LR-) FNR/TNR		

[**Sources:** *(121-124)*]

**Abbreviations:* *D: disease; DOR: diagnostic odds ratio; FDR: false discovery rate; FN: false negative; FNR: false negative rate; FOR: false omission rate; FP: false positive; FPR: false positive rate; LR+: positive likelihood ratio; LR-: negative likelihood ratio; N: total population; ND: no disease; NPV: negative predictive value; PPV: positive predictive value; T+: positive test; T-: negative test; TN: true negative; TNR: true negative rate; TP: true positive; TPR: true positive rate*

***Description:* *Sensitivity/specificity are calculated along the first two columns, while PPV/NPV along the first two rows. LRs are calculated along the sensitivity and specificity lines, while DOR by dividing LR+ by LR-.*

2.2.4 Study Designs

2.2.4.1 Hierarchy of Evidence

Various study designs can be used for the evaluation of health technologies and the validity of their results can be ordered in a hierarchy of evidence (**Figure 6**). Starting from the bottom, the least reliable evidence comes from *expert opinion*, *anecdotes* and *editorials*, which are based on personal ideas, experience and/or observations. Going up one level, there are the *purely descriptive/non-analytic* studies, which inform on the causes of the disease, the participants' characteristics, and the place/time that the events occurred. These studies can be further divided into *ecologic* studies (i.e. compare outcomes between groups) or *case series/reports* and *cross-sectional* studies (i.e. follow patients over time). In the levels above lie the analytical study types, i.e. *observationa*l and *experimental* studies. Observational studies contain *case-control* (i.e. start with a disease/outcome and look backwards in time for its causes) and *cohort* (i.e. follow the logical sequence from exposure to outcomes), but also cross-sectional studies. Observational studies are prone to *selection bias* (i.e. participants are not randomly allocated to exposure or treatment). To avoid this, *randomised controlled trials (RCTs)* are performed. RCTs are conducted under a well-described protocol and conditions, and patients are randomly assigned into different groups with each group receiving a different intervention or diagnostic testing strategy (14, 88, 128, 131, 132). Although RCTs are often referred to as the *'gold standard'* in research since they avoid selection bias (133), all study designs may suffer from *internal* and *external validity* problems. Internal validity is the degree to which study results depend on the independent variable (e.g. exposure) and no other factors (e.g. confounders – factors that influence both the independent/cause and dependent/effect variables). On the other hand, external validity is the degree to which study results can be generalised or applied more broadly to actual clinical practice (121, 134, 135).

Well-conducted *evidence syntheses* and *systematic reviews* can bring together data from different study designs to answer specific research questions. Here, authors normally critically appraise the way the analyses were conducted, the results that were found, and the strength and weaknesses of the included studies. Systematic reviews can also include *meta-analyses*, where researchers pool results from different studies using statistical methods (14, 88, 128,

131, 132). *Decision-analytic models (DAMs)* are a type of evidence synthesis commonly used by health economists to draw together diverse evidence on the accuracy of diagnostic tests, the impact on treatment selection, the outcomes of treatment and the costs of all care. DAMs then use this evidence to examine and compare different interventions in terms of their costs and effects, assisting with decisions regarding the most efficient allocation of resources. There are several types of DAMs, each having different advantages and disadvantages. When compared to experimental studies, DAMs have also several advantages which makes their role, especially in EE, essential (e.g. comparing all possible alternatives). However, DAMs are also associated with many methodological challenges (e.g. decision on the optimal model structure), while their quality is only as good as the sources of the data that they use. For this reason, they should be considered as complements and not substitutes of the other study designs (17, 19, 120, 129, 136-139) [Further details on DAMs are provided in **Chapter VI**].

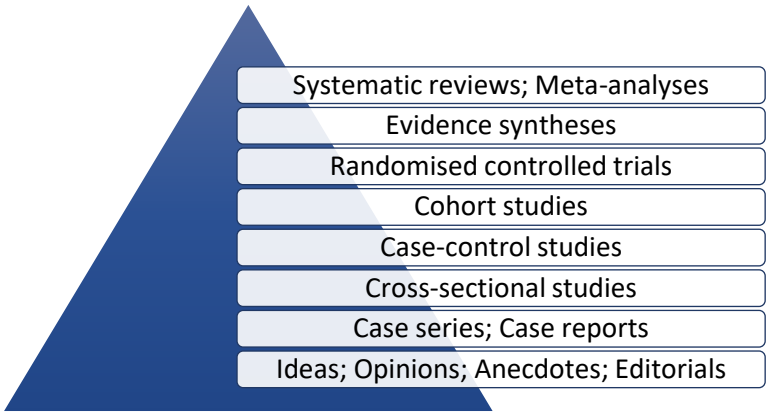


Figure 6: Hierarchy of clinical and economic evidence

2.2.4.2 Diagnostic Evaluation Studies

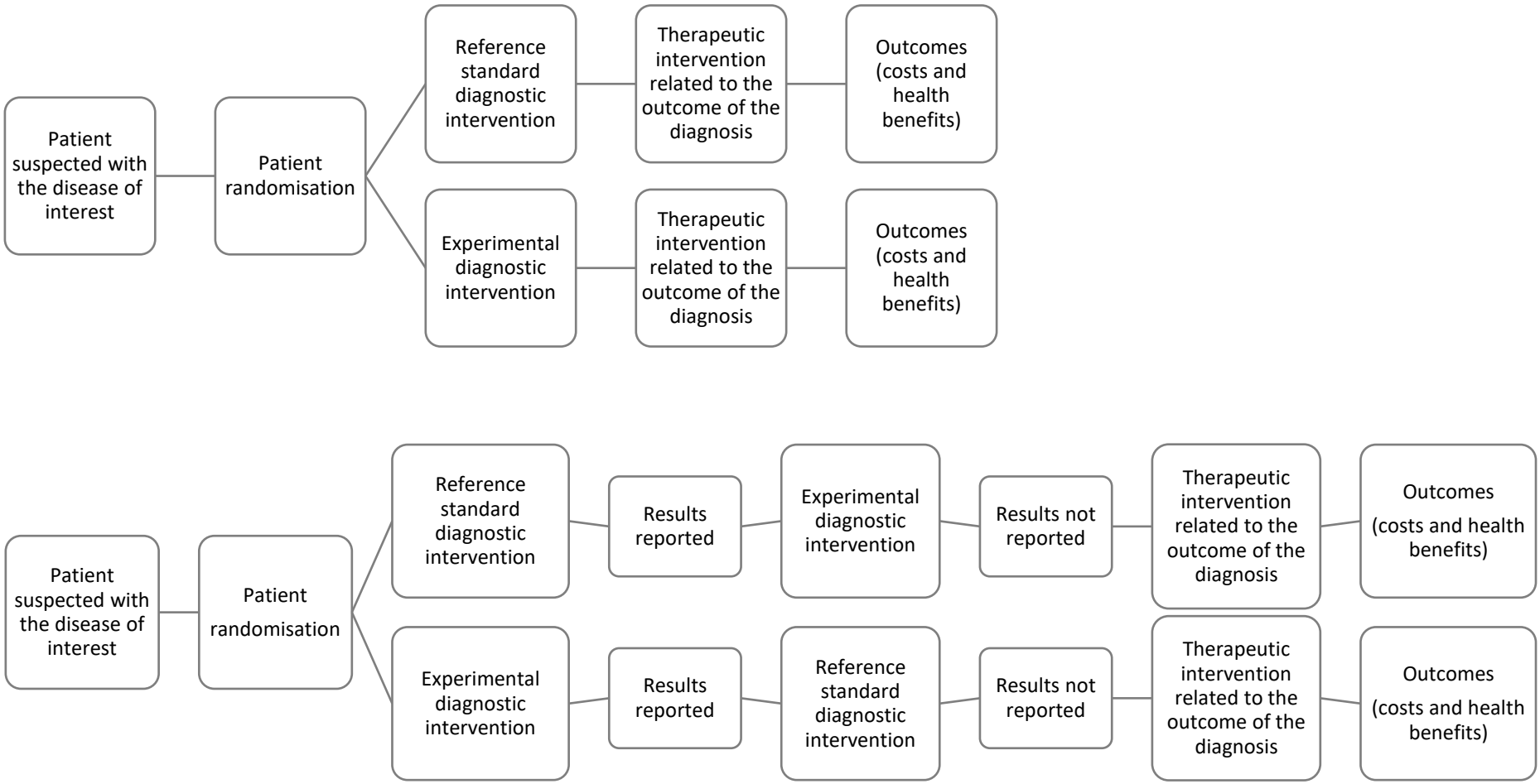
Diagnostic (accuracy) cohort studies are commonly performed to assess the characteristics of diagnostic interventions (**Figure 7**). In these studies, a patient suspected with the disease of interest undergoes both the index and reference standard tests in a sequential way (13, 16, 118, 121). Diagnostic cohort studies are simple and inexpensive to conduct, well-accepted by clinicians, and can be followed by further observational ('*before-after*') studies to examine whether physicians would change the patient's management plan after seeing the index test results (12, 13). Diagnostic cohort studies can measure diagnostic accuracy but cannot directly measure therapeutic impact, effectiveness or cost-effectiveness because the counterfactual

(i.e. what would have happened in the absence of the index test) cannot be observed (12, 13, 121). Sometimes, diagnostic case-control studies are used instead to estimate the diagnostic performance of a test (e.g. *ULTRADIAN*). In these studies, the index test is used in confirmed *cases* (with disease) and *controls* (healthy or those with other diseases) individuals. Although this study design is intuitively appealing, it is more prone to *spectrum bias* (i.e. the experimental test accuracy is influenced by the disease severity, and participant demographic and other characteristics) than diagnostic cohort studies. Therefore, these studies are more often used in the early evaluation of tests where the goal is to quickly explore the potential of an index test in the diagnosis of the disease of interest (140).

[**Source:** Image developed based on a figure presented in *Rodger et al. (2012) (13)*]

Figure 7: Prospective diagnostic accuracy cohort study

A more internally valid study design is an open or blinded diagnostic RCT (**Figure 8**). In RCTs, patient randomisation means that the health outcomes observed can be attributed to the test that has been used, while different tests can be directly compared in terms of their cost-effectiveness (13, 120, 129). Despite being a robust single study source of evidence, RCTs are not commonly used for tests. One reason is that it is not always feasible to blind the clinicians involved in the study, especially when they need to be aware of the procedure to administer the appropriate treatment (12, 13, 121). RCTs are typically more expensive than observational studies since they require a multidisciplinary research team, while to add statistical power in their results, they may require a large number of participants (from multiple sites to maximise generalisability) and/or a long duration of follow-up (12, 13, 16, 118). Lastly, RCTs require previous experience with using the technology and are time-consuming, which is a drawback given the fact that technology progresses rapidly and the examined test can soon be replaced by newer tests (12, 13).



[**Sources:** Image developed based on figures presented in *Rodger et al. (2012) (13)*, *Lord et al. (2006) (118)* and *Bossuyt et al. (2012) (120)*]

Figure 8: Open (*upper*) and blinded (*lower*) diagnostic randomised controlled trial

2.3 Economic Evaluation of Diagnostic Technologies

Until now, **Chapter II** has mainly focused on how diagnostic technologies can and should be clinically evaluated. However, just knowing that a test is effective is not enough for clinicians and policy makers to decide whether or not it should be adopted in clinical practice and be reimbursed given the growing demand for healthcare services and resource scarcity. To do so, an EE is required (17).

2.3.1 Defining Economic Evaluation

According to *Drummond et al. (2005) (17)*, EE is *"the comparative analysis of alternative courses of action in terms of both their costs and consequences"*. In other words, EE is a systematic procedure that aims to identify, measure and compare the costs (inputs) and health benefits (outputs) of two or more alternative healthcare products, procedures, services or programmes in order to find out which one is the most cost-effective. Given that the resources that are dedicated to health care are always limited (e.g. human, capital), choices need to be made (17-19, 138). To assist with these decisions, EE is often conducted in parallel with clinical evaluation using patient-level data (e.g. *'trial-based EE'*). However, EE can also be based on evidence synthesis from multiple sources estimating the effectiveness (e.g. meta-analysis of trial data), harms, and longer-term costs and outcomes of health conditions affected by the healthcare product, procedure, service or programme (i.e. DAMs). EE based on DAMs can reduce the uncertainty in the adoption/rejection decision as it is based on the totality of evidence rather than just one study (15, 17, 19).

2.3.2 The Role of Economic Evaluation in Health Care

To understand EE's role in health care, it is essential to introduce the concept of *'efficiency'*. Efficiency examines the relation between healthcare resource costs (i.e. labour, capital, equipment), and intermediate and/or final health outcomes (e.g. numbers treated, life-years gained) (141), and implies that society aims to allocate scarce resources in the best way to increase health benefits (142). There are two main types of efficiency in health care: *'technical'* and *'allocative'*. Technical efficiency is concerned with comparing and choosing

different combinations of resources to achieve maximum health benefit for a given cost or minimum cost for a given health outcome within one area of health care. However, it cannot inform resource allocation decisions at a broader level (i.e. across all and/or outside health care) since outcomes are incommensurate (143, 144). Allocative efficiency accounts not only for the way that resources are used to produce health benefits but also for how resources should be spent in different areas of health care to maximise health for the entire population. Both technical and allocative efficiency are based upon the *'Pareto improvement'* principle, which states that social welfare (wellbeing) increases when an individual's welfare increases without another person's falling, or when individuals' welfare increases sufficiently so that they would be hypothetically willing to compensate those who lose and still remain better off. When all Pareto improvements have been made (i.e. not possible to make anyone better off without making someone else worse off), social efficiency *('Pareto optimal'*) has been achieved (144-146).

A 'perfect' market is assumed to tend towards an equilibrium between supply and demand (e.g. for healthcare services), meaning that prices go up or down depending on whether there is a shortage or abundance of goods, respectively. The equilibrium price is the price at which the quantity demanded equals that supplied. Rational providers (e.g. device manufacturers) would supply their products at this price when there is an efficient way to maximise their profits, while rational consumers (e.g. patients) would be willing to pay this when the gain (i.e. health) from doing so exceeds the extra cost incurred. Perfect markets would lead to a socially efficient allocation of resources (Pareto optimal) when there are no costs or benefits to society other than those experienced by individual consumers and suppliers. Cases in which Pareto optimality is not achieved are described by the term *'market failure'*. In health care, potential causes of market failure include: a) externalities (i.e. costs and benefits that are not borne by the consumer or producer involved); b) monopoly (e.g. high prices of patented products); c) public goods (i.e. provided to all individuals when needed); and d) imperfect information (i.e. uncertainty in the demand and supply of health care) (145). Because of market failure, governments around the world intervene in the funding and provision of health care. Therefore, EE is often used by health system funders to analyse information on cost-effectiveness, mimic the function of the market and ensure that healthcare resources are allocated in the most efficient way.

There are several forms of analysis that can be performed in any EE study. The decision of the most appropriate type should be made based on the: 1) research question; 2) institutional framework (e.g. healthcare system organisation); 3) practical measurement challenges (e.g. ability to estimate effectiveness with specific measures); and 4) economic perspective (i.e. payer's viewpoint) of the analysis (17). To define the role of EE in health care, the analyst can follow one of the three theoretical and methodological approaches that are described below. The choice of approach should be made very carefully since each one has different benefits and limitations and may lead to different results given the way that outcomes are identified, measured and valued. Here, it should be noted that there is currently no consensus among health economists about the exact distinctions between these frameworks. Therefore, the definitions provided are based on some of the most relevant sources (17, 144, 147-152):

- Firstly, there is the *'welfarist'* approach, which focuses on the maximisation of the social welfare subject to a societal budget constraint, and subsequently values the output of health care based on its contribution to the overall welfare. This framework considers the values (in monetary terms) that individuals place on outcomes and states of the world according to their preferences (i.e. desirability judgements on different options) on the assumption that they are the best judges of their own welfare given their income restraints. These values are usually measured by calculating the total amount that individuals are *willing to pay (WTP)* for the evaluated healthcare programme (i.e. sacrifice in terms of other commodities) or are *willing to accept* in compensation to forego the product's benefits. Alternative approaches to monetary valuation include discrete choice experiments (see *Section 4.4.4* below), where cost is one of several attributes of health care which respondents are asked to trade off. These amounts can then be compared to the cost of the programme to examine whether it is worthwhile to adopt it and replace the existing alternative(s), or use this budget for other healthcare- and/or non-healthcare goods/purposes.

- The second approach is called *'extra-welfarist'* and its aim is to maximise health in a resource-constrained healthcare system. For this reason, this approach places emphasis on the limited healthcare budget and how it can be spent most efficiently among alternative programmes to maximise the health of the population. Here, the output of

health care is valued in terms of its contribution to health itself and not by comparing preferences between health- and non-health-related products/attributes. More precisely, 'extra-welfarism' compares the healthcare resources that are used with the health benefits (status) that are obtained from each alternative programme measured in natural units (i.e. health effects not valued by the individual's preferences) and/or health-related preference-based weighting of outcomes (e.g. quality-adjusted life-years). This rationale avoids some of the problems of 'welfarism' (e.g. ability to pay; discrimination against those outside the labour force; monetising health effects), and moves beyond the Pareto principle by allowing interpersonal wellbeing comparisons in a variety of dimensions (e.g. health-related quality of life; life expectancy; capability). Nevertheless, because of its narrow health-oriented goals, it may not provide sufficient information to decision-makers about how to allocate scarce resources in the society.

- The third approach is that of the pragmatic *'decision-maker*' and constitutes a broader societal perspective with a strong preference for "equal access for equal need", especially when decisions are made in a fairly centralised and public environment. Here, the analyst estimates a wide range of costs and consequences, and informs decision-makers on how to allocate resources to the optimal alternative. This approach allows health benefits to be measured in both natural and monetary units, without favouring one valuation method over the other, while it takes into account the differences in the individuals' economic background, health status and life expectancy when they make decisions. Here, it should be noted that despite their differences on a theoretical basis, the 'extra-welfarist' and 'decision-maker' perspectives have a common objective: to maximise health output from available resources. This is why they are often referred to as *'non-welfarist'* approaches, are adopted by several health technology assessment bodies (e.g. the National Institute for Health and Care Excellence), and are widely implemented in health economics.

2.3.3 Perspectives in Economic Evaluation

In any EE, it is essential to specify the study perspective from the very beginning to determine the types of costs and health benefits that will be included in the analysis, and subsequently the viewpoint from which the decisions will be made. The three main perspectives are those of the *healthcare provider* (i.e. healthcare system or hospital/clinic), *patient* and *society*. The healthcare provider viewpoint includes all costs associated with the delivery of health service to the patient. Specifically, it accounts for the costs of screening and diagnosis (e.g. medical devices, equipment and consumables), treatment (e.g. drug acquisition, administration and monitoring), other healthcare resource use related to the management of the disease (e.g. primary care visits, hospital admissions, adverse events), and the salaries of healthcare professionals. However, it does not consider the costs incurred by the patient for obtaining care, such as transport to healthcare facility, time off work, over-the-counter purchases, and co-payments. Furthermore, the health outcomes that are used under the healthcare provider perspective are measured based on valuations made by the general population (obtained through interviews/surveys) and not patients' valuations of their own health states (patient viewpoint). The societal is the broadest perspective since it involves a full range of social opportunity costs associated with the use of the examined interventions (e.g. healthcare use and patient costs; productivity losses due to the patients' inability to work and how these are affected after the introduction of a new technology). However, it is often challenging, time-consuming and expensive to measure the value of lost productivity and the impact of an intervention on every affected individual in the society. For this reason, many EEs use the narrower healthcare provider perspective (17, 153, 154).

2.3.4 Types of Economic Evaluation

There are five main EE types that can be performed, all of which are based on the different theories described in the previous section. Although all analyses measure costs in monetary units, they differ on how they measure and value outcomes (**Table 8**) (17-19).

2.3.4.1 Cost-Minimisation Analysis

Cost-minimisation analysis (CMA) is a type of EE that is used when two interventions have equivalent effects but are different in terms of their costs. In CMA, it should have previously been established that the comparator technologies are clinically equivalent (e.g. drugs of the same pharmacologic category). However, in practice, it is more common for effectiveness to vary between different programmes (17-19).

2.3.4.2 Cost-Effectiveness Analysis

Cost-effectiveness analysis (CEA) is an EE form that is used when alternative technologies differ in both their costs and consequences. Consequences can be measured using any appropriate non-monetary natural/physical units/effects that are specific to the programme that is examined (e.g. 'life-years gained', 'disease cases correctly diagnosed'). For CEA to be performed, a single, common measure of output (i.e. cost per unit of effect) should be considered (17-19, 138).

2.3.4.3 Cost-Consequence Analysis

Sometimes, consequences occur in more than one group (e.g. patients and carers) or are diverse and impossible to summarise in one primary outcome measure. In this case, a *cost-consequence analysis* may be used since an array of outcomes are presented in the analysis. As in CEA, policy makers should then value these outcomes and compare the most important ones before making decisions on how to best allocate their limited resources (17, 19).

2.3.4.4 Cost-Utility Analysis

Cost-utility analysis (CUA) is an EE form that is interested in both the quantity and the quality of life that will be produced if a healthcare programme is reimbursed. CUA uses a measure called *'utility'*, which in economic theory is defined as the personal gain that is expected after consuming a good/service (17-19, 138). Utility can be interpreted as the preference of an individual or society to have one set of health outcomes instead of another (i.e. the more preferred an outcome is, the higher its utility). Utility/preference is a relative value that is placed on a specific health state (e.g. diseased, non-diseased) in the form of a score (weight) to compare it with other states. Crucially, this weight is anchored at '0' (health state as bad as death) and '1' (perfect health) allowing improvements in health-related quality of life

(HRQoL) (morbidity) to be valued against improvements in life expectancy. Measuring preferences is done by first defining the health states of interest and then valuing them using techniques, such as *'rating scale'*, *'category scaling'* or *'visual analogue scale'*, *'standard gamble'* and *'time trade-off'*. Alternatively, this can be done by using validated generic health questionnaires, usually administered to patients or carers, that have pre-specified valuation tariffs derived from general population surveys (e.g. EuroQoL EQ-5D) (17, 19, 138, 155-157).

2.3.4.4.1 Quality-Adjusted Life-Years

By using utility/preference analysis, CUA provides a generic measure of effects that can be used in comparisons across different areas of health care (i.e. diseases, settings). This generic measure is commonly expressed using quality-adjusted life-years (QALYs), which combine both the quantity (i.e. length of time; gains from reduced mortality) and quality (i.e. utility value; gains from reduced morbidity) of life into a single index (17, 18, 88, 138, 155). CUA results are expressed in terms of cost per QALY gained. In QALYs, HRQoL weights are again anchored at 'zero' (death) and 'one' (perfect health). Health states worse than death can also be characterised by receiving a negative utility value. The preference scores are then multiplied by the length of time spent in each disease state. The area under the curve (AUC_{QALY}) that is created by plotting a graph with HRQoL values on the y-axis and length of time on the x-axis gives the total number of QALYs (**Figure 9**) (17, 138, 155).

One concern that has been raised with QALYs is that they are considered of equal social value meaning that they are 'blind' to health conditions and personal characteristics (e.g. age, gender, disease severity). However, some researchers view this as a virtue since it treats all alike regardless of their characteristics. Additionally, QALYs do not capture all the benefits associated with a healthcare intervention (e.g. patient's wellbeing outside health) and are difficult to obtain when health is valued using condition-specific measures/questionnaires since results need to be mapped to a preference-based measure, such as the EQ-5D (155).

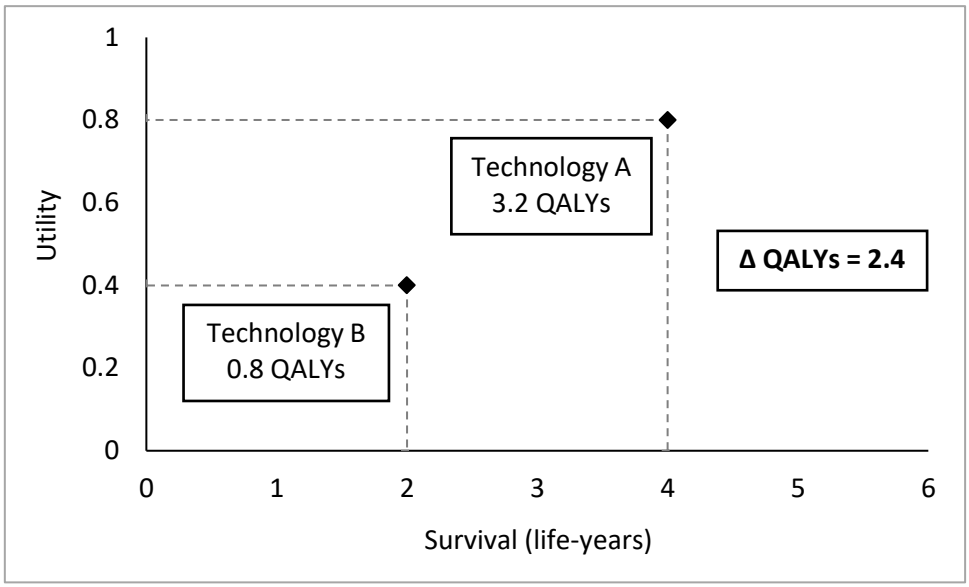


Figure 9: Quality-adjusted life-years example

2.3.4.5 Cost-Benefit Analysis

Cost-benefit analysis (CBA) is used when both the costs and consequences of the examined interventions are measured in monetary units. In comparison to CEA and CUA that are concerned with how to efficiently allocate an existing healthcare budget between alternative programmes by focusing only on their comparative cost-effectiveness, CBA examines whether it is worth expanding or even using the budget for this purpose (i.e. value of resources within and outside the healthcare sector). This is the reason why CBA is often considered as the broadest form of EE. In CBA, results are normally presented as ratios of costs to benefits or the net benefit (loss) of one health technology over another. A comprehensive CBA might use human capital, revealed preferences and/or WTP to value the health benefits. However, one problem of this form of analysis is the fact that the range of effects that could be measured in monetary units is quite limited, while translating them into their monetary benefit is not always an easy task. Consequently, although CBA is widely used in other areas (e.g. transport, environment, education) and some health economists strongly endorse it, its use in health care is more restricted than the use of the other EE types (17-19, 138).

2.3.4.6 Choosing the Type of Analysis based on Approach

Based on the definitions given above, a 'welfarist' analyst would most probably use CBA to estimate the individuals' WTP for a service. 'Extra-welfarist' analysts would most likely use CEA/CUA since they would be more interested in the total health improvement. Lastly, a 'decision-maker' would prefer to collect all the available information on costs and consequences. Therefore, cost-consequence analysis or alternatively, CEA/CUA would be the form of EE that would be chosen (17, 138, 147, 149, 150).

Table 8: Characteristics of the different types of economic evaluation

Type of Analysis	Measurement/ Valuation of Costs	Measurement/Valuation of Health Benefits
Cost-minimisation	Monetary units	Health benefits assumed/demonstrated to be equivalent between alternatives
Cost-effectiveness & Cost-consequence	Monetary units	Natural units common for all alternatives (e.g. life-years gained, cases diagnosed)
Cost-utility	Monetary units	Healthy years (e.g. quality-adjusted life-years)
Cost-benefit	Monetary units	Monetary units

[**Source:** Table developed based on a table presented in *Drummond et al. (2005) (17)*]

2.3.5 Challenges in the Economic Evaluation of Diagnostics

EE has been widely used as a tool when making resource allocation decisions in health care, including the use of medicinal products (e.g. drugs). However, its use in diagnostics has been quite limited (26, 158). In 1997, *Sassi et al. (158)* identified some major methodological challenges when conducting an EE for diagnostics that can be grouped into five areas:

1. **Identifying appropriate alternatives:** Often, multiple diagnostic strategies exist for each condition. Diagnostic technologies have also multiple applications in the same disease and in different diseases which need to be assessed individually and together. Moreover, since

a new test can be used either in combination or as a replacement of an existing procedure, an oversimplification of diagnostic processes in trials and DAMs is not representative of actual clinical practice.

2. **Costing:** A challenge with costing diagnostics is that many of the resources that they use may be shared (e.g. capital/fixed costs, like hospital and equipment; variable costs, like human resources). Therefore, it is important to identify only the costs that are associated with the use of the technology that is examined when comparing its cost to that of the existing alternatives and measuring its opportunity cost.

3. **Evaluating health outcomes:** As shown in *Section 2.2.2*, the value of a test does not only depend on its accuracy/performance (intermediate outcome) but also on the choice and effectiveness of the subsequent treatment and management plan. Additionally, a medical test may also have an intrinsic value if it provides reassurance to patients and physicians. Conversely, the test may cause distress (e.g. if it results in a false positive finding) and/or harm due to adverse events or by triggering a cascade of unwarranted further tests and treatments.

4. **Selecting appropriate study designs:** *Section 2.2.4.2* presented the different study designs that can be used when evaluating the effectiveness and cost-effectiveness of diagnostics. In all cases, a large sample size is required, while difficulties conducting the studies (e.g. ethical problems; identification of homogenous patient groups; differences between actual and ideal clinical practice) should be handled appropriately.

5. **Other general issues:** These include rapid technology advances that might make newly marketed tests become quickly obsolete or bring new applications and/or performance improvements to existing tests, and the development of new therapies.

2.3.6 Early Economic Evaluation in Diagnostics

This thesis focuses on a new diagnostic medical device (MD) that is currently at an early development stage (i.e. technological capability; before large clinical trials). One distinction between 'early' and conventional EE is that rather than simply examining whether a new technology is cost-effective and should be adopted, it informs and determines whether further investment and development (i.e. technical improvement) is needed, and/or whether further research (i.e. clinical trials) should be conducted to reduce existing uncertainty and support the decisions made. This is because the aim of this early assessment is to reduce the failure rate and costs at each stage of product development (99, 159, 160). Appropriate methods for early EE have begun to receive more attention in recent years. For this reason, an exploratory literature review was undertaken in August 2018, in which PubMed® (https://pubmed.ncbi.nlm.nih.gov/) was searched for peer-reviewed publications, while internet searches using Google™ Scholar (https://scholar.google.com/) were carried out to identify grey literature in this area. Search terms, such as 'early economic evaluation', 'premarket assessment' and 'early cost-effectiveness', were used to identify potentially relevant publications. The citations of the eligible papers were then hand-searched to find other relevant literature. Below there is a brief description of the methods that were found.

2.3.6.1 Multi-Criteria Decision Analysis

Multi-criteria decision analysis (MCDA) is used to make/support decisions between discrete alternatives. Although it has several validated methods (e.g. *'conjoint analysis'*, *'contingent valuation'*), the *'analytic hierarchy process' (AHP)* is the most commonly used and the one that has been applied to diagnostics and MDs. In AHP, researchers make a series of pairwise comparisons (e.g. novel vs current practice technologies; sets of hypothetical new products) in terms of similar/common attributes (criteria). They then attach priority weights to these criteria and create a hierarchy structure with a range from the most to the least important element based on preferences and/or personal experiences from different stakeholder/user groups. In other words, this method resembles the way that individuals implicitly make decisions in their daily lives by evaluating their consequences (30, 31, 34, 88, 161, 162). This tool has the advantage of including in the analysis additional factors apart from clinical and cost effectiveness (e.g. patient preferences on device characteristics). Therefore, it is a useful

method for the early EE of diagnostic MDs since it considers all the factors that can affect the relative value of a product and can later introduce them to an economic model to inform decisions regarding further development, market approval and/or coverage (30, 34). Obviously, this method has limitations since it is challenging to weigh each criterion and convert priorities to absolute estimations for each parameter (30).

Hummel et al. (2011) (163) used MCDA-AHP to elicit expert knowledge/estimation regarding the diagnostic accuracy, patient comfort and safety of a new imaging technology (photoacoustic mammography) for the diagnosis of breast cancer relative to magnetic resonance imaging (reference standard test). The authors then converted the relative diagnostic performance information to estimations of the sensitivity and specificity for the novel test, and estimated the indirect impact of patient comfort and safety on patient health utility. The data were used to populate a previously published DAM that compared the cost-effectiveness (cost/QALY) of the two technologies.

2.3.6.2 Decision-Analytic Modelling

DAMs are commonly used in early comparisons of diagnostic interventions since they provide a simplified representation of real-life, are inexpensive and can be easily updated as evidence becomes available (14, 17, 29, 30). Modelling synthesises available data to examine the cost-effectiveness of a new test under different scenarios, and understand its potential role and value in clinical practice (14, 29, 30, 33). Nevertheless, during the development of a new test, it is difficult to obtain data that fully inform on its potential accuracy, risks and benefits, and consider all different patient cases and a full range of the test's future outcomes (88, 99). Therefore, in early models for diagnostics it may be necessary to stratify patients in one or two groups, examine one or two promising applications of the test, consider only the main outcomes, and use a limited number of alternative diagnostic pathways or health states based on data availability (88). Sensitivity analysis is also valuable to examine the key parameters that could potentially affect the test's cost-effectiveness and identify the areas where further research is needed (14, 22-25, 30, 34-36).

An example of an early DAM was developed by *Buisman et al. (2016) (22)* who aimed to assess the early cost-effectiveness (cost/QALY) and potential role in clinical practice (i.e. add-on or

replacement) of various new diagnostic testing strategies for rheumatoid arthritis. The authors used data from published literature, patient cohorts and trials, and based on expert opinion to populate the model, while different types of sensitivity analysis were performed to explore the impact of changing important parameters (e.g. diagnostic accuracy, costs) on the cost-effectiveness outcomes.

2.3.6.3 Headroom Approach

The *'headroom'*, or '*effectiveness gap*', method is a simple threshold approach that is used to rapidly estimate the maximum/ceiling cost (headroom) that an intervention can have in order to be cost-effective (i.e. reimbursed) at a given WTP threshold (e.g. £20,000-£30,000 per QALY). Here, a simple cost-utility mathematical model (e.g. ΔC/ΔE, where ΔC and ΔE the incremental costs and health benefits associated with the technology compared to existing alternatives) and univariate sensitivity analysis are used to examine important parameters and the range of prices at which the new product would be more cost-effective. This tool can be used throughout the innovation process to inform go/no-go investment, research and development (R&D) decisions and can be updated as more evidence becomes available. However, it requires an explicit WTP threshold, which limits it to specific markets/countries. In addition, it requires good knowledge of where in the clinical pathway the new technology will sit in order to build the mathematical model. This may be more straightforward for therapeutic than diagnostic technologies, which, as previously discussed, might be used in different roles in clinical practice (e.g. triage, add-on). Furthermore, the deterministic character of this tool may easily result in false-stop decisions if other key parameters change rapidly. The method also focuses only on the product's potential cost-effectiveness (e.g. based on the performance of the device), while actual reimbursement and uptake may be based on other factors (e.g. ease of use) (14, 23-25, 30, 34, 164, 165).

After running the DAM to compare the cost-effectiveness of different tests for breast cancer, *Buisman et al. (2016) (22)* used the headroom approach to find the maximum cost for each test below which the respective test was likely to be the most cost-effective strategy at the €20,000/QALY WTP threshold. These costs were measured by considering the role of each test in clinical practice, its diagnostic accuracy, and whether it was aimed to be given to all or

some (e.g. intermediate risk) of the patients. The authors then identified the tests for which the unit costs should decrease in order to become more efficient alternatives.

2.3.6.4 Bayesian Framework and Value of Information Analysis

The *Bayesian analytical framework* is used to combine prior evidence or assumptions about a parameter to posterior knowledge when new information becomes available. It is used to examine how robust current evidence is, where more information is needed, the appropriate design for future clinical trials, and the value of investing in further R&D (25, 30, 32, 33, 79). VOI analysis uses Bayesian methodologies to compare the costs and benefits of acquiring additional information on a parameter or set of parameters. VOI methods include the *'expected value of perfect information (EVPI)'*, the *'expected value of partial perfect information (EVPPI)'*, the *'expected value of sample information (EVSI)'* and the *'expected net benefit of sampling (ENBS)'*. EVPI is an abstract concept that reflects the difference between using current knowledge (uncertainty) and perfect information (no uncertainty) in the population of patients who might benefit from the test. EVPPI is a similar concept that focuses on a specific subset of parameters where uncertainty can be reduced through further research (e.g. larger study). EVSI estimates the value of new evidence from a study of a given sample size, and by definition, is less than EVPI and EVPPI. If the cost of additional research is less than EVSI, then ENBS is positive (24, 25, 30, 33, 137, 166). EVSI is possibly the most useful VOI concept at the early stages of development when the manufacturer is still deciding whether to persist with the technology and if so, which study to do next (166). One limitation of all VOI methods is that they require a probabilistic DAM to be built, which requires more mature data and/or assumptions (25, 30, 32, 33).

Van de Wetering et al. (2012) (35) used a DAM to examine the cost-effectiveness (cost/QALY) of a new point-of-care (i.e. at or near the point of patient care) test for the detection and monitoring of abnormal potassium levels in high-risk patients with heart failure requiring antihypertensive medications. The authors then estimated the EVPPI of specific parameters (shown to be vital to the cost-effectiveness outcomes in sensitivity analysis) and compared them to the cost of additional research to examine whether continuing with the R&D of the new technology would be worthwhile.

2.3.6.5 Clinical Trial Simulation

Clinical trial simulation (CTS) uses mathematical models to synthesise available data on a product. It uses computer simulation (e.g. Monte Carlo) to represent a real-world situation and is used to estimate efficiency and clinical response before clinical data are available. CTS also explores *'what-if'* scenarios about parameters and study designs before real clinical trials are conducted to identify any limitations. Its aims are to use previous information as efficiently as possible and improve the protocols of future trials as well as their results. The estimates derived from its outputs can provide more information on parameters that would otherwise be unavailable at early development stages and can prevent manufacturers from committing resources to potentially unnecessary studies (25, 32, 33). So far, CTS has mainly been used in drugs to optimise their dosage and plan future trials (32, 33).

2.3.6.6 Choosing between Early Economic Evaluation Methods

The choice of the most appropriate early EE method mainly depends on the stage of the product's development and the available clinical and economic evidence. Their use also depends on the target audience for these early assessments, i.e. manufacturers, regulators or policy makers (25, 159). Early phases of product development are characterised by limited clinical and economic data (e.g. no/small previous studies; unit cost estimations). Therefore, implementing methods that require a lot of data (e.g. CTS, sophisticated DAMs) is challenging and is better to be avoided until later stages when more evidence becomes available. MCDA would be useful at this stage since it can produce some initial data (e.g. diagnostic accuracy; product characteristics and costs) by simply asking different stakeholders/experts to make some plausible assumptions based on their experience and/or preferences. This analysis can indicate the health technologies that have more potential to be accepted by different user groups and prioritise the areas where further investment would be worthwhile (25).

Another equally applicable method would be the headroom approach since it uses broader estimates to determine the maximum reimbursable price of the new technology (adoption decision) under the optimistic scenario/assumption that the product is effective: [e.g. H = (Net cost reduction ignoring the device's price) + R_T*(additional QALYs produced)]. Based on its results, a promising idea/product can proceed to further development or be abandoned (25, 159, 160, 164, 165). The headroom approach can also be used later in the development cycle.

However, in this case, more realistic costs (e.g. production), revenues and benefits should be considered: [e.g. V = M*(H-U), where *V* are the revenues, *M* the number of items sold over a time horizon, *H* the maximum reimbursable price, and *U* the cost of production per unit)] (165).

Early models are useful tools when a decision for further development of a product has been made. By using models, manufacturers can make an early market assessment (i.e. competing products), decide the areas where they need to invest, and make some first estimations of the product's future price and reimbursement scenarios (79). Models can also be updated once more clinical and economic evidence becomes available, and can be more robust closer to the product's launch (79, 160). Furthermore, models can be combined with VOI methods for manufacturers to identify the impact of different types of uncertainty in the product's development (R&D decision) (79, 159, 160). These methods can also be used by decision-makers at later stages of development since the situation that they are facing is not dissimilar to that faced by manufacturers at earlier stages (i.e. are there benefits from collecting further information?) (164).

In some cases, collecting additional evidence before deciding on further product development might be the best strategy since it may prevent manufacturers from wasting a considerable amount of money on a technology that is not cost-effective. However, waiting may also result in health benefits forgone if the emerging technology can produce higher health benefits than the existing alternatives. Therefore, a trade-off between the costs and benefits of further R&D should be made (160, 167). A method proposed to assist with this is called *real options analysis (ROI)*. ROI considers both the uncertainty surrounding the decision of the technology adoption (**Figure 10**) and the cost/inability to reverse the decision to adopt it (160, 165, 168). Early EE should be an iterative process that re-runs as more evidence becomes available (29, 160). For this reason, some health technology assessment bodies approve technologies under the condition that further evidence will be collected (*'coverage with evidence development'*). The problem here is that once a product is in the market and covered, it might be politically difficult to withdraw it if it is found to be ineffective, while manufacturers might lose incentive to conduct any further research (101).

	Type of Uncertainty
Resolvable	Functionality (e.g. technological/design)
	Production costs (e.g. labour, capital)
	Health service issues (e.g. service costs, clinical pathways)
	Treatment effectiveness (e.g. comparator treatment)
	Competition (i.e. other products)
Irresolvable	Market penetration (e.g. service policy, local purchasing)

[**Source:** Image developed based on figure presented in *Girling et al. (2015) (165)*]

Figure 10: Uncertainties in product development

2.4 Applying Early Economic Evaluation Methods to U-Rhythm

Chapter II has provided a summary of the methods that can be used for the clinical and (early) economic evaluation of diagnostic technologies and could potentially be implemented in this thesis for the early EE of the *U-Rhythm* device. In general, applying the methods described above to *U-Rhythm* is challenging for many reasons. Firstly, the device has clinical applications in a number of relatively rare endocrine disorders with little pre-existing research. This is because different clinical sites and even clinicians in the same centre tend to use different approaches to diagnose these diseases. Secondly, the device is still at a (very) early stage of development, since its size, design and components are evolving to optimise it and work towards marketing approval (CE marking). The evolution in the design of the device as well as the number and type of assays that are used to analyse samples makes it difficult to estimate its final cost and diagnostic accuracy when in real-world use. The test's unit cost will evolve and may decrease once the technology stabilises and moves from research to production at greater scale. The technical capabilities of the device and the assays used to analyse the hormone profiles are also evolving.

At the outset of this PhD project, it was not clear in which of the six endocrine disorders examined in *ULTRADIAN U-Rhythm* would demonstrate technical capability and diagnostic promise. Therefore, the author decided to first conduct a broad systematic literature review

(**Chapter III**) to identify and appraise the quality of previous EEs of diagnostic tests across all six conditions. The aim was to find RCTs and/or model structures that have been used for these diseases to inform a future DAM for one or more of the conditions. For three disorders (primary aldosteronism, Cushing's syndrome, Addison's disease) where it became evident that *U-Rhythm* had higher diagnostic potential, *ULTRADIAN* (**Chapter IV**) collected limited data on participants' satisfaction with the device, recent healthcare resource use and impact of their condition on their work productivity, and HRQoL scores, which could be used as input parameters for cost-effectiveness models and could help improve the device characteristics. For one condition (primary aldosteronism) the author then used routine data (**Chapter V**) to explore healthcare resource use during the pre-, peri- and post-diagnostic period to provide more context about current diagnostic and therapeutic practice. For the same condition, the author used the economic models identified in **Chapter III** and the wider literature to develop a DAM to perform an early EE of *U-Rhythm* (**Chapter VI**). Based on the literature review of the EE types and methods reported in **this Chapter** and given the limited availability of data, the author decided that the early economic model would take a narrow UK National Health Service perspective and the 'extra-welfarist' approach to identify the role of *U-Rhythm* in current practice, and estimate the potential cost and QALYs of the device compared to existing alternatives. The author used and contrasted the headroom approach and VOI in estimating the value in continuing to develop the technology and conduct further research.

CHAPTER III

SYSTEMATIC LITERATURE REVIEW

ECONOMIC EVIDENCE ON OPTIMAL APPROACHES TO DIAGNOSIS AND MONITORING IN RARE ENDOCRINE DISORDERS

CHAPTER III OVERVIEW

Chapter I provided information on the rationale and objectives of this PhD project mentioning that the overall aim is to apply health economic methods to summarise current knowledge on the diagnosis and monitoring of several endocrine disorders, and examine the potential clinical and economic value of a new diagnostic device (*U-Rhythm*) in one of them (primary aldosteronism). **Chapter II** gave a brief overview of the methods that can be used in the clinical and (early) economic evaluation of diagnostic tests providing also information on the most common diagnostic accuracy and effectiveness measures, and study designs that are used for this purpose.

Chapter III reports a systematic literature review that was undertaken to identify and appraise the quality of the available economic evidence evaluating the diagnosis, monitoring and treatment of six selected endocrine conditions: acromegaly, Addison's disease, congenital adrenal hyperplasia, Cushing's syndrome, growth hormone deficiency, and primary aldosteronism. The review also provided an overview of the methods which have been used to assess the cost-effectiveness of some of the diagnostic tests that are used in these diseases to inform a model-based economic evaluation that is conducted later in **Chapter VI**.

Economic Evidence on Optimal Approaches to Diagnosis and Monitoring in Rare Endocrine Disorders

3.1 Background

Chapter I mentioned that one of the objectives of this PhD thesis is to explore the potential value of *U-Rhythm* in the diagnosis of endocrine conditions. To achieve this goal, a systematic literature review of the current economic evidence on diagnostic and monitoring methods for endocrine disorders was first conducted. There is currently little information regarding the cost-effectiveness of the different diagnostic tests/strategies for these diseases. There is also limited evidence on the cost of the general diagnosis and management of these conditions, meaning that it is hard to know the impact that they have on the healthcare expenditure of a country (169-175). Therefore, the aim of this systematic review (**Chapter III**) is to assess the extent to which the current clinical care and practice guidelines for the diagnosis and management of endocrine disorders are based on high-quality economic evidence (41-46).

The review findings will give a better idea of the diagnostic and monitoring tests that are normally used in clinical practice, and the methodology that is followed when examining their cost-effectiveness. Additionally, the outcomes from this review will be used to design and conduct a new economic analysis for the optimal diagnosis and management of endocrine diseases (primary aldosteronism), in which a decision-analytic model will be developed and populated (**Chapter VI**). Given that there is a large number of endocrine disorders, ranging from diseases with a higher prevalence (e.g. diabetes mellitus) to rare diseases (e.g. Addison's disease), this review only focuses on the six selected disorders where the dynamic diagnostic testing of hormones over a 24-hour period (using *U-Rhythm*) is thought to have the most potential to improve diagnosis and treatment (**Chapter I**). To the author's knowledge, this is the first systematic review examining this research question for these six conditions (37, 176). Details of the protocol for this systematic review were registered on *PROSPERO* and can be accessed at: https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=73860.

3.2 Methods

3.2.1 Review Question and Objectives

The primary objective of this systematic review is to examine the available economic evidence evaluating diagnostic testing and monitoring tools for six endocrine disorders: acromegaly (AC), Addison's disease (AD), congenital adrenal hyperplasia (CAH), Cushing's syndrome (CS), growth hormone deficiency (GHD), and primary aldosteronism (PA). This includes studies where the costs, or the costs and health benefits of tests currently used in clinical practice are reported.

3.2.2 Inclusion and Exclusion Criteria

To answer the review question, the *PICO* framework (i.e. P: Population; I: Intervention(s); C: Comparator(s); and O: Outcome(s)) was followed (176). Below, the inclusion and exclusion criteria of this review are listed (**Table 9**).

3.2.2.1 Population

Studies of human participants with a suspected diagnosis of any one of the abovementioned endocrine diseases (i.e. AC, AD, CAH, CS, GHD or PA) were included. Studies of animals with these disorders or human patients with suspected (endocrine) diseases other than the ones mentioned above were excluded.

3.2.2.2 Interventions and Comparators

Any laboratory or radiological test used in current practice (primary, secondary or tertiary care) for the diagnosis and/or monitoring of these endocrine disorders was included. Examples of diagnostic/monitoring tools that were included are: oral glucose tolerance test; short Synacthen stimulation test; midnight serum cortisol test; insulin tolerance test; saline infusion test. Studies evaluating only treatment were excluded.

3.2.2.3 Outcomes

Any studies reporting the costs of diagnostic procedures used for the abovementioned endocrine disorders, including the cost of the tests and/or treatments, were included. Studies that reported both costs and patient outcomes following diagnosis were also included. Costs could be measured in any currency, while patient outcomes could be summarised using any intermediate (e.g. cases accurately diagnosed) or health outcome measure (e.g. life-years; quality-adjusted life-years, QALYs). Studies measuring the cost of treatment and not of the preceding diagnostic method, or evaluating patient outcomes without costs were excluded (177).

3.2.2.4 Study Design, Perspective and Time Horizon

All different study designs that contained a partial or full economic evaluation (EE) (e.g. cost, cost-effectiveness or cost-benefit analysis) of the diagnostic tools were included. There was no restriction on the perspective (e.g. societal, healthcare provider) or the methodology (e.g. modelling techniques) used, or the time horizon or country where the study was conducted.

3.2.2.5 Study Language and Publication Status

Studies published in all languages were eligible for inclusion. For non-English language papers: a) when the abstract was available, this was translated and screened, and if it appeared to meet the inclusion criteria of the review, the full text of the paper was requested; b) when the abstract was not available and the full text was not in English, the paper was excluded. Only studies published in the peer-reviewed literature were included since these were the sources that were expected to present the most relevant and high-quality results. Any relevant conference abstracts or grey literature (i.e. not peer-reviewed) were also recorded but not included in the evidence synthesis since they were expected to provide little and low-quality information to address the review question. Lastly, the analysis was restricted to studies published after 1990 to focus on more recent evidence.

Table 9: Summary of the systematic review inclusion and exclusion criteria

Criteria	Included	Excluded
Population	-Human male/female -Suspected diagnosis of: Acromegaly; Addison's disease; Congenital adrenal hyperplasia; Cushing's syndrome; Growth hormone deficiency; or Primary aldosteronism	-Animals -Other (endocrine) disease
Interventions	-Any diagnostic test (± treatment) -Any diagnostic and monitoring tool (± treatment)	-Only treatment
Outcomes	-Cost of diagnostic test (± treatment) -Both costs and patient outcomes (e.g. cases diagnosed, life-years gained)	-Only cost of treatment -Only patient outcomes
Study design	-All partial/full economic evaluations -No restriction on perspective, methodology, time horizon or country	-No partial/full economic evaluation
Publication status	-No restriction on language -Only peer-reviewed literature -Published after 1990	-Not available abstract and non-English full text -Conference abstracts -Grey literature -Published before 1990

3.2.3 Identifying Research Evidence

In July 2017, the bibliographic databases Medline® and Embase™ were searched using the Ovid® (Wolters Kluwer, Alphen aan den Rijn, The Netherlands) database search interface (https://www.ovid.com/) since these sources provide up-to-date access to publications in health care. The Cochrane Library (http://www.cochranelibrary.com/) was also searched since it includes publications of a large database of clinical trials as well as health technology and economic assessments. Furthermore, the citations of the included papers were hand-searched using forwards and backwards citation tracking (i.e. articles that cited the papers included and articles that were cited in the papers included) to identify any relevant studies that were not found during the initial search. Specifically, the former was conducted by using the relevant tools that are available on Google™ Scholar (https://scholar.google.com/), PubMed® (https://www.ncbi.nlm.nih.gov/pubmed/), Scopus® (https://www.scopus.com/), and the Web of Science™ (https://apps.webofknowledge.com/), while the latter was performed by manually screening the titles and abstracts of the studies that were referenced in the included papers.

The search strategy (***Appendix 2***) was developed following the PICO framework (176), and included clinical, diagnostic and economic text words and subject terms (e.g. Medical Subject Headings, MeSH; Embase™ Subject Headings, EMTREE). During the search, filters (human participants; publication date) were also applied to obtain only the most recent literature that was most relevant to the review's objectives. The search strategy clinical and diagnostic terms were based on keywords and text terms used in related clinical papers and recommendations from endocrinology experts, while the economic terms were based on search filters from the *Centre for Reviews and Dissemination (https://www.york.ac.uk/crd/)*, the *InterTASC Information Specialists' Sub-Group (https://www.intertasc.org.uk/subgroups/issg/)*, the *Canadian Agency for Drugs and Technologies in Health (https://www.cadth.ca/)*, the *Health Information Research Unit (http://hiru.mcmaster.ca/hiru/)*, and the *Scottish Intercollegiate Guidelines Network for healthcare improvement (http://www.sign.ac.uk/)* (176).

3.2.4 Study Selection, Quality Assessment, Data Extraction and Synthesis

Studies were extracted from the electronic databases into EndNote™ X8 (Clarivate Analytics, Philadelphia, USA) for screening. Titles and abstracts were initially screened against the inclusion criteria to identify potentially relevant papers. The full texts of these studies were then screened and assessed, and the selection of the eligible papers was made. A flow chart (*PRISMA, http://prisma-statement.org/*) was developed to document the study selection process (*Section 3.3*) (176). Data were extracted on the disease group; study design and objectives; type and perspective of the economic analysis; diagnostic strategies compared; location and setting; cost and outcome measures; analytical methods used (e.g. analysis of uncertainty); results; and the authors' conclusions (***Appendix 3***). Afterwards, a preliminary descriptive summary of the characteristics and findings of the included studies was conducted. Given the diversity in study design, patient populations, interventions, outcomes and locations, a structured narrative synthesis was considered to be more appropriate than a quantitative synthesis (e.g. meta-analysis). This was used to describe the differences between studies, comment on their outcome measures, discuss strengths and weaknesses in the existing literature, draw conclusions, and make recommendations, where appropriate.

The included papers were critically appraised using the *Drummond et al. (2005) (17)* 10-item checklist for assessing health economic evaluation studies, while the quality of reporting of the available evidence was assessed using the ISPOR[4] *Consolidated Health Economic Evaluation Reporting Standards (CHEERS)* statement (178, 179), a commonly used 24-item checklist. The assessment focused on the quality of the methods that were used for the EEs that were conducted and the way that these were reported in the papers. Where economic decision models had been used, their quality was examined based on the guidelines for good practice in decision-analytic modelling (DAM) in health technology assessment (*Philips et al. (2004)* checklist) (180, 181).

The study search and selection, the data extraction and synthesis, and the study quality assessment were conducted by two researchers at the University of Bristol, UK (AEC, WH) who worked independently in order to reduce any potential biases. Both reviewers screened

[4] ISPOR: The Professional Society for Health Economics and Outcomes Research (https://www.ispor.org/)

the first 200 titles and abstracts (random selection of papers) to ensure that the results were consistent. Studies were categorised as 'include', 'possibly include' or 'exclude'. The first reviewer (AEC) screened the rest of the titles and abstracts of the papers and the second reviewer (WH) screened those that were considered by the first researcher to potentially meet the eligibility criteria for the review. Afterwards, the first reviewer screened the full texts of the papers selected based on title and abstract screening, and the second researcher read the full texts of the studies that were considered potentially eligible for the review to reach consensus on inclusion/exclusion. The first researcher then extracted the data and assessed the quality of the papers that were included, while the second researcher checked the data for accuracy and completeness. In the cases of disagreement, a discussion between the two reviewers took place in order to reach consensus. No substantial differences of opinion were found.

3.2.5 Initial Screening Results

After screening the first 200 titles and abstracts, both reviewers agreed that only one study should be included in the full-text screening (**Table 10**). The first reviewer (AEC) considered that 33 articles possibly met inclusion criteria, whereas the second reviewer (WH) considered only 15 papers. Most of the discrepancies (n=13) in the 20 papers where there was disagreement occurred because the second reviewer excluded conference abstracts, letters, notes and editorials from the beginning as well as studies evaluating screening in healthy populations instead of diagnosing patients with signs and symptoms.

Table 10: Summary of outcomes after screening the first 200 titles and abstracts

		Reviewer 2*			
		Include	**Possibly Include**	**Exclude**	**Total**
Reviewer 1*	**Include**	1	0	0	**1**
	Possibly Include	0	14	19	**33**
	Exclude	0	1	165	**166**
	Total	**1**	**15**	**184**	**200**

**Reviewer 1: AEC; Reviewer 2: WH*

Table 11: Inter-rater agreement based on initial screening outcomes

Agreement	Expected Agreement	Kappa Coefficient	Standard Error	Z-Score	Probability > Z
Unweighted					
90.00%	77.60%	0.5536	0.0623	8.89	<0.001
Weighted					
95.00%	88.36%	0.5704	0.0621	9.18	<0.001

**Description:* This table shows the reviewers' study selection agreement when screening the first 200 titles and abstracts. The weighted and unweighted Cohen's kappa coefficients are used to measure the inter-rater agreement showing good correlation in the final results (182).*

Table 11 presents the decision agreement between the two reviewers, when screening the first 200 titles and abstracts, by using the unweighted and weighted *Cohen's kappa coefficients (κ)*. Cohen's kappa is a statistic that is used to assess the reliability and variability of two raters/observers when they make decisions regarding categorical data for which the response variable is nominal or ordinal. The definition of κ is:

$$\kappa = \frac{Po - Pe}{1 - Pe} = 1 - \frac{1 - Po}{1 - Pe} \qquad \textbf{(Equation 1)}$$

where p_o is the observed agreement among raters and p_e is the expected agreement among raters due to chance (i.e. hypothetical probability that the raters agree by chance). Depending on the level of agreement, κ can take values from '0' (no agreement) to '1' (complete agreement), but it can also take negative values (≥ -1) when agreement is less/worse than expected by chance. **Table 11** provides both the unweighted and weighted kappa; the latter assigns different weights depending on the extent of disagreement (**Table 12**). During the initial screening, the unweighted κ was 0.5536, while the weighted κ was 0.5704. According to *Landis et al. (1977) (182)*, a kappa statistic between 0.41-0.60 represents a moderate strength of agreement.

Table 12: Weights given for each screening outcome

		Reviewer B		
		Include	**Possibly Include**	**Exclude**
Reviewer A	**Include**	1.0000	0.5000	0.0000
	Possibly Include	0.5000	1.0000	0.5000
	Exclude	0.0000	0.5000	1.0000

3.2.5.1 Modifications in Screening Rules

The two reviewers established additional eligibility criteria before continuing screening. Firstly, conference abstracts, letters, notes and editorials were excluded from the review. Secondly, all papers whose titles and abstracts clearly did not meet the inclusion criteria were excluded even if they listed potentially relevant keywords (e.g. 'cost-effectiveness'). Lastly, if a paper discussed screening healthy populations instead of diagnosing patients with signs and symptoms, this was also excluded since the *ULTRADIAN* study examined a new diagnostic test and not a screening device. After applying these criteria, 18 of the 20 'possibly include' articles were excluded.

3.2.5.2 Further Screening Results

After screening the titles and abstracts of the rest of the studies (4181 records) using the modified rules, the first reviewer ended up with 15 and 163 papers in the 'include' and 'possibly include' categories, respectively. Afterwards, the second reviewer screened again the titles and abstracts of the 'possibly include' category and excluded 140 of 163. At the end, 16 and 22 papers remained in the 'include' and 'possibly include' categories, respectively (overall, 38 records).

3.3 Results

The search strategy identified 5191 studies (905 in Medline®, 3218 in Embase™ and 1068 in the Cochrane Library). From these, 4381 remained after removing duplicate records. After screening the titles and abstracts, 4343 were excluded as they did not meet the eligibility criteria. Thirty-eight (38) full-text articles were assessed for eligibility, from which, 31 were excluded (***Appendix 4***). From these, 28 papers were excluded mainly because they did not include an EE, two because of the diseases that they examined, and one because it discussed the treatment and not the diagnosis of endocrine disease. Therefore, only seven studies were considered eligible for inclusion. Furthermore, the hand-searching of the references of these papers and the articles that had cited them did not identify any other papers that met the criteria. The data from these seven papers were then extracted and their quality of reporting was assessed (**Figure 11**).

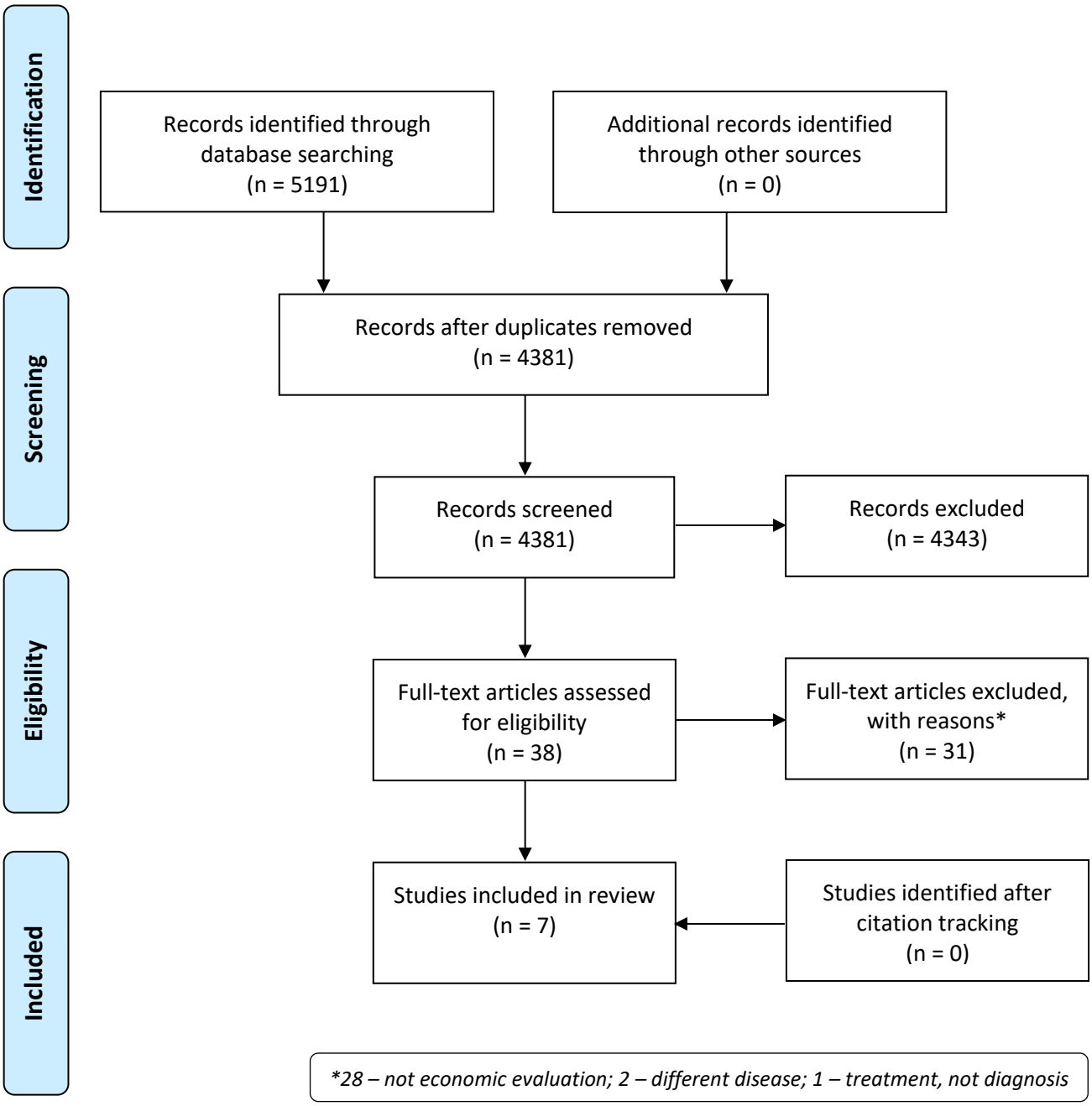


Figure 11: PRISMA flow diagram

3.3.1 Overview of Included Studies

Below, a brief summary of the seven studies that met the eligibility criteria of this review is given (**Table 13**). These studies evaluated a variety of diagnostic and monitoring tools in various patient groups using different outcome measures:

- *Ben-Shlomo et al. (2016) (183)* compared the performance and costs of the cosyntropin stimulation test (CST) for the diagnosis of AD before and after implementing a new electronic medical record (EMR) system protocol.

- *León-Justel et al. (2011) (50)* evaluated and compared the hospital budget impact, reproducibility and diagnostic accuracy of using midnight salivary cortisol (MSVC) instead of midnight serum cortisol (MSC) in the confirmatory diagnosis of CS.

- *Midgette et al. (1995) (51)* compared the costs and cost-effectiveness of using simultaneous bilateral inferior petrosal sinus sampling (IPSS) instead of high-dose dexamethasone (HDD) suppression test followed by IPSS (in patients whose adrenocorticotropic hormone (ACTH) levels did not suppress) in the differential diagnosis of CS from occult ectopic ACTH syndrome.

- *Dekkers et al. (2016) (184)* evaluated and compared the costs and outcomes of managing PA patients after diagnosing and identifying the subtype of the disease with adrenal computerised tomography (CT) compared to bilateral adrenal vein sampling (AVS) preceded by CT.

- *Lubitz et al. (2015) (185)* compared the costs and health benefits of six recommended and/or commonly used screening strategies for PA to identify hypertensive patients with unilateral, surgically correctable hyperaldosteronism (aldosterone-producing adenoma). The authors also compared the abovementioned strategies with treating with medication all patients with resistant hypertension (RH) without testing for PA.

- *Sato et al. (2015) (186)* examined and compared the cost-effectiveness of diagnosing (captopril challenge test, CT and AVS) and treating PA in the hypertensive population to just managing hypertension by medication unless there are clear signs that PA or other complications are present.

- *Velasco et al. (2015) (187)* compared the costs and disease detection rates of using therapeutic drug monitoring (TDM) to a non-selective approach for diagnosing PA in patients with treatment-resistant hypertension (TRH).

Table 13: Summary of included studies

Authors	Year	Title	Location
Ben-Shlomo et al. (183)	2016	Enhanced cosyntropin stimulation test performance enabled by electronic medical record	USA
León-Justel et al. (50)	2011	Budget impact of using midnight salivary cortisol in the diagnosis of hypercortisolism	Spain
Midgette et al. (51)	1995	High-dose dexamethasone suppression testing versus inferior petrosal sinus sampling in the differential diagnosis of adrenocorticotropin-dependent Cushing's syndrome: A decision analysis	USA
Dekkers et al. (184)	2016	Adrenal vein sampling versus CT scan to determine treatment in primary aldosteronism: an outcome-based randomised diagnostic trial	Netherlands, Poland
Lubitz et al. (185)	2015	Cost-effectiveness of screening for primary aldosteronism and subtype diagnosis in the resistant hypertensive patients	USA
Sato et al. (186)	2015	Cost-effectiveness analysis of the diagnosis and treatment of primary aldosteronism in Japan	Japan
Velasco et al. (187)	2015	Cost-effectiveness of therapeutic drug monitoring in diagnosing primary aldosteronism in patients with resistant hypertension	USA

3.3.2 Descriptive Summary of Included Studies

In this section, an extended description of the seven included studies is made to provide more details on the population that was examined, the interventions that were investigated, the analytic methods that were used and the main results that were derived (**Table 14**).

Ben-Shlomo et al. (2016) (183)

This was an observational study conducted between 2013-2015 in the USA that aimed to compare the performance and costs of CST for the diagnosis of AD, in hospitalised patients, before and after implementing a new EMR system protocol. 406 CSTs were performed, 279 before and 127 after implementing the new protocol. Costs were calculated using the unit costs of a single cosyntropin injection and a single serum cortisol measurement. When estimating the total cost of CST, the loss due to incorrectly performed and uninterpretable tests was considered. The authors compared the proportion of CSTs performed correctly; incorrectly, but interpretable; and incorrectly and uninterpretable based on the standard and EMR protocol. The findings of the study showed that correct CST performance improved significantly (16.1% to 53.5%; $p<0.001$), while the number of incorrect CSTs decreased significantly after implementing the EMR protocol. The cost per interpretable test decreased by ≈$500.00 and the authors estimated that $50,414.00 would be saved for every 100 tests performed under the new protocol. The authors concluded that using the EMR system could lead to a more efficient use of staff time and resources.

León-Justel et al. (2011) (50)

This was a diagnostic cohort study conducted in 2009 in Spain that evaluated and compared the hospital budget impact, reproducibility and diagnostic accuracy of using MSVC instead of MSC in the confirmatory diagnosis of CS. The analysis used the healthcare provider perspective. The study involved patients suspected of having CS (n=50) and patients referred to the unit to manage proven CS (n=27), who received both MSC and MSVC tests. To determine diagnostic performance, MSC and MSVC results were compared to the final clinical diagnosis based on all test results (i.e. serum and salivary cortisol levels). To estimate the costs of the two procedures, the direct costs of hospitalisation and confirmatory testing were used (unit costs obtained from a regional pricelist). The findings of the study did not indicate

statistically significant differences in terms of the diagnostic accuracy of the two tests (90.9% for MSVC vs 96.1% for MSC; $p<0.05$). On the contrary, differences in costs were high, meaning that a replacement of MSC by MSVC would lead to cost savings of €21,885.09 to €109,425.27 per year depending on the number of patients that took each test. Results remained robust in sensitivity analysis. The authors concluded that MSVC is a low-cost test with high reproducibility, effectiveness and accuracy that could replace MSC in future clinical practice.

Midgette et al. (1995) (51)

This study was conducted in the USA and used evidence synthesis methods (DAM) to compare the cost-effectiveness of using IPSS instead of HDD followed by IPSS (in patients whose ACTH levels did not suppress after HDD) in the differential diagnosis of CS from occult ectopic ACTH syndrome. The analysis was performed using the healthcare provider perspective. For the model, data on the diagnostic accuracy of the two tests, CS prevalence and mortality rates after treatment were collected from literature, while costs were calculated from hospital charges. By estimating the number of patients alive from 10,000 hypothetical patients who were evaluated and treated, the study results indicated that 18 additional lives and $1,753.00 would be saved if opting for the IPSS strategy. Results were mainly sensitive to the pre-test CS probability, and the tests' characteristics (accuracy) and costs. The authors concluded that IPSS needs to have a high diagnostic accuracy and be followed by low mortality rates to be used initially for the differential diagnosis of CS.

Dekkers et al. (2016) (184)

This was a randomised controlled trial (RCT) that was undertaken between 2010-2015 in twelve Dutch and one Polish medical sites. The study aimed to compare the cost-effectiveness of adrenal CT versus AVS preceded by CT in the diagnosis and management of adult patients with PA. Both diagnostic strategies were followed by treatment with adrenalectomy or mineralocorticoid receptor antagonists (MRAs) depending on the subtype of the disease diagnosed. The study randomised and followed up for one year 184 patients (92 in the CT-group; 92 in the AVS-group). Randomisation was stratified by study centre, and minimised by sex, age, blood pressure and the intensity of drug treatment. Costs were calculated from the time of randomisation until the end of follow-up, while effects (in QALYs) were measured using the SF-6D algorithm for describing health states based on data collected from the SF-36

health questionnaire at three timepoints (baseline, six and twelve months). The findings of the study indicated non-significant differences in the intensity of antihypertensive drug use, biochemical outcomes, adverse events and changes in the patients' health-related quality of life (HRQoL) after one year of follow-up (0.05 QALYs more for the AVS-group). Additionally, AVS-based management had significantly higher costs (€6,746.30) than CT-based management (€4,227.80). Therefore, AVS-based management was unlikely (probability <0.2) to be a cost-effective use of resources. Post-hoc analyses also showed no sensitivity to uncertainty in the final results.

Lubitz et al. (2015) (185)

This study was conducted in the USA and used evidence synthesis methods (DAM) to compare the cost-effectiveness of six screening strategies for identifying unilateral, surgically correctable PA among patients with RH. The screening strategies used tests such as CT and AVS, and were also compared to a strategy of treating all RH patients with MRAs without testing for PA. The analysis used the healthcare system perspective, and compared immediate-after-diagnosis and lifetime costs and benefits for all seven strategies. Costs and probabilities were collected from the literature, while effectiveness was measured in QALYs that were derived from national primary survey data (836 hypertensive patients) based on changes in systolic blood pressure (SBP) applied on a cardiovascular simulation model. This model used utility weights based on EQ-5D utility scores associated with chronic diseases derived from a national sample of cases, and data from the SF-12v1 questionnaire (converted using the SF-6D algorithm) that was administered to a cohort of patients with confirmed PA before and after surgery or treatment with MRAs. Both costs and QALYs were discounted at 3% per annum. The study results indicated that using AVS-only led to the greatest SBP reduction and cost, while treating all patients without previous testing was the least costly strategy but with the lowest SBP reduction. CT-AVS was shown to be the most cost-effective strategy unless unilateral PA prevalence increased above 50%.

Sato et al. (2015) (186)

This study was conducted in Japan and used evidence synthesis methods (DAM) to compare the cost-effectiveness of diagnosing and treating PA (comprehensive diagnostic strategy, CPD) in the over-50-year-old hypertensive male population to managing hypertension by

medication unless there were clear signs of PA or other complications. The analysis was performed using the payer's perspective and investigated lifetime costs and consequences. Input data (i.e. cost, effectiveness and probabilities) were obtained from expert opinion, hospital records, literature and national sources (e.g. national statistics). Bost costs and effects (in life-years) were discounted at 3% per annum. According to the outcomes of the study, CPD increased expected costs and life-years by ¥64,004.00 and 0.013, respectively, and was the most cost-effective strategy. The result was sensitive to medication costs, and changes in screening and treatment outcomes.

Velasco et al. (2015) (187)

This study was conducted between 2009-2014 in the USA and used evidence synthesis methods (DAM) and data collected retrospectively from hospital records to compare the costs and disease detection rates of using a TDM-guided to a non-selective approach for screening PA in TRH patients. In both strategies the same PA screening tests were used. The study used data from 225 patients who were referred to the researchers' hospital. Seventy-eight (78) patients received TDM, from whom 43 were shown not to be adherent to at least one medication, and therefore were not continued to PA screening. The costs used in the analysis were based on the Medicare fee schedule. The findings of the study indicated that using the TDM-approach led to lower costs compared to the non-selective alternative ($1,042.02 vs $1,632.71, respectively) as well as a lower rate of PA detection (11.0% vs 14.8%, respectively) and unnecessary PA screening (35.0% vs 85.2%, respectively). The TDM-approach was also associated with improved medication adherence. Results were sensitive to the prevalence of medication non-adherence.

Table 14: Detailed overview of included studies

Authors & Year	Population & Disease Group	Location(s) & Setting	Study Design	Type of Economic Analysis/Perspective	Intervention(s)/ Comparator(s)	Outcome Measure(s)	Main Results
Ben-Shlomo et al. (2016) (183)	Inpatients diagnosed with Addison's disease	USA (TC)	Observational study (2013-2015); IPD	Cost analysis; Healthcare provider (hospital) perspective [*Not clearly stated*]	Standard CST; EMR-protocol CST; Less validated CSTs	Test performance; Test costs	Improvement of CST correct performance after using EMR: from 16.1% to 53.5%; Cost savings: $50,414.00/100 tests
León-Justel et al. (2011) (50)	Patients with suspected CS; Patients referred to hospital with proven CS	Spain (TC)	Observational study (2009); IPD	Cost and budget impact analysis; Healthcare provider (hospital) perspective	MSC (inpatient); MSVC (outpatient)	Test costs; Test reproducibility and diagnostic performance	Potential annual savings after replacing MSC for MSVC: €21,885.09 to €109,425.27; MSVC showed high reproducibility, effectiveness, accuracy and low cost; Results not sensitive to uncertainty
Midgette et al. (1995) (51)	Patients with ACTH-dependent hypercortisolism: CS or occult ectopic ACTH syndrome	Hypothetical cohort; USA (TC)	Evidence synthesis: DAM	Cost-effectiveness analysis; Healthcare provider (clinician-hospital) perspective	IPSS for ACTH levels in response to corticotropin-releasing hormone; HDD followed by IPSS (in patients with negative HDD)	Test costs; Cost-effectiveness: Cost per life saved	Average survival rate: 0.9878 for IPSS, 0.9860 for HDD-IPSS (number alive/10,000 patients); Average cost per patient: $24,823.00 for IPSS, $23,070.00 for HDD-IPSS; ICER for IPSS: ≈$1,000,000 per additional life saved; Results sensitive to pre-test CS probability, test characteristics (accuracy) and costs; IPSS must have extremely high diagnostic accuracy and low mortality to be chosen

Authors & Year	Population & Disease Group	Location(s) & Setting	Study Design	Type of Economic Analysis/Perspective	Intervention(s)/ Comparator(s)	Outcome Measure(s)	Main Results
Dekkers et al. (2016) (184)	Adult patients with PA confirmed by an oral or intravenous salt-loading test	Netherlands and Poland (TC)	RCT (2010-2015); IPD	Cost-effectiveness analysis; Healthcare provider perspective [*Not clearly stated*]	Adrenal CT- vs AVS- (preceded by CT to determine adrenal vein anatomy) based management	Drug use intensity; Biochemical outcomes; HRQoL (SF-36); Test costs; QALYs (SF-36 via SF-6D algorithm); Cost-effectiveness: Cost per QALY; Adverse events	No significant differences in drug treatment intensity; No significant difference in HRQoL; No significant difference (0.05) in QALYs for AVS-based management; Significant increase in costs: €2,285.00/patient for the AVS-group; ICER for AVS: €45,700.00/QALY; Probability <0.2 that AVS is more efficient when compared to a €30,000/QALY WTP threshold; No differences in adverse events; Results not sensitive to uncertainty
Lubitz et al. (2015) (185)	RH patients (risk of PA)	Hypothetical cohort; USA (TC)	Evidence synthesis: DAM	Cost-effectiveness analysis; Healthcare system perspective	SIT-CT-AVS; CT-AVS; SIT-AVS; AVS-only; SIT-CT; CT-only; MRAs-only with no previous testing	Test costs: Health gains (changes in SBP converted into QALYs); Cost-effectiveness: Cost per QALY (lifetime)	ICER for CT-AVS compared with MRAs-only: $82,000.00/QALY; At a recommended WTP threshold of $150,000/QALY, screening for PA in the RH population is cost-effective in comparison with MRAs-only; Results not sensitive to uncertainty

Authors & Year	Population & Disease Group	Location(s) & Setting	Study Design	Type of Economic Analysis/Perspective	Intervention(s)/ Comparator(s)	Outcome Measure(s)	Main Results
Sato et al. (2015) (186)	Hypothetical 50-year-old male patient diagnosed with stage I-III hypertension	Hypothetical cohort; Japan (TC)	Evidence synthesis: DAM	Cost-effectiveness analysis; Payer's perspective	CPD for PA; SMA to manage hypertension by medication (unless the typical signs of PA or other complications were recorded)	Test costs; Effectiveness (life-years); Cost-effectiveness: Cost per life-years	Expected costs: ¥3,432,339 for SMA, ¥3,496,343 for CPD; Expected life-years: 20.512 for SMA, 20.525 for CPD; ICER for CPD: ¥4,923,385/year (recommended WTP threshold of ¥5-6,000,000 – found in literature); Results sensitive to medication costs as well as changes in the discount rates and outcomes of screening and treatment
Velasco et al. (2015) (187)	TRH patients	USA (TC)	Evidence synthesis; Retrospective collection of hospital records (2009-2014); DAM and IPD	Cost analysis; Healthcare provider (hospital) perspective [*Not clearly stated*]	TDM-guided vs non-selective approach for PA screening	Test costs; PA detection rates; PA prevalence	Average cost per patient: $1,042.02 for TDM-guided PA screening, $1,632.71 for unselective PA screening; Compared to non-selective PA screening, TDM-guided PA screening led to: a) a lower rate of PA detection (11.0% vs 14.8%, respectively); b) a lower rate of unnecessary PA screening (35.0% vs 85.2%, respectively); Results sensitive to the prevalence of medication non-adherence

**Abbreviations: ACTH: adrenocorticotropic hormone; AVS: adrenal venous sampling; CPD: comprehensive diagnostic strategy; CS: Cushing's syndrome; CST: cosyntropin stimulation test; CT: computerised tomography; DAM: decision-analytic modelling; EMR: electronic medical record system; HDD: high-dose dexamethasone suppression test; HRQoL: health-related quality of life; ICER: incremental cost-effectiveness ratio; IPD: individual patient data; IPSS: inferior petrosal sinus sampling; MRAs: mineralocorticoid receptor antagonists; MSC: midnight serum cortisol; MSVC: midnight salivary cortisol; PA: primary aldosteronism; QALYs: quality-adjusted life-years; RCT: randomised controlled trial; RH: resistant hypertension; SBP: systolic blood pressure; SF-36: Short-form 36-item health questionnaire; SF-6D: Short-form six-dimension algorithm; SIT: saline infusion test; SMA: suboptimal management approach; TC: tertiary care; TDM: therapeutic drug monitoring; TRH: treatment-resistant hypertension; USA: United States of America; WTP: willingness-to-pay*

***Symbols: €: Euros; $: United States Dollar; ¥: Japanese Yen*

3.3.3 Narrative Synthesis of Included Studies

3.3.3.1 Patient Population, Setting and Year of Publication

Four studies (184-187) examined the diagnosis and management of PA, two (50, 51) of CS and one (183) of AD, while no study investigating the diagnosis and treatment of AC, CAH or GHD was found. Additionally, from the studies that evaluated the diagnosis of PA or CS, no two studies addressed identical research questions. Four of the studies (51, 183, 185, 187) were conducted in the USA, one (186) in Japan, one (184) in the Netherlands and Poland, and one (50) in Spain. All papers were published after 2011 apart from one (51) that was published in mid 90s.

3.3.3.2 Study Design

Three studies (51, 185, 186) used a hypothetical cohort in their analyses, while three (50, 183, 187) used observational hospital data and one (184) data from an RCT. Moreover, three studies (51, 185, 186) used evidence synthesis methods and included a DAM, three (50, 183, 184) used individual patient data (IPD) in their economic analyses, and one (187) used both DAM and IPD. From the four modelling studies, two (51, 187) used only a decision tree, while the other two (185, 186) used a decision tree followed by a Markov model.

3.3.3.3 Perspective and Type of Economic Evaluation

Five studies (50, 51, 183, 184, 187) used the healthcare provider (hospital) as the perspective for their analysis. From these, three studies (183, 184, 187) did not clearly state that this was the perspective used but this was assumed based on the analysis and results. The remaining two studies (185, 186) used the healthcare system and payer's perspectives, respectively. Four studies (51, 184-186) conducted a cost-effectiveness analysis (CEA), while from the other three studies, two (183, 187) performed a cost analysis, and one (50) a cost and budget impact analysis.

3.3.4 Critical Appraisal of Included Studies

The included papers were critically appraised using the *Drummond et al. (2005) (17)* checklist to examine the quality of the methods used in their EEs (***Appendix 6***):

Ben-Shlomo et al. (2016) (183) clearly described the different diagnostic strategies that were examined. However, they only compared the costs and performance of testing and not its impact on patient management or outcomes. Moreover, the perspective of the analysis was not specified. Although the EMR protocol increased the percentage of correctly performed tests, the findings of the study were not sufficient to prove that the new system would be a more cost-effective option. Furthermore, the study used observational data from one medical centre and did not combine the findings with data from other clinical studies. Despite using a reasonably large sample of 406 CSTs, the authors did not clearly describe the number of tests per patient or the characteristics of the participants. The study did not also give any details about the source of the test cost data, while no sensitivity analysis was performed to assess the uncertainty in the final outcomes.

León-Justel et al. (2011) (50) described in detail the two diagnostic procedures that were examined, but only compared their costs and diagnostic accuracy and not their effectiveness in improving patient management or outcomes. All test accuracy and cost data were obtained from a small sample of 77 participants at a single centre, while in their conclusions the authors also tried to combine their results with those found in the relevant literature. In terms of costs, all the important data were identified for both alternatives, while their source was also defined. Nevertheless, costs were not inflated to the year of the study. Moreover, sensitivity analysis was performed to estimate cost savings under different clinical scenarios, but not how other parameters could affect the final outcomes. Although costs were clearly estimated, to influence decision-making in this area, the authors should have also measured the health benefits derived from the two diagnostic pathways since similarities in their diagnostic accuracy do not necessarily make them equally effective.

Midgette et al. (1995) (51) clearly described the two diagnostic strategies that were compared and the assumptions made for the DAM that was used. For their analysis, the researchers

examined a hypothetical cohort of patients using disease probability, diagnostic accuracy and mortality data collected in a non-systematic way from the literature. Effectiveness data were also obtained from various clinical studies, while costs were calculated using hospital charges for which the actual sources were not reported. Different types of sensitivity analysis were performed for many key parameters to explore the robustness of the results. Nevertheless, key limitations of the analysis were that its time horizon was not mentioned, and costs and outcomes were not discounted to adjust for future values.

Dekkers et al. (2016) (184) described in detail the two diagnostic and management strategies that were compared. Nevertheless, the perspective of the analysis was not specified. In the study, the effectiveness of both strategies was at low risk of bias since data came from a well-conducted RCT, involving a moderately large sample of 184 patients from different medical centres. Effectiveness was measured and compared in QALYs which were derived from a validated and commonly used health questionnaire (SF-36). Furthermore, the researchers identified all relevant costs associated with each strategy and gave more information on the data used in the supplementary material of the paper. Results were compared to those from the PA clinical guidelines. A post-hoc analysis was also performed to explore uncertainty in the final outcomes.

Lubitz et al. (2015) (185) clearly described the seven strategies that were compared and the assumptions made during the analysis. The effectiveness of the different alternatives was established using SBP primary data from a national survey of a reasonably large sample of 836 hypertensive patients to derive QALYs. All healthcare-related cost data were retrieved from the official governmental cost source (Medicare). Effectiveness data mainly came from observational studies, while probability data (e.g. test accuracy) were identified from both literature and a recent meta-analysis of relevant studies. Therefore, the DAM that was used was based on good-quality data. Both costs and consequences seem to have been valued credibly, while they were also discounted using a commonly used 3% rate to adjust for differential timing. Different types of sensitivity analysis were also performed to assess uncertainty in the final decisions. The study results were compared to data from clinical guidelines, other relevant studies and clinical practice.

Sato et al. (2015) (186) described in detail the two strategies that were compared. However, the assumptions made when developing the DAM were not reported. The model used quite weak data. Effectiveness and disease prevalence data were obtained from literature and national statistics, while diagnostic accuracy data were derived from expert opinion, the researchers' hospital records and selected observational studies. Costs were also calculated based on a limited amount of data from ≈50 patients of the researchers' hospital and from literature using non-systematic methods. Both costs and consequences were discounted using a national and commonly used discounting rate (3%). According to the authors, the use of QALYs instead of life-years as an outcome measure might have been more appropriate and meaningful. However, these were difficult to be calculated. The decision about the more cost-effective strategy was made based on a recommended and not a national willingness-to-pay threshold, reducing the credibility of the results. Moreover, sensitivity and scenario analyses were performed to assess the robustness of the results. At the end, results were compared to those from other studies and guidelines for the disease.

Velasco et al. (2015) (187) clearly described the two strategies that were compared but did not state the perspective of the analysis. Cost components were clearly reported in a table, while effectiveness was only expressed in the form of a PA detection rate. Although the study used both data from literature and data collected from a moderately large sample of 225 patients, data for the TDM-approach came only from 35 patients. Additionally, only basic sensitivity analysis was performed to find the key parameters that could affect the decision regarding the optimal strategy. At the end, results were compared to those of other studies and recommendations from clinical guidelines.

3.3.5 Further Quality Assessment of Included Studies

3.3.5.1 Quality of Reporting

The ISPOR CHEERS statement (178, 179) was used to assess the quality of reporting of the included papers. Below, their similarities and differences are described following the order of the different items of the checklist (***Appendix 7***).

3.3.5.1.1 Title, Abstract and Introduction (Items 1-3)

In their titles, all studies apart from two (183, 184) used more specific terms to describe the type of EE that was conducted (e.g. CEA), while all apart from one (183) mentioned the disease that was examined in the paper. Five of the included studies (50, 51, 183, 184, 187) also gave information about the interventions that were compared in the paper. When looking at their abstracts, five papers (50, 183-186) included a more structured summary of the objectives, methods and results of the study, with (50) also giving information on the perspective that was followed. The remaining two papers (51, 187) described only the background/context and provided a brief summary of the results of the study. Lastly, in their introductory sections, the authors of all seven papers described the broader context of the study and presented the research question and aims of the study.

3.3.5.1.2 Methods (Items 4-17)

The 'Methods' section of all seven papers gave information about the population of the study that was examined and how this was chosen. In the cases of the studies that used participants' hospital data (50, 183, 184, 186, 187), the location of the medical units and the way that these data were collected were also reported. Three studies (183, 184, 187) did not clearly state that the healthcare provider perspective was used in the analysis, with the latter giving more information when mentioning that the Medicare fee schedule was used. When it comes to the interventions examined, all studies clearly described the comparators. The time horizon that was used in the analysis was only mentioned in four studies (50, 184-186), while only two papers (185, 186) mentioned using a discounting rate. All the other five studies (50, 51, 183, 184, 187) did not discount their costs and consequences, presumably because they did not track patients beyond the immediate diagnosis and treatment. Nevertheless, it would be

meaningful for the *Midgette et al. (1995) (51)* study to use a discounting rate since it examined treatment and lives saved.

Three studies (50, 183, 187) did not use any measures of health benefit in their analysis, while the remaining four (51, 184-186) used commonly used measures, i.e. QALYs and life-years. To measure effectiveness, three studies (51, 185, 186) used data from the available literature (i.e. synthesis-based estimates), with one of them (51) not providing information on the sources used and assigning an arbitrary disutility value in one of the treatment options. From the remaining four studies, two (184, 187) used data from the hospital in which the study was conducted (i.e. single study-based estimates) and two (50, 183) did not measure effectiveness at all. Additionally, only two studies (184, 185) used methods to elicit preferences for outcomes by using QALYs as outcome measures.

Regarding the estimation of resources used and their respective costs, all seven papers gave information on how costs were calculated for the analysis, without mentioning any adjustments made to approximate opportunity costs. The studies that were conducted in the USA (51, 183, 185, 187) used US dollars as their currency for their analysis, while (50, 184) used Euros and (186) the Japanese Yen. From these studies, only one (186) gave the exchange rate that was used for converting costs into a common currency. The remaining six either did not mention how costs were converted or used the official price lists from their countries. In addition, the price date was only mentioned in two papers (50, 187).

Only four studies (51, 185-187) used DAM in their analysis. All of them clearly described the model used and its pathways, while only three of them (51, 185, 187) stated the assumptions made. As far as the analytical methods used in the studies are concerned, all model studies (51, 185-187) conducted a base-case and sensitivity analyses. From the non-model studies, one (183) conducted only a statistical analysis, while the other two (50, 184) performed both statistical and sensitivity analyses.

3.3.5.1.3 Results (Items 18-21)

In their 'Results' sections, five studies (50, 184-187) gave more details on the input data that were used in their analyses (e.g. ranges, parameters), while from the remaining two, one

(183) only gave some basic information and one (51) no information at all. All studies described the incremental costs and outcomes (if used) that were calculated, and all apart from one (183) used sensitivity analysis to characterise uncertainty in their results. When it comes to exploring heterogeneity, only four papers (51, 184, 186, 187) reported how costs could be affected, but not much detail was given.

3.3.5.1.4 Discussion and Other (Items 22-24)

At the end of the paper, all studies summarised and discussed the key findings of the analyses conducted. However, only six of them (51, 183-187) reported any limitations when analysing data. Lastly, five papers (50, 184-187) gave some information on their source of funding for the study, while only four (183, 184, 186, 187) provided a statement on any conflicts of interest.

3.3.5.2 Good Practice in Decision-Analytic Modelling

The quality assessment of the modelling techniques used in the model-based economic studies (51, 185-187) was undertaken using the *Philips et al. (2004)* checklist (180, 181), a consistent guideline that was developed based on the available evidence on best practice in creating DAMs. This practical framework contains many attributes which can be categorised into three main statements: structure, data and consistency. This order is followed below when conducting a critical appraisal of the models used in the included studies (***Appendix 8***).

3.3.5.2.1 Structure

All four studies clearly stated the decision problem that was investigated in their analysis. This included information about the condition that was evaluated, the patient population that was examined, and the different diagnostic and (when used) treatment pathways that were compared. Additionally, in all four papers, the objective of the EE was consistent with the decision question of the study. All but one article (187) clearly mentioned the perspective of the analysis and conducted a CEA, while (187) performed only a cost analysis. The analyses conducted in the other three studies (51, 185, 186) as well as the input of the data were consistent with the perspective and the overall objective of each study.

Regarding the structure and type of the model, all studies described a series of causal relationships in their models by using a number of diagnostic (and treatment) pathways that were currently used or that were intended to be introduced in clinical practice. In general, the choice of the model was justified by the research objectives of each study. Two studies (185, 186) also combined a decision tree with a Markov model to show the progression of the disease after the diagnosis and treatment. Both studies used an appropriate cycle length for their model health states which was justified by the condition that was examined. All papers apart from (186) gave a clear description and justification of the assumptions made when their models were developed, and all of them provided to some extent the sources of the input data (e.g. costs, probabilities) that were used.

3.3.5.2.2 Data

The methods that were used to identify their input data were relatively transparent in all four studies and the data used were consistent with the objectives of the studies. Data (e.g. costs, probabilities, diagnostic accuracy) were obtained either from the available literature or the hospital experience or both sources, but the quality of the data was not always well-described. Moreover, only two papers (185, 186) used discounting methods in their analyses. The remaining two papers (51, 187) did not mention discounting techniques at all, although this was possibly needed. The sources of the costs were most of the time the hospital prices of the country where the studies were conducted, while the sources of the health benefits of the tests came either from literature or hospital data. Only one study (185) used HRQoL weights by using QALYs in its analysis, and the way that these were calculated was clearly stated in the report. Again, all papers apart from one (186) gave a clear description and justification of the assumptions made when incorporating their data in their models.

Uncertainty in the parameters used and how a change in them would affect the final outcomes were explored in all four studies. Three studies (51, 185, 186) examined more or less analytically how heterogeneity would affect their results, while one paper (185) examined partially the structural uncertainty in its model by adding the costs of an additional test in the diagnostic pathway. Nevertheless, none of the studies conducted a sensitivity analysis to examine methodological uncertainty.

3.3.5.2.3 Consistency

Internal consistency of their models (i.e. how a change in a parameter could predictably change the final outcomes) was examined more or less thoroughly in all four studies using sensitivity analysis (i.e. one- or two-way sensitivity analysis; probabilistic sensitivity analysis; and/or best- or worst-case scenarios). However, no mathematical logic of the models was tested in any of the studies. In order to have external consistency in their results, all four studies described their results and conclusions made based on them. The studies also mentioned which tests other papers or guidelines might consider cost-effective, although they did not directly compare their results to them.

3.4 Discussion

3.4.1 Main Findings

A systematic literature review was conducted to identify the most recent economic evidence on diagnostic and monitoring tools for six selected endocrine disorders (AC, AD, CAH, CS, GHD and PA). Seven studies were included in this review, from which three (51, 185, 186) conducted a CEA using modelling techniques and four (50, 183, 184, 187) used data from hospital records or literature to perform a cost and/or budget impact analysis. Four studies (184-187) examined tools for the diagnosis and monitoring of PA, two (50, 51) of CS, and one (183) of AD, while no studies were found for AC, CAH and GHD. All studies compared different tests using different outcome measures in their analysis. Given the large diversity in the included studies in terms of the diseases and the tests that were examined, only a narrative summary could be conducted to describe their characteristics. From the analysis, it was shown that one study used data based on an RCT (184), three (50, 183, 187) mainly used data collected from hospital records, and three (51, 185, 186) used a DAM based on secondary data. Five of the studies (50, 51, 183, 184, 187) were conducted from the healthcare provider perspective, although not always clearly stated in their reports, while one (186) used the payer's perspective and one (185) that of the healthcare system.

When assessing the quality of the EE conducted and the quality of reporting of the different papers, there were aspects in the *Drummond et al. (2005) (17)* and CHEERS (178, 179) checklists that could not be answered based on the information provided. For instance, some papers (51, 183) did not give much information regarding the resources used and how the costs of the interventions were calculated, some (50, 183, 187) did not measure effectiveness at all, while some of them (50, 183, 185) did not explore heterogeneity in their final results. Regarding the four studies that included modelling techniques in their analysis (51, 185-187), all of them reported and justified most of the information associated with model structure and data identification. Nevertheless, in one study (186), the assumptions made for the model were not reported. All studies used sensitivity analysis to explore parameter uncertainty in their models; however, none of these studies addressed methodological uncertainty. In most cases, the conclusions made in all the included studies appeared able to answer their research questions. Most of the studies tried to compare their findings with information from other papers or clinical guidelines for the diagnosis of the diseases that were examined; however, none of them did that thoroughly. Lastly, all but one of the papers (183) described some limitations in their analyses and results.

3.4.2 Strengths and Limitations

To the author's knowledge, this is the first systematic literature review seeking to identify economic evidence on the diagnosis and management of the six abovementioned disorders. This review followed good practice guidelines for systematic reviews (176) and used a well-informed search strategy to find papers relevant to the research question. The review identified evidence from three comprehensive databases, and final decisions about inclusion and exclusion were made by two researchers in order to limit selection bias.

Nevertheless, there were also several limitations in this review that should be reported. Firstly, although this review was conducted by two researchers, only one reviewer (AEC) read through all the titles and abstracts of the papers. As shown in *Section 3.2.5*, there were several differences between the two reviewers regarding the papers that should be considered eligible for this review. The likely reason for these differences was that the first researcher

(AEC) had less experience of conducting systematic reviews, and therefore was more tolerant when using the eligibility criteria to avoid mistakenly excluding potentially eligible articles. Furthermore, only one reviewer (AEC) screened all the full texts of the studies that were considered to meet the eligibility criteria of the review after the initial screening. This could have led to the omission of relevant papers. Only one reviewer (AEC) extracted all the data, and assessed the quality of reporting and the quality of the DAMs used. As quality assessment is subjective this might have reduced the objectivity of the assessment of study quality.

Moreover, there is currently no standard checklist for assessing the quality of EEs, which means that there is no accurate way to evaluate how well the studies were conducted. For this reason, the *Drummond et al. (2005)* checklist (17), the CHEERS statement (178, 179) and the *Philips et al. (2004)* checklist (180, 181) were all used since they constitute the best available tools for assessing the quality of methods and reporting for EEs. Another limitation is that papers with no English abstracts and full texts were not included in this review. Although this might have reduced the final number of studies that were examined, there was no non-English language study based on the title that seemed able to answer the research question. Lastly, during the screening stages of the papers, several conference abstracts and editorials were found that seemed to meet the inclusion criteria of this review. This means that probably more unpublished economic analyses in this area have been conducted in later years, and therefore more results might become available in the future.

3.4.3 Implications for Clinical Practice and Future Research

Only a small number of economic studies were found for three of the diseases that were examined (AD, CS, PA), all of which evaluated different diagnostic strategies using different and sometimes low-quality study designs. The highest-quality single study was *Dekkers et al. (2016) (184)* that used a relatively small RCT to suggest that AVS-based management was more expensive and not a cost-effective addition to CT-based management in the diagnosis and treatment of PA. *Lubitz et al. (2015) (185)* used good-quality evidence synthesis methods to indicate that the combination of CT and AVS is a cost-effective strategy when screening for PA in the hypertensive population. Small and single-centre observational studies (50, 183)

were also found suggesting that more efficient protocols of currently used diagnostic tests or less invasive tests could potentially improve the cost-effectiveness of diagnosing rare endocrine diseases. Obviously, these observational studies may be prone to bias, and need replication in multi-centre studies and/or studies of a different design (e.g. RCTs). However, these studies can provide valuable initial evidence for diagnostic strategies that are potentially much less expensive than the existing ones. Furthermore, no economic evidence on AC, CAH and GHD was found to help clinicians decide on their most cost-effective diagnostic and management pathway. This is perhaps not surprising given that these are relatively rare disorders. However, it emphasises the need for clinicians in tertiary centres, where patients are diagnosed and treated, to evaluate the costs and outcomes of alternative diagnostic strategies to improve the care of patients in the future.

Given the absence of a reference standard diagnostic test for the six diseases, it is currently difficult to diagnose them and monitor their treatment. Diagnostic uncertainty means that clinicians need to either repeat the same tests or use different tests (sometimes not available in all countries) sequentially or in parallel to reach diagnosis. The latter in combination with the high cost of some of the tests that are currently being used (e.g. invasive) can lead to an inaccurate or belated diagnosis of the disease as well as an increase in waiting times, staff time and the use of healthcare resources. This can be burdensome for the patient and constitutes a problem for the healthcare systems of the different countries, especially where this is publicly funded and the country has a limited budget allocated to health services. Therefore, further research should be conducted in this area and perhaps, the development of a new diagnostic tool or pathway is needed to make sure that patients are diagnosed accurately, on time and without the need of unnecessary resources.

This systematic review informed this thesis in several ways. First, it helped to identify all the EEs of diagnostic tests that have been conducted across all six conditions to examine the amount, type and quality of health economic research that has been performed; understand the medical care that is normally provided; and assist with the decision about which diseases the rest of the analyses (**Chapters IV-VI**) should focus on. This was useful since at the outset of this PhD project, it was unclear in which of these conditions *U-Rhythm* would have the greatest diagnostic potential. Second, the studies found were used to inform the structure of

a DAM developed for PA (**Chapter VI**). More precisely, the model was influenced by the DAMs presented in *Lubitz et al. (2015) (185)* and *Sato et al. (2015) (186)*, which although they used different research questions, PICOs and methods, they followed a similar logic in the way that diagnosis and treatment were represented in their models (i.e. according to the true disease aetiology and progression). Lastly, the data that were provided in the EEs included in this review and relevant citations were used to ensure that the best available data were used in the PA model.

3.5 Conclusion

In conclusion, the findings of this systematic review show the lack of high-quality economic evidence on diagnosing and monitoring AC, AD, CAH, CS, GHD and PA. Seven studies were found that addressed the research question, from which four examined PA, two CS and one AD. Although these studies contained useful information regarding the diagnosis of these diseases, they had a lot of limitations when it comes to how the EE was conducted or reported in the paper.

CHAPTER IV

ULTRADIAN DATA ANALYSIS

DESCRIPTION OF ULTRADIAN PARTICIPANT

DEVICE SATISFACTION, HEALTHCARE RESOURCE USE AND HEALTH-RELATED QUALITY OF LIFE

CHAPTER IV OVERVIEW

Chapter I briefly described the *ULTRADIAN* study mentioning that this was a European multi-centre study that aimed to evaluate a new sampling device (*U-Rhythm*) for the diagnosis and monitoring of six rare endocrine disorders: acromegaly, Addison's disease, congenital adrenal hyperplasia, Cushing's syndrome, growth hormone deficiency, and primary aldosteronism. As part of the study, participants (patients and healthy volunteers) were requested to respond to three questionnaires asking about their satisfaction with the device; their most recent healthcare usage and the impact of their health condition on their work productivity; and their current health-related quality of life. The aim of **Chapter IV** is to describe and compare the results from the individuals' responses to these questionnaires.

Description of ULTRADIAN Participant Device Satisfaction, Healthcare Resource Use and Health-Related Quality of Life

4.1 Background

Chapter I gave a brief overview of the *Dynamic Hormone Diagnostics (ULTRADIAN)* clinical study, providing information about its rationale, aims and design. As mentioned, *ULTRADIAN* (protocol in ***Appendix 9***) was a multi-centre diagnostic accuracy case-control study that aimed to further develop and evaluate an innovative, portable, minimally invasive, 24-hour hormone sampling device (*U-Rhythm*) to improve the diagnosis, management and monitoring of rare endocrine diseases in Europe. To do so, *U-Rhythm* was planned to be tested in patients with acromegaly (AC), Addison's disease (AD), congenital adrenal hyperplasia (CAH), pituitary or adrenal Cushing's syndrome (CS), growth hormone deficiency (GHD), and unilateral or bilateral primary aldosteronism (PA) since results from pilot studies had suggested technical promise in these conditions. Additionally, samples from healthy volunteers were planned to be used as controls. The overall aim of the study was to obtain detailed information on the physiological rhythm of hormones throughout a 24-hour period, compare it with pathophysiological rhythms caused by the six diseases, and define the limits of normality allowing the early and accurate diagnosis and management of endocrine disorders.

Given that *ULTRADIAN* was the first clinical study examining *U-Rhythm* (device and analytical methods) on patients with the abovementioned endocrine conditions (compared to healthy individuals), it was decided that it would be useful to collect and compare some data on the participants' satisfaction regarding the use of the evolving versions of the device. This would help to identify the key areas that should be improved in the future for *U-Rhythm* to be widely acceptable for its users. In addition, given that during the study the device was used within several healthcare systems and across different populations, this was a good opportunity to measure costs related to its use, and its diagnostic accuracy. Moreover, current literature provides little information on the relative impact of these diseases on patient's healthcare usage and general health-related quality of life (HRQoL), with evidence focusing on one or two patient groups (e.g. CS, PA) (188-208) and, in the case of HRQoL, estimating it using

various health measures which are often not validated and/or not sensitive, and yield non-comparable results (209-214). Therefore, it was decided that *ULTRADIAN* was a good opportunity to measure and compare the most recent healthcare resources used and HRQoL between (e.g. CS vs PA vs healthy) and within (e.g. PA pre- and post-operatively) diseases, and test the validity of a widely used health questionnaire (i.e. EQ-5D-5L) within these groups. It was hoped that this preliminary evidence on the costs and diagnostic accuracy of the *U-Rhythm* device and the characteristics of patients would inform the parameter estimates for the early economic evaluation (EE) model reported later in this thesis (**Chapter VI**).

The aim of **this Chapter** is to analyse and compare the results that were produced from the responses of the *ULTRADIAN* patients and healthy volunteers to three questionnaires asking about their satisfaction after using *U-Rhythm*, their most recent (within three months before sampling) healthcare usage and the impact of their health condition on their productivity at work, and their current (before sampling) HRQoL. Knowledge of these elements is an essential prerequisite for estimating the potential cost-effectiveness of *U-Rhythm* as a new method for diagnosing these diseases, and guiding and monitoring their therapy (**Chapter VI**); and providing evidence for policy makers, clinicians and patients on its potential value in the area.

4.1.1 ULTRADIAN Study Information

4.1.1.1 U-Rhythm Rationale

Diagnosing rare endocrine disorders with conventional methods is often difficult, inaccurate and time-consuming due to the short half-life of hormones (i.e. time until they lose half of their physiological activity), and the inability of existing tests to effectively assess dynamic changes in hormone levels over the day. Current procedures predominately rely on single hormone analyses from blood samples that are taken in the hospital setting during office hours, and most often early in the morning, followed by a second stage of dynamic testing of samples. The problem with these tests is that they usually have high false positive or negative results and require the patient to be admitted to hospital (41-46). The latter factor can sometimes lead to stress-induced hormonal changes which in combination with other

parameters (e.g. age, sex, food, time of day) can complicate the interpretation of results. Therefore, the development of an (automated) ambulatory test (e.g. *U-Rhythm*) that patients can use in their own environment without changing their daily routine could be a potential solution. Specifically, *ULTRADIAN* hypothesised that the performance of a 24-hour sampling of hormone profiles of each individual could potentially reveal abnormal levels and disorders of hormone rhythmicity much earlier in the course of the disease or when the disease is present in a mild form, in less time and with higher accuracy. Additionally, the 24-hour sampling could be employed after definitive treatment (e.g. surgery) to evaluate its success and to identify disease relapses at an early stage, or to individualise treatment. Overall, it was anticipated that *U-Rhythm* would simplify the diagnostic procedure, improve treatment, and decrease the morbidity and mortality in endocrine patients.

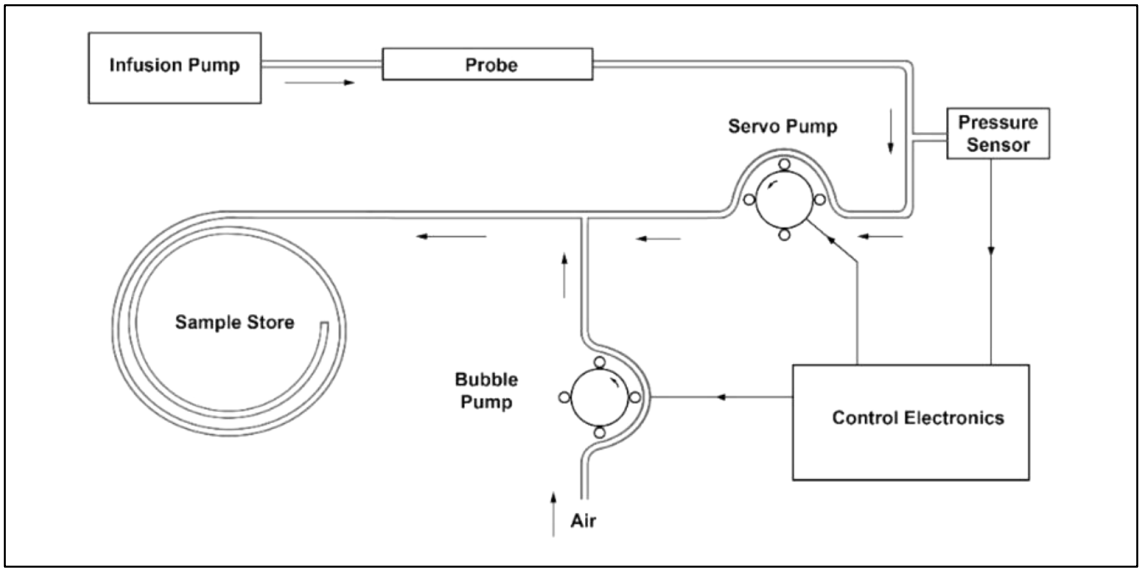
4.1.1.2 U-Rhythm Components and Analytical Techniques

U-Rhythm (**Figures 12-13**) is an automated ambulatory diagnostic/monitoring device that was conceptualised, developed and initially used at the University of Bristol (United Kingdom) to study in detail the physiology and pathophysiology of the hypothalamus-pituitary-adrenal axis[5] rhythmicity. The device is currently under development by an external manufacturer (Designworks, https://www.designworks.studio/). *U-Rhythm* is a collection system that can easily be attached to a belt allowing users to continue with their daily activities (including sleep) whilst undergoing sampling. The device uses a unique *'microdialysis'* technique, in which a sterile catheter (*'microdialysis probe'*) of <1mm diameter, made of a semi-permeable membrane with tiny 'pores', is subcutaneously injected at the lower abdominal area of the patient without the need for local/topical anaesthetic. In *U-Rhythm*, the probe is connected to an external infusion pump which perfuses fluid through the microdialysis catheter and into the *U-Rhythm* collecting device (into an internal fraction collector, *'spool'*). The procedure, including connecting the patient to the sampling unit, lasts ≈30-60 minutes and is performed by trained healthcare professionals. Microdialysate fluid is collected and time fractionated (range between 10-80 minutes) for 24 hours. Separation of samples in the spool is achieved with the injection of air bubbles by a peristaltic pump as the column of microdialysate moves

[5] A complex set of organs and glands that is related to the release and function of hormones (39, 40, 215).

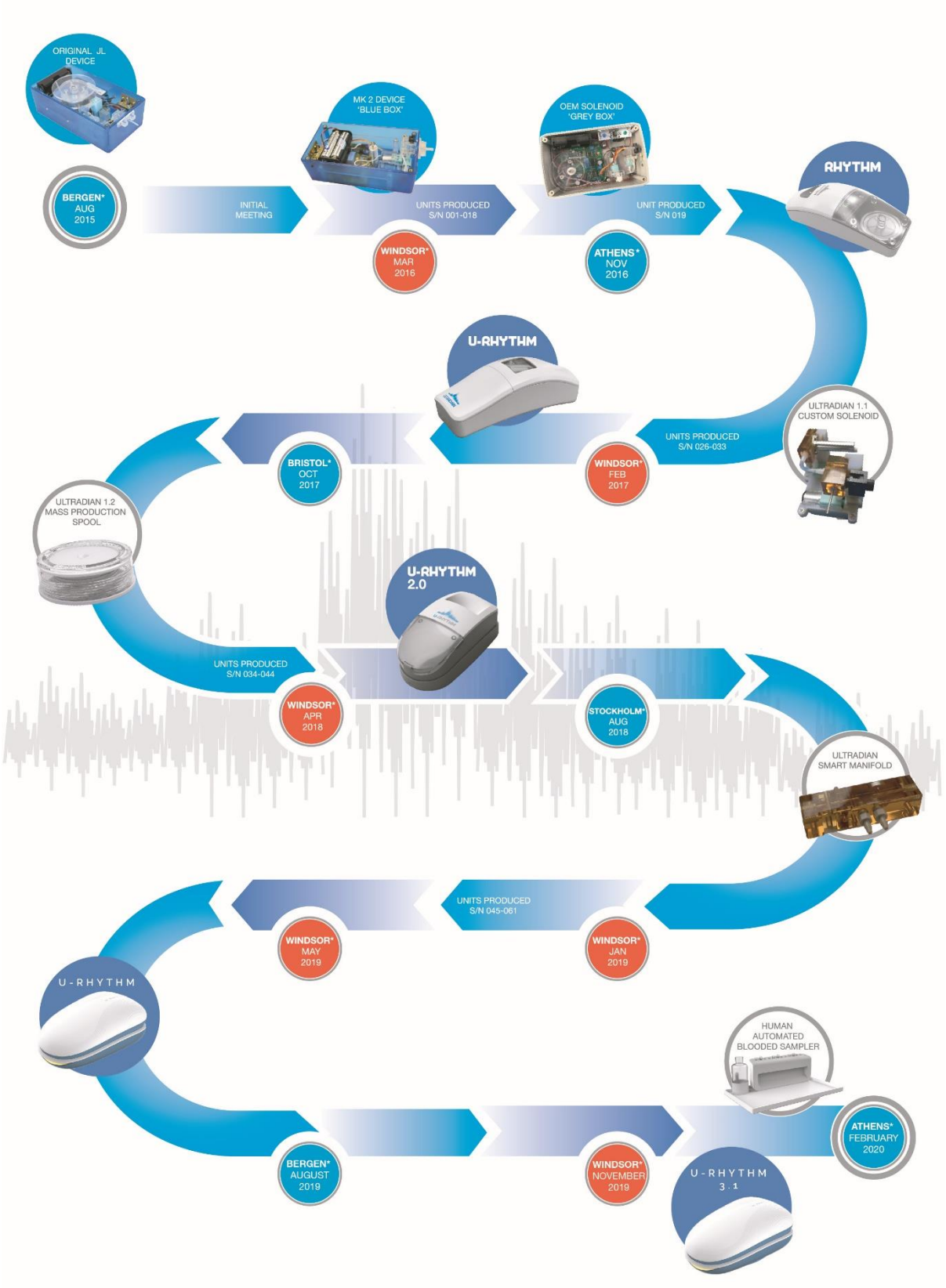
along the tubing (i.e. as new samples are introduced). The current device works with rechargeable batteries that last longer than the 24-hour collection period.

When the 24-hour sampling is completed, the patient is asked to return to the clinical site to have the spool removed, the device cleaned and the samples sent for analysis in the laboratory. In *ULTRADIAN*, disease-related hormones, proteins and peptides were analysed using *'ultrasensitive liquid chromatography tandem mass spectroscopy (LCMS/MS)'*, a method developed at the University of Bergen (Norway) for the analysis of steroid hormones, and the *'proximity extension assay'*, a method developed at the Karolinska Institute (Sweden) that allows the simultaneous assay of up to 96 analytes in only 1µL fluid (**Figure 14**). The latter method was performed by OLINK (www.olink.com) using Proseek® Multiplex immunoassays but did not provide useable results for this study due to technical reasons. Therefore, it was decided that this method will not be used in the future for the analysis of samples.



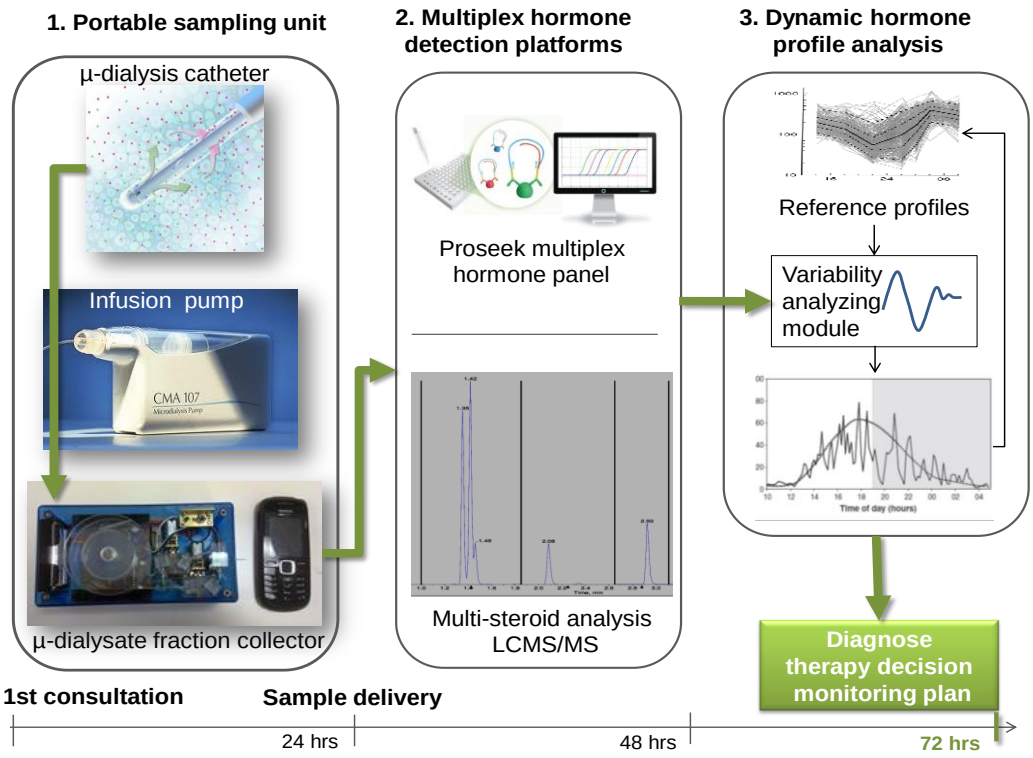
[**Source:** Figure created by Designworks. Permission to reproduce this figure in this thesis was given by Robin Crossley (Head of Engineering, Designworks)]

Figure 12: Diagram of the microdialysis sampling system



[**Source:** Figure created by Designworks. Permission to reproduce this figure in this thesis was given by Robin Crossley (Head of Engineering, Designworks)]

Figure 13: The U-Rhythm device (at the beginning, during and at the end of ULTRADIAN)



[**Source:** Figure created by *ULTRADIAN*. Permission to reproduce this figure in this thesis was given by Stafford Lightman (Professor of Medicine, University of Bristol)]

Figure 14: The U-Rhythm dynamic diagnostic system

4.1.1.3 Participant Recruitment

ULTRADIAN involved four European clinical sites (University of Bergen, Norway; Karolinska Institute, Sweden; University of Bristol, UK; and the Evangelismos General Hospital, Greece). After obtaining appropriate ethical and local approval for all study procedures by the relevant health research regulatory bodies, each centre was responsible for the enrolment of healthy volunteers (Part A) and endocrine patients (Part B) throughout the study duration (i.e. years 2016-2020). The healthy cohort was divided into three subgroups depending on the investigation that took place, while there were six endocrine condition patient subgroups (**Table 15**). Recruitment aimed to include individuals of an appropriate variation in age and gender. However, due to the rarity of the diseases this was harder for the patient cohorts. Additionally, due to demographics and local disease prevalence, it was expected that it would

be harder for some centres to enrol participants for some of the patient subgroups compared to others.

Table 15: Planned ULTRADIAN participant subgroups

	Subgroup Name	Study Centre	Target Number of Participants
Part A Definition of the normal circadian and/or ultradian profiles of pituitary and adrenal hormones in healthy volunteers	**1A** Normal 24-hour circadian and/or ultradian profile across age and sex	Athens, Bergen, Bristol, Stockholm	200 (≈50 from each centre)
	2A Day to day variability	Athens, Bergen, Bristol, Stockholm	20 (≈5 from each centre)
	3A Comparison of tissue and blood concentrations of hormones	Bristol	20
Part B Definition of the circadian and/or ultradian rhythm in patients with pituitary and adrenal disorders	**1B** Diagnosis of Cushing's syndrome by *U-Rhythm* dynamic cortisol measurements	Athens, Bergen, Bristol, Stockholm	40 (≈5 pituitary and ≈5 adrenal from each centre)
	2B Monitoring of Addison's disease by *U-Rhythm* dynamic cortisol and adrenocorticotropic hormone measurements	Bergen, Bristol, Stockholm	20 (≈6-8 from each centre)

	Subgroup Name	Study Centre	Target Number of Participants
	3B Monitoring of congenital adrenal hyperplasia by *U-Rhythm* dynamic cortisol, adrenocorticotropic hormone and androgen measurements	Bergen, Bristol, Stockholm	20 (≈6-8 from each centre)
	4B Diagnosis of primary aldosteronism by *U-Rhythm* dynamic aldosterone and renin measurements	Athens, Bergen, Bristol, Stockholm	30 (≈7-8 from each centre)
	5B Diagnosis of acromegaly by *U-Rhythm* dynamic growth hormone measurements	Athens, Bergen, Bristol, Stockholm	20 (≈5 from each centre)
	6B Diagnosis of growth hormone deficiency by *U-Rhythm* dynamic growth hormone measurements	Athens, Bergen, Bristol, Stockholm	20 (≈5 from each centre)

[**Source:** Table developed based on a table presented in the *ULTRADIAN* protocol]

4.1.1.3.1 Healthy Cohort

The study involved adult healthy individuals (aged 18-68 years) who could understand English or the local centre language. ***Appendix 9*** provides more information on the eligibility criteria. Recruitment was facilitated by local (i.e. within the University/hospital) advertisement via emails, telephone calls and/or in print. Interested individuals contacted the research team. Potentially eligible individuals were provided with a healthy volunteer's information sheet either personally or electronically, where possible. Participants had the opportunity to discuss

any part of the study at any point in time. Volunteers were compensated for their travel expenses and time away from work, and were free to withdraw from the study at any time.

4.1.1.3.2 Patient Cohorts

The study enrolled adult endocrine patients (18-68 years old) at any stage in the course of their disease (i.e. from recently diagnosed to already receiving treatment for some time) who could understand English or the local centre language. ***Appendix 9*** provides full information on the eligibility criteria for each patient subgroup. Each centre recruited patients either from its local patient databases or outpatient clinics. Recruitment was conducted by sending an invitation letter in print, by email or through the outpatient clinic. Patients were provided with a patient information sheet and the study local phone number/email. Interested individuals contacted the research team. A face-to-face consultation was then offered to discuss study details and any queries. Patients had at least one week to consider participation. During this time, they were encouraged to discuss any concerns with their routine caregivers (e.g. general practitioner, specialist physician). If surgery was required promptly (e.g. PA, CS) and the patient was keen to participate in the study, a decision was exceptionally made within a shorter time period. Patients were contacted at the end of that week if their final decision had not been received before that time. Patients were compensated for their travel expenses and time away from work, and were free to withdraw from the study at any time.

4.1.1.4 Screening and Sampling Procedures

Each participant was invited for a screening visit. This appointment occurred in the morning and the individual was asked to arrive fasting (i.e. no food or drink except for water after midnight). During the visit, the participant's informed written consent was signed, details of the study were explained, and any questions were answered by a study investigator. For participants who met the screening criteria, demographic characteristics and medical history were recorded/retrieved through an interview, and height, weight, waist-hip ratio and blood pressure measurements were taken. In healthy volunteers, a blood sample was also collected and analysed locally unless valid results were available within the last two months. Breakfast was then supplied. Participants were asked to abstain from alcohol and smoking for 48 hours before and during sampling, while they were also advised to maintain a regular bedtime and avoid strenuous physical activity for the same period. Furthermore, individuals were given an

activity diary to complete the day prior to and during the sampling period, and important information about the study (e.g. appointment dates, sampler handling instructions).

Sampling commenced at a subsequent pre-arranged date. Participants were asked to arrive fasting, and height, weight, waist-hip ratio and blood pressure measurements were taken. Blood was collected to establish a baseline hormone and biochemistry/haematology profile. In healthy volunteers, a urine sample was also taken to test for drugs of abuse, while females of childbearing age underwent a pregnancy test. Breakfast was then given. Although local anaesthetic was not used in Bristol, at the other centres, one hour prior to the microdialysis catheter insertion, participants either used local anaesthetic cream or local anaesthetic was injected subcutaneously at a localised site on the abdomen. The sterile probe was then inserted using aseptic precautions and *U-Rhythm* was attached. Afterwards, participants were free to return to their daily activities.

Participants were informed that they should not shower whilst wearing the device. Healthy volunteers and CS/AD/CAH patients also received three salivette tubes with instructions for saliva collection (at around 11pm or just before bedtime; after waking up; 30 minutes post awakening). Participants were provided with a telephone number in case of any problems and were contacted after 12 hours by a researcher to check their condition. At the end of the sampling period, *U-Rhythm* was disconnected and a plaster was applied at the insertion site. All procedures (apart from sampling) and appointments took place in the Endocrinology department, outpatient clinic or research unit and were undertaken by trained healthcare professionals. Apart from subgroup 1B, in which four sampling sessions could be needed, in all other cases, individuals participated up to a maximum of three occasions unless sampling issues or treatment adjustments (for patients) were reported. In these cases, sampling was repeated with the patient's consent. If the patient's routine care involved a surgical procedure, they were asked to participate once pre-operatively and then up to a maximum of two times post-operatively (2-3 months, 6-12 months).

4.1.1.5 Data and Sample Collection and Handling

The primary endpoint of the study was to obtain 24-hour hormone profiles in all participants. Secondary endpoints varied between the subgroups (***Appendix 9***). All data (e.g. screening

information; sampling details; time and reason of participant discontinuation) and samples were collected, used, stored and disposed in accordance with the local and international good clinical practice guidelines and data protection regulations (216-218). Data/Samples were also kept anonymised and were labelled using a unique centre and study identification number. All study-related data were recorded using a case report form and were entered into a local and a study database. All data and study results were uploaded to a central server (University of Bergen) for analysis.

4.1.2 Work Declaration

Before moving to the main analysis, it should be mentioned that the author of this thesis did not contribute to any of the procedures described in *Section 4.1.1*. Specifically, the concept and design of the *ULTRADIAN* study was agreed before the start of this PhD project. All device development, participant recruitment, data collection, sampling and sample analyses were performed by other *ULTRADIAN* partners. The author only contributed to the design of the device satisfaction and healthcare resource use questionnaire, and the selection of the HRQoL measure which were completed by the participants, and the analysis of these data. Moreover, the author assisted with the estimation of the device and sampling costs (***Appendix 10***), while he discussed with the statisticians involved in *ULTRADIAN* the diagnostic accuracy information (***Appendix 11***) that he needed for the early economic model that is analysed in **Chapter VI**.

4.2 Methods

4.2.1 Study Objectives

The aim of the present study is to explore the variation in the results that were yielded from the *ULTRADIAN* participants' responses to three questionnaires: i) Device satisfaction; ii) Use of healthcare and impact of health on work; and iii) the EuroQol EQ-5D-5L (219). To do so, the demographic and clinical characteristics of the *ULTRADIAN* participants at screening date, information on their most recent screening and diagnostic testing (for patients only), and their sampling details, issues and results are first described and compared. The responses to

the three questionnaires are then analysed to identify any issues that the participants had when using the *U-Rhythm* device, and the impact of the examined endocrine disorders on the patients' healthcare use/needs, productivity at work and HRQoL compared to healthy individuals. The key parameters that were likely to affect participants' responses to the questionnaires and HRQoL are also investigated.

4.2.2 Study Population

The present study uses data from the majority of the *ULTRADIAN* participants who underwent sampling at least once between October 2016 and October 2019. Data for some participants were not available on the central database at the time that this analysis was conducted (date of data extraction: 1 November 2019). In this study, participants are grouped into two wider cohorts: a) healthy volunteers (controls), and b) endocrine patients (cases). The patient group is further divided into patients with PA, CS or AD. Here, it should be highlighted that only a few patients with AC, CAH or GHD were recruited in *ULTRADIAN* due to technical difficulties in analysing samples from these patients. In addition, the duration that the patients had their condition for was not collected. Therefore, these data are not included in this analysis.

4.2.3 Study Questionnaires

Three questionnaires were provided to each participant at each *U-Rhythm* sampling session: i) Device satisfaction; ii) Use of health care and impact of health on work; and iii) the EQ-5D-5L (health status) (219). The first two questionnaires were co-developed by the author of this thesis and the *ULTRADIAN* partners. These are two non-validated questionnaires generated for the purposes of this study, and after considering the characteristics of the device and the healthcare resources that are commonly used in the examined conditions. The EQ-5D-5L was developed by the EuroQol Group and is a standardised well-validated preference-based instrument for measuring generic health status (HRQoL) (219). Participants were asked to complete the 'device satisfaction' questionnaire after sampling, whereas the other two questionnaires were completed before sampling. All questionnaires were administered in the native language of the recipient. Below, a brief description of each questionnaire is provided.

4.2.3.1 'Device Satisfaction' Questionnaire

As seen in **Figure 15**, the 'device satisfaction' questionnaire consisted of eight scale questions that asked the respondent to rate, from '1' to '5', how much they agreed or disagreed with the following statements for *U-Rhythm*: 1) comfortability of the sampling probe; 2) easiness of carrying the device; 3) ability to continue with their usual daily activities; 4) experience of pain from the sampling probe; 5) problems with noise from the sampler during the day; 6) easiness to sleep when wearing the device; 7) disturbance during sleep due to noise from the sampler; and 8) whether the user would be happy to wear the device again, if asked. The questionnaire also included a final free-text question in which the respondent could provide any other comments on their experience with the device.

4.2.3.2 'Use of Health Care and Impact of Health on Work' Questionnaire

The 'use of health care and impact of health on work' questionnaire (**Figure 16**) included six questions. The first three questions asked the participant about their use of health care over the last three months before sampling. The respondent provided the number of nights that they were admitted to hospital (if any), and the number of times that a specialist or other doctor/nurse were visited (if applicable). If this could not be estimated, the respondent had the option to check the 'Not known' box. For the same time period, the next three questions asked the participant about the impact that their health condition had on their working times and responsibilities[6]. The respondent reported whether they were currently employed and if yes, they provided the number of days that they missed from work (if any and if known) and rated how much their health problems had affected their productivity at work on a scale from '0' (no effect) to '10' (completely prevented from working).

4.2.3.3 EQ-5D-5L Questionnaire

4.2.3.3.1 EQ-5D-5L Description

The EQ-5D comprises two parts (219): a) a short (one-page) descriptive system, and b) a visual analogue scale (EQ VAS). The first part provides a health profile for the respondent which can be converted to a single summary index value (utility score) using an algorithm. The second

[6] Questions were adapted from the 'Work Productivity and Activity Impairment' questionnaire (220).

part asks the individual to rate their own overall current ('today') health on a vertical scale from '0' (worst health) to '100' (perfect health). Results can then be used to calculate quality-adjusted life-years (QALYs) to inform EEs. In this analysis, the EQ-5D-5L (latest) version of the EQ-5D instrument was used since this has been reported to be more reliable and sensitive compared to the previous version (EQ-5D-3L) (219, 221-227). Its use in the *ULTRADIAN* study was registered with the EuroQol Group (219). The EQ-5D-5L descriptive system consists of five health-related dimensions: 1) mobility; 2) self-care; 3) usual activities; 4) pain/discomfort; and 5) anxiety/depression. Each dimension is further divided into five levels: i) no problems; ii) slight problems; iii) moderate problems; iv) severe problems; and v) unable to/extreme problems (219). Participants were asked to indicate their current health state by checking the appropriate level-box of each one of the five dimensions. The EQ VAS was not used in every centre and is not reported in this study since it is less commonly used in health economic research due to concerns that it is inferior to preference-based methods (228).

In EQ-5D-5L, each answer to the five dimensions is coded as a single-digit number expressing the severity level selected (e.g. '3' means 'moderate problems'). The digits obtained from all responses are combined to form a five-digit number that represents the respondent's health state (e.g. 12134)[7]. In *ULTRADIAN*, this number was converted into the EQ-5D index value using the *van Hout et al. (2012) (229)* 'crosswalk' algorithm, a value set[8] that was developed for the EQ-5D-3L using respondents from six European countries, including the UK, and was then mapped to fit the EQ-5D-5L descriptive system. Given the similarities in the utility scores produced by European populations, the same algorithm was used for calculating utilities for all countries. The 'crosswalk' algorithm was preferred to the actual EQ-5D-5L value set since it is currently recommended by the National Institute for Health and Care Excellence (NICE) (230). Since the aim of the study is to present and compare the utility scores associated with each health condition (i.e. healthy, PA, CS, AD) at one point in time, no QALYs were calculated.

[7] These numbers are labels used to describe a health profile and have no arithmetic properties (e.g. 23451 is not better than 12156). A total of 3,125 possible health states can be defined (219).

[8] Value sets provide weights for each EQ-5D health state which are produced by asking a representative sample of the general population of the examined country/region to place a value on it. The standardised valuation is performed by using different techniques (e.g. visual analogue scale; time trade-off valuation) (17, 219).

4.2.3.3.2 Rationale for Using EQ-5D-5L in ULTRADIAN

The EQ-5D (219) is widely used in research, and is recommended by several health technology assessment bodies (including NICE (117, 230)) for measuring the impact of new diagnostic and therapeutic technologies on patient health. This is because it is easy to administer; is available in both paper and digital versions; is cognitively undemanding and can be self-completed by the respondent within a few minutes; has been developed for both child and adult populations; is standardised and well-validated; is free to use in research (non-commercial use); and has been translated into most major languages (including those used in *ULTRADIAN*). Moreover, its results can be used to describe, value and compare health states across a wide range of diseases and populations; and to easily calculate QALYs. Lastly, the fact that the utilities produced represent the preferences of the general population of a country/region means that the societal perspective is accounted for, which is important when informing health policy and making resource allocation decisions.

Nevertheless, the EQ-5D has several limitations. Firstly, research has shown that its sensitivity to detect health status changes varies depending on the condition. Additionally, the fact that it is so simple may lead to *ceiling effects*[9], leaving little room for health improvement. Responsiveness to clinical change is a critical property for every instrument used in EEs since its absence could result in incorrect conclusions regarding the health economic value of the examined interventions (231-234). Secondly, the EQ-5D includes a small number of broad health-related dimensions. However, some conditions may affect other HRQoL areas that are not measured (e.g. dexterity, cognition, wellbeing), which raises questions about the extent to which it represents all aspects of the pragmatic HRQoL of patients (231, 233-238). Thirdly, unrepresentative utility scores may also be produced because the HRQoL valuation is performed by asking the general population to state their preferences for each health state. In practice, patients may place different emphasis on each health aspect and may be able to adapt to some problems that the public is difficult to imagine (e.g. pain/discomfort, mobility). Similarly, it may be hard for patients with chronic health problems to imagine having their symptoms alleviated. The EQ-5D does not also capture the benefits of an intervention for the patients' families/carers (231, 233, 234, 239, 240).

[9] A lot of participants scoring highly (i.e. 'best' health), despite seeking health care.

The issues described above are common problems of generic preference-based instruments, and this is why disease-specific questionnaires are often preferred as effectiveness measures. However, the latter are not typically designed to allow comparability across different health conditions and produce preference-based HRQoL weights for EEs. Therefore, they need to be mapped to a preference-based measure (often the EQ-5D) to do so (232, 233, 235-237). This requires a degree of overlap between the descriptive systems of the two measures, while a mapping algorithm should be available and validated in large clinical trials to avoid likely uncertainty/error in the HRQoL measurement (236, 237, 241). This is typically not the case, especially in rare diseases, where limited research with small and heterogenous samples has been undertaken. The mapping process may also ignore some significant dimensions of the disease-specific questionnaire, causing the latter to lose its value and purpose (233, 241). In addition, not all diseases (e.g. CAH) have an existing and translated measure, which makes the use of a generic instrument inevitable (233). Despite the limited evidence on the appropriateness and validity of EQ-5D on the conditions examined in *ULTRADIAN*, research has indicated that in most diseases, its results have a good correlation between deteriorations in health and lower utility scores (231, 232). The EQ-5D has also been responsive to clinical change when the treatment effect is large (e.g. surgery) or in moderate-to-severe conditions (231, 232). Moreover, its dimensions have been proven to incorporate different aspects of health (e.g. mental health), while attempts to add more domains (e.g. fatigue) had little impact on utilities (234). Finally, the purpose of *ULTRADIAN* is to provide some additional evidence on the HRQoL of the examined patient groups, and use a common way to measure and compare outcomes across health conditions. Therefore, the EQ-5D-5L was preferred for this study compared to other generic or disease-specific questionnaires (209-214, 242).

Dynamic hormone diagnostics (ULTRADIAN) *REC Number:*

Local Subject ID		International subject ID	

ULTRADIAN

Thinking back over the time using the sampling device, please read the following statements and rank your response on the scale:

	Strongly disagree	Disagree	Undecided	Agree	Strongly agree
1. Having the sampling probe fitted was not too uncomfortable	1.	2.	3.	4.	5.
2. Carrying the sampling device was easy	1.	2.	3.	4.	5.
3. I was able to continue with my normal daily activities	1.	2.	3.	4.	5.
4. I did not experience any pain from the sampling probe	1.	2.	3.	4.	5.
5. During the day I was not bothered by noise from the sampler	1.	2.	3.	4.	5.
6. It was easy to sleep wearing the sampling device	1.	2.	3.	4.	5.
7. During sleep, noise from the sampler did not disturb me	1.	2.	3.	4.	5.
8. I would be happy to wear the sampler again if I was asked	1.	2.	3.	4.	5.

Do you have any comments about your experience wearing the sampling device?

Researcher completing form (capitals) _____

Signature _____ Date d d / m m / y y y y

Version 1.1/04/04/2017 *Page 2 of 2*

Figure 15: 'Device satisfaction' questionnaire

Local ID		International ID	

Dynamic Hormone Diagnostics (ULTRADIAN)

Use of health care and impact of health on work

The following questions ask about your need to use any health care resources because of any health problems that you had over the last three months and the impact that your health condition had on your working times and responsibilities.

Please fill in the blank spaces that are provided or check the box with the answer that best applies to your case.

1. In the **last three months**, how many nights have you spent in hospital?

   _____ nights ☐ Not known

2. In the **last three months**, how many times have you seen a specialist doctor?

   _____ times ☐ Not known

3. In the **last three months**, how many times have you seen another doctor or nurse (for example a GP)?

   _____ times ☐ Not known

4. Are you currently employed (working for pay)? If **not**, please check the '*No*' box and leave the rest of the questions blank.

   ☐ No ☐ Yes

5. If yes, in the **last three months**, how many days did you miss from work because of any health problems?

   _____ days ☐ Not known

6. If yes, in the **last three months**, how much did health problems affect your productivity while you were working? Please circle the number that you think that best applies to your case.

Health problems had no effect on my work	0 1 2 3 4 5 6 7 8 9 10	Health problems completely prevented me from working

Figure 16: 'Use of health care and impact of health on work' questionnaire

4.2.4 Analysis

This study includes a descriptive comparison and a regression analysis. Both analyses were conducted using the statistical software Excel 365 (Microsoft, Washington, USA) and STATA® MP 15.1 (StataCorp, Texas, USA).

4.2.4.1 Descriptive Analysis

A descriptive analysis is performed to examine and compare the number of participants that were involved in each *ULTRADIAN* cohort (i.e. healthy volunteers, PA, CS, AD); their main demographic and clinical characteristics at screening date; their number and type of screening, diagnostic and monitoring tests that they had received before sampling (for patients only); and their sampling details, issues and results. Specifically, demographic and clinical characteristics include information on the participant's gender, age, smoking status, body mass index (BMI), and disease cause (when applicable). Screening and diagnostic testing contain information on the tests and procedures that were provided to each patient group. Sampling data comprise information on the participants' blood pressure measurement before sampling; the number of samples taken and the sampling timepoint(s) (i.e. at baseline, pre- or post-operatively); the number of times that each version of *U-Rhythm* (Version 1.0 or 2.0) was used and whether users had adverse skin reactions or other issues; and the microdialysis results per device version. Moreover, this analysis examines and compares the proportion of each cohort's participants who responded to the three questionnaires described in *Section 4.2.3* and their average responses. Where a participant underwent several sampling sessions, only the questionnaires completed at the first sampling session are reported for simplicity. The only exception to this is that for PA and CS patients who had both pre- and post-operative data, in which case pre- and post-operative EQ-5D-5L utility scores are also compared.

For continuous variables, sample means, standard deviations and minimum-maximum value ranges are calculated, while cohort differences are explored using the F-test since multiple groups are compared. For categorical variables, data are summarised using proportions and percentages, while group differences are estimated using the Pearson's chi-squared test. In both cases, statistical significance is set at a p-value of 0.05. Furthermore, the proportions of

missing data for each variable are presented. Here, imputation (e.g. mean/median or multiple imputation) was not performed due to the relatively low prevalence of missing data.

4.2.4.2 Regression Analysis

A regression analysis is conducted to assess the impact of important individual and sampling characteristics on participants' responses to the three questionnaires and utility scores. More precisely, ordered logistic regression models are developed for each scale question of the 'device satisfaction' questionnaire in which the gender of the respondent together with their age at sampling date, health state (i.e. healthy, diseased) and the *U-Rhythm* device version with which they were sampled are the independent variables. For the 'use of health care and impact of health on work' questionnaire, ordinary least squares (OLS) linear regression is used for all questions apart from that asking about the participant's current employment status, where logistic regression is used. Here, the respondent's gender, age at sampling date, and health state are used as the explanatory variables. Lastly, OLS linear regression with the same regressors is used for the EQ-5D-5L utility scores, while ordered logistic regression models are built for their responses to each dimension of the questionnaire.

4.3 Results

4.3.1 Descriptive Analysis

219 healthy, 65 PA, 49 CS and 24 AD participants were recruited and sampled (at least once) in *ULTRADIAN* between the years 2016-2019[10]. **Table 16** displays the number of individuals that were enrolled by each site together with their demographic and clinical characteristics, diagnostic testing information (when applicable), and sampling details. As shown, Bergen recruited the majority of healthy volunteers (36.53%) and PA patients (67.69%), while Athens enrolled 65.31% of CS patients and Bristol, 45.83% of AD patients. In all cohorts, most participants were females with their average age at screening date being ≈39 years old for healthy individuals and ≈50 years for all other groups. Most participants were non-smokers and had an average BMI between 23-30 (with higher mean BMI particularly in the PA and CS

[10] Total recruitment numbers in *ULTRADIAN*: 234 healthy, 65 PA, 56 CS, 48 AD and 3 CAH (excluded).

groups). Regarding the disease aetiology in PA, 24 patients had unilateral and 32 bilateral PA, while for 9 patients the cause was not yet identified, most probably due to not having received a subtype test or undergone surgery. CS was mainly caused by issues in the pituitary gland (67.35%), while adrenal causes were present instead in 30.61% of the patients and a non-defined aetiology was recorded in one patient.

Table 16 presents a non-comprehensive list of disease-specific tests that the patients received within a median period of ≈1 year for PA and CS or ≈8 years for AD patients before *ULTRADIAN* screening. Most PA patients underwent the aldosterone-renin ratio (95.38%) and serum potassium screening tests (96.92%), the saline infusion confirmatory test (84.62%), and both the adrenal computerised tomography (CT) and venous sampling subtyping tests (87.69% and 78.46%, respectively). In CS, the serum cortisol and adrenocorticotropin hormone (ACTH), dexamethasone suppression and 24-hour urine cortisol diagnostic tests were performed in 93.88%, 91.84% and 85.71% of all patients, respectively, with adrenal and pituitary imaging also used frequently (30.61% and 59.18%, respectively). For AD patients, adrenal antibodies (screening) were checked in all patients, while low basal plasma cortisol or elevated basal plasma ACTH were used for confirming the disease in ≈80% of the patients.

Before the first sampling with the *U-Rhythm* device, blood pressure was recorded elevated in all patient cohorts compared to the healthy volunteers, with PA patients having the highest mean systolic (≈140) and diastolic (≈87) measurements. In all groups, the majority of participants had one sample taken during the study. 54 PA and 27 CS patients were sampled pre-operatively only, while 11 and 20, respectively, both pre- and post-operatively. Most participants were sampled using the second version of *U-Rhythm*, which was mainly used after the end of 2017. In general, in >92% of all sampling cases (at all sampling timepoints and using any device version) (n=357) had no skin reactions during/after sampling (**Table 17**), while issues with the sampler or other issues were reported <15% of the time. Regarding the quality of the samples collected, these were good or usable in 77.28% and 83.88% of all cases that versions 1.0 and 2.0 were used, respectively.

Table 16: ULTRADIAN participant demographic and clinical characteristics, diagnostic testing information, and sampling details

Participant Characteristics & Sampling Details	Participant Subgroups				Difference between Cohorts (p-value)
	Healthy (n = 219)	Primary Aldosteronism (n = 65)	Cushing's Syndrome (n = 49)	Addison's Disease (n = 24)	
Number per clinical site; n (%)					
Athens	48 (21.92)	20 (30.77)	32 (65.31)	0 (0.00)	$chi^2(9) = 100.420$
Bergen	80 (36.53)	44 (67.69)	14 (28.57)	6 (25.00)	$p < 0.001$
Bristol	53 (24.20)	1 (1.54)	3 (6.12)	11 (45.83)	
Stockholm	38 (17.35)	0 (0.00)	0 (0.00)	7 (29.17)	
Gender; n (%)					
Female	128 (58.45)	35 (53.85)	41 (83.67)	18 (75.00)	$chi^2(3) = 14.521$
Male	91 (41.55)	30 (46.15)	8 (16.33)	6 (25.00)	$p = 0.002$
Age at screening date (years); mean (SD, range)	39.05 (12.97, 19.10-68.18)	51.93 (10.80, 21.70-70.36)	50.94 (14.07, 17.97-78.48)	48.26 (11.78, 20.39-69.27)	$F = 25.26$ $p < 0.001$
Smoking status; n (%)					
Non-smoker	191 (87.21)	57 (87.69)	35 (71.43)	20 (83.33)	$chi^2(6) = 17.914$
Smoker	28 (12.79)	8 (12.31)	12 (24.49)	4 (16.67)	$p = 0.006$
Missing	0 (0.00)	0 (0.00)	2 (4.08)	0 (0.00)	

Participant Characteristics & Sampling Details	Participant Subgroups				Difference between Cohorts (p-value)
	Healthy (n = 219)	**Primary Aldosteronism (n = 65)**	**Cushing's Syndrome (n = 49)**	**Addison's Disease (n = 24)**	
Body mass index at screening date; mean (SD, range)[11]	23.74 (3.15, 16.79-34.14)	29.30 (4.96, 17.58-42.76)	29.94 (6.29, 19.36-43.95)	25.05 (3.52, 19.87-35.55)	F = 50.76 p < 0.001
Disease aetiology; n (%)[12]					
Adrenal	N/A	N/A	15 (30.61)	N/A	-
Bilateral	N/A	32 (49.23)	N/A	N/A	
Pituitary	N/A	N/A	33 (67.35)	N/A	
Unilateral	N/A	24 (36.92)	N/A	N/A	
Undefined	N/A	9 (13.85)	1 (2.04)	N/A	
Screening and diagnostic testing; n (%)[13]					
Aldosterone-renin ratio	N/A	62 (95.38)	N/A	N/A	-
Serum potassium	N/A	63 (96.92)	N/A	N/A	
Saline infusion	N/A	55 (84.62)	N/A	N/A	
Fludrocortisone suppression	N/A	1 (1.54)	N/A	N/A	
Oral sodium loading	N/A	1 (1.54)	N/A	N/A	

[11] Missing values in 1 healthy and 1 PA participants.
[12] 'Undefined' aetiology means either that information was not entered in the database or that the disease cause was not yet identified.
[13] Tests received by the patient before their inclusion in *ULTRADIAN*. Missing values in 1 PA and 1 CS participants.

Participant Characteristics & Sampling Details	Participant Subgroups				Difference between Cohorts (p-value)
	Healthy (n = 219)	Primary Aldosteronism (n = 65)	Cushing's Syndrome (n = 49)	Addison's Disease (n = 24)	
Captopril challenge	N/A	0 (0.00)	N/A	N/A	
Other dynamic (for primary aldosteronism)	N/A	7 (10.77)	N/A	N/A	
Adrenal computerised tomography	N/A	57 (87.69)	15 (30.61)	N/A	
Adrenal magnetic resonance imaging	N/A	5 (7.69)	N/A	N/A	
Other imaging	N/A	2 (3.08)	N/A	N/A	
Adrenal venous sampling	N/A	51 (78.46)	N/A	N/A	
Serum cortisol and ACTH	N/A	N/A	46 (93.88)	N/A	
Dexamethasone suppression	N/A	N/A	45 (91.84)	N/A	
24-hour urine cortisol	N/A	N/A	42 (85.71)	N/A	
Late-night salivary cortisol	N/A	N/A	11 (22.45)	N/A	
Late-night serum cortisol	N/A	N/A	31 (63.27)	N/A	
Pituitary imaging	N/A	N/A	29 (59.18)	N/A	
Bilateral inferior petrosal sinus sampling	N/A	N/A	16 (32.65)	N/A	
Corticotropin-releasing hormone	N/A	N/A	25 (51.02)	N/A	
High-dose dexamethasone suppression	N/A	N/A	16 (32.65)	N/A	
Other Cushing's-related	N/A	N/A	2 (4.08)	N/A	
Positive adrenal antibodies 21-hydroxylase	N/A	N/A	N/A	24 (100.00)	

Participant Characteristics & Sampling Details	Participant Subgroups				Difference between Cohorts (p-value)
	Healthy (n = 219)	Primary Aldosteronism (n = 65)	Cushing's Syndrome (n = 49)	Addison's Disease (n = 24)	
Low basal plasma cortisol	N/A	N/A	N/A	20 (83.33)	
Elevated basal plasma ACTH	N/A	N/A	N/A	18 (75.00)	
Abnormal Synacthen	N/A	N/A	N/A	11 (45.83)	
Other Addison's-related	N/A	N/A	N/A	6 (25.00)	
Blood pressure before first sampling; mean (SD, range)[14]					
Systolic	113.86 (14.63, 85-155)	139.76 (18.06, 100-200)	129.88 (18.74, 90-182)	121.17 (15.51, 95-145)	F = 47.89 p < 0.001
Diastolic	71.71 (9.63, 50-100)	86.73 (12.57, 65-130)	81.33 (12.18, 60-128)	78.54 (10.44, 60-104)	F = 37.16 p < 0.001
Number of samples; n (%)					
1	180 (82.19)	52 (80.00)	28 (57.14)	23 (95.83)	$chi^2(6) = 37.266$
2	17 (7.76)	11 (16.92)	18 (36.73)	1 (4.17)	p < 0.001
3+	22 (10.05)	2 (3.08)	3 (6.12)	0 (0.00)	

[14] Missing values in 5 healthy and 3 PA participants.

Participant Characteristics & Sampling Details	Participant Subgroups				Difference between Cohorts (p-value)
	Healthy (n = 219)	Primary Aldosteronism (n = 65)	Cushing's Syndrome (n = 49)	Addison's Disease (n = 24)	
Sampling timepoint; n (%)					
Pre-operative only	N/A	54 (83.08)	27 (55.10)	N/A	
Post-operative only	N/A	-	1 (2.04)	N/A	-
Pre- and post-operative	N/A	11 (16.92)	20 (40.82)	N/A	
Missing	N/A	-	1 (2.04)	N/A	
***U-Rhythm* device at first sampling; n (%)**					
Version 1.0	5 (2.28)	10 (15.38)	6 (12.24)	1 (4.17)	$chi^2(3) = 18.559$
Version 2.0	214 (97.72)	55 (84.62)	43 (87.76)	23 (95.83)	$p < 0.001$

**Abbreviations:* *ACTH: adrenocorticotropic hormone; N/A: not applicable; SD: standard deviation*

Table 17: User adverse reactions/issues during sampling and microdialysis sample quality

Details*	*U-Rhythm* Sampler Type	
	Version 1.0 (n = 22)	**Version 2.0 (n = 335)**
Skin reactions; n (%)		
No	22 (100.00)	311 (92.84)
Yes	-	13 (3.88)
Missing	-	11 (3.28)
Sampler issues; n (%)		
No	19 (86.36)	299 (89.25)
Yes	3 (13.64)	22 (6.57)
Missing	-	14 (4.18)
Other issues; n (%)		
No	18 (81.82)	271 (80.90)
Yes	2 (9.09)	38 (11.34)
Missing	2 (9.09)	26 (7.76)
Microdialysis sample quality; n (%)		
Poor	2 (9.09)	53 (15.82)
Usable	3 (13.64)	41 (12.24)
Good	14 (63.64)	240 (71.64)
Missing	3 (13.64)	1 (0.30)

**Information is based on the total number of samples collected (n=357).*

Table 18: Questionnaire completion numbers

Questionnaire*	Participant Subgroups			
	Healthy (n = 219)	**Primary Aldosteronism (n = 65)**	**Cushing's Syndrome (n = 49)**	**Addison's Disease (n = 24)**
Device satisfaction; n (%)				
No	6 (2.74)	9 (13.85)	5 (10.20)	-
Yes	213 (97.26)	56 (86.15)	44 (89.80)	24 (100.00)
Use of health care and impact of health on work; n (%)				
No	12 (5.48)	11 (16.92)	6 (12.24)	-
Yes	207 (94.52)	54 (83.08)	43 (87.76)	24 (100.00)
EQ-5D-5L; n (%)				
No	11 (5.02)	5 (7.69)	5 (10.20)	-
Yes	208 (94.98)	60 (92.31)	44 (89.80)	24 (100.00)

**Only the first time of completion is reported. 'No' means that the questionnaire was either missing in the database or not completed by the participant.*

Table 18 indicates that in the majority of cases, the three questionnaires were partially or fully completed the first time that sampling was undertaken. **Table 19** shows the responses of each group to each 'device satisfaction' questionnaire question. Specifically, individuals strongly agreed that having the sampling probe fitted was not too uncomfortable and did not cause any pain, while they were not bothered by the noise of the device during the day or at night. Participants also agreed that carrying the device was easy and that they would be happy to wear the device again. Being able to continue with their daily activities and sleep normally during the night were the two elements with which all groups agreed less. The latter issues were also reported in the comment section of the questionnaire, with one third of those who completed this part (n=126) mentioning that it was difficult to do their usual activities, especially those that involved bending (e.g. sitting, exercising/cycling, driving). Sleep was also a problem given that the individuals lay on the bed on the opposite side to that of the device with restricted movement due to concern about disconnecting or damaging it. Many respondents also said that they had problems with the tubes of the device and that it would

be better if all of them were contained within the device/belt. The number of tubes hanging caused participants worries since they needed to be careful not to damage/kink/disconnect them during the day or when sleeping. Another problem that respondents had was the size and weight of *U-Rhythm*. Many found it too big and heavy to transfer or too bulky to hide under their clothes, feeling weird when being around people. Additionally, some problems with noise from the pump, especially in silent occasions (e.g. meetings; cinema; at night), were reported. Lastly, participants made some suggestions on improvements that they would like to see in the future versions (e.g. more adjustable and stable belt to avoid having the device moving around; waterproof; less sticky and warm to avoid sweating; more user-friendly).

Table 20 provides information on the healthcare resources that participants had used within three months from their first sampling session. As shown, CS and PA patients had the largest number of hospital stays, and specialist and other doctor/nurse visits, followed by AD patients and healthy individuals. Moreover, in all groups except CS, most individuals were employed. CS patients reported that they had missed on average ≈15 days from work and that their work productivity had been affected by ≈40% due to their health condition in the last three months. PA reported missing ≈5 days from work, with their productivity being deteriorated to the same extent. Healthy and AD participants missed on average <1 day from work in that time period, but AD patients had also productivity issues similar to CS and PA patients. **Figure 17** presents the mean utility scores at first sampling for each study group (208 healthy, 59 PA, 40 CS, 24 AD). The figure illustrates that healthy volunteers had the highest HRQoL (0.93), with AD patients showing slightly lower scores (0.85), followed by PA (0.76) and CS (0.53) patients. When looking at the health utilities of patients that completed the EQ-5D-5L both pre- and post-operatively (**Figure 18**), HRQoL increased from 0.66 to 0.74 in PA patients (t=0.69; p=0.499), while it increased from 0.56 to 0.64 in CS patients (t=0.88; p=0.385).

Table 19: 'Device satisfaction' questionnaire responses

Question°		Participant Subgroups				Difference between Cohorts (p-value)
		Healthy (n = 219)	Primary Aldosteronism (n = 65)	Cushing's Syndrome (n = 49)	Addison's Disease (n = 24)	
Mean (SD, range)	**Q1**¤	4.43 (0.83, 1-5)	4.76 (0.43, 4-5)	4.75 (0.49, 3-5)	4.71 (0.46, 4-5)	chi^2 (12) = 16.824 p = 0.156
	Q2¤	3.94 (0.85, 1-5)	3.96 (1.13, 1-5)	4.17 (1.12, 1-5)	3.74 (0.96, 2-5)	chi^2 (12) = 34.003 p = 0.001
	Q3¤	3.77 (0.97, 1-5)	3.47 (1.09, 2-5)	4.09 (1.13, 2-5)	3.91 (0.95, 2-5)	chi^2 (12) = 27.367 p = 0.007
	Q4¤	4.47 (0.86, 1-5)	4.69 (0.70, 1-5)	4.64 (0.85, 1-5)	4.57 (0.66, 3-5)	chi^2 (12) = 10.993 p = 0.530
	Q5¤	4.38 (0.91, 1-5)	4.71 (0.65, 2-5)	4.57 (0.89, 1-5)	4.61 (0.66, 3-5)	chi^2 (12) = 15.900 p = 0.196
	Q6¤	3.89 (0.98, 1-5)	3.69 (1.21, 1-5)	4.00 (1.14, 1-5)	3.87 (1.06, 1-5)	chi^2 (12) = 12.874 p = 0.378
	Q7¤	4.40 (0.95, 1-5)	4.46 (0.88, 2-5)	4.51 (0.90, 2-5)	4.39 (0.72, 3-5)	chi^2 (12) = 9.582 p = 0.653
	Q8¤	4.17 (0.94, 1-5)	4.22 (1.03, 1-5)	4.21 (1.20, 1-5)	4.61 (0.50, 4-5)	chi^2 (12) = 18.205 p = 0.110

**Abbreviations:* SD: standard deviation*

°Only the first time of completion is reported. Although the responses have no arithmetic properties, the measures of central tendency are presented to indicate how strongly participants agreed or disagreed with each question/statement.

¤Description and number of missing responses:

Q1 (comfortability of the sampling probe) *– 6 healthy, 10 PA, 5 CS*

Q2 (easiness of carrying the sampler) *– 7 healthy, 11 PA, 7 CS, 1 AD*

Q3 (ability to continue with daily activities) *– 7 healthy, 10 PA, 6 CS, 1 AD*

Q4 (pain from the sampling probe) *– 6 healthy, 11 PA, 7 CS, 1 AD*

Q5 (noise from the sampler during the day) *– 7 healthy, 9 PA, 7 CS, 1 AD*

Q6 (easiness to sleep with the sampler) *– 9 healthy, 11 PA, 8 CS, 1 AD*

Q7 (disturbance during sleep due to noise) *– 7 healthy, 10 PA, 8 CS, 1 AD*

Q8 (willingness to wear the sampler again) *– 8 healthy, 10 PA, 7 CS, 1 AD*

Table 20: 'Use of health care and impact of health on work' questionnaire responses

Question°		Participant Subgroups				Difference between Cohorts (p-value)
		Healthy (n = 219)	Primary Aldosteronism (n = 65)	Cushing's Syndrome (n = 49)	Addison's Disease (n = 24)	
n (%) / Mean (SD, range)	**Q1¤**	**206 (94.06)**	**52 (80.00)**	**42 (85.71)**	**23 (95.83)**	F = 1.13 p = 0.336
		0.01 (0.07, 0-1)	1.77 (5.26, 0-33)	4.17 (7.92, 0-50)	0.13 (0.46, 0-2)	
	Q2¤	**206 (94.06)**	**48 (73.85)**	**41 (83.67)**	**23 (95.83)**	F = 1.65 p = 0.178
		0.07 (0.36, 0-3)	2.14 (2.10, 0-7.5)	3.35 (2.65, 0-12.5)	0.65 (0.98, 0-4)	
	Q3¤	**205 (93.61)**	**52 (80.00)**	**40 (81.63)**	**23 (95.83)**	F = 1.55 p = 0.201
		0.34 (0.59, 0-3)	2.97 (2.19, 0-10.5)	2.54 (2.01, 0-10)	0.87 (1.18, 0-4)	
	Q4¤	**207 (94.52)**	**54 (83.08)**	**42 (85.71)**	**24 (100.00)**	
	No	46 (22.22)	21 (38.89)	26 (61.90)	7 (29.17)	chi^2 (3) = 28.003
	Yes	161 (77.78)	33 (61.11)	16 (38.10)	17 (70.83)	p < 0.001
	Q5¤	**163 (74.43)**	**32 (49.23)**	**17 (34.69)**	**16 (66.67)**	F = 1.74 p = 0.160
		0.99 (4.93, 0-60)	5.09 (8.23, 0-30)	15.38 (27.71, 0-90)	0.81 (1.87, 0-6)	
	Q6¤	**131 (59.82)**	**34 (52.31)**	**25 (51.02)**	**16 (66.67)**	F = 1.33 p = 0.266
		0.85 (2.07, 0-10)	3.59 (3.05, 0-10)	3.92 (3.21, 0-10)	3.81 (2.69, 0-8)	

**Abbreviations:* SD: standard deviation*

°Only the first time of completion is reported.

¤Description and number of missing/'not known'/'not applicable' responses:

Q1 (number of nights in hospital) *– 13 healthy, 13 PA, 7 CS, 1 AD*

Q2 (number of specialist visits) *– 13 healthy, 17 PA, 8 CS, 1 AD*

Q3 (number of other doctor/nurse visits) *– 14 healthy, 13 PA, 9 CS, 1 AD*

Q4 (current employment status) *– 12 healthy, 11 PA, 7 CS*

Q5 (number of days missed from work) *– 56 healthy, 33 PA, 32 CS, 8 AD*

Q6 (impact of health on work productivity) *– 88 healthy, 31 PA, 24 CS, 8 AD*

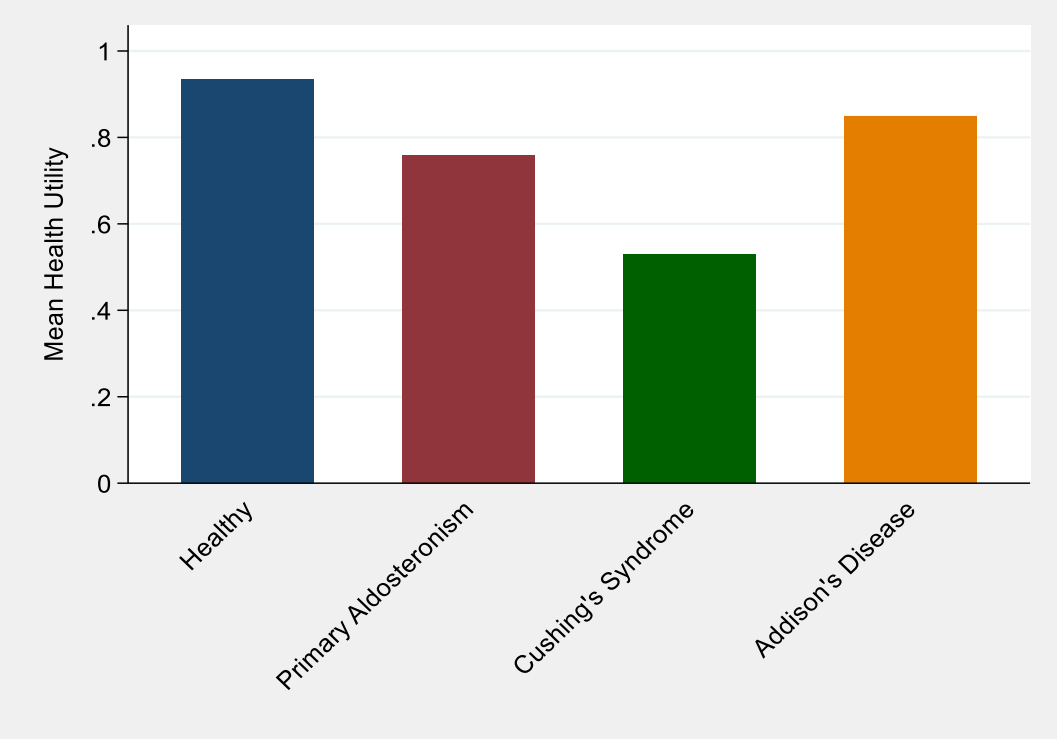


Figure 17: Baseline mean health utility per study group

**Description:* Utility scores for 208 healthy, 59 PA, 40 CS and 24 AD participants.*

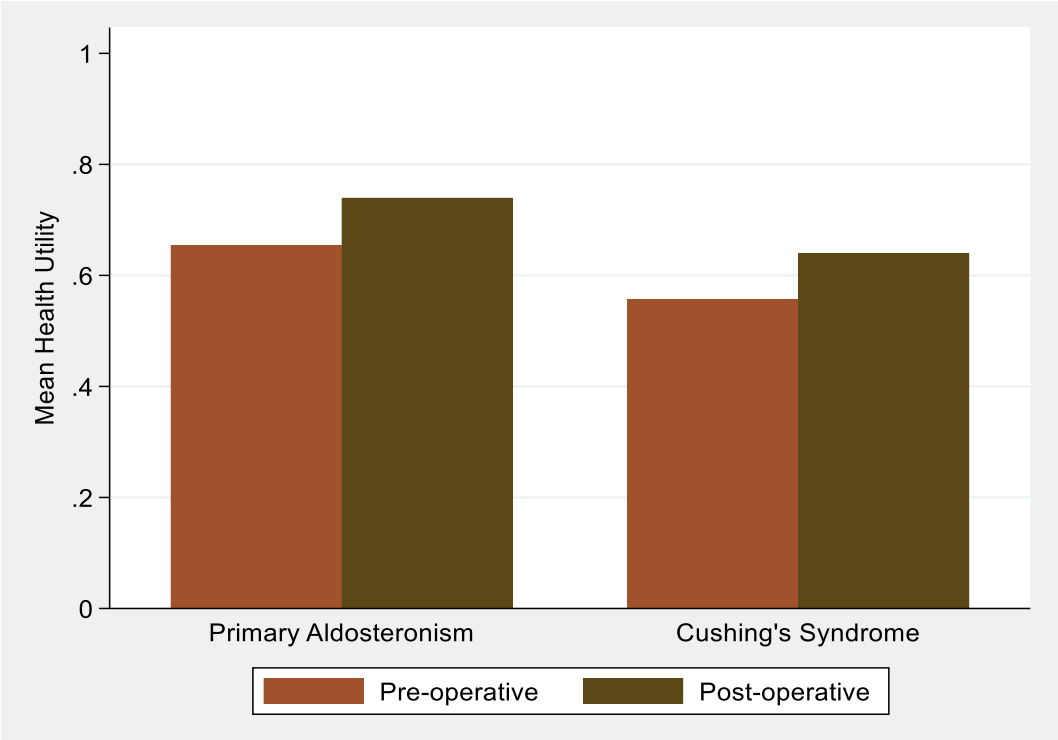


Figure 18: Pre- and post-operative mean health utilities per disease group

**Description:* Utility scores for 8 PA and 15 CS patients with pre- and post-operative utility scores.*

4.3.2 Regression Analysis

Table 21 presents the outcomes of the ordered logistic regression models for each question of the 'device satisfaction' questionnaire. Here, it should be mentioned that the ordered logit models estimate one equation over all the levels of each question, which means that for one unit increase in each predictor, they derive the odds of having the next level response (e.g. undecided) versus the combined previous levels (i.e. strongly disagreed and disagreed), or the combined responses (e.g. agreed and strongly agreed) versus the previous level response (i.e. undecided), given that all other variables are held constant. As shown, in all cases apart from questions 6 (easiness to sleep) and 7 (disturbance from noise during sleep), age was the most significant factor associated with the participant's responses, with older patients reporting less problems with the device. This may be because *U-Rhythm* interferes with vigorous activity that younger individuals might engage in more. The sampler version had a non-significant effect on the participant's satisfaction. Furthermore, the health condition of the respondent was the most significant parameter that affected the healthcare resources that they used and their productivity at work (**Table 22**). The health condition of the participant also played an important role in their HRQoL (**Table 23**), with utility scores dropping by 0.08, 0.17 and 0.39 for AD, PA and CS patients, respectively, when compared to those of the healthy individuals ($p<0.05$). This was also shown when examining the parameters that affected their responses to each dimension of the EQ-5D-5L questionnaire (***Appendix 12***).

Table 21: 'Device satisfaction' questionnaire regression analysis results

Question 1: Comfortability of the sampling probe				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	1.31	0.32	1.09	0.277
Age	1.04	0.01	4.18	<0.001
Condition[b]				
Primary aldosteronism	1.67	0.62	1.37	0.170
Cushing's syndrome	1.82	0.75	1.46	0.144
Addison's disease	1.50	0.72	0.85	0.397
Sampler version[c]	1.17	0.71	0.25	0.801
Number of observations = 336		Likelihood ratio $chi^2(6)$ = 34.10		p-value < 0.001
Pseudo R^2 = 0.060		Log likelihood = -265.532		
Question 2: Easiness of carrying the sampler				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.95	0.21	-0.22	0.823
Age	1.02	0.01	2.09	0.037
Condition[b]				
Primary aldosteronism	1.19	0.38	0.53	0.593
Cushing's syndrome	2.03	0.74	1.93	0.054
Addison's disease	0.59	0.25	-1.26	0.208
Sampler version[c]	2.32	1.39	1.40	0.160
Number of observations = 331		Likelihood ratio $chi^2(6)$ = 15.55		p-value = 0.016
Pseudo R^2 = 0.019		Log likelihood = -393.966		

Question 3: Ability to continue with daily activities				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.85	0.18	-0.75	0.454
Age	1.02	0.01	2.63	0.009
Condition[b]				
Primary aldosteronism	0.45	0.14	-2.67	0.008
Cushing's syndrome	1.82	0.64	1.69	0.091
Addison's disease	1.02	0.42	0.05	0.957
Sampler version[c]	0.66	0.35	-0.80	0.425
Number of observations = 333		Likelihood ratio chi^2(6) = 21.24		p-value = 0.002
Pseudo R^2 = 0.025		Log likelihood = -422.704		

Question 4: Pain from the sampling probe				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.99	0.25	-0.03	0.974
Age	1.03	0.01	3.35	0.001
Condition[b]				
Primary aldosteronism	1.10	0.41	0.26	0.797
Cushing's syndrome	1.34	0.57	0.69	0.491
Addison's disease	0.78	0.37	-0.52	0.601
Sampler version[c]	0.15	0.16	-1.81	0.071
Number of observations = 332		Likelihood ratio chi^2(6) = 21.83		p-value = 0.001
Pseudo R^2 = 0.037		Log likelihood = -288.005		

Question 5: Noise from the sampler during the day				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.91	0.22	-0.41	0.684
Age	1.03	0.01	3.11	0.002
Condition[b]				
Primary aldosteronism	1.68	0.64	1.38	0.169
Cushing's syndrome	1.27	0.51	0.60	0.550
Addison's disease	1.23	0.60	0.43	0.670
Sampler version[c]	0.14	0.15	-1.87	0.062
Number of observations = 333		Likelihood ratio chi^2(6) = 26.49		p-value < 0.001
Pseudo R^2 = 0.042		Log likelihood = -304.506		

Question 6: Easiness to sleep with the sampler				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.92	0.20	-0.40	0.689
Age	1.02	0.01	1.93	0.054
Condition[b]				
Primary aldosteronism	0.67	0.21	-1.30	0.195
Cushing's syndrome	1.18	0.41	0.46	0.644
Addison's disease	0.86	0.36	-0.36	0.715
Sampler version[c]	1.36	0.74	0.57	0.572
Number of observations = 328		Likelihood ratio chi^2(6) = 6.85		p-value = 0.335
Pseudo R^2 = 0.008		Log likelihood = -433.542		

Question 7: Disturbance during sleep due to noise				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.82	0.19	-0.86	0.390
Age	1.01	0.01	0.50	0.617
Condition[b]				
Primary aldosteronism	1.00	0.33	-0.01	0.994
Cushing's syndrome	1.28	0.50	0.64	0.520
Addison's disease	0.71	0.30	-0.82	0.412
Sampler version[c]	0.41	0.27	-1.36	0.174
Number of observations = 331		Likelihood ratio chi^2(6) = 4.93		p-value = 0.553
Pseudo R^2 = 0.007		Log likelihood = -334.791		

Question 8: Willingness to wear the sampler again				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	1.13	0.25	0.54	0.589
Age	1.03	0.01	3.25	0.001
Condition[b]				
Primary aldosteronism	0.80	0.26	-0.70	0.483
Cushing's syndrome	1.07	0.39	0.19	0.846
Addison's disease	1.78	0.78	1.33	0.183
Sampler version[c]	0.58	0.31	-1.04	0.298
Number of observations = 331		Likelihood ratio chi^2(6) = 16.09		p-value = 0.013
Pseudo R^2 = 0.020		Log likelihood = -385.784		

*Reference categories: **a** – 'female'; **b** – 'healthy'; **c** – 'Version 1.0'*

***Only the first time of completion is reported.*

Table 22: 'Use of health care and impact of health on work' questionnaire regression analysis results

Question 1: Number of nights in hospital				
Variable	**Coefficient**	**Standard Error**	**t**	**p-value**
Gender[a]	-0.28	0.41	-0.67	0.505
Age	-0.01	0.02	-0.56	0.578
Condition[b]				
Primary aldosteronism	1.88	0.59	3.20	0.001
Cushing's syndrome	4.19	0.64	6.60	<0.001
Addison's disease	0.15	0.79	0.19	0.850
Constant	0.46	0.69	0.67	0.503
Number of observations = 323		F(5, 317) = 10.73		p-value < 0.001
R^2 = 0.145		Adjusted R^2 = 0.131		Root MSE = 3.543
ESS(5) = 673.78		RSS(317) = 3,979.85		TSS(322) = 4,653.63
Question 2: Number of specialist visits				
Variable	**Coefficient**	**Standard Error**	**t**	**p-value**
Gender[a]	0.12	0.15	0.76	0.447
Age	-0.01	0.01	-2.46	0.014
Condition[b]				
Primary aldosteronism	2.26	0.22	10.16	<0.001
Cushing's syndrome	3.48	0.23	14.86	<0.001
Addison's disease	0.72	0.29	2.49	0.013
Constant	0.57	0.26	2.23	0.026
Number of observations = 318		F(5, 312) = 56.61		p-value < 0.001
R^2 = 0.476		Adjusted R^2 = 0.467		Root MSE = 1.293
ESS(5) = 473.43		RSS(312) = 521.83		TSS(317) = 955.26

Question 3: Number of other doctor/nurse visits				
Variable	**Coefficient**	**Standard Error**	**t**	**p-value**
Gender[a]	0.03	0.15	0.18	0.861
Age	-0.00	0.01	-0.56	0.574
Condition[b]				
Primary aldosteronism	2.67	0.21	12.77	<0.001
Cushing's syndrome	2.25	0.23	9.71	<0.001
Addison's disease	0.56	0.28	1.99	0.047
Constant	0.45	0.25	1.81	0.071

Number of observations = 320	F(5, 314) = 48.06	p-value < 0.001
R^2 = 0.434	Adjusted R^2 = 0.425	Root MSE = 1.267
ESS(5) = 385.85	RSS(314) = 504.20	TSS(319) = 890.05

Question 4: Current employment status				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	1.35	0.36	1.10	0.270
Age	0.99	0.01	-1.42	0.155
Condition[b]				
Primary aldosteronism	0.53	0.19	-1.82	0.069
Cushing's syndrome	0.22	0.08	-3.99	<0.001
Addison's disease	0.83	0.41	-0.38	0.704

Number of observations = 327	Likelihood ratio $chi^2(5)$ = 29.95	p-value < 0.001
Pseudo R^2 = 0.074	Log likelihood = -186.36	

Question 5: Number of days missed from work				
Variable	**Coefficient**	**Standard Error**	**t**	**p-value**
Gender[a]	-0.43	1.25	-0.35	0.728
Age	0.09	0.05	1.74	0.082
Condition[b]				
Primary aldosteronism	3.30	1.82	1.81	0.072
Cushing's syndrome	13.64	2.36	5.78	<0.001
Addison's disease	-0.83	2.40	-0.34	0.731
Constant	-2.28	2.24	-1.02	0.309
Number of observations = 228		F(5, 222) = 9.13		p-value < 0.001
R^2 = 0.171		Adjusted R^2 = 0.152		Root MSE = 9.029
ESS(5) = 3,722.19		RSS(222) = 18,099.00		TSS(227) = 21,821.19

Question 6: Impact of health on work productivity				
Variable	**Coefficient**	**Standard Error**	**t**	**p-value**
Gender[a]	-0.78	0.36	-2.13	0.035
Age	0.02	0.01	1.06	0.291
Condition[b]				
Primary aldosteronism	2.67	0.49	5.47	<0.001
Cushing's syndrome	2.62	0.57	4.63	<0.001
Addison's disease	2.65	0.66	4.02	<0.001
Constant	0.61	0.63	0.96	0.337
Number of observations = 206		F(5, 200) = 14.83		p-value < 0.001
R^2 = 0.271		Adjusted R^2 = 0.252		Root MSE = 2.434
ESS(5) = 439.31		RSS(200) = 1,184.75		TSS(205) = 1,624.06

**Abbreviations:* *ESS: explained sum of squares; MSE: mean square residual (error); RSS: residual sum of squares; TSS: total sum of squares*

Reference categories:* ***a *– 'female';* ***b*** *– 'healthy'*

****Only the first time of completion is reported.*

Table 23: Baseline utility score regression analysis results

Variable	Coefficient	Standard Error	t	p-value
Gender[a]	0.02	0.02	1.18	0.237
Age	-0.00	0.00	-0.80	0.424
Condition[b]				
Primary aldosteronism	-0.17	0.02	-6.96	<0.001
Cushing's syndrome	-0.39	0.03	-13.47	<0.001
Addison's disease	-0.08	0.03	-2.21	0.028
Constant	0.95	0.03	31.39	<0.001
Number of observations = 331		F(3, 325) = 51.51		p-value < 0.001
R^2 = 0.442		Adjusted R^2 = 0.434		Root MSE = 0.154
ESS(5) = 6.105		RSS(325) = 7.704		TSS(330) = 13.809

**Abbreviations:* *ESS: explained sum of squares; MSE: mean square residual (error); RSS: residual sum of squares; TSS: total sum of squares*

Reference categories:* ***a *– 'female';* ***b*** *– 'healthy'*

****Only the first time of completion is reported.*

4.4 Discussion

A descriptive and regression analysis was conducted using data collected from *ULTRADIAN* participants (healthy, PA, CS, AD) after being sampled with *U-Rhythm* and completing three questionnaires asking about their experience with the device; the healthcare resources that they had used and the impact of their health condition on their work productivity within three months before sampling; and their HRQoL on the day of sampling (at baseline, pre- and post-operatively).

4.4.1 Key Findings

219 healthy, 65 PA, 49 CS and 24 AD participants were recruited and sampled between October 2016 and October 2019, exceeding the initial expectation on recruitment numbers for each cohort. Overall, participants were satisfied with the experience of using the sampler,

mentioning that they would be happy to wear the device again and making suggestions on elements that could improve in future versions (e.g. dimensions, belt, waterproof). PA and CS patients used substantially more healthcare resources and had worse HRQoL compared to the other two groups. CS had by far the largest impact on whether the individual was employed, while the work productivity of all patient groups was affected by ≈40% due to their health condition. Regression analysis results indicated that age was the significant parameter in participants' responses to the 'device satisfaction' questionnaire, while, as expected, their health condition was the factor with the most significant impact on their healthcare use, work productivity and HRQoL.

4.4.2 Comparison with Related Studies

After searching the literature for studies that have examined the healthcare utilisation in PA, CS and AD patients, no studies on PA could be found showing the scarcity of comparable data for this disease. For CS, five retrospective cohort studies were identified, from which four were conducted in the United States (USA) using data from large health insurance databases (188-193) and one in the UK using Clinical Practice Research Datalink linked with Hospital Episode Statistics (HES) data[15] (194). This is a main difference to the current study that used a small number of CS patients and data were based on their responses to a bespoke questionnaire. Therefore, the data collected are not expected to be as precise as routine data. Although all studies used higher-quality and more detailed data than the current study, only three compared CS data with data from (matched) healthy and/or CS-free individuals (188, 191-193). All studies mainly focused on the number and type of the comorbidities of CS and the healthcare use associated with its treatment (e.g. prescriptions, hospitalisations) and remission, while none of them examined the impact of the disease on work productivity. All studies also assessed the healthcare utilisation of CS patients in a period larger than one year. In the studies, the average age of CS patients (≈45 years) and the proportion of females (≈78%) were slightly lower than those of the current study. In the studies that compared CS with non-CS data, CS patients were found to have used 2-4 times more resources, which was much lower to what was estimated in the current study. This is expected since the CS patients

[15] Details on these last two databases are provided in **Chapter V**.

in these studies had been diagnosed and treated for many years compared to the current study that examined patients that were more recently diagnosed.

For AD, two studies assessing healthcare utilisation were identified, from which one was conducted in the USA using health insurance data (196) and one in the UK using HES data (195). The former study followed up AD and matched non-AD endocrine patients (e.g. CAH) for >12 months, while the latter estimated the resource use of AD patients over a one-year period. The main focus of both studies was on the treatment and hospitalisations related to AD, while the second study also measured the costs that were associated with reduced work productivity due to the disease. The US study used AD patients of similar age to the current study. However, the proportion of females was much lower (64.4%). The UK study was a conference abstract and did not provide any demographic information. In the US study, the annual number of hospital days were ≈4.2 (i.e. ≈1 every three months), which was much higher than the number found in the current study. The UK study estimated that the annual cost per patient due to reduced productivity is ≈£590. Here, it should be noted that most of the CS and AD studies identified aimed to measure the cost of the disease, whereas the current study only provided descriptive data on resource usage.

Regarding HRQoL, results from a prospective randomised controlled trial that was conducted in Europe and in which the EQ-5D-3L questionnaire was completed by PA patients indicated that the utility scores of those who were surgically treated increased from 0.85 to 0.92, while the scores of those who were medically treated increased from 0.85 to 0.89 over a period of one year from initiation of therapy (197). Studies that used other generic or disease-specific health questionnaires confirmed that the HRQoL of PA patients was significantly lower than that of the general population (199, 200), while it improved significantly after surgery (198, 214) or as the patient continued receiving medications (199). Two studies were found to use EQ-5D-3L and CushingQoL data from the European Registry on Cushing's syndrome to evaluate the HRQoL of CS patients (202) and differences in utility scores due to the aetiology of the disease (i.e. pituitary or adrenal) before and after surgery (203). The first study reported a mean utility score very close to that reported in the current study (0.55). In the second study, EQ-5D-3L responses were similar between the two groups before and after treatment (no utility scores reported), while CushingQoL responses were slightly worse for pituitary CS

patients after >1 year from surgery. This was in contrast to a previous study that did not find any differences between the two groups when the CushingQoL questionnaire was used (211). Other studies using generic health questionnaires confirmed that the HRQoL of CS patients is worse than that of the general population even after surgery (204, 205). For AD, one cross-sectional study was found to compare the HRQoL between patients and healthy controls, in which AD patients scored worse in the 'activity' dimension of the EQ-5D-3L (207). Other studies that used other generic health questionnaires also confirmed an impaired HRQoL for AD patients compared to the general population (206, 208).

4.4.3 Strengths and Limitations

This is the first study to compare the clinical features, recent healthcare use and testing, and HRQoL in PA, CS and AD patients (compared to healthy individuals) simultaneously. This is important since until now this has been an under-researched clinical area with only a few studies looking at the resource utilisation, costs and health outcomes associated with these diseases (50, 51, 183-208). This analysis has also the benefit that it included patients from different European countries and subsequently healthcare jurisdictions, giving a broad picture of the number and types of screening, diagnostic and monitoring tests that are most commonly used for each condition. Additionally, this study used a well-validated health questionnaire (EQ-5D-5L) to measure utility scores across different populations, which can be easily compared to values from other diseases. In contrast, current evidence has used other (non-preference-based) generic and/or disease-specific questionnaires on single populations (209-214, 242), producing results that are less easily interpreted and compared. All these aspects are essential when examining the clinical area and market in which a new technology can be beneficial in order to know the interventions with which it will compete and identify the areas where further research should be conducted to develop a more efficient alternative.

Nevertheless, this analysis included only a small number of patients. This was partially due to the low prevalence of the examined diseases in the different populations but also due to not all *ULTRADIAN* participants having data available to include in the analysis. Having said that, except AD where data from half of the participants were not available, only a few patients

enrolled in the other *ULTRADIAN* groups were not included in the analysis. Therefore, it is anticipated that the final dataset and results would not differ too much. Missing values were also caused by the fact that this was a multi-centre study, so different sites used different databases and employed different numbers of staff to collect, enter and upload the data on the *ULTRADIAN* server. Moreover, the fact that the study used combined data from different countries might have affected its findings. This is because each country may use different diagnostic tools and treatments depending on the structure of its healthcare system, its resource availability, its technological capacity, and the skillset of its medical personnel. Therefore, participants might have used different resources and might have had different expectations of *U-Rhythm* based on their experience with other medical tests. Given the small recruitment numbers and the limited amount of data collected from *ULTRADIAN*, it was difficult to account for the differences between the populations, while combining their data was needed to enhance the robustness of the results. Nevertheless, the countries included in the study have similar demographic and socioeconomic characteristics, and a publicly financed healthcare system that uses products approved in a common (European) market. Hence, it is expected that the outcomes would not have changed too much if each site was examined separately.

Another limitation of this study was that it used two simple non-validated questionnaires to collect information on the participants' satisfaction with *U-Rhythm* and their healthcare use. Although the applicability and validity of these instruments could have been checked in pilot work, this was not possible given the time constraints needed to get ethical approval for the *ULTRADIAN* project. Therefore, the two questionnaires were developed to obtain basic data that could be easily compared between the patient groups. Furthermore, as mentioned in *Section 4.2.3.3.2*, when deciding to use the EQ-5D-5L to measure the participants' HRQoL, there were some questions raised regarding its appropriateness for these diseases (e.g. responsiveness). However, given that not all the examined conditions have an existing, translated and validated HRQoL instrument, and the fact that the aim of this study was to compare health outcomes between all patient groups, it was thought that the EQ-5D-5L was the best alternative. One problem that may have affected the utility scores produced was the use of the UK EQ-5D-3L 'crosswalk' algorithm instead of using an actual EQ-5D-5L value set from individual countries. However, given that most countries have not yet developed EQ-

5D-5L value sets and the 'crosswalk' algorithm is recommended by NICE for use in the UK, this was considered the most appropriate method.

Another problem that may have influenced the findings of this analysis is that *ULTRADIAN* did not collect information on how long the patients had their conditions for. This is important since patients that are very early on or just diagnosed with a disease may respond differently compared to those who have had the condition for a longer period (e.g. healthcare usage, HRQoL). Since the *ULTRADIAN* study design and case report forms were already developed before this PhD began, this was an intractable limitation of the study. Lastly, regarding the regression analysis of the EQ-5D data, although OLS linear and (ordered) logistic regression models were produced to explore the parameters that affect the individual's responses and subsequently HRQoL, other types of models (e.g. adjusted limited dependent variable (beta-based) mixture regression models) have been proposed as better alternatives (243). These models were not used in this analysis due to time constraints but also because of the small number of participants and variety in their characteristics, which would have made their development challenging.

4.4.4 Implications for Clinical Practice and Further Research

Information about the patients' demographic and clinical characteristics, healthcare use and HRQoL; estimations of the plausible cost and diagnostic accuracy of a new medical test; and a preliminary evaluation of its potential role in future practice are useful for the developers of a test at earlier stages of development. By examining these elements early, manufacturers can understand more about the clinical area; identify key drivers of the total test unit cost; and begin to draw comparisons with competitor technologies already available in clinical practice. Nevertheless, these analyses must be conducted with appropriate caution as the components (i.e. equipment, consumables, laboratory analysis) of the diagnostic test and therefore its cost, and its use and diagnostic accuracy are likely to evolve/change rapidly during development. This may result in different clinical uses than originally expected as well as cost increases (e.g. as more expensive components are chosen to achieve CE marking) or decreases (e.g. as larger production runs yield economies of scale and competition among

different laboratories offering similar assays). All these should be assessed by both developers and policy makers before deciding whether they should adopt/reject the new technology.

Within *ULTRADIAN*, the goal was to disseminate the study outputs to the study collaborators and relevant scientific communities (e.g. through national and international congresses, and peer-reviewed publications) to deepen knowledge in this clinical area and raise awareness to stakeholders (e.g. healthcare authorities and providers; patient organisations), clinicians and other experts/scientists, and the general public about the development of a potentially more efficient test for the diagnosis and monitoring of the examined endocrine disorders. The plan was to first evaluate the device in a substantial number of patients from different European countries. Further work might then include a larger multi-centre study (e.g. prospective diagnostic cohort study), which could potentially indicate whether the device has advantages compared to current testing. This could lead to a faster and more accurate diagnosis of the diseases in the future as well as the formation of new and more established clinical guidelines. The larger multi-centre study could also be used to examine the validity of EQ-5D-5L for these diseases as well as compare its results to those produced by disease-specific questionnaires. Additionally, a validated healthcare use questionnaire for each condition could be developed.

Following the *ULTRADIAN* participants' responses to the 'device satisfaction' questionnaire, the third version of *U-Rhythm*, which was developed close to the end of the study, came with several improvements. Specifically, the third version was smaller in size and lighter to allow greater movement, be less noticeable and fit better in a belt; less noisy to avoid disturbing the user during quiet hours; more compact to reduce the chance of the tubes being damaged or disconnected; and was covered with higher-quality materials to give a soft-touch finish to the product. Nevertheless, to improve the design, function and performance of *U-Rhythm*'s future versions, the 'device satisfaction' questionnaire is not sufficient since it only provides limited information on the user's preferences. Additionally, these aspects of the value of new medical devices cannot be captured by measuring costs and effects. Therefore, alternative approaches, such as discrete choice experiments and/or willingness-to-pay studies, might be more appropriate for valuing these non-cost and non-health attributes (244). In these methods, respondents are given sets of hypothetical scenarios/alternatives/products and are asked to state their preferences or how much they would be willing to pay for them. Each

option is described by several attributes (e.g. speed of diagnosis, invasiveness, accuracy) and attribute levels, and responses are used to infer importance, and subsequently the value placed on each attribute and option (17, 245). Perhaps one or both these methods could be employed in a future diagnostic accuracy study of the device to understand the aspects that need to improve before it takes its final form and its final cost-effectiveness is evaluated.

4.5 Conclusion

In conclusion, the current analysis described and compared the demographic and clinical characteristics, healthcare use and diagnostic testing, HRQoL and sampling details after using a new diagnostic device (*U-Rhythm*) in healthy individuals and patients with PA, CS or AD. The results indicated differences in all these elements between the cohorts but also a general satisfaction when using the device. In the future, it is important that further studies should be conducted (using a greater number of patients and comparing the device to existing tests) to explore whether *U-Rhythm* would be a more effective option in the diagnosis of these diseases.

CHAPTER V

HEALTHCARE DATA ANALYSIS

HEALTHCARE USE AND COST OF CARE BEFORE AND AFTER THE DIAGNOSIS OF PRIMARY ALDOSTERONISM IN UK PATIENTS

CHAPTER V OVERVIEW

Chapter IV provided some limited information on the healthcare services that patients with primary aldosteronism had used three months before sampling with the *U-Rhythm* device. In addition, the systematic review that was conducted in **Chapter III** did not identify any study measuring the long-term healthcare utilisation and costs that are related to the condition. Therefore, **Chapter V** presents a costing study that aimed to describe the resources that are used in primary and secondary care for the diagnosis, monitoring and treatment of primary aldosteronism; estimate its incremental costs during the pre-, peri- and post-diagnostic phases of care; and identify the parameters that mainly affect the cost for the UK National Health Service. To do so, the study used routinely collected UK national healthcare data, and compared the healthcare use and costs that were found for primary aldosteronism patients to those of an age, gender and GP practice matched general population.

Healthcare Use and Cost of Care before and after the Diagnosis of Primary Aldosteronism in UK Patients

5.1 Background

Chapter IV briefly described the healthcare resources that endocrine patients had used three months before sampling with the *U-Rhythm* device, with their usage being compared to that of the healthy volunteers. Two limitations of the *ULTRADIAN* study were the small number of participants in all four clinical centres for all the examined endocrine disorders and the fact that data were collected using a short, self-reported, bespoke questionnaire. Therefore, the results of the abovementioned descriptive analysis are not expected to be sufficient and very robust to estimate the burden of these diseases. To examine the healthcare resources that are typically used before and after the diagnosis of an endocrine disorder as well as their associated costs for the payer, a costing study was designed. This study focuses on patients with primary aldosteronism (PA) and the United Kingdom (UK) publicly funded healthcare system (National Health Service, NHS) since it aims to provide the context and data on healthcare costs that could be used in a later decision-analytic model representing the diagnosis and treatment of the disease (**Chapter VI**). PA was selected because *U-Rhythm* has demonstrated technical promise in this clinical area.

In the UK, almost the whole population (>98%) is registered under the NHS with a primary care general practitioner (GP), who can be visited free at the point of consumption. GPs are the first point of contact for any non-emergency health-related issues. After consultation, the patient can be either treated within primary care and/or referred to secondary care, if necessary (246, 247). Patient data – identified by a unique patient NHS number – are routinely (daily basis) recorded on an electronic health record (EHR) system (e.g. Vision® or EMIS Web® system software) by the practice staff (246, 248). A subset of these (de-identified coded) clinical data are routinely (monthly basis) collected by the *Clinical Practice Research Datalink (CPRD) (248)*, formerly known as the *General Practice Research Database* (246), a governmental and not-for-profit UK health research service (248), funded by the *UK National Institute for Health Research (NIHR) (249)* and the *Medicines and Healthcare products Regulatory Agency (MHRA) (250)*. CPRD produces a large longitudinal electronic database (i.e.

GOLD or Aurum, if the data come from the Vision® or EMIS® software, respectively) that contains representative primary care data for the total UK population (246-248).

This study makes use of the CPRD GOLD database. Based on the results of the systematic review (**Chapter III**) and an exploratory literature search that was conducted (August 2020), no study answering the same question in the UK or another country was identified. However, several studies were found to use PA registry data for different purposes, such as to describe patient characteristics, the general test-treatment pathway, and the clinical outcomes and mortality following PA diagnosis/treatment (200, 251-283). Therefore, to the author's knowledge, this is the first study that estimates the healthcare use and costs for this disease, adding further economic evidence to the current literature. This information is important to help clinicians and policy makers understand the current UK clinical practice, and identify potential ways to make PA diagnosis, monitoring and treatment more efficient and cost-effective.

5.2 Methods

5.2.1 Study Design

This is a retrospective, primarily descriptive and exploratory observational cohort study (CPRD protocol number: 17_119R) whose main objective is to describe the healthcare resources that are used in primary and secondary care for the diagnosis, monitoring and treatment of PA. To do so, CPRD GOLD data are linked with data from three other English national data sources; the *Hospital Episode Statistics (HES) (284)*, the *Index of Multiple Deprivation (IMD) (285)*, and the *Office for National Statistics (ONS) mortality data (286)*. These data were requested for PA patients, and are compared to those of an age, gender and GP practice matched general population (i.e. patients without any of the endocrine diseases examined in *ULTRADIAN*). This aims to help decision-makers understand the type and number of health services that are currently used, the laboratory and imaging tests that are frequently performed, and the medications that are typically administered in the UK before and after the diagnosis of PA. The average NHS resource use and its associated cost are also measured to examine the

differences between the two cohorts, while the key parameters that affect the total cost are identified.

5.2.2 Data Source

5.2.2.1 CPRD GOLD

The source population of this study comprises all UK patients with 'research quality' data who are registered with GP surgeries that participate in CPRD GOLD and whose practices as well as themselves have also consented to participate in the data linkage scheme (248). CPRD GOLD contains anonymised data from >17 million patients from >720 GP practices across the UK (November 2017) that are collected prospectively from routine care since 1987. The records for >15 million patients meet the CPRD standards of acceptable quality for research use (i.e. registration status; recording of events; valid age and gender). In CRPD, patients with specific diseases are identified using Read Codes, which code (among other things) patient history, examination, signs, symptoms, procedures, diagnoses and therapy, and have been used in the NHS since 1985. Practices and patients are identified by a pseudo-identifier[16] to preserve anonymity and confidentiality (246, 248, 287, 288).

CPRD GOLD contains demographic and clinical data separated into ten different categories (246, 248, 288): 1) Patient (i.e. patient demographic and registration status); 2) Practice (i.e. practice region and administrative data); 3) Staff (i.e. gender and role of the staff that entered the data); 4) Consultation (i.e. administrative information about primary care consultation); 5) Clinical details (i.e. medical history/conditions); 6) Additional clinical details (i.e. additional information regarding clinical events); 7) Referral (i.e. referrals to secondary care or medical specialists/services); 8) Immunisation (i.e. patient immunisation/vaccination data); 9) Test (i.e. medical test referrals made in primary care and test results); and 10) Therapy (i.e. medications prescribed and other treatments provided in primary care).

[16] A patient that moves between different GP surgeries (registered with CPRD) receives a new identifier and their records are no longer linked with their previous data (246, 248, 287, 288).

5.2.2.2 Linkage Source Data

CPRD GOLD data are linked[17] to data from three other health-related and socioeconomic data sources; the HES (secondary care inpatient/hospitalisation data), the IMD (social deprivation data related to the patient's small area of residence and practice location) and the ONS (mortality statistics and cause of death data). The linkage is currently limited to practices in England, where almost 75% of the English practices contributing to CPRD GOLD have consented to participate, meaning that >10,550,000 patients are eligible for linkage (246, 248).

5.2.2.3 Data Access

Access to CPRD GOLD and linked datasets was provided after approval of the research protocol by the *MHRA Independent Scientific Advisory Committee (ISAC)* (246, 248). Data were provided by the University of Bristol *NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) West (290)*. Since no funding was available for obtaining the data, the author applied for the CPRD data under the 'fee exemption' scheme, an option that applies for any PhD student of the Bristol Medical School when no external funding is available. For the analysis, the most recently available data from integrated HES (set 14), patient- and practice-level IMD data for the year 2015 (stratified by quintiles), and ONS mortality data (April 2017) were obtained (***Appendix 13***). These data were linked to the most recent CPRD monthly snapshot (November 2017). The total time that CPRD (January 1987-November 2017) and HES (April 1997-March 2016) data were available is considered as the study period for CPRD and CPRD-HES data analyses, respectively.

[17] Patient-level linkage is done via a trusted third party (*NHS Digital*, formerly the *Health and Social Care Information Centre*) (246, 288). Linkage methodology is described in *Padmanabhan et al. (2019) (289)*.

5.2.3 Cohort Identification

This study contains two patient cohorts; the exposed and the unexposed (control) groups.

5.2.3.1 Exposed Cohort

The target population of this study contains patients diagnosed with PA. Male and female patients of all ages were eligible for inclusion since PA can sometimes appear early in life (e.g. genetically inherited) (46, 291). This study aimed to identify an incident cohort of patients who had a first relevant diagnostic Read Code after a period of 12 months from their start date. The start date was defined as the latest of the practice up-to-standard (UTS) date (the date that the registered GP practice starts meeting the CPRD standards) or the patient's current registration date (CRD). An exception was when the first relevant Read Code occurred within twelve months from birth and the patient was registered with an UTS practice since birth. Patients were eligible if: a) their CPRD records were 'acceptable', meaning that their primary care records satisfied a minimum set of the CPRD data quality criteria; b) had a PA diagnosis (index) date between January 1987-November 2017; and c) had an index date and CRD which occurred after the UTS practice registration for that patient. The exposed group of patients was identified in the databases using the most relevant Read Codes for PA (***Appendix 14***). Read Codes were found by searching the CPRD code browser using terms relevant to PA (e.g. 'primary aldosteronism', 'hyperaldosteronism', 'saline infusion test') and confirming that the codes found are related to the disease with the help of the endocrinologists involved in *ULTRADIAN*. Data were extracted by designated University of Bristol CPRD data holders after importing the relevant Read Codes in the CPRD web browser.

5.2.3.2 Unexposed Cohort

The control group includes a randomly selected age, sex, GP practice matched unexposed group of patients, who were also matched on eligibility for HES, IMD and ONS linkage. This general population cohort contains individuals with other conditions, including endocrine diseases other than those examined in *ULTRADIAN* (e.g. diabetes mellitus). These other diseases may have also been present as co-morbidities in the exposed cohort. The start date for data collection for each unexposed individual was within six months before or after the start date of the matching patient in the exposed group. Additionally, the age of the

unexposed patient was within ±3 years of the age of the matched exposed patient. The follow-up period for both cohorts includes the total time that CPRD and linked data were available (i.e. until death, last GP record or other specific date, e.g. transfer-out from the GP practice). When matching data, there were no restrictions in relation to the end date and post follow-up period as long as these were after or included the matching index date (i.e. date of PA diagnosis for the matched exposed patient), respectively. For the matching of the groups, a 2:1 (control:case) ratio without replacement was used to increase precision of results (292).

5.2.4 Main Analysis

Four types of analysis are conducted in this study; a descriptive, a healthcare use prediction, a cost and a cost prediction analysis. CPRD and linked data were imported and analysed using the statistical software STATA® MP 15.1 (StataCorp, Texas, USA). More information for each analysis is provided below.

5.2.4.1 Descriptive Analysis

A descriptive analysis is performed to examine the number of PA and non-PA patients; their demographic, socioeconomic and clinical characteristics; and the average number and types of healthcare resources that they used in primary and secondary care within a period of ten years before and after their index date. This twenty-year period is further divided into six timepoints (0-1 year, 1-5 years and 5-10 years before and after index date) to examine and compare trends in patients' average healthcare usage. Parametric and non-parametric tests are used to compare cohorts. For continuous variables, the sample mean, standard deviation (SD) and minimum-maximum value range are calculated, while the Student's t-test is used to test differences between the two groups. When time is measured, the median, interquartile range (IQR) and minimum-maximum value range are estimated, while the differences between the cohorts are calculated by using the Mann-Whitney U (or Wilcoxon rank sum) test. For categorical variables, the total number and overall percentages are estimated, while differences are calculated using the Pearson's chi-squared test. In both parametric and non-parametric tests, a p-value of 0.05 is used [**NOTE:** The same level of significance is used in the rest of the analyses].

5.2.4.1.1 Patient Characteristics

Demographic characteristics contain information on the patient's gender, age at index date, and marital status, while for those with linked HES data, the ethnic group is also reported. For English patients, the socioeconomic status based on the IMD of their area of residence and the location of their practice is presented, while for all patients the practice region is reported. The total time registered with the current GP practice and the time registered since index date are estimated. To do so, the start date is considered as the patient's CRD, while the end date is the earliest date of: a) that the patient was transferred out of the practice; b) the date of death; or c) the practice's latest UTS date of data collection. For both cohorts, the number of individuals that have been transferred out of the surgery is calculated, while the number of deaths and the average age at death are also estimated. ONS mortality data linkage provides additional information on the underlying and other recorded causes of death through ICD-10 diagnostic codes[18] (293). Lastly, the number and types of comorbidities are identified by using the disease categories proposed for calculating the Charlson Comorbidity Index (CCI) (294, 295), using the STATA® codes provided by *Khan et al. (2010) (296)* and *Walker et al. (2019) (297)*.

5.2.4.1.2 Healthcare Data

Healthcare resource data are divided into five categories: 1) primary care consultations; 2) medical tests/procedures used/referred to in primary care; 3) drugs prescribed in primary care; 4) referrals to secondary care/medical specialists; and 5) hospital admissions (when HES data are available). **Table 24** provides detailed information on the data included in each of these categories. Moreover, several data cleaning rules were followed before the analysis. Events that were undated or occurred more than ten years before or after the index date, before the CRD date, or after the end date were excluded. Duplicate events (e.g. more than one consultation on the same date) were removed. During each time period (e.g. 10-5 years before index date) patients who were registered for less than 20% of the time period were excluded from that analysis. For each time period, annual healthcare use was estimated.

[18] The *International Classification of Diseases and Related Health Problems (10th version, ICD-10)* is a medical classification list (e.g. diseases; signs and symptoms) published by the World Health Organization (293).

Table 24: Healthcare resource categories

Category	Sub-categories	
Primary care consultations	-Clinic -Home -Telephone -Out-of-hour -Other administrative tasks	
Diagnostic testing	<u>General:</u> -Biochemistry -Diagnostic imaging -Other diagnostic -Haematology -Microbiology -Other pathology -Serology & immunology -Other	<u>For primary aldosteronism:</u> -Blood -Creatinine blood -Serum aldosterone & renin -Urine -Chest X-ray -Blood pressure -Electrocardiogram -Other
Drug prescriptions	-Mineralocorticoid receptor antagonists -Other diuretics -Other antihypertensives -Other	
Secondary care referrals	-Cardiologist -Endocrinologist -General physician -Nephrologist -Urologist -General pathologist -Radiologist -Other specialist	
Hospital admissions (cause)	-Cardiovascular -Endocrine -Renal -Other	

5.2.4.2 Healthcare Use Prediction Analysis

Many individuals make use of a small number of healthcare resources (usually close to or at zero), with only a few patients requiring many of them. Therefore, healthcare use is expressed by non-negative integer-valued count data, which are severely right-skewed, intrinsically heteroskedastic, with variance that increases with the mean. When predicting the conditional mean or the response of the conditional mean to a parameter, it is tempting to ignore the discreteness and skewness of the data and use ordinary least squares (OLS) linear regression methods. If skewness is a concern, but not discreteness, generalised linear models (GLMs) (i.e. a flexible generalisation of OLS linear regression) can be used instead. However, when both discreteness and skewness are present, these regression methods can be quite inefficient, meaning that they suffer from substantial losses in statistical power. Additionally, there may also be a substantive interest in the estimation of event probabilities (e.g. that the number of GP visits is zero or greater than zero). In this case, a type of models, called *'count'* models, are more appropriate (298, 299).

To identify the change in healthcare use that is associated with the presence of PA (ten years before and after index date), five Poisson regression models are developed for the resource categories described in *Section 5.2.4.1.2*. As the name implies, the outcome variable is assumed to follow the Poisson distribution (i.e. its mean is equal to its variance), making them robust in estimating count parameters. Given that the healthcare resource data in this analysis are panel/longitudinal data, where the time of the event is independent of the time of the prior event, the time between events follows an exponential distribution. The Poisson distribution can inherit this property and estimate parameters that are generally consistent under the relatively weak assumption that the conditional mean is correctly specified. Poisson regression uses maximum likelihood estimation to obtain parameter estimates. This is an iterative procedure that aims to maximise a likelihood function to identify a combination of model parameter values that best represent the sample of data (298, 299).

In this study, the models developed use (the logarithm of) the number of GP consultations, tests, prescriptions, referrals or hospitalisations as the respective dependent variable and the patient characteristics that are considered likely to affect resource use as the independent variables (*'covariates'*). These include the event time, the patient's endocrine diagnosis (i.e.

PA vs non-PA), the interaction between the event time and the presence of PA, the age at index date, the gender, the number of comorbidities within two years before index date, and the patient's socioeconomic status based on IMD scores[19]. All models use the STATA® time exposure feature to account for exposure varying between individuals in each time period (e.g. due to start date, death or transfer during that period).

Poisson regression models constitute the canonical regression model type for count data and are always the starting point of any count-data analysis. However, when measuring healthcare use, they are sometimes considered inefficient and imprecise (e.g. over-dispersed data due to unobserved heterogeneity, where the variance exceeds the mean). For this reason, negative binomial regression is often considered more efficient, given that it relaxes the restrictive mean-variance property of the Poisson regression. Nevertheless, this method does not generally inherit the robustness property of the Poisson, meaning that there is often a tension between consistency under general conditions and efficiency of the estimates. It also requires the data-generating process to follow a negative binomial distribution to ensure that the parameter estimates are consistent (298, 299). In this analysis, the results of the negative binomial methods are compared to those from the Poisson models to examine whether there are any differences in the marginal effects estimated.

5.2.4.3 Cost Analysis

In this part of the analysis, the healthcare resources used are costed for each cohort using the NHS perspective. To do so, the appropriate unit costs for each resource, resource category, or event are taken from official UK national sources [**NOTE:** The specific unit costs used in the analysis are available upon request from the author]. The total (20-year) mean (SD, range) and median (IQR, range) NHS costs for each cohort are then estimated and compared, while the Student's t-test is used to identify the differences between the two groups.

[19] Given that IMD data are only available for English patients and to avoid model misspecification from omitting a large number of patient data, the average deprivation score was used as an imputed value in non-English individuals.

5.2.4.3.1 Cost Identification

GP consultations are grouped and costed based on the member of staff that entered the data into the system (expected to be the same professional that made the consultation) and the mode of contact using the *2018 Personal Social Services Research Unit (PSSRU) health and social care unit costs (300)*, a UK healthcare cost report whose development is funded by the NIHR and is updated yearly since 1993. For simplicity, all attendances are costed as single-consultant appointments, while any associated administrative task is not costed separately since these costs are already included in the health professional's overhead cost. Normal-working-hour clinic visits are costed using the costs directly taken from the PSSRU 2018 (300), while home visit costs are estimated by multiplying the respective 2018 resource costs by the home-clinic visit ratio[20] found in the PSSRU 2010 (301). For out-of-hour visits, the costs are multiplied by a day- or night-time 'home care worker' multiplier depending on the time of the visit. Furthermore, for telephone consultations, the GP- and nurse-led costs are used for GPs and nurses, respectively, while the average cost of the two is used for all other professionals. In the types of consultation where no cost is found in the PSSRU, a weighted average of all costs for the respective mode-of-contact category is used. Afterwards, unit costs are multiplied by the duration of the visit/call reported in the CPRD dataset. When this information is not available, the median duration for each professional is calculated by mode of contact. In case that both the role of staff and duration are missing, the weighted average cost and median duration per mode of contact is used.

Tests provided in primary care, secondary care referrals and hospital admissions are costed using the national tariffs collected annually from NHS providers/trusts *(NHS reference costs 2017/2018) (302)*. To do so, tests other than diagnostic imaging, which are costed separately in the NHS dataset ('unbundled' elements of care), are costed based on the groupings provided in the 'Directly Accessed Pathology Services' section (e.g. biochemistry, microbiology), while secondary referrals are costed using the costs provided in the 'Consultant-led' and 'Non-consultant-led' sections depending on whether they constituted a first or a follow-up visit to the specialist, respectively. Hospital admissions are costed based on the ICD-10 codes that were recorded at each episode. To do so, the *NHS Digital Healthcare*

[20] For GPs and all nurses, the GP-consultant and GP-nurse ratios are used, respectively, while for all other professionals, the most common community-based professional ratio is used.

Resource Group (HRG4+) Local Payment Grouper application (303) is used to combine ICD-10 with HRG4+ codes (i.e. UK's disease reimbursement codes used to reflect similar patient activity). Each ICD-10 code is given a cost that comes from the weighted average of the 'Elective Inpatients, 'Elective Inpatients Excess Bed Days', 'Non-Elective Inpatients', 'Non-Elective Inpatients Excess Bed Days', 'Non-Elective Short Stay', 'Day Case' and 'Regular Day or Night Admissions' sections based on the total annual NHS activity that is reported at each case. Since the main reason for the hospital admission cannot be found in the integrated HES dataset, the average cost of the respective HRG4+ codes is used to cost each episode. For all episodes where costs cannot be found in the NHS reference costs, a weighted average HRG4+ cost is used.

The cost of medication prescriptions is calculated using the prices that are provided in the *British National Formulary (BNF) (304)*. To do so, drugs are first classified based on the BNF chapter, section or, if possible, paragraph that they belong (e.g. spironolactone, diuretics), and then costs are attached based on a weighted average of the cost per prescription derived from the 2018 NHS Digital *Prescription Cost Analysis (PCA) (305, 306)*. If no prescription cost is found, the weighted average of all prescription costs derived from PCA is used.

5.2.4.4 Cost Prediction Analysis

To estimate the patient-average adjusted predicted NHS costs, two GLMs are developed. These models use the individual's NHS cost of care as the dependent variable and the patient's endocrine diagnosis (i.e. PA vs non-PA), age at index date, gender, number of comorbidities within two years before index date, GP registration time, and socioeconomic status as the independent variables. As in the count models described in *Section 5.2.4.2*, the covariates were chosen based on the patient characteristics that are likely to affect the patient's NHS costs.

According to *Deb et al. (2017) (299)*, GLMs are more appropriate models for predicting healthcare costs compared to simpler regression models, such as OLS linear or log-linear (i.e. the natural logarithm of the dependent variable is used) models. This is because healthcare cost outcomes, like healthcare use data, typically have three key statistical features. First, they have a distribution that is highly positively skewed (i.e. long thin right tails), reflecting

the small fraction of the population that uses a high amount of healthcare resources, leading to very high expenditures. Second, data may be heteroscedastic (i.e. non-constant standard errors) due to population variability. Third, these data are always non-negative and often have a substantial point mass (modal value) very close to or at zero. Although it would be tempting to drop observations with extreme values, it is not possible to ensure that these patients are actual outliers (e.g. individuals with coding errors in their data). Therefore, these observations should be incorporated into the analysis to accurately compute the effects of covariates on healthcare expenditures (298, 299).

When using OLS linear models for healthcare expenditures, one problem is that they may predict negative costs for some patients, since it is expected that the response variable follows a normal distribution. Additionally, linear models may not be able to estimate mean effects without bias in long-tail distributions given the large sample-to-sample variations that exist in costs, while they are likely to be inefficient in the presence of heteroskedasticity given that the error term in expenditure models does not follow a normal distribution (i.e. non-constant variance across the sample observations). Using logarithmic transformations of cost outcomes can partially address the skewness issue. However, log-linear models are not good for handling zero-cost observations, while they also require the regression output to be transformed (i.e. exponentiated) back to the original scale, which may cause bias in the presence of heteroskedasticity (298, 299).

GLMs are more flexible than OLS models since they provide the ability to use a range of functional forms (*'link functions'*) to estimate the relationship between the dependent and the linear index[21] of the independent variables. The link function allows the magnitude of the variance of the cost outcome to be a function of its predicted (mean) value by choosing the appropriate distribution – which belongs to the exponential family (e.g. gamma distribution) – from which the dependent variable is assumed to be generated. Therefore, GLMs can accommodate skewness and incorporate heteroskedasticity in natural ways. Additionally, GLMs can consider zero-value observations. Consequently, GLMs give researchers greater modelling flexibility and can fit healthcare costs much better (298, 299).

[21] This index is linear in the coefficients of the regressors but may be non-linear in the underlying covariates.

To identify the type of GLM that best fits the data of this study, several specification tests were applied. Firstly, to find the appropriate link function, the Box-Cox test (307) on positive cost observations was used. This approach tests the scalar power of the dependent variable (y^{δ}) that results in the most symmetric transformed distribution[22], since the method targets skewness in the error term. Secondly, to identify the relationship between the variance of the error term and the expected value of the cost-outcome variable (GLM's distribution family), a modified version of the Park test (308) was used. This test is conducted after deciding (based on assumptions, e.g. from literature) on the link function and distribution family that will be used in the GLM. Here, the expected value (conditional on the covariates) and the squared error (variance) for each observation are first calculated, and then a regression of the logarithm of the variance on the mean value is performed[23].

Thirdly, to test the specification of the explanatory variables, a Pregibon-type link test (309, 310) with a traditional 5% significance level was performed. Although this test is typically done in OLS models, a modified version can be used in GLMs. The Pregibon test examines whether a model that includes a squared cost variable as a covariate has more explanatory power than a model excluding this term (298, 299). Lastly, to check the quality and appropriateness of the chosen model (i.e. both link function and family distribution simultaneously) relative to each of the other models that could be used for the data, the Akaike information criterion (AIC) (311) and Bayesian information criterion (BIC) (312) were estimated, since these have shown robustness to model misspecification (313). The AIC and BIC are used to estimate the amount of information that is lost by a given model. The less information that is lost (lower AIC/BIC results), the better the quality of the model. All models used robust standard errors (i.e. robust to potential misspecification given that the observations are independent).

[22] A power of $\delta=1$ corresponds to a linear model, $\delta=0.5$ to the square root transformation, and $\delta=0$ to the natural log transformation model. When $\delta<1$, the Box-Cox transformation makes right-skewed data more symmetric; when $\delta>1$, it makes left-skewed data more symmetric.

[23] Depending on the coefficient on the expected value, the appropriate distribution is decided (≈0: Gaussian; ≈1: Poisson-type; ≈2: gamma; and ≈3: inverse-Gaussian).

5.2.5 Sensitivity Analysis

To examine the robustness of the results, two types of sensitivity analysis (SA) are conducted. First, the impact of the number of 'ghost' patients on the final outcomes of all the components of the main analysis is measured. 'Ghost' patients are patients registered with a GP surgery and reported in its EHR, although, in reality, they have died or moved away (e.g. registered with another practice). In England, more than three million 'ghost' patients have recently been reported based on NHS information (314). In this study, 'ghost' patients are considered the individuals that had not had any GP consultations in i) the last one, or ii) the last five years before index date. Second, to examine the robustness of all outcomes, these are re-examined after removing patients with 'unreasonably' high usage (potential outliers) in each one of the five resource categories (i.e. consultations, tests, prescriptions, referrals, hospitalisations). To do so, the 1% (of all patients) with the highest number of resources per category per year over the entire study period (i.e. 20 years) is removed and the analyses are re-run.

5.3 Results

5.3.1 Cohort Numbers

475 PA patients were found to meet the eligibility criteria, from whom two had no matches in the general population and were omitted. Therefore, 1,419 patients (473 and 946 for the exposed and unexposed group, respectively) were included. From these, data from 292 PA and 584 non-PA patients (overall 876 patients) were linked with HES, IMD and ONS data.

5.3.2 Descriptive Analysis

5.3.2.1 Patient Characteristics

By design, the two cohorts were perfectly matched in terms of their GP practice (**Table 25**) and gender (**Table 26**), while there was only a slight difference in age (≈55 years old; p=0.916). No significant variations were also seen in their marital status (p=0.277) and socioeconomic status based on the individual's residence area (p=0.496). For patients with HES data, there

was a significant difference in their ethnic group, with the exposed group containing a higher percentage of black patients (8.22% vs 0.86%; p<0.001). Significant differences between the two cohorts were also found when examining the number of comorbidities, with a higher percentage of individuals within the PA group having at least one comorbidity within two years before index date (21.99% vs 13.32%; p<0.001). Specifically, circulatory/heart and renal diseases were more prevalent in the PA group.

When looking at patient characteristics after index date (**Table 27**), a similar percentage of cases transferred out of the GP surgery and deaths were demonstrated, with the age at death being slightly lower for the non-PA cohort (76.59 vs 74.40; p=0.188). A higher proportion of PA patients had circulatory, endocrine and/or renal as their underlying and/or other recorded cause of death, although when compared to the non-PA cohort, variations were not statistically significant. Non-statistically significant differences were also shown in the total time and time after index date that the patients were registered with their GP practice (**Figure 19**). **Figure 20** presents the Kaplan-Meier survival estimates after index date for the two cohorts indicating that survival dropped moderately faster at ≈6 and ≈16 years for PA patients. **Table 28** presents the results of the Cox regression which showed no significant difference in survival between the PA and non-PA groups, and indicated that age at index date had the greatest impact on predicting survival.

Table 25: General practice characteristics

General Practice Characteristics	PA (n = 473)	Non-PA (n = 946)	Difference between Cohorts (p-value)
Small area practice region; n (%)			
East Midlands	10 (2.11)	20 (2.11)	
East of England	46 (9.73)	92 (9.73)	
London	57 (12.05)	114 (12.05)	
North East	9 (1.90)	18 (1.90)	
North West	47 (9.94)	94 (9.94)	
Northern Ireland	12 (2.54)	24 (2.54)	$chi^2(12) = 0.000$
Scotland	54 (11.42)	108 (11.42)	p = 1.000
South Central	47 (9.94)	94 (9.94)	
South East Coast	48 (10.15)	96 (10.15)	
South West	45 (9.51)	90 (9.51)	
Wales	40 (8.46)	80 (8.46)	
West Midlands	40 (8.46)	80 (8.46)	
Yorkshire and the Humber	18 (3.81)	36 (3.81)	
Socioeconomic status based on practice area (quintiles)[24]; n (%)			
<20% (least deprived)	51 (17.11)	102 (17.11)	
20-40%	52 (17.45)	104 (17.45)	$chi^2(4) = 0.000$
40-60%	65 (21.81)	130 (21.81)	p = 1.000
60-80%	59 (19.80)	118 (19.80)	
>80% (most deprived)	71 (23.83)	142 (23.83)	

**Abbreviations:* Non-PA: without primary aldosteronism; PA: primary aldosteronism*

[24] Only for patients with IMD data (292 PA, 584 non-PA).

Table 26: Patient characteristics before index date

Patient Characteristics	PA (n = 473)	Non-PA (n = 946)	Difference between Cohorts (p-value)
Gender; n (%)			
Female	187 (39.53)	374 (39.53)	$chi^2(1) = 0.000$
Male	286 (60.47)	572 (60.47)	p = 1.000
Age at index date (years)[25]; mean (SD, range)	55.20 (13.94, 0-93)	55.12 (13.93, 0-93)	t(1417) = -0.105 p = 0.916
Ethnicity[26]; n (%)			
Black	24 (8.22)	5 (0.86)	$chi^2(3) = 66.852$ p < 0.001
White	224 (76.71)	395 (67.64)	
Other	17 (5.82)	22 (3.77)	
Unknown	27 (9.25)	162 (27.74)	
Marital status; n (%)			
Co-habiting/Married	86 (18.18)	169 (17.86)	
Divorced/Separated/Widowed	11 (2.33)	9 (0.95)	$chi^2(4) = 5.107$
Single	18 (3.81)	33 (3.49)	p = 0.277
Data not entered	87 (18.39)	163 (17.23)	
Unknown	271 (57.29)	572 (60.47)	
Socioeconomic status based on residence area (quintiles)[27]; n (%)			
<20% (least deprived)	73 (24.91)	144 (24.66)	
20-40%	60 (20.48)	137 (23.46)	$chi^2(4) = 3.382$
40-60%	63 (21.50)	126 (21.58)	p = 0.496
60-80%	57 (19.45)	88 (15.07)	
>80% (most deprived)	40 (13.65)	89 (15.24)	

[25] Two PA and four non-PA patients were <12 months old. Age '0' means <12 months old at index date.

[26] Only for patients with HES data (292 PA, 584 non-PA).

[27] Only for patients with IMD data (292 PA, 584 non-PA).

Patient Characteristics	PA (n = 473)	Non-PA (n = 946)	Difference between Cohorts (p-value)
Number of comorbidities (based on Charlson's categories)[28]; n (%)			
0	349 (73.78)	801 (84.67)	$chi^2(3) = 24.889$ $p < 0.001$
1	104 (21.99)	126 (13.32)	
2	17 (3.59)	16 (1.69)	
3	3 (0.63)	3 (0.32)	
Type of comorbidities (based on Charlson's categories); n (%)			
AIDS	-	1 (0.10)	
Cancer	9 (1.81)	19 (1.96)	
Cerebrovascular disease	14 (2.82)	6 (0.62)	
Chronic pulmonary disease	26 (5.24)	42 (4.34)	
Congestive heart disease	10 (2.02)	4 (0.41)	
Dementia	1 (0.20)	1 (0.10)	
Diabetes	47 (9.48)	51 (5.27)	
Diabetes with complications	7 (1.41)	7 (0.72)	$chi^2(14) = 56.497$ $p < 0.001$
Hemiplegia	-	-	
Metastatic tumour	1 (0.20)	1 (0.10)	
Mild liver disease	-	2 (0.21)	
Moderate liver disease	-	-	
Myocardial infarction	6 (1.21)	2 (0.21)	
Peptic ulcer disease	-	-	
Peripheral vascular disease	3 (0.60)	8 (0.83)	
Renal disease	21 (4.23)	15 (1.55)	
Rheumatological disease	2 (0.40)	8 (0.83)	
None	349 (70.36)	801 (82.75)	

**Abbreviations:* *Non-PA: without primary aldosteronism; PA: primary aldosteronism; SD: standard deviation*

[28] The comorbidities that were present within two years before index date are demonstrated.

Table 27: Patient characteristics after index date

Patient Characteristics	PA (n = 473)	Non-PA (n = 946)	Difference between Cohorts (p-value)
Number of cases transferred out of the GP; n (%)	143 (30.23)	314 (33.19)	$chi^2(1) = 1.265$ p = 0.261
Number of deaths[29]; n (%)	71 (15.01)	133 (14.06)	$chi^2(1) = 0.232$ p = 0.630
Age at death (years); mean (SD, range)	76.59 (10.40, 44-99)	74.40 (11.76, 41-96)	t(202) = -1.320 p = 0.188
Underlying cause of death (ONS data only); n (%)			
Circulatory, endocrine or renal	21 (44.68)	27 (29.35)	$chi^2(1) = 3.235$
Other disease	26 (55.32)	65 (70.65)	p = 0.072
Other recorded cause of death (ONS data only); n (%)			
Circulatory, endocrine or renal	30 (63.83)	51 (55.43)	$chi^2(1) = 0.902$
Other disease	17 (36.17)	41 (44.57)	p = 0.342
Total time registered with GP (years); median (IQR, range)	20.92 (19.09, 1.95-85.34)	21.71 (18.69, 1.25-85.81)	z = -0.127 p = 0.899
Time registered with GP since index date (years); median (IQR, range)	5.31 (7.01, 0.03-25.90)	5.40 (6.55, 0.02-25.90)	z = -0.066 p = 0.947

**Abbreviations: GP: general practice; IQR: interquartile range; Non-PA: without primary aldosteronism; ONS: Office for National Statistics; PA: primary aldosteronism; SD: standard deviation*

[29] CPRD: 45 (9.51%) PA and 102 (10.78%) non-PA deaths
ONS: 47 (9.94%) PA and 92 (9.73%) non-PA deaths

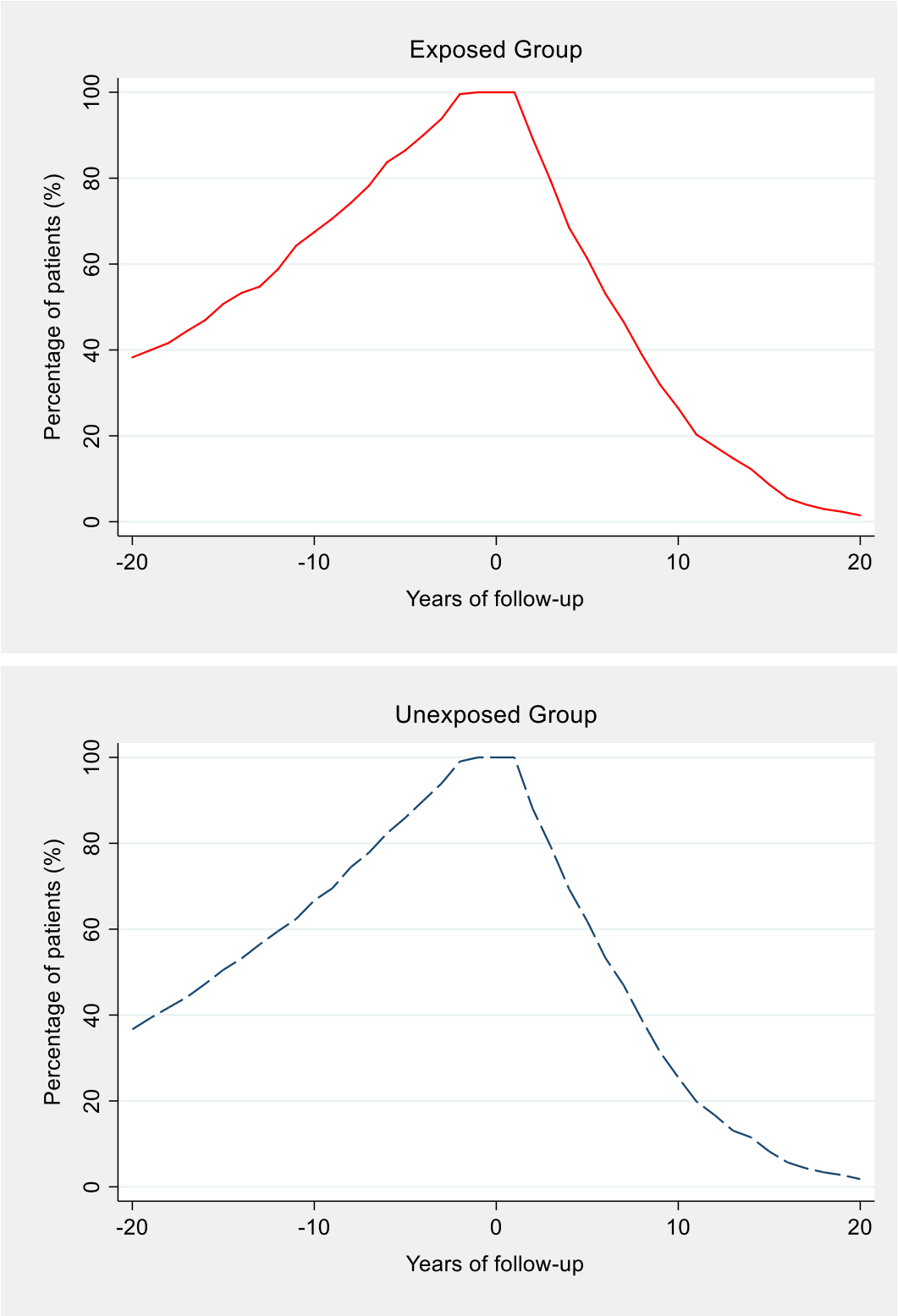


Figure 19: Exposed and unexposed cohort follow-up 20 years before and after index date

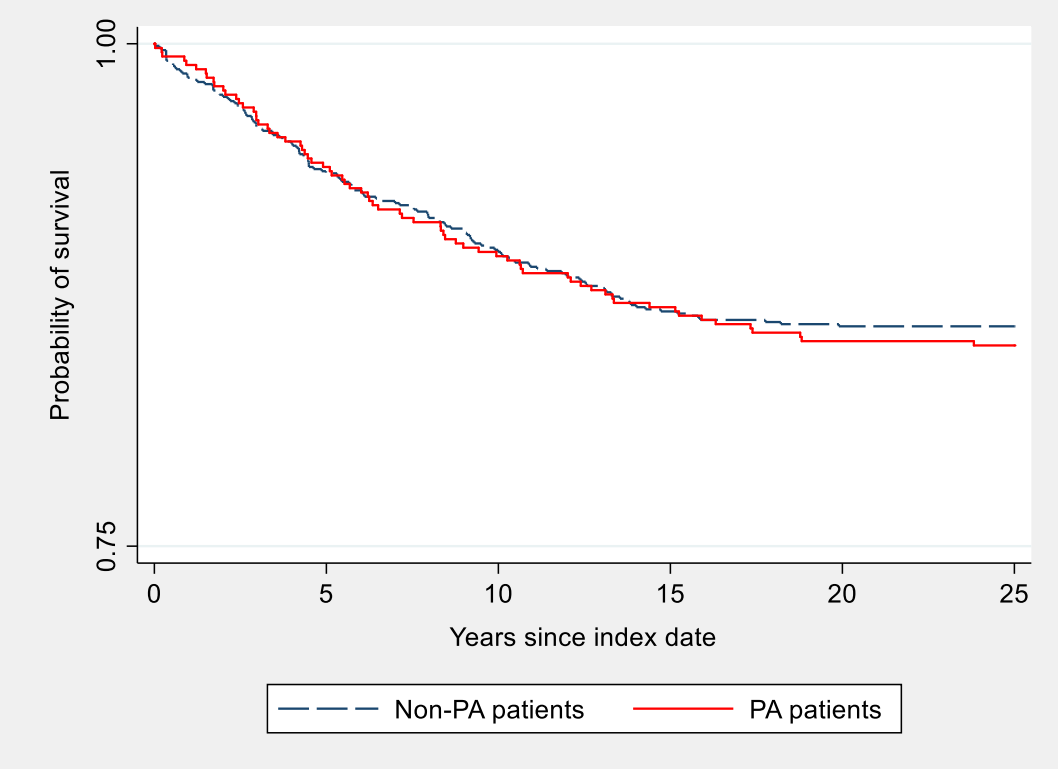


Figure 20: Kaplan-Meier survival estimates for non-PA and PA patients

Table 28: Cox regression model

Variables	Hazard Ratio	Standard Error	z	p-value
Age at index date	1.10	0.01	14.71	<0.001
PA	1.12	0.17	0.76	0.445
Number of observations = 1,419		Likelihood ratio chi^2(2) = 256.26		p-value < 0.001
Number of failures/deaths = 204		Log likelihood = -1,337.10		
Time at risk = 31,798.07				
Test of proportional-hazards assumptions (global test):				
chi^2(2) = 2.55		p-value = 0.280		

Abbreviations: PA: primary aldosteronism

5.3.2.2 Healthcare Use

Tables 29-30 provide a descriptive summary of the twenty-year (ten years before and after index date) average annual healthcare use (per resource category) for PA and non-PA patients (***Appendix 15*** gives more details). For PA patients, healthcare use was increasing before diagnosis, peaking in the peri-diagnosis period (i.e. one year before and after index date). After diagnosis, numbers slightly dropped but remained higher than the pre-diagnosis period, while they started increasing again after five years post-diagnosis. For non-PA patients, the annual use of healthcare resources increased relatively linearly and almost doubled in the five- to ten-year post-index-date period compared to the five- to ten-year pre-index-date period for almost all resource categories (hospital admissions increased by seven times). Overall, healthcare utilisation was higher in the PA group on all measures and at all timepoints, with the difference being most pronounced in the peri-diagnosis period. Additionally, the numbers of both GP consultations and drug prescriptions were higher for PA patients even 5-10 years before the first diagnosis of the disease.

Table 29: Average annual healthcare resource use (SD, range) for PA patients before and after index date

Time / Resource Category	5-10 years before index date (n = 370)	1-5 years before index date (n = 447)	0-1 years before index date (n = 471)	0-1 years after index date (n = 468)	1-5 years after index date (n = 383)	5-10 years after index date (n = 220)
Primary care consultations	6.84 (6.50, 0-47.2)	9.58 (6.93, 0-59.75)	14.90 (10.24, 0-73)	14.40 (12.49, 0-123)	10.55 (8.15, 0-46.25)	10.63 (8.78, 0-50)
Diagnostic testing	2.14 (2.83, 0-15.21)	4.84 (5.07, 0-32.5)	9.28 (9.31, 0-72)	8.17 (9.55, 0-86)	6.96 (6.95, 0-42)	7.72 (6.58, 0-31.98)
Drug prescriptions	9.18 (10.90, 0-68.4)	12.55 (10.90, 0-64.25)	18.67 (14.55, 0-86)	18.68 (16.08, 0-104.44)	16.76 (15.34, 0-86.25)	18.48 (19.35, 0-139.22)
Secondary care referrals	0.47 (0.64, 0-4)	0.76 (1.02, 0-8.25)	1.30 (1.98, 0-27)	0.89 (1.71, 0-20)	0.71 (0.92, 0-5.25)	0.76 (0.99, 0-5.81)
Hospital admissions (HES patients only)	**(n = 228)**	**(n = 276)**	**(n = 292)**	**(n = 288)**	**(n = 235)**	**(n = 139)**
	0.13 (0.28, 0-1.6)	0.28 (0.64, 0-8)	0.89 (1.30, 0-8)	1.10 (2.41, 0-22)	0.68 (2.15, 0-30.75)	1.68 (11.59, 0-133.25)

**Abbreviations:* HES: Hospital Episode Statistics*

Table 30: Average annual healthcare resource use (SD, range) for non-PA patients before and after index date

Time / Resource Category	5-10 years before index date (n = 736)	1-5 years before index date (n = 898)	0-1 years before index date (n = 945)	0-1 years after index date (n = 932)	1-5 years after index date (n = 774)	5-10 years after index date (n = 444)
Primary care consultations	4.43 (4.92, 0-43)	5.39 (5.61, 0-49)	5.76 (6.60, 0-44)	6.17 (7.60, 0-61)	6.39 (6.82, 0-50.5)	7.54 (7.82, 0-65.08)
Diagnostic testing	1.41 (2.62, 0-24.8)	2.33 (3.61, 0-25.25)	3.00 (5.28, 0-60)	3.30 (5.60, 0-43)	3.76 (5.05, 0-44.42)	5.03 (6.81, 0-73.45)
Drug prescriptions	5.02 (9.82, 0-126.8)	6.53 (9.98, 0-81)	7.44 (11.67, 0-93)	7.93 (12.25, 0-102.95)	8.73 (12.86, 0-126.47)	11.17 (13.37, 0-102.56)
Secondary care referrals	0.38 (0.70, 0-6.6)	0.41 (0.66, 0-6.75)	0.44 (0.99, 0-11)	0.46 (1.08, 0-12.03)	0.48 (0.85, 0-13.15)	0.51 (0.70, 0-3.92)
Hospital admissions (HES patients only)	**(n = 445)** 0.10 (0.26, 0-2.2)	**(n = 558)** 0.16 (0.47, 0-7.32)	**(n = 584)** 0.20 (0.67, 0-6)	**(n = 575)** 0.36 (1.29, 0-14.22)	**(n = 474)** 0.35 (1.00, 0-11.25)	**(n = 267)** 0.73 (3.41, 0-52.2)

**Abbreviations:* HES: Hospital Episode Statistics*

5.3.3 Healthcare Use Prediction Analysis

Tables 31-35 present the exponentiated outcomes of the healthcare use Poisson regression models. Here, the coefficients are expressed as the percent change in the use of the examined healthcare resource (e.g. diagnostic testing) for every category/unit change in the respective independent variable (e.g. event time, age), if all other variables are held constant. As shown, in all resource categories apart from hospital admissions where numbers were lower (n=876), the model parameters were estimated from 5,691 observations in 1,419 patients. Moreover, in all models, as the clinical event time moved further from 5-10 years before index date and age increased, the number of resources used increased as well ($p<0.05$). A significant increase in healthcare utilisation for all resource categories apart from hospital admissions (increase with $p=0.468$) was also indicated when PA was present.

When looking at the effect of the interaction between the event time and the presence of PA, this was found to be positive and significant in the peri-diagnosis period, showing that PA patients required more healthcare than the general population during that period. This interaction became less significant as the time moved away from the index date. Furthermore, gender seemed to affect the number of healthcare resources that were required, with men using slightly less primary care services ($p<0.003$) but being admitted slightly more times in hospital ($p=0.825$). Another factor that was in most cases significant was the number of comorbidities, with the models indicating that their presence was related to more healthcare resources being used. In contrast, in all models, the socioeconomic status of the patient seemed to not have a significant impact on the number of health services that were used. When the outcomes of the Poisson regression were compared to those of the negative binomial (***Appendix 16***), no differences in the variables that had a significant effect on healthcare usage were identified.

Table 31: Prediction of total primary care consultations using Poisson regression

Variable	IRR	Standard Error	z	p-value
Event time[a]				
1-5 years before	1.243	0.043	6.27	<0.001
1 year before & after	1.378	0.056	7.90	<0.001
1-5 years after	1.518	0.065	9.75	<0.001
5-10 years after	1.769	0.082	12.30	<0.001
PA[b]	1.480	0.090	6.43	<0.001
Event time # PA[c]				
1-5 years before	1.128	0.062	2.18	0.029
1 year before & after	1.595	0.102	7.32	<0.001
1-5 years after	1.058	0.071	0.85	0.396
5-10 years after	0.942	0.072	-0.79	0.430
Age at index date	1.016	0.001	11.18	<0.001
Gender[d]	0.743	0.029	-7.62	<0.001
Comorbidities[e]				
One	1.506	0.074	8.36	<0.001
Two	1.437	0.161	3.25	0.001
Three	2.123	0.401	3.98	<0.001
Deprivation[f]				
Less deprived	1.035	0.072	0.50	0.620
Average	1.012	0.055	0.23	0.820
Deprived	1.011	0.079	0.14	0.887
Most deprived	1.084	0.082	1.07	0.283
Constant	1.889	0.191	6.31	<0.001
Observation time[g]	1 (exposure)			
Number of observations = 5,691		Number of groups = 1,419		
Wald $chi^2(18)$ = 1,160.69		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Reference categories:* ***a, c *– '5-10 years before index date';* ***b*** *– 'non-PA';* ***d*** *– 'female';* ***e*** *– 'no comorbidities';* ***f*** *– 'least deprived';* ***g*** *– consideration of patient's GP registration time*

Table 32: Prediction of total diagnostic testing using Poisson regression

Variable	IRR	Standard Error	z	p-value
Event time[a]				
1-5 years before	1.726	0.094	10.08	<0.001
1 year before & after	2.365	0.159	12.77	<0.001
1-5 years after	2.969	0.203	15.91	<0.001
5-10 years after	4.142	0.313	18.80	<0.001
PA[b]	1.404	0.134	3.56	<0.001
Event time # PA[c]				
1-5 years before	1.357	0.101	4.08	<0.001
1 year before & after	1.869	0.174	6.73	<0.001
1-5 years after	1.208	0.115	1.99	0.047
5-10 years after	1.080	0.111	0.75	0.451
Age at index date	1.023	0.002	11.65	<0.001
Gender[d]	0.757	0.040	-5.27	<0.001
Comorbidities[e]				
One	1.654	0.105	7.96	<0.001
Two	1.679	0.274	3.17	0.002
Three	1.483	0.444	1.32	0.188
Deprivation[f]				
Less deprived	1.006	0.090	0.06	0.949
Average	1.033	0.075	0.45	0.654
Deprived	0.920	0.092	-0.83	0.407
Most deprived	0.966	0.095	-0.35	0.725
Constant	0.375	0.055	-6.64	<0.001
Observation time[g]	1 (exposure)			
Number of observations = 5,691		Number of groups = 1,419		
Wald $chi^2(18)$ = 1,443.06		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Reference categories:* ***a, c *– '5-10 years before index date';* ***b*** *– 'non-PA';* ***d*** *– 'female';* ***e*** *– 'no comorbidities';* ***f*** *– 'least deprived';* ***g*** *– consideration of patient's GP registration time*

Table 33: Prediction of total drug prescriptions using Poisson regression

Variable	IRR	Standard Error	z	p-value
Event time[a]				
1-5 years before	1.327	0.084	4.49	<0.001
1 year before & after	1.572	0.105	6.79	<0.001
1-5 years after	1.772	0.123	8.28	<0.001
5-10 years after	2.262	0.164	11.24	<0.001
PA[b]	1.735	0.160	5.97	<0.001
Event time # PA[c]				
1-5 years before	1.036	0.080	0.46	0.644
1 year before & after	1.335	0.114	3.39	0.001
1-5 years after	1.068	0.097	0.73	0.467
5-10 years after	0.926	0.101	-0.71	0.481
Age at index date	1.024	0.002	11.41	<0.001
Gender[d]	0.845	0.049	-2.92	0.003
Comorbidities[e]				
One	1.708	0.110	8.35	<0.001
Two	1.419	0.201	2.47	0.013
Three	1.806	0.408	2.62	0.009
Deprivation[f]				
Less deprived	1.086	0.111	0.80	0.424
Average	1.096	0.089	1.14	0.255
Deprived	1.137	0.118	1.24	0.215
Most deprived	1.320	0.151	2.43	0.015
Constant	1.087	0.174	0.52	0.602
Observation time[g]	1 (exposure)			
Number of observations = 5,691		Number of groups = 1,419		
Wald $chi^2(18)$ = 936.88		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Reference categories:* ***a, c *– '5-10 years before index date';* ***b*** *– 'non-PA';* ***d*** *– 'female';* ***e*** *– 'no comorbidities';* ***f*** *– 'least deprived';* ***g*** *– consideration of patient's GP registration time*

Table 34: Prediction of total secondary care referrals using Poisson regression

Variable	IRR	Standard Error	z	p-value
Event time[a]				
1-5 years before	1.105	0.074	1.48	0.139
1 year before & after	1.219	0.107	2.26	0.024
1-5 years after	1.316	0.111	3.26	0.001
5-10 years after	1.390	0.120	3.81	<0.001
PA[b]	1.229	0.120	2.11	0.035
Event time # PA[c]				
1-5 years before	1.455	0.153	3.56	<0.001
1 year before & after	1.958	0.254	5.18	<0.001
1-5 years after	1.179	0.142	1.37	0.170
5-10 years after	1.178	0.156	1.24	0.215
Age at index date	1.014	0.002	5.92	<0.001
Gender[d]	0.780	0.044	-4.43	<0.001
Comorbidities[e]				
One	1.342	0.096	4.12	<0.001
Two	1.307	0.254	1.38	0.169
Three	1.631	0.673	1.19	0.236
Deprivation[f]				
Less deprived	0.997	0.092	-0.03	0.975
Average	0.911	0.070	-1.22	0.221
Deprived	0.771	0.077	-2.60	0.009
Most deprived	0.951	0.120	-0.39	0.693
Constant	0.192	0.033	-9.70	<0.001
Observation time[g]	1 (exposure)			
Number of observations = 5,691		Number of groups = 1,419		
Wald chi^2(18) = 267.95		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Reference categories:* ***a, c *– '5-10 years before index date';* ***b*** *– 'non-PA';* ***d*** *– 'female';* ***e*** *– 'no comorbidities';* ***f*** *– 'least deprived';* ***g*** *– consideration of patient's GP registration time*

Table 35: Prediction of total hospital admissions using Poisson regression

Variable	IRR	Standard Error	z	p-value
Event time[a]				
1-5 years before	1.517	0.197	3.21	0.001
1 year before & after	2.807	0.469	6.18	<0.001
1-5 years after	3.661	0.675	70.4	<0.001
5-10 years after	8.254	3.219	5.41	<0.001
PA[b]	1.167	0.249	0.73	0.468
Event time # PA[c]				
1-5 years before	1.463	0.297	1.87	0.061
1 year before & after	2.872	0.668	4.53	<0.001
1-5 years after	1.515	0.569	1.11	0.268
5-10 years after	0.952	0.578	-0.08	0.936
Age at index date	1.025	0.011	2.29	0.022
Gender[d]	1.045	0.208	0.22	0.825
Comorbidities[e]				
One	1.961	0.680	1.94	0.052
Two	1.513	0.399	1.57	0.116
Three	1.564	0.875	0.80	0.425
Deprivation[f]				
Less deprived	0.965	0.178	-0.20	0.845
Average	1.301	0.458	0.75	0.455
Deprived	1.352	0.481	0.85	0.397
Most deprived	1.222	0.240	1.02	0.309
Constant	0.016	0.014	-4.89	<0.001
Observation time[g]	1 (exposure)			
Number of observations = 3,508		Number of groups = 876		
Wald chi^2(18) = 440.81		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Reference categories:* ***a, c *– '5-10 years before index date';* ***b*** *– 'non-PA';* ***d*** *– 'female';* ***e*** *– 'no comorbidities';* ***f*** *– 'least deprived';* ***g*** *– consideration of patient's GP registration time*

5.3.4 Cost Analysis

Estimates of the total (twenty-year) mean and median healthcare costs and their components (i.e. resource categories) for PA and non-PA patients are shown in **Tables 36-37**. As indicated, drug prescription costs dominated healthcare costs for all CPRD patients, comprising 55-70% of total costs (hospital admissions excluded). For PA patients, mean therapy costs were almost double that of non-PA individuals ($t_{1,417}$=-9.439; p<0.001), while median medication costs were approximately three times higher (z=-13.318; p<0.001). The second resource category with the highest costs for all patients was the 'primary care consultations'. Here, mean and median PA costs were twice as high as non-PA costs (mean/median: $t_{1,417}$=-11.114/z=-13.288; p<0.001) indicating the higher number of GP visits that were needed before and after PA was diagnosed. An approximately similar pattern of variations was seen in mean and median diagnostic testing and secondary care referral costs, where again the differences between the two groups were significant. In patients with HES data, the difference in their mean inpatient costs was not significant (t_{874}=-1.589; p=0.112), while that in their median costs was significant (z=-9.629; p<0.001). This possibly indicated that there were a few patients in both groups that have very high inpatient costs, so this was examined in SA (see *Section 5.3.6* below). Overall, during a ten-year period before and after diagnosis, a PA patient costs the NHS £22,352.70, while a non-PA patient costs £12,188.82 (p<0.001). When hospitalisation costs are also measured, mean costs rise to £45,263.66 and £26,175.01 for PA and non-PA individuals, respectively.

Table 36: Average costs (SD; range) per healthcare resource category

Healthcare Resource Category	PA	Non-PA	Difference between Cohorts (p-value)
Primary care consultations	£3,964.83 (3,250.94; 0-25,278.61)	£2,189.16 (2,605.79; 0-41,740.34)	t(1417) = -11.114 p < 0.001
Diagnostic testing	£1,133.29 (1,237.63; 0-7,878.94)	£708.96 (999.22; 0-8,730.16)	t(1417) = -6.948 p < 0.001
Drug prescriptions	£15,670.89 (15,989.55; 0-112,841.60)	£8,360.16 (12,488.69; 0-103,273.70)	t(1417) = -9.439 p < 0.001
Secondary care referrals	£1,583.69 (1,641.95; 0-12,013.75)	£930.54 (1,174.19; 0-11,236.24)	t(1417) = -8.603 p < 0.001
Hospital admissions (HES patients only)	£21,889.33 (59,952.78; 0-708,984.50)	£14,304.05 (69,673.96; 0-1,439,457.00)	t(874) = -1.589 p = 0.112
Total costs (all patients) (without hospital admissions)	**£22,352.70 (19,720.21; 695.99-135,365.40)**	**£12,188.82 (15,424.48; 0-123,878.30)**	**t(1417) = -10.632 p < 0.001**
Total costs (HES patients only)	**£45,263.66 (66,473.30; 2,381.61-755,263.70)**	**£26,175.01 (73,955.50; 0-1,480,879.00)**	**t(874) = -3.722 p < 0.001**

**Abbreviations:* HES: Hospital Episode Statistics; Non-PA: without primary aldosteronism; PA: primary aldosteronism*

Table 37: Median costs (IQR; range) per healthcare resource category

Healthcare Resource Category	PA	Non-PA	Difference between Cohorts (p-value)
Primary care consultations	£3,013.38 (3,561.81; 0-25,278.61)	£1,452.52 (2,419.83; 0-41,740.34)	z = -13.288 p < 0.001
Diagnostic testing	£754.50 (1,139.33; 0-7,878.94)	£349.91 (832.18; 0-8,730.16)	z = -9.513 p < 0.001
Drug prescriptions	£10,581.87 (15,071.10; 0-112,841.60)	£3,592.08 (9,818.88; 0-103,273.70)	z = -13.318 p < 0.001
Secondary care referrals	£1,127.44 (1,541.76; 0-12,013.75)	£555.46 (1,145.65; 0-11,236.24)	z = -10.458 p < 0.001
Hospital admissions (HES patients only)	£10,938.05 (17,304.07; 0-708,984.50)	£2,736.43 (10,352.78; 0-1,439,457.00)	z = -9.629 p < 0.001
Total costs (all patients) (without hospital admissions)	**£16,047.64 (19,746.87; 695.99-135,365.40)**	**£6,597.48 (14,324.40; 0-123,878.30)**	**z = -13.628 p < 0.001**
Total costs (HES patients only)	**£28,243.87 (35,637.78; 2,381.61-755,263.70)**	**£10,708.54 (24,131.64; 0-1,480,879.00)**	**z = -11.481 p < 0.001**

**Abbreviations:* HES: Hospital Episode Statistics; Non-PA: without primary aldosteronism; PA: primary aldosteronism*

5.3.5 Cost Prediction Analysis

To predict the patient-average adjusted NHS costs that are related to PA ten years before and after the diagnosis of the disease, a GLM that uses the square root as a link function and the gamma distribution for the data was used. This type of GLM was found to fit the data best based on the four specification tests described in *Section 5.2.4.4*. Specifically, the Box-Cox test showed that a scalar power of 0.5 (i.e. square root) led to a more symmetric transformed distribution; the Park test indicated that a square root link function and a gamma distribution should be used; the Pregibon test demonstrated that a squared cost variable did not give more explanatory power to the model; and AIC/BIC took lower values than those found when other link functions (e.g. log) and distributions (e.g. Gaussian) were explored. **Tables 38-39** present the marginal and incremental effects of the independent variables on healthcare expenditure for patients with and without HES data, respectively. For comparison purposes and given the small variations in parameters between the two types of models, ***Appendix 17*** provides the parameter values if a GLM with a log link function and a gamma distribution was to be used instead.

Table 38 illustrates that the presence of PA added £9,013.77 ($p<0.001$) to the total cost of care of the patient. One additional year of age and GP registration costed an extra £294.77 ($p<0.001$) and £212.81 ($p<0.001$), respectively, for the NHS. Men tended to cost the NHS less than women, with the average incremental difference being £4,445.62 ($p<0.001$). The presence of one comorbidity was related to an additional cost of £9,416.70 ($p<0.001$), while this cost rose to £10,830.89 ($p=0.073$) when the patient had three comorbidities. The socioeconomic status of the patients did not significantly affect the total cost of health care. When looking at the results of the GLM developed from patients with HES data (**Table 39**), the pattern was similar. More precisely, the presence of PA led to an additional cost of £18,593.46 ($p<0.001$) for the NHS, while one year more of age and registration with the GP caused a marginal effect of £833.01 ($p<0.001$) and £228.65 ($p=0.049$), respectively. Gender again played an important role, with men now costing the NHS £9,799.29 ($p=0.001$) less than women. Moreover, patients with one or more comorbidities had higher costs than those without comorbidities recorded, while patients with two or three comorbidities did not cost

more than those with one. Lastly, the individual's socioeconomic status based on their residential area was not significantly associated with healthcare costs.

Table 38: Prediction of total healthcare costs (without hospital admissions) using GLM with square root link function and gamma distribution

Variable	dy/dx	Standard Error	z	p-value
PA[a]	9,013.77	837.04	10.77	<0.001
Age at index date	294.77	22.37	13.18	<0.001
Gender[b]	-4,445.62	945.60	-4.70	<0.001
Comorbidities[c]				
One	9,416.70	1,316.41	7.15	<0.001
Two	9,435.36	2,776.83	3.40	0.001
Three	10,830.89	6,046.33	1.79	0.073
Years of GP registration	212.81	37.82	5.63	<0.001
Deprivation[d]				
Less deprived	316.47	1,274.42	0.25	0.804
Average	557.88	1,152.21	0.48	0.628
Deprived	397.20	1,615.64	0.25	0.806
Most deprived	3,265.19	1,781.49	1.83	0.067
Number of observations = 1,419		Residual df = 1,407		
Scale parameter = 1.30				
Deviance = 1,412.86		(1/df) Deviance = 1.00		
Pearson = 1,829.57		(1/df) Pearson = 1.30		
AIC = 20.93		BIC = -8,798.74		
Log pseudolikelihood = -14,837.35				

Abbreviations: AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism

***Reference categories: **a** – 'non-PA'; **b** – 'female'; **c** – 'No comorbidities'; **d** – 'least deprived'*

Table 39: Prediction of total healthcare costs (with hospital admissions) using GLM with square root link function and gamma distribution

Variable	dy/dx	Standard Error	z	p-value
PA[a]	18,593.46	3,091.82	6.01	<0.001
Age at index date	833.01	110.69	7.53	<0.001
Gender[b]	-9,799.29	2,913.96	-3.36	0.001
Comorbidities[c]				
One	23,206.19	10,761.54	2.16	0.031
Two	16,812.52	7,371.72	2.28	0.023
Three	10,861.25	19,279.33	0.56	0.573
Years of GP registration	228.65	116.26	1.97	0.049
Deprivation[d]				
Less deprived	3,997.38	5,750.14	0.70	0.487
Average	-1,403.46	3,314.74	-0.42	0.672
Deprived	1,338.51	4,368.82	0.31	0.759
Most deprived	6,259.76	4,562.15	1.37	0.170
Number of observations = 876		Residual df = 864		
Scale parameter = 4.01				
Deviance = 1,203.46		(1/df) Deviance = 1.39		
Pearson = 3,460.08		(1/df) Pearson = 4.01		
AIC = 22.39		BIC = -4,650.46		
Log pseudolikelihood = -9,794.77				

**Abbreviations:* AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism*

***Reference categories:* **a** – 'non-PA'; **b** – 'female'; **c** – 'No comorbidities'; **d** – 'least deprived'*

5.3.6 Sensitivity Analysis

The robustness of the results presented in the sections above was assessed using two types of SA. Firstly, after removing potential 'ghost' patients with no GP consultations one year prior to index date (***Appendix 18***), 470 PA and 858 non-PA patients remained. When comparing to the main analysis, no significant differences in the patients' characteristics before and after index date were identified. The same thing applied when looking at the average number of annual healthcare resources used. Here, the usage trend for both groups was similar as before, with the difference that, this time, there was a small increase in the number of resources used at each timepoint. This increase led the average and median costs per category to rise slightly. The similarity in costs between the SA and the main analysis was also obvious from the GLMs that were developed, where the same parameters remained significant, with only a small change in their marginal effects found. Moreover, after removing patients with no GP consultations five years before index date, 472 PA and 929 non-PA patients remained. Here, the results were even closer, if not identical in some cases, to the results of the main analysis. In the second SA (***Appendix 19***), after removing patients with an unreasonably high healthcare use, the data from 441 PA and 855 non-PA patients were analysed. Again, no significant changes in the individuals' characteristics were found. Although the average annual healthcare usage dropped at each timepoint for all resource categories, its trend remained similar as in the main analysis for both cohorts. These changes had again only a small impact on the cost prediction GLMs, with the same regressors as in the main analysis remaining important.

5.4 Discussion

A costing study was conducted to examine the healthcare resources that are normally used by UK patients ten years before and ten years after the diagnosis of PA as well as their associated costs for the NHS. To do so, the study made use of CPRD data linked with data from the HES, IMD and ONS databases. These data were obtained for PA patients, and were compared to those of an age, gender and GP practice matched general population using a 2:1 (control:case) ratio. For the purposes of the study, a descriptive and regression analysis was undertaken, while the uncertainty in the results was tested in SA.

5.4.1 Key Findings

473 PA and 946 non-PA patients were included in the study, from whom 292 PA and 584 non-PA patients had linked data in the HES, IMD and ONS databases. Patients of both groups were perfectly matched in terms of their gender and GP practice, while they had a slight variation in their age at index date. Although a significant difference was found in their ethnicity (higher proportion of black PA patients), number and types of comorbidities (two years before index date), no significant variations in the other patient characteristics (e.g. socioeconomic status; age at death; number and cause of death) were identified. According to literature, there is no current evidence that PA is more prevalent within the black population, although the latter group has a slightly higher tendency to develop hypertension due to low renin levels (315-317). Additionally, the survival curve that was drawn based on the data of the current study (survival dropped faster for PA at ≈6 and ≈16 years) had a similar trend to that presented in *Reincke et al. (2012) (272)* for German PA patients.

Regarding healthcare utilisation, a similar pattern of average annual use for both cohorts was shown in all resource categories (i.e. GP consultations; diagnostic testing; drug prescriptions; secondary care referrals; hospital admissions), with PA patients using more resources at all timepoints. For PA patients, annual usage increased before diagnosis, then dropped after one year post-diagnosis, and finally increased again after five years post-diagnosis. For non-PA patients, the annual resource use increased relatively linearly throughout the twenty-year period. This pattern was expected given that as age increases, people tend to require more health care (become sicker), while the NHS has changed over time with GPs prescribing more drugs, providing more tests and/or referring more patients to secondary care. As anticipated, the most pronounced difference between the two groups was found in the peri-diagnosis period, and in the drug prescriptions and hospitalisations 5-10 years post-diagnosis. However, numbers of both GP consultations and drug prescriptions were higher for PA patients even at 5-10 years pre-diagnosis, suggesting that healthcare usage was elevated long before the diagnosis was made.

The regression models indicated that the presence of PA, the age at index date, the gender and the number of comorbidities were the most predictive variables of healthcare utilisation

(and expenditure). The fact that men were found to use slightly less resources compared to women agrees with the current literature on gender differences in healthcare utilisation (318). The healthcare use prediction models also showed that the time of the clinical event and the interaction between the time and PA exposure were two other significant factors. This is sensible given that the study focused on a period when PA patients were expected to require more health care compared to the general population. When analysing costs, drug prescriptions were found to dominate total healthcare costs for all CPRD patients, and together with hospitalisation costs they were also found to be the largest cost component for HES patients. Overall, the twenty-year primary care costs for the PA group were £22,352.70, while for the non-PA group, £12,188.82. When hospital admission costs were measured, costs increased to £45,263.66 and £26,175.01 for PA and non-PA patients, respectively. Results did not substantially change in an SA that explored several scenarios.

5.4.2 Comparison with Related Studies

As mentioned in **Chapter IV**, no study examining the healthcare utilisation and subsequently costs of PA patients were found in the literature. Additionally, no study using CPRD and/or its linked data for any other purposes was identified. However, a good number of studies using German, Japanese, Taiwanese, United States and international registry data were found (200, 251-283). These studies focused on other aspects associated with PA, such as estimating the prevalence/risk of diseases (e.g. diabetes mellitus) and cardiovascular or other comorbidities in patients, and assessing the different types of tests and treatments as well as their success rates, clinical outcomes and impact on the patients' quality of life. Moreover, as mentioned above, *Reincke et al. (2012) (272)* presented a similar Kaplan-Meier survival plot for PA patients to that provided in the current study. However, in the study, survival estimates were compared between PA and normotensive and hypertensive controls. These survival curves suggest that PA has only a small impact on survival in the first 10-20 years after diagnosis. Therefore, a longer follow-up should probably be examined in a future study to identify bigger differences in mortality between PA, non-PA and essential hypertensive patients. This study could potentially use registry data to ensure that its results will be more accurate and will avoid the limitations of GP data (see *Section 5.4.4* below).

5.4.3 Strengths and Limitations

This study used data obtained from the CPRD database, one of the few large primary care databases in the world. Specifically, CPRD GOLD covers >9% of the total UK population, and its data are representative of the general population in terms of the patients' demographic, socioeconomic and clinical information. CPRD GOLD data have a median patient follow-up of almost ten years for active patients and 5-6 years overall. These data are regularly validated and checked for their quality before being entered in the CPRD database. They also have the benefit that they can be linked (as in this study) for over half of the patients with secondary care (HES), socioeconomic (IMD) and mortality (ONS) records to provide a fuller picture of the patients' characteristics, hospital care and outcomes (246, 247). The ONS Death Registration data are considered the gold standard for mortality data in the UK, with studies showing that 98.2% of deaths in the ONS dataset are recorded in the CPRD GOLD (319), while 76.8% of death dates are perfectly matched (320). Moreover, in this study, a twenty-year period (ten years before and after index date) was examined. This was a sufficiently large period for estimating the use and cost of the healthcare resources that PA patients require, and compare them with those needed by the general population. Costs were assigned using official UK national sources, while the models that were developed to predict healthcare utilisation and expenditure included both continuous and categorical variables, and were checked for appropriateness using several recommended specification tests. Findings were found to be robust after removing potential 'ghost' patients and outliers in SA.

Nevertheless, the study had the limitations that come with CPRD data: e.g. missing data due to incomplete information entered by GPs in primary care records; how comprehensively and specifically GPs code each disease (here, PA); variations in coding between practices; free-text data received by GPs from secondary care facilities are not available to CPRD for data governance reasons; existence of duplicate records from patients who have moved between different GPs (246, 247). Additionally, there are data that are not captured by the CPRD, such as social support, over-the-counter drug use and adherence to treatments, while certain patient groups are also missing, such as prisoners, private patients and the homeless (246). Furthermore, data linkage is only available for English patients, so it is not possible to cross-validate deaths, calculate secondary care resource usage and identify the socioeconomic

status of non-English patients. The use of integrated and not full HES data also meant that complete information on hospitalisations (e.g. admission dates, duration of hospitalisation, diagnostic procedures), outpatient appointments, and accident and emergency events was not available for analysis. An additional limitation was the unavailability of HES data before April 1997. To overcome most of these issues and ensure that all important data were considered in the analysis, the dataset received was checked for duplicate records; the study included a sufficiently large period of time were most patient had data; results were reported separately for patients with and without linked data; and several SAs were conducted to examine the uncertainty in the final outcomes.

Although the study was designed in a way that could reduce any potential bias (e.g. exclusion, matching), some confounding bias, which is common in observational studies due to omitted variables, could have been present (321, 322). In this analysis, omitted variables refer to parameters that could have affected the predictions in healthcare use and costs but were not included because they were missing from the CPRD dataset (e.g. body mass index, smoking). Obviously, this adds to the general question of how comprehensively GPs record patient data, and subsequently how inclusive the CPRD databases are. However, given that there is currently no UK PA registry from which this study could have obtained its data, CPRD was the next best available source for getting the necessary data for estimating the burden of PA for the NHS. Additionally, although in this study CCI was used as a measure of comorbidities, other indices (e.g. Chronic Disease Score) have also been proposed (323). Nevertheless, CCI was selected as the most commonly used index for this type of analysis.

Another limitation was that this was a costing and not a cost-of-illness analysis. Costing analyses often use the healthcare provider perspective and measure the direct medical costs that are associated with the examined disease (e.g. specialist visits, medications). In contrast, cost-of-illness studies typically use a broader societal approach, and hence they also consider other cost components, such as direct non-medical (e.g. patient transportation), indirect (i.e. productivity loss), and intangible (i.e. patient's psychological pain/discomfort) costs (324-327). Although cost-of-illness analyses have long been used in health economics (324), there is currently no standard methodology on how they should be conducted (325, 326), with different studies using different perspectives (e.g. patient) and including different types of

costs (325, 327). Therefore, there is some uncertainty on how accurately they have estimated costs as well as some questions regarding their reliability, validity and usefulness in healthcare decision-making (324, 325, 327, 328). Despite this study focusing only on (a limited range of) healthcare costs, it used methods that are employed in higher-quality cost-of-illness studies, such as matching populations with and without the disease; using a detailed bottom-up costing approach; measuring incremental costs between the examined cohorts; and using advanced multivariable regression/econometric models to find the parameters that affect patient costs (324, 325, 327).

Moreover, although the Poisson and negative binomial regressions naturally accommodate zeros, other models such as the hurdle (two-part model for zero and positive count data) and zero-inflated models sometimes do so in a more explicit way (298, 299). However, given that the results from both Poisson and negative binomial models were similar, no differences in the significant parameters were expected when using the other two types of models, so these were not tested in this analysis. Modelling costs was also challenging since misspecification is possible despite the use of robust standard errors in the models. Additionally, the tests that were used to identify the appropriate link and family functions in GLMs have limitations and require a degree of contextual interpretation. For example, the Pregibon test is a diagnostic test that tells that the simpler model specification should be rejected without informing on the source of the problem (i.e. missing interactions or squared terms in the covariates, misspecification of the dependent variable, or all the above) (299, 309, 310). An alternative to the Pregibon test could have been the Ramsey's Regression Equation Specification Error Test (RESET). The difference between the two is that the Pregibon test examines whether a model including a squared cost variable as a covariate has more explanatory power than a model excluding this term, while the Ramsey's RESET jointly tests the squared, cubed and fourth-order terms. Nevertheless, if a model does not show significance in squared-order terms, by rule there is no point in checking higher-order terms (298, 299, 329). Therefore, the Ramsey's RESET was not used in this analysis.

Furthermore, in the health economics literature (298, 299), instead of using a 'single index' GLM to estimate a cost-outcome variable, two-part models are sometimes considered as a better alternative. As the name implies these models contain two parts, one that models the

probability that a person has non-zero healthcare use and/or costs (i.e. logit or probit), and a second linear, log-linear or GLM on the subset of the population that has any expenditures. In this analysis, two-part models as cost prediction models were explored. However, given that the first part of the models omitted all PA patients (since they always had non-zero costs), and dropped the 'number of comorbidities' variable due to collinearity, a two-part model was not considered appropriate. Other options to replace GLMs could be to use an OLS or a tobit model. The former was tested in this analysis showing no differences in the significant parameters, while the latter, according to *Deb et al. (2017) (299)*, is not commonly used for healthcare expenditure data, so it was not examined.

5.4.4 Implications and Future Research

As possibly the first analysis that has been conducted to measure the cost of PA diagnosis and treatment, this study is contributing to an under-researched clinical area. The next step would be to combine CPRD GOLD with CPRD Aurum data and link them with full HES data in order to get a more accurate estimate of the primary and secondary care use and costs of UK PA patients. Given the limitations of the CPRD dataset described in the previous section, it would also be useful to start creating a UK registry for PA which would include only verified patients with the disease (recorded by specialists) instead of patients suspected of having the disease (GP records). A registry would provide more detailed and accurate information (e.g. patient characteristics; time to diagnosis; test results; treatment and monitoring outcomes) that could give researchers more confidence in the data that they use in this type of analysis and the results that are yielded. Given that a UK PA registry does not currently exist (August 2020), data from other PA registries (e.g. the German Conn registry (272)) could be used.

Moreover, the findings of this analysis could be combined with the outcomes of a future observational study or clinical trial that would ask PA patients and non-PA individuals about their work productivity and how this has been affected by their current health condition (e.g. 'Work Productivity and Activity Impairment' questionnaire) as well as any healthcare costs that they needed to pay (e.g. using a bespoke disease-specific questionnaire). This would help to provide more accurate estimates of the total burden of PA for the society and inform health

policy in this clinical area. Lastly, given the results of this study, it is obvious that PA patients require a large number of primary care consultations and medications even 5-10 years (if not more) before the diagnosis of PA. This highlights the difficulty to diagnose this disease; and the need for an alternative more efficient test that could potentially reduce the time and cost of diagnosis and assist clinicians with making better decisions on the appropriate treatment, improving the patients' quality of life.

5.5 Conclusion

In conclusion, this was a study that aimed to describe the demographic, socioeconomic and clinical characteristics of PA patients; provide information on the healthcare services that they require ten years before and after their diagnosis; calculate how much these cost the NHS; and identify the parameters that mainly affect healthcare utilisation and expenditure. Data were compared to an age, gender and GP practice matched general population. Results indicated a higher use of health care for PA patients throughout the examined period with differences between the groups being more pronounced in the peri-diagnosis period. PA patients were also found to cost the NHS twice as much as non-PA patients during this twenty-year period. The factors that had a significant impact on healthcare utilisation and costs were the presence of PA, the time before and after the index date, the age at index date, the gender, and the number of comorbidities. In the future, it is important that further studies should be undertaken combining data from primary and secondary care databases with data from registries in order to get better estimates of the burden of PA.

CHAPTER VI

EARLY ECONOMIC EVALUATION

EARLY ECONOMIC EVALUATION OF U-RHYTHM FOR THE DIAGNOSIS AND MANAGEMENT OF PRIMARY ALDOSTERONISM

CHAPTER VI OVERVIEW

Chapters I-II provided a brief description of the ways that primary aldosteronism can be diagnosed and treated, and the methods that have been proposed for the early economic evaluation of diagnostic tests and medical devices. **Chapter III** demonstrated that there is only limited available evidence evaluating the cost-effectiveness of different tests for primary aldosteronism. **Chapter IV** showed that after screening for the disease, the saline infusion test is probably the most common confirmatory test, while it reported a general satisfaction of the patients after using *U-Rhythm* for this purpose. **Chapter V** indicated that the diagnosis and management of primary aldosteronism require a high number of resources, highlighting the need for a new test that would reduce the time to diagnosis and cost for the payer.

The aim of **Chapter VI** is to present an early economic evaluation that was conducted to show the potential place of *U-Rhythm* in the diagnosis of primary aldosteronism and compare its cost-effectiveness to that of the saline infusion test. In addition, the headroom approach was used to identify the highest price that the device can have and be cost-effective, while a value of information analysis was performed to examine how further evidence could affect the final decision of adopting/rejecting the innovative device for the diagnosis of the disease.

Early Economic Evaluation of U-Rhythm for the Diagnosis and Management of Primary Aldosteronism

6.1 Background

Chapters I-V have shown that the diagnosis and monitoring of the six endocrine disorders examined in the *ULTRADIAN* study is challenging, time-consuming and expensive since there is currently no clearly optimal test that can be used in each case. Various diagnostic strategies are followed depending on cost; the patient's characteristics and compliance; country; clinical site; and/or physician. More precisely, different patients might need different types and/or number of tests (repeated or not) depending on how clear their signs and symptoms are. Different countries may reimburse and/or have the technological capacity for performing different tests, while different clinical centres might have different resource availability, specialist skills and/or laboratory routines. Therefore, clinicians would have to choose the best diagnostic method(s) based on their experience, expertise and preferences.

Chapter III indicated the limited number and low quality of economic studies that have been conducted in this area of endocrinology, with only four studies (184-187) examining the diagnosis and management of primary aldosteronism (PA), two (50, 51) of Cushing's syndrome (CS), one (183) of Addison's disease (AD), and none of acromegaly, congenital adrenal hyperplasia and growth hormone deficiency. **Chapter IV** mentioned that the technical performance of the earlier versions of the *U-Rhythm* sampling device and analytical tools has been better for PA, CS and AD, and for this reason only a couple of patients were recruited for the other three diseases. Specifically, PA is an area in which the device has demonstrated diagnostic promise based on initial results. Additionally, the potential place for *U-Rhythm* in the PA diagnostic pathway is relatively straightforward (i.e. confirmatory testing), while it is theoretically possible that the device might also allow PA subtyping, avoiding the need for other (invasive) classification tests.

The abovementioned factors led the author to decide to perform a decision analysis for PA diagnosis to explore the potential cost-effectiveness of *U-Rhythm* in this context. Therefore,

the aim of **this Chapter** is to identify the potential role of *U-Rhythm* in the future clinical practice, use decision-analytic modelling (DAM) to compare its cost-effectiveness to that of the most commonly used diagnostic methods in the United Kingdom (UK) and Europe, and estimate the value of collecting additional information to support decision-making using value of information (VOI) analysis. To better understand the rationale and methods of this analysis, *Sections 6.1.1* and *6.1.2* provide some information on the aetiology, diagnosis and treatment of PA, and the use of DAMs/VOI in the economic evaluation (EE) of health technologies.

6.1.1 Primary Aldosteronism

6.1.1.1 Aetiology and Epidemiology

Chapter I provided a brief overview of the aetiology, epidemiology, diagnosis and treatment of PA. As mentioned, PA is a rare endocrine disorder that occurs when the adrenal glands autonomously produce an excess amount of aldosterone. Given that aldosterone's release is regulated by the renin-angiotensin system, its autonomous overproduction leads to renin suppression, and sodium retention and increased urinary potassium excretion by the kidneys. Prolonged or severe potassium excretion may cause hypokalaemia. Sodium retention leads to increased plasma volume, resulting in raised blood pressure. PA is often misdiagnosed as essential (primary) hypertension (EH) (46, 53, 56-58, 330-332) in spite of the fact that PA patients show a substantially higher risk profile compared to age, sex and blood pressure matched EH individuals (333, 334). PA can be caused by either aldosterone-producing adenomas (unilateral PA, UPA) or bilateral adrenal hyperplasia (bilateral PA, BPA) (46, 53, 56-58, 330, 335-338)[30]. PA is considered to be prevalent in 5-10% of the resistant hypertensive population (46, 54, 55, 330). However, based on some evidence (54, 343-346), this might be an underestimation, with a figure of 20% being more realistic.

[30] Unilateral adrenal hyperplasia, adrenal carcinomas and familial hyperaldosteronism are other PA causes not examined in this thesis due to their rarity (<5% of PA cases), and the different methods used for their diagnosis and treatment (46, 330, 335-342).

6.1.1.2 Screening

Evidence from around the world has shown that only 1% of the total PA population are correctly diagnosed and even less are appropriately treated (55). Timely diagnosis is essential to reduce the risk of missing cases (especially UPA patients which can be surgically cured), which leads to potentially poorer response to treatment once PA is finally diagnosed (46). According to the most recent Endocrine Society clinical guidelines (46), individuals should be suspected and screened for PA if they have: a) sustained blood pressure (BP) >150/100 mmHg on each one of the three measurements collected on different days, treatment-resistant hypertension (BP>140/90 mmHg) after receiving three ordinary antihypertensive drugs, or stable BP<140/90 mmHg and currently receiving ≥4 antihypertensives; b) hypertension and hypokalaemia (spontaneous or due to diuretics); c) hypertension and adrenal incidentaloma (347); d) hypertension and sleep apnoea; e) hypertension and family history of hypertension or cerebrovascular accident at an age <40 years old; or f) hypertensive first-degree relatives who have PA. In all cases, the clinician (e.g. general practitioner, cardiologist) should screen the patient by measuring their plasma aldosterone-renin ratio (ARR) (***Appendix 20***). ARR is the best available screening test for PA[31], although its results might sometimes be inconclusive or difficult to interpret, in which case it needs to be repeated (46, 338, 349). Antihypertensive drugs need to be withdrawn two to four weeks before ARR screening to reduce the number of false cases detected (46, 58, 330, 348, 350). Individuals with an ARR above the local cut-off values (ARR-positive) and at least modestly elevated plasma aldosterone concentration (PAC) should be considered for further confirmation/exclusion and subtyping of the disease (46, 54).

6.1.1.3 Confirmatory Diagnosis

After a positive ARR, the patient should be referred to secondary care to undergo one or more confirmatory (aldosterone-suppression) tests to definitively confirm/exclude the diagnosis. Failure of plasma aldosterone to be suppressed in response to these tests means that PA is present (46, 58, 330-332, 336, 348). There are four commonly used confirmatory tests (***Appendix 20***), none of which can be considered a reference standard: i) captopril challenge test (CCT); ii) fludrocortisone suppression test (FST); iii) oral sodium loading test (OSLT); and

[31] Other methods include measuring serum potassium, aldosterone or renin independently (46, 330, 338, 348).

iv) saline infusion test (SIT) (46, 330, 331). The choice of the appropriate test depends on its cost and availability, the patient's characteristics and compliance, and the doctor's expertise and preference. Clinicians might opt to proceed to subtype classification after one positive confirmatory test or perform multiple tests before confirming/excluding PA[32] (46).

6.1.1.4 Subtype Classification

The identification (i.e. UPA or BPA) and lateralisation (i.e. right/left or both adrenal glands) of the source of the excessive aldosterone release is essential in guiding the treatment of PA (i.e. surgery or medication). Subtype classification starts with the use of adrenal computerised tomography (CT), which excludes the presence of an adrenocortical carcinoma and images the anatomy of the adrenals (***Appendix 20***). CT has good diagnostic power to identify unilateral or bilateral micro- (≤1 cm) and/or macro- (>1 cm) adenomas. Nevertheless, non-functioning, incidental, unilateral macroadenomas are quite common in patients >35 years old and can be incorrectly interpreted as UPA. Therefore, if the patient has agreed to undergo surgery and a unilateral adenoma is observed on CT, (bilateral) adrenal venous sampling (AVS) should be performed as a more accurate test for diagnosing a unilateral source of aldosterone secretion and to avoid unnecessary surgery due to falsely positive CT (***Appendix 20***). Furthermore, CT is useful for guiding the cannulation of the adrenal veins during AVS and later the adrenalectomy. AVS is invasive, expensive, and requires expertise that is not widely available (46, 330, 331, 336, 338, 351). Other tests (46, 330, 352-356) can also be used for PA classification but they are not discussed further as they are not considered as accurate as CT and AVS (46, 330).

6.1.1.5 Treatment

The aim in treating PA is to normalise hypokalaemia, resolve or improve hypertension, and mitigate cardiovascular and renal morbidity (e.g. myocardial infarction, stroke, chronic kidney disease) caused by excess aldosterone (46, 330, 335, 336, 338). For UPA patients who are able/willing to undergo surgery, unilateral laparoscopic adrenalectomy has been proven to be more effective (i.e. improved blood pressure and quality of life; lower need for antihypertensive drugs; less medication-related side effects) (46, 58, 330, 336, 338) and cost-

[32] In patients with spontaneous hypokalaemia, plasma renin below detection levels and PAC>20 ng/dL (550 pmol/L), confirmatory testing is not needed and the patient can proceed directly to subtyping (46, 336).

effective (46, 330, 357, 358) than medication. If the patient is unable/unwilling to undergo surgery, lifelong treatment with mineralocorticoid receptor antagonists (MRAs) (i.e. spironolactone or eplerenone) and other antihypertensive drugs should be administered (46, 330, 331). The same medications are also provided after a partially successful or unsuccessful surgery (i.e. PA and hypertension partially or not cured). The decision on the appropriate drug regimen is taken after measuring plasma aldosterone-renin levels shortly after surgery (biochemical response) (46, 330, 335). For BPA patients, antihypertensives including MRAs are the recommended mode of treatment (46, 330, 331, 336).

6.1.2 Decision-Analytic Modelling in Economic Evaluation

6.1.2.1 General Information

Chapter II provided an overview of the different types of study designs that can be used in the evaluation of diagnostic tests and the EE of health technologies. As shown, randomised controlled trials (RCTs) are considered to provide high-quality evidence because, if well designed, they avoid selection bias (17, 133). However, there are challenges to conducting RCTs of diagnostic tests (12, 13, 16, 118, 121, 359) and limitations when they are used as a single study design for EEs (17, 136). Contrarily, DAMs have the advantage of being able to bring together several sources of evidence to address a health problem for a particular healthcare system, point in time and jurisdiction (17, 129, 137, 138). Additionally, they can be particularly useful in the early phase of technology development in exploring uncertainties in parameters (e.g. diagnostic accuracy, effectiveness, cost) and scenarios for how the technology/test might fit into future clinical practice (14, 29, 30, 33).

To understand the role of DAMs in EE, it should be clarified that any EE involves two important activities: measurement and decision analysis. Measurement focuses on collecting data on resource use, unit costs, diagnostic accuracy, effectiveness and health utilities from different sources of evidence; estimating and testing hypotheses relating to these parameters and relationships between these parameters; and exploring any uncertainties in their values (17). Decision analysis focuses on identifying the optimal course of action for a specific group of patients based on the expected costs and outcomes of the different alternatives, and the

objectives and constraints of the respective healthcare provision; and informing healthcare decisions based on best available evidence considering that these decisions will always be taken under conditions of uncertainty. DAMs constitute a valuable tool for clinical decision-making under uncertainty. They comprise a set of mathematical relationships between different clinical pathways/states and their parameters, showing how a disease progresses over time, and the impact that alternative interventions have on the payer's budget, and the patients' health-related quality of life (HRQoL) and survival. In general, the advantages of DAM as an analytic method for EE can be summarised in five key elements (17, 137):

1. **Structure:** Ability to represent many of the alternative clinical pathways that the individual can follow and/or possible disease prognoses depending on the patient's characteristics. Ability to estimate the impact of the compared technologies depending on the strategy that is chosen or on how the disease progresses over time.
2. **Evidence:** Ability to combine different sorts of evidence in the model and estimate its input parameters to reach clinical decisions. Ability to use evidence from different jurisdictions (e.g. countries) to make decisions.
3. **Evaluation:** Ability to use the available evidence to estimate the costs and health outcomes that are associated with each clinical strategy, and use appropriate decision rules to identify the most cost-effective alternative.
4. **Uncertainty and heterogeneity:** Ability to examine the uncertainty relating to the model's structure, input parameters and results, and characterise heterogeneity across different populations (e.g. age/gender subgroups).
5. **Future research:** Ability to identify the key areas that need further investigation to make the 'adopt' or 'reject' decision more robust.

In comparison to RCTs which typically study a subset of alternative courses of action, sometimes using placebos as comparators, DAMs can provide the framework for comparing the cost-effectiveness of all relevant strategies for particular patient groups combining information from various sources. To do so, data should be collected systematically, and then synthesised using appropriate statistical methods (i.e. meta-analysis) or selected using pre-defined rules (e.g. highest quality evidence) to be used as input parameters in the model. Additionally, DAMs can link intermediate (e.g. diagnostic accuracy) to final endpoints (e.g. mortality, HRQoL), while they can extrapolate beyond the follow-up period of a clinical study

by making reasonable assumptions on how parameters will change over time. The choice of the appropriate time horizon is important in EE since this is the period over which the costs and health benefits of the interventions that are compared are expected to differ (e.g. patient's lifetime for chronic conditions, such as EH and PA). RCTs have the disadvantage that they rarely follow up their participants for sufficiently long periods of time, mainly due to their expense. Moreover, results from DAMs can become applicable to different decision-making contexts by adjusting the structure of the model and its parameters to make them more relevant. Lastly, as mentioned in **Chapter II**, although DAMs are good at informing clinical decisions, they still require high-quality evidence (e.g. from RCTs) to ensure that these decisions are based on robust evidence. Therefore, they should be used in combination and not as an alternative to RCTs (17, 137).

6.1.2.2 Key Elements of Decision-Analytic Modelling

All DAMs include two key elements to decision analysis. The first one is the use of probabilities to reflect the likelihood, frequency or strength of belief that an event or a change in health will occur in a given population based on previous knowledge and experience (e.g. disease prevalence; test diagnostic accuracy). The second element is the expected values of the costs and effects of each alternative (i.e. costs and effects weighted by the probability that an event will take place). Expected costs are expressed in monetary units (e.g. British Pound Sterling), while expected outcomes can be expressed using any effectiveness measure (e.g. quality-adjusted life-years, QALYs) depending on the type of EE that is conducted (**Chapter II**). Information on each parameter can be obtained from different sources, including formal studies, national data (e.g. pricelists, statistics) or expert opinion. Although some types of evidence might be imperfect, decisions can only be taken based on the currently available data. However, a comprehensive description of the methods used and the assumptions made in the decision analysis should always be provided to make limitations explicit (17, 137).

6.1.2.3 Developing a Decision-Analytic Model

To develop a DAM, several steps should be followed (17, 137). First, is the specification of the decision problem (i.e. research question). Here, the targeted group of individuals and the interventions that are to be compared should be defined. Details on the location (e.g. the UK National Health Service, NHS) and setting (e.g. secondary care) in which the alternatives are

being delivered should also be provided. The second stage is to specify the boundaries of the model. All DAMs are simplifications of reality. Therefore, it is important to decide from the beginning which perspective will be chosen in the analysis (e.g. publicly funded healthcare system), the time horizon that will be examined (e.g. lifetime), and the appropriate outcomes measure(s). The decisions on what to include in the model are normally based on the expected availability of data, the complexity of the clinical pathways, the impact on cost-effectiveness, and the availability of resources (e.g. modellers' time) (17, 137).

The third step in developing a DAM is its conceptualisation and structure. In part, this is done based on the nature of the interventions evaluated and the natural history of the condition examined (i.e. biological/clinical process). The appropriate model structure can be decided by answering several key questions (17, 137):

a. What are the clinical events of interest (e.g. tests, treatment, disease progression)?
b. How many times do these events occur over the selected time horizon?
c. Do future events depend on previous events that appear in the model?
d. Are there any risks that might appear over time (e.g. adverse events)?
e. Are all important events included and are there any events that could be combined to simplify the model?
f. What are the model parameters that are needed?
g. How easy would it be to estimate these parameters based on available evidence?
h. Do these parameters change over time?
i. What is the durability of the examined intervention's effectiveness compared to the alternatives?
j. Is the model appropriate for the disease that is examined (e.g. long-term condition)?
k. Is the clinical prognosis/condition of a patient affected by the clinical status of other patients (e.g. infectious disease)?

The final stage in a model's development is its implementation. There are two types of DAMs that are commonly used in the EE of health technologies: decision trees and Markov (transition) models. Decision trees are simple to use and analyse, and easy to interpret. They include a number of clinical strategies that consist of a series of mutually exclusive events/branches (e.g. diagnostic tests, treatment). Moving from left to right, each event is

represented by a probability (baseline for initial events; conditional for subsequent events) and a cost, while the final event of each strategy (*'end point'*) is associated with a health benefit value (e.g. QALYs). To find the most cost-effective strategy, overall expected costs and benefits (i.e. sum of the costs and outcomes of all relevant pathways weighted by the respective pathway probabilities) are compared. Although useful, decision trees have several limitations; events occur over a discrete period, and they become too complicated and time-consuming to program and analyse when representing the progression of complex long-term conditions (17, 137).

To overcome these limitations, Markov models are often used in combination with decision trees. Markov models contain a series of mutually exclusive health states that a person can occupy at a given point in time. In contrast to decision trees, states may contain a range of combined events, while the probability of future events does not typically depend on previous ones (Markov or *'memoryless'* property). Nevertheless, there are work arounds (e.g. tunnel states[33]) that can provide a Markov model with some control over patient history. Over time, individuals can move between these states at discrete periods of time (*'cycles'*). The duration of each cycle depends on the disease and interventions that are evaluated. The direction and speed with which a person moves between the Markov states is described by the *'transition'* probabilities. Transition probabilities can either remain stable or change over time depending on the condition that is examined. Each health state is also associated with a mean cost and effect which may change at each cycle (17, 137).

To calculate expected values, the costs and outcomes associated with each Markov state are weighted by the time that a patient spends in that state using *'cohort simulation'*. This method assumes a hypothetical number of individuals (e.g. 1,000) that start in one or more health states of the model and produces a *'Markov trace'* that shows the proportion of the cohort that occupies each state at a given time. The model runs for many cycles until the majority of patients ends up in the last (absorbing) state (e.g. death). The overall expected costs and benefits are then calculated by adding the expected values of each cycle and dividing their sum by the arbitrary number of patients that entered the model. Costs and effectiveness can

[33] Additional states with transition probabilities, costs and effects dependent on how long an individual has been in a particular health condition (137).

be adjusted for differential timing using *'discounting'* methods. Discounting is done to reflect the lower value that individuals place on costs and outcomes in the future (17, 137).

Decision trees and Markov models are both examples of *'cohort models'* which estimate the expected values for the average patient without considering any variations between individuals. To estimate the costs and benefits of individual patients, an alternative approach is to use *'micro/patient-level simulation'* or *'individual sampling'* models. These DAMs are more flexible than cohort models since, by providing the baseline characteristics and tracking the history of events for each patient, they can predict the risk of future events that the individual can experience. Individual patient simulations can deal with patient heterogeneity more easily than cohort models. However, running these simulations to estimate expected values and any sensitivity analyses to quantify uncertainty can be time-consuming. In addition, these models require a high number of parameters to be estimated and therefore richer data (17, 137, 360-362). Given that finding this amount of evidence would be infeasible for rare diseases and early-stage technologies, these types of models are not explored further. Similarly, the use of other non-static models, called *'dynamic transmission models'* (363), are not considered relevant to the research question since these are mainly used to model infectious diseases.

6.1.2.4 Dealing with Uncertainty and Heterogeneity

In every EE, several sources of uncertainty are present, relating to the methodological assumptions made; the data required for the analysis; the need to extrapolate data or results over time; the desire to generalise results to another jurisdiction; and for models, the parameters and structure chosen. Various forms of sensitivity analysis (SA) are typically employed to assess the uncertainty within an EE and its impact on results. To run an SA, any uncertain parameters should first be identified. The plausible ranges for each parameter should then be determined and justified based on literature, expert opinion or specified confidence intervals (CIs) around the mean (for stochastic data) (17). The third step is to decide the appropriate form of SA. The simplest and most common type of SA is to perform a *'one-way'* analysis, which involves varying one parameter at a time to explore its implications for results. Nevertheless, this does not take into account that the overall uncertainty in the results depends on the combined uncertainty in several variables.

Therefore, a more realistic approach is to run a *'multi-way'* SA, varying the values of two or more parameters simultaneously. However, as the number of variables assessed increases, their potential combinations increase as well, making the computation of their impact on outcomes tedious. One way to handle this is to examine only the combinations that are important based on previous experimental studies. Alternatively, *'scenario'* analysis can be used, in which a series of scenarios representing a subset of potential multi-way analyses are investigated. Typically, these include a base-case, a best-case and a worst-case scenario or other expected scenarios. *'Threshold'* analysis (e.g. headroom approach) is another method, where the critical/threshold value(s) of parameter(s) that are important for the decision is/are identified to explore whether the results on cost-effectiveness would change (17).

Although the abovementioned SAs are useful, their results are partial, they fail to provide information on the joint parameter uncertainty, and they ignore variables that are correlated (e.g. test sensitivity and specificity). To overcome these limitations, another form of SA, called '*probabilistic sensitivity analysis (PSA)'*, is used. Here, an appropriate probability distribution is specified for each input parameter or multivariate distributions for correlated variables. The implications of the uncertainty in all parameters for the expected costs and effects are then evaluated simultaneously using repeated simulation. In Monte Carlo simulation, random values from each parameter distribution are drawn a large number of times (e.g. 10,000) to estimate the distribution of expected cost-effectiveness. The results of the PSA are presented using estimates of means and CIs. PSA can also be used to evaluate structural uncertainty by weighting competing structural assumptions to reflect their relative plausibility. However, this is rarely done in the EE literature. To deal with the heterogeneity in the characteristics of patients (e.g. gender) and/or their location, the model can be re-run separately for different subgroups of patients using relevant values for each parameter (17, 137).

6.1.2.5 Assessing the Value of Additional Research

Given the uncertainty, there is a probability that a decision on cost-effectiveness based on current evidence is incorrect. This decision uncertainty leads to the question of whether additional research is required before adopting/rejecting a new technology to examine whether it remains/becomes cost-effective when more information becomes available. To address this question, PSA results can be used to quantify the health benefit forgone and the

cost of wasted resources if a wrong decision is taken. The combination of this cost and the probability of making an error represents the expected opportunity losses or cost of uncertainty surrounding the adoption/rejection decision, which is equivalent to the value of the additional research that could reduce/remove this uncertainty. The latter is the *'expected value of perfect information (EVPI)'*. If the cost of doing more research is greater than the EVPI, the additional research is a poor investment and current evidence is sufficient to make the adoption/rejection decision (17, 137, 364, 365).

EVPI represents the cost of uncertainty associated with all parameters, but it is possible to measure the value of information of a subset of parameters, i.e. the *'expected value of partial perfect information (EVPPI)'*. EVPPI can focus on those parameters that are expected to have a major impact on the final decision (i.e. more uncertain; closely related to the difference in the cost-effectiveness between alternative strategies) and for which a further research study (e.g. a larger diagnostic accuracy study) could provide more precise estimates. Although the EVPI and the EVPPI can indicate whether additional evidence is needed, they cannot inform on the optimal design of the future study. For this, the *'expected value of sample information (EVSI)'* and its net payoff, i.e. the *'expected net benefit of sampling (ENBS)'*, need to be measured. These methods can compare the marginal cost of choosing one study design over another or adding more participants to a study to the consequent marginal reduction in the cost of uncertainty. They can also identify appropriate outcomes, follow-up and stopping rules (17, 137, 364, 365).

6.1.2.6 Critical Appraisal of Decision-Analytic Models

Despite being a valuable framework for EE, the quality of a model's outcomes is conditional on the evidence and assumptions that are used. In other words, it is essential to critically review the evidence that is used for finding the model's input parameters and the methods employed to yield its results (17). This could be done by following the 'good practice' guidelines and recommendations in the literature (137, 362, 363, 366-372), and by using quality appraisal checklists for EE and DAM (17, 181).

6.2 Methods

6.2.1 Decision Problem

The aim of this analysis is to examine the potential clinical and economic value of *U-Rhythm* in the diagnosis and management of PA. To do so, the potential cost-effectiveness of the device if it was used to replace SIT is explored. Additionally, the cost-effectiveness of another theoretically possible strategy where *U-Rhythm* provides sufficiently detailed diagnostic information to replace both confirmatory (i.e. PA or not) and subtype (i.e. UPA or BPA) tests is explored. As a secondary objective, the headroom approach is employed to estimate the maximum price or the minimum diagnostic accuracy that *U-Rhythm* can have and be/become cost-effective, while a VOI analysis is performed to identify key areas of uncertainty in the decision to use the device in PA diagnosis and treatment. The suitability of these methods in early EE is also investigated and compared.

6.2.2 Decision Analysis

6.2.2.1 Patient Profile

The DAM developed follows a 55-year-old hypothetical hypertensive patient suspected of having UPA or BPA after a positive ARR test. The patient is assumed to have already been receiving three or more antihypertensive drugs. Additionally, the patient is willing/able to proceed to surgery, if appropriate. These patient characteristics were selected to assume a patient similar to that presented in **Chapter V** and exclude cases that follow a different diagnostic and management pathway (see Endocrine Society guidelines (46, 330)). The model starts with the patient being referred to secondary care for confirmatory testing and follows them over their lifetime considering cardiovascular morbidity and all-cause mortality.

6.2.2.2 Comparators

The DAM compares the long-term cost-effectiveness of three diagnostic strategies for the confirmatory diagnosis, subtype differentiation and treatment of PA: a) SIT as a confirmatory test; b) *U-Rhythm-current* as a confirmatory test; or c) *U-Rhythm-theoretical* as a confirmatory

and subtype test. The first two strategies follow the current PA diagnostic pathway proposed by the Endocrine Society (46, 330), in which adrenal CT and AVS are used sequentially for final PA subtype classification. The third strategy assumes that *U-Rhythm* can provide sufficient information regarding the type of PA (without the need to use AVS as a subtype test); CT is still required to image the adrenals prior to surgery. *U-Rhythm-theoretical* is proposed as an alternative pathway to current practice which would be used if *U-Rhythm* is sufficiently accurate to diagnose the subtype of PA. For this reason, it is compared to SIT rather than *U-Rhythm-current*. In all three strategies, treatment is based on diagnosis (i.e. adrenalectomy for UPA; antihypertensive drugs including MRAs for BPA; antihypertensives without MRAs for EH). Following surgery in UPA, pharmacotherapy depends on the success of surgery in treating hypertension.

6.2.2.3 Model Conceptualisation

The model was conceptualised based on the Endocrine Society clinical practice guideline (46) combined with suggestions made by endocrinologists from the four *ULTRADIAN* clinical sites. The structure of the model was also influenced by the *Lubitz et al. (2015) (185)* and *Sato et al. (2015) (186)* DAMs identified in the systematic review presented in **Chapter III** since both addressed a similar decision question and provided detailed information on input data. The model has two parts: i) a decision tree that represents the diagnostic and initial treatment pathway; and ii) a three-state Markov model that tracks cardiovascular events and survival conditional upon the aetiology and treatment selected. PA diagnosis and treatment selection constitutes a sequential procedure that runs in a 'discrete' period of time, and therefore it can be adequately represented by a decision tree. On the contrary, the measurement of the longer-term costs and outcomes of treatment is more complicated since the patient's health state changes over their remaining lifetime. For this reason, the use of a Markov model was considered more appropriate.

6.2.2.4 Model Structure

6.2.2.4.1 Decision Tree

In the model, PA diagnosis starts in secondary care, when a hypertensive patient, after receiving clinical examination and having a positive ARR test, is suspected of having PA and is

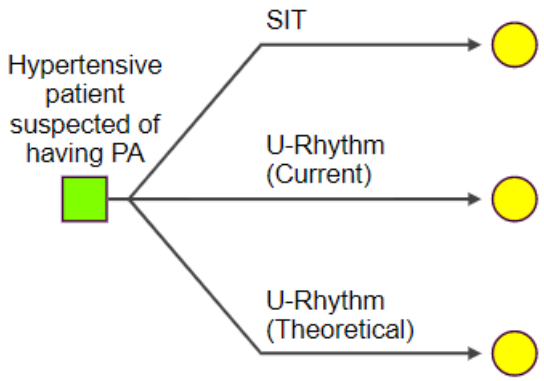
referred for confirmatory testing (**Figure 21A**). The square on the left (*'decision node'*) represents the decision of the specialist doctor (e.g. cardiologist, endocrinologist) to confirm/exclude PA presence using: a) SIT; b) *U-Rhythm-current* (confirmatory); or c) *U-Rhythm-theoretical* (confirmatory and subtype). For clarity, the decision tree tracks patients separately according to the true aetiology: UPA, BPA or EH. The circles (*'chance nodes'*) represent the probability of an event occurring. In reality, the true aetiology cannot be observed and must be inferred based on test results. Starting with SIT (**Figure 21B**), results can be either conclusive or inconclusive. An inconclusive finding means that SIT may be repeated unless the physician decides to treat the patient empirically with antihypertensive medications. For PA patients, a conclusive SIT result can be true positive (TP), i.e. PA correctly confirmed, or false negative (FN), i.e. PA wrongly excluded. For EH patients, this can be true negative (TN), i.e. PA rightly excluded, or false positive (FP), i.e. PA incorrectly confirmed. Positive results lead to further PA subtyping to determine treatment, whereas negative results lead to the prescription of antihypertensive drugs.

Adrenal CT is used as the initial subtype test to distinguish UPA from BPA. CT also provides imaging of the adrenal anatomy to guide adrenal vein cannulation for AVS and, if required, the adrenalectomy. AVS is performed whether the CT result suggests UPA (CT-positive) or not (CT-negative) to confirm the source of the excessive aldosterone secretion and reduce the risk of unnecessary surgery. The cannulation of the adrenal veins during AVS can be difficult, and its success and complication rate (i.e. adrenal haemorrhage) mainly depend on the expertise of the angiographer (46, 330). An unsuccessful or negative AVS (i.e. PA not lateralised or BPA rather than UPA indicated, respectively) leads to the administration of antihypertensive medications. If CT and AVS indicate UPA, adrenalectomy is conducted to remove the adrenal adenoma. There is a small probability of peri-operative mortality (373). Death, represented by a triangle (*'end node*') in the model, is termed an *'absorbing health state'* since once entered, it cannot be left. In UPA, surgery can be completely successful (i.e. hypertension resolved; no antihypertensive medications are required); partially successful (i.e. hypertension improved; same or less antihypertensives compared to before diagnosis are required); or unsuccessful (i.e. hypertension not improved; same or more antihypertensive drugs compared to before diagnosis are needed). Surgery is assumed to be unsuccessful in patients who have a true diagnosis of BPA or EH (i.e. inappropriate adrenalectomy).

Except for patients with UPA successfully treated with surgery, all patients require lifelong medications to treat hypertension. The drug regimen is complex and determined empirically on a case-by-case basis. For BPA patients, MRAs are the first line of pharmacotherapy, but other antihypertensive medications are usually required as well. Similarly, for UPA patients who have unsuccessful surgery, MRAs are re-introduced alongside other antihypertensives post-surgery. Antihypertensive medications may include angiotensin II receptor blockers, angiotensin-converting enzyme inhibitors, calcium channel blockers, diuretics and beta-blockers. For simplicity, the model assumes that patients with a presumed diagnosis of PA receive the abovementioned antihypertensive drugs including MRAs, while in patients with a presumed diagnosis of EH, antihypertensives without MRAs are administered. In patients with partially successfully treated UPA, antihypertensives including 50% MRAs are given. These assumptions are largely arbitrary and are tested in SA.

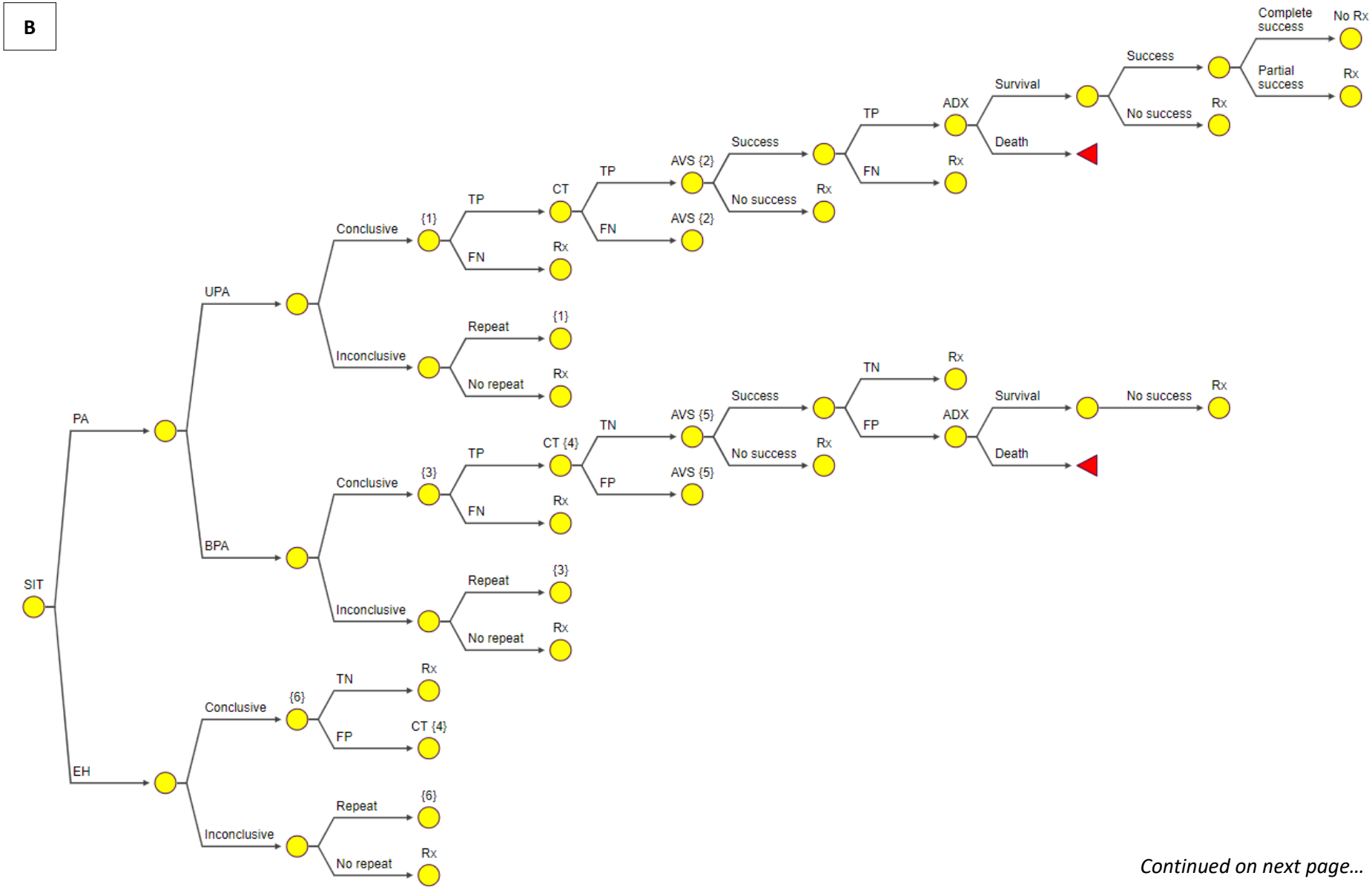
Although the *U-Rhythm* device is still at an early stage of development, the wealth of data collected on hormones over a 24-hour period means that it has the potential to replace more than one test in the PA diagnostic pathway. Firstly, it could replace the confirmatory test (SIT) (**Figure 21C**). Secondly, it could confirm PA and identify UPA, replacing both SIT and the need for AVS before surgery (**Figure 21D**). Here, a positive *U-Rhythm-theoretical* result indicates that UPA is present, so surgery should be performed after anatomical imaging confirmation with CT without the need for AVS. If CT is negative, AVS is used to determine the appropriate treatment as in current practice. A negative *U-Rhythm-theoretical* result indicates BPA or EH, so appropriate pharmacotherapy should be prescribed.

A



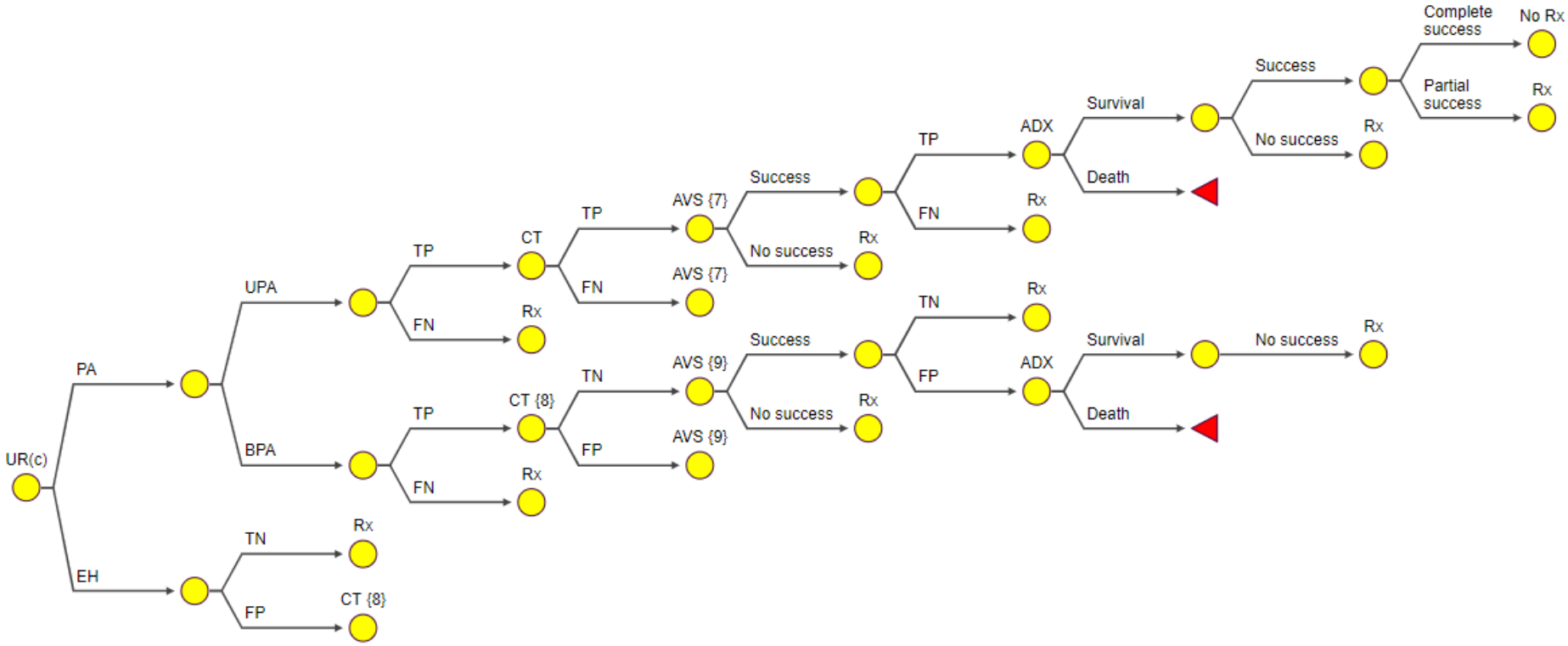
Continued on next page…

B



Continued on next page…

C



Continued on next page…

D

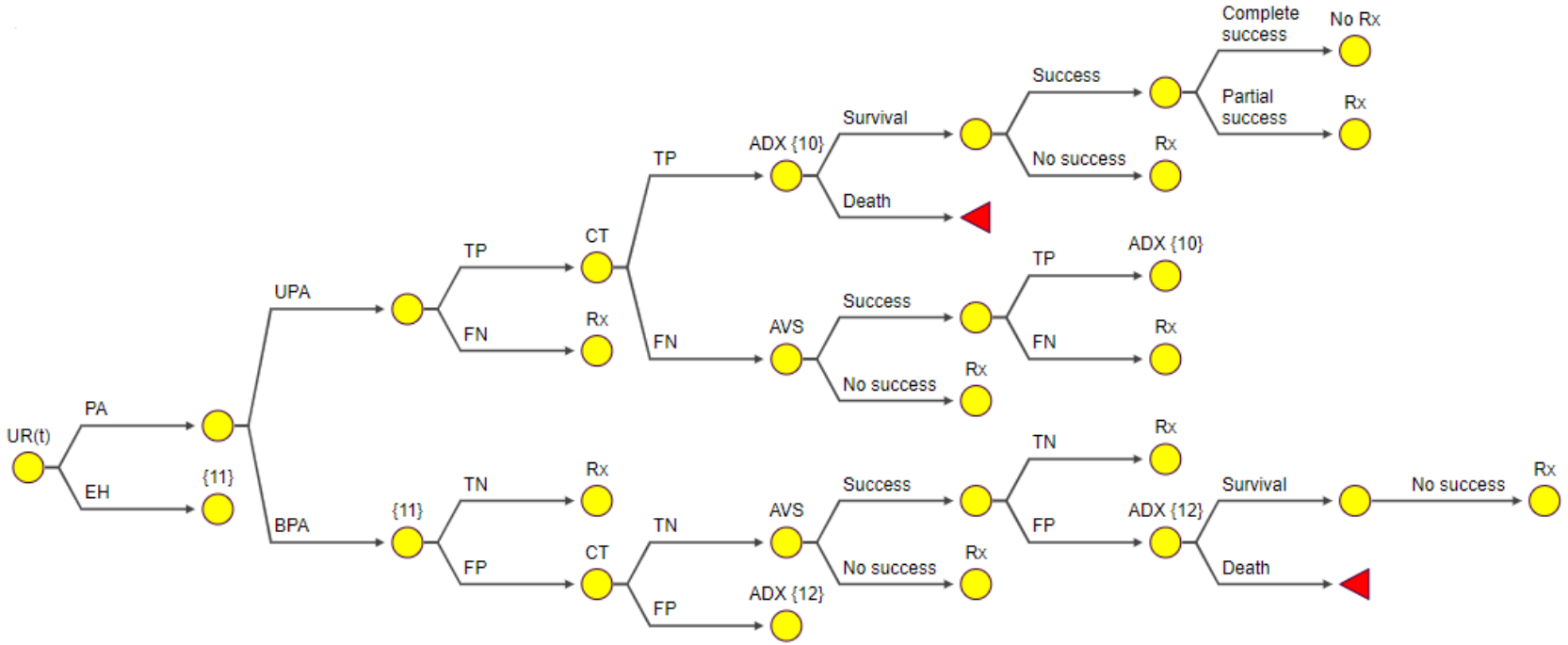


Figure 21: Diagrammatic representation of the decision tree

**Abbreviations:* ADX: laparoscopic adrenalectomy; AVS: adrenal venous sampling; BPA: bilateral primary aldosteronism; CT: computerised tomography; EH: essential hypertension; FN: false negative; FP: false positive; PA: primary aldosteronism; Rx: pharmacotherapy; SIT: saline infusion test; TN: true negative; TP: true positive; UPA: unilateral primary aldosteronism; UR(c): U-Rhythm (current practice); UR(t): U-Rhythm (theoretical practice)*

***Notes:* The numbers in the braces represent the subtree used; The decision tree leads to the Markov model when 'No Rx' or 'Rx' is administered.*

6.2.2.4.2 Markov Model

To assess long-term healthcare costs, HRQoL, cardiovascular events (CVEs) and mortality, the decision tree is combined with a Markov model (**Figure 22**). Given that the epidemiological literature on PA is sparse and often of low quality, it would be difficult to identify all the necessary parameters for even a simple Markov model. Additionally, the Markov models used in previous studies, e.g. *Lubitz et al. (2015) (185)* and *Sato et al. (2015) (186)*, either required access to complex cardiovascular risk models or were not reported transparently enough to replicate. For these reasons, a simple three-state Markov model was developed ('Alive without CVEs'; 'Alive with CVEs'; and 'Dead'). The model quantifies the cost and HRQoL implications of living with and without major CVEs (i.e. heart failure, myocardial infarction, stroke). Apart from those who die peri-operatively, patients enter the Markov model in the 'Alive; Without CVEs' health state. After each cycle, the patient can remain in the same state continuing receiving the same medication regime (if any) or move to either the 'Alive; With CVEs' or 'Dead' state. Patients in the 'Alive; With CVEs' state have additional treatment costs and after each cycle, they can either remain in the same state or die. Death is a terminal state. The arrows in **Figure 22** show the transitions between states that are permitted in each cycle. Cycles last for a year and the model follows patients over lifetime.

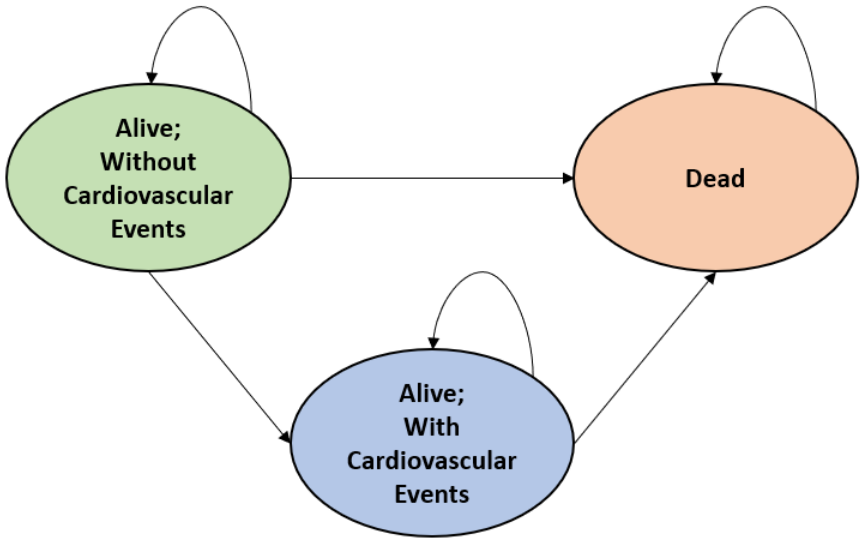


Figure 22: Diagrammatic representation of the Markov model

6.2.2.5 Model Assumptions

To develop the DAM described above, the following key assumptions were made:

1. The model uses a similar starting age to that used in the *Sato et al. (2015) (186)* model and shown in the *Reincke et al. (2012) (272)* observational study. In practice, the diagnosis for younger patients might be faster due to clearer indications (e.g. adrenal nodules), while the use of surgery in older patients depends on the patient's characteristics.
2. No diagnosis or treatment for causes other than those described is considered.
3. Although the Endocrine Society recommends CCT, FST, OSLT or SIT for confirming PA, none is clearly superior (46, 330). SIT is chosen as the only comparator given that it is easy and inexpensive to perform and is commonly used in the UK and Europe (374).
4. In practice, when SIT is inconclusive, a repetition of the test or a combination of several confirmatory tests may be used before proceeding to subtyping and treatment. The choice of test(s) is based on the test availability and clinician's preference.
5. *U-Rhythm* is assumed to provide conclusive results. This is because the device collects a greater amount of hormone data over a 24-hour period which should provide better differentiation. *U-Rhythm* is at an early stage of development. Therefore, larger studies than *ULTRADIAN* are required to examine this assumption.
6. CT-positive results for patients with UPA can still lead to incorrect diagnosis if nodules are localised on the wrong adrenal. This as well as its inconclusiveness percentage and the chance of identifying adrenocortical carcinomas are not presented in the model for simplification purposes.
7. An alternative to CT could be the use of magnetic resonance imaging. However, this is a more expensive test and has less spatial resolution, making it less accurate (46, 330).
8. Various AVS techniques are available (e.g. sequential or simultaneous; unstimulated or cosyntropin-stimulated) (46, 330), but these are not tested for simplification purposes.
9. In practice, the clinician might opt to repeat AVS if unsuccessful. This alternative is not presented in the model.
10. Posture stimulation test, iodocholesterol scintigraphy, 18-hydroxycoricosterone levels, and c-metomidate positron emission tomography-computerised tomography are known to be alternatives to CT-AVS in PA subtyping (46, 330, 352-356). However, given that these tests are not widely used due to not being as accurate as CT-AVS (46, 330), they are not examined.

11. Adrenalectomy can be undertaken using an open or laparoscopic transperitoneal or retroperitoneal approach (331). For simplicity, these types are not examined separately.
12. Before surgery, hypokalaemia and hypertension should be controlled using potassium supplements and/or MRAs (46, 58, 330, 335, 336). Since regimens would vary among patients, the costs of these drugs are not measured separately and are considered to be included in the total cost of surgery.
13. Clinicians may request post-operative testing to ensure that PA is cured (46, 330). Given the limited availability of data, post-operative testing is not examined.
14. Apart from MRAs, epithelial sodium channel antagonists (i.e. amiloride, triamterene) can also be used for treating PA. These drugs are less effective than MRAs and, if used, are usually given in combination with the latter (46, 330). For this reason, they are not expected to have an impact on final outcomes and are not shown in the model.
15. To examine the success of PA treatment, the cure of autonomous aldosterone secretion, hypokalaemia (biochemical outcome) and hypertension (clinical outcome) is tested. Given that in most cases the first two outcomes are ameliorated after surgery/medication (46, 330), only the treatment's clinical success is investigated in the model.
16. The simplified Markov model that is used is very crude and does not separately model all cardiovascular and renal risk factors and events (e.g. atrial fibrillation, diabetes, chronic kidney disease) (333, 336, 375, 376) due to lack of data.

6.2.2.6 Model Parameters

At the outset of this PhD project, the structure of this DAM was unclear since little was known about the role of *U-Rhythm* and other existing tests in the diagnosis and management of PA. For this reason, a systematic literature review (**Chapter III**) was conducted to identify available EEs of different tests for PA that could potentially inform the structure and input parameters of this model (i.e. through forwards and backwards citation tracking). Moreover, the data collected from the *ULTRADIAN* study (**Chapter IV**) were anticipated to provide some estimates of the diagnostic accuracy and cost of *U-Rhythm*, and the patients' utility scores before and after treatment (i.e. surgery and medications). Additionally, by using the Clinical Practice Research Datalink (CPRD) and Hospital Episode Statistics (**Chapter V**), it was expected that some information on the types and costs of the healthcare resources that are commonly employed to diagnose and treat PA would become available and be used to populate the

model. Despite the initial intention of the author to use as much of these data as possible in the analysis, it became clear that some of the evidence from **earlier Chapters** could not inform the parameters required for this model (e.g. medication costs), while in other cases (e.g. utilities), other more relevant data were available in the literature. Therefore, it was decided that when higher-quality data were available in the literature (e.g. from large observational studies or RCTs) or could be obtained from UK national sources (e.g. test costs), these would be preferred to inform the parameters of the model since they would increase the robustness of the cost-effectiveness results.

6.2.2.6.1 Probabilities

The probabilities can be divided into four categories: a) disease prevalence; b) test attributes; c) surgery outcomes; and d) transitions between health states. Disease probabilities include data on the prevalence of PA within ARR-positive hypertensive patients, and UPA in the PA population. Test features contain information on the inconclusiveness, success and repetition (if needed) percentage and diagnostic accuracy of tests. Diagnostic accuracy is expressed in terms of sensitivity (TP/TP+FN) and specificity (TN/TN+FP). Moreover, surgery outcomes can be grouped into peri-operative mortality after adrenalectomy, and complete, partial or no success in treating hypertension. Lastly, transition probabilities represent the patient's moves between health states based on their likelihood to experience CVEs or die in each cycle.

The Endocrine Society PA guidelines (46, 330) and the PA papers included in the systematic review (**Chapter III**) (184-187) were hand-searched (i.e. forwards and backwards citation tracking) to identify the probabilities for the model. Additionally, several targeted/rapid searches were conducted on PubMed® using appropriate clinical and diagnostic terms (e.g. 'saline infusion test', 'adrenal venous sampling', 'adrenalectomy'). Relevant studies were selected after first screening their titles and abstracts and then screening their full texts. Given the heterogeneity between studies (e.g. in patient characteristics; test/treatment procedures; test laboratory assays and cut-off points; healthcare setting and country), no meta-analysis of the data found could be conducted. Instead, the evidence for each parameter was selected based on: i) most relevant to the decision problem; ii) strongest study design (e.g. meta-analyses, RCTs, multi-centre observational studies); iii) largest sample; and iv) studies after 1990. Only studies published in peer-reviewed literature were searched since

these sources were expected to provide higher-quality results. Furthermore, non-English language studies were excluded due to time constraints. Where several probability estimates were available, those not used in the base-case analysis were used in the SA (see *Section 6.2.3* below). Where no evidence was found for a parameter, expert opinion from the *ULTRADIAN* clinicians was used. Last, *U-Rhythm*'s diagnostic accuracy was estimated using the method presented in ***Appendix 11***.

Table 40 shows the values and sources of the probabilities used in the DAM. For the decision tree, the numbers were: a) measured from the proportion of events (e.g. TP/TN cases) that were observed in the highest-quality study; b) estimated by taking the average of the values obtained from several studies, if these were of similar quality; or c) based on *ULTRADIAN* expert opinion. Given that the model examines collectively exhaustive and mutually exclusive events (i.e. at least one of them must occur and they cannot coincide), the total probability of each two-way split of a branch of the decision tree must always add to 1. Therefore, the probability of an event's complement was given by:

$$P(A^c) = 1 - P(A) \qquad \textbf{(Equation 2)}$$

where *P(A)* the probability of the event and $P(A^c)$ the probability of the complementary event. The same rule applied to the 'Alive; With CVEs' state of the Markov model for all transitions made. However, for the 'Alive; Without CVEs' state, where three complementary events can take place, two transition probabilities needed to be known.

Given the structure of the Markov model, it is expected that the annual probabilities of CVEs and death increase over time (as the cohort ages), while in the case of 'death', they also depend on the state from which the patient comes. To deal with this, cause and time-dependency were incorporated into the model's transition probabilities using survival data from three sources. *Hundemer et al. (2018) (266)* provided information on the number of cardiovascular and all-cause mortality events for EH patients for eight and eleven years, respectively. Therefore, the annual event rates were calculated by (137):

$$r = -[\ln(1 - p)]/t \qquad \textbf{(Equation 3)}$$

where *r* is the hazard rate, *p* the probability/proportion of events and *t* the time period of interest. The study also provided the multivariable adjusted CVE and mortality hazard ratios (compared to EH patients) for PA patients who were surgically or medically treated[34]. The transition probabilities were then calculated by (137):

$$p = 1 - \exp(-rt) \qquad \textbf{(Equation 4)}$$

where the EH rates were multiplied by the respective hazard ratios for the two PA cohorts.

The patient-group-specific transition probabilities from without to with CVEs were assumed to be constant over time since PA and EH are chronic conditions that put patients at risk of CVEs throughout their lifetime. In contrast, the probability of death increases in later years of life. Therefore, the transition probabilities to 'death' were assumed to be constant for only the first eleven cycles where *Hundemer et al. (2018) (266)* provided data. For each one of these years, the excess mortality from cardiovascular causes was estimated by subtracting the UK average age-specific death rate for both males and females – taken from the UK national life tables (377) – from the proportion of patients estimated to die that year based on *Hundemer et al. (2018) (266)*. The average of these excess mortality percentages was then added to the UK annual age-specific death rate to estimate the transition probabilities of death from cycle twelve until the end of the patient's lifetime. Since *Hundemer et al. (2018) (266)* did not distinguish between the causes of death, the annual mortality rate from CVEs was taken from *Kaczmarski et al. (2019) (378)*, who estimated this for all treatment resistant-hypertensive (TRH) patients. Its value was assumed to be constant over time, and this or its complement rate were multiplied by the patient-group-specific transition probability for death to give the number of individuals who survived/stayed in the 'Alive; With CVEs' and 'Alive; Without CVEs' states, respectively.

[34] Given that the mortality hazard ratio for PA surgically treated patients was not provided in the study, this was assumed to be equal to the CVE hazard ratio for this PA group since the respective two ratios for PA medically treated patients had small differences.

6.2.2.6.2 Costs

For *U-Rhythm*, the potential cost was measured from the average costs of the equipment, consumables, staff time and laboratory analysis, as estimated by the *ULTRADIAN* partners, using the process described in ***Appendix 10***. NHS reference costs 2018/2019 (379) were used to estimate the cost of SIT (cost of uncomplicated day case admission for endocrine disorders) and the weighted average costs of the adrenal CT and adrenalectomy based on the relevant Healthcare Resource Groups (HRGs)[35], while the AVS cost was measured by taking the average of the costs reported in six studies (184, 185, 214, 358, 380, 381). The annual cost of antihypertensive drugs came from *Belsey et al. (2012) (382)*, who estimated these costs for UK hypertensive patients (using patient records from a national primary care database) for 2-5 years after initiation of therapy. To calculate the annual cost of MRAs, the average dose per day for each type (i.e. 100mg spironolactone, 100mg eplerenone) was estimated based on the Endocrine Society guidelines (46) and related literature (336, 348). Doses were selected to represent typical doses while minimising adverse events. The British National Formulary (BNF) (304) was then used to calculate the number of tablets and subsequently packages needed per year. The annual cost for each drug was measured and the mean of the two is used in the analysis. Lastly, the initial and subsequent costs associated with CVEs were obtained from *Danese et al. (2016) (383)* who used CPRD data to estimate these costs for UK patients with first and second CVE-related hospitalisations (**Table 40**).

Costs reported in currencies other than the British Pound Sterling (£) were converted using the purchasing power parities (PPPs) for the relevant year (384). PPPs, which equalise the purchasing power of currencies – in terms of essential goods and services – by controlling for price level differences across countries, are considered to be more accurate currency conversion rates than market exchange rates, which fluctuate throughout the year. The UK's gross domestic product deflator (385) was used to adjust all costs to 2019 values.

[35] Groups of diagnoses and interventions that use similar levels of NHS resources (379).

6.2.2.6.3 Health Benefits

Clinical effectiveness was measured using QALYs since this is the outcome measure that is widely used by the NHS when comparing different healthcare programmes (117). To identify data on utility scores (**Table 40**), several targeted searches were conducted. *Velema et al. (2018) (197)* was the best available source for providing comparative HRQoL data in the form of EQ-5D scores (12-month follow-up) for PA surgically and medically treated patients (i.e. 'Alive; Without CVEs' state). EH patients were assumed to have similar scores to those of PA medically treated patients given that both groups have hypertension and receive similar treatment. To estimate the scores for the 'Alive; With CVEs' state, the mean utility decrement (0.05) reported in *Briggs et al. (2017) (386)* was subtracted from the 'Alive; Without CVEs' utility score. Although this paper compared UK diabetes mellitus patients with and without CVEs, the negative impact of CVEs on HRQoL was assumed to be similar in patients with PA or EH. This assumption was supported by a study that examined the HRQoL of elderly Korean patients with hypertension, diabetes and/or CVEs, which reported similar utility scores (387).

6.2.2.6.4 Assigning Parameters

To represent all possible transitions and their associated probabilities, costs and outcomes, six Markov models are used in the analysis (**Table 41**).

Table 40: Model input parameters used in the base-case and sensitivity analyses

Model Parameter	Value/Mean (SE)	Distribution Parameters[a]	Distribution	Source(s)
Decision Tree Probabilities				
PA prevalence in ARR-positive hypertensive population	0.112 (0.009)	126/999	Beta	(388)
UPA prevalence in PA population	0.429 (0.044)	54/72	Beta	(388)
SIT inconclusiveness percentage	0.417 (0.030)	115/161	Beta	(389)
SIT repetition percentage if initially inconclusive	0.500 (-)	0.250/0.750	Uniform[b]	Expert opinion
SIT sensitivity	0.818 (-)	0.739/0.896	Uniform[c]	(390-395)
SIT specificity	0.876 (-)	0.751/1.000	Uniform[c]	(390-395)
U-Rhythm sensitivity	0.900 (0.065)[d]	18/2	Beta	*ULTRADIAN*
U-Rhythm specificity	0.862 (0.031)[d]	106/17	Beta	*ULTRADIAN*
Adrenal CT sensitivity	0.871 (0.059)	27/4	Beta	(396)
Adrenal CT specificity	0.718 (0.071)	28/11	Beta	(396)
AVS success percentage	0.958 (0.020)	92/4	Beta	(184)
AVS sensitivity	0.883 (-)	0.795/0.970	Uniform[c]	(397-400)
AVS specificity	0.875 (-)	0.750/1.000	Uniform[c]	(397-400)
Adrenalectomy peri-operative morality percentage	0.004 (0.002)	4/1079	Beta	(401, 402)
Adrenalectomy hypertension improved percentage	0.855 (0.013)	649/110	Beta	(403, 404)
Adrenalectomy hypertension cured percentage	0.441 (0.019)	286/363	Beta	(403, 404)

Model Parameter	Value/Mean (SE)	Distribution Parameters[a]	Distribution	Source(s)
Markov Model Transition Probabilities, Cardiovascular and Survival Data				
Annual CVE hazard rate – EH	0.030 (0.002)	8,600/33,253	Beta	(266)
Adjusted CVE hazard ratio – EH vs PA-ADX	0.580 (0.260)	0.35/0.97	Log-normal	(266)
Adjusted CVE hazard ratio – EH vs PA-Rx	1.910 (0.082)	1.63/2.25	Log-normal	(266)
Annual all-cause mortality hazard rate – EH	0.016 (0.002)	6,443/35,410	Beta	(266)
Adjusted mortality hazard ratio – EH vs PA-ADX	0.580 (0.260)[e]	0.35/0.97	Log-normal	(266)
Adjusted mortality hazard ratio – EH vs PA-Rx	1.340 (0.122)	1.06/1.71	Log-normal	(266)
Mortality rate from CVEs – TRH	0.373 (0.020)	211/354	Beta	(378)
'Alive; Without CVEs' to 'Alive; With CVEs' – EH	0.029 (-)[f]	-	-	-
'Alive; Without CVEs' to 'Alive; With CVEs' – PA-ADX	0.017 (-)[f]	-	-	-
'Alive; Without CVEs' to 'Alive; With CVEs' – PA-Rx	0.055 (-)[f]	-	-	-
'Alive; Without/With CVEs' to 'Dead' – EH	0.016 (-)[f]	-	-	-
'Alive; Without/With CVEs' to 'Dead' – PA-ADX	0.009 (-)[f]	-	-	-
'Alive; Without/With CVEs' to 'Dead' – PA-Rx	0.021 (-)[f]	-	-	-
Costs (£)				
SIT	407.11 (-)[g]	203.56/610.67	Uniform[b]	(379)
U-Rhythm	1,101.02 (-)	1,008.07/1,193.08	Uniform[c]	*ULTRADIAN*

Model Parameter	Value/Mean (SE)	Distribution Parameters[a]	Distribution	Source(s)
Adrenal CT	88.89 (-)	83.23/107.16	Uniform[h]	(379)
AVS	2,036.60 (-)[i]	558.80/4,682.47	Uniform[c]	(184, 185, 214, 358, 380, 381)
Adrenalectomy	5,318.21 (-)	4,605.29/6,040.24	Uniform[h]	(379)
Antihypertensive drugs (annual cost)	95.20 (0.60)[i,j,k]	25,406.15/0.00	Gamma	(382)
MRAs (annual cost)	90.99 (-)[l]	68.24/113.74	Uniform[m]	(304)
CVE treatment initial cost	5,091.81 (45.09)[l,k,n]	12,753/0.40	Gamma	(383)
CVE treatment subsequent cost	2,557.61 (64.28)[l,k,o]	1,583.36/1.62	Gamma	(383)
Health Utilities (per Cycle)				
EH	0.890 (-)[p]	0.015/0.045	Uniform[q]	(197)
PA-ADX	0.920 (-)	0.874/0.966	Uniform[r]	(197)
PA-Rx	0.890 (-)	0.015/0.045	Uniform[q]	(197)
EH with CVEs	0.840 (-)[p,s]	0.025/0.075	Uniform[t]	(197, 386)
PA-ADX with CVEs	0.870 (-)[s]			
PA-Rx with CVEs	0.840 (-)[s]			

**Abbreviations:* AVS: adrenal venous sampling; CT: computerised tomography; CVE(s): cardiovascular event(s); EH: essential hypertension; MRAs: mineralocorticoid receptor antagonists; PA: primary aldosteronism; PA-ADX: PA surgically treated; PA-Rx: PA medically treated; SE: standard error; SIT: saline infusion test; TRH: treatment-resistant hypertension; UPA: unilateral PA*

***<u>Notes:</u> [a]Distribution parameters represent 'a' and 'b' values for beta, uniform and gamma distributions, and confidence intervals for log-normal distributions; [b]Arbitrary fit of the uniform distribution using the -/+50% of the deterministic value to calculate the lower and upper bounds; [c]The lower and upper values found/estimated are used to fit the uniform distribution. The average is used as the deterministic value; [d]U-Rhythm's sensitivity and specificity are based on its ability to distinguish PA from EH. Given the lack of data, the same values are assumed when the device is used to distinguish UPA from BPA; [e]Assumed to be equal to the CVE hazard ratio for PA surgically treated patients due to lack of data; [f]Estimated using the respective hazard rate and hazard ratio (if needed). CVE probabilities remain stable over time. Death probabilities change after eleven years, after which they get higher every year; [g]Assumed to be equal to the cost of uncomplicated day case admission for endocrine disorders (Code: KA08C); [h]The lower and upper values found are used to fit the uniform distribution. The weighted average is used as the deterministic value (CT codes: RD201, RD21A and RD22Z; Adrenalectomy codes: KA04A and KA04B); [i]Costs and standard errors inflated to 2019; [j]Annual cost for 2-5 years after initiation of therapy; [k]Assumed to apply to all hypertensive patients; [l]Estimated from the average of spironolactone 100mg/28 tablets (cost = £2.22) and eplerenone 50mg/28 tablets (cost = £5.87) considering that the effective dose for both drugs is 100mg/day; [m]Arbitrary fit of the uniform distribution using the +/-25% of the deterministic value to calculate the lower and upper bounds of the distribution; [n]Cost of first and second CVEs combined for months 1-6 (in the paper); [o]Cost of first and second CVEs combined for months 7-36 (in the paper); [p]Assumed to be equal to the utility score for PA medically treated patients found in the study; [q]Arbitrary fit of the uniform distribution using the -/+50% of the difference between the PA surgical and PA medical treated utility scores to calculate the lower and upper bounds; [r]Arbitrary fit of the uniform distribution using the -/+5% of the deterministic value to calculate the lower and upper bounds; [s]The utility decrement found in Briggs et al. 2017 (386) was subtracted from the respective utility score found in Velema et al. (2018) (197) to find the respective utility score after experiencing CVEs. The utility decrement was assumed to apply to hypertensive patients; [t]Arbitrary fit of the uniform distribution using the -/+50% of the utility decrement to calculate the lower and upper bounds.*

Table 41: Information on the six Markov models used in the analysis

Model	Patient Characteristics	Probabilities	Costs	Utilities
1	-PA with false negative SIT or *U-Rhythm-current* -PA with unrepeated SIT	PA-Rx	Antihypertensives; CVEs	PA-Rx
2	-UPA with false negative *U-Rhythm-theoretical* -BPA with true negative *U-Rhythm-theoretical* -PA with unsuccessful AVS -PA with negative AVS -PA with unsuccessful adrenalectomy	PA-Rx	Antihypertensives; MRAs; CVEs	PA-Rx
3	-UPA with complete surgical success	PA-ADX	CVEs	PA-ADX
4	-UPA with partial surgical success	EH	Antihypertensives; 50% MRAs; CVEs	PA-ADX
5	-EH with true negative confirmatory test -EH with unrepeated SIT	EH	Antihypertensives; CVEs	EH
6	-EH with unsuccessful AVS -EH with negative AVS -EH with unsuccessful adrenalectomy	EH	Antihypertensives; MRAs; CVEs	EH

**Abbreviations: AVS: adrenal venous sampling; BPA: bilateral PA; CVEs: cardiovascular events; EH: essential hypertension; MRAs: mineralocorticoid receptor antagonists; PA: primary aldosteronism; PA-ADX: PA surgically treated; PA-Rx: PA medically treated; SIT: saline infusion test; UPA: unilateral PA*

6.2.2.7 Expected Costs and Outcomes

To estimate the expected cost of each pathway of the decision tree, the costs of all resources used were added and then multiplied by the branch probability (i.e. product of the initial and subsequent event probabilities). To calculate the total costs for each Markov model, the costs of each health state were weighted by the probability of a patient being in a particular state at each cycle (estimated using the *'cohort simulation'* method). Here, the models started with 1,000 patients entering the 'Alive; Without CVEs' state. For each cycle, the proportion of the cohort in each state was derived by using the relevant transition probabilities. When 100 years of age was reached (after 45 cycles/years), all but a very small number of individuals had died. Once the proportion in each state for each year was calculated (*'Markov trace'*), the expected cost per cycle was estimated by summing the cost of each cycle weighted by the number of patients in each state. The total lifetime cost for the patient was then measured by dividing the sum of the expected costs of all cycles by 1,000. The expected cost of each branch of the decision tree was added to the appropriate Markov model cost, which had been weighted by the branch probability, to give the total expected cost for the pathway. By summing the total expected costs of all pathways associated with a diagnostic strategy, the expected cost of that strategy was estimated.

The total outcomes (i.e. life-years, QALYs) for each Markov model were calculated in a similar way to that described for costs. However, for survival duration, this involved multiplying the number of patients in each state per cycle by 1 or 0 (depending on whether they were alive or dead, respectively), while in the case of QALYs, this involved weighting the proportions by the utility values associated with each state. In both lifetime costs and outcomes, half-cycle correction was applied to account for patients moving between states at different times during a given cycle. Additionally, the NHS discount rate (3.5%) (117), set by the UK HM Treasury's guidance on appraisal and evaluation (405), was used to adjust for differential timing. To do so, costs and outcomes were divided by the factor $(1 + r)^{-n}$, where r the discount rate and n the number of years passed (137).

6.2.2.8 Decision Tree Intermediate Analysis

To estimate the efficiency of *U-Rhythm* in diagnosis and treatment selection, an intermediate analysis is performed that measures the incremental diagnostic and initial treatment costs

between the strategies and the incremental number of cases appropriately diagnosed and treated. To do so, the expected cost of each strategy was first calculated using the method described in *Section 6.2.2.7* without summing the expected Markov model costs. Afterwards, the cases that were correctly diagnosed and treated for each strategy were identified and their probabilities per strategy were added. Lastly, the differences between the strategies' costs, cases appropriately diagnosed and cases appropriately treated were measured.

6.2.2.9 Cost-Effectiveness Analysis

The cost-effectiveness analysis (CEA) is performed under the NHS perspective and compares the alternative strategies in terms of their long-term costs and QALYs. The expected costs and QALYs of each diagnostic strategy are used to measure the incremental cost-effectiveness ratios (ICERs) between them (i.e. *U-Rhythm-current* or *U-Rhythm-theoretical* versus SIT) using the formula (17):

$$ICER = \frac{C2-C1}{E2-E1} \qquad \textbf{(Equation 5)}$$

where C_2-C_1 the difference in costs and E_2-E_1 the difference in QALYs between the examined strategies. These are then compared to the current NHS willingness-to-pay (WTP) threshold (£20,000-30,000/QALY) (117) to identify the strategy that would be more cost-effective for the healthcare system. Using ICERs to make healthcare decisions can be problematic since negative ratios do not indicate whether there is a negative difference in costs or effects; positive ratios can be produced when both costs and outcome differences are negative; and very small differences in health benefits can lead to unstable ratios close to infinity. In addition, ICERs can only be used to compare two alternatives at a time, making interpretation difficult when multiple options are compared (17, 137).

To overcome these issues, the net-benefit framework is used. The net benefit is a linearised version of the ICER for a given value of the WTP threshold (e.g. net monetary benefit, $NMB = R_T\Delta E - \Delta C > 0$, where R_T is the payer's WTP threshold, ΔE is the difference in effects and ΔC is the difference in costs) and avoids many problems associated with ICERs. In comparison to ICERs, where the difference between two (average) ratios is not equal to the ratio of the

differences, the difference between two (average) net benefits is equal to the incremental net benefit. Decisions made using incremental net benefits are equivalent to those made using ICERs, with positive values suggesting that the intervention is good value for money and negative values suggesting that it is not cost-effective. Net benefits are also informative when comparing multiple alternatives, with the option having the highest (average) value being the most cost-effective. Therefore, there is no need to consider dominance (*strict* or *extended*)[36] or to specify the appropriate comparator (17, 137). In this analysis, both ICERs and NMBs are presented to help readers identify the diagnostic strategy that is more cost-effective.

6.2.3 Sensitivity Analysis

Different types of SA are conducted to estimate the uncertainty in the decision analysis. First, a PSA is performed to examine the stochastic uncertainty. To do so, a distribution for each parameter was selected (**Table 40**) based on any logical constraints on values, the data informing its estimations, and the method of estimation itself (137). Specifically, given that probabilities must be constrained on the zero-one interval, a beta distribution was considered more appropriate. This distribution fits the data naturally since it has two parameters, called *'alpha'* and *'beta'*, which can be interpreted as the 'number of events' (e.g. TP cases) and the 'number of non-events' (e.g. FN cases), respectively. The beta distribution was also used for the rates that were used to estimate transition probabilities, while the log-normal distribution was chosen for hazard ratios given that these were measured on the natural log scale. For probabilities where high uncertainty was present (e.g. different values from different sources of same quality; values based on expert opinion), the uniform distribution was used bounded between the lowest (alpha) and highest (beta) values found in literature or created arbitrarily from the deterministic value.

In a similar way, the uniform distribution was fitted for all the costs of tests, surgery and MRAs given the variety of costs found in different sources. For costs that came from large longitudinal studies (i.e. other antihypertensive medications; CVEs), the gamma distribution

[36] A *'dominant'* strategy is an intervention that costs less and provides more health benefits. An *'extended dominant'* strategy is an intervention that has a higher ICER than the next more effective option (17, 137).

was preferred since this fits the properties of cost data (i.e. restricted on the interval zero to positive infinity). The gamma distribution has again an alpha and a beta parameter which can be used to fit the distribution. These were measured from the standard errors of the costs found in the respective source. Lastly, for utility scores, the uniform distribution was used since no moments were provided by *Velema et al. (2018) (197)* to fit beta distributions. The latter are considered the best option given that utilities are typically bounded from zero to one (137). After fitting all distributions, 10,000 Monte Carlo simulations were undertaken, each one drawing random values from the respective distribution of each model parameter to estimate the impact of the change in values on the incremental costs, number of cases appropriately diagnosed and treated, life-years and QALYs.

Second, a univariate SA is conducted in which the impact of the extreme values of all model parameters (95% upper/lower CI bounds for parameters with beta/gamma distributions; 95% percentiles for parameters with uniform distributions) on incremental diagnostic costs and cases appropriately diagnosed and treated is measured. Third, for aspects of uncertainty related to *U-Rhythm*'s cost, a threshold SA (headroom approach) analogous to the one described in **Chapter II** is conducted. In this deterministic SA, single point estimates of the device's cost are used to examine the maximum cost that the device can have in order to be more cost-effective compared to SIT. Similarly, different values of its sensitivity and specificity (independently and in combination) are used to explore their implications for NMBs.

6.2.4 Value of Information Analysis

A VOI analysis is conducted to investigate whether further research is needed to support this decision in the future. Given that this is an early EE, this analysis also informs on whether it is worth to invest further on the development of *U-Rhythm* and if yes, which parameters should be explored in a future clinical study. To examine whether further research is potentially worthwhile, the EVPI surrounding the adoption decision is initially estimated by calculating the difference between the expected NMB with perfect and current information. Specifically, the average of the maximum NMB between the two strategies evaluated for each one of the 10,000 PSA simulations was first measured and then the maximum of their average NMB

based on the PSA was subtracted. Given that this informs the decision for the individual patient, the next step was to express it for the total population of patients who could benefit from additional information (*'effective population'*) over the expected lifetime of *U-Rhythm* (137, 364, 365).

In the analysis, two types of UK adult patients are considered: a) those with TRH, and b) those with suspected PA after a positive ARR test. To estimate the number of TRH patients (upper effective population), the UK adult population (52,403,344) (406) was multiplied by the prevalence of hypertension (30%) (407) and the incidence of TRH (1.2%) in the hypertensive population (408), while for PA patients (lower effective population), these numbers were also multiplied by the prevalence of PA after a positive ARR test (11.2%) (388). Given that SIT has been used as a PA confirmatory test for the last 50 years (409-411) and to take into account the faster advances of technology these days, three effective lifetimes for *U-Rhythm* are considered: 10, 20 and 50 years. For each one of these, future patients were estimated by discounting the number of the effective population examined by 3.5% (117). The total EVPI for current and future patients for each case was measured by multiplying the respective total effective population by the EVPI for the individual patient.

Afterwards, the EVPPI is estimated to identify the individual importance/contribution of different parameters to the overall decision uncertainty. Similar to the EVPI, the EVPPI is the difference between the expected NMB between two strategies with perfect and existing information about their important parameters. To measure this, the parameters of interest were first selected and 1,000 different values (outer loop) for each one of them were drawn from their respective distribution. 1,000 Monte Carlo simulations (inner loop) were then run by keeping each one of these 1,000 values of the parameter of interest constant (i.e. assumed to be known with certainty) and varying all other variables. For each iteration, the expected NMB for each strategy and the maximum NMB between the two strategies assessed were calculated. At the end, the difference between the average of the maximum NMB and the maximum of the mean NMB between the strategies from all 1,000 simulations was measured for the lower and upper effective populations (137, 364, 365).

6.2.5 Quality Assessment

To ensure the quality of the DAM developed, the Professional Society for Health Economics and Outcomes Research (ISPOR) good research practices for conceptualising and developing decision trees and state-transition models, estimating parameters and their uncertainty, and conducting VOI analysis were followed (364-369). Additionally, the analysis of the model was based on the methods proposed by *Briggs et al. (2006) (137)*, while the *Philips et al. (2004) (181)* checklist was used to identify strengths and limitations. The internal validation of the model was checked using 'extreme value' SA and calculating a Markov trace (370). To assess the quality of the EE conducted, the principles listed in the *Drummond et al. (2005) (17)* and ISPOR Consolidated Health Economic Evaluation Reporting Standards (CHEERS) (178, 179) checklists were followed throughout the analysis.

6.2.6 Analysis Software

To depict the decision tree, the SilverDecisions (http://silverdecisions.pl/) free, online, open-source software was used. The schematic representation of the Markov model as well as the cost-effectiveness, sensitivity/scenario and VOI analyses were conducted using Microsoft® Excel 365 (Microsoft Corporation, Washington, USA) and its Visual Basic for Applications programming language.

6.3 Results

6.3.1 Main Analysis

6.3.1.1 Decision Tree Intermediate Analysis

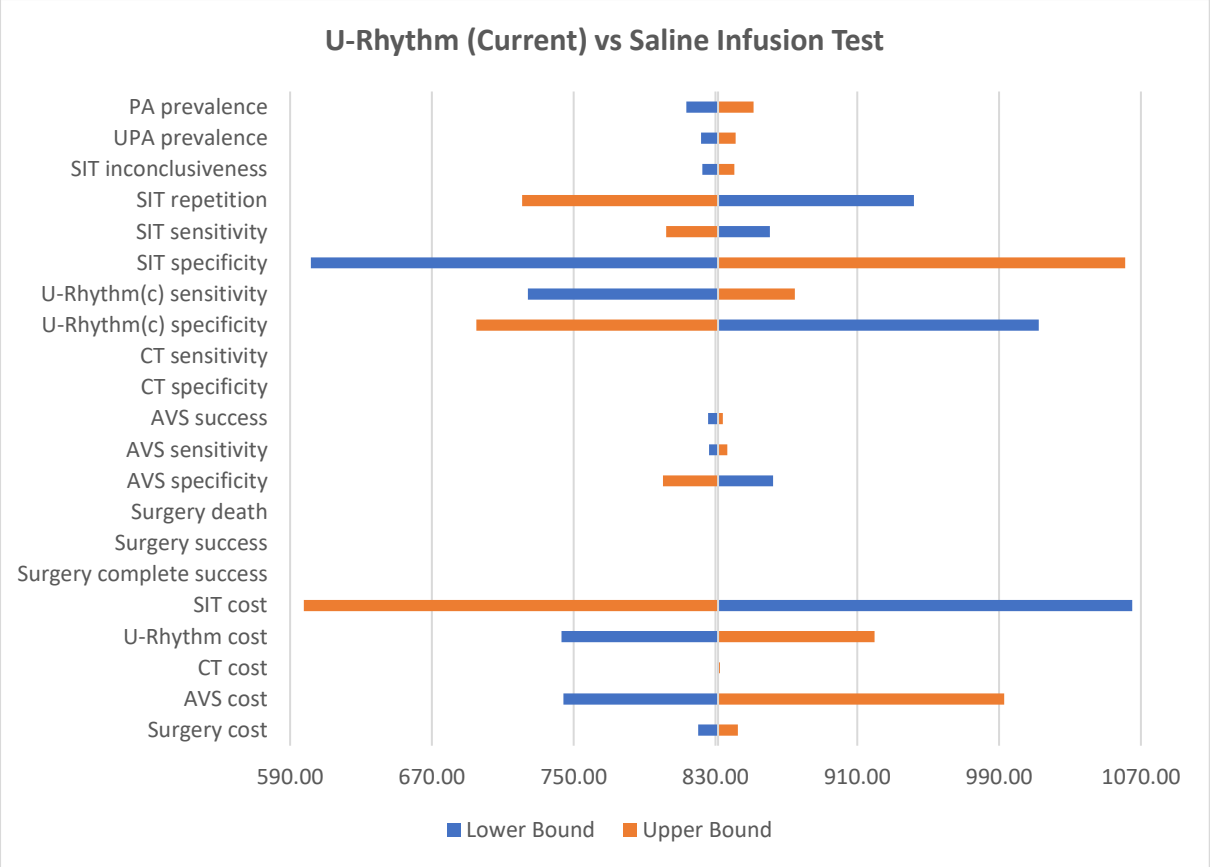
The deterministic analysis showed that the expected costs for SIT, *U-Rhythm-current* and *U-Rhythm-theoretical*, when long-term costs were excluded, were £1,053.90, £1,885.33 and £1,803.02, respectively. Moreover, using SIT led to 67.65% of cases being correctly diagnosed, while the proportions were higher after *U-Rhythm-current* and *U-Rhythm-theoretical* (85.01% and 86.81%, respectively). Although, the percentage of cases appropriately treated was also

lower for SIT when compared to *U-Rhythm-theoretical* (86.32% vs 86.84%), this was higher when compared to *U-Rhythm-current* (85.25%). **Table 42** shows that except for costs, which were slightly higher for each strategy, the rest of the values remained almost identical in the PSA. The distributions of the intermediate incremental costs and outcomes between SIT and *U-Rhythm* (current or theoretical) are presented in ***Appendix 21***. Furthermore, **Figures 23-25** summarise the results of the one-way SAs when extreme values for each tree parameter were used. The confirmatory test diagnostic accuracies and SIT repetition percentage had the biggest impact on both incremental costs and outcomes, while incremental costs were also affected to a large extent by the SIT, *U-Rhythm* and AVS costs.

Table 42: Probabilistic sensitivity analysis results on intermediate costs and outcomes

Strategy / **Outcome**	**Saline Infusion Test**	**U-Rhythm-Current**	**U-Rhythm-Theoretical**
Cost (£) mean (95% CI)	1,148.37 (633.83-1,855.77)	2,015.96 (1,508.40-2,620.53)	1,862.48 (1,539.24-2,266.27)
Cases correctly diagnosed (%) mean (95% CI)	67.69 (54.41-82.25)	85.06 (79.09-90.16)	86.81 (80.87-91.80)
Cases correctly treated (%) mean (95% CI)	86.34 (77.69-94.86)	85.30 (79.31-90.42)	86.83 (80.90-91.81)

**Abbreviations:* CI: confidence interval*



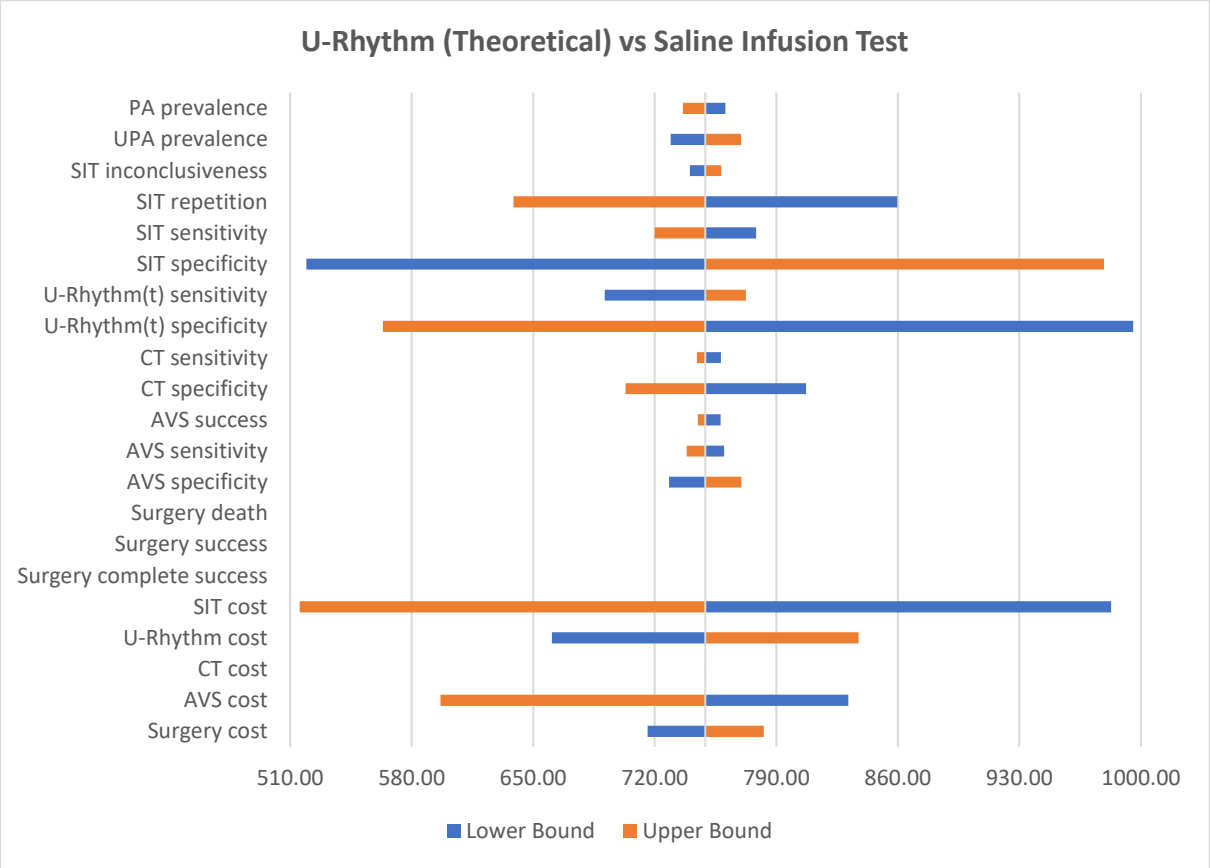
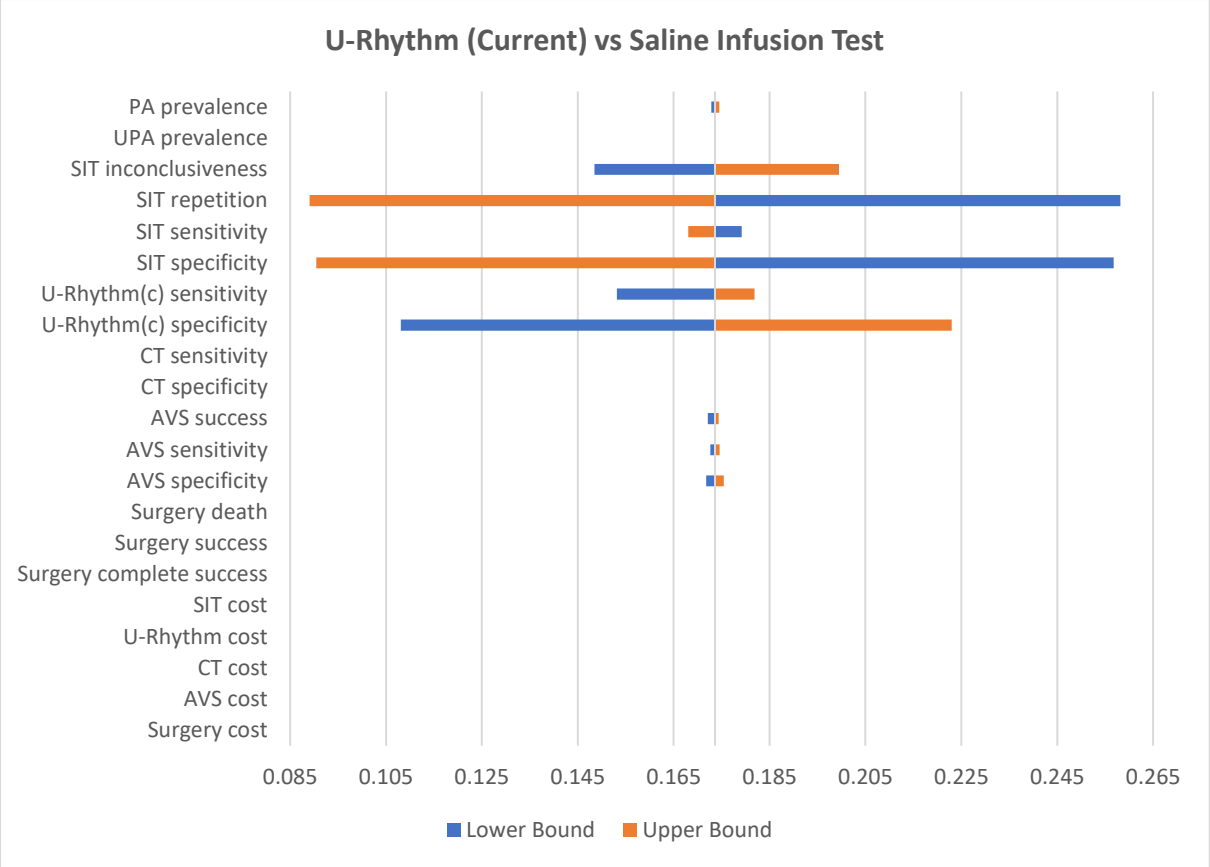


Figure 23: Tornado diagrams of incremental short-term costs



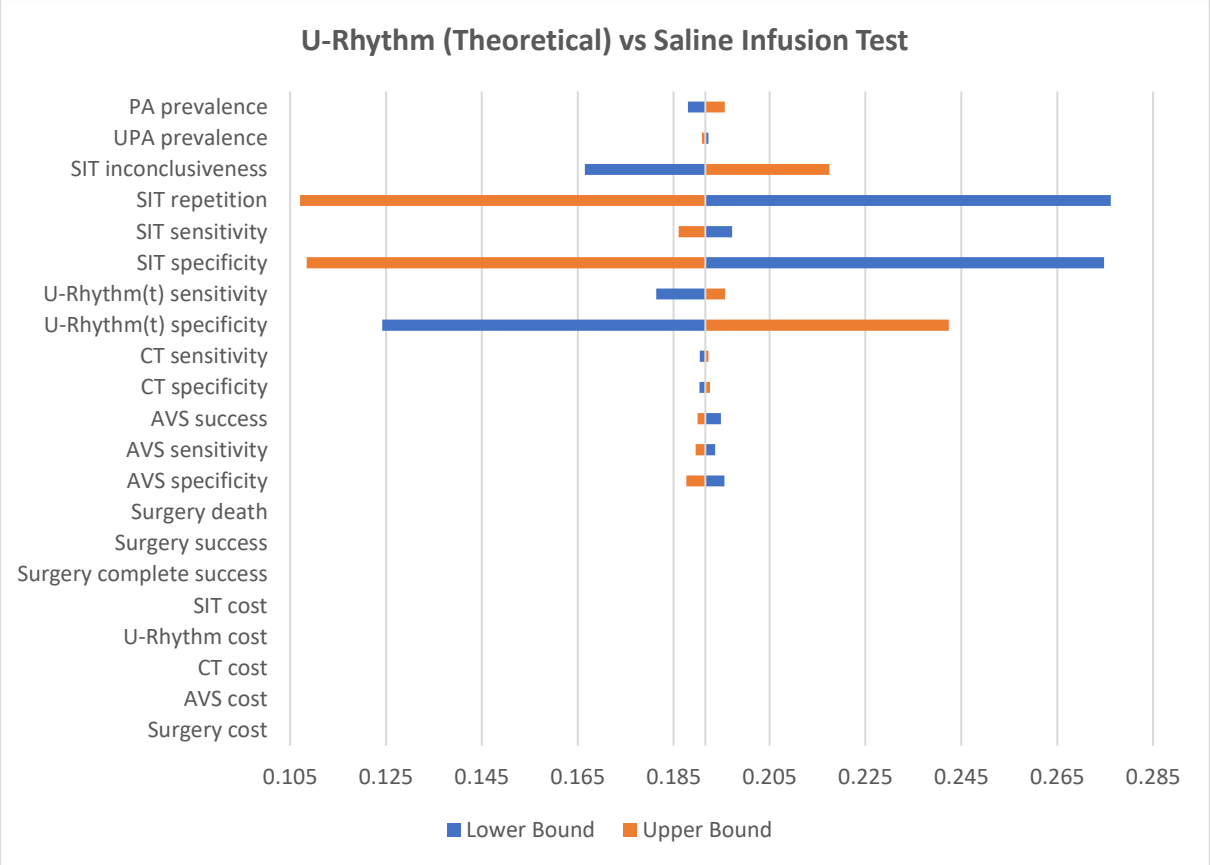
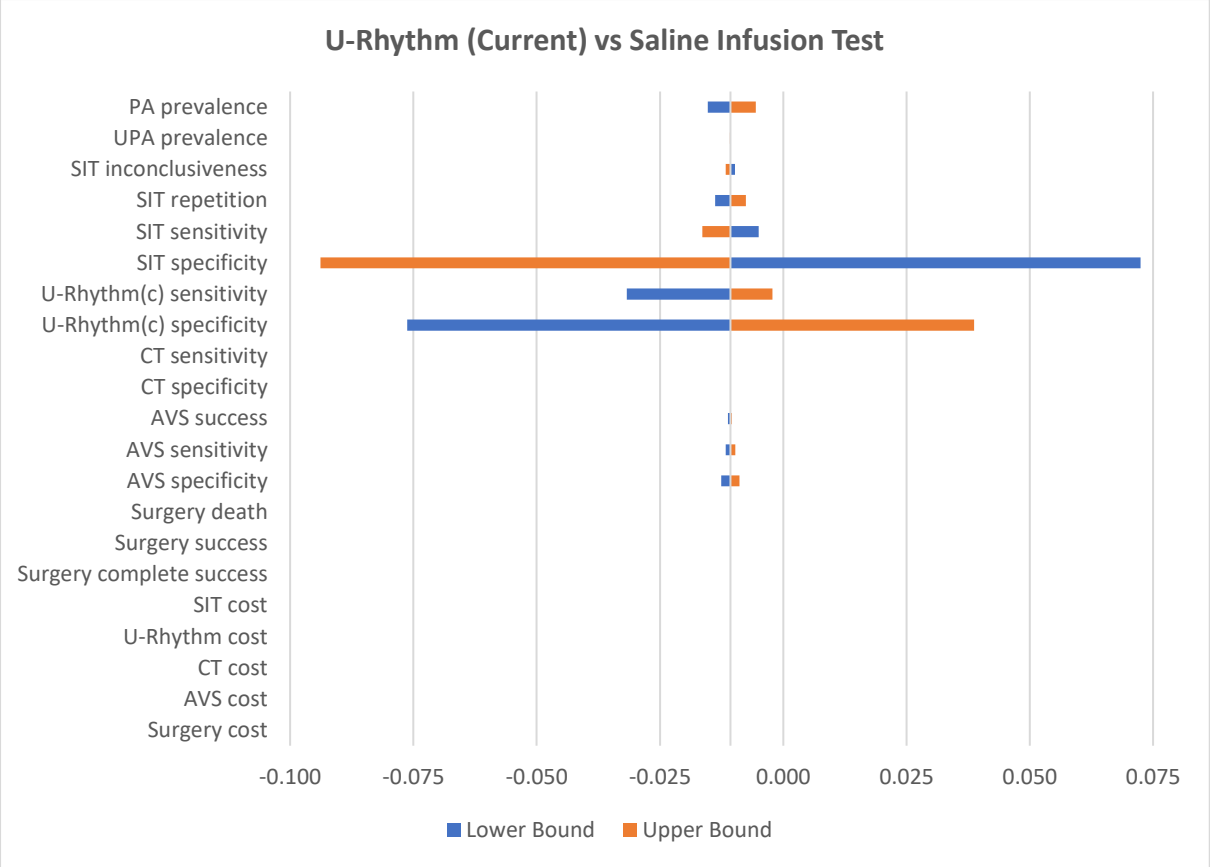


Figure 24: Tornado diagrams of incremental cases appropriately diagnosed



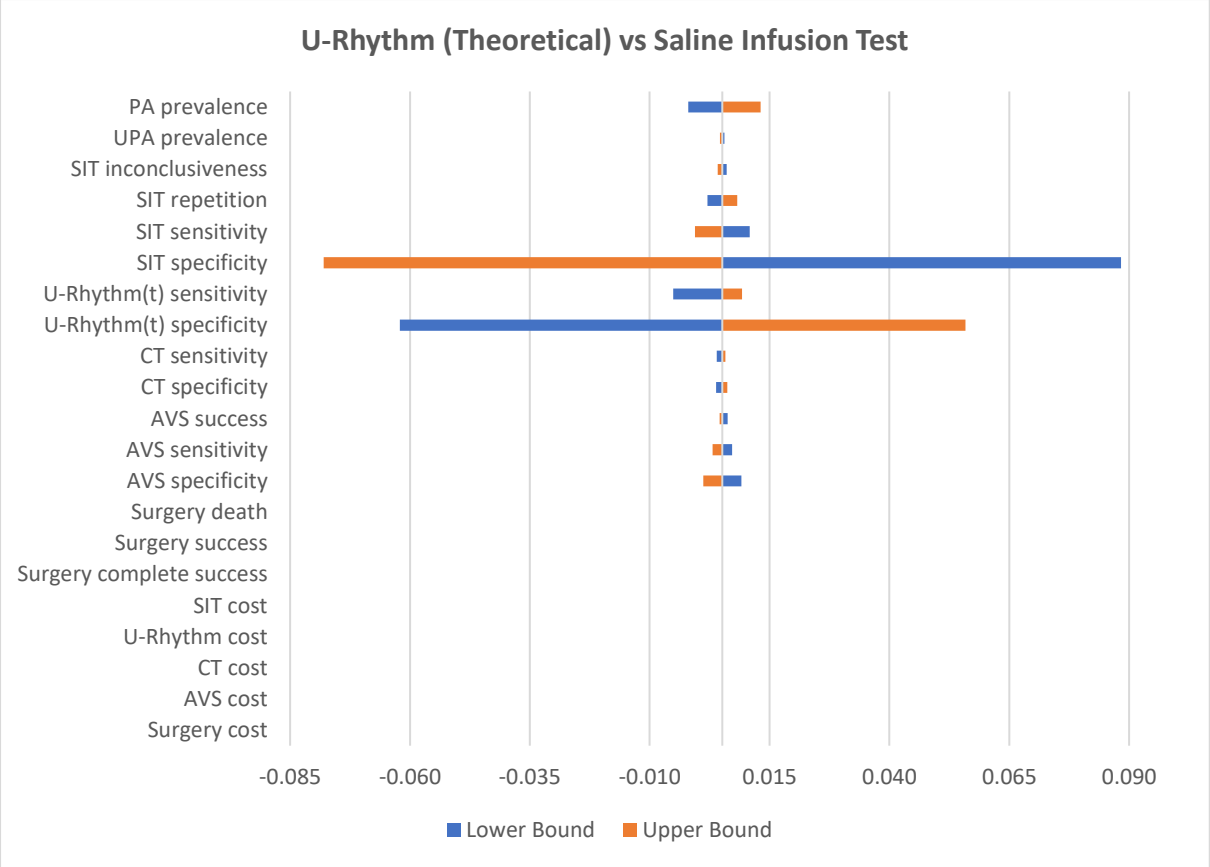


Figure 25: Tornado diagrams of incremental cases appropriately treated

6.3.1.2 Cost-Effectiveness Analysis

The results from the deterministic CEA indicated that the expected long-term costs and QALYs for the SIT strategy were £13,613.75 and 13.24, respectively. Using *U-Rhythm* in the current clinical practice was more costly (£14,429.36) and more effective (13.26). Similarly, the device led to higher healthcare costs (£14,313.34) and QALYs (13.27) in the theoretical strategy. The ICER produced between *U-Rhythm-current* and SIT was £50,705/QALY, while that between *U-Rhythm-theoretical* and SIT was £29,125/QALY. This means that at the lower bound of the NHS WTP threshold (£20,000/QALY), SIT is more cost-effective, while at the higher bound (£30,000/QALY), SIT is a more cost-effective option only when compared to *U-Rhythm-current*. This was also confirmed when measuring the NMBs of SIT, *U-Rhythm-current* and *U-Rhythm-theoretical* (at £20,000/QALY: £251,212, £250,718 and £250,992, respectively; at £30,000/QALY: £383,624, £383,291 and £383,645, respectively). These results remained stable in the PSA (**Table 43**) with the incremental NMB between *U-Rhythm-current* and SIT being -£531 (95% CI: -1231, 93) and that between *U-Rhythm-theoretical* and SIT being -£187 (95% CI: -895, 578) at the lower bound, and -£372 (95% CI: -1107, 335) and £52 (95% CI: -770, 941), respectively, at the upper bound of the NHS WTP threshold. Lastly, both deterministic and probabilistic analyses showed that when diagnosis is made using SIT, the notional patient is expected to live 15.10 years, whereas when it is made with *U-Rhythm-current* or *U-Rhythm-theoretical*, individuals are expected to live 15.11 or 15.12 years, respectively.

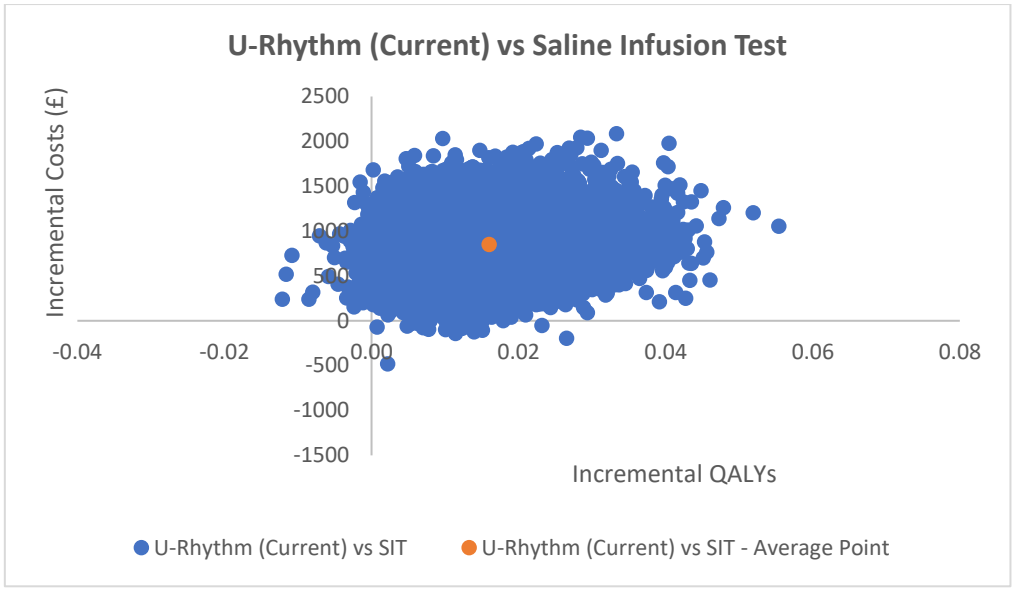
Figure 26 presents the joint distribution of incremental costs and QALYs on cost-effectiveness planes for each pairwise strategy comparison based on PSA results. As shown, the simulations cover more than one quadrant, making ICER and CI results difficult to interpret. For this reason, the probability of each strategy being cost-effective at different WTP thresholds is presented using cost-effectiveness acceptability curves (**Figure 27**). As demonstrated, when all strategies are compared, SIT remains the most cost-effective alternative until the WTP threshold gets around £28,500/QALY, with *U-Rhythm-theoretical* becoming more cost-effective thereafter. On the contrary, the device does not become cost-effective even at extremely high WTP thresholds when used only as a confirmatory test to replace SIT. When SIT is only compared to *U-Rhythm-current*, the former is the more value for money option until the WTP threshold reaches £54,000/QALY.

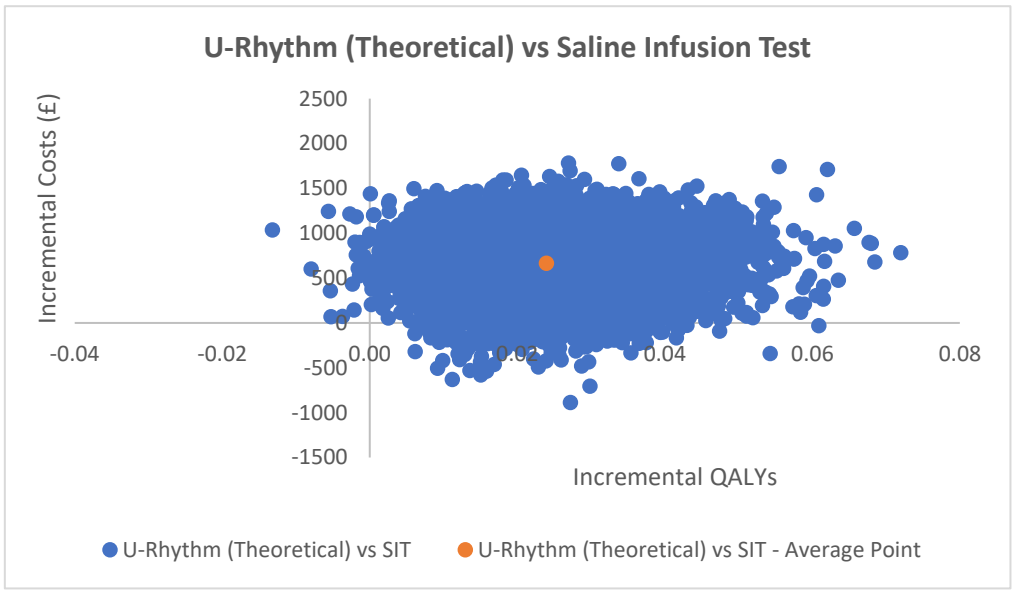
Table 43: Probabilistic sensitivity analysis results on total cost-effectiveness

Strategy / Outcome	Saline Infusion Test	U-Rhythm-Current	U-Rhythm-Theoretical
Expected Cost (£) mean (95% CI)	13,695.16 (12,756.69-14,710.19)	14,545.64 (13,658.77-15,500.94)	14,361.42 (13,581.53-15,187.53)
Expected QALYs mean (95% CI)	13.24 (12.50-13.98)	13.25 (12.51-14.00)	13.26 (12.52-14.01)
Expected Life-Years mean (95% CI)	15.10 (14.99-15.20)	15.11 (15.01-15.21)	15.12 (15.02-15.21)
ICER[¤] (vs SIT)	-	£53,298/QALY	£27,831/QALY
£20,000/QALY Willingness-to-pay Threshold			
NMB (£) mean (95% CI)	251,043 (236,233-266,058)	250,512 (235,720-265,522)	250,855 (236,069-265,826)
iNMB[¤] (vs SIT) mean (95% CI)	-	-531 (-1,231 to 93)	-187 (-895 to 578)
£30,000/QALY Willingness-to-pay Threshold			
NMB (£) mean (95% CI)	383,412 (361,186-405,826)	383,040 (360,837-405,479)	383,464 (361,305-405,871)
iNMB[¤] (vs SIT) mean (95% CI)	-	-372 (-1,107 to 335)	52 (-770 to 941)

**Abbreviations:* CI: confidence interval; ICER: incremental cost-effectiveness ratio; iNMB: incremental net monetary benefit; NMB: net monetary benefit; QALY(s): quality-adjusted life-year(s); SIT: saline infusion test*

***Notes:* Both ICERs and iNMBs are presented to indicate the differences in cost-effectiveness between the examined comparators (i.e. U-Rhythm-current vs SIT; U-Rhythm-theoretical vs SIT); ICERs were calculated from the average expected costs and QALYs of each strategy; Any negative iNMB values represent that SIT had a higher NMB.*





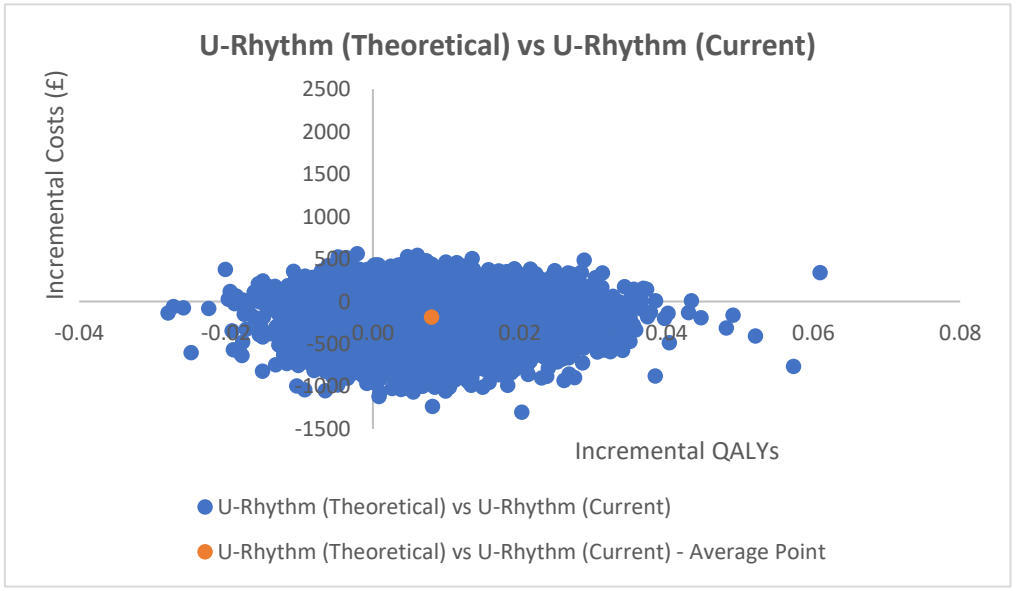
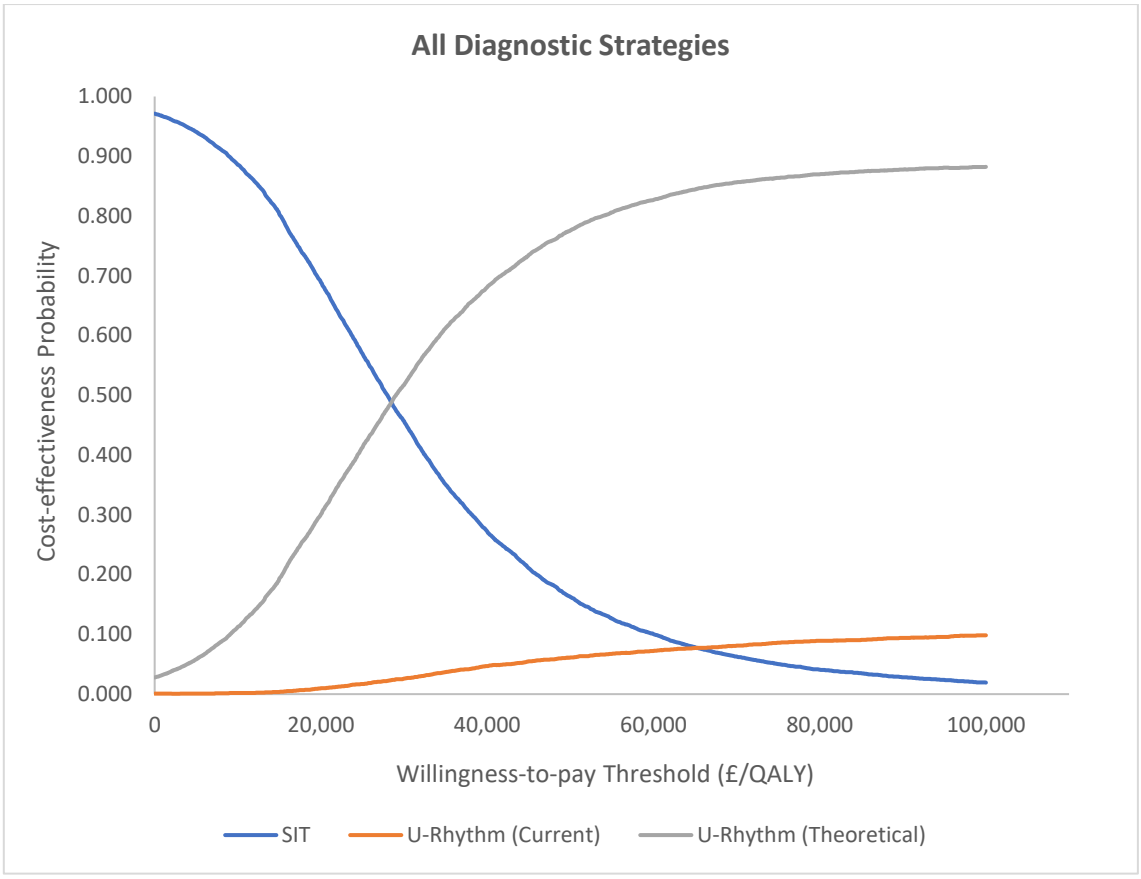


Figure 26: Cost-effectiveness planes



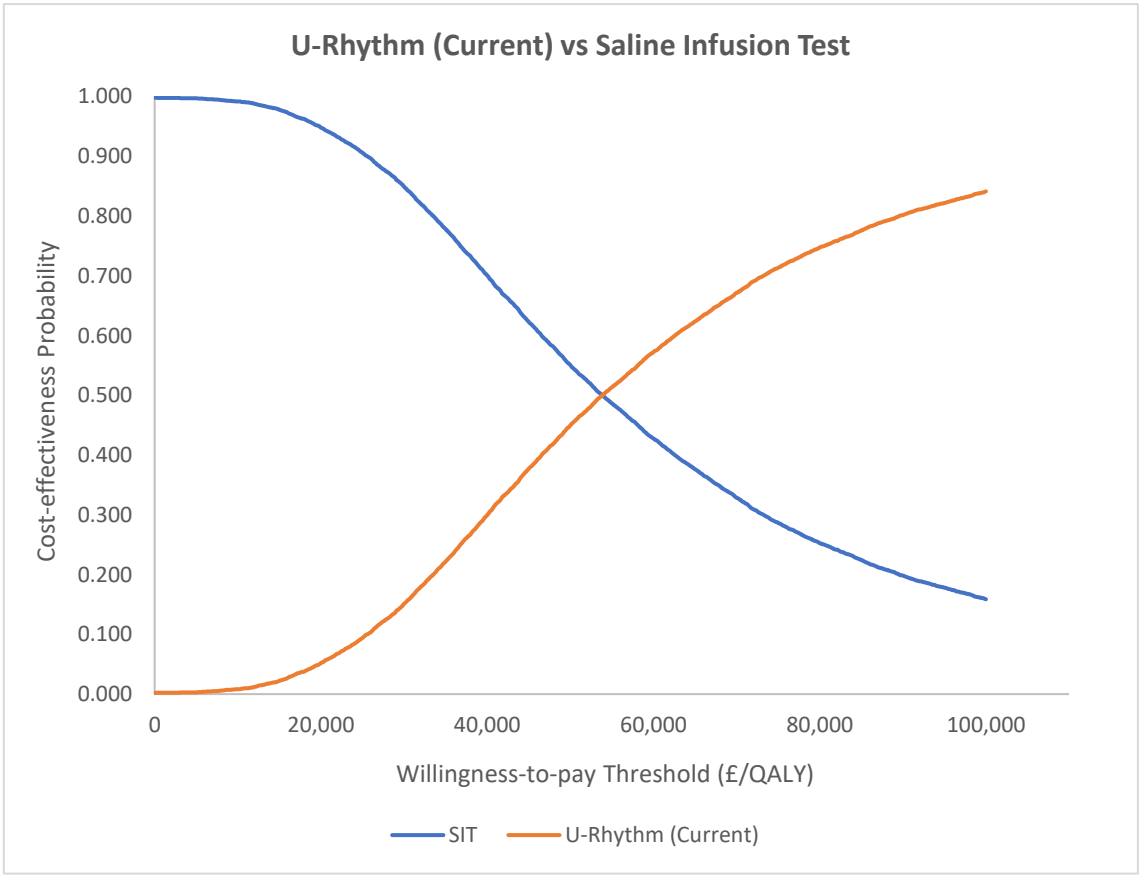


Figure 27: Cost-effectiveness acceptability curves

6.3.2 Headroom Approach

Since *U-Rhythm* is a novel technology that is aimed for use in a relatively rare condition, the £30,000/QALY NHS WTP threshold was considered more appropriate as a ceiling ratio when measuring NMBs in the headroom approach. This is because the National Institute for Health and Care Excellence (NICE) is more flexible with thresholds when assessing new technologies and rare diseases, in which case it considers additional criteria, such as the innovation of the intervention, its position in the care pathway and the availability of alternatives (see *Section 1.7.3.3*) (87, 115, 117). NICE's threshold is not related to the NHS budget and is arbitrary since it was not based on empirical research. Therefore, there is currently a debate about whether NICE's thresholds should change based on the intervention's use and cost to ensure that resources are allocated efficiently (412-416). Especially for rare diseases, a higher threshold would give incentives to developers to invest in new technologies for these diseases instead of other more common conditions. To address this, NICE has produced guidance for highly specialised technologies, where the threshold is higher (£100,000/QALY) (417, 418). Although these guidelines refer to drugs, NICE is likely to follow the same process for (diagnostic) devices to improve patient outcomes (87, 115).

Figure 28 presents how an increase in *U-Rhythm*'s diagnostic accuracy (sensitivity and specificity tested alone and in combination) and a decrease in its cost would influence its NMB in both current and theoretical practice. Given that *U-Rhythm-theoretical* was more cost-effective than SIT at this WTP threshold, its changes in NMBs are not discussed further. Regarding *U-Rhythm-current*, as its sensitivity increased, its NMB increased linearly but never exceeded SIT's NMB. On the contrary, when its specificity was around 0.957, it became the more cost-effective strategy between the two. When its sensitivity and specificity were examined simultaneously, a ≈52% increase in their current values (measured on a scale from their respective deterministic value to '1') was needed for *U-Rhythm* to become more cost-effective. As far as the device's cost is concerned, this needed to drop by ≈£333.50 from its current value for *U-Rhythm* to become more cost-effective.

6.3.3 Value of Information Analysis

Figures 29-30 illustrate the lower and upper population EVPI surrounding the decision to choose the *U-Rhythm-strategies* over SIT at different ceiling ratios. As shown, the EVPI curves related to each pairwise comparison had a similar trend for all effective lifetimes with the maximum value of EVPI increasing as the number of years that *U-Rhythm* was expected to be in use increased. Furthermore, the figures indicate that at £30,000/QALY, where *U-Rhythm-current* was not expected to be more cost-effective than SIT based on current information, EVPI was relatively low (<£20 million in the lower effective population; <£150 million in the upper effective population). In this case, the EVPI increased rapidly with the threshold until it reached its peak at ≈£53,300/QALY and then fell at a declining rate. This is because, at lower thresholds, the probability of error increased and the value of a decision error was higher. On the contrary, at higher thresholds (e.g. >£60,000/QALY), where *U-Rhythm-current* was expected to be more cost-effective, there was more certainty that the adoption decision was the correct one. For *U-Rhythm-theoretical*, the EVPI for both lower and upper effective populations reached a maximum at ≈£27,800/QALY indicating that around this value there was a high uncertainty in whether the device should be adopted or rejected. This is because this threshold was almost equal to the ICER between the device and SIT.

Figure 31 indicates that the specificity of SIT was the parameter that had the greater EVPPI in all cases. SIT's sensitivity and both the sensitivity and specificity of *U-Rhythm* had a zero EVPPI when the device was used as a confirmatory test only, meaning that further research on these parameters would not be cost-effective. In contrast, these parameters affected the expected costs and effects of each alternative when the device was used in theoretical practice. More precisely, *U-Rhythm*'s sensitivity was the second most important parameter, while its specificity followed. Of all parameters, SIT's sensitivity had the lowest contribution to the overall decision uncertainty. The high EVPPI showed that undertaking further research on *U-Rhythm*'s diagnostic accuracy would be worthwhile if it was for use in the theoretical context.

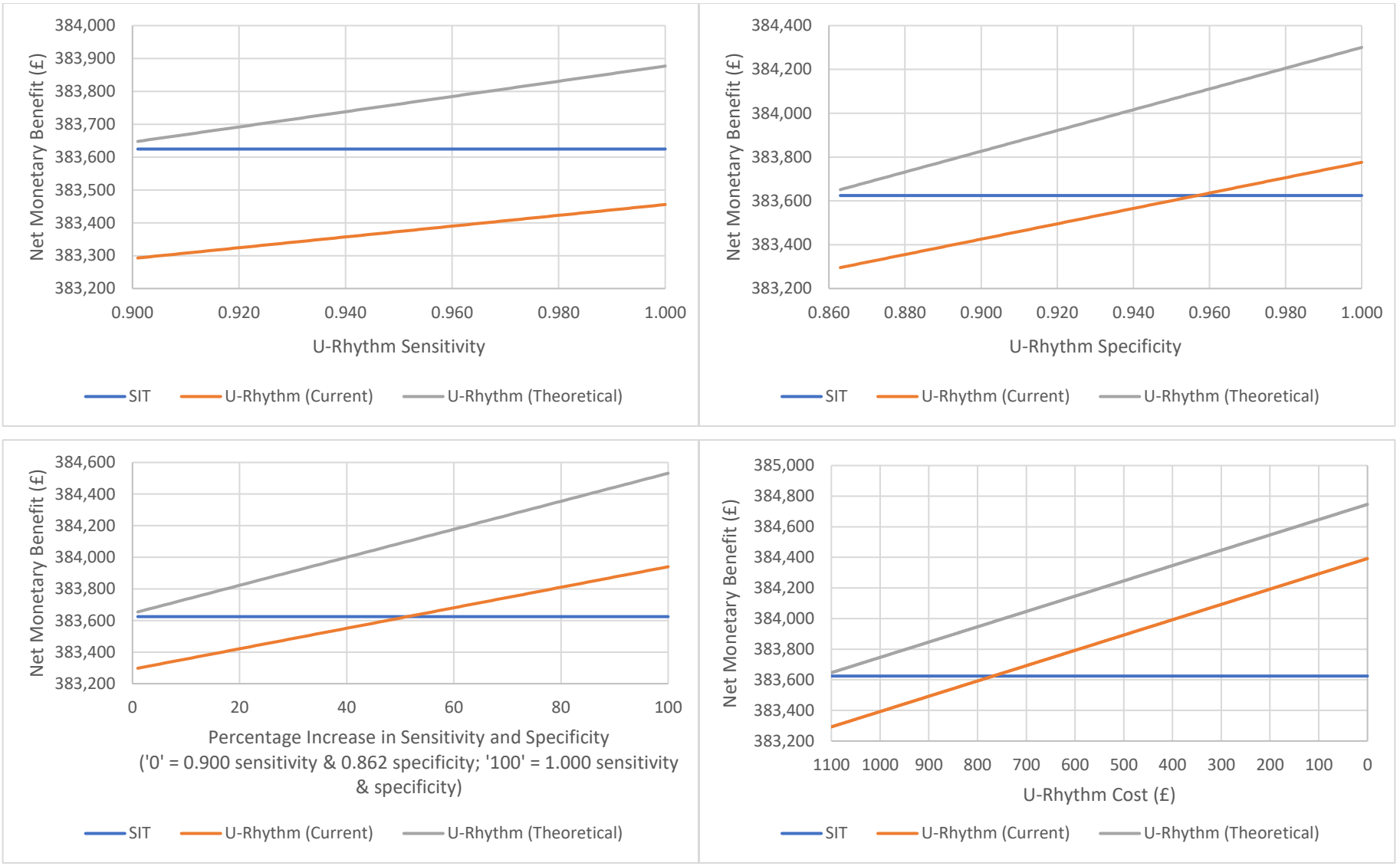
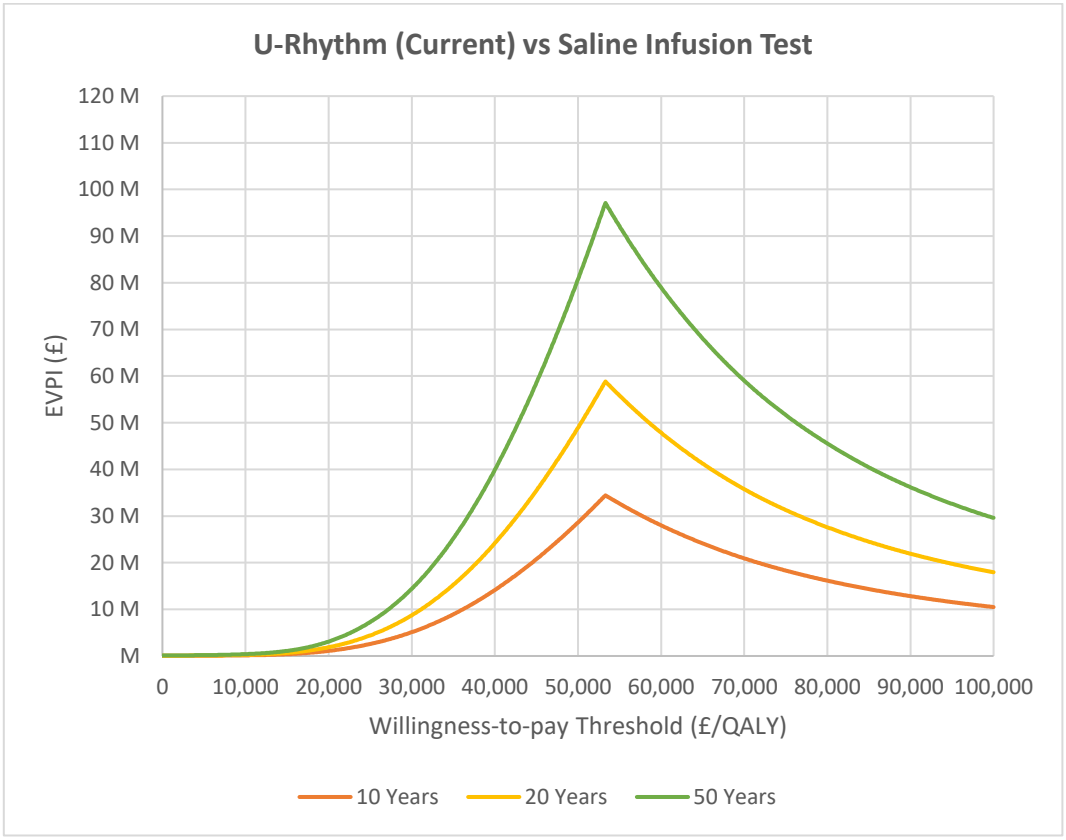


Figure 28: Headroom approach results



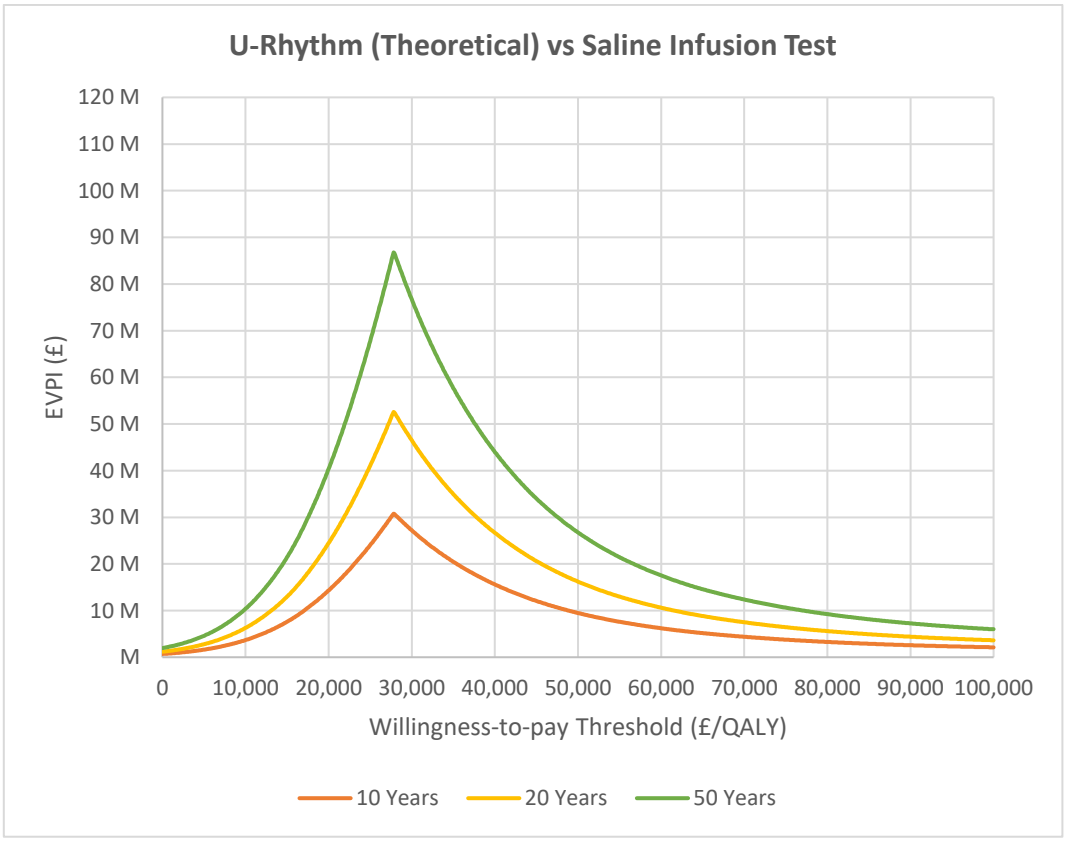
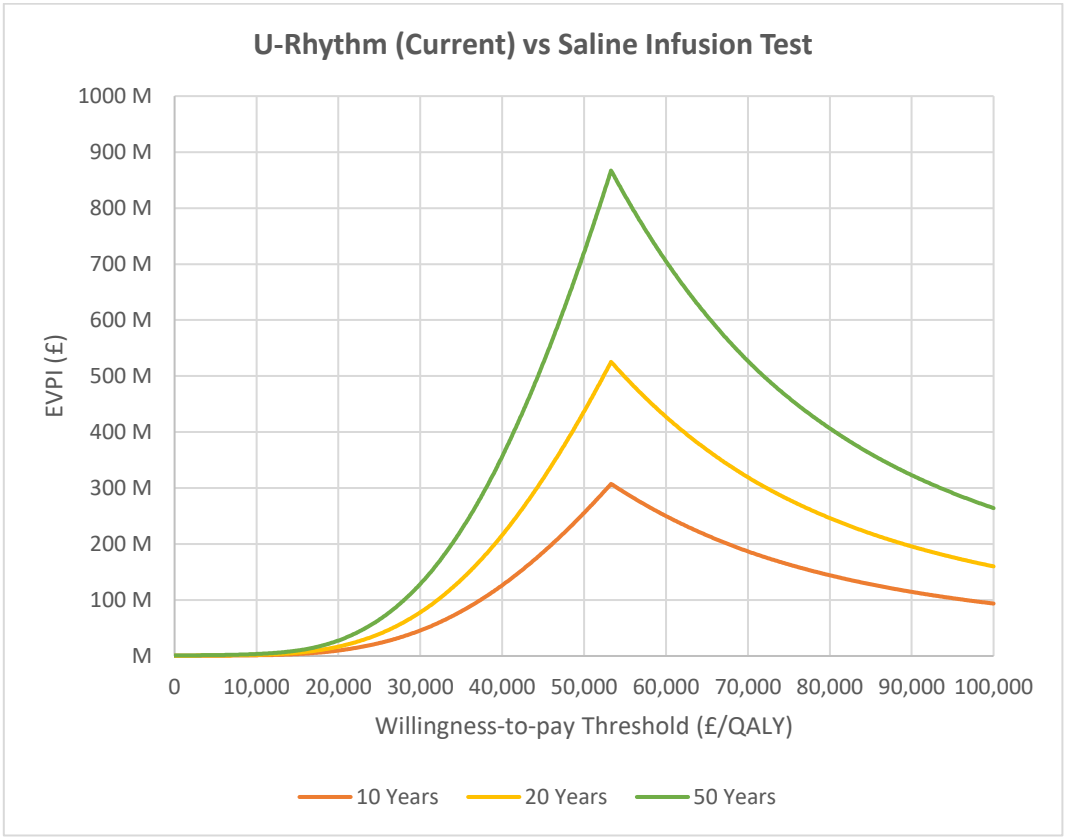


Figure 29: Expected value of perfect information curves (lower effective population)



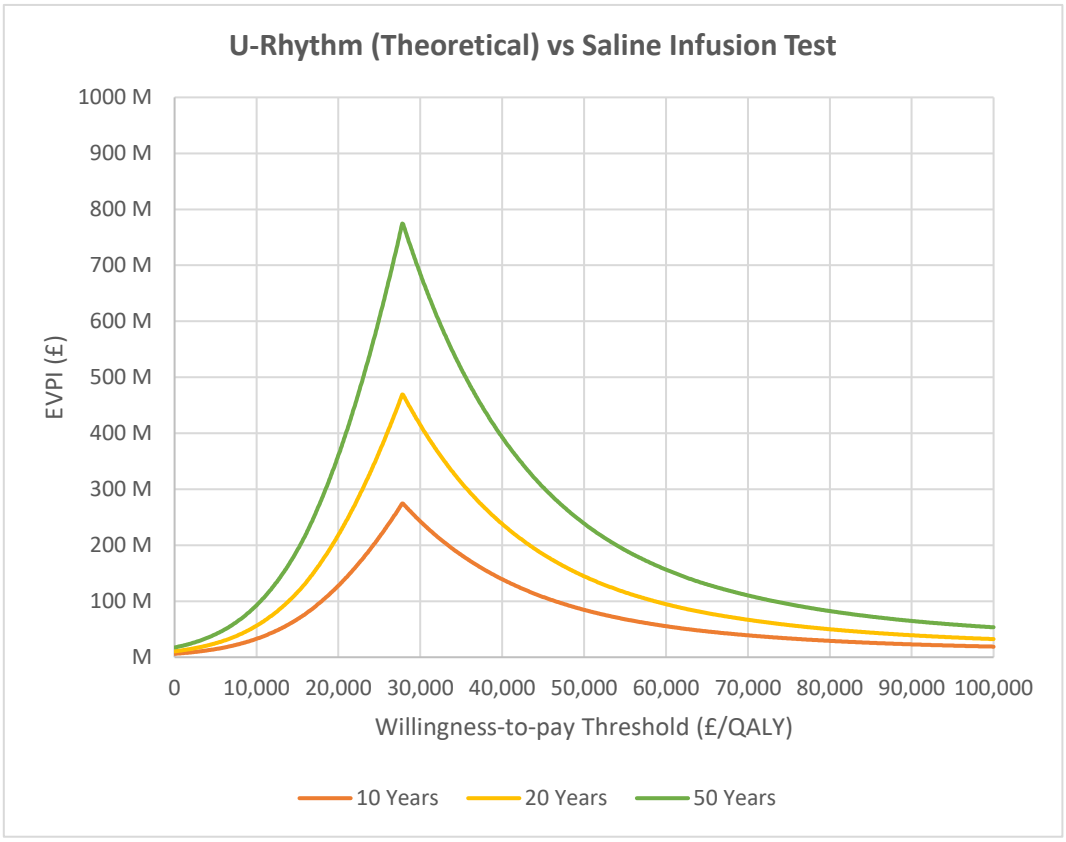


Figure 30: Expected value of perfect information curves (upper effective population)

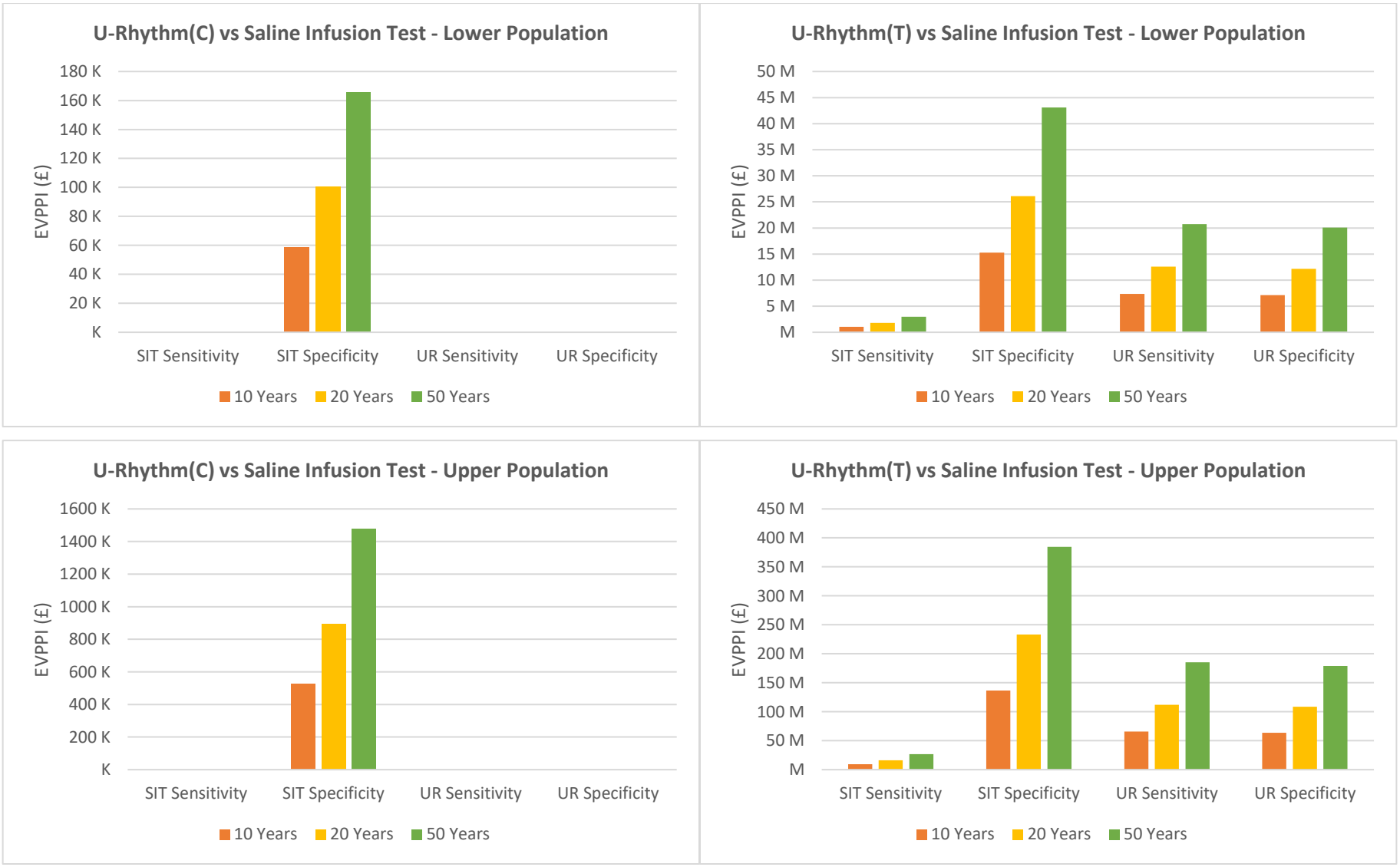


Figure 31: Expected value of partial perfect information results

6.4 Discussion

An early EE (CEA) using DAM techniques was conducted to examine whether *U-Rhythm* could be a more cost-effective alternative for confirming (current clinical practice) and potentially subtyping (theoretical practice) PA compared to the most commonly used PA confirmatory test in the UK and Europe (i.e. SIT). The aim of the study was also to identify the potential impact of *U-Rhythm*'s diagnostic accuracy and cost, which are expected to evolve as the technology matures, on the overall adoption decision. For this purpose, several SAs, the headroom approach and a VOI analysis were implemented. To conduct the analyses, a DAM was developed based on the latest Endocrine Society PA guidelines (46), suggestions made by the *ULTRADIAN* endocrinologists, and two previous cost-effectiveness models (185, 186). The analyses were performed and critically appraised using recommendations from ISPOR (364-370), *Drummond et al. (2005) (17)* and *Briggs et al. (2006) (137)*.

6.4.1 Key Findings

The outcomes of the deterministic analysis indicated that based on current evidence and at the £20,000/QALY WTP threshold (NHS lower ceiling ratio), SIT was the most cost-effective strategy. The second most cost-effective option was to use *U-Rhythm* as a confirmatory and subtyping test, making the use of the device as a confirmatory test only the least efficient alternative. At the £30,000/QALY WTP threshold (NHS upper ceiling ratio), the device was the most cost-effective option when used in the theoretical context, while it remained the least efficient when used as a confirmatory test only. Furthermore, *U-Rhythm* led to more cases being correctly diagnosed in both current and theoretical practice, but more cases being appropriately treated only when the device was used in the theoretical context. All strategies led to similar life expectancies for patients. Results remained relatively robust in PSA, while the univariate SA indicated that incremental costs and outcomes were mainly affected by SIT's and *U-Rhythm*'s diagnostic accuracy and the costs of the confirmatory tests and AVS. The headroom approach showed that, at the £30,000/QALY WTP threshold, the cost of *U-Rhythm* would need to drop by ≈30%, or its sensitivity and specificity should simultaneously increase by ≈52% from their current values for the device to be more cost-effective in current practice. At the same threshold, the VOI analysis demonstrated that there was a lot of

uncertainty surrounding the decision to choose *U-Rhythm* over SIT in the theoretical practice (i.e. further research would be worthwhile). Additionally, the analysis showed that if the cost of *U-Rhythm* does not decrease, the use of the device to replace SIT as a confirmatory test could be rejected with confidence (i.e. further research would not be cost-effective).

The results of the main analysis were expected since the actual cost of testing was increased when *U-Rhythm* was used, with the vast majority of patients receiving the same diagnosis and treatment. Even when the device replaced the expensive AVS, the total test-treatment costs remained high, with only a small proportion of patients getting more appropriate diagnosis and treatment. The fact that the probability of SIT repetition and the diagnostic accuracies of both confirmatory tests were shown to be the parameters with the highest influence on both costs and benefits was anticipated given that all patients in the model got a confirmatory test, whereas other parameters (e.g. CT) were only experienced by a minority of individuals. Additionally, there was a lot of uncertainty when estimating the values of these parameters. Specifically, for SIT, a variety of estimates on its diagnostic accuracy were found in literature due to the fact that there is currently no consensus on the optimal laboratory assays and cut-off points that should be used (389-395, 419-422). For *U-Rhythm*, uncertainties are large because it is still at an early phase of development and all values were estimates based on current evidence and/or expectations.

The headroom approach confirmed the outcomes of the main analysis highlighting that both the diagnostic accuracy and cost of *U-Rhythm* need to be improved in order to become more cost-effective when used as a confirmatory test only. The VOI analysis agreed showing that there is a low value in conducting additional research on the device's diagnostic accuracy. Here, it should be noted that any EVPI and EVPPI found in the analysis represent the maximum benefit of further research. In reality, a study could only reach these 'perfect' values if it enrolled an infinite number of participants, but this is unrealistic. Interestingly, SIT's specificity EVPPI was high despite the fact that the test has been used as a PA confirmatory test for the last 50 years (409-411). This illustrated how much uncertainty is present in PA diagnosis. Moreover, *U-Rhythm* has the potential to be used in several endocrine disorders, for different purposes and across a range of different patient groups. Given that *U-Rhythm* is still evolving as a technology, to decide whether further research and investment on it would be cost-

effective, it would first be necessary to have its cost and analytical techniques stabilised and then conduct similar analyses for all its uses. For example, the cost of using *U-Rhythm* is expected to reduce dramatically should it be accepted in practice. However, there is no easy way to parameterise the uncertainty about how much the cost will reduce by in the future. Therefore, this is ignored in the VOI analysis leading to a possible false negative finding to reject the need for further research. Stabilising these parameters is essential before any decision on the device is taken to avoid rejecting a potentially beneficial technology.

6.4.2 Model Parameter Estimation and Uncertainty

In early EEs, and especially in studies like the present where new technologies for rare diseases are assessed, uncertainty in all its forms is expected to be more prevalent. This is because data on parameters are obtained from smaller size and potentially lower-quality studies (e.g. due to lack of a reference standard), and often conducted at earlier stages of product development, meaning that any estimates are likely to evolve over time. Additionally, any assumptions made on the model structure are based on limited information and more likely, expert opinion. Methods that rely on probability distributions (e.g. PSA, EVPPI) are well-suited to explore the uncertainty arising from studies with small sample sizes. However, they are less well-suited for exploring non-stochastic uncertainty or bias that may be more prevalent at the early stage of technology development. Bias-adjusted meta-analysis methods are being developed to address this issue but are not widely applied in DAMs of diagnostic tests to date (423). Therefore, deterministic methods (e.g. univariate/multivariate SA, threshold SA, the headroom approach) are still needed to explore the model robustness and the potential value of further research.

Regarding the parameter uncertainty in the model, as shown in ***Appendix 11***, the preliminary diagnostic accuracy of *U-Rhythm* in the confirmatory diagnosis of PA was estimated by using data from the healthy volunteers and PA patients of *ULTRADIAN*. Nevertheless, differentiating between a healthy individual and a PA patient is likely to be easier than distinguishing between a PA and an EH patient, or a UPA and a BPA patient. In addition, the test will be used in TRH patients who are suspected of having PA after a positive ARR test (46). Therefore,

current sensitivity and specificity values might be an overestimation of *U-Rhythm*'s actual diagnostic accuracy for PA. Furthermore, *ULTRADIAN* included a small number of PA patients, with this analysis estimating sensitivity and specificity values from an even smaller number of participants since not all samplings were available at the time of analysis. Therefore, the values used in the model are likely to change in the future.

The estimation of the current costs of the device, sampling and laboratory analysis treated *U-Rhythm*'s research and development (R&D) costs as 'sunk costs' and instead focused on the equipment, staffing, consumables and overhead costs that would be required to use the test and analyse the samples in future practice. In reality, the development of *U-Rhythm* to this point in time has required substantial investment. Had this R&D been funded with private capital by a for-profit company, these R&D costs would be reflected in the final price of the device. However, as the intellectual property has been developed within a University setting using public funds (European and UK Biotechnology and Biological Sciences Research Council), the price of *U-Rhythm* (and therefore the cost to public healthcare systems) does not need to recoup these R&D investments. Additionally, as mentioned above, the reduction in the costs when the device will be widely used in clinical practice cannot be predicted at this stage of development. Therefore, the CEA results are anticipated to be different when this analysis is conducted before/after launching the final product.

6.4.3 Comparison with Related Studies

Although the structure of the model used in this study resembled that of the DAMs presented in *Lubitz et al. (2015) (185)* and *Sato et al. (2015) (186)*, these studies had several differences in their research question and methods. Specifically, *Lubitz et al. (2015) (185)* compared six diagnostic strategies for identifying UPA after a positive ARR test for TRH patients in the United States: a) SIT/CT/AVS; b) CT/AVS; c) SIT/AVS; d) AVS-only; e) SIT/CT; and f) CT-only. A seventh strategy in which all patients received MRAs without further testing was also evaluated. The DAM combined a decision tree with an existing Markov model of seven CVE risk factors. The DAM considered parameters not investigated in the current model (e.g. ARR; adrenal incidentaloma; location of nodules on the adrenals) and ignored other (e.g. AVS after

a negative CT; hypertension improvement levels after surgery). Lifetime CVE-related costs and QALYs were measured, and parameter uncertainty was assessed using various SAs. CT/AVS was found to be the most cost-effective diagnostic strategy with all strategies including SIT not being cost-effective.

Sato et al. (2015) (186) compared the recommended Japanese PA diagnostic and treatment approach (i.e. ARR/CCT/CT/AVS for diagnosis; surgery or antihypertensive medications for UPA and BPA treatment, respectively) to a drug-only strategy for hypothetical, 50-year-old, male hypertensive patients. A decision tree was combined with a three-state Markov model ('Alive', 'Stroke', Dead') to examine long-term costs and outcomes expressed in life-years. The DAM ignored several parameters used in the current model (e.g. test inconclusiveness, peri-operative mortality). The effects of patient heterogeneity and parameter uncertainty on final outcomes were assessed using several deterministic SAs. The CEA results showed that the conventional strategy was more cost-effective. However, the ICER was compared to an informal WTP threshold estimated in a previous Japanese study. Here, it should be noted that both *Lubitz et al. (2015) (185)* and *Sato et al. (2015) (186)* compared more established diagnostic strategies, meaning that they faced different challenges in their analyses. What makes the current study unique is that it evaluated the potential role of a new test in an already uncertain environment.

The same thing applies when the current analysis is compared to the other two studies identified in the systematic review (**Chapter III**). Specifically, *Velasco et al. (2015) (187)* used a DAM to compare the cost of therapeutic drug monitoring (TDM) guided PA screening (i.e. testing only TRH patients adherent to antihypertensive medications) to that of unselective PA screening (i.e. testing all TRH patients). Diagnostic testing consisted of SIT, CT and AVS. A cost analysis was conducted and deterministic SAs were employed to evaluate its results. The authors concluded that TDM-guided screening is a cost-saving strategy with a lower ability to detect PA and a lower rate of unnecessary testing. *Dekkers et al. (2016) (184)* conducted an RCT to compare the cost-effectiveness of CT and AVS in determining the appropriate treatment for PA. Statistical analysis was used to address the data uncertainty and detect any differences between the two patient groups. The results showed that there are no statistically

significant or clinically important differences between the two strategies, with both tests being imperfect in identifying UPA patients.

Compared to the *ULTRADIAN* study (**Chapter IV**), the analysis reported in **this Chapter** used higher PA utility values to populate the model and estimate long-term QALY gains for the patients. In *ULTRADIAN*, it was unclear how long the patients had their condition for; whether they were experiencing any cardiovascular events; what medications they were receiving (if any); and how recently any surgery was performed. For this reason, the author decided to use utility data from a large Dutch-Polish RCT (197) since after searching the literature, this was the best available source for providing EQ-5D scores for PA patients treated with surgery or medications. Given that studies have shown some small differences in self-reported health between the UK, the Netherlands and Poland (e.g. due to demographic and socioeconomic factors) (424), some slight HRQoL variations between the PA populations might be present. However, utility scores are expected to be high. This is because PA is similar to hypertension, meaning that it does not affect the HRQoL of the patients, unless it is poorly controlled and they experience adverse events, where HRQoL drops (425).

6.4.4 Strengths and Limitations

The current study used a DAM that was conceptualised using various sources to ensure that the diagnosis and management pathways drawn represented the current PA practice in the UK and Europe. A combination of a decision tree examining the efficiency of diagnosis and a Markov model estimating lifelong costs and health benefits was used. To identify the best available data to populate the DAM, several targeted reviews were undertaken, including searches on UK national sources. To deal with parameter and decision uncertainty, several SAs (including a PSA of 10,000 Monte Carlo simulations), the headroom approach and a VOI analysis were performed. The latter two analyses also helped to identify areas where further research should be conducted to make *U-Rhythm* a potentially more competitive alternative for PA diagnosis. The model was developed, analysed and critically appraised using methods proposed by high-quality and widely acceptable sources (17, 137, 181, 364-369). At the end,

the internal validation of the DAM was examined to make sure that the results of the analysis were correct (***Appendix 22***).

Nevertheless, the study did not examine all PA confirmatory tests due to time constraints and the difficulty to identify data on their cost and diagnostic accuracy. Having said that, the use of SIT (most commonly used test in the UK and Europe (374)), makes it an appropriate comparator for *U-Rhythm*. Despite trying to reflect the current practice, the use of a decision tree cannot easily represent the delay in diagnosis which is present when confirmatory tests are inconclusive and need to be repeated. In addition, the Markov model that was developed included only three health states. In reality, more health states based on cardiovascular, renal or other disease events occur (333, 336, 375, 376). These could have been represented with more health states in the model. However, given the lack of evidence to populate the model, this three-state model was considered the best alternative.

Time constraints were also the reason why a full systematic review and a meta-analysis were not undertaken to find data for the model parameters. However, since this was an early EE, the goal was to quickly identify whether it would be worthwhile to continue with the development of *U-Rhythm*, and what its cost and diagnostic accuracy should be to potentially become the more cost-effective diagnostic strategy. Methods have been developed to allow the appropriate meta-analysis of diagnostic accuracy studies accounting for the negative correlation between sensitivity and specificity (426). However, these methods rely on having a sufficient number of studies that have evaluated the examined tests against a common reference standard. Given that such studies are not available in the present clinical scenario, high-quality diagnostic accuracy studies are needed to be conducted in the future.

The fact that *U-Rhythm* is still at an early stage of development means that estimates of diagnostic accuracy are expected to change over time as more data become available. In addition, the estimates used in this analysis were based only on a small number of PA patients and healthy individuals examined in *ULTRADIAN*. Diagnostic accuracy studies (especially case-control studies) are sensitive to spectrum bias. Therefore, there is a need for a future diagnostic cohort study or even RCT to be conducted to obtain better estimates. Furthermore, the pathway that considered *U-Rhythm* as a potential replacement of both confirmatory and

subtyping tests is purely theoretical and based on the assumption that the device is expected to have more conclusive results regarding the cause of PA due to the amount of hormone data that it collects. The pharmacotherapy proposed in the model was also a crude approximation given that the choice of the drugs that are administered is made by a clinician based on the individual's characteristics before and after diagnosis. Another limitation is that CVE costs and their utility decrements came from studies examining non-PA or even non-hypertensive patients. However, given the similarities in the patient characteristics, these parameters are not anticipated to differ much.

To account for patient heterogeneity, individual patient simulation models would have been a more suitable type of models compared to cohort models (17, 137, 360-362). However, as mentioned in *Section 6.1.2.3*, such models require richer data, something not available in this clinical scenario. An alternative would have been to run subgroup analyses based on patient factors (e.g. age, gender) expected to be influential for cost-effectiveness (369). However, these were not explored further given the lack of data and time. Additionally, these more in-depth analyses are more likely to be appropriate once the potential cost-effectiveness of *U-Rhythm* in the whole population is established. Moreover, when running the PSA, the correlation between the sensitivity and specificity of each confirmatory and subtype test were not considered. This could have been done using the Cholesky decomposition method (137). However, in this case, a covariance between these parameters would be needed, which was not available from the analysis of the *ULTRADIAN* diagnostic accuracy data or the data found in literature for the other tests.

The lack of evidence and time were also considerations in the decision not to conduct EVSI. The fact that *U-Rhythm* is still evolving in terms of its cost and diagnostic accuracy meant that an EVSI would not be appropriate currently. In the future, once the cost and assay techniques of *U-Rhythm* have stabilised, EVSI might be used to inform optimal study design (e.g. diagnostic cohort study; RCT), including sample size, length of follow-up and outcomes of interest. If no design has a positive ENBS, current evidence should be considered sufficient for taking decisions (17, 137, 364, 365). Furthermore, the EVPPI measured in this analysis provided a good first step in establishing whether more research would be cost-effective. Given that its values were low or even zero in some cases, it is expected that the EVSI values

would be very low as well, until the cost of *U-Rhythm* is reduced. One limitation in the use of EVPPI was that sensitivity and specificity were not grouped to examine their correlation. Grouping correlated parameters is essential since exploring each parameter independently can lead to under- or over-estimations in the EVPPI values depending on the type and strength of correlation, and the direction of the association with the incremental NMBs. In some cases, correlations between parameters can also lead to EVPPIs even higher than the decision EVPI (17, 137, 364, 365). Nevertheless, in this analysis, it was considered important to evaluate which parameter has the biggest impact on results. Lastly, data identification and analyses were conducted by one analyst, but the whole process was supervised by a more experienced analyst to identify any potential errors.

6.4.5 Implications for Clinical Practice and Future Research

The decision analysis showed that with its current preliminary cost and diagnostic accuracy, *U-Rhythm* is more cost-effective than SIT only when used in the theoretical context and at the upper NHS WTP threshold. The headroom approach indicated that it is the cost of the device that is most influential in determining cost-effectiveness. The VOI analysis confirmed that any research would not be worthwhile until *U-Rhythm*'s cost has first been substantially reduced closer to the costs of current confirmatory tests, such as SIT. Nevertheless, there is currently high uncertainty regarding SIT's diagnostic accuracy despite this being commonly used in current practice. Therefore, it is essential that more high-quality studies, with larger sample sizes and stronger study designs than *ULTRADIAN* (e.g. *Dekkers et al. (2016) (184))*, are conducted in the future to investigate these parameters and compare them to those of other PA confirmatory tests. By doing so, stronger evidence on the optimal test for PA will become available, which can potentially reduce the healthcare costs associated with the disease, reduce the time to diagnosis and treatment, and improve the health benefits for the patients.

6.5 Conclusion

In conclusion, the CEA results indicated that SIT is currently a more cost-effective diagnostic strategy for PA when compared to *U-Rhythm* used in current practice. Nevertheless, with its improved diagnostic accuracy, *U-Rhythm* is associated with more cases being appropriately diagnosed and a slightly higher number of expected life-years and QALYs for the patient. Results were sensitive to changes in the diagnostic accuracy and costs of both confirmatory tests, with the headroom approach showing that the sensitivity and specificity of *U-Rhythm* should increase by ≈52%, while its cost should reduce by ≈30% in order to be more cost-effective in current practice. The VOI analysis suggested that the use of the device as a PA confirmatory test can be rejected with confidence, so no further research should be conducted until some of *U-Rhythm*'s non-stochastic parameters (i.e. cost, laboratory analytic methods) become more firmly established. Once this is done and more information on *U-Rhythm*'s use for other purposes becomes available, it would be worth re-running similar analysis to provide more robust estimates.

CHAPTER VII

CONCLUSION

DISCUSSION ON KEY FINDINGS AND LIMITATIONS, IMPLICATIONS FOR CLINICAL PRACTICE AND FURTHER RESEARCH, AND CONCLUSIONS

CHAPTER VII OVERVIEW

Chapters I-II provided the rationale, research question and analytical methods that would be used in this PhD project. **Chapters III-VI** presented the outcomes of the analyses that were conducted. **This last Chapter** discusses the key findings of all the analyses, their limitations, and the implications of their results for clinical practice and future research.

Discussion on Key Findings and Limitations, Implications for Clinical Practice and Further Research, and Conclusions

7.1 PhD Rationale, Research Question and Methods

Diagnosis is key to the determination of the optimal management and treatment of a disease. Diagnosis can be performed using several types of tests, from simple clinical examination and chemical/pathological measurements to more advanced procedures, scans and/or medical devices (MDs). The economic evaluation (EE) of diagnostic technologies and MDs has lately received increasing attention since many national and international authorities require their extensive examination before they receive marketing and reimbursement approval. However, evidence and guidelines on the best ways to conduct EE of these technologies have been limited and are not well-established. Furthermore, until recently, health technology assessment (HTA) has only been undertaken late in a product's development or just before its regulatory approval, when most clinical and economic data are available. However, EEs at earlier stages of development could help manufacturers to reduce development and production costs, and policy makers to plan how to allocate resources more efficiently.

The aim of this PhD project was to identify the methodologies that have been recommended and/or used for the early EE of diagnostics and MDs, and explore the challenges of applying them to a new diagnostic device that is currently at an early stage of development. To do so, an innovative, portable, minimally invasive, 24-hour hormone collection device (*U-Rhythm*), with a potential to be used in the diagnosis, management and monitoring of a range of endocrine disorders, was examined. In parallel, the device was being further developed and evaluated in a multi-centre (European) diagnostic accuracy case-control study (*ULTRADIAN*), which aimed to collect hormone samples from healthy volunteers and endocrine patients in order to compare physiological to pathophysiological hormone rhythms throughout a 24-hour period, and define the limits of normality as a first step towards the early and accurate diagnosis of the examined conditions.

To achieve the objectives of this PhD project, some research on the clinical area and existing tests, and the current European and United Kingdom (UK) regulations and guidelines for the market approval and HTA of diagnostic tests and MDs was initially conducted. A literature review was also performed to identify the methods that have been proposed and/or used for their clinical and early economic evaluation. Afterwards, a systematic review was undertaken to retrieve economic studies exploring and comparing the efficiency of different tests for the examined endocrine diseases. The *ULTRADIAN* participant demographic, clinical and sampling data together with their responses to three questionnaires asking about their experience with *U-Rhythm*, their recent healthcare usage and the impact of their health condition on their work productivity, and their current health-related quality of life (HRQoL) were then analysed and compared. Moreover, given that the device has shown technical capability and diagnostic promise for primary aldosteronism (PA), a costing study using UK healthcare data was performed to better understand the demographic and clinical characteristics of PA patients, measure their cost for the National Health Service (NHS), and identify the parameters that affect this. A cost-effectiveness analysis using decision-analytic modelling (DAM) techniques was also conducted (using the NHS perspective) to compare *U-Rhythm* to the most commonly used test for the diagnosis of PA (saline infusion test, SIT). Lastly, the headroom approach and a value of information (VOI) analysis were undertaken to find the maximum price that the device can have and be/become more cost-effective than SIT, and examine whether further research in the area would be worthwhile.

At the outset of this PhD research, the only elements that were predefined were the concept and design of *ULTRADIAN*. For this reason, the author used the HTA guidelines and evidence found on early EE methods for diagnostic tests and MDs (**Chapter I**) to select the analytical methods that he would later use to evaluate the cost-effectiveness of *U-Rhythm* (i.e. DAM, the headroom approach, VOI) (**Chapter VI**). The systematic review (**Chapter III**) gave an overview of the economic research that has been conducted in the area, indicating that more evidence is available for PA than the other health conditions. Here, two previous cost-effectiveness models were found that were used to conceptualise and develop the DAM used in **Chapter VI**, while all PA studies identified were used as sources for input data for the model. The *ULTRADIAN* study (**Chapter IV**) proved that *U-Rhythm* has technical capability and diagnostic promise for PA. This in combination with the findings of the systematic review led

the author to decide to examine the cost-effectiveness of the device in this disease (**Chapter VI**). Before doing so, the author used the information on different study designs (**Chapter II**) to design a retrospective cohort study that would give more information on the healthcare costs that are associated with PA (**Chapter V**). Although the data obtained from *ULTRADIAN* (apart from the diagnostic accuracy and cost of *U-Rhythm*) and the costing study were not ultimately used to populate the model as more appropriate data were available in the literature, they provided novel evidence on the impact of PA on healthcare costs and patient HRQoL. All analyses (**Chapters III-VI**) helped to understand the challenges of conducting an early EE of a new diagnostic medical device and inform on the implications for future research.

7.2 PhD Key Findings

The initial research indicated that diagnosing endocrine disorders with conventional methods is challenging, inexact, time-consuming and expensive due to their inability to effectively assess dynamic hormone release patterns in ambulatory settings, and the need for multiple testing and hospital visits (41-46). Moreover, although the European and UK regulations on the marketing approval of diagnostic tests and MDs are well-established (71, 73, 89, 92-96), the guidelines on the optimal methods for their EE (28, 77, 81-85, 87, 110-112) are still developing showing the difficulty in their evaluation and the fact that their efficiency depends on several other parameters (e.g. efficacy of the subsequent treatment; user's expertise). Furthermore, no optimal study design for their clinical evaluation was identified since cohort and case-control studies, and randomised controlled trials have different limitations when used for this purpose (12, 13, 16, 118, 121, 140). Five methods for early EE were found in the literature: a) DAM (14, 23-25, 29, 30, 33, 34, 427); b) the headroom approach (14, 23-25, 30, 34, 164, 165, 427); c) VOI analysis (Bayesian framework) (24, 25, 30, 32, 33, 79); d) clinical trial simulation (CTS) (25, 32, 33); and e) multi-criteria decision analysis (MCDA) (30, 31, 34, 88, 427). Some of them have only been tested in pharmaceuticals, while none of them is clearly superior and their selection mainly depends on the data that are available (i.e. stage of product development).

Only seven studies assessing and comparing the costs, health benefits and diagnostic accuracy of the existing diagnostic tests for three of the examined diseases (PA; Cushing's syndrome, CS; Addison's disease, AD) were found indicating the limited economic evidence in this clinical area (50, 51, 183-187). The studies compared a variety of tests/procedures illustrating the absence of a reference standard method for the diagnosis of these diseases. The *ULTRADIAN* data analysis showed the general satisfaction of the participants (healthy volunteers, PA, CS, AD) after using *U-Rhythm*, while differences were identified when estimating their healthcare use, the impact of their condition on their work productivity and their HRQoL. Specifically, healthcare use was much higher in PA and CS patients, while their HRQoL was much worse when compared to the other two groups. The costing study found that the total (ten years before and after its diagnosis) NHS costs for PA were £22,352.70 or £45,263.66 (without and with hospital admissions, respectively), more than double those for non-PA patients. Lastly, *U-Rhythm* was found to be less cost-effective than SIT when used as a confirmatory test (current practice), while it was the more efficient option when used as a confirmatory and subtyping test (theoretical practice), and compared against the upper NHS willingness-to-pay threshold (£30,000/quality-adjusted life-year). The headroom approach indicated that the current cost of the device would need to decrease by ≈30%, or its sensitivity and specificity would need to increase by ≈52% to become the more cost-effective diagnostic strategy, while the VOI analysis suggested that given its current parameters and ignoring its multiple uses in other endocrine diseases, no further research would be worthwhile.

7.3 PhD Strengths and Areas of Innovative Research

The main contribution of this PhD project to the wider clinical and health economic research is that it used the best available information and data to conduct a mixed-method early EE of a new diagnostic and monitoring device for rare endocrine disorders, which is currently at a (very) early stage of development. Specifically, it not only performed an EE for a combination of a diagnostic test and an MD, two already underdeveloped methodological areas of HTA, but also did that prior to its marketing authorisation, when data are limited, in a clinical area which is under-researched. To do so, the current European and UK regulations and guidelines for the marketing approval and HTA of both diagnostic technologies and MDs were described

in one place, probably for the first time. The proposed methods for the early EE of health technologies in general, and diagnostics and MDs more specifically, were also identified, reported and compared; something that, to the knowledge of the author, has only been done twice before to this extent (30, 79).

Three of these methods were then chosen and combined with two other methods that are widely used in health economic research: a) a systematic review to retrieve the available economic evidence in the clinical area and indicate the disease on which an early EE would be most feasible (given data availability); and b) a costing study on the chosen disease to produce new data on its diagnosis and treatment costs. Moreover, the project used data from a multi-national case-control study, probably the first one examining all these endocrine disorders together. This information was analysed to understand more about the patients' preferences regarding the design of the device and produce good-quality data on the patients' healthcare use and HRQoL. Using the limited available data from *ULTRADIAN*, *U-Rhythm*'s cost and diagnostic accuracy were estimated by the *ULTRADIAN* partners and were used in a DAM that compared the device to a test whose diagnostic accuracy is still uncertain despite being in use for decades. In this context, this PhD project applied the three early EE methods exploring their limitations and making suggestions on the data that are needed before deciding on whether it would be worthwhile to conduct further research and/or invest more on the new product. Throughout the analysis, the author aimed to use the best research practices to provide a complete and high-quality early economic study, the first for *U-Rhythm* and probably the first in this clinical area.

7.4 PhD Key Limitations

The first limitation of this PhD project is that it did not explore all the early EE methods that were found. However, this would have been difficult due to: a) the programme's time limits, the number of additional analyses (e.g. systematic review) and research-related activities that needed to take place (e.g. protocols and data application forms; data collection and cleaning), and because all these analyses needed to be done by only one researcher; b) the fact that CTS and MCDA have mainly been used in medications, with no clear guidelines for diagnostics,

and the time and challenges of carrying out tasks related to their performance. Moreover, a costing study was not conducted for all endocrine disorders, especially CS and AD for which the device has also shown some technical capability and diagnostic promise, while for PA, it did not have access to all the relevant hospital/medical data, meaning that costs would be underestimated. Additionally, non-medical, indirect and intangible costs were not considered in the analysis, making this study limited compared to a comprehensive cost-of-illness study. Nonetheless, a costing (or cost-of-illness) study for all diseases would have been infeasible for the reasons described above, while additional funding would have been needed to obtain more data.

Furthermore, the decision analysis used data from an early phase clinical trial with a case-control study design and limited recruitment numbers conducted in different countries (e.g. different organisation systems, databases, resources and areas of expertise; heterogeneity between participants). Obviously, these factors are expected to have had a small impact on the accuracy and precision of the data produced. However, to have a clinical study like this running in parallel was useful since any estimates in key parameters were based on real data and not expert opinion or assumptions, while the ease of incorporating new data (as they were being produced) in the analysis was also tested. Of course, one limitation was that the diagnostic accuracy of the device was measured only for PA patients versus healthy individuals and not between PA (unilateral or bilateral) and essential hypertensive patients, which would have been more reflective of actual practice. However, given that this was an early EE of the device and results were only aimed to explore its potential in the clinical area using current data and not to make final adoption/rejection decisions, this limited information was sufficient to conduct the analyses. Similarly, the fact that the decision analysis did not compare all tests for the diagnosis of PA and did not use data obtained from a systematic review and/or meta-analysis, and the fact that the VOI analysis ignored the other uses of *U-Rhythm* were necessary limitations that allowed the assessment of the potential of the device in one clinical area of promise.

Lastly, discrete choice experiments (DCEs) and monetary valuation methods (e.g. willingness-to-pay) would have been two useful tools to better understand the preferences of the users (i.e. patients, clinicians) on key aspects of *U-Rhythm* (e.g. design; size; weight; consumer

experience and satisfaction), and explore the trade-offs between these factors and other attributes of the device (e.g. ability to test in an ambulatory setting; diagnostic accuracy; ease and/or speed of diagnosis; reassurance provided by diagnosis; and health outcomes) (428). These methods can be used to incorporate health and non-health attributes in HTAs. For example, in a cost-consequence analysis, different programmes can be clearly and explicitly compared (e.g. in a tabular form) in terms of their costs, health benefits, and health and non-health attributes of health care (e.g. collected from DCEs), allowing decision-makers to decide how to weigh each element. Alternatively, if monetary valuation methods are used to value non-health attributes in monetary terms, then all elements can be summarised together in a cost-benefit analysis. In both cases, the efficiency of the examined alternatives would depend on the implicit and explicit (utility or monetary) values that decision-makers attach to the different elements of costs and outcomes (148, 429).

DCEs and monetary valuation methods can assist with decisions made beyond HRQoL and willingness-to-pay thresholds by providing additional information about implications for equity, need and other relevant objectives. However, their use in EE has been limited since there are questions about which factors should be valued, whose values should be used, and when they should be elicited (148, 429). Although, at the beginning, DCEs were considered as part of this PhD project, the plan was revised given the time and tasks that would be needed in order to organise and run an additional study. Another reason was the slow recruitment of patients in *ULTRADIAN*. Therefore, enrolling new patients to an additional study or putting more tasks to *ULTRADIAN* participants would have been challenging. Nevertheless, some of the preferences of the *ULTRADIAN* participants on *U-Rhythm* were elicited using the 'device satisfaction' questionnaire and could probably inform the design of a future DCE or willingness-to-pay study.

7.5 Comparisons with Related Health Economic Evidence

As indicated in **Chapter VI**, except for some variations in the study design, analytical methods and populations used, the main difference between this PhD project and the PA economic studies identified in the systematic literature review (**Chapter III**) is that here, an early EE of a non-established diagnostic test/strategy was conducted. Therefore, this project had different objectives and the author faced different challenges when performing the analysis (e.g. role of index test in clinical practice; limited/uncertain data; selection/application of analytical methods). After re-running the systematic review in September 2020, no new economic studies were found (for all examined endocrine disorders) demonstrating the slow pace that new economic evidence is produced in this area and the need for further investment in research to improve the diagnosis and treatment of these diseases, and reduce their associated costs for the payer. Based on the literature review on early EE methods (**Chapter II**) and several targeted searches performed in the area in September 2020, only one study was found to have used all three EE methods (applied in this project) for a new diagnostic device that was currently at an early phase of development (35). The study compared the new device to usual practice for monitoring heart failure patients with high risk of hyperkalaemia. The authors used the headroom approach before developing a DAM to quickly check the maximum reimbursement price of the MD using assumptions on its health benefits. VOI analysis was also used to examine the uncertainty in specific parameters and whether additional research would be worthwhile to reduce/remove this. The authors reported the challenge of dealing with limited data/research in the area and the importance of sensitivity analysis to partially address this. They also mentioned that there are factors (e.g. technology's learning effects; acceptance by the clinicians) that should be considered but are hard to quantify. Lastly, that early EE can provide useful advice on whether a new technology should be further developed but not on its final cost-effectiveness.

From the other early EE studies that were identified in literature (22, 36, 160, 430-450), most used a DAM, alone or in combination with the headroom approach or VOI analysis, to assess tests/MDs that were at various stages of development (from ideas to pre-market). Depending on the product's stage of development, data came from various sources (e.g. clinical trials, national sources, literature, expert opinion, assumptions). Three studies (36, 431, 445) also

used belief elicitation methods (e.g. questionnaires, interviews) to obtain experts' beliefs on important test parameters (e.g. diagnostic accuracy, cost). Interestingly, in these studies the examined products were at later stages of development showing that this method is probably better applied once the role/use of the technology is more established and more data are available. In general, all studies used some sort of sensitivity analysis (e.g. one- or two-way; probabilistic) to examine the robustness of their results. In addition, they were conducted in clinical areas where more research has been performed compared to this project (e.g. cancer, cardiovascular diseases). Lastly, most studies were undertaken in Europe, especially in the Netherlands and the UK, and after 2012, demonstrating the increasing interest in conducting early EE of diagnostics in the last decade.

7.6 Implications for Clinical Practice and Further Research

This PhD project indicated that the endocrine disorders examined, although rare within the general population, cost the healthcare system a significant amount of money due to them not being promptly and accurately diagnosed. The fact that a) patients may need to visit specialists and/or other doctors/nurses, be admitted to hospital, undergo various testing and change their treatment several times before and after being finally diagnosed; b) the choice of current testing is based on the patient's characteristics and compliance, its availability and/or the physician's preference; and c) the existing tests have different cut-off points depending on the clinical site illustrates the need for a new diagnostic/monitoring test that can dynamically assess the underlying causes of these diseases using acceptable and easily comparable cut-off values. A test like this would not only lead to a better diagnosis and lower healthcare costs but would also reduce the danger that is associated with a patient receiving inappropriate surgery or pharmacotherapy, and increase their life expectancy and HRQoL.

Nevertheless, for policy makers, clinicians and other experts, and patients to be convinced that *U-Rhythm* is a better diagnostic/monitoring method than the existing alternatives, it is essential that its use is tested more thoroughly in further studies, which would recruit a greater number of participants, patient groups and sites, and would compare the device to a range of commonly used tests for each disease (e.g. SIT for PA). This would either require

funding from a national or international research agency (e.g. the European Committee; the UK National Institute for Health Research) or from a device manufacturer that aims to invest in and develop the product further before releasing it to the market. Given its current cost, *U-Rhythm* is more likely to be initially used as a confirmatory, subtype and/or monitoring test than a screening method. As mentioned in **Chapter VI**, in the long run, the costs of the device, sampling and hormone assays are expected to decrease due to economies of scale in manufacture. Nevertheless, the costs of sampling and analytics are likely to predominate initially. Whilst there may be some ways to streamline these processes, current estimates suggest that *U-Rhythm* will be more costly than some confirmatory laboratory (e.g. SIT) and imaging (e.g. computerised tomography) tests but less expensive than others (e.g. adrenal vein sampling). To be competitive, *U-Rhythm* will need to demonstrate advantages over other similarly priced confirmatory tests or to preclude the need for more expensive and invasive additional tests. These potential advantages include greater diagnostic accuracy leading to faster diagnosis and treatment, fewer inconclusive or conflicting confirmatory test results, and increased patient satisfaction with the device.

Here, it is important to mention that current research on *U-Rhythm* (outside *ULTRADIAN*) has shown that the device can have multiple other uses apart from diagnosing/monitoring the endocrine disorders examined in this project (https://www.u-rhythm.co.uk/). More precisely, the *Peacock* paediatric cortisol study (http://www.isrctn.com/ISRCTN98258655), a study funded by the *British Heart Foundation*, ran in parallel to *ULTRADIAN* and aimed to investigate how a child's body responds (stress) to cardiac surgery and cardiac catheterisation (patients aged 0-5 or 10-16 years). The *bioRHYTHM* (https://biorhythm.blogs.bristol.ac.uk/) study used data collected from *U-Rhythm* and other wearable devices, while individuals continue with their daily activities (including sleep), to measure rhythms in life. At the time that this thesis was written, an award from the *Medical Research Council Confidence in Concept* scheme was also given for investigating thyroid hormone levels in patients with thyroid disease (using *U-Rhythm*). Further grant applications related to the use of the device in monitoring patients with diabetes, sleep disorders and cardiovascular diseases were also in process. Additionally, during meetings with external experts as part of *U-Rhythm*'s commercialisation strategy, some interest in using the device for investigating CS in dogs and horses had been expressed, while tracking circadian rhythms in reindeers had also been conceived as an idea for a future

study. The wide range of potential applications demonstrates the challenge facing developers in predicting the eventual impact and value of an MD, such as *U-Rhythm*.

Apart from the value of *U-Rhythm* in endocrine disease, this project also explored the benefit of conducing EE early at a product's development for manufacturers, policy makers, clinicians and patients. As indicated, performing an early EE, even in a highly under-researched area, is a feasible and inexpensive task that can provide useful information on the potential role of a product in current/future practice and the worthiness of investing in it instead of other technologies. This information can help manufacturers to understand the market and focus on the areas of the device that need further development; policy makers to make decisions considering the potential of new technologies being launched in the market soon; and clinicians and patients to be more informed and prepared when new technologies are released. Nevertheless, it was also demonstrated that there are several challenges that need to be addressed (e.g. incomplete and evolving information on key parameters), which might mean that the early EE may have reached the wrong conclusion. Regarding diagnostics and MDs, this project illustrated that an (early) EE is more challenging and data demanding than that for pharmaceutics, but it is feasible if their use in the diagnostic-treatment pathway, the effectiveness of the subsequent treatment and their comparators are explicit. Therefore, an early EE of diagnostic technologies and MDs using at least the three methods described in this thesis should be conducted in order for stakeholders to have a clear guide when making go/no-go decisions on further investment in research and development.

7.7 Conclusion

In conclusion, this PhD project aimed to explore the use of early EE, especially for diagnostic MDs, demonstrating some of the potential benefits and challenges of current methods. This is an area of EE that is likely to grow in the future to meet the needs of technology developers and regulators. Because as *Benjamin Franklin*, one of the Founding Fathers of the United States, once said:

"By failing to prepare, you are preparing to fail"

References

1. Fischer MA, Fabing H, Marr R. Fischerisms. Baltimore, Md: Springfield, Ill., C.C. Thomas; 1944.

2. Rakel RE, Rakel D. Textbook of family medicine. 8th ed. ed. Philadelphia: Elsevier/Saunders; 2011.

3. Kongstvedt PR. The managed health care handbook. 4th ed. ed. Gaithersburg, Md.: Aspen Publishers; 2000.

4. Rakel RE. Diagnosis. November 23, 2018 [cited February 14, 2019]. In: Encyclopædia Britannica [Internet]. Encyclopædia Britannica, inc., [cited February 14, 2019]. Available from: https://www.britannica.com/science/diagnosis.

5. Berger D. A brief history of medical diagnosis and the birth of the clinical laboratory. Part 1--Ancient times through the 19th century. MLO Med Lab Obs. 1999;31(7):28-30, 2, 4-40.

6. Berger D. A brief history of medical diagnosis and the birth of the clinical laboratory. Part 2--Laboratory science and professional certification in the 20th century. MLO Med Lab Obs. 1999;31(8):32-4, 6, 8.

7. Engle RL, Jr., Davis BJ. Medical Diagnosis: Present, Past, and Future. I. Present Concepts of the Meaning and Limitations of Medical Diagnosis. Arch Intern Med. 1963;112:512-9.

8. Engle RL, Jr. Medical Diagnosis: Present, Past, and Future. Ii. Philosophical Foundations and Historical Development of Our Concepts of Health, Disease, and Diagnosis. Arch Intern Med. 1963;112:520-9.

9. Engle RL, Jr. Medical Diagnosis: Present, Past, and Future. Iii. Diagnosis in the Future, Including a Critique on the Use of Electronic Computers as Diagnostic Aids to the Physician. Arch Intern Med. 1963;112:530-43.

10. Walker HK. The Origins of the History and Physical Examination. In: rd, Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. Boston1990.

11. Moore RE. An Historical Perspective on the Clinical Diagnostic Laboratory. 2006 [cited February 14, 2019]. In: Molecular Diagnostics [Internet]. Humana Press, [cited February 14, 2019]. Available from: https://doi.org/10.1385/1-59259-928-1:003.

12. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ. 1986;134(6):587-94.

13. Rodger M, Ramsay T, Fergusson D. Diagnostic randomized controlled trials: the final frontier. Trials. 2012;13:137.

14. Buisman LR, Rutten-van Molken MP, Postmus D, Luime JJ, Uyl-de Groot CA, Redekop WK. The Early Bird Catches the Worm: Early Cost-Effectiveness Analysis of New Medical Tests. Int J Technol Assess Health Care. 2016;32(1-2):46-53.

15. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making. 1991;11(2):88-94.

16. Van den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J Clin Epidemiol. 2007;60(11):1116-22.

17. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the economic evaluation of health care programmes. 3rd ed. ed. Oxford: Oxford University Press; 2005.

18. Jefferson T, Demicheli V, Mugford M. Elementary economic evaluation in health care. 2nd ed. ed. London: BMJ books; 2000.

19. Kielhorn A, Graf von der Schulenburg J-M. The Health Economics Handbook. Second ed. England: Adis International Limited; 2000. 215 p.

20. Paul JE, Trueman P. 'Fourth hurdle reviews', NICE, and database applications. Pharmacoepidemiol Drug Saf. 2001;10(5):429-38.

21. Rawlins MD. Crossing the fourth hurdle. Br J Clin Pharmacol. 2012;73(6):855-60.

22. Buisman LR, Luime JJ, Oppe M, Hazes JM, Rutten-van Molken MP. A five-year model to assess the early cost-effectiveness of new diagnostic tests in the early diagnosis of rheumatoid arthritis. Arthritis Res Ther. 2016;18(1):135.

23. Frempong SN, Sutton AJ, Davenport C, Barton P. Economic evaluation of medical tests at the early phases of development: a systematic review of empirical studies. Expert Rev Pharmacoecon Outcomes Res. 2018;18(1):13-23.

24. Steuten LM, Ramsey SD. Improving early cycle economic evaluation of diagnostic technologies. Expert Rev Pharmacoecon Outcomes Res. 2014;14(4):491-8.

25. Vallejo-Torres L, Steuten LM, Buxton MJ, Girling AJ, Lilford RJ, Young T. Integrating health economics modeling in the product development cycle of medical devices: a Bayesian approach. Int J Technol Assess Health Care. 2008;24(4):459-64.

26. Soares MO, Walker S, Palmer SJ, Sculpher MJ. Establishing the Value of Diagnostic and Prognostic Tests in Health Technology Assessment. Med Decis Making. 2018;38(4):495-508.

27. Busby J, Schroeder K, Woltersdorf W, Sterne JA, Ben-Shlomo Y, Hay A, et al. Temporal growth and geographic variation in the use of laboratory tests by NHS general practices: using routine data to identify research priorities. Br J Gen Pract. 2013;63(609):e256-66.

28. National Institute for Health and Clinical Excellence. Diagnostics Assessment Programme Manual. NICE Process and Methods Guides. London2011.

29. Sculpher M, Drummond M, Buxton M. The iterative use of economic evaluation as part of the process of health technology assessment. J Health Serv Res Policy. 1997;2(1):26-30.

30. Redekop K, Mikudina B. Early medical technology assessments of medical devices and tests. JHPOR. 2013(1):26-37.

31. Miquel-Cases A, Schouten PC, Steuten LM, Retel VP, Linn SC, van Harten WH. (Very) Early technology assessment and translation of predictive biomarkers in breast cancer. Cancer Treat Rev. 2017;52:117-27.

32. Hartz S, John J. Contribution of economic evaluation to decision making in early phases of product development: a methodological and empirical review. Int J Technol Assess Health Care. 2008;24(4):465-72.

33. Miller P. Role of pharmacoeconomic analysis in R&D decision making: when, where, how? PharmacoEcon. 2005;23(1):1-12.

34. IJzerman MJ, Koffijberg H, Fenwick E, Krahn M. Emerging Use of Early Health Technology Assessment in Medical Product Development: A Scoping Review of the Literature. PharmacoEcon. 2017;35(7):727-40.

35. van de Wetering G, Steuten LMG, von Birgelen C, Adang EMM, IJerman MJ. Early Bayesian modeling of a potassium lab-on-a-chip for monitoring of heart failure patients at increased risk of hyperkalaemia. Technol Forecast Soc Change. 2012;79(7):1268-79.

36. Khoudigian-Sinani S, Blackhouse G, Levine M, Thabane L, O'Reilly D. The premarket assessment of the cost-effectiveness of a predictive technology "Straticyte" for the early detection of oral cancer: a decision analytic model. Health Econ Rev. 2017;7(1):35.

37. Golden SH, Robinson KA, Saldanha I, Anton B, Ladenson PW. Prevalence and Incidence of Endocrine and Metabolic Disorders in the United States: A Comprehensive Review. J Clin Endocrinol Metab. 2009;94(6):1853-78.

38. Lightman S, Terry JR. The importance of dynamic signalling for endocrine regulation and drug development: relevance for glucocorticoid hormones. Lancet Diabetes Endocrinol. 2014;2(7):593-9.

39. Lightman SL, Conway-Campbell BL. The crucial role of pulsatile activity of the HPA axis for continuous dynamic equilibration. Nat Rev Neurosci. 2010;11(10):710-8.

40. Chrousos GP, Gold PW. The concepts of stress and stress system disorders. Overview of physical and behavioral homeostasis. JAMA. 1992;267(9):1244-52.

41. Katznelson L, Laws ER, Melmed S, Molitch ME, Murad MH, Utz A, et al. Acromegaly: An endocrine society clinical practice guideline. J Clin Endocrinol Metab. 2014;99(11):3933-51.

42. Bornstein SR, Allolio B, Arlt W, Barthel A, Don-Wauchope A, Hammer GD, et al. Diagnosis and Treatment of Primary Adrenal Insufficiency: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab. 2016;101(2):364-89.

43. Speiser PW, Azziz R, Baskin LS, Ghizzoni L, Hensle TW, Merke DP, et al. Congenital adrenal hyperplasia due to steroid 21-hydroxylase deficiency: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab. 2010;95(9):4133-60.

44. Nieman LK, Biller BM, Findling JW, Newell-Price J, Savage MO, Stewart PM, et al. The diagnosis of Cushing's syndrome: an Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab. 2008;93(5):1526-40.

45. Molitch ME, Clemmons DR, Malozowski S, Merriam GR, Vance ML. Evaluation and treatment of adult growth hormone deficiency: An endocrine society clinical practice guideline. J Clin Endocrinol Metab. 2011;96(6):1587-609.

46. Funder JW, Carey RM, Mantero F, Murad MH, Reincke M, Shibata H, et al. The Management of Primary Aldosteronism: Case Detection, Diagnosis, and Treatment: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab. 2016;101(5):1889-916.

47. Schlaff WD. Dynamic testing in reproductive endocrinology. Fertil Steril. 1986;45(5):589-606.

48. Kirkland J, Saenger P, MacGillivray M, LaFranchi S, Rosenfield R. Physician and clinic charges for diagnosing growth hormone deficiency. J Pediatr. 1996;128(5 Pt 2):S61-2.

49. Knutzen R, Ezzat S. The cost of medical care for the acromegalic patient. Neuroendocrinology. 2006;83(3-4):139-44.

50. Leon-Justel A, Mangas MA, Infante Fontan R, Castro Luque J, Venegas Moreno E, Madrazo Atutxa A, et al. Budget impact of using midnight salivary cortisol in the diagnosis of hypercortisolism. Clin Chim Acta. 2011;412(23-24):2248-53.

51. Midgette AS, Aron DC. High-dose dexamethasone suppression testing versus inferior petrosal sinus sampling in the differential diagnosis of adrenocorticotropin-dependent Cushing's syndrome: a decision analysis. Am J Med Sci. 1995;309(3):162-70.

52. Puig J, Wagner A, Caballero A, Rodriguez-Espinosa J, Webb SM. Cost-effectiveness and accuracy of the tests used in the differential diagnosis of Cushing's syndrome. Pituitary. 1999;1(2):125-32.

53. Conn JW. Presidential address. I. Painting background. II. Primary aldosteronism, a new clinical syndrome. J Lab Clin Med. 1955;45(1):3-17.

54. Funder JW. Primary Aldosteronism: Seismic Shifts. J Clin Endocrinol Metab. 2015;100(8):2853-5.

55. Funder JW. Primary aldosteronism as a public health issue. Lancet Diabetes Endo. 2016;4(12):972-3.

56. Conn JW, Cohen EL, Rovner DR, Nesbit RM. Normokalemic Primary Aldosteronism. A Detectable Cause of Curable "Essential" Hypertension. JAMA. 1965;193:200-6.

57. Rossi GP. A comprehensive review of the clinical aspects of primary aldosteronism. Nat Rev Endocrinol. 2011;7(8):485-95.

58. Schirpenbach C, Reincke M. Primary aldosteronism: current knowledge and controversies in Conn's syndrome. Nat Clin Pract Endocrinol Metab. 2007;3(3):220-7.

59. Cushing H. The basophil adenomas of the pituitary body and their clinical manifestations (pituitary basophilism). 1932. Obes Res. 1994;2(5):486-508.

60. Lindholm J, Juul S, Jorgensen JO, Astrup J, Bjerre P, Feldt-Rasmussen U, et al. Incidence and late prognosis of cushing's syndrome: a population-based study. J Clin Endocrinol Metab. 2001;86(1):117-23.

61. Sharma ST, Nieman LK, Feelders RA. Cushing's syndrome: epidemiology and developments in disease management. Clin Epidemiol. 2015;7:281-93.

62. Newell-Price J, Bertagna X, Grossman AB, Nieman LK. Cushing’s syndrome. The Lancet. 2006;367:1605-17.

63. Nieman LK, Ilias I. Evaluation and treatment of Cushing's syndrome. Am J Med. 2005;118(12):1340-6.

64. Fernandez-Rodriguez E, Stewart PM, Cooper MS. The pituitary-adrenal axis and body composition. Pituitary. 2009;12(2):105-15.

65. Burton C, Cottrell E, Edwards J. Addison's disease: identification and management in primary care. Br J Gen Pract. 2015;65(638):488-90.

66. Vaidya B, Chakera AJ, Dick C. Addison's disease. BMJ. 2009;339:b2385.

67. Willis AC, Vince FP. The prevalence of Addison's disease in Coventry, UK. Postgrad Med J. 1997;73(859):286-8.

68. Chakera AJ, Vaidya B. Addison disease in adults: diagnosis and management. Am J Med. 2010;123(5):409-13.

69. Wallace I, Cunningham S, Lindsay J. The diagnosis and investigation of adrenal insufficiency in adults. Ann Clin Biochem. 2009;46(Pt 5):351-67.

70. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Adrenal Insufficiency and Addison's Disease. National Institutes of Health (NIH) Publication. 2014;No. 14–3054:1-16.

71. Council of the European Communities. Council Directive 93/42/EEC of 14 June 1993 concerning medical devices. In: European Union, editor. Official Journal of the European Communities, 1993. p. 1-60.

72. European Parliament and Council of the European Union. Directive 2007/47/EC of the European Parliament and of the Council of 5 September 2007 amending Council Directive 90/385/EEC on the approximation of the laws of the Member States relating to active implantable medical devices, Council Directive 93/42/EEC concerning medical devices and Directive 98/8/EC concerning the placing of biocidal products on the market. In: European Union, editor. Official Journal of the European Union, 2007. p. L 247/21-55.

73. European Parliament and Council of the European Union. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. In: European Union, editor. Official Journal of the European Union, 2017. p. L 117/1-75.

74. Clarkson DM. Medical Device Guidebook: A browser information resource for medical device users. Med Eng Phys. 2017;41:97-102.

75. Geremia F. Quality aspects for medical devices, quality system and certification process. Microchem J. 2018;136:300-6.

76. Taylor RS, Iglesias CP. Assessing the clinical and cost-effectiveness of medical devices and drugs: are they that different? Value Health. 2009;12(4):404-6.

77. Cookson R, Hutton J. Regulating the economic evaluation of pharmaceuticals and medical devices: a European perspective. Health Policy. 2003;63(2):167-78.

78. Henschke C, Panteli D, Perleth M, Busse R. Taxonomy of Medical Devices in the Logic of Health Technology Assessment. Int J Technol Assess Health Care. 2015;31(5):324-30.

79. Craig JA, Carr L, Hutton J, Glanville J, Iglesias CP, Sims AJ. A review of the economic tools for assessing new medical devices. Appl Health Econ Health Policy. 2015;13(1):15-27.

80. Drummond M, Griffin A, Tarricone R. Economic evaluation for devices and drugs--same or different? Value Health. 2009;12(4):402-4.

81. Schnell-Inderst P, Mayer J, Lauterberg J, Hunger T, Arvandi M, Conrads-Frank A, et al. Health technology assessment of medical devices: What is different? An overview of three European projects. Z Evid Fortbild Qual Gesundhwes. 2015;109(4-5):309-18.

82. Tarricone R, Torbica A, Drummond M. Challenges in the Assessment of Medical Devices: The MedtecHTA Project. Health Econ. 2017;26 Suppl 1:5-12.

83. Tarricone R, Callea G, Ogorevc M, Prevolnik Rupel V. Improving the Methods for the Economic Evaluation of Medical Devices. Health Econ. 2017;26 Suppl 1:70-92.

84. Tarricone R, Torbica A, Drummond M, Medtec HTAPG. Key Recommendations from the MedtecHTA Project. Health Econ. 2017;26 Suppl 1:145-52.

85. Ciani O, Wilcher B, van Giessen A, Taylor RS. Linking the Regulatory and Reimbursement Processes for Medical Devices: The Need for Integrated Assessments. Health Econ. 2017;26 Suppl 1:13-29.

86. Rothery C, Claxton K, Palmer S, Epstein D, Tarricone R, Sculpher M. Characterising Uncertainty in the Assessment of Medical Devices and Determining Future Research Needs. Health Econ. 2017;26 Suppl 1:109-23.

87. National Institute for Health and Clinical Excellence (NICE). Medical Technologies Evaluation Programme, Methods guide. https://www.nice.org.uk/process/pmg33 April 2011.

88. Pecchia L, Craven MP, editors. Early stage Health Technology Assessment (HTA) of biomedical devices. The MATCH experience2013; Berlin, Heidelberg: Springer Berlin Heidelberg.

89. Medical Device Certification (MDC) GmbH. Basic Information about the European Directive 93/42/EEC on Medical Devices. https://www.mdc-ce.de/ October 2009.

90. European Parliament and Council of the European Union. Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on in vitro diagnostic medical devices. In: European Union, editor. Official Journal of the European Communities, 1998. p. L 331/1-37.

91. Council of the European Communities. Council Directive of 20 June 1990 on the approximation of the laws of the Member States relating to active implantable medical devices (90/385/EEC). In: European Union, editor. Official Journal of the European Communities, 1990. p. No L 189/17-36.

92. French-Mowat E, Burnett J. How are medical devices regulated in the European Union? J R Soc Med. 2012;105 Suppl 1:S22-8.

93. National Patient Safety Agency, Service NRE, National Health Service (NHS). Approval for medical devices research. Guidance for researchers, manufacturers, research ethics committees and NHS R&D offices. https://www.nihcollaboratory.org/ March 2008.

94. Medicines & Healthcare products Regulatory Agency (MHRA). Clinical investigations of medical devices - guidance for manufacturers. https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency March 2016.

95. Medicines & Healthcare products Regulatory Agency (MHRA). Medicines & Medical Devices Regulation: What you need to know. https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency April 2008.

96. UK Parliament, Secretatry of State. Statutory Instrument 2002 No. 618, Consumer Protection, The Medical Devices Regulations 2002. p. 1-40.

97. EU Health Technology Assessment Network. Strategy for EU Cooperation on Health Technology Assessment. https://ec.europa.eu/health/technology_assessment/policy/network_en October 2014.

98. Velasco Garrido M, Borlum Kristensen F, Palmhoj Nielsen C, Busse R. Health Technology Assessment and Health Policy-Making in Europe. Current status, challenges and potential. http://www.euro.who.int/2008.

99. Steuten LM. Early Stage Health Technology Assessment for Precision Biomarkers in Oral Health and Systems Medicine. OMICS. 2016;20(1):30-5.

100. Drummond MF. Economic appraisal of health technology in the European Community. Oxford: Oxford University Press; 1987.

101. Hartz S, John J. Public health policy decisions on medical innovations: what role can early economic evaluation play? Health Policy. 2009;89(2):184-92.

102. Nestler-Parr S, Korchagina D, Toumi M, Pashos CL, Blanchette C, Molsen E, et al. Challenges in Research and Health Technology Assessment of Rare Disease Technologies: Report of the ISPOR Rare Disease Special Interest Group. Value Health. 2018;21(5):493-500.

103. European Parliament and Council of the European Union. Directive 2011/24/EU of the European Parliament and of the Council of 9 March 2011 on the application of patients' rights in cross-border healthcare. In: European Union, editor. Official Journal of the European Union, 2011. p. L 88/45-65.

104. European Network for Health Technology Assessment (EUnetHTA). Methods for health economic evaluations - A guideline based on current practices in Europe. https://eunethta.eu/methodology-guidelines/ May 2015.

105. European Network for Health Technology Assessment (EUnetHTA). EUnetHTA JA2 WP8 Deliverable. HTA Core Model Version 3.0 for the full assessment of Diagnostic Technologies, Medical and Surgical Interventions, Pharmaceuticals and Screening Technologies. https://www.eunethta.eu/ January 2016.

106. World Health Organization (WHO). Health technology assessment of medical devices. https://www.who.int/2011.

107. Kobelt G. Health economics : an introduction to economic evaluation. 2nd ed. ed: Association of the British Pharmaceutical Industry, Office of Health Economics; 2002.

108. Drummond M, Banta D. Health technology assessment in the United Kingdom. Int J Technol Assess Health Care. 2009;25 Suppl 1:178-81.

109. National Institute for Health and Care Excellence (NICE). Guide to the processes of technology appraisal 2014. Available from: http://nice.org.uk/process/pmg9.

110. Keltie K, Bousfield DR, Cole H, Sims AJ. Medical Technologies Evaluation Programme: A review of NICE progression decisions, 2010-2013. Health Policy Techn. 2016;5(3):243-50.

111. Sprange K, Clift M. The NICE Medical Technologies Evaluation Programme (MTEP): manufacturer submission challenges. J R Soc Med. 2012;105 Suppl 1:S4-11.

112. National Institute for Health and Clinical Excellence (NICE). Medical Technologies Evaluation Programme, Process guide. https://www.nice.org.uk/process/pmg33 April 2011.

113. National Institute for Health and Care Excellence (NICE). Medical Technologies Evaluation Programme, Sponsor submission of evidence form. www.nice.org.uk/mtep March 2013.

114. Brockis E, Marsden G, Cole A, Devlin N. A Review of NICE Methods Across Health Technology Assessment Programmes: Differences, Justifications and Implications. https://www.ohe.org/publications/review-nice-methods-across-health-technology-assessment-programmes-differences; April 2016.

115. National Institute for Health and Clinical Excellence (NICE). Diagnostics Assessment Programme manual. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-diagnostics-guidance December 2011.

116. Newland A. NICE diagnostics assessment programme. Ann R Coll Surg Engl. 2011;93(5):412-3.

117. National Institute for Health and Care Excellence (NICE). Guide to the methods of technology appraisal 2013. Available from: http://nice.org.uk/process/pmg9.

118. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. 2006;144(11):850-5.

119. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332(7549):1089-92.

120. Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem. 2012;58(12):1636-43.

121. Deeks JJ. Using evaluations of diagnostic tests: understanding their limitations and making the most of available evidence. Ann Oncol. 1999;10(7):761-8.

122. Simundic AM. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC. 2009;19(4):203-11.

123. Eusebi P. Diagnostic accuracy measures. Cerebrovasc Dis. 2013;36(4):267-72.

124. Mandrekar JN. Simple statistical measures for diagnostic accuracy assessment. J Thorac Oncol. 2010;5(6):763-4.

125. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. BMJ. 2002;324(7335):477-80.

126. Brodersen J, Kramer BS, Macdonald H, Schwartz LM, Woloshin S. Focusing on overdiagnosis as a driver of too much medicine. BMJ. 2018;362:k3494.

127. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. J Clin Epidemiol. 2010;63(8):883-91.

128. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ. 2012;344:e686.

129. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Making. 2009;29(5):E22-9.

130. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-5.

131. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. 2002;359(9300):57-61.

132. Schaafsma JD, van der Graaf Y, Rinkel GJ, Buskens E. Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. J Clin Epidemiol. 2009;62(12):1248-52.

133. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.

134. Slack MK, Draugalis JR. Establishing the internal and external validity of experimental studies. Am J Health Syst Pharm. 2001;58(22):2173-81; quiz 82-3.

135. Torre DM, Picho K. Threats to Internal and External Validity in Health Professions Education Research. Acad Med. 2016;91(12):e21.

136. Sculpher MJ, Claxton K, Drummond M, McCabe C. Whither trial-based economic evaluation for health care decision making? Health Econ. 2006;15(7):677-87.

137. Briggs AH, Claxton K, Sculpher MJ. Decision modelling for health economic evaluation. Oxford: Oxford University Press; 2006.

138. Edlin R, McCabe C, Hulme C, Hall P, Wright J. Cost effectiveness modelling for health technology assessment : a practical course.

139. Caro JJ, Briggs AH, Siebert U, Kuntz KM, Force I-SMGRPT. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--1. Value Health. 2012;15(6):796-803.

140. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51(8):1335-41.

141. Mooney GH, Russell EM, Weir RD. Choices for health care : a practical introduction to the economics of health provision. 2nd ed. ed: Macmillan; 1986.

142. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. N Engl J Med. 1977;296(13):716-21.

143. Palmer S, Torgerson DJ. Economic notes: definitions of efficiency. BMJ. 1999;318(7191):1136.

144. Culyer AJ. The normative economics of health care finance and provision. Oxford Rev Econ Pol. 1989;5(1):34-58.

145. Morris S, Devlin NJ, Parkin D. Economic analysis in health care. Chichester: John Wiley; 2007.

146. McGuire A, Fenn P, Mayhew K. Providing health care : the economics of alternative systems of finance and delivery. Oxford: Oxford University Press; 1991.

147. Brouwer WB, Koopmanschap MA. On the economic foundations of CEA. Ladies and gentlemen, take your positions! J Health Econ. 2000;19(4):439-59.

148. Coast J. Is economic evaluation in touch with society's health values? BMJ. 2004;329(7476):1233-6.

149. Brouwer WB, Culyer AJ, van Exel NJ, Rutten FF. Welfarism vs. extra-welfarism. J Health Econ. 2008;27(2):325-38.

150. Coast J, Smith RD, Lorgelly P. Welfarism, extra-welfarism and capability: the spread of ideas in health economics. Soc Sci Med. 2008;67(7):1190-8.

151. Buchanan J, Wordsworth S. Welfarism versus extra-welfarism: can the choice of economic evaluation approach impact on the adoption decisions recommended by economic evaluation studies? PharmacoEcon. 2015;33(6):571-9.

152. Lancsar E, Savage E. Deriving welfare measures from discrete choice experiments: inconsistency between current methods and random utility and welfare theory. Health Econ. 2004;13(9):901-7.

153. Fakhri MAB, Hanafiah Juni M, Rosliza AM, Faisal I. Societal perspective in economic evaluation. International Journal of Public Health and Clinical Sciences. 2017;4(4):41-50.

154. York Health Economics Consortium. Perspective 2016. Available from: https://yhec.co.uk/glossary/perspective/.

155. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. Br Med Bull. 2010;96:5-21.

156. Prieto L, Sacristan JA. Problems and solutions in calculating quality-adjusted life years (QALYs). Health Qual Life Outcomes. 2003;1:80.

157. Flynn TN. Using Conjoint Analysis and Choice Experiments to Estimate QALY Values Issues to Consider. PharmacoEcon. 2010;28(9):711-22.

158. Sassi F, McKee M, Roberts JA. Economic evaluation of diagnostic technology. Methodological challenges and viable solutions. Int J Technol Assess Health Care. 1997;13(4):613-30.

159. Markiewicz K, van Til JA, MJ IJ. Medical devices early assessment methods: systematic literature review. Int J Technol Assess Health Care. 2014;30(2):137-46.

160. Retel VP, Grutters JP, van Harten WH, Joore MA. Value of research and value of development in early assessments of new medical technologies. Value Health. 2013;16(5):720-8.

161. Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criteria Decision Analysis for Health Care Decision Making--An Introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. Value Health. 2016;19(1):1-13.

162. Oliveira MD, Mataloto I, Kanavos P. Multi-criteria decision analysis for health technology assessment: addressing methodological challenges to improve the state of the art. Eur J Health Econ. 2019;20(6):891-918.

163. Hummel M, Steuten L, Groothuis-Oudshoorn K, Ijzerman M. Using the analytic hierarchy process to filling missing gaps in early health economic modeling. Meeting of the Lowlands Health Economics Study Group 2011; The Netherlands.

164. Girling A, Young T, Brown C, Lilford R. Early-stage valuation of medical devices: the role of developmental uncertainty. Value Health. 2010;13(5):585-91.

165. Girling A, Lilford R, Cole A, Young T. Headroom Approach to Device Development: Current and Future Directions. Int J Technol Assess Health Care. 2015;31(5):331-8.

166. Chen MH, Willan AR. Value of information methods for assessing a new diagnostic test. Stat Med. 2014;33(11):1801-15.

167. Eckermann S, Willan AR. Expected value of information and decision making in HTA. Health Econ. 2007;16(2):195-209.

168. Grutters JPC, Abrams KR, de Ruysscher D, Pijls-Johannesma M, Peters HJM, Beutner E, et al. When to Wait for More Evidence? Real Options Analysis in Proton Therapy. Oncologist. 2011;16(12):1752-61.

169. Broder MS, Neary MP, Chang E, Cherepanov D, Ludlam WH. Burden of illness, annual healthcare utilization, and costs associated with commercially insured patients with Cushing disease in the United States. Endocr Pract. 2015;21(1):77-86.

170. Didoni G, Grottol S, Gasco V, Battistini M, Ferone D, Giusti M, et al. Cost-of-illness study in acromegalic patients in Italy. J Endocrinol Invest. 2004;27(11):1034-9.

171. Ehrnborg C, Hakkaart-Van Roijen L, Jonsson B, Rutten FF, Bengtsson BA, Rosen T. Cost of illness in adult patients with hypopituitarism. PharmacoEcon. 2000;17(6):621-8.

172. Placzek H, Xu Y, Mu Y, Begelman SM, Fisher M. Clinical and Economic Burden of Commercially Insured Patients with Acromegaly in the United States: A Retrospective Analysis. J Manag Care Spec Pharm. 2015;21(12):1106-12.

173. Roset M, Merino-Montero S, Luque-Ramirez M, Webb SM, Lopez-Mondejar P, Salinas I, et al. Cost of clinical management of acromegaly in Spain. Clin Drug Investig. 2012;32(4):235-45.

174. Swearingen B, Wu N, Chen SY, Pulgar S, Biller BMK. Health care resource use and costs among patients with cushing disease. Endocr Pract. 2011;17(5):681-90.

175. Wilson LS, Shin JL, Ezzat S. Longitudinal assessment of economic burden and clinical outcomes in acromegaly. Endocr Pract. 2001;7(3):170-80.

176. Centre for Reviews and Dissemination, University of York. Systematic Reviews: CRD's guidance for undertaking reviews in health care. York: York Publishing Services Ltd; 2009. 1-294 p.

177. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. In: Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, editors. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. AHRQ Methods for Effective Health Care. Rockville (MD)2013.

178. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. BMJ. 2013;346:f1049.

179. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)-Explanation and Elaboration: A Report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Health. 2013;16(2):231-50.

180. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. PharmacoEcon. 2006;24(4):355-71.

181. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess. 2004;8(36):iii-iv, ix-xi, 1-158.

182. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

183. Ben-Shlomo A, Guzman J, Mirocha J. Enhanced cosyntropin stimulation test performance enabled by electronic medical record. Pituitary. 2016;19(5):503-6.

184. Dekkers T, Prejbisz A, Kool LJ, Groenewoud HJ, Velema M, Spiering W, et al. Adrenal vein sampling versus CT scan to determine treatment in primary aldosteronism: an outcome-based randomised diagnostic trial. The Lancet Diabetes & Endocrinology. 2016;4(9):739-46.

185. Lubitz CC, Economopoulos KP, Sy S, Johanson C, Kunzel HE, Reincke M, et al. Cost-Effectiveness of Screening for Primary Aldosteronism and Subtype Diagnosis in the Resistant Hypertensive Patients. Circulation Cardiovascular Quality & Outcomes. 2015;8(6):621-30.

186. Sato M, Morimoto R, Seiji K, Iwakura Y, Ono Y, Kudo M, et al. Cost-Effectiveness Analysis of the Diagnosis and Treatment of Primary Aldosteronism in Japan. Horm Metab Res. 2015;47(11):826-32.

187. Velasco A, Chung O, Raza F, Pandey A, Brinker S, Arbique D, et al. Cost-Effectiveness of Therapeutic Drug Monitoring in Diagnosing Primary Aldosteronism in Patients With Resistant Hypertension. J Clin Hypertens. 2015;17(9):713-9.

188. Swearingen B, Wu N, Chen SY, Pulgar S, Biller BM. Health care resource use and costs among patients with cushing disease. Endocr Pract. 2011;17(5):681-90.

189. Broder M, Neary MP, Chang E, W. L, Cherepanov D, editors. Annual health care utilization and costs in Cushing's disease patients in the United States. The Professional Society for Health Economics and Outcomes Research (ISPOR); 2013: Value in Health.

190. Broder MS, Neary MP, Chang E, Cherepanov D, Ludlam WH. Burden of illness, annual healthcare utilization, and costs associated with commercially insured patients with Cushing disease in the United States. Endocr Pract. 2015;21(1):77-86.

191. Broder MS, Neary MP, Chang E, Ludlam WH. Incremental healthcare resource utilization and costs in US patients with Cushing's disease compared with diabetes mellitus and population controls. Pituitary. 2015;18(6):796-802.

192. Burton TM, LeNestour E, Neary MP, Ludlam WH, editors. Economic burden of Cushing disease in a large United States managed care health plan. The Professional Society for Health Economics and Outcomes Research (ISPOR); 2014: Value in Health.

193. Burton T, Le Nestour E, Neary M, Ludlam WH. Algorithm development and the clinical and economic burden of Cushing's disease in a large US health plan database. Pituitary. 2016;19(2):167-74.

194. Murray RD, Forsythe A, Siva V, Oliver N, Rojas-Farreras S, Roset M, editors. Treatment patterns and burden of illness of Cushing's disease in the United Kingdom: Real world (RW) data from clinical practice research datalink (CPRD). The Professional Society for Health Economics and Outcomes Research (ISPOR); 2013: Value in Health.

195. Chauhan R, editor Adrenal insufficiency: Burden of disease and cost of illness. The Professional Society for Health Economics and Outcomes Research (ISPOR); 2013: Value in Health.

196. Gunnarsson C, Ryan MP, Marelli C, Baker ER, Stewart PM, Johannsson G, et al. Health Care Burden in Patients With Adrenal Insufficiency. J Endocr Soc. 2017;1(5):512-23.

197. Velema M, Dekkers T, Hermus A, Timmers H, Lenders J, Groenewoud H, et al. Quality of Life in Primary Aldosteronism: A Comparative Effectiveness Study of Adrenalectomy and Medical Treatment. J Clin Endocrinol Metab. 2018;103(1):16-24.

198. Sukor N, Kogovsek C, Gordon RD, Robson D, Stowasser M. Improved quality of life, blood pressure, and biochemical status following laparoscopic adrenalectomy for unilateral primary aldosteronism. J Clin Endocrinol Metab. 2010;95(3):1360-4.

199. Ahmed AH, Gordon RD, Sukor N, Pimenta E, Stowasser M. Quality of life in patients with bilateral primary aldosteronism before and during treatment with spironolactone and/or amiloride, including a comparison with our previously published results in those with unilateral disease treated surgically. J Clin Endocrinol Metab. 2011;96(9):2904-11.

200. Kunzel HE, Apostolopoulou K, Pallauf A, Gerum S, Merkle K, Schulz S, et al. Quality of life in patients with primary aldosteronism: gender differences in untreated and long-term treated patients and associations with treatment and aldosterone. J Psychiatr Res. 2012;46(12):1650-4.

201. Citton M, Viel G, Torresan F, Rossi GP, Iacobone M. Effect of unilateral adrenalectomy on the quality of life of patients with lateralized primary aldosteronism. BMC Surg. 2019;18(Suppl 1):105.

202. Badia X, Roset M, Valassi E, Franz H, Forsythe A, Webb SM. Mapping CushingQOL scores to EQ-5D utility values using data from the European Registry on Cushing's syndrome (ERCUSYN). Qual Life Res. 2013;22(10):2941-50.

203. Valassi E, Feelders R, Maiter D, Chanson P, Yaneva M, Reincke M, et al. Worse Health-Related Quality of Life at long-term follow-up in patients with Cushing's disease than patients with cortisol producing adenoma. Data from the ERCUSYN. Clin Endocrinol (Oxf). 2018;88(6):787-98.

204. Hawn MT, Cook D, Deveney C, Sheppard BC. Quality of life after laparoscopic bilateral adrenalectomy for Cushing's disease. Surgery. 2002;132(6):1064-8; discussion 8-9.

205. Thompson SK, Hayman AV, Ludlam WH, Deveney CW, Loriaux DL, Sheppard BC. Improved quality of life after bilateral laparoscopic adrenalectomy for Cushing's disease: a 10-year experience. Ann Surg. 2007;245(5):790-4.

206. De Bucy C, Guignat L, Niati T, Bertherat J, Coste J. Health-related quality of life of patients with hypothalamic-pituitary-adrenal axis dysregulations: a cohort study. Eur J Endocrinol. 2017;177(1):1-8.

207. Tiemensma J, Andela CD, Kaptein AA, Romijn JA, van der Mast RC, Biermasz NR, et al. Psychological morbidity and impaired quality of life in patients with stable treatment for primary adrenal insufficiency: cross-sectional study and review of the literature. Eur J Endocrinol. 2014;171(2):171-82.

208. Kluger N, Matikainen N, Sintonen H, Ranki A, Roine RP, Schalin-Jantti C. Impaired health-related quality of life in Addison's disease--impact of replacement therapy, comorbidities and socio-economic factors. Clin Endocrinol (Oxf). 2014;81(4):511-8.

209. Webb SM, Prieto L, Badia X, Albareda M, Catala M, Gaztambide S, et al. Acromegaly Quality of Life Questionnaire (ACROQOL) a new health-related quality of life questionnaire for patients with acromegaly: development and psychometric properties. Clin Endocrinol (Oxf). 2002;57(2):251-8.

210. Oksnes M, Bensing S, Hulting AL, Kampe O, Hackemann A, Meyer G, et al. Quality of life in European patients with Addison's disease: validity of the disease-specific questionnaire AddiQoL. J Clin Endocrinol Metab. 2012;97(2):568-76.

211. Webb SM, Badia X, Barahona MJ, Colao A, Strasburger CJ, Tabarin A, et al. Evaluation of health-related quality of life in patients with Cushing's syndrome with a new questionnaire. Eur J Endocrinol. 2008;158(5):623-30.

212. Badia X, Valassi E, Roset M, Webb SM. Disease-specific quality of life evaluation and its determinants in Cushing's syndrome: what have we learnt? Pituitary. 2014;17(2):187-95.

213. McKenna SP, Doward LC, Alonso J, Kohlmann T, Niero M, Prieto L, et al. The QoL-AGHDA: an instrument for the assessment of quality of life in adults with growth hormone deficiency. Qual Life Res. 1999;8(4):373-83.

214. Christakis I, Livesey JA, Sadler GP, Mihai R. Laparoscopic Adrenalectomy for Conn's Syndrome is Beneficial to Patients and is Cost Effective in England. J Invest Surg. 2018;31(4):300-6.

215. Walker JJ, Terry JR, Lightman SL. Origin of ultradian pulsatility in the hypothalamic-pituitary-adrenal axis. Proc Biol Sci. 2010;277(1688):1627-33.

216. International Conference on Harmonization. Good Clinical Practice Guidelines. July 2020. Available from: https://ichgcp.net/.

217. UK Human Tissue Authority. Human Tissue Act 2004. July 2020. Available from: https://www.hta.gov.uk/policies/human-tissue-act-2004.

218. UK Government. Data Protection Act 1998. July 2020. Available from: https://www.legislation.gov.uk/ukpga/1998/29/contents.

219. EuroQol. EQ-5D Instruments. November 2019. Available from: https://euroqol.org/.

220. Reilly Associates. Work Productivity and Activity Impairment (WPAI) Questionnaire. August 2020. Available from: http://www.reillyassociates.net/WPAI_GH.html.

221. Devlin NJ, Brooks R. EQ-5D and the EuroQol Group: Past, Present and Future. Appl Health Econ Health Policy. 2017;15(2):127-37.

222. Pickard AS, Kohlmann T, Janssen MF, Bonsel G, Rosenbloom S, Cella D. Evaluating equivalency between response systems: application of the Rasch model to a 3-level and 5-level EQ-5D. Med Care. 2007;45(9):812-9.

223. Pickard AS, De Leon MC, Kohlmann T, Cella D, Rosenbloom S. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care. 2007;45(3):259-63.

224. Janssen MF, Birnie E, Bonsel GJ. Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods. Qual Life Res. 2008;17(3):463-73.

225. Janssen MF, Birnie E, Haagsma JA, Bonsel GJ. Comparing the standard EQ-5D three-level system with a five-level version. Value Health. 2008;11(2):275-84.

226. Buchholz I, Janssen MF, Kohlmann T, Feng YS. A Systematic Review of Studies Comparing the Measurement Properties of the Three-Level and Five-Level Versions of the EQ-5D. PharmacoEcon. 2018;36(6):645-61.

227. Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L Better Than EQ-5D-3L? A Head-to-Head Comparison of Descriptive Systems and Value Sets from Seven Countries. PharmacoEcon. 2018;36(6):675-97.

228. Gray AM. Applied methods of cost-effectiveness analysis in health care. Oxford: Oxford University Press; 2011.

229. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012;15(5):708-15.

230. Position statement on use of the EQ-5D-5L value set for England [press release]. October 2019.

231. Tordrup D, Mossman J, Kanavos P. Responsiveness of the EQ-5D to clinical change: is the patient experience adequately represented? Int J Technol Assess Health Care. 2014;30(1):10-9.

232. Payakachat N, Ali MM, Tilford JM. Can The EQ-5D Detect Meaningful Change? A Systematic Review. PharmacoEcon. 2015;33(11):1137-54.

233. Meregaglia M, Nicod E, Drummond M. The estimation of health state utility values in rare diseases: overview of existing techniques. Int J Technol Assess Health Care. 2020;36(5):469-73.

234. Efthymiadou O, Mossman J, Kanavos P. Health related quality of life aspects not captured by EQ-5D-5L: Results from an international survey of patients. Health Policy. 2019;123(2):159-65.

235. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med. 2005;60(7):1571-82.

236. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ. 2010;11(2):215-25.

237. Lin FJ, Longworth L, Pickard AS. Evaluation of content on EQ-5D as compared to disease-specific utility measures. Qual Life Res. 2013;22(4):853-74.

238. Knoble N, Nayroles G, Cheng C, Arnould B. Illustration of patient-reported outcome challenges and solutions in rare diseases: a systematic review in Cushing's syndrome. Orphanet J Rare Dis. 2018;13(1):228.

239. Mann R, Brazier J, Tsuchiya A. A comparison of patient and general population weightings of EQ-5D dimensions. Health Econ. 2009;18(3):363-72.

240. Silva EN, Sousa TR. Economic evaluation in the context of rare diseases: is it possible? Cad Saude Publica. 2015;31(3):496-506.

241. Brazier J, Tsuchiya A. Preference-based condition-specific measures of health: what happens to cross programme comparability? Health Econ. 2010;19(2):125-9.

242. Rand Health Care. 36-Item Short Form Survey (SF-36). July 2020. Available from: https://www.rand.org/health-care/surveys_tools/mos/36-item-short-form.html.

243. Hernandez Alava M, Wailoo AJ, Ara R. Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. Value Health. 2012;15(3):550-61.

244. Wilkinson G, Drummond M. Alternative approaches for assessing the socioeconomic benefits of medical devices: a systematic review. Expert Rev Med Devices. 2015;12(5):629-48.

245. Ryan M, Gerard K, Amaya-Amaya M. Using Discrete Choice Experiments to Value Health and Health Care: Springer, Dordrecht; 2008.

246. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36.

247. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol. 2019.

248. Clinical Practice Research Datalink (CPRD). September 2019. Available from: https://www.cprd.com/.

249. National Institute for Health Research (NIHR). National Institute for Health Research (NIHR). September 2019. Available from: https://www.nihr.ac.uk/.

250. Medicines and Healthcare products Regulatory Agency (MHRA). Medicines and Healthcare products Regulatory Agency (MHRA). September 2019. Available from: https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency.

251. Adolf C, Asbach E, Dietz AS, Lang K, Hahner S, Quinkler M, et al. Worsening of lipid metabolism after successful treatment of primary aldosteronism. Endocrine. 2016;54(1):198-205.

252. Adolf C, Heinrich DA, Holler F, Lechner B, Nirschl N, Sturm L, et al. Patients With Primary Aldosteronism Respond to Unilateral Adrenalectomy With Long-Term Reduction in Salt Intake. J Clin Endocrinol Metab. 2020;105(3).

253. Asbach E, Bekeran M, Konig A, Lang K, Hanslik G, Treitl M, et al. Primary and Secondary Hyperparathyroidism in Patients with Primary Aldosteronism - Findings From the German Conn's Registry. Exp Clin Endocrinol Diabetes. 2020;128(4):246-54.

254. Born-Frontsberg E, Reincke M, Beuschlein F, Quinkler M, Participants of German Conn's R. Tumor size of Conn's adenoma and comorbidities. Horm Metab Res. 2009;41(10):785-8.

255. Born-Frontsberg E, Reincke M, Rump LC, Hahner S, Diederich S, Lorenz R, et al. Cardiovascular and cerebrovascular comorbidities of hypokalemic and normokalemic primary aldosteronism: results of the German Conn's Registry. J Clin Endocrinol Metab. 2009;94(4):1125-30.

256. Er LK, Chen L, Tsai YC, Lin YH, Huang WC, Chang CC, et al. Risk of new-onset autoimmune diseases in primary aldosteronism: a nation-wide population-based study. J Hypertens. 2020;38(4):745-54.

257. Fischer E, Adolf C, Pallauf A, Then C, Bidlingmaier M, Beuschlein F, et al. Aldosterone excess impairs first phase insulin secretion in primary aldosteronism. J Clin Endocrinol Metab. 2013;98(6):2513-20.

258. Fischer E, Beuschlein F, Degenhart C, Jung P, Bidlingmaier M, Reincke M. Spontaneous remission of idiopathic aldosteronism after long-term treatment with spironolactone: results from the German Conn's Registry. Clin Endocrinol (Oxf). 2012;76(4):473-7.

259. Fischer E, Reuschl S, Quinkler M, Rump LC, Hahner S, Bidlingmaier M, et al. Assay characteristics influence the aldosterone to renin ratio as a screening tool for primary aldosteronism: results of the German Conn's registry. Horm Metab Res. 2013;45(7):526-31.

260. Fujii Y, Takeda Y, Kurihara I, Itoh H, Katabami T, Ichijo T, et al. Historical changes and between-facility differences in adrenal venous sampling for primary aldosteronism in Japan. J Hum Hypertens. 2020;34(1):34-42.

261. Handgriff L, Adolf C, Heinrich DA, Braun L, Nirschl N, Sturm L, et al. The Impact of Glucocorticoid Co-Secretion in Primary Aldosteronism on Thyroid Autoantibody Titers During the Course of Disease. Horm Metab Res. 2020;52(6):404-11.

262. Hanslik G, Wallaschofski H, Dietz A, Riester A, Reincke M, Allolio B, et al. Increased prevalence of diabetes mellitus and the metabolic syndrome in patients with primary aldosteronism of the German Conn's Registry. Eur J Endocrinol. 2015;173(5):665-75.

263. Hanusch FM, Fischer E, Lang K, Diederich S, Endres S, Allolio B, et al. Sleep quality in patients with primary aldosteronism. Hormones (Athens). 2014;13(1):57-64.

264. Heinrich DA, Adolf C, Holler F, Lechner B, Schneider H, Riester A, et al. Adrenal Insufficiency After Unilateral Adrenalectomy in Primary Aldosteronism: Long-Term Outcome and Clinical Impact. J Clin Endocrinol Metab. 2019;104(11):5658-64.

265. Heinrich DA, Adolf C, Rump LC, Quack I, Quinkler M, Hahner S, et al. Primary aldosteronism: key characteristics at diagnosis: a trend toward milder forms. Eur J Endocrinol. 2018;178(6):605-11.

266. Hundemer GL, Curhan GC, Yozamp N, Wang M, Vaidya A. Cardiometabolic outcomes and mortality in medically treated primary aldosteronism: a retrospective cohort study. Lancet Diabetes Endocrinol. 2018;6(1):51-9.

267. Katabami T, Fukuda H, Tsukiyama H, Tanaka Y, Takeda Y, Kurihara I, et al. Clinical and biochemical outcomes after adrenalectomy and medical treatment in patients with unilateral primary aldosteronism. J Hypertens. 2019;37(7):1513-20.

268. Kawashima J, Araki E, Naruse M, Kurihara I, Takahashi K, Tamura K, et al. Baseline Plasma Aldosterone Level and Renin Activity Allowing Omission of Confirmatory Testing in Primary Aldosteronism. J Clin Endocrinol Metab. 2020;105(5).

269. Kobayashi H, Nakamura Y, Abe M, Kurihara I, Itoh H, Ichijo T, et al. Effect of cosyntropin during adrenal venous sampling on subtype of primary aldosteronism: analysis of surgical outcome. Eur J Endocrinol. 2020;182(3):265-73.

270. Morisaki M, Kurihara I, Itoh H, Naruse M, Takeda Y, Katabami T, et al. Predictors of Clinical Success After Surgery for Primary Aldosteronism in the Japanese Nationwide Cohort. J Endocr Soc. 2019;3(11):2012-22.

271. Ohno Y, Sone M, Inagaki N, Yamasaki T, Ogawa O, Takeda Y, et al. Prevalence of Cardiovascular Disease and Its Risk Factors in Primary Aldosteronism: A Multicenter Study in Japan. Hypertension. 2018;71(3):530-7.

272. Reincke M, Fischer E, Gerum S, Merkle K, Schulz S, Pallauf A, et al. Observational study mortality in treated primary aldosteronism: the German Conn's registry. Hypertension. 2012;60(3):618-24.

273. Reincke M, Rump LC, Quinkler M, Hahner S, Diederich S, Lorenz R, et al. Risk factors associated with a low glomerular filtration rate in primary aldosteronism. J Clin Endocrinol Metab. 2009;94(3):869-75.

274. Remde H, Dietz A, Emeny R, Riester A, Peters A, de Las Heras Gala T, et al. The cardiovascular markers copeptin and high-sensitive C-reactive protein decrease following specific therapy for primary aldosteronism. J Hypertens. 2016;34(10):2066-73.

275. Riester A, Fischer E, Degenhart C, Reiser MF, Bidlingmaier M, Beuschlein F, et al. Age below 40 or a recently proposed clinical prediction score cannot bypass adrenal venous sampling in primary aldosteronism. J Clin Endocrinol Metab. 2014;99(6):E1035-9.

276. Rossi GP, Rossitto G, Amar L, Azizi M, Riester A, Reincke M, et al. Clinical Outcomes of 1625 Patients With Primary Aldosteronism Subtyped With Adrenal Vein Sampling. Hypertension. 2019;74(4):800-8.

277. Rossitto G, Amar L, Azizi M, Riester A, Reincke M, Degenhart C, et al. Subtyping of Primary Aldosteronism in the AVIS-2 Study: Assessment of Selectivity and Lateralization. J Clin Endocrinol Metab. 2020;105(6).

278. Schirpenbach C, Segmiller F, Diederich S, Hahner S, Lorenz R, Rump LC, et al. The diagnosis and treatment of primary hyperaldosteronism in Germany: results on 555 patients from the German Conn Registry. Dtsch Arztebl Int. 2009;106(18):305-11.

279. Umakoshi H, Tsuiki M, Takeda Y, Kurihara I, Itoh H, Katabami T, et al. Significance of Computed Tomography and Serum Potassium in Predicting Subtype Diagnosis of Primary Aldosteronism. J Clin Endocrinol Metab. 2018;103(3):900-8.

280. Vonend O, Ockenfels N, Gao X, Allolio B, Lang K, Mai K, et al. Adrenal venous sampling: evaluation of the German Conn's registry. Hypertension. 2011;57(5):990-5.

281. Weigel M, Riester A, Hanslik G, Lang K, Willenberg HS, Endres S, et al. Post-saline infusion test aldosterone levels indicate severity and outcome in primary aldosteronism. Eur J Endocrinol. 2015;172(4):443-50.

282. Wu CH, Wu V, Yang YW, Lin YH, Yang SY, Lin PC, et al. Plasma Aldosterone After Seated Saline Infusion Test Outperforms Captopril Test at Predicting Clinical Outcomes After Adrenalectomy for Primary Aldosteronism. Am J Hypertens. 2019;32(11):1066-74.

283. Wu VC, Chueh SJ, Chen L, Chang CH, Hu YH, Lin YH, et al. Risk of new-onset diabetes mellitus in primary aldosteronism: a population study over 5 years. J Hypertens. 2017;35(8):1698-708.

284. NHS Digital. Hospital Episode Statistics (HES). September 2019. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics.

285. UK Government. English indices of deprivation 2015. September 2019. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015.

286. Office for National Statistics (ONS). ONS Mortality Data. September 2019. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths.

287. Clinical Practice Research Datalink (CPRD). CPRD GOLD Release Notes. London: National Institute for Health Research (NIHR) and Medicines & Healthcare products Regulatory Agency (MHRA); 2017.

288. NHS Digital. September 2019. Available from: https://digital.nhs.uk.

289. Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. Eur J Epidemiol. 2019;34(1):91-9.

290. NIHR Collaboration for Leadership in Applied Health Research and Care West (CLAHRC West). September 2009. Available from: https://clahrc-west.nihr.ac.uk/.

291. Vaidya A, Hamrahian AH, Auchus RJ. Genetics of primary aldosteronism. Endocr Pract. 2015;21(4):400-5.

292. Hennessy S, Bilker WB, Berlin JA, Strom BL. Factors influencing the optimal control-to-case ratio in matched case-control studies. Am J Epidemiol. 1999;149(2):195-7.

293. World Health Organization (WHO). ICD-10 Version:2010. September 2019. Available from: http://apps.who.int/classifications/icd10/browse/2010/en.

294. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. J Clin Epidemiol. 1994;47(11):1245-51.

295. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373-83.

296. Khan NF, Perera R, Harper S, Rose PW. Adaptation and validation of the Charlson Index for Read/OXMIS coded databases. BMC Fam Pract. 2010;11:1.

297. Walker VM, Davies NM, Martin RM, Kehoe PG. Comparison of antihypertensive drug classes for dementia prevention. BioRxiv: The Preprint Server for Biology. January 2019.

298. Deb P, Norton EC. Modeling Health Care Expenditures and Use. Annu Rev Public Health. 2018;39:489-505.

299. Deb P, Norton EC, Manning WG. Health econometrics using Stata. College Station, Texas: Stata Press; 2017. xvi, 264 pages p.

300. Curtis L, Burns A. Unit Costs of Health and Social Care 2018 (PSSRU). Canterbury; 2018.

301. Curtis L. Unit Costs of Health and Social Care 2010 (PSSRU). Canterbury; 2010.

302. NHS Improvement. NHS Reference Costs 2017/2018. Available from: https://improvement.nhs.uk/resources/reference-costs/.

303. NHS Digital. HRG4+ 2017/2018 Local Payment Grouper. May 2018. Available from: https://digital.nhs.uk/services/national-casemix-office/downloads-groupers-and-tools/payment-hrg4-2017-18-local-payment-grouper.

304. Royal Pharmaceutical Society of Great Britain and British Medical Association. British National Formulary (BNF). BMJ Publishing Group Ltd and Royal Pharmaceutical Society; June 2019.

305. NHS Digital. Prescription Cost Analysis (PCA), England. 2018. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/prescription-cost-analysis/prescription-cost-analysis-england-2017.

306. NHS Business Services Authority (NHSBSA). Prescription Cost Analysis (PCA) data. 2018. Available from: https://www.nhsbsa.nhs.uk/prescription-data/dispensing-data/prescription-cost-analysis-pca-data.

307. Box GEP, Cox DR. An Analysis of Transformations. J Roy Stat Soc B. 1964;26(2):211-52.

308. Park RE. Estimation with Heteroscedastic Error Terms. Econometrica. 1966;34(4):888-&.

309. Pregibon D. Logistic-Regression Diagnostics. Ann Stat. 1981;9(4):705-24.

310. Pregibon D. Goodness of Link Tests for Generalized Linear Models. Journal of the Royal Statistical Society Series C (Applied Statistics). 1980;29(1):15-24.

311. Akaike H. Statistical Predictor Identification. Ann I Stat Math. 1970;22(2):203-&.

312. Schwarz G. Estimating Dimension of a Model. Ann Stat. 1978;6(2):461-4.

313. Leroux BG. Consistent Estimation of a Mixing Distribution. Ann Stat. 1992;20(3):1350-60.

314. O'Dowd A. Watchdog claims more accurate GP lists will save NHS 6.1m pounds sterling. BMJ. 2012;344:e1366.

315. Calhoun DA, Nishizaka MK, Zaman MA, Thakkar RB, Weissmann P. Hyperaldosteronism among black and white subjects with resistant hypertension. Hypertension. 2002;40(6):892-6.

316. Mattson DL, Kunert MP, Kaldunski ML, Greene AS, Roman RJ, Jacob HJ, et al. Influence of diet and genetics on hypertension and renal disease in Dahl salt-sensitive rats. Physiol Genomics. 2004;16(2):194-203.

317. Brown MJ. Hypertension and ethnic group. BMJ. 2006;332(7545):833-6.

318. Bertakis KD, Azari R, Helms LJ, Callahan EJ, Robbins JA. Gender differences in the utilization of health care services. J Fam Pract. 2000;49(2):147-52.

319. Gallagher AM, Dedman D, Padmanabhan S, Leufkens HGM, de Vries F. The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations. Pharmacoepidemiol Drug Saf. 2019;28(5):563-9.

320. Harshfield A, Abel GA, Barclay S, Payne RA. Do GPs accurately record date of death? A UK observational analysis. BMJ Support Palliat Care. 2018.

321. Delgado-Rodriguez M, Llorca J. Bias. J Epidemiol Community Health. 2004;58(8):635-41.

322. Steiner PM, Kim Y. The Mechanics of Omitted Variable Bias: Bias Amplification and Cancellation of Offsetting Biases. J Causal Inference. 2016;4(2).

323. Ou HT, Mukherjee B, Erickson SR, Piette JD, Bagozzi RP, Balkrishnan R. Comparative performance of comorbidity indices in predicting health care-related behaviors and outcomes among Medicaid enrollees with type 2 diabetes. Popul Health Manag. 2012;15(4):220-9.

324. Tarricone R. Cost-of-illness analysis. What room in health economics? Health Policy. 2006;77(1):51-63.

325. Akobundu E, Ju J, Blatt L, Mullins CD. Cost-of-illness studies : a review of current methods. PharmacoEcon. 2006;24(9):869-90.

326. Larg A, Moss JR. Cost-of-illness studies: a guide to critical evaluation. PharmacoEcon. 2011;29(8):653-71.

327. Onukwugha E, McRae J, Kravetz A, Varga S, Khairnar R, Mullins CD. Cost-of-Illness Studies: An Updated Review of Current Methods. PharmacoEcon. 2016;34(1):43-58.

328. Rice DP. Cost-of-illness studies: fact or fiction? Lancet. 1994;344(8936):1519-20.

329. Ramsey JB. Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. J Roy Stat Soc B. 1969;31(2):350-&.

330. Funder JW, Carey RM, Fardella C, Gomez-Sanchez CE, Mantero F, Stowasser M, et al. Case detection, diagnosis, and treatment of patients with primary aldosteronism: an endocrine society clinical practice guideline. J Clin Endocrinol Metab. 2008;93(9):3266-81.

331. Nishikawa T, Omura M, Satoh F, Shibata H, Takahashi K, Tamura N, et al. Guidelines for the diagnosis and treatment of primary aldosteronism--the Japan Endocrine Society 2009. Endocr J. 2011;58(9):711-21.

332. Mysliwiec J, Gorska M. Primary aldosteronism: a common and important problem. A practical guide to the diagnosis and treatment. Endokrynol Pol. 2012;63(4):324-36.

333. Milliez P, Girerd X, Plouin PF, Blacher J, Safar ME, Mourad JJ. Evidence for an increased rate of cardiovascular events in patients with primary aldosteronism. J Am Coll Cardiol. 2005;45(8):1243-8.

334. Rossi GP, Sechi LA, Giacchetti G, Ronconi V, Strazzullo P, Funder JW. Primary aldosteronism: cardiovascular, renal and metabolic implications. Trends Endocrinol Metab. 2008;19(3):88-90.

335. Young WF. Primary aldosteronism: renaissance of a syndrome. Clin Endocrinol (Oxf). 2007;66(5):607-18.

336. Young WF, Jr. Diagnosis and treatment of primary aldosteronism: practical clinical perspectives. J Intern Med. 2019;285(2):126-48.

337. Mattsson C, Young WF, Jr. Primary aldosteronism: diagnostic and treatment strategies. Nat Clin Pract Nephrol. 2006;2(4):198-208; quiz, 1 p following 30.

338. Chao CT, Wu VC, Kuo CC, Lin YH, Chang CC, Chueh SJ, et al. Diagnosis and management of primary aldosteronism: an updated review. Ann Med. 2013;45(4):375-83.

339. Aglony M, Martinez-Aguayo A, Carvajal CA, Campino C, Garcia H, Bancalari R, et al. Frequency of familial hyperaldosteronism type 1 in a hypertensive pediatric population: clinical and biochemical presentation. Hypertension. 2011;57(6):1117-21.

340. Mulatero P, Tizzani D, Viola A, Bertello C, Monticone S, Mengozzi G, et al. Prevalence and characteristics of familial hyperaldosteronism: the PATOGEN study (Primary Aldosteronism in TOrino-GENetic forms). Hypertension. 2011;58(5):797-803.

341. Pallauf A, Schirpenbach C, Zwermann O, Fischer E, Morak M, Holinski-Feder E, et al. The prevalence of familial hyperaldosteronism in apparently sporadic primary aldosteronism in Germany: a single center experience. Horm Metab Res. 2012;44(3):215-20.

342. Ng L, Libertino JM. Adrenocortical carcinoma: diagnosis, evaluation and treatment. J Urol. 2003;169(1):5-11.

343. Douma S, Petidis K, Doumas M, Papaefthimiou P, Triantafyllou A, Kartali N, et al. Prevalence of primary hyperaldosteronism in resistant hypertension: a retrospective observational study. Lancet. 2008;371(9628):1921-6.

344. Gouli A, Kaltsas G, Tzonou A, Markou A, Androulakis, II, Ragkou D, et al. High prevalence of autonomous aldosterone secretion among patients with essential hypertension. Eur J Clin Invest. 2011;41(11):1227-36.

345. Markou A, Sertedaki A, Kaltsas G, Androulakis, II, Marakaki C, Pappa T, et al. Stress-induced Aldosterone Hyper-Secretion in a Substantial Subset of Patients With Essential Hypertension. J Clin Endocrinol Metab. 2015;100(8):2857-64.

346. Papanastasiou L, Markou A, Pappa T, Gouli A, Tsounas P, Fountoulakis S, et al. Primary aldosteronism in hypertensive patients: clinical implications and target therapy. Eur J Clin Invest. 2014;44(8):697-706.

347. Grumbach MM, Biller BMK, Braunstein GD, Campbell KK, Carney JA, Godley PA, et al. Management of the clinically inapparent adrenal mass ("incidentaloma"). Ann Intern Med. 2003;138(5):424-9.

348. Pimenta E, Calhoun DA. Primary aldosteronism: diagnosis and treatment. J Clin Hypertens (Greenwich). 2006;8(12):887-93.

349. Stowasser M, Ahmed AH, Pimenta E, Taylor PJ, Gordon RD. Factors affecting the aldosterone/renin ratio. Horm Metab Res. 2012;44(3):170-6.

350. Seifarth C, Trenkel S, Schobel H, Hahn EG, Hensen J. Influence of antihypertensive medication on aldosterone and renin concentration in the differential diagnosis of essential hypertension and primary aldosteronism. Clin Endocrinol (Oxf). 2002;57(4):457-65.

351. Rossi GP. Diagnosis and treatment of primary aldosteronism. Rev Endocr Metab Disord. 2011;12(1):27-36.

352. Ganguly A, Dowdy AJ, Luetscher JA, Melada GA. Anomalous postural response of plasma aldosterone concentration in patients with aldosterone-producing adrenal adenoma. J Clin Endocrinol Metab. 1973;36(2):401-4.

353. Conn JW, Morita R, Cohen EL, Beierwaltes WH, McDonald WJ, Herwig KR. Primary aldosteronism. Photoscanning of tumors after administration of 131 I-19-iodocholesterol. Arch Intern Med. 1972;129(3):417-25.

354. Biglieri EG, Schambelan M. The significance of elevated levels of plasma 18-hydroxycorticosterone in patients with primary aldosteronism. J Clin Endocrinol Metab. 1979;49(1):87-91.

355. Hennings J, Sundin A, Hagg A, Hellman P. 11C-metomidate positron emission tomography after dexamethasone suppression for detection of small adrenocortical adenomas in primary aldosteronism. Langenbecks Arch Surg. 2010;395(7):963-7.

356. Zettinig G, Mitterhauser M, Wadsak W, Becherer A, Pirich C, Vierhapper H, et al. Positron emission tomography imaging of adrenal masses: (18)F-fluorodeoxyglucose and the 11beta-hydroxylase tracer (11)C-metomidate. Eur J Nucl Med Mol Imaging. 2004;31(9):1224-30.

357. Sywak M, Pasieka JL. Long-term follow-up and cost benefit of adrenalectomy in patients with primary hyperaldosteronism. Br J Surg. 2002;89(12):1587-93.

358. Reimel B, Zanocco K, Russo MJ, Zarnegar R, Clark OH, Allendorf JD, et al. The management of aldosterone-producing adrenal adenomas--does adrenalectomy increase costs? Surgery. 2010;148(6):1178-85; discussion 85.

359. Ferrante di Ruffano L, Dinnes J, Sitch AJ, Hyde C, Deeks JJ. Test-treatment RCTs are susceptible to bias: a review of the methodological quality of randomized trials that evaluate diagnostic tests. BMC Med Res Methodol. 2017;17(1):35.

360. Karnon J. Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. Health Econ. 2003;12(10):837-48.

361. Caro JJ. Pharmacoeconomic analyses using discrete event simulation. PharmacoEcon. 2005;23(4):323-32.

362. Karnon J, Stahl J, Brennan A, Caro JJ, Mar J, Moller J. Modeling Using Discrete Event Simulation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Med Decis Making. 2012;32(5):701-11.

363. Pitman R, Fisman D, Zaric GS, Postma M, Kretzschmar M, Edmunds J, et al. Dynamic Transmission Modeling: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-5. Value Health. 2012;15(6):828-34.

364. Fenwick E, Steuten L, Knies S, Ghabri S, Basu A, Murray JF, et al. Value of Information Analysis for Research Decisions-An Introduction: Report 1 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. Value Health. 2020;23(2):139-50.

365. Rothery C, Strong M, Koffijberg HE, Basu A, Ghabri S, Knies S, et al. Value of Information Analytical Methods: Report 2 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. Value Health. 2020;23(3):277-86.

366. Caro JJ, Briggs AH, Siebert U, Kuntz KM, Pract I-SMGR. Modeling Good Research Practices-Overview: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Med Decis Making. 2012;32(5):667-77.

367. Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M, et al. Conceptualizing a Model: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-2. Value Health. 2012;15(6):804-11.

368. Siebert U, Alagoz O, Bayoumi AM, Jahn B, Owens DK, Cohen DJ, et al. State-Transition Modeling: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. Med Decis Making. 2012;32(5):690-700.

369. Briggs AH, Weinstein MC, Fenwick EAL, Karnon J, Sculpher MJ, Paltiel AD, et al. Model Parameter Estimation and Uncertainty: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. Value Health. 2012;15(6):835-42.

370. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB, et al. Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Med Decis Making. 2012;32(5):733-43.

371. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models. A suggested framework and example of application. PharmacoEcon. 2000;17(5):461-77.

372. Kaltenthaler E, Tappenden P, Paisley S. Reviewing the evidence to inform the population of cost-effectiveness models within health technology assessments. Value Health. 2013;16(5):830-6.

373. Caiazzo R, Marciniak C, Lenne X, Clement G, Theis D, Menegaux F, et al. Adrenalectomy Risk Score: An Original Preoperative Surgical Scoring System to Reduce Mortality and Morbidity After Adrenalectomy. Ann Surg. 2019;270(5):813-9.

374. Wolley MJ, Stowasser M. New Advances in the Diagnostic Workup of Primary Aldosteronism. J Endocr Soc. 2017;1(3):149-61.

375. Catena C, Colussi G, Lapenna R, Nadalini E, Chiuch A, Gianfagna P, et al. Long-term cardiac effects of adrenalectomy or mineralocorticoid antagonists in patients with primary aldosteronism. Hypertension. 2007;50(5):911-8.

376. Catena C, Colussi G, Nadalini E, Chiuch A, Baroselli S, Lapenna R, et al. Cardiovascular outcomes in patients with primary aldosteronism after treatment. Arch Intern Med. 2008;168(1):80-5.

377. Office for National Statistics (ONS). UK national life tables. September 2019. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables.

378. Kaczmarski KR, Sozio SM, Chen J, Sang Y, Shafi T. Resistant hypertension and cardiovascular disease mortality in the US: results from the National Health and Nutrition Examination Survey (NHANES). BMC Nephrol. 2019;20(1):138.

379. NHS Improvement. NHS Reference Costs 2018/2019. Available from: https://improvement.nhs.uk/resources/national-cost-collection/#ncc1819.

380. Stewart PM, Allolio B. Adrenal vein sampling for Primary Aldosteronism: time for a reality check. Clin Endocrinol (Oxf). 2010;72(2):146-8.

381. Rossi GP, Barisa M, Allolio B, Auchus RJ, Amar L, Cohen D, et al. The Adrenal Vein Sampling International Study (AVIS) for identifying the major subtypes of primary aldosteronism. J Clin Endocrinol Metab. 2012;97(5):1606-14.

382. Belsey JD. Optimizing adherence in hypertension: a comparison of outcomes and costs using single tablet regimens vs individual component regimens. J Med Econ. 2012;15(5):897-905.

383. Danese MD, Gleeson M, Kutikova L, Griffiths RI, Azough A, Khunti K, et al. Estimating the economic burden of cardiovascular events in patients receiving lipid-modifying therapy in the UK. BMJ Open. 2016;6(8):e011805.

384. Organisation for Economic Co-operation and Development (OECD). Purchasing Power Parities (PPP). 2019. Available from: https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm.

385. UK Government. GDP deflators at market prices, and money GDP. March 2020. Available from: https://www.gov.uk/government/statistics/gdp-deflators-at-market-prices-and-money-gdp-march-2020-budget.

386. Briggs AH, Bhatt DL, Scirica BM, Raz I, Johnston KM, Szabo SM, et al. Health-related quality-of-life implications of cardiovascular events in individuals with type 2 diabetes mellitus: A subanalysis from the Saxagliptin Assessment of Vascular Outcomes Recorded in Patients with Diabetes Mellitus (SAVOR)-TIMI 53 trial. Diabetes Res Clin Pract. 2017;130:24-33.

387. Chin YR, Lee IS, Lee HY. Effects of hypertension, diabetes, and/or cardiovascular disease on health-related quality of life in elderly Korean individuals: a population-based cross-sectional survey. Asian Nurs Res (Korean Soc Nurs Sci). 2014;8(4):267-73.

388. Rossi GP, Bernini G, Caliumi C, Desideri G, Fabris B, Ferri C, et al. A prospective study of the prevalence of primary aldosteronism in 1,125 hypertensive patients. J Am Coll Cardiol. 2006;48(11):2293-300.

389. Velema MS, Linssen EJM, Hermus A, Groenewoud H, van der Wilt GJ, van Herwaarden AE, et al. A prediction model for primary aldosteronism when the salt loading test is inconclusive. Endocr Connect. 2018;7(12):1308-14.

390. Giacchetti G, Ronconi V, Lucarelli G, Boscaro M, Mantero F. Analysis of screening and confirmatory tests in the diagnosis of primary aldosteronism: need for a standardized protocol. J Hypertens. 2006;24(4):737-45.

391. Mulatero P, Milan A, Fallo F, Regolisti G, Pizzolo F, Fardella C, et al. Comparison of confirmatory tests for the diagnosis of primary aldosteronism. J Clin Endocrinol Metab. 2006;91(7):2618-23.

392. Rossi GP, Belfiore A, Bernini G, Desideri G, Fabris B, Ferri C, et al. Prospective evaluation of the saline infusion test for excluding primary aldosteronism due to aldosterone-producing adenoma. J Hypertens. 2007;25(7):1433-42.

393. Rossi GP, Belfiore A, Bernini G, Desideri G, Fabris B, Ferri C, et al. Comparison of the captopril and the saline infusion test for excluding aldosterone-producing adenoma. Hypertension. 2007;50(2):424-31.

394. Willenberg HS, Vonend O, Schott M, Gao X, Blondin D, Saleh A, et al. Comparison of the saline infusion test and the fludrocortisone suppression test for the diagnosis of primary aldosteronism. Horm Metab Res. 2012;44(7):527-32.

395. Song Y, Yang S, He W, Hu J, Cheng Q, Wang Y, et al. Confirmatory Tests for the Diagnosis of Primary Aldosteronism: A Prospective Diagnostic Accuracy Study. Hypertension. 2018;71(1):118-24.

396. Mulatero P, Bertello C, Rossato D, Mengozzi G, Milan A, Garrone C, et al. Roles of clinical criteria, computed tomography scan, and adrenal vein sampling in differential diagnosis of primary aldosteronism subtypes. J Clin Endocrinol Metab. 2008;93(4):1366-71.

397. Young WF, Stanson AW, Thompson GB, Grant CS, Farley DR, van Heerden JA. Role for adrenal venous sampling in primary aldosteronism. Surgery. 2004;136(6):1227-35.

398. Rossi GP, Pitter G, Bernante P, Motta R, Feltrin G, Miotto D. Adrenal vein sampling for primary aldosteronism: the assessment of selectivity and lateralization of aldosterone excess baseline and after adrenocorticotropic hormone (ACTH) stimulation. J Hypertens. 2008;26(5):989-97.

399. Lim V, Guo Q, Grant CS, Thompson GB, Richards ML, Farley DR, et al. Accuracy of adrenal imaging and adrenal venous sampling in predicting surgical cure of primary aldosteronism. J Clin Endocrinol Metab. 2014;99(8):2712-9.

400. Mailhot JP, Traistaru M, Soulez G, Ladouceur M, Giroux MF, Gilbert P, et al. Adrenal Vein Sampling in Primary Aldosteronism: Sensitivity and Specificity of Basal Adrenal Vein to Peripheral Vein Cortisol and Aldosterone Ratios to Confirm Catheterization of the Adrenal Vein. Radiology. 2015;277(3):887-94.

401. Brunt LM. Minimal access adrenal surgery. Surg Endosc. 2006;20(3):351-61.

402. Steichen O, Zinzindohoue F, Plouin PF, Amar L. Outcomes of adrenalectomy in patients with unilateral primary aldosteronism: a review. Horm Metab Res. 2012;44(3):221-7.

403. Williams TA, Lenders JWM, Mulatero P, Burrello J, Rottenkolber M, Adolf C, et al. Outcomes after adrenalectomy for unilateral primary aldosteronism: an international consensus on outcome measures and analysis of remission rates in an international cohort. Lancet Diabetes Endocrinol. 2017;5(9):689-99.

404. Williams TA, Burrello J, Sechi LA, Fardella CE, Matrozova J, Adolf C, et al. Computed Tomography and Adrenal Venous Sampling in the Diagnosis of Unilateral Primary Aldosteronism. Hypertension. 2018;72(3):641-9.

405. HM Treasury. The Green Book: Central Government Guidance on Appraisal and Evaluation. www.gov.uk: Open Government Licence v3.0; 2018. Available from: https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-governent.

406. Office for National Statistics (ONS). Projected UK adult population for 2018. June 2018. Available from: https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/projectedukadultpopulationfor2018.

407. Joffres M, Falaschetti E, Gillespie C, Robitaille C, Loustalot F, Poulter N, et al. Hypertension prevalence, awareness, treatment and control in national surveys from England, the USA and Canada, and correlation with stroke and ischaemic heart disease mortality: a cross-sectional study. BMJ Open. 2013;3(8):e003423.

408. Sinnott SJ, Smeeth L, Williamson E, Douglas IJ. Trends for prevalence and incidence of resistant hypertension: population based cohort study in the UK 1995-2015. BMJ. 2017;358:j3984.

409. Streeten DH, Tomycz N, Anderson GH. Reliability of screening methods for the diagnosis of primary aldosteronism. Am J Med. 1979;67(3):403-13.

410. Holland OB, Brown H, Kuhnert L, Fairchild C, Risk M, Gomez-Sanchez CE. Further evaluation of saline infusion for the diagnosis of primary aldosteronism. Hypertension. 1984;6(5):717-23.

411. Stowasser M. Is It the Beginning of the End for the Recumbent Saline Infusion Test? Hypertension. 2016;68(4):857-8.

412. Towse Ae, Pritchard CDE, Devlin Ne. Cost-effectiveness thresholds : economic and ethical issues. London: King's Fund; 2002.

413. Devlin N, Parkin D. Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. Health Econ. 2004;13(5):437-52.

414. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. PharmacoEcon. 2008;26(9):733-44.

415. Raftery JP. NICE's cost-effectiveness range: should it be lowered? PharmacoEcon. 2014;32(7):613-5.

416. Rawlins MD, Culyer AJ. National Institute for Clinical Excellence and its value judgments. BMJ. 2004;329(7459):224-7.

417. National Institute for Health and Care Excellence (NICE). Interim Process and Methods of the Highly Specialised Technologies Programme 2017. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-highly-specialised-technologies-guidance.

418. Weinstein N, Martin M, Campbell R. Orphan drugs in the UK, do they meet the NICE Highly Specialised Technology threshold? Value Health. 2017;20:A660.

419. Schirpenbach C, Seiler L, Maser-Gluth C, Rudiger F, Nickel C, Beuschlein F, et al. Confirmatory testing in normokalaemic primary aldosteronism: the value of the saline infusion test and urinary aldosterone metabolites. Eur J Endocrinol. 2006;154(6):865-73.

420. Nanba K, Tamanaha T, Nakao K, Kawashima ST, Usui T, Tagami T, et al. Confirmatory testing in primary aldosteronism. J Clin Endocrinol Metab. 2012;97(5):1688-94.

421. Kayser SC, Deinum J, de Grauw WJ, Schalk BW, Bor HJ, Lenders JW, et al. Prevalence of primary aldosteronism in primary care: a cross-sectional study. Br J Gen Pract. 2018;68(667):e114-e22.

422. Ahmed AH, Cowley D, Wolley M, Gordon RD, Xu S, Taylor PJ, et al. Seated saline suppression testing for the diagnosis of primary aldosteronism: a preliminary study. J Clin Endocrinol Metab. 2014;99(8):2745-53.

423. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: heterogeneity--subgroups, meta-regression, bias, and bias-adjustment. Med Decis Making. 2013;33(5):618-40.

424. Olsen KM, Dahl SA. Health differences between European countries. Soc Sci Med. 2007;64(8):1665-78.

425. Lindgren P, Kahan T, Poulter N, Buxton M, Svarvar P, Dahlof B, et al. Utility loss and indirect costs following cardiovascular events in hypertensive patients: the ASCOT health economic substudy. Eur J Health Econ. 2007;8(1):25-30.

426. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239-51.

427. Ijzerman MJ, Steuten LM. Early assessment of medical technologies to inform product development and market access: a review of methods and applications. Appl Health Econ Health Policy. 2011;9(5):331-47.

428. Soekhai V, de Bekker-Grob EW, Ellis AR, Vass CM. Discrete Choice Experiments in Health Economics: Past, Present and Future. PharmacoEcon. 2019;37(2):201-26.

429. Tinelli M, Ryan M, Bond C. What, who and when? Incorporating a discrete choice experiment into an economic evaluation. Health Econ Rev. 2016;6(1):31.

430. Bosch JL, Beinfeld MT, Muller JE, Brady T, Gazelle GS. A cost-effectiveness analysis of a hypothetical catheter-based strategy for the detection and treatment of vulnerable coronary plaques with drug-eluting stents. J Interv Cardiol. 2005;18(5):339-49.

431. Cao Q, Postmus D, Hillege HL, Buskens E. Probability elicitation to inform early health economic evaluations of new medical technologies: a case study in heart failure disease management. Value Health. 2013;16(4):529-35.

432. Huang W, Gaydos CA, Barnes MR, Jett-Goheen M, Blake DR. Comparative effectiveness of a rapid point-of-care test for detection of Chlamydia trachomatis among women in a clinical setting. Sex Transm Infect. 2013;89(2):108-14.

433. Retel VP, Joore MA, Drukker CA, Bueno-de-Mesquita JM, Knauer M, van Tinteren H, et al. Prospective cost-effectiveness analysis of genomic profiling in breast cancer. Eur J Cancer. 2013;49(18):3773-9.

434. Markiewicz K, van TJ, Steuten LM, MJ IJ. Combining Headroom and Return on Investment Analysis To Rank Potential Commercial Value of Six Medical Devices in Development. Value Health. 2014;17(7):A443.

435. Turner KM, Round J, Horner P, Macleod J, Goldenberg S, Deol A, et al. An early evaluation of clinical and economic costs and benefits of implementing point of care NAAT tests for Chlamydia trachomatis and Neisseria gonorrhoea in genitourinary medicine clinics in England. Sex Transm Infect. 2014;90(2):104-11.

436. Vaidya A, Joore MA, ten Cate-Hoek AJ, ten Cate H, Severens JL. Cost-effectiveness of risk assessment and tailored treatment for peripheral arterial disease patients. Biomark Med. 2014;8(8):989-99.

437. Van Nimwegen KJ. Feasibility of the Headroom Analysis in Early Economic Evaluation of Innovative Diagnostic Technologies With no Immediate Treatment Implications. Value Health. 2014;17(7):A550.

438. Miquel-Cases A, Steuten LM, Retel VP, van Harten WH. Early stage cost-effectiveness analysis of a BRCA1-like test to detect triple negative breast cancers responsive to high dose alkylating chemotherapy. Breast. 2015;24(4):397-405.

439. Wu AC, Gay C, Rett MD, Stout N, Weiss ST, Fuhlbrigge AL. Pharmacogenomic test that predicts response to inhaled corticosteroids in adults with asthma likely to be cost-saving. Pharmacogenomics. 2015;16(6):591-600.

440. Kolominsky-Rabas PL, Kriza C, Djanatliev A, Meier F, Uffenorde S, Radeleff J, et al. Health Economic Impact of a Pulmonary Artery Pressure Sensor for Heart Failure Telemonitoring: A Dynamic Simulation. Telemed J E Health. 2016;22(10):798-808.

441. Luime JJ, Buisman LR, Oppe M, Hazes JM, Rutten-van Molken MP. Cost-Effectiveness Model for Evaluating New Diagnostic Tests in the Evaluation of Patients With Inflammatory Arthritis at Risk of Having Rheumatoid Arthritis. Arthritis Care Res (Hoboken). 2016;68(7):927-35.

442. de Windt TS, Sorel JC, Vonk LA, Kip MMA, Ijzerman MJ, Saris DBF. Early health economic modelling of single-stage cartilage repair. Guiding implementation of technologies in regenerative medicine. J Tissue Eng Regen Med. 2017;11(10):2950-9.

443. Hall PS, Smith A, Hulme C, Vargas-Palacios A, Makris A, Hughes-Davies L, et al. Value of Information Analysis of Multiparameter Tests for Chemotherapy in Early Breast Cancer: The OPTIMA Prelim Trial. Value Health. 2017;20(10):1311-8.

444. Hall PS, Mitchell ED, Smith AF, Cairns DA, Messenger M, Hutchinson M, et al. The future for diagnostic tests of acute kidney injury in critical care: evidence synthesis, care pathway analysis and research prioritisation. Health Technol Assess. 2018;22(32):1-274.

445. Kip MM, Steuten LM, Koffijberg H, MJ IJ, Kusters R. Using expert elicitation to estimate the potential impact of improved diagnostic performance of laboratory tests: a case study on rapid discharge of suspected non-ST elevation myocardial infarction patients. J Eval Clin Pract. 2018;24(1):31-41.

446. Sutton AJ, Lamont JV, Evans RM, Williamson K, O'Rourke D, Duggan B, et al. An early analysis of the cost-effectiveness of a diagnostic classifier for risk stratification of haematuria patients (DCRSHP) compared to flexible cystoscopy in the diagnosis of bladder cancer. PLoS One. 2018;13(8):e0202796.

447. Abel L, Dakin HA, Roberts N, Ashdown HF, Butler CC, Hayward G, et al. Is stratification testing for treatment of chronic obstructive pulmonary disease exacerbations cost-effective in primary care? an early cost-utility analysis. Int J Technol Assess Health Care. 2019;35(2):116-25.

448. Abel L, Shinkins B, Smith A, Sutton AJ, Sagoo GS, Uchegbu I, et al. Early Economic Evaluation of Diagnostic Technologies: Experiences of the NIHR Diagnostic Evidence Co-operatives. Med Decis Making. 2019;39(7):857-66.

449. Grutters JPC, Govers T, Nijboer J, Tummers M, van der Wilt GJ, Rovers MM. Problems and Promises of Health Technologies: The Role of Early Health Economic Modeling. Int J Health Policy Manag. 2019;8(10):575-82.

450. Kapoor R, So JBY, Zhu F, Too HP, Yeoh KG, Yoong JS. Evaluating the Use of microRNA Blood Tests for Gastric Cancer Screening in a Stratified Population-Level Screening Program: An Early Model-Based Cost-Effectiveness Analysis. Value Health. 2020;23(9):1171-9.

451. Lavrentaki A, Paluzzi A, Wass JA, Karavitaki N. Epidemiology of acromegaly: review of population studies. Pituitary. 2017;20(1):4-9.

452. Burton T, Le Nestour E, Neary M, Ludlam WH. Incidence and prevalence of acromegaly in a large US health plan database. Pituitary. 2016;19(3):262-7.

453. Dal J, Feldt-Rasmussen U, Andersen M, Kristensen LO, Laurberg P, Pedersen L, et al. Acromegaly incidence, prevalence, complications and long-term prognosis: a nationwide cohort study. Eur J Endocrinol. 2016;175(3):181-90.

454. Chanson P, Salenave S. Acromegaly. Orphanet J Rare Dis. 2008;3:17.

455. Holdaway IM, Rajasoorya RC, Gamble GD. Factors influencing mortality in acromegaly. J Clin Endocrinol Metab. 2004;89(2):667-74.

456. Melmed S. Medical progress: Acromegaly. N Engl J Med. 2006;355(24):2558-73.

457. Melmed S. Acromegaly pathogenesis and treatment. J Clin Invest. 2009;119(11):3189-202.

458. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Acromegaly. National Institutes of Health (NIH) Publication. 2008;No. 08–3924:1-10.

459. Hannah-Shmouni F, Morissette R, Sinaii N, Elman M, Prezant TR, Chen W, et al. Revisiting the prevalence of nonclassic congenital adrenal hyperplasia in US Ashkenazi Jews and Caucasians. Genet Med. 2017;19(11):1276-9.

460. Khalid JM, Oerton JM, Dezateux C, Hindmarsh PC, Kelnar CJ, Knowles RL. Incidence and clinical features of congenital adrenal hyperplasia in Great Britain. Arch Dis Child. 2012;97(2):101-6.

461. Merke D, Kabbani M. Congenital adrenal hyperplasia: epidemiology, management and practical drug treatment. Paediatr Drugs. 2001;3(8):599-611.

462. Speiser PW, White PC. Congenital adrenal hyperplasia. N Engl J Med. 2003;349(8):776-88.

463. New MI. Diagnosis and management of congenital adrenal hyperplasia. Annu Rev Med. 1998;49:311-28.

464. Turcu AF, Auchus RJ. The next 150 years of congenital adrenal hyperplasia. J Steroid Biochem Mol Biol. 2015;153:63-71.

465. Feldt-Rasmussen U, Klose M. Adult Growth Hormone Deficiency Clinical Management. In: Feingold KR, Anawalt B, Boyce A, Chrousos G, Dungan K, Grossman A, et al., editors. Endotext. South Dartmouth (MA)2000.

466. Stanley T. Diagnosis of growth hormone deficiency in childhood. Curr Opin Endocrinol Diabetes Obes. 2012;19(1):47-52.

467. Stochholm K, Gravholt CH, Laursen T, Jorgensen JO, Laurberg P, Andersen M, et al. Incidence of GH deficiency - a nationwide study. Eur J Endocrinol. 2006;155(1):61-71.

468. Clemmons DR. The diagnosis and treatment of growth hormone deficiency in adults. Curr Opin Endocrinol Diabetes Obes. 2010;17(4):377-83.

469. Sonksen PH, Christiansen JS. Consensus guidelines for the diagnosis and treatment of adults with growth hormone deficiency. Growth Hormone Research Society. Growth Horm IGF Res. 1998;8 Suppl B:89-92.

470. Kargi AY, Merriam GR. Diagnosis and treatment of growth hormone deficiency in adults. Nat Rev Endocrinol. 2013;9(6):335-45.

471. Glynn N, Agha A. Diagnosing growth hormone deficiency in adults. Int J Endocrinol. 2012;2012:972617.

472. Ambrosi B, Passini E, Re T, Barbetta L. The clinical evaluation of silent adrenal masses. J Endocrinol Invest. 1997;20(2):90-107.

473. Brooke AM, Monson JP. Adult growth hormone deficiency. Clin Med (Northfield Il). 2003;3(1):15-9.

474. Carr CE, Cope C, Cohen DL, Fraker DL, Trerotola SO. Comparison of sequential versus simultaneous methods of adrenal venous sampling. J Vasc Interv Radiol. 2004;15(11):1245-50.

475. Craig D, Fayter D, Stirk L, Crott R. Growth monitoring for short stature: Update of a systematic review and economic model. Health Technol Assess. 2011;15(11).

476. Danilowicz K, Fainstein Day P, Manavela MP, Herrera CJ, Deheza ML, Isaac G, et al. Implementing a screening program for acromegaly in Latin America: necessity versus feasibility. Pituitary. 2016;19(4):370-4.

477. Doppman JL. Petrosal sinus sampling and corticotropin-releasing hormone in Cushing's syndrome. Endocrinologist. 1997;7(1 SUPPL.):24S-9S.

478. Durham E. Growth hormone deficiency in children. A change in diagnostic approach. Adv Nurse Pract. 2003;11(1):41-2, 68.

479. Duskova M, Simunkova K, Vitku J, Sosvorova L, Jandikova H, Pospisilova H, et al. A Comparison of Salivary Steroid Levels during Diagnostic Tests for Adrenal Insufficiency. Prague Med Rep. 2016;117(1):18-33.

480. Evans AJ. Screening tests for growth hormone deficiency. J R Soc Med. 1995;88(3):161P-5P.

481. Gibney J, Healy ML, Smith TP, McKenna TJ. A simple and cost-effective approach to assessment of pituitary adrenocorticotropin and growth hormone reserve: combined use of the overnight metyrapone test and insulin-like growth factor-I standard deviation scores. J Clin Endocrinol Metab. 2008;93(10):3763-8.

482. Gross MD, Shapiro B, Shreve P. Radionuclide imaging of the adrenal cortex. Q J Nucl Med. 1999;43(3):224-32.

483. Gunnala V, Guo R, Minutti C, Durazo-Arvizu R, Laporte C, Mathews H, et al. Measurement of salivary cortisol level for the diagnosis of critical illness-related corticosteroid insufficiency in children. Pediatr Crit Care Med. 2015;16(4):e101-6.

484. Isidori AM, Kaltsas GA, Mohammed S, Morris DG, Jenkins P, Chew SL, et al. Discriminatory value of the low-dose dexamethasone suppression test in establishing the diagnosis and differential diagnosis of Cushing's syndrome. J Clin Endocrinol Metab. 2003;88(11):5299-306.

485. Jabbar J, Ghani F, Siddiqui I, Omair A. Diagnostic efficacy of 0, 30, 45, 60, 90 and 120 min growth hormone samples in insulin tolerance test: utility of growth hormone measurement at different time-points and a cost-effective analysis. Scand J Clin Lab Invest. 2009;69(3):359-64.

486. Juul A, Bernasconi S, Chatelain P, Hindmarsh P, Hochberg Z, Hokken-Koelega A, et al. Diagnosis of growth hormone (GH) deficiency and the use of GH in children with growth disorders. Hormone Res. 1999;51(6):284-99.

487. Kievit J, Haak HR. Diagnosis and treatment of adrenal incidentaloma: A cost-effectiveness analysis. Endocrinol Metab Clin North Am. 2000;29(1):69-88.

488. Kohek MB, Nicolau W, Mendonca BB. Non-conjugated versus total urinary cortisol in the diagnosis of Cushing's syndrome. Revista do Hospital das Clinicas; Faculdade de Medicina Da Universidade de Sao Paulo. 1998;53(5):222-4.

489. Kumar B, Swee M. Aldosterone-renin ratio in the assessment of primary aldosteronism. JAMA - Journal of the American Medical Association. 2014;312(2):184-5.

490. Lee MT, Won JG, Lee TI, Yang HJ, Lin HD, Tang KT. The relationship between morning serum cortisol and the short ACTH test in the evaluation of adrenal insufficiency. Chung Hua I Hsueh Tsa Chih (Taipei). 2002;65(12):580-7.

491. Lesen E, Granfeldt D, Houchard A, Dinet J, Berthon A, Olsson DS, et al. Comorbidities, treatment patterns and cost-of-illness of acromegaly in Sweden: a register-linkage population-based study. Eur. 2017;176(2):203-12.

492. Lichtenauer UD, Gerum S, Asbach E, Manolopoulou J, Fourkiotis V, Quinkler M, et al. The Clinical Value of Salivary Aldosterone in Diagnosis and Follow-Up of Primary Aldosteronism. Horm Metab Res. 2016;48(10):638-43.

493. Lin DD, Loughlin KR. Diagnosis and management of surgical adrenal diseases. Urology. 2005;66(3):476-83.

494. Livingston M, Twomey PJ, Basu A, Smellie S, Kane JW, Heald A. Should Free Thyroxine Go Back into the Routine Thyroid Profile? Exp Clin Endocrinol Diabetes. 2015;123(10):594-7.

495. Mysliwiec J, Zukowski L, Grodzka A, Pilaszewicz A, Dragowski S, Gorska M. Diagnostics of primary aldosteronism: is obligatory use of confirmatory tests justified? J Renin Angiotensin Aldosterone Syst. 2012;13(3):367-71.

496. Quinkler M, Lepenies J, Diederich S. Primary hyperaldosteronism. Exp Clin Endocrinol Diabetes. 2002;110(6):263-71.

497. Randall BR, Kraus KL, Simard MF, Couldwell WT. Cost of evaluation of patients with pituitary incidentaloma. Pituitary. 2010;13(4):383-4.

498. Vidal-Rios P, Caixas A, Cabezas R, Cajas P, Garcia-Patterson A, Rodriguez-Espinosa J, et al. Critical assessment of the efficacy and cost of the megatest in the management of pituitary tumors. [Spanish]. Endocrinologia. 1994;41(3):75-80.

499. Zidek W. Diagnosis of hypertension, an update - Efficient and cost-saving. [German]. MMW-Fortschritte der Medizin. 2004;146(19):43-6.

500. Grumbach MM, Biller BM, Braunstein GD, Campbell KK, Carney JA, Godley PA, et al. Management of the clinically inapparent adrenal mass ("incidentaloma"). Ann Intern Med. 2003;138(5):424-9.

501. Young WF, Jr. Clinical practice. The incidentally discovered adrenal mass. N Engl J Med. 2007;356(6):601-10.

502. Curtis L, Burns A. Unit Costs of Health and Social Care 2017. Canterbury: Personal Social Services Research Unit, University of Kent; 2017.

503. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second ed: Springer, New York, NY; 2009.

504. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in R.: Springer New York Heidelberg Dordrecht London; 2013.

505. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008;28(5):1-26.

506. The R Development Core Team. R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing, Vienna, Austria; 2018. Available from: https://www.R-project.org/.

507. NHS Digital. National Clinical Coding Standards OPCS-4: Terminology and Classifications Delivery Service, NHS Digital; April 2018. Available from: https://hscic.kahootz.com/gf2.ti/f/762498/33700805.1/PDF/-/NCCSOPCS420185.pdf.

508. Morera J, Reznik Y. MANAGEMENT OF ENDOCRINE DISEASE: The role of confirmatory tests in the diagnosis of primary aldosteronism. Eur J Endocrinol. 2019;180(2):R45-R58.

List of Abbreviations

Abbreviation	Definition
17-OHP	17-hydroxyprogesterone
ΔC	Incremental costs
ΔE	Incremental effects
A&E	Accident and Emergency
AC	Acromegaly
ACE	Angiotensin-converting enzyme
ACTH	Adrenocorticotropic hormone
AD	Addison's disease
ADX	Laparoscopic adrenalectomy surgery
AHP	Analytic hierarchy process
AIC	Akaike information criterion
APA	Aldosterone-producing adenoma
APC	Hospital Admitted Patient Care
ARR	Aldosterone-renin ratio
AUC	Area under the curve
AVS	Adrenal venous/vein sampling
BIC	Bayesian information criterion
BMI	Body mass index
BNF	The British National Formulary
BP	Blood pressure
BPA	Bilateral primary aldosteronism
CAH	Congenital adrenal hyperplasia
CBA	Cost-benefit analysis
CCI	Charlson Comorbidity Index
CCT	Captopril challenge/suppression test
CE	Conformité Européenne (i.e. European Conformity)
CEA	Cost-effectiveness analysis

Abbreviation	Definition
CHEERS	Consolidated Health Economic Evaluation Reporting Standards
CI(s)	Confidence interval(s)
CLAHRC	The NIHR Collaboration for Leadership in Applied Health Research and Care
CMA	Cost-minimisation analysis
CPD	Comprehensive diagnostic strategy
CPRD	Clinical Practice Research Datalink
CRD	Current registration date
CRF(s)	Case report form(s)
CS	Cushing's syndrome
CST	Cosyntropin stimulation test
CT	Computerised tomography
CTS	Clinical trial simulation
CUA	Cost-utility analysis
CVE(s)	Cardiovascular event(s)
DAC	NICE's Diagnostics Advisory Committee
DAM(s)	Decision-analytic modelling/model(s)
DAP	NICE's Diagnostics Assessment Programme
DCE(s)	Discrete choice experiment(s)
DF	Degrees of freedom
DHSC	The UK Department of Health and Social Care
DNA	Deoxyribonucleic acid
DOR	Diagnostic odds ratio
EE	Economic evaluation
EH	Essential (primary) hypertension
EHR	Electronic health record
EMR	Electronic medical record system protocol
EMTREE	Embase™ Subject Headings
ENBS	Expected net benefit of sampling
EQ-5D-3L	The EuroQoL EQ-5D health questionnaire (five dimensions; three levels)

Abbreviation	Definition
EQ-5D-5L	The EuroQoL EQ-5D health questionnaire (five dimensions; five levels)
EQ VAS	The EQ-5D visual analogue scale
ERG	Evidence Review Group
EU	European Union
EUnetHTA	The European Network for Health Technology Assessment
EVPI	Expected value of perfect information
EVPPI	Expected value of partial perfect information
EVSI	Expected value of sample information
FDR	False discovery rate
FN	False negative
FNR	False negative rate
FOR	False omission rate
FP	False positive
FPR	False positive rate
FST	Fludrocortisone suppression test
GCP	Good Clinical Practice
GH	Growth hormone
GHD	Growth hormone deficiency
GHRH	Growth hormone-releasing hormone
GLM(s)	Generalised linear model(s)
GP	General practice/practitioner
GRO	The UK General Register Office
HDD	High-dose dexamethasone suppression test
HES	Hospital Episode Statistics
HPA	Hypothalamus-pituitary-adrenal axis
HRG(s)	Healthcare Resource Group(s)
HRQoL	Health-related quality of life
HTA	Health technology assessment
ICD	The International Classification of Diseases and Related Health Problems

Abbreviation	Definition
ICER(s)	Incremental cost-effectiveness ratio(s)
ICH	The International Conference of Harmonization
IGF-1	Insulin-like growth factor 1
IMD	Index of Multiple Deprivation
iNMB(s)	Incremental net monetary benefit(s)
IPD	Individual patient data
IPSS	Bilateral inferior petrosal sinus sampling
IQR	Interquartile range
IRR	Incident rate ratio
ISAC	Independent Scientific Advisory Committee
ISPOR	The Professional Society for Health Economics and Outcomes Research
ITT	Insulin tolerance test
JCRU	Joint Clinical Research Unit
LCMS/MS	Ultrasensitive liquid chromatography tandem mass spectroscopy
LR	Likelihood ratio
LSOA	Lower layer super output area
LY(s)	Life-year(s)
MCDA	Multi-criteria decision analysis
MD(s)	Medical device(s)
MDD	The European Medical Devices Directive
MDR	The UK Medical Devices Regulations
MeSH	Medical Subject Headings
MHRA	The UK Medicines and Healthcare products Regulatory Agency
MRA(s)	Mineralocorticoid receptor antagonist(s)
MRI	Magnetic resonance imaging
MSC	Midnight serum cortisol
MSVC	Midnight salivary cortisol
MTAC	NICE's Medical Technologies Advisory Committee
MTEP	NICE's Medical Technology Evaluation Programme

Abbreviation	Definition
N/A	Not applicable
NCCAH	Non-classic congenital adrenal hyperplasia
NHS	The UK National Health Service
NICE	The UK National Institute for Health and Care Excellence
NIHR	The UK National Institute for Health Research
NMB(s)	Net monetary benefit(s)
NPV	Negative predictive value
OCP	Oral contraceptive pill
OLS	Ordinary least squares
ONS	The UK Office for National Statistics
OSLT	Oral sodium loading test
PA	Primary aldosteronism
PAC	Plasma aldosterone concentration
PCA	Prescription Cost Analysis
PICO(s)	P: Population; I: Intervention(s); C: Comparator(s); O: Outcome(s)
PPPs	Purchasing power parities
PPV	Positive predictive value
PSA	Probabilistic sensitivity analysis
PSSRU	The Personal Social Services Research Unit
QALY(s)	Quality-adjusted life-year(s)
R&D	Research and development
RCT(s)	Randomised controlled trial(s)
REC	Research Ethics Committee
RESET	Ramsey's Regression Equation Specification Error Test
RH	Resistant hypertension
ROC	Receiver-operating characteristic curve
ROI	Real options analysis
RT	Radiation therapy
Rx	Pharmacotherapy

Abbreviation	Definition
SA(s)	Sensitivity analysis/analyses
SBP	Systolic blood pressure
SD	Standard deviation
SE	Standard error
SF-6D	The Short-Form Six-Dimension algorithm
SF-12v1	The 12-item Short-Form Health Survey questionnaire (Version 1)
SF-36	The 36-item Short-Form Health Survey questionnaire
SIT	Saline infusion test
SMA	Suboptimal management approach
SRL(s)	Somatostatin receptor ligand(s)
SRT	Stereotactic radiotherapy
SSHSC	The UK Secretary of State for Health and Social Care
TAP	NICE's Technology Appraisal Programme
TC	Tertiary care
TDM	Therapeutic drug monitoring
TN	True negative
TNR	True negative rate
TP	True positive
TPR	True positive rate
TRH	Treatment-resistant hypertension
UH Bristol	University Hospitals Bristol NHS Foundation Trust
UiB	University of Bergen
UK	United Kingdom
UPA	Unilateral primary aldosteronism
URc / URt	*U-Rhythm* (current practice / theoretical practice)
USA	United States of America
UTS	Up-to-standard
VOI	Value of information analysis
WTP	Willingness-to-pay

Appendices

Appendix 1: Other Endocrine Disorders Examined in ULTRADIAN

Acromegaly

Acromegaly (AC) (pathology and clinical features first described by *Andrea Verga*, 1864, and *Pierre Marie*, 1886) is a chronic endocrine disorder with an annual prevalence of 2.8-13.7 cases per million population (451-454). AC is characterised by disproportionate skeletal, tissue and organ growth, and is caused when tumorous pituitary somatotroph cells proliferate and produce an excess amount of *growth hormone (GH)*. The circulation of GH stimulates the liver and systemic extrahepatic tissues to synthesise *insulin-like growth factor I (IGF-1)*, a hormone that has a similar molecular structure to insulin and plays an essential role in the body's physical growth, especially in childhood. Prolonged GH hypersecretion leads to IGF-1 overproduction causing symptoms such as acral overgrowth (e.g. hands), physical disfigurement (e.g. prognathism), headaches and multisystem-associated morbidities (e.g. arthritis), while it is also associated with increased mortality (41, 455-458).

Over 95% of AC patients present with a benign tumour on the pituitary gland which produces excess GH. Adenomas grow slowly in patients >50 years old and faster in younger patients, resulting in accelerated growth and gigantism (i.e. abnormal body growth). Depending on their size, smaller or larger than 1cm, these tumours are called *micro-* or *macro-adenomas*, respectively. AC is rarely caused by non-pituitary tumours (e.g. pancreas, lungs, adrenal glands) or other parts of the brain (hypothalamus). These tumours cause GH hypersecretion either by producing GH themselves (ectopically) or by producing *growth hormone-releasing hormone (GHRH)*, which induces GH secretion from the pituitary gland (41, 455-458).

The diagnosis of acromegaly relies on the demonstration of elevated or equivocal IGF-1 levels as well as dysregulated and enhanced GH release in patients with typical AC clinical symptoms (e.g. somatic growth), patients without them but who have several of the AC-associated conditions (e.g. sleep apnoea syndrome), or patients with a pituitary mass. For this purpose, dynamic biochemical testing is performed to examine whether IGF-1 levels are above an age-

and sex-adjusted reference point. The diagnosis of acromegaly is later confirmed with the use of the oral glucose tolerance test that checks whether GH is not suppressed. Magnetic resonance imaging (MRI) is then used to detect and locate the tumour's size and appearance and the brain's parasellar region. Computerised tomography (CT) can also be provided when MRI is contraindicated/unavailable, while a formal visual field testing is recommended if the tumour is shown to touch the optic chiasm (41, 455-458).

The management of acromegaly involves the evaluation of the AC-associated comorbidities and the control of IGF-1 and GH levels against age and sex matched normal/common ranges. AC is mainly treated using transsphenoidal surgery, if the tumour is resectable, which is repeated in patients with residual intrasellar disease following initial surgery. Pre-operative medications should not be provided unless the patient has severe comorbidities (e.g. heart failure). Here, somatostatin receptor ligands (SRLs) are used to reduce surgical risk (e.g. octreotide, lanreotide). After twelve weeks from surgery, IGF-1 and GH levels should be measured, while imaging testing should be used to examine the residual tumour. Patients with persistent acromegaly should be administered SRLs, or alternatively GH receptor antagonists (pegvisomant) or dopamine agonists (e.g. cabergoline) as their initial adjuvant medical therapy. If a residual tumour is still present after initial surgery and medical therapy is unavailable, unsuccessful or not tolerated, stereotactic radiotherapy (SRT) or simple radiation therapy (RT) should be used. After SRT/RT, the annual IGF-1/GH levels should be monitored since they slowly return to normal ranges (41, 455-458).

Congenital Adrenal Hyperplasia

Congenital adrenal hyperplasia (CAH) (first described by *Luigi de Crecchio*, 1865) is a family of autosomal recessive disorders with an annual incidence of 1:10,000 to 1:20,000 births (459-461). CAH is caused by the deficiency of one of the five enzymes that are required to produce cortisol in the adrenal cortex. The most common CAH form (≈95%) is characterised by the deficiency of the *steroid 21-hydroxylase cytochrome enzyme* due to mutations in the gene that encodes it. This enzyme catalyses the conversion of *17-hydroxyprogesterone (17-OHP)* to 11-deoxycortisol and progesterone to deoxycorticosterone, which are precursors of cortisol and aldosterone, respectively. Due to the loss of this enzyme's function, patients cannot produce cortisol efficiently, and consequently the adrenal cortex is stimulated by

corticotropin and synthesises large amounts of cortisol precursors, some of which are diverted to the production of sex hormones. This may result in androgen excess, which is evident through genital ambiguity in new-born girls and rapid postnatal somatic growth in both boys and girls. Additionally, 75% of patients present with aldosterone deficiency which can lead to salt loss, failure to thrive, fatal hypovolemia and shock (salt-wasting form). In rare cases (0.2%), a mild, non-classic CAH (NCCAH) form can be present, which is sometimes asymptomatic or shows signs of postnatal androgen excess because of mild-to-moderate overproduction of sex hormone precursors. Opposite to CAH, NCCAH patients can produce normal amounts of cortisol and aldosterone (43, 462-464).

For the diagnosis of CAH/NCCAH symptomatic patients, an early morning measurement of the serum 17-OHP with the use of immunoassays or liquid chromatography/tandem mass spectrometry is recommended. Furthermore, the cosyntropin stimulation test (CST) should be used to obtain a complete adrenocortical profile and distinguish 21-hydroxylase from other enzyme deficiencies. Genotyping (i.e. determination of genetic differences) is also suggested when the results after CST are equivocal or genetic counselling is needed. CAH is treated with the use of orally administered glucocorticoids (e.g. hydrocortisone) or mineralocorticoids (e.g. fludrocortisone and sodium chloride supplements). The aim here is to replace the missing hormones and reduce the undesirable effects of the hormone overproduction. When hydrocortisone is given, mineralocorticoids are normally not needed since the former can activate the mineralocorticoid receptors sufficiently. Additionally, treatment monitoring is suggested by taking hormone measurements at specific timepoints (43, 462-464).

Symptomatic NCCAH must be treated with glucocorticoids if an inappropriately early onset and rapid progression of pubic hair or bone age (in children), overt virilisation (in adolescents), or hyperandrogenism and/or infertility (in adults) is present. Asymptomatic NCCAH is usually not treated. Severely virilised females should be counselled for clitoral and perineal reconstruction, while bilateral adrenalectomy may rarely performed in selected cases (e.g. infertile females) (43, 462-464).

Growth Hormone Deficiency

Growth hormone deficiency (GHD) is an endocrine disorder with an annual prevalence of 2-3:10,000 (465-467) and is characterised by insufficient GH production, which often leads to one of its most noticeable symptoms, growth failure or short height. GHD can occur in both children and adults. Depending on the aetiology of the disease, children patients are usually grouped into two categories: a) those with organic causes, and b) those with idiopathic GHD (i.e. unknown cause). Adult patients are normally divided into three groups: a) those who had GHD in childhood; b) those who have secondary GHD due to structural lesions/trauma on the hypothalamus-pituitary; and c) those with idiopathic GHD. GHD is commonly caused by pituitary tumours, parasellar masses, or treatment of such tumours with cranial surgery or radiotherapy. Acquired GHD can be caused by traumatic brain injuries, infiltrative or granulomatous disease, or infections or haemorrhage of the central nervous system. GHD affects multiple organs and systems, makes changes in body composition (i.e. increased body fat; reduced muscle mass) and cardiovascular function, reduces strength-exercise capacity, and impairs quality of life (45, 468-471).

Since its signs and symptoms are non-specific, GHD is normally diagnosed biochemically with the use of provocative/stimulation tests. More precisely, the insulin tolerance test (ITT) and the GHRH-arginine test are mainly used to measure insulin-induced hypoglycaemia. However, the GHRH-arginine test can be ambiguous when GHD is due to hypothalamic causes (e.g. irradiation of the area). In these cases as well as when both ITT and GHRH-arginine tests are contraindicated, unavailable or impractical (e.g. patients with a history of ischemic heart disease), the glucagon stimulation test can be used. Diagnosis is recommended for patients with structural hypothalamic-pituitary lesions, history of cranial surgery or irradiation, head trauma, and pituitary hormone deficiencies. Patients who had childhood-onset GHD and have achieved their final adult height should be retested for GHD if they do not have mutations or embryopathic lesions. Idiopathic GHD is present in rare cases, and for this reason its diagnosis is made using stricter criteria (e.g. two tests before confirming diagnosis) (45, 468-471).

In patients with irreversible structural lesions and multiple hormone deficiencies as well as those with proven genetic GHD causes, a measurement of serum IGF-1 levels after at least one month without GH replacement therapy (i.e. biosynthetic recombinant human GH) is

recommended. If IGF-1 levels are normal, provocative testing is needed to confirm GHD diagnosis. If IGF-1 levels are low, GHD is confirmed and the patient undergoes GH stimulation testing to examine whether GH treatment would be beneficial. The prerequisite here is that no fasting takes place and no liver diseases or catabolic conditions (e.g. diabetes mellitus) are present. Additionally, if three or more other pituitary hormone deficiencies are present, GHD is confirmed and the use of stimulation testing is not necessary (45, 469-471).

GHD is treated with GH replacement therapy which improves body composition, exercise capacity and physical performance, skeletal integrity, and cardiac function. The dose of GH should be adjusted to the individual's gender, oestrogen status and age, and not be weight-based. Therapy should start with lower doses that will increase gradually according to clinical response, side effects and IGF-1 levels. For this reason, dose monitoring is essential. When GHD is persistent, GH therapy should be continued even after the achievement of the potential maximum adult height to ensure the maturation of the bones and muscles. GH treatment should not be administered when an active malignancy is present, while it may require the adjustment of antidiabetic drugs in patients with diabetes mellitus. Lastly, during treatment, the thyroid and adrenal function should be closely monitored (45, 468-470).

Table 44: Summary information on other endocrine diseases examined in ULTRADIAN

Endocrine Disease	Prevalence/Incidence	Key Signs and Symptoms	Diagnosis	Treatment
Acromegaly	2.8-13.7 cases per million people per year	• Disproportionate skeletal, tissue and organ growth • Physical disfigurement • Multisystem-associated morbidities	<u>Diagnosis:</u> • Dynamic biochemical testing (IGF-1 levels) <u>Confirmatory testing:</u> • Oral glucose tolerance test <u>Tumour imaging:</u> • MRI or CT (when MRI is contraindicated/unavailable) • Formal visual field testing	• Pituitary transsphenoidal surgery • SRLs used before surgery • SRLs or GH receptor antagonists or dopamine agonists as initial adjuvant medical therapy for patients with persistent disease • Stereotactic radiotherapy or simple radiation therapy if after surgery medical therapy is unavailable, unsuccessful or not tolerated

Appendix 1: Other Endocrine Disorders Examined in ULTRADIAN

Endocrine Disease	Prevalence/Incidence	Key Signs and Symptoms	Diagnosis	Treatment
Congenital adrenal hyperplasia	1:10,000 to 1:20,000 per year	• Genital ambiguity in new-born girls • Rapid postnatal somatic growth (in boys and girls)	Diagnosis: • 17-hydroxyprogesterone measurement Confirmatory testing: • 17-hydroxyprogesterone during short Synacthen stimulation test	• Orally administered glucocorticoid (and/or mineralocorticoid) therapy • Treatment monitoring by taking hormone measurements
Growth hormone deficiency	2-3:10,000 per year	• Growth failure or short height • Other pituitary hormone deficiencies	Diagnosis: • ITT and GHRH-arginine test • Glucagon stimulation test (when ITT and GHRH-arginine are contraindicated, unavailable or impractical) Confirmatory testing: • One or more of the above tests • IGF-1 immunoassay after >1 month with GH replacement therapy	• GH replacement therapy adjusted to the gender, oestrogen status and age (not weight-based) • Adjustment of antidiabetic drugs in patients with diabetes mellitus • Thyroid and adrenal function monitoring during treatment

[**Sources:** (41, 43, 45, 451-471)]

**Abbreviations:* *CT: computerised tomography; GH: growth hormone; GHRH: growth hormone-releasing hormone; IGF-1: insulin-like growth factor 1; ITT: insulin tolerance test; MRI: magnetic resonance imaging; SRLs: somatostatin receptor ligands*

Appendix 2: Systematic Review Search Strategy

Below the search strategy that was used in the systematic literature review is presented including a list of index terms as well as concepts and their synonyms, abbreviations and spelling variants that were used for its purposes. This strategy was written for Medline® and Embase™ (using the Ovid® database interface), and for the Cochrane Library.

Search Terms in Medline® (via Ovid®)

Endocrine Disorders:

1. Endocrine System/
2. Endocrine Glands/
3. Endocrine System Diseases/
4. (endocrine adj3 disorder*).mp.
5. (endocrine adj3 disease*).mp.
6. 1 or 2 or 3 or 4 or 5

Cushing's Syndrome:

7. Cushing Syndrome/
8. Pituitary ACTH Hypersecretion/
9. Cushing* syndrome*.mp.
10. Cushing* disease*.mp.
11. Cushing*.mp.
12. hypercortisolism.mp.
13. hypercortisol?emia.mp.
14. pituitary acth hypersecretion.mp.
15. Adrenocortical Hyperfunction/
16. adrenocortical hyperfunction.mp.
17. adrenal cortex hyperfuction.mp.
18. adrenal gland hyperfunction.mp.
19. hyperadrenalism.mp.
20. 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19

Acromegaly:

21. Acromegaly/
22. acromegaly.mp.
23. excess* growth hormone*.mp.
24. growth hormone excess*.mp.
25. 21 or 22 or 23 or 24

Primary Aldosteronism:

26. Hyperaldosteronism/
27. primary hyperaldosteronism.mp.
28. primary aldosteronism.mp.
29. hyperaldosteron?emia.mp.
30. aldosteronism.mp.
31. aldosteron?emia.mp.
32. Conn* syndrome*.mp.
33. (excess* adj3 aldosteron*).mp.
34. hyporeninemic hyperaldosteronism.mp.
35. 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34

Addison's Disease:

36. Addison Disease/
37. Addison* disease*.mp.
38. Adrenal Insufficiency/
39. adrenal insufficienc*.mp.
40. adrenal cortex insufficiency.mp.
41. primary adrenal insufficiency.mp
42. chronic adrenal insufficiency.mp.
43. primary adrenocortical insufficiency.mp.
44. primary adrenocortical deficiency.mp.
45. hypocortisolism.mp.
46. hypoadrenalism.mp.
47. 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46

Congenital Adrenal Hyperplasia:

48. Adrenal Hyperplasia, Congenital/
49. congenital adrenal hyperplasia.mp.
50. adrenal hyperplasia.mp.
51. Hyperandrogenism/
52. hyperandrogenism.mp.
53. Adrenogenital Syndrome/
54. adrenogenital syndrome*.mp.
55. 21 hydroxylase deficiency.mp.
56. 11 beta hydroxylase deficiency.mp.
57. 17 alpha hydroxylase deficiency.mp.
58. 48 or 49 or 50 or 51 or 52 or 53 or 54 or 55 or 56 or 57

Growth Hormone Deficiency:

59. exp Growth Hormone/df
60. growth hormone deficiency.mp.
61. exp Hypopituitarism/
62. hypopituitarism.mp.
63. panhypopituitarism.mp.
64. adult growth hormone deficiency.mp.
65. 59 or 60 or 61 or 62 or 63 or 64

All Endocrine Disorders:

66. 6 or 20 or 25 or 35 or 47 or 58 or 65

Diagnosis:

67. exp Diagnosis/
68. “Diagnostic Techniques and Procedures”/
69. exp Clinical Laboratory Techniques/
70. exp Diagnostic Imaging/
71. exp Diagnostic Techniques, Endocrine/
72. Diagnostic Tests, Routine/

73. Diagnosis, Differential/
74. diagnostic technique*.mp.
75. diagnostic procedure*.mp.
76. diagnos*.mp.
77. diagnostic*.mp.
78. exp Monitoring, Physiologic/
79. monitor*.mp.
80. screen*.mp.
81. 67 or 68 or 69 or 70 or 71 or 72 or 73 or 74 or 75 or 76 or 77 or 78 or 79 or 80

Economic Studies:

82. Economics/
83. health economic*.mp.
84. exp Economics, Hospital/
85. exp Economics, Medical/
86. Economics, Nursing/
87. Economics, Pharmaceutical/
88. Financial Management/
89. Financial Management, Hospital/
90. "Costs and Cost Analysis"/
91. "Cost Allocation"/
92. exp Resource Allocation/
93. cost* allocation*.mp.
94. resource* allocation*.mp.
95. Cost-Benefit Analysis/
96. (cost* adj2 (effective* or utilit* or benefit* or minimi* or consequence* or analys* or outcome*)).ti,ab,kf.
97. "Cost Control"/
98. Cost Savings/
99. cost* control*.mp.
100. cost* saving*.mp.
101. "cost of Illness"/

102. (cost* adj2 (illness or disease or sickness)).mp.
103. (burden? adj2 (illness or disease? or condition? or economic*)).mp.
104. “Cost Sharing”/
105. cost* sharing.mp.
106. “Deductibles and Coinsurance”/
107. Medical Savings Accounts/
108. Health Care Costs/
109. (health* adj cost*).mp.
110. Direct Service Costs/
111. Drug Costs/
112. Employer Health Costs/
113. Hospital Costs/
114. Health Expenditures/
115. Capital Expenditures/
116. “Value of Life”/
117. exp “Fees and Charges”/
118. exp Budgets/
119. budget*.ti,ab,kf.
120. (low* adj cost*).mp.
121. (high* adj cost*).mp.
122. (cost* adj estimat*).mp.
123. (cost* adj variable*).mp.
124. (unit* adj cost*).mp.
125. (economic* or cost* or costing or price* or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure* or expense* or financial or finance* or fiscal or funding).ti,ab,kf.
126. (value adj2 (money or monetary)).ti,ab,kf.
127. (out-of-pocket adj2 (payment* or expenditure* or cost* or spending or expense*)).ti,ab,kf.
128. (expenditure* adj3 (health or direct or indirect)).mp.
129. budget impact.ti,ab,kf.
130. exp models, economic/

131. economic model*.ti,ab,kf.
132. econometric model*.ti,ab,kf.
133. markov chains/
134. markov.ti,ab,kf.
135. monte carlo method/
136. monte carlo.ti,ab,kf.
137. exp Decision Theory/
138. (decision* adj2 (tree* or analys* or model* or theory)).mp.
139. exp Technology Assessment, Biomedical/
140. health technology assessment.ti,ab,kf.
141. economic evaluation*.ti,ab,kf.
142. economic analys*.ti,ab,kf.
143. 82 or 83 or 84 or 85 or 86 or 87 or 88 or 89 or 90 or 91 or 92 or 93 or 94 or 95 or 96 or 97 or 98 or 99 or 100 or 101 or 102 or 103 or 104 or 105 or 106 or 107 or 108 or 109 or 110 or 111 or 112 or 113 or 114 or 115 or 116 or 117 or 118 or 119 or 120 or 121 or 122 or 123 or 124 or 125 or 126 or 127 or 128 or 129 or 130 or 131 or 132 or 133 or 134 or 135 or 136 or 137 or 138 or 139 or 140 or 141 or 142

Endocrine, Diagnostic and Economic Studies:

144. 66 and 81 and 143

Limit for Humans:

145. Animal/ not Human/
146. 144 not 145

Limit for Time of Publication:

147. (1990* or 1991* or 1992* or 1993* or 1994* or 1995* or 1996* or 1997* or 1998* or 1999* or 2000* or 2001* or 2002* or 2003* or 2004* or 2005* or 2006* or 2007* or 2008* or 2009* or 2010* or 2011* or 2012* or 2013* or 2014* or 2015* or 2016* or 2017*).ed,dc,yr.
148. 146 and 147
149. remove duplicates from 148

Search Terms in Embase™ (via Ovid®)

Endocrine Disorders:

1. endocrine system/
2. endocrine gland/
3. endocrine tumor/
4. endocrine disease/
5. (endocrine adj3 disorder*).mp.
6. (endocrine adj3 disease*).mp.
7. 1 or 2 or 3 or 4 or 5 or 6

Cushing's Syndrome:

8. Cushing disease/
9. Cushing syndrome/
10. Cushing* disease*.mp.
11. Cushing* syndrome*.mp.
12. Cushing*.mp.
13. hypercortisolism/
14. hypercortisolism.mp.
15. hypercortisol?emia.mp.
16. pituitary acth hypersecretion.mp.
17. adrenal cortex hyperfunction/
18. adrenal cortex hyperfunction.mp.
19. adrenocortical hyperfunction.mp.
20. adrenal gland hyperfunction.mp.
21. hyperadrenalism.mp.
22. moon face/
23. 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22

Acromegaly:

24. acromegaly/
25. acromegaly.mp.
26. excess* growth hormone*.mp.
27. growth hormone excess*.mp.
28. 24 or 25 or 26 or 27

Primary Aldosteronism:

29. hyperaldosteronism/
30. primary hyperaldosteronism/
31. primary hyperaldosteronism.mp.
32. primary aldosteronism.mp.
33. hyperaldosteron?emia.mp.
34. aldosteronism.mp.
35. aldosteron?emia.mp.
36. Conn* syndrome*.mp.
37. (excess* adj3 aldosteron*).mp.
38. hyporeninemic hyperaldosteronism.mp
39. 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38

Addison's Disease:

40. Addison disease/
41. Addison* disease*.mp.
42. adrenal insufficiency/
43. adrenal insufficienc*.mp.
44. adrenal cortex insufficiency/
45. adrenal cortex insufficiency.mp.
46. primary adrenal insufficiency.mp.
47. chronic adrenal insufficiency.mp.
48. primary adrenocortical insufficiency.mp.
49. primary adrenocortical deficiency.mp.
50. hypocortisolism.mp.

51. hypoadrenalism.mp.
52. 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47 or 48 or 49 or 50 or 51

Congenital Adrenal Hyperplasia:

53. congenital adrenal hyperplasia/
54. congenital adrenal hyperplasia.mp.
55. adrenal hyperplasia/
56. adrenal hyperplasia.mp.
57. hyperandrogenism/
58. hyperandrogenism.mp.
59. adrenogenital syndrome*.mp.
60. steroid 21 monooxygenase deficiency/
61. 21 hydroxylase deficiency.mp.
62. 11 beta hydroxylase deficiency.mp.
63. 17 alpha hydroxylase deficiency.mp.
64. 53 or 54 or 55 or 56 or 57 or 58 or 59 or 60 or 61 or 62 or 63

Growth Hormone Deficiency:

65. growth hormone deficiency/
66. pituitary dwarfism/
67. growth hormone deficiency.mp.
68. hypopituitarism/
69. hypopituitarism.mp.
70. panhypopituitarism.mp.
71. adult growth hormone deficiency.mp.
72. 65 or 66 or 67 or 68 or 69 or 70 or 71

All Endocrine Disorders:

73. 7 or 23 or 28 or 39 or 52 or 64 or 72

Diagnosis:

74. exp diagnosis/
75. exp diagnostic procedure/
76. differential diagnosis/
77. exp diagnostic imaging/
78. exp diagnostic test/
79. exp endocrine system examination/
80. diagnostic technique*.mp.
81. diagnostic procedure*.mp.
82. diagnos*.mp.
83. exp diagnostic equipment/
84. diagnostic*.mp.
85. exp monitoring/
86. exp monitor/
87. monitor*.mp.
88. exp screening/
89. screen*.mp.
90. 74 or 75 or 76 or 77 or 78 or 79 or 80 or 81 or 82 or 83 or 84 or 85 or 86 or 87 or 88 or 89

Economic Studies:

91. exp economics/
92. exp health economics/
93. exp pharmacoeconomics/
94. health economic*.mp.
95. “cost”/
96. “cost benefit analysis”/
97. “cost effectiveness analysis”/
98. “cost minimization analysis”/
99. “cost utility analysis”/
100. “cost of illness”/
101. “program cost effectiveness”/

102. “global disease burden/
103. (cost* adj2 (effective* or utilit* or benefit* or minimi* or consequence* or analys* or outcome*)).ti,ab,kw.
104. (cost* adj2 (illness or disease or sickness)).mp.
105. (burden? adj2 (illness or disease? or condition? or economic*)).mp.
106. economic aspect/
107. financial management/
108. “health care cost”/
109. health care financing/
110. exp “hospital cost”/
111. “nursing cost”/
112. (health?care adj cost*).mp.
113. “cost control”/
114. cost* control*.mp.
115. cost* allocation*.mp.
116. cost* saving*.mp.
117. cost* sharing.mp.
118. exp resource management/
119. resource* allocation*.mp.
120. “drug cost”/
121. exp fee/
122. budget/
123. budget*.ti,ab,kw.
124. (low* adj cost*).mp.
125. (high* adj cost*).mp.
126. (cost* adj estimat*).mp.
127. (cost* adj variable*).mp.
128. (unit* adj cost*).mp.
129. (economic* or cost* or costing or price* or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure* or expense* or financial or finance* or fiscal or funding).ti,ab,kw.
130. (value adj2 (money or monetary)).ti,ab,kw.

131. (out-of-pocket adj2 (payment? or expenditure? or cost? or spending or expense?)).ti,ab,kw.
132. (expenditure? adj3 (health or direct or indirect)).mp.
133. budget impact.ti,ab,kw.
134. economic model/
135. statistical model/
136. economic model*.ti,ab,kw.
137. econometric model*.ti,ab,kw.
138. markov chain/
139. markov.ti,ab,kw.
140. monte carlo method/
141. monte carlo.ti,ab,kw.
142. decision theory/
143. "decision tree"/
144. (decision* adj2 (tree* or analys* or model* or theory)).mp
145. biomedical technology assessment/
146. health technology assessment.ti,ab,kw.
147. economic evaluation/
148. economic evaluation*.ti,ab,kw.
149. economic analys*.ti,ab,kw.
150. 91 or 92 or 93 or 94 or 95 or 96 or 97 or 98 or 99 or 100 or 101 or 102 or 103 or 104 or 105 or 106 or 107 or 108 or 109 or 110 or 111 or 112 or 113 or 114 or 115 or 116 or 117 or 118 or 119 or 120 or 121 or 122 or 123 or 124 or 125 or 126 or 127 or 128 or 129 or 130 or 131 or 132 or 133 or 134 or 135 or 136 or 137 or 138 or 139 or 140 or 141 or 142 or 143 or 144 or 145 or 146 or 147 or 148 or 149

Endocrine, Diagnostic and Economic Studies:

151. 73 and 90 and 150

Limit for Humans:

152. Animal/ not Human/
153. 151 not 152

Limit for Time of Publication:

154. (1990* or 1991* or 1992* or 1993* or 1994* or 1995* or 1996* or 1997* or 1998* or 1999* or 2000* or 2001* or 2002* or 2003* or 2004* or 2005* or 2006* or 2007* or 2008* or 2009* or 2010* or 2011* or 2012* or 2013* or 2014* or 2015* or 2016* or 2017*).ed,dc,yr.
155. 153 and 154
156. remove duplicates from 155

Search Terms in the Cochrane Library

Endocrine Disorders:

#1. MeSH descriptor: [Endocrine System] this term only
#2. MeSH descriptor: [Endocrine Glands] this term only
#3. MeSH descriptor: [Endocrine System Diseases] this term only
#4. endocrine near/3 disorder*
#5. endocrine near/3 disease*
#6. #1 or #2 or #3 or #4 or #5

Cushing's Syndrome:

#7. MeSH descriptor: [Cushing Syndrome] this term only
#8. MeSH descriptor: [Pituitary ACTH Hypersecretion] this term only
#9. Cushing* syndrome*
#10. Cushing* disease*
#11. Cushing*
#12. hypercortisolism
#13. hypercortisol*emia
#14. pituitary acth hypersecretion
#15. adrenal cortex hyperfunction
#16. MeSH descriptor: [Adrenocortical Hyperfunction] this term only
#17. adrenocortical hyperfunction
#18. adrenal gland hyperfunction

#19. hyperadrenalism

#20. #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18 or #19

Acromegaly:

#21. MeSH descriptor: [Acromegaly] this term only

#22. acromegaly

#23. excess* growth hormone*

#24. growth hormone* excess*

#25. #21 or #22 or #23 or #24

Primary Aldosteronism:

#26. MeSH descriptor: [Hyperaldosteronism] this term only

#27. primary hyperaldosteronism

#28. primary aldosteronism

#29. aldosteronism

#30. aldosteron*emia

#31. Conn syndrome* or Conns syndrome* or Conn's syndrome*

#32. excess* near/3 aldosteron*

#33. #26 or #27 or #28 or #29 or #30 or #31 or #32

Addison's Disease:

#34. MeSH descriptor: [Addison Disease] this term only

#35. Addison* disease*

#36. MeSH descriptor: [Adrenal Insufficiency] this term only

#37. adrenal insufficienc*

#38. adrenal cortex insufficiency

#39. primary adrenal insufficiency

#40. chronic adrenal insufficiency

#41. primary adrenocortical insufficiency

#42. primary adrenocortical deficiency

#43. hypocortisolism

#44. hypoadrenalism

#45. #34 or #35 or #36 or #37 or #38 or #39 or #40 or #41 or #42 or #43 or #44

Congenital Adrenal Hyperplasia:

#46. MeSH descriptor: [Adrenal Hyperplasia, Congenital] this term only
#47. congenital adrenal hyperplasia
#48. adrenal hyperplasia
#49. MeSH descriptor: [Hyperandrogenism] this term only
#50. hyperandrogenism
#51. MeSH descriptor: [Adrenogenital Syndrome] explode all trees
#52. adrenogenital syndrome*
#53. 21 hydroxylase deficiency
#54. 11 beta hydroxylase deficiency
#55. 17 alpha hydroxylase deficiency
#56. #46 or #47 or #48 or #49 or #50 or #51 or #52 or #53 or #54 or #55

Growth Hormone Deficiency:

#57. MeSH descriptor: [Dwarfism, Pituitary] this term only
#58. growth hormone near/6 deficiency
#59. MeSH descriptor: [Hypopituitarism] explode all trees
#60. hypopituitarism
#61. panhypopituitarism
#62. adult near/3 growth hormone near/6 deficiency
#63. #57 or #58 or #59 or #60 or #61 or #62

All Endocrine Disorders:

#64. #6 or #20 or #25 or #33 or #45 or #56 or #63

Diagnosis:

#65. MeSH descriptor: [Diagnosis] explode all trees
#66. MeSH descriptor: [Diagnostic Techniques and Procedures] this term only
#67. MeSH descriptor: [Diagnostic Techniques, Endocrine] explode all trees
#68. MeSH descriptor: [Clinical Laboratory Techniques] explode all trees

#69. MeSH descriptor: [Diagnostic Imaging] explode all trees
#70. MeSH descriptor: [Diagnostic Tests, Routine] this term only
#71. MeSH descriptor: [Diagnostic Equipment] explode all trees
#72. MeSH descriptor: [Diagnosis, Differential] this term only
#73. diagnostic technique*
#74. diagnostic procedure*
#75. diagnos*
#76. diagnostic*
#77. MeSH descriptor: [Monitoring, Physiologic] explode all trees
#78. monitor*
#79. screen*
#80. #65 or #66 or #67 or #68 or #69 or #70 or #71 or #72 or #73 or #74 or #75 or #76 or #77 or #78 or #79

Economic Studies:

#81. MeSH descriptor: [Economics] this term only
#82. health economic*
#83. MeSH descriptor: [Costs and Cost Analysis] this term only
#84. MeSH descriptor: [Cost Allocation] this term only
#85. cost* allocation*
#86. MeSH descriptor: [Cost-Benefit Analysis] this term only
#87. cost* near/2 (effective* or utilit* or benefit* or minimi* or consequence* or analys* or outcome*)
#88. MeSH descriptor: [Cost Control] this term only
#89. MeSH descriptor: [Cost Savings] this term only
#90. cost* control*
#91. cost* saving*
#92. MeSH descriptor: [Health Expenditures] this term only
#93. MeSH descriptor: [Capital Expenditures] this term only
#94. MeSH descriptor: [Cost Sharing] explode all trees
#95. cost* sharing
#96. MeSH descriptor: [Health Care Costs] this term only

#97. health near/1 cost*

#98. MeSH descriptor: [Direct Service Costs] this term only

#99. MeSH descriptor: [Drug Costs] this term only

#100. MeSH descriptor: [Employer Health Costs] this term only

#101. MeSH descriptor: [Hospital Costs] this term only

#102. MeSH descriptor: [Cost of Illness] this term only

#103. cost* near/2 (illness or disease or sickness)

#104. burden* near/2 (illness or disease* or condition* or economic*)

#105. MeSH descriptor: [Economics, Hospital] explode all trees

#106. MeSH descriptor: [Economics, Medical] explode all trees

#107. MeSH descriptor: [Economics, Nursing] this term only

#108. MeSH descriptor: [Economics, Pharmaceutical] this term only

#109. MeSH descriptor: [Fees and Charges] explode all trees

#110. MeSH descriptor: [Financial Management] this term only

#111. MeSH descriptor: [Budgets] explode all trees

#112. budget*

#113. MeSH descriptor: [Financial Management, Hospital] this term only

#114. MeSH descriptor: [Health Care Economics and Organizations] this term only

#115. MeSH descriptor: [Technology Assessment, Biomedical] explode all trees

#116. health technology assessment

#117. economic evaluation*

#118. economic analys*

#119. MeSH descriptor: [Resource Allocation] explode all trees

#120. resource* allocation*

#121. value near/2 (money or monetary)

#122. out-of-pocket adj2 (payment* or expenditure* or cost* or spending or expense*)

#123. expenditure* near/3 (health or direct or indirect)

#124. MeSH descriptor: [Value of Life] this term only

#125. MeSH descriptor: [Healthcare Financing] this term only

#126. low* near/1 cost*

#127. high* near/1 cost*

#128. cost* near/1 estimat*

#129. cost* near/1 variable*

#130. unit* near/1 cost*

#131. economic* or cost* or costing or price* or pricing or pharmacoeconomic* or pharmaco-economic* or expenditure* or expense* or financial or finance* or fiscal or funding

#132. budget impact

#133. MeSH descriptor: [Models, Economic] explode all trees

#134. economic model*

#135. econometric model*

#136. MeSH descriptor: [Markov Chains] this term only

#137. Markov (model* or chain*)

#138. MeSH descriptor: [Monte Carlo Method] this term only

#139. monte carlo method

#140. MeSH descriptor: [Decision Theory] explode all trees

#141. decision (tree*or analys* or model* or theory)

#142. #81 or #82 or #83 or #84 or #85 or #86 or #87 or #88 or #89 or #90 or #91 or #92 or #93 or #94 or #95 or #96 or #97 or #98 or #99 or #100 or #101 or #102 or #103 or #104 or #105 or #106 or #107 or #108 or #109 or #110 or #111 or #112 or #113 or #114 or #115 or #116 or #117 or #118 or #119 or #120 or #121 or #122 or #123 or #124 or #125 or #126 or #127 or #128 or #129 or #130 or #131 or #132 or #133 or #134 or #135 or #136 or #137 or #138 or #139 or #140 or #141

Endocrine, Diagnostic and Economic Studies:

#143. #64 and #80 and #142

#144. #64 and #80 and #142

Publication Year from Jan 1990 to May 2017

Appendix 3: Systematic Review Data Extraction Form

Table 45: Systematic review data extraction form

Item	Information Extracted
First Author	
Publication Year	
Population(s)/Disease Group(s)	
Study Design	
Type of Economic Analysis	
Perspective(s)	
Study Objective(s)	
Location(s)/Setting(s)	
Intervention(s)	
Primary Outcome Measure(s)	
Secondary Outcome Measure(s)	
Analytical Method(s)	
Primary Results	
Secondary Results	
Conclusions	

[**Source:** (176)]

Appendix 4: Excluded Studies from the Systematic Review

		Study	Ambrosi et al. (1997) (472)	Brooke et al. (2003) (473)	Carr et al. (2004) (474)	Craig et al. (2011) (475)	Danilowicz et al. (2016) (476)	Doppman et al. (1997) (477)	Durham et al. (2003) (478)	Duskova et al. (2016) (479)	Evans et al. (1995) (480)	Gibney et al. (2008) (481)	Gross et al. (1999) (482)	Gunnala et al. (2015) (483)	Isidori et al. (2003) (484)	Jabar et al. (2009) (485)	Juul et al. (1999) (486)	Kievit et al. (2000) (487)	Kirkland et al. (1996) (48)	Knutzen et al. (2006) (49)	Kohek et al. (1998) (488)	Kumar et al. (2014) (489)	Lee et al. (2002) (490)	Lesen et al. (2017) (491)	Lichtenauer et al. (2016) (492)	Lin et al. (2005) (493)	Livingston et al. (2015) (494)	Mysliwiec et al. (2012) (495)	Puig et al. (1999) (52)	Quinkler et al. (2002) (496)	Randall et al. (2010) (497)	Vidal-Rios et al. (1994) (498)	Zidek et al. (2004) (499)	Total
Reason for Exclusion	1	Published before 1990																															0
	2	No human participants							1				1																	1			1	4
	3	No diseases												1				1					1				1						1	5
	4	No diagnosis		1		1	1										1	1						1									1	7
	5	No costs		1	1		1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1		1		1	1	1	1	26
	6	No economic evaluation	1	1	1		1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1		1	1	1	1	1	1	28
	7	Not English																															1	1
		Number of Reasons	1	3	2	1	3	2	3	2	2	2	3	3	2	2	3	2	2	2	2	2	3	3	2	2	1	2	1	3	2	2	6	

Appendix 5: Studies that Nearly Met the Systematic Review Inclusion Criteria

Several studies were excluded from the review although they nearly met the eligibility criteria. Below are the reasons of their exclusion from this systematic review:

Craig et al. (2011) (475)

This is a very good economic evaluation study for the growth monitoring of short stature. This paper was excluded because the authors did not use a radiological or laboratory test in order to conduct the screening (criterion 4). Instead, they were measuring the height of the children and how fast it changed. Additionally, the study was more interested in monitoring short stature and not specifically growth hormone deficiency (GHD).

Gibney et al. (2008) (481)

This study seemed to meet all seven eligibility criteria apart from criteria 5 (measuring costs) and 6 (containing a partial or full economic evaluation). More precisely, this is an English-written paper, published after 1990, that discussed human patients diagnosed with hypopituitarism (including GHD). It is also a study that compared two laboratory tests: overnight metyrapone test / IGF-I measurement and insulin tolerance test (ITT). In terms of the analysis that was performed, this study could be characterised as a cost-effectiveness analysis measuring costs and number of cases diagnosed for each one of the two tests. However, when it comes to effectiveness, there was only a descriptive comparison of the number of cases that were diagnosed when using one of the two tests, while when it comes to costs, these were not measured using monetary units but only by comparing the cost-savings in the form of percentages. Therefore, if this study was to be considered as a partial economic evaluation study, it would not include a proper cost description or cost analysis, whereas if the study was to be considered as a full economic evaluation, it would not include a proper cost-effectiveness analysis.

Jabar et al. (2009) (485)

As with the study above, this study seemed to meet all seven eligibility criteria apart from criteria 5 and 6. More specifically, this is an English-written study, published after 1990, that included human patients with GHD. This study compared two different ways of using the ITT: one using six samples (as in clinical practice) and one using four samples. In terms of the analysis performed, this study could be characterised as a cost-effectiveness analysis measuring costs and growth hormone (GH) levels. However, in the study it was not very clear how GH levels were compared between the two procedures, and when it comes to costs, these were not measured using monetary units but only compared using the percentages of cost reduction. Therefore, if this study was to be considered as a partial economic evaluation study, it would not include a proper outcome comparison, cost description or cost analysis. If this study was to be considered as a full economic evaluation, it would not include a proper cost-effectiveness analysis.

Kievit et al. (2000) (487)

This is a good economic evaluation study that examined different strategies for identifying the different causes of adrenal incidentaloma. The reason for excluding this paper was that firstly, no clear diagnostic/monitoring test was examined. Moreover, adrenal incidentalomas are adrenal masses that are normally found after using a scanning test (e.g. computerised tomography, magnetic resonance imaging). These masses can be either benign or malignant. In the former case, adrenal incidentalomas can be related to one or more of the diseases that were examined in this review (e.g. Cushing's syndrome, CS), while in the latter case, they are related to cancer. However, the reason for their diagnosis is mainly to divide them into one of these two categories. Whether a diagnosis of one of the diseases that were examined in this review will follow highly depends on whether other signs and symptoms are present (500, 501).

Livingston et al. (2015) (494)

This is a good economic evaluation study for the diagnosis of hypopituitarism. This study was excluded because it was more interested in another form of hypopituitarism (i.e. thyroid hormone deficiency) instead of the cause that is examined in the *ULTRADIAN* study (i.e. GHD).

Puig et al. (1999) (52)

This study seemed to meet all seven eligibility criteria apart from criterion 6. This is an English-written study, published after 1990, that included human CS patients. This study compared the combinations of different laboratory tests for CS in order to find the best approach. In terms of the analysis performed, this study could be characterised as a cost-effectiveness analysis measuring costs in US dollars and outcomes in test performance/results. Although test results and costs were described, it is not clear how the conclusions of this study had been made. There was no proper analysis comparing the alternative in terms of costs and health outcomes, just a pure description of them.

Appendix 6: Quality Assessment of the Included Economic Evaluations

Table 46: Quality assessment of the economic evaluations conducted in the included studies

Question*	Ben-Shlomo et al. (2016) (183)	León-Justel et al. (2011) (50)	Midgette et al. (1995) (51)	Dekkers et al. (2016) (184)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
1. Was a well-defined question posed in answerable form?	P	P	Y	Y	Y	Y	P
1.1. Did the study examine both costs and effects of the service(s) or programme(s)?	N	N	Y	Y	Y	Y	N
1.2. Did the study involve a comparison of alternatives?	Y	Y	Y	Y	Y	Y	Y
1.3. Was a viewpoint for the analysis stated and was the study placed in any particular decision-making context?	P	Y	Y	P	Y	Y	P
2. Was a comprehensive description of the competing alternatives given (that is, can you tell who did what to whom, where, and how often)?	P	Y	Y	Y	Y	Y	Y
2.1. Were there any relevant alternatives omitted?	NK	Y	Y	P	P	P	P
2.2. Was (Should) a *do-nothing* alternative (be) considered?	N	N	N	N	Y	Y	Y
3. Was the effectiveness of the programmes or services established?	N	N	Y	Y	Y	Y	N
3.1. Was this done through a randomised controlled clinical trial? If so, did the trial protocol reflect what would happen in regular practice?	N	N	N	Y	N	N	N

Appendix 6: Quality Assessment of the Included Economic Evaluations

Question*	Ben-Shlomo et al. (2016) (183)	León-Justel et al. (2011) (50)	Midgette et al. (1995) (51)	Dekkers et al. (2016) (184)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
3.2. Were effectiveness data collected and summarised through a systematic overview of clinical studies? If so, were the search strategy and rules for inclusion or exclusion outlined?	N	N	P	N	P	P	N
3.3. Were observational data or assumptions used to establish effectiveness? If so, what are the potential biases in results?	N	N	P	N	Y	Y	N
4. Were all the important and relevant costs and consequences for each alternative identified?	**N**	**P**	**NK**	**Y**	**Y**	**P**	**P**
4.1. Was the range wide enough for the research question at hand?	N	P	NK	Y	Y	P	P
4.2. Did it cover all relevant viewpoints? (Possible viewpoints include the community or social viewpoint, and those of patients and third-party payers. Other viewpoints may also be relevant depending upon the particular analysis).	N	Y	Y	Y	Y	Y	Y
4.3. Were the capital costs, as well as operating costs, included?	N	Y	NK	P	Y	P	P
5. Were costs and consequences measured accurately in appropriate physical units (for example, hours of nursing time, number of physician visits, lost work-days, gained life-years)?	**P**	**P**	**NK**	**P**	**P**	**P**	**P**
5.1. Were the sources of resource utilisation described and justified?	P	Y	P	Y	Y	Y	Y
5.2. Were any of the identified items omitted from measurement? If so, does this mean that they carried no weight in the subsequent analysis?	P	P	NK	NK	NK	Y	P

Appendix 6: Quality Assessment of the Included Economic Evaluations

Question*	Ben-Shlomo et al. (2016) (183)	León-Justel et al. (2011) (50)	Midgette et al. (1995) (51)	Dekkers et al. (2016) (184)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
5.3. Were there any special circumstances (for example, joint use of resources) that made measurement difficult? Were these circumstances handled appropriately?	P	NK	Y	Y	Y	Y	Y
6. Were the costs and consequences valued credibly?	**P**	**P**	**P**	**Y**	**Y**	**P**	**P**
6.1. Were the sources of all values clearly identified? (Possible sources include market values, patient or client preferences and views, policy-makers' views, and health professionals' judgements).	P	P	P	Y	Y	Y	P
6.2. Were market values employed for changes involving resources gained or depleted?	P	Y	P	Y	Y	P	P
6.3. Where market values were absent (for example, volunteer labour) or market values did not reflect actual values (such as clinic space donated at a reduced rate), were adjustments made to approximate market values?	P	N	NK	P	P	N	NK
6.4. Was the valuation of consequences appropriate for the question posed (that is, has the appropriate type or types of analysis – cost-effectiveness, cost-utility, cost-benefit – been selected)?	N	N	Y	Y	Y	Y	N
7. Were costs and consequences adjusted for differential timing?	**N**	**N**	**N**	**N**	**Y**	**Y**	**N**
7.1. Were costs and consequences that occur in the future 'discounted' to their present values?	N	N	N	N	Y	Y	N
7.2. Was there any justification given for the discount rate used?	N	N	N	N	Y	Y	N

Appendix 6: Quality Assessment of the Included Economic Evaluations

Question*	Ben-Shlomo et al. (2016) (183)	León-Justel et al. (2011) (50)	Midgette et al. (1995) (51)	Dekkers et al. (2016) (184)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
8. Was an incremental analysis of costs and consequences of alternatives performed?	**N**	**P**	**Y**	**Y**	**Y**	**Y**	**P**
8.1. Were the additional (incremental) costs generated by one alternative over another compared to the additional effects, benefits or utilities generated?	N	N	Y	Y	Y	Y	P
9. Was allowance made for uncertainty in the estimates of costs and consequences?	**N**	**P**	**Y**	**Y**	**Y**	**Y**	**P**
9.1. If patient-level data on costs or consequences were available, were appropriate statistical analyses performed?	P	Y	N	Y	N	Y	P
9.2. If a sensitivity analysis was employed, was justification provided for the ranges or distributions of values (for key study parameters), and the form of sensitivity analysis used?	N	P	Y	Y	Y	Y	P
9.3. Were the conclusions of the study sensitive to uncertainty in the results, as quantified by the statistical and/or sensitivity analysis?	N	N	Y	N	P	Y	Y
10. Did the presentation and discussion of study results include all issues of concern to users?	**P**	**P**	**Y**	**Y**	**Y**	**Y**	**Y**
10.1. Were the conclusions of the analysis based on some overall index or ratio of costs to consequences (for example, cost-effectiveness ratio)? If so, was the index interpreted intelligently or in a mechanistic fashion?	N	N	P	Y	Y	P	P

Appendix 6: Quality Assessment of the Included Economic Evaluations

Question*	Ben-Shlomo et al. (2016) (183)	León-Justel et al. (2011) (50)	Midgette et al. (1995) (51)	Dekkers et al. (2016) (184)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
10.2. Were the results compared with those of others who have investigated the same question? If so, were allowances made for potential differences in study methodology?	N	P	P	Y	Y	Y	Y
10.3. Did the study discuss the generalisability of the results to other settings and patient/client groups?	P	P	Y	Y	Y	Y	Y
10.4. Did the study allude to, or take account of, other important factors in the choice or decision under consideration (for example, distribution of costs and consequences, or relevant ethical issues)?	N	N	Y	Y	Y	Y	P
10.5. Did the study discuss issues of implementation, such as the feasibility of adopting the 'preferred' programme given existing financial or other constraints, and whether any freed resources could be redeployed to other worthwhile programmes?	P	Y	Y	Y	Y	P	P
Total Number of 'Yes' Responses**	**1**	**10**	**19**	**26**	**31**	**27**	**10**
Total Number of 'Partially' Responses**	**15**	**13**	**8**	**5**	**5**	**10**	**19**
Total Number of 'No' Responses**	**22**	**15**	**6**	**7**	**2**	**2**	**9**
Total Number of 'Not Known' Responses**	**1**	**1**	**6**	**1**	**1**	**0**	**1**

[**Source:** Table developed based on the *Drummond et al. (2005)* checklist (17)]

**Scale*: *Yes (Y); Partially (P); No (N); Not known (NK)*

***The total number of each response is only given for reference and does not reflect an actual quality score since the different items of the checklist are not of equal importance.*

Appendix 7: Quality of Reporting for the Included Studies

Table 47: Quality of reporting for the included studies

#	Selection/Item	Recommendation*	Ben-Shlomo, et al. (2016) (183)	León-Justel, et al. (2011) (50)	Midgette, et al. (1995) (51)	Dekkers, et al. (2016) (184)	Lubitz, et al. (2015) (185)	Sato, et al. (2015) (186)	Velasco, et al. (2015) (187)
Title and Abstract									
1	Title	Identify the study as an economic evaluation, or use more specific terms such as "cost-effectiveness analysis" and describe the interventions compared.	N	Y	P Y	N	Y	Y	Y
2	Abstract	Provide a structured summary of objectives, perspective, setting, methods (including study design and inputs), results (including base-case and uncertainty analyses), and conclusions.	P	Y	P	Y	Y	Y	P
Introduction									
3	Background and objectives	Provide an explicit statement of the broader context for the study. Present the study question and its relevance for health policy or practice decisions.	Y	Y	Y	Y	Y	Y	Y

Appendix 7: Quality of Reporting for the Included Studies

#	Selection/Item	Recommendation*	Ben-Shlomo, et al. (2016) (183)	León-Justel, et al. (2011) (50)	Midgette, et al. (1995) (51)	Dekkers, et al. (2016) (184)	Lubitz, et al. (2015) (185)	Sato, et al. (2015) (186)	Velasco, et al. (2015) (187)
Methods									
4	Target population and subgroups	Describe characteristics of the base-case population and subgroups analysed including why they were chosen.	PN	Y	P	Y	Y	Y	Y
5	Setting and location	State relevant aspects of the system(s) in which the decision(s) need(s) to be made.	Y	Y	P	Y	P Y	P Y	Y
6	Study perspective	Describe the perspective of the study and relate this to the costs being evaluated.	N	Y	Y	PY	Y	Y	N
7	Comparators	Describe the interventions or strategies being compared and state why they were chosen.	Y	Y	Y	Y	Y	Y	Y
8	Time horizon	State the time horizon(s) over which costs and consequences are being evaluated and say why appropriate.	N	Y	N	Y	Y	Y	N
9	Discount rate	Report the choice of discount rate(s) used for costs and outcomes and say why appropriate.	N	N	N	N	Y	Y	N
10	Choice of health outcomes	Describe what outcomes were used as the measure(s) of benefit in the evaluation and their relevance for the type of analysis performed.	N	N	P	Y	Y	Y	N
11a	Measurement of effectiveness	*Single study-based estimates:* Describe fully the design features of the single effectiveness study and why the single study was a sufficient source of clinical effectiveness data.	PN	N	N/A	Y	N/A	N/A	P

Appendix 7: Quality of Reporting for the Included Studies

#	Selection/Item	Recommendation*	Ben-Shlomo, et al. (2016) (183)	León-Justel, et al. (2011) (50)	Midgette, et al. (1995) (51)	Dekkers, et al. (2016) (184)	Lubitz, et al. (2015) (185)	Sato, et al. (2015) (186)	Velasco, et al. (2015) (187)
11b		*Synthesis-based estimates:* Describe fully the methods used for identification of included studies and synthesis of clinical effectiveness data.	N/A	N/A	N	N/A	Y	Y	PN
12	Measurement and valuation of preference-based outcomes	If applicable, describe the population and methods used to elicit preferences for outcomes.	N/A	N/A	N/A	Y	Y	N/A	N/A
13a	Estimating resources and costs	*Single study-based economic evaluation:* Describe approaches used to estimate resource use associated with the alternative interventions. Describe primary or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.	PN	PY	N/A	P	N/A	N/A	N/A
13b		*Model-based economic evaluation:* Describe approaches and data sources used to estimate resource use associated with model health states. Describe primary or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.	N/A	N/A	P	N/A	Y	PY	PY

Appendix 7: Quality of Reporting for the Included Studies

#	Selection/Item	Recommendation*	Ben-Shlomo, et al. (2016) (183)	León-Justel, et al. (2011) (50)	Midgette, et al. (1995) (51)	Dekkers, et al. (2016) (184)	Lubitz, et al. (2015) (185)	Sato, et al. (2015) (186)	Velasco, et al. (2015) (187)
14	Currency, price date, and conversion	Report the dates of the estimated resource quantities and unit costs. Describe methods for adjusting estimated unit costs to the year of reported costs if necessary. Describe methods for converting costs into a common currency base and the exchange rate.	N	PN	P	PY	PY	PY	PY
15	Choice of model	Describe and give reasons for the specific type of decision-analytic model used. Providing a figure to show model structure is strongly recommended.	N/A	N/A	Y	N/A	Y	Y	Y
16	Assumptions	Describe all structural or other assumptions underpinning the decision-analytic model.	N/A	N/A	Y	N/A	Y	N	Y
17	Analytic methods	Describe all analytic methods supporting the evaluation. This could include methods for dealing with skewed, missing or censored data; extrapolation methods; methods for pooling data; approaches to validate or make adjustments (e.g. half-cycle corrections) to a model; and methods for handling population heterogeneity and uncertainty.	P	Y	P	Y	Y	Y	Y

Appendix 7: Quality of Reporting for the Included Studies

#	Selection/Item	Recommendation*	Ben-Shlomo, et al. (2016) (183)	León-Justel, et al. (2011) (50)	Midgette, et al. (1995) (51)	Dekkers, et al. (2016) (184)	Lubitz, et al. (2015) (185)	Sato, et al. (2015) (186)	Velasco, et al. (2015) (187)
Results									
18	Study parameters	Report the values, ranges, references, and if used, probability distributions for all parameters. Report reasons or sources for distributions used to represent uncertainty where appropriate. Providing a table to show the input values is strongly recommended.	P	PY	N	Y	P	Y	Y
19	Incremental costs and outcomes	For each intervention, report mean values for the main categories of estimated costs and outcomes of interest as well as mean differences between the comparator groups. If applicable, report incremental cost-effectiveness ratios.	PY	PY	PY	Y	Y	Y	Y
20a	Characterising uncertainty	*Single study-based economic evaluation:* Describe the effects of sampling uncertainty for the estimated incremental cost, incremental effectiveness and incremental cost-effectiveness, together with the impact of methodological assumptions (such as discount rate, study perspective).	N	P	N/A	Y	N/A	N/A	N/A
20b		*Model-based economic evaluation:* Describe the effects on the results of uncertainty for all input parameters, and uncertainty related to the structure of the model and assumptions.	N/A	N/A	Y	N/A	Y	Y	Y

Appendix 7: Quality of Reporting for the Included Studies

#	Selection/Item	Recommendation*	Ben-Shlomo, et al. (2016) (183)	León-Justel, et al. (2011) (50)	Midgette, et al. (1995) (51)	Dekkers, et al. (2016) (184)	Lubitz, et al. (2015) (185)	Sato, et al. (2015) (186)	Velasco, et al. (2015) (187)
21	Characterising heterogeneity	If applicable, report differences in costs, outcomes or cost-effectiveness that can be explained by variations between subgroups of patients with different baseline characteristics or other observed variability in effects that are not reducible by more information.	N	PN	PY	PY	P	P	P
Discussion									
22	Study findings, limitations, generalisability and current knowledge	Summarise key study findings and describe how they support the conclusions reached. Discuss limitations and the generalisability of the findings and how the findings fit with current knowledge.	P	Y	Y	Y	Y	Y	Y
Other									
23	Source of funding	Describe how the study was funded and the role of the funder in the identification, design, conduct and reporting of the analysis. Describe other nonmonetary sources of support.	N	P	N	Y	Y	Y	Y

Appendix 7: Quality of Reporting for the Included Studies

#	Selection/Item	Recommendation*	Ben-Shlomo, et al. (2016) (183)	León-Justel, et al. (2011) (50)	Midgette, et al. (1995) (51)	Dekkers, et al. (2016) (184)	Lubitz, et al. (2015) (185)	Sato, et al. (2015) (186)	Velasco, et al. (2015) (187)
24	Conflicts of interest	Describe any potential for conflict of interest of study contributors in accordance with journal policy. In the absence of a journal policy, we recommend authors comply with International Committee of Medical Journal Editors' recommendations.	Y	N	N	Y	N	Y	Y
Total Number of 'Yes' Responses**			**4**	**10**	**7**	**16**	**19**	**18**	**14**
Total Number of 'Partially Yes' Responses**			**1**	**3**	**3**	**3**	**2**	**3**	**2**
Total Number of 'Partially' Responses**			**4**	**2**	**7**	**1**	**2**	**1**	**3**
Total Number of 'Partially No' Responses**			**3**	**2**	**0**	**0**	**0**	**0**	**1**
Total Number of 'No' Responses**			**9**	**4**	**6**	**2**	**1**	**1**	**4**
Total Number of 'Not Applicable' Responses**			**6**	**6**	**4**	**5**	**3**	**4**	**3**

[**Source:** Table developed based on the CHEERS checklist (178, 179)]

**Scale*: *Yes (Y); Partially Yes (PY); Partially (P); Partially No (PN); No (N); Not applicable (N/A)*

***The total number of each response is only given for reference and does not reflect an actual quality score since the different items of the checklist are not of equal importance.*

Appendix 8: Quality Assessment of the Included Model-Based Economic Studies

Table 48: Quality assessment of the included model-based economic studies

#	Dimension of Quality	Questions for Critical Appraisal*	Midgette et al. (1995) (51)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
Structure						
S1	Statement of decision problem/objective	Is there a clear statement of the decision problem?	Y	Y	Y	Y
		Is the objective of the evaluation and model specified and consistent with the stated decision problem?	P	Y	Y	Y
		Is the primary decision-maker specified?	Y	P	P	P
S2	Statement of scope/perspective	Is the perspective of the model stated clearly?	Y	Y	Y	N
		Are the model inputs consistent with the stated perspective?	P	Y	Y	Y
		Has the scope of the model been stated and justified?	Y	Y	Y	Y
		Are the outcomes of the model consistent with the perspective, scope and overall objective of the model?	P	Y	Y	P

Appendix 8: Quality Assessment of the Included Model-Based Economic Studies

#	Dimension of Quality	Questions for Critical Appraisal*	Midgette et al. (1995) (51)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
S3	Rationale for structure	Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	P	Y	Y	Y
		Are the sources of data used to develop the structure of the model specified?	P	Y	Y	P
		Are the causal relationships described by the model structure justified appropriately?	Y	Y	Y	Y
S4	Structural assumptions	Are the structural assumptions transparent and justified?	Y	Y	N	Y
		Are the structural assumptions reasonable given the overall objective, perspective and scope of the model?	Y	Y	N/A	Y
S5	Strategies/comparators	Is there a clear definition of the options under evaluation?	Y	Y	Y	Y
		Have all feasible and practical options been evaluated?	N	N	N	N
		Is there justification for the exclusion of feasible options?	P	P	P	N
S6	Model type	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?	P	Y	Y	P
S7	Time horizon	Is the time horizon of the model sufficient to reflect all important differences between options?	Y	Y	Y	P
		Are the time horizon of the model, the duration of treatment and the duration of treatment effect described and justified?	P	P	P	N/A

Appendix 8: Quality Assessment of the Included Model-Based Economic Studies

#	Dimension of Quality	Questions for Critical Appraisal*	Midgette et al. (1995) (51)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
S8	Disease states/pathways	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?	Y	Y	Y	Y
S9	Cycle length	Is the cycle length defined and justified in terms of the natural history of disease?	N/A	P	P	N/A
Data						
D1	Data identification	Are the data identification methods transparent and appropriate given the objectives of the model?	P	Y	P	P
		Where choices have been made between data sources, are these justified appropriately?	P	P	P	P
		Has particular attention been paid to identifying data for the important parameters in the model?	N	Y	N	N
		Has the quality of the data been assessed appropriately?	NK	NK	N	N
		Where expert opinion has been used, are the methods described and justified?	N/A	N/A	P	N
D2	Data modelling	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?	P	Y	Y	P

Appendix 8: Quality Assessment of the Included Model-Based Economic Studies

#	Dimension of Quality	Questions for Critical Appraisal*	Midgette et al. (1995) (51)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
D2a	Baseline data	Is the choice of baseline data described and justified?	P	Y	Y	P
		Are transition probabilities calculated appropriately?	P	Y	Y	P
		Has a half-cycle correction been applied to both cost and outcome?	N/A	N	N	N/A
		If not, has this omission been justified?	N/A	N	N	N/A
D2b	Treatment effects	If relative treatment effects have been derived from trial data, have they been synthesised using appropriate techniques?	P	P	P	N/A
		Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified?	P	Y	P	N/A
		Have alternative assumptions been explored through sensitivity analysis?	Y	Y	Y	N/A
		Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified? Have alternative assumptions been explored through sensitivity analysis?	P	P	P	N/A
D2c	Costs	Are the costs incorporated into the model justified?	P	Y	Y	P
		Has the source for all costs been described?	P	Y	P	Y
		Have discount rates been described and justified given the target decision-maker?	N	Y	Y	N/A

Appendix 8: Quality Assessment of the Included Model-Based Economic Studies

#	Dimension of Quality	Questions for Critical Appraisal*	Midgette et al. (1995) (51)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
D2d	Quality of life weights (utilities)	Are the utilities incorporated into the model appropriate?	N/A	Y	N/A	N/A
		Is the source for the utility weights referenced?	N/A	Y	N/A	N/A
		Are the methods of derivation for the utility weights justified?	N/A	Y	N/A	N/A
D3	Data incorporation	Have all data incorporated into the model been described and referenced in sufficient detail?	P	Y	Y	P
		Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)?	N/A	N/A	N/A	N/A
		Is the process of data incorporation transparent?	Y	Y	Y	Y
		If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?	N/A	N/A	N/A	N/A
		If data have been incorporated as distributions, is it clear that second order uncertainty is reflected?	N/A	N/A	N/A	N/A
D4	Assessment of uncertainty	Have the four principal types of uncertainty been addressed?	N	N	N	N
		If not, has the omission of particular forms of uncertainty been justified?	N	N	N	N
D4a	Methodological	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?	N	P	P	N

Appendix 8: Quality Assessment of the Included Model-Based Economic Studies

#	Dimension of Quality	Questions for Critical Appraisal*	Midgette et al. (1995) (51)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
D4b	Structural	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?	N	P	N	N
D4c	Heterogeneity	Has heterogeneity been dealt with by running the model separately for different subgroups?	P	N	P	N
D4d	Parameter	Are the methods of assessment of parameter uncertainty appropriate?	P	Y	Y	Y
		If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?	N	Y	Y	N
Consistency						
C1	Internal consistency	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?	N	P	P	N
C2	External consistency	Are any counterintuitive results from the model explained and justified?	N/A	Y	N/A	P
		If the model has been calibrated against independent data, have any differences been explained and justified?	P	P	P	N/A
		Have the results of the model been compared with those of previous models and any differences in results explained?	N	N	N	N

Appendix 8: Quality Assessment of the Included Model-Based Economic Studies

#	Dimension of Quality	Questions for Critical Appraisal*	Midgette et al. (1995) (51)	Lubitz et al. (2015) (185)	Sato et al. (2015) (186)	Velasco et al. (2015) (187)
Total Number of 'Yes' Responses**			12	33	23	13
Total Number of 'Partially' Responses**			22	11	15	13
Total Number of 'No' Responses**			10	7	10	14
Total Number of 'Not Applicable' Responses**			11	4	8	16
Total Number of 'Not Known' Responses**			1	1	0	0

[**Source:** Table developed based on the *Philips et al. (2004)* checklist (180, 181)]

**Scale*: *Yes (Y); Partially (P); No (N); Not applicable (N/A); Not Known (NK)*

***The total number of each response is only given for reference and does not reflect an actual quality score since the different items of the checklist are not of equal importance.*

Appendix 9: ULTRADIAN Study Protocol

This Appendix provides the most important sections of the *ULTRADIAN* protocol used at the University of Bristol (United Kingdom). Non-UK sites used a similar protocol, but information was applied to their local populations. The parts of the protocol that are discussed in the **main Chapter** or are non-essential for describing the study are omitted (reported in 'brackets').

Full Title: Dynamic Hormone Diagnostics (ULTRADIAN)

Short Title: ULTRADIAN

Pan-study protocol for Bristol and non-UK sites. Each country will be seeking individual ethical and local approvals, and consequently will include local site-specific information – protocol for application for Bristol site only.

[…*Sponsor and investigator contact details per site; List of collaborations; General information on the study; Details on endocrine disease; Study hypothesis; Proposed study; Objectives and outcomes; Device components; Study design; Participant recruitment*…]

Part A: Definition of the normal circadian and ultradian profiles of pituitary and adrenal hormones in healthy volunteers

Aims and Objectives: The primary objective is to define the normal circadian and ultradian profiles of pituitary and adrenal hormones obtained from microdialysate samplings in the subcutaneous tissue of healthy individuals of different age and sex taken over 24 hours. Secondary objectives are to assess intra-individual variability and the accuracy of levels in microdialysates compared to blood sampling.

We will determine the 24-hour profile of cortisol and adrenocorticotropic hormone (ACTH), adrenal androgens and its precursors, 17-hydroxyprogesterone (17-OHP), androstenedione, dehydroepiandrosterone, testosterone, dihydrotestosterone and "backdoor" pathway metabolites pdiol and androsterone, aldosterone, renin, growth hormone (GH) and insulin-like growth factor-I (IGF-I), and other related hormones, proteins and peptides.

To achieve these aims, the study will be divided into 3 subgroups. We will consent healthy individuals for each sub-study separately. Each healthy volunteer will be asked at the first visit to participate in two studies, the single sampling study (**Study 1A**) and the repetitive collection study (**Study 2A**). At the end of the first study, agreement to participate in the repetitive collection study will be confirmed.

Study Population: Each clinical centre (Athens, Bergen, Bristol, Stockholm) will invite healthy individuals based on a random selection from the background population, ensuring an appropriate variation in age and sex.

Study Participants: Healthy adult male and female individuals aged between 18 to 68 years.

Inclusion Criteria:

- Healthy male and female individuals aged 18-68 years;
- Non-smokers or regular smokers of >6 cigarettes a day;
- Body mass index (BMI) of 16-29;
- Study subgroup 3A: Comparison of tissue and blood concentrations of hormones only: no oral contraceptive pill (OCP) usage for 6 weeks prior to sampling (females only).

Exclusion Criteria:

- Undergoing or planning pregnancy, or lactating women;
- Presence of any unstable pathological condition in the past 3 months;
- On any regular prescribed medications (except contraception);
- Prior or current history of an endocrine disorder;
- Known allergy to lidocaine and/or plasters;

- Current or past steroid therapies (oral, inhaled, parenteral or topical) or other interfering medication in the last 3 months;
- Regular alcohol intake of >26 units of alcohol per week;
- Taking of any investigational drug within the past 2 months;
- Abuse of illicit drugs;
- Occasional smokers of cigarettes not able to abstain during the sampling period or smoking <6 cigarettes a day, or individuals actively trying to stop smoking;
- Needle phobia;
- Interfering diet/over the counter herbal remedies in the last 14 days (St John's wart, liquorice and/or grapefruit juice).

Sample Size: We plan inclusion of 200 healthy volunteers aged between 18 to 68 years, with 20 individuals for each of the five age intervals of 18 to <28, 28 to <38, 38 to <48, 48 to <58, and 58 to 68 distributed equally between sexes. The number is based on requirements to be able for a dynamic secretion of hormones. This means that each of the four centres will recruit approximately 5 male and 5 female subjects from each of the age intervals of 18 to <28, 28 to <38, 38 to <48, 48 to <58, and 58 to 68. All recruited participants will be uploaded to a central server to ensure fluidity between centres as it may be easier/harder for certain centres to recruit from different demographics in comparison to others. This is a pilot study. As the *U-Rhythm* technology is new, it is difficult to estimate the numbers needed. The aim is to define normal profiles. Power analysis cannot be used to determine the sample size for this purpose.

Study subgroup 1A: **Definition of the normal circadian and ultradian profiles of pituitary and adrenal hormones in healthy subjects**

Study Design: Multi-centre observational study (Athens, Bergen, Bristol, Stockholm).
Each individual (total number 200, anticipated 50 per study centre) will be sampled by the *U-Rhythm* sampling device for 27 hours as described above (3 hours acclimatisation and 24 hours of sampling). Partner University of Bergen (UiB) and Proseek® immunoassays (OLINK).

Study subgroup 2A: **Day to day hormonal variability**

Study Design: Multi-centre observational study (Athens, Bergen, Bristol, Stockholm).
A subgroup of 20 participants will be asked to undergo sampling on three occasions to assess reproducibility of hormonal levels over time. These volunteers will be sampled over 27 hours (3 hours acclimatisation and 24 hours of sampling) once a week over 3 consecutive weeks by the same procedure described above to obtain information on day-to-day variations. Healthy individuals will be asked after each session whether they agree to participate in the next session. If the healthy volunteer agrees, a new informed consent will be sought prior to each new sampling session.

Study subgroup 3A: Comparison of tissue and blood concentrations of hormones

Study Design: Single-centre observational study (Bristol).
This study will be performed only at the University Hospitals Bristol NHS Foundation Trust (UH Bristol). 20 individuals will be asked to participate in the study comparing hormonal tissue and blood levels. Each healthy volunteer attending the screening visit at UH Bristol will be asked whether he/she would like to attend either the simple 27 hours of the *U-Rhythm* system sampling or the simultaneous blood and *U-Rhythm* system sampling. If these participants cannot be recruited from the original cohort, up to an additional 20 individuals will be recruited specifically for this subgroup. They will undergo the same 27 hours sampling by the *U-Rhythm* system as described above whilst simultaneously undergoing blood sampling via an automated blood sampling system to compare the tissue and blood levels of hormones. After the probe insertion, an intravenous cannula will be placed into the cubital vein of the arm and connected to the automated blood collector system. The participant will arrive at a mutually agreed time to the Joint Clinical Research Unit (JCRU) at UH Bristol. A blood sample will be automatically collected at pre-programed time intervals (anticipated range: 10-20 minutes) for 24 hours from the back of the indwelling cannula. The total blood volume per sample will be about 2ml. Participants will be required to stay at the research unit for the duration of the blood sampling period. A standardised routine will be kept and all meals will be provided.

Blood Collector System: This system painlessly collects blood samples at pre-programmed time intervals (anticipated range: 10-20 minutes) from the back of a venous cannula in the forearm. The system was developed by the research group of the University of Bristol and has been safely used in many studies including healthy volunteers and patients [*Reference*]. The system is currently utilised by the endocrine group of the University of Bristol in a study examining different patterns of glucocorticoid replacement in healthy volunteers and patients with adrenal insufficiency (NHS REC number –14/SW/1050, 11/SW/0078, University of Bristol REC number 2525). This system is safe, well tolerated, and has been used continuously for up to 24 hours in a series of healthy human volunteers and patients with the permission of the NHS ethics committee in Bristol, United Kingdom. This technique has allowed measurement of hormones, including during the important period of sleep, without disturbing or waking up the individual [*Reference*]. Collecting frequent timed samples automatically for prolonged periods makes it possible to investigate hormone rhythm physiology in greater detail without causing the stress or typical loss of blood associated with repeated standard venesection.

In both the healthy volunteer and patient studies, the collected blood samples will be analysed for cortisol and ACTH, followed by adrenal androgens and precursors, namely 17-OHP, androstenedione, dehydroepiandrosterone, testosterone, dihydrotestosterone and "backdoor" pathway metabolites pdiol and androsterone, followed by aldosterone, renin, GH and IGF-I, and other related hormones, proteins and peptide profiles analysed by LCMS/MS (partner UiB) and Proseek® immunoassays (OLINK).

Part B: Definition of the circadian and/or ultradian rhythm in patients with pituitary and adrenal disorders

Study subgroup 1B: Diagnosis of Cushing's syndrome by U-Rhythm dynamic cortisol measurements

Aims and Objectives: The primary objective is to establish circadian and ultradian hormonal profiles of patients with Cushing's syndrome (CS) from 24-hour ambulatory sampling of

subcutaneous fluid. A secondary aim is to compare the pre- and post-operative hormonal profiles of patients with CS, and to compare these results to age/sex matched controls.

Study Design: Multi-centre observational study (Athens, Bergen, Bristol, Stockholm).

Study Participants: Patients with established clinical and biochemical CS (ACTH-producing pituitary adenoma or ACTH-independent adrenal source).

Primary Endpoint: Defining circadian free cortisol profiles in the subcutaneous tissues of patients with confirmed CS.

Secondary Endpoints:

- Comparison with age/sex matched healthy volunteers;
- Variation in circadian subcutaneous ACTH, free cortisol and cortisone, and other related hormones, proteins and peptide profiles following surgical treatment for CS;
- Salivary cortisol and cortisone measurements at bedtime, on waking and 30 minutes post awakening for comparison with the free microdialysate cortisol;
- Comparison of circadian and ultradian cortisol dynamics in adrenal and pituitary CS.

Study Procedure: Each patient will be sampled by the *U-Rhythm* sampling device for 27 hours (3 hours acclimatisation and 24 hours of sampling) on up to four occasions, once before operative treatment and up to three times post-operatively (2-3 months and at 6 months, if evidence of hypothalamus-pituitary-adrenal (HPA) axis recovery, and up to 18 months regardless of HPA recovery). In the post-operative sampling sessions, patients must be taken off steroids the night before (no glucocorticoids after 6pm the night prior to sampling) and during the sampling period. All patients will be supplied with a steroid emergency kit and emergency telephone number. If required, the patient will be admitted to the local research unit/endocrinology department to undergo sampling. The collected samples will be analysed for ACTH, cortisol, cortisone, and other related hormones, proteins and peptide profiles by LCMS/MS (partner UiB) and Proseek® immunoassays (OLINK).

Study Population: Patients with established clinical and biochemical CS, either an ACTH-producing pituitary adenoma or an ACTH-independent adrenal adenoma, will be recruited at each study centre from a patient registry or its outpatient clinics. The patients will give written informed consent and will be free to withdraw at any time.

Inclusion Criteria:

- Male and female patients aged 18-68 years;
- Biochemically confirmed CS based on cortisol and ACTH measurements, dexamethasone suppression test, and one of the following positive investigations (24-hour urine cortisol or salivary/serum cortisol profile);
- Visible tumour in the pituitary and/or one of the following positive tests pointing to pituitary source for ACTH overproduction (bilateral petrosal sinus sampling, corticotropin-releasing hormone test or high-dose dexamethasone suppression test), or adrenal tumour considered fit for surgery;
- No OCP usage for 6 weeks prior to sampling (females only);
- Patients on medical therapy for CS to have a 2-week washout off treatment prior to sampling.

Exclusion Criteria:

- Undergoing or planning pregnancy (females only);
- Known allergy to lidocaine;
- Adrenal cancer (post-operative histology diagnosis), cyclic CS, squamous cell lung carcinoma, bronchial carcinoid and occult ectopic CS;
- Concurrent use of glucocorticoid for any other medical condition (oral, inhaled, parenteral or topical);
- Anticoagulation treatment except low-dose low-molecular-weight heparin and/or low-dose aspirin.

Sample Size: A total of 40 patients comprising approximately 20 individuals with a pituitary cause of CS and 20 individuals with an adrenal cause of CS will be recruited to the study. If a patient is no longer deemed eligible in the post-operative period due to tissue histology

diagnosis, then another patient will be recruited in their place. This patient's data will still be included in primary endpoint data analysis. It is difficult to use power analysis for estimating sample sizes because the primary aim of these studies is to define or measure profiles for these sub-populations.

Study subgroup 2B: Monitoring of primary adrenal insufficiency (Addison's disease) by U-Rhythm dynamic cortisol and ACTH measurements

Aims and Objectives: To compare hormonal profiles of patients with Addison's disease (AD) on different replacement regimes to age/sex matched controls.

Study Design: Multi-centre observational study (Bergen, Bristol, Stockholm).

Study Participants: Patients with established AD.

Primary Endpoint: Measurement of circadian and ultradian free cortisol, ACTH, and other related hormones, proteins and peptide profiles on different forms of replacement therapy.

Secondary Endpoints:

- Comparison with age/sex matched healthy individuals;
- Salivary cortisol and cortisone measurements at bedtime, on waking and 30 minutes post awakening.

Study Procedure: Each subject will be sampled by the *U-Rhythm* sampling device for 27 hours (3 hours acclimatisation and 24 hours of sampling) on their current replacement regime. If an adjustment is made to their replacement regime for clinical reasons, then, with the patient's consent, *U-Rhythm* sampling will be repeated on their new regime. Patients will be sampled up to a maximum of three occasions. The collected samples will be analysed for cortisol, cortisone, ACTH, and other related hormones, proteins and peptide profiles by LCMS/MS (partner UiB) and Proseek® immunoassays (OLINK).

Study Population: Patients with established AD will be recruited at each study centre from a registry, database or via their outpatient clinics. The participants will give written informed consent and are free to withdraw at any time.

Inclusion Criteria:

- Male and female patients aged 18 -68 years;
- Biochemically confirmed AD based on basal cortisol and ACTH measurements, and/or Synacthen test with positive adrenal antibodies;
- On oral hydrocortisone or cortisone acetate glucocorticoid replacement therapy.

Exclusion Criteria:

- Undergoing or planning pregnancy (females only);
- Known allergy to lidocaine;
- Use of any glucocorticoid medications other than their standard glucocorticoid and fludrocortisone replacement;
- Other interfering medication or diet within 2 weeks of sampling (St. John's wart, liquorice and/or grapefruit juice);
- Anticoagulation treatment except low-dose low-molecular-weight heparin and low-dose aspirin;
- Any other form of glucocorticoid replacement, such as prednisone, prednisolone, dexamethasone, plenadren® and pump treatment.

If the participant meets all eligibility criteria but testing for adrenal antibodies cannot be located in the clinical record, the study investigators may ask permission from the participant to take a further blood sample for this purpose.

Sample Size: A total of 20 patients will be recruited to the study. It is difficult to use power analysis for estimating sample sizes because the primary aim of these studies is to define or measure profiles for these sub-populations.

Study subgroup 3B: Monitoring of congenital adrenal hyperplasia by U-Rhythm dynamic cortisol, ACTH and androgen measurements

Aims and Objectives: To compare hormonal profiles of patients with congenital adrenal hyperplasia (CAH) on different replacement regimes to age/sex matched controls.

Study Design: Multi-centre observational study (Bergen, Bristol, Stockholm).

Study Participants: Patients with established CAH, either salt-wasting or simple virilisation forms, on glucocorticoid replacement therapy.

Primary Endpoint: Measurement of circadian and ultradian simultaneous free cortisol, androgen or androgen metabolite profiles (including 17-OHP, androstenedione, dehydroepiandrosterone, pdiol, androsterone), and other related hormones, proteins and peptide profiles in the subcutaneous tissues of patients with CAH.

Secondary Endpoints:

- Comparison with age/sex matched healthy individuals;
- Salivary cortisol and cortisone measurements at bedtime, on waking and 30 minutes post awakening.

Study Procedure: Each patient will be sampled by the *U-Rhythm* sampling device for 27 hours (3 hours acclimatisation and 24 hours of sampling) on their current replacement regime. If an adjustment is made to their replacement regime for clinical reasons, then, with the patient's consent, *U-Rhythm* sampling will be repeated on their new regime. Patients will be sampled up to a maximum of three occasions. The collected samples will be analysed for ACTH, cortisol, cortisone, androgens, metabolites, and other related hormones, proteins and peptide profiles by LCMS/MS (partner UiB) and Proseek® immunoassays (OLINK).

Study Population: Patients with established CAH, either salt-wasting or simple virilisation form, will be recruited. The participants will have to give written informed consent and will be free to withdraw at any time.

Inclusion Criteria:

- Male and female patients aged 18-68 years;
- Biochemically confirmed salt-wasting or simple virilising CAH based on 17-OHP, cortisol, androgen, renin and ACTH measurements, disease-causing mutation in CYP21A2, and/or Synacthen testing;
- All current steroid replacement treatment and each combination.

Exclusion Criteria:

- Undergoing or planning pregnancy (females only);
- Known allergy to lidocaine;
- Use of other steroid medications other than their standard glucocorticoid and fludrocortisone replacement;
- Other interfering medication or diet within 2 weeks of sampling (St. John's wart, liquorice and/or grapefruit juice);
- Anticoagulation treatment except low-dose low-molecular-weight heparin and low-dose aspirin.

Sample Size: A total of 20 participants will be recruited to the study. It is difficult to use power analysis for estimating sample sizes because the primary aim of these studies is to define or measure profiles for these sub-populations.

Study subgroup 4B: Diagnosis of primary aldosteronism by U-Rhythm dynamic aldosterone and renin measurements

Aims and Objectives: The primary objective is to establish circadian and ultradian profiling of free aldosterone and renin in the subcutaneous tissue of primary aldosteronism (PA) patients. Secondary objectives are: (1) to compare pre- and post-operative profiles; (2) to identify profiles typical for adenoma as opposed to bilateral hyperplasia; and (3) to compare profiles to age/sex matched controls.

Study Design: Multi-centre observational study (Athens, Bergen, Bristol, Stockholm).

Study Participants: Patients with PA.

Primary Endpoint: Measurement of simultaneous circadian and ultradian free aldosterone, renin, and other related hormones, proteins and peptide profiles in the subcutaneous tissue of patients with PA.

Secondary Endpoints:

- Comparison of circadian and ultradian subcutaneous free aldosterone, renin, and other related hormones, proteins and peptide profiles of patients with proven PA compared to suspected PA and healthy individuals;
- Change in circadian and ultradian subcutaneous free aldosterone, renin, and other related hormones, proteins and peptide profiles following surgical treatment for PA.

Study Procedure: Patients with PA will be sampled by the *U-Rhythm* sampling device for 27 hours (3 hours acclimatisation and 24 hours of sampling) after biochemical confirmation of PA. As part of their routine clinical care, patients on antihypertensive treatment will be changed to non-interfering antihypertensives for 2-4 weeks (dependent upon the type of antihypertensives) prior to dynamic testing. If sampling for ultradian cannot be performed within this time window, participants will be asked to continue on the modified regime for up to 2-3 weeks post dynamic testing. If biochemical diagnosis is confirmed, all patients will undergo further radiological plus/minus dynamic testing to confirm unilateral versus bilateral disease as part of their usual care. Patients with confirmed unilateral disease who undergo operative treatment will be re-sampled during the second or third post-operative month. Patients with bilateral hyperplasia will start medical treatment with aldosterone antagonists and not tested further. Patients who do not have confirmed biochemical PA will not be sampled further. The collected samples will be analysed for aldosterone, renin, and other related hormones, proteins and peptide profiles by LCMS/MS (partner UiB) and Proseek® immunoassays (OLINK).

Study Population: Patients with suspected clinical PA will be recruited at each study centre from a patient registry or their outpatient clinics. The participants will have to give written informed consent and will be free to withdraw at any time.

Inclusion Criteria:

- Male and female patients aged 18-68 years;
- Biochemical confirmation of PA based on a valid pathological aldosterone-renin ratio and non-suppressible aldosterone levels on one of the currently used confirmatory tests (saline infusion or fludrocortisone suppression test);
- Diagnostic computerised tomography and/or adrenal vein sampling;
- Potassium within normal range at the time of sampling.

Exclusion Criteria:

- Undergoing or planning pregnancy (females only);
- Known allergy to lidocaine;
- Known adrenal failure and/or on steroid therapy (oral, inhaled, parenteral or topical);
- Concurrent use of specified interfering medication or diet within 2-4 weeks of ultradian sampling (time scale dependent upon type of interfering medication);
- Anticoagulation treatment except low-dose low-molecular-weight heparin and low-dose aspirin.

Sample Size: Participants will be recruited until a total of 30 patients with confirmed PA are recruited to the study. It is difficult to use power analysis for estimating sample sizes because the primary aim of these studies is to define or measure profiles for these sub-populations.

Study subgroup 5B: Diagnosis and treatment of acromegaly by U-Rhythm dynamic growth hormone measurements

Aims and Objectives: The primary objective is to establish hormonal profiles of patients with acromegaly (AC). A secondary objective is to compare pre- and post-operative profiles, and to compare these profiles to age/sex matched controls.

Study Design: Multi-centre observational study (Athens, Bergen, Bristol, Stockholm).

Study Participants: Patients with established clinical and biochemical acromegaly by current diagnostic criteria.

Primary Endpoint: Measurement of circadian and ultradian GH, and other related hormones, proteins and peptide profiles in the subcutaneous tissues of patients with confirmed acromegaly, before and after definitive treatment with surgery.

Secondary Endpoints:

- Changes in IGF-I levels, comparison with age, sex and body weight matched healthy individuals;
- Change in circadian and ultradian subcutaneous GH, and other related hormones, proteins and peptide profiles following surgical treatment for acromegaly.

Study Procedure: Each patient will be sampled by the *U-Rhythm* sampling device on a maximum of two occasions for 27 hours (3 hours acclimatisation and 24 hours of sampling), once at diagnosis and on a second occasion after surgical treatment (during the second or third post-operative month). The collected samples will be analysed for GH, IGF-I, and other related hormones, proteins and peptide profiles by LCMS/MS (partner UiB) and Proseek® immunoassays (OLINK).

Study Population: Patients with established clinical and biochemical acromegaly will be recruited at each study centre from a registry or their outpatient clinic. The participants will give written informed consent and will be free to withdraw at any time.

Inclusion Criteria:

- Male and female patients aged 18-68 years;
- Biochemically confirmed acromegaly (oral glucose tolerance test, or GH day curve and diagnostic IGF1 levels) with radiological evidence of a pituitary adenoma.

Exclusion Criteria:

- Undergoing or planning pregnancy (females only);
- Known allergy to lidocaine;
- Treatment with somatostatin analogues for 2-3 months (baseline sampling only) and other interfering medication (e.g. cabergoline) for 6 weeks prior to sampling (all sampling sessions);
- Anticoagulation treatment except low-dose low-molecular-weight heparin and low-dose aspirin;
- Pegvisomant treatment.

Sample Size: Recruitment will continue until a total of 20 patients who receive operative treatment is reached. If a patient, who is initially recruited to the study, does not receive surgical treatment, then another will be recruited in their place. This patient's data will still be included in primary endpoint data analysis. It is difficult to use power analysis for estimating sample sizes because the primary aim of these studies is to define or measure profiles for these sub-populations.

Study subgroup 6B: Diagnosis of growth hormone deficiency by U-Rhythm dynamic growth hormone measurements

Aims and Objectives: The primary objective is to establish hormonal circadian and ultradian profiles of adult growth hormone deficiency (GHD) patients by analysing the GH profile in the subcutaneous tissue fluid.

Study Design: Multi-centre observational study (Athens, Bergen, Bristol, Stockholm).

Study Participants: Adult patients with established clinical and biochemical GHD.

Primary Endpoint: Measurement of 24-hour GH, and other related hormone, proteins and peptide profiles in the subcutaneous tissue of patients with confirmed GHD.

Secondary Endpoints:

- Change in IGF-I levels;
- Comparison of circadian GH, and other related hormones, proteins and peptide profiles with age, sex and body weight matched healthy individuals.

Study Procedure: Patients with confirmed GHD will be asked to stop their GH treatment for the day prior to and the day of sampling. Each patient will be sampled by the *U-Rhythm* sampling device once for 27 hours (3 hours acclimatisation and 24 hours of sampling). The collected samples will be analysed for GH, IGF-I, and other related hormones, proteins and peptide profiles by LCMS/MS (partner UiB) and Proseek® immunoassays (OLINK).

Study Population: Patients with established clinical and biochemical GHD will be recruited at each study centre from a registry, database or their outpatient clinics. The participants will give written informed consent and will be free to withdraw at any time.

Inclusion Criteria:

- Male and female patients aged 18-68 years;
- Biochemically confirmed GHD (arginine GH releasing hormone test, insulin tolerance test, known pituitary disease with confirmed pan hypopituitarism).

Exclusion Criteria:

- Undergoing or planning pregnancy (females only);
- Known allergy to lidocaine;
- Anticoagulation treatment except low-dose low-molecular-weight heparin and low-dose aspirin.

Sample Size: A total of 20 patients will be recruited to the study. It is difficult to use power analysis for estimating sample sizes because the primary aim of these studies is to define or measure profiles for these sub-populations.

[…*General endpoints; Procedures within the trial*…]

11. Participant Compliance

The trial does not involve administration of any treatment. Compliance with the wearing of the collection device can only be ascertained by self-report by the wearer. Compliance related to alcohol consumption, caffeine intake, medication and other lifestyle factors will be evaluated before the study.

If a patient chooses to discontinue their trial participation at any time during the study, the time of discontinuation and the reason will be recorded in the data sheet, and only the samples collected will be analysed. They will automatically be excluded from the next part of the trial, if applicable. They will be provided with contact details of the chief investigator for any queries they may have subsequent to their discontinuation.

12. Withdrawal of Participants

Participants are free to withdraw from the study at any time. Should a participant withdraw from the study, another will be recruited in their place, using the same recruitment method as detailed above. It is not anticipated that withdrawn participants will require follow-up. If a patient withdraws from the study, they will be offered a follow-up interview with one of the study investigators on their request to address any queries that may have arisen.

In the best interest of the patient or healthy volunteer, the investigator and the sponsor can decide to withdraw them from the study. If a patient develops conditions meeting the exclusion criteria, the patient will be withdrawn from the study. In case of serious local adverse events, the patient may be withdrawn from the study. Any data/samples collected from the patients will still be analysed pending their consent.

13. Data Collection

The chief investigator will collect data from the following sources:

- Patients notes and coding for confirmation of diagnosis for patients with AC, AD, CAH, CS, GHD and/or PA;
- Healthy volunteer interview;
- Weight, height, BMI, blood pressure and waist-hip ratio;
- Activity diary;
- 10-minutely up to 80-minutely tissue microdialysate samples for:
    - Hypothalamic hormones: corticotropin-releasing hormone, melatonin;
    - Pituitary hormones: ACTH, GH, prolactin;
    - Adrenal/gonadal hormones: cortisol, cortisone, aldosterone, renin, 17-OHP, androstenedione, dehydroepiandrosterone, testosterone, dihydrotestosterone, and "backdoor" pathway metabolites pdiol and androsterone;
    - Others: IGF-1 and any other subgroup specific related hormones, proteins and peptides.
- 10-minutely up to 60-minutely blood samples for:
    - Hypothalamic hormones: corticotrophin-releasing hormone, melatonin;
    - Pituitary hormones: ACTH, GH, prolactin;
    - Adrenal/gonadal hormones: cortisol, cortisone, aldosterone, renin, 17-OHP, androstenedione, dehydroepiandrosterone, testosterone, dihydrotestosterone, and "backdoor" pathway metabolites pdiol and androsterone;
    - Other: IGF-1 and any other subgroup specific related hormones, proteins and peptides.
- Three salivary cortisol samples per 24-hour period;
- Basal hormonal blood samples.

All the endocrine investigations, including blood tests and radiological investigations, will be accessed and recorded in the data collection form. This will help to correlate the observed hormonal profiles with known endocrine abnormalities found during routine investigations of patients in day-to-day clinical practice that form the basis of clinical decision-making. All data will be collected and recorded by the study investigators and their team.

14. Sample Collection, Storage, Shipping, Analysis and Disposal

All samples will be collected, used, stored and disposed of in accordance with the Human Tissue Act 2004 and International Conference on Harmonization (ICH) Good Clinical Practice (GCP) requirements. In each participant, samples will be collected for a total of 27 hours (3 hours of baseline and 24 hours of sample collection proper) on each occasion. This will occur on a maximum of three occasions: at baseline (patient and healthy volunteer); following surgery (2-3 post-operative months and 6-12 months, if clinical signs of recovery); post routine treatment modification (if appropriate) and repetitive sampling of healthy volunteers over 3 weeks (81 hours in total, 27 hours each week), with the exception of **Study 1B** in which a fourth sampling session may occur. Records of all samples will be kept in the site file log.

A. Samples for Transfer to Bergen

All samples will be labelled with a unique study ID, transferred to 96-well plates, divided by laboratory robots and stored at -80 degrees Celsius at each study centre project storage area**.** Saliva, microdialysis and blood hormonal samples will be shipped on dry ice via courier service to the *ULTRADIAN* biobank at Bergen, and will be analysed for corticosteroids and metabolites using LCMS/MS (partner UiB). ACTH, renin, GH, melatonin and IGF-I profiles, and other related hormones, proteins and peptide profiles will be assayed with Proseek® immunoassays (OLINK). Prior to shipping, Bergen/OLINK will receive notification that the samples are due to be sent and their anticipated arrival date. Samples will remain the responsibility of each study centre until their arrival has been logged and confirmed by Bergen/OLINK. Samples from each study centre will be analysed by LCMS/MS at University of Bergen and with Proseek® immunoassays (OLINK). Any residual sample will be stored for up to five years at the biobank at Bergen in case of repeated or additional analysis to allow for completion of the study. After that time, all residual samples will be destroyed on site at Bergen in accordance with the Human Tissue Act 2004 and ICH GCP requirements.

B. Locally Processed Samples

Screening blood tests (healthy volunteers), and baseline biochemistry and haematology tests (patient groups) will be processed at the UH Bristol pathology laboratory. Samples will be anonymised and labelled with the unique study ID before transfer by the study investigators to the laboratory. Any residual sample will be stored for up to five years at the study centre

project storage area in case of repeated or additional analysis to allow for completion of the study. After that time, all residual samples will be destroyed on site in accordance with the Human Tissue Act 2004 and ICH GCP requirements.

15. Data Handling and Record Keeping

Study researchers are responsible for data collection, recording and quality. All data will be recorded in the study site file. Data collected in the site file will be entered into a database. All data relating to study results will be held in anonymised form using a unique identification number, which will be linked to the case report form (CRF). A participant identification code list will be kept in the investigator's binder. All CRFs will be filled out with an archive-proof pen. Corrections are to be made by striking through the incorrect information and writing the revision alongside. All corrections will be initiated and dated.

The results of the samples processed in Bergen (blood, saliva and microdialysis) will be stored on a University of Bergen password restricted computer and/or the project database, which may be accessed via an interactive webpage with a protected domain. Locally processed sample results will be accessible to investigators via the UH Bristol secure intranet results server. Sample results from external laboratories will be sent electronically on password encrypted files or via courier. Results will then be downloaded and stored on a University of Bergen password restricted computer, and/or the project database and interactive webpage with a protected domain. Data for the activity diary will all be processed and stored on a university password protected study computer.

Site-specific personal data will be kept securely as paper records and entered onto a local computer database. Access to computer databases will be password protected. Data will be stored on each study centre server and be accessible to designated study personnel only. Data will be collected and retained in association with the Data Protection Act of 1998. No personal data will be stored on laptops. Any data that needs to be transferred off site will be done so either electronically, or on disks and couriered. This data is fully anonymous. A copy of the consent form(s) will be stored in the master file (stored in a secure location at each study site - the Dorothy Hodgkin Building, University of Bristol) and site files at the research units (JCRU - restricted entry, access to designated personnel). Local study documents (paper and

electronic) will be retained in a secure location during and after the trial has finished by the University of Bristol. The study documentation and research data will be stored for five years after the termination of the study. Where trial-related information is documented in medical records, those records will be identified by a **'Do not destroy before dd/mm/yyyy'** label, where date is 15 years after last patient visit.

16. Access to Source Documents

In accordance with applicable regulatory requirements including participant consent, patient records will be made available to the applicable regulatory agency, monitor and auditor.

17. Sample Size Calculation and Statistical Analysis

The study is multi-centre observational pilot study, limited in numbers by rare diseases. As a pilot study with limited number of participants, statistical power calculation is difficult. Force analysis has been performed to determine the necessary number of participants accounting for presumed differences between groups and an estimated variance of key parameters. The results indicated that around 20-60 participants should be sufficient for presumed difference detection at significance level $p<0.01$.

Mathematical modelling techniques (deconvolution analysis) will be used to estimate hormonal secretory measures by the study researchers using Pulse_XP software (University of Virginia). Additionally, we will be using simple statistical tests including measures of area under the curve (AUC) and paired t-testing.

[…*Potential risks; Safety assessments*…]

20. Stopping/Discontinuation Rules

The criteria for discontinuation of the study may arise from participant or device failure:

- The participant is free to withdraw at any time.

The investigator will consider withdrawal if:

- The participant is having major difficulties regarding probe insertion;
- The participant is having major difficulties regarding wearing the fraction collector;
- Significant protocol deviation or violation;
- Significant failure to comply with study requirements;
- If the participant meets the exclusion criteria;
- In case of a serious local adverse event;
- If the participant failed to cooperate with the study investigators;
- Death of the participant.

A device failure is anticipated as a very rare event and may arise from:

- Any part of the fraction collector, microdialysis pump (battery failure; obtaining damaged samples due to device failure);
- Discontinuation of any part of the fraction collector;
- A probe falling/pulled out from the skin.

In this case, the participant will be asked to either repeat the study after reconsideration of the principal investigator or discontinue study participation.

The study will be completed when the last participant's samples have been fully analysed. It is not anticipated that the study will be terminated early as this is an observational study. As this study does not involve a randomisation study, there is no need for a policy for breaking randomisation. After the termination of the study, the local Regional Committee for medical and health research ethics will be notified. If needed to terminate the study earlier, we will notify the Local Regional Committee for medical and health research ethics 15 days before, and the reason of the premature termination will be clarified and kept as a record in the study site file and in CRF form.

[…*International collaborators; Monitoring and audit; Ethical consideration; Research governance; Ethics and R&D approval; Finance; Indemnity; Reporting and dissemination; References; Appendices*…]

Appendix 10: U-Rhythm Cost Analysis

As mentioned in **Chapter IV**, the *U-Rhythm* diagnostic procedure involves a number of steps, each with an associated cost. To provide a preliminary estimate of the cost of this diagnostic test, if used in routine clinical practice, these steps are described and costed using current estimates.

The U-Rhythm Collection Device

The sampling device includes both custom-made and commercially sourced components. In brief, the device is housed in a custom mechanical enclosure which is subtractive prototyped from solid block acrylonitrile butadiene styrene plastic. The housings are then hand-finished and painted. This was deemed the most cost-efficient process for manufacturing small batches for the devices used in the *ULTRADIAN* study. The production version will ultimately be injection moulded from a plastic polymer. Initial investment in injection mould tooling will be high, but the unit cost for enclosure plastics will be massively reduced versus the subtractive prototype process. Within the mechanical enclosure, there is a mixture of sub-assemblies. These sub-assemblies include a fluid and air handling manifold, a printed circuit board and a rechargeable battery pack. The main high-price point items in the device are sited in the general manifold assembly. The two-port solenoid, two miniature peristaltic pumps, pressure sensor and custom polymethyl methacrylate manifold, and polyether ether ketone barbs represent ≈20% of the device cost. The population and assembly of the printed circuit board contribute ≈15% to its cost. Enclosure plastics and other ancillary items equate to ≈15% of the total cost. The remainder of the cost is attributed to assembly, component finishing, and verification and validation testing prior to device dispatch. The current estimated cost price of the *U-Rhythm* sampling device is €6,837.00. The price includes material, labour, and verification and validation testing. This is a prototype device and its final version will have somewhat different component selection and manufacture price. It is expected that, once in production, some of the device component costs may increase due to additional design features (e.g. charging dock) and general regulatory requirements for CE marking. As currently designed, the device sits in a waist belt (≈€18.00) while sampling occurs in an ambulatory participant.

The device is designed to work with a reusable microdialysis infusion pump (M Dialysis AB, http://www.mdialysis.com/pumps/-107), a portable battery-driven pump. The pump syringe is filled with 2.5mL of sterile perfusion fluid, connected to the microdialysis catheter and then placed in the pump. When the pump lid is closed, a five-minute flush cycle begins and is followed by an automatic decrease to the pre-set operating rate. The estimated cost of the microdialysis pump is approximately €2,773.00. The collection device, the waist belt and the pump are multi-use items that would be reused throughout their technologically useful life with presumably little or no resale value at the end of this period. At this point in the development of the prototype device, it is very difficult to estimate both the useful lifespan of the equipment and the number of patients who might use the equipment in any given time period. Assuming, as a starting point, a total cost of multi-use equipment of €10,000, a technologically useful lifespan of one to five years, and a throughput of between 50 and 150 patients per year, the equipment cost per patient use of the diagnostic device are estimated in **Table 49**. If used widely in clinical practice, it is possible that the device price might also include the cost of a service agreement whereby after a fixed period or number of uses, the device is returned for service and recalibration. For example, it is recommended that the microdialysis pump is returned for a 'control check' every two years at a cost of ≈10% of the purchase price. As the future business model for servicing *U-Rhythm* devices is unclear, this cost has not been included in the calculations, but it is likely to increase the cost of long-term use. There is considerable uncertainty about the estimated device cost per patient and this is unlikely to be fully resolved until the device is CE marked and is in routine clinical use.

Table 49: Estimate of the U-Rhythm device cost per patient per year

Technologically Useful Timespan	Patients per Year	Estimated Device Cost per Patient*
1 year	50	€200
	100	€100
	150	€67
3 years	50	€67
	100	€33
	150	€22
5 years	50	€40
	100	€20
	150	€13

**Assuming no resale value.*

Sampling Procedure

The sampling procedure involves a clinic room, clinician time (**Table 50**), and a number of single- or limited-use items of equipment (**Table 51**). A sampling session typically takes place over a two-day period (e.g. from morning of day 1 to midday of day 2), and requires a clinic room bed for ≈40 minutes on day 1 for the insertion of the probe, and fitting of the pump and device, and for a shorter period (≈10 minutes) on day 2 for probe and device removal. The costs of the sampling procedure are dominated by the cost of the disposable microdialysis catheter (€159.36), and the time spent by the technician in preparing the samples (€78.89) and the nurse practitioner or doctor in fitting the device (€43.48). The total cost of the sampling procedure is estimated to be in the region of €360. As the salaries and the cost of living in the UK are fairly high relative to some other European countries, the cost of the sampling procedure may be lower in other countries.

Table 50: Staff time and overheads for sampling and probe removal

Clinical Time Description	Time	Allocation	Estimated Cost*
Prepare sampler prior to participant arrival	00:20	Technician (Band 4)	€10.52
Setup trolley in clinical unit prior to participant arrival	00:05	Healthcare assistant (Band 2)	€1.93
Probe fitting (insert; attach to sampler; put in bag; reminders to participant etc.)	00:40	Nurse practitioner (Band 8a)/Doctor	€43.48
Probe removal	00:10	Nurse/Healthcare assistant (Band 4)	€5.26
Sampler cleaning and sample preparation	02:30	Technician (Band 4)	€78.89
Total per Sample	**03:45**		**€ 140.08**

**Costs include salary, oncosts, qualifications and overheads. They are estimated from the UK Unit Costs of Health and Social Care (2017) (502).*

Table 51: Single- or limited-use items needed during the sampling procedure

Item	Cost per Patient
Pump syringe	€12.68
Linear microdialysis catheter	€159.36
Energizer lithium AA sampler battery	€2.31
CMA 107 pump battery	€2.10
1m FEP tubing	€3.46
MAB connector	€2.73
FEP tubing adapter	€1.89
T1 perfusion fluid	€6.31
Prewound spool 150 spools/unit	€33.66
Total Cost per Sample	**€ 224.50**

Analysis of the Collected Samples

The analysis of the microdialysate fluid samples is performed using ultrasensitive liquid chromatography tandem mass spectroscopy (LCMS/MS) for steroid hormones. The LCMS/MS procedure was conducted at the University of Bergen but could in routine clinical practice be performed in any hospital laboratory with the required equipment. The process involves equipment, consumables and personnel costs. The LCMS/MS machine at the University of Bergen requires one full-time engineer and has the potential to process six plates per week or 252 plates per year (assuming a 42-week working year). Each plate contains 72 samples collected every 20 minutes over a 24-hour period from one participant. Sample processing is robotised and carried out on a Hamilton STAR™ robot since this improves the analytical quality and allows for a high throughput on one instrument. It is assumed that the LCMS/MS machine is fully allocated to *U-Rhythm* samples and that one robot could service three LCMS/MS machines. The predominant costs of the process (**Table 52**) are the consumables, machine depreciation and service contract costs, and the staff costs. The annual cost of the LCMS/MS analysis process is estimated to be in the region of €160,000 which equates to ≈€640.00 per plate (i.e. per participant). This excludes some infrequent overhead costs (e.g. analytes, internal standards, calibrators, quality control) and consumables (e.g. nitrogen, argon, electricity) which are shared over a large number of laboratory processes. The estimated cost includes 25% value added tax for many of the equipment and consumables that are imported to Norway from the European Economic Area and other countries. As salaries and the cost of living in Norway are high relative to most other European countries, the cost of the LCMS/MS process may be lower in other countries.

Table 52: Single- or limited-use items/services needed for the analysis of samples

Item/Service	Quantity	Annual Cost°
Consumables	**Per Month**	
Methanol	14L	€4,425.21
Ethyl acetate	2L	€1,106.30
Hexane	1L	€590.03
Pipette tips 50uL	9,216 tips	€5,162.75

Item/Service	Quantity	Annual Cost°
Pipette tips 1000uL	4,608 tips	€5,900.29
Chromacol™ vials 500uL	4 boxes/500-vial pack	€7,375.36
Chromacol™ vials 1mL	4 boxes/500-vial pack	€6,195.30
Cortex columns	2	€11,579.31
VanGuard pre-columns	3	€4,759.32
	Per Year	
Mats	20	€958.80
Impactor pin	3	€156.95
Cone	1	€392.96
Propanol	2	€78.18
Acetonitrile	2	€88.50
Staff	**Per Year**	
Engineer	1	€47,939.82
Equipment	**Per Year**	
LCMS/MS machine¤	1	€34,335.09
Hamilton Microlab® STAR™ robot¤	0.5	€3,745.30
LCMS/MS service contract	1	€22,722.37
Hamilton service contract	0.5	€2,805.09
Total Annual Cost		**€160,316.93**
Total Cost per Plate (252 per Year)		**€636.18**

**Abbreviations: LCMS/MS: ultrasensitive liquid chromatography tandem mass spectroscopy*

°Assuming machine in use for 10 months within the year.

¤Assuming 8-year technologically useful lifespan and straight-line depreciation. Assuming that one robot can service three LCMS/MS machines.

U-Rhythm Total Cost Estimate

The total cost of the *U-Rhythm* diagnostic procedure is estimated to be in the region of €1,100 if LCMS/MS is used to analyse steroid hormones (e.g. cortisol).

Table 53: Total cost of the U-Rhythm diagnostic procedure

Item	Estimated Cost per Patient
Cost of the *U-Rhythm* device	€13 to €200 (depending on throughput and lifespan)
Cost of the sampling procedure	€365
Cost of LCMS/MS	€636
Total Cost	**€1,014.00 to €1,201.00** **(£1,008.07 to £1,193.98)**

**Abbreviations: LCMS/MS: ultrasensitive liquid chromatography tandem mass spectroscopy*

Appendix 11: U-Rhythm Diagnostic Accuracy Analysis

Diagnostic Accuracy Analysis Methods

To estimate the diagnostic accuracy (i.e. sensitivity/specificity) of *U-Rhythm* as a confirmatory test for the diagnosis of primary aldosteronism (PA), a logistic regression model was developed using *ULTRADIAN* data from PA patients and their matched controls. The model predicted whether a device 'sampling' (which included many 'samples' taken at 20-minute intervals) was from a healthy volunteer or a pre-operative PA patient. It was built using 10-times repeated 5-fold cross-validation with logistic regression (503, 504) and the R-package *'caret'* (505, 506). The hormones included in this model were cortisol, cortisone, aldosterone and 18-hydroxycortisol. The hormone profile data on which the model was trained consisted of the individual samples within the samplings scaled by a fixed value for each hormone to bring the data (roughly) into the '0-1' range suitable for logistic regression. The means of each hormone within the sampling was also included. Other predictors which were included were the sleep/wake times recorded in the participant's activity diary plus their age and sex. The target metric for the cross-validation was balanced accuracy, which is an average of the sensitivity and specificity. As part of the cross-validation procedure, the probability cut-off or threshold was also tuned. This threshold was used to determine whether a class probability predicted healthy or not. The final model of the cross-validation process was obtained using all the data and the tuned parameters from the training process. The estimated diagnostic accuracy was based on predictions from this final regression model. Once the class probabilities had been obtained for the individual samples from a sampling, they were combined to give class probabilities for the sampling as a whole. A separate threshold (also determined during the cross-validation training process) was used to predict the class of the sampling from the combined probability.

Diagnostic Accuracy Analysis Results

The logistic regression model developed used data from 143 *ULTRADIAN* participants (**Table 54**). The sample and sampling threshold used had a value of 0.15. The sensitivity and specificity of *U-Rhythm* in diagnosing PA were estimated to be 0.900 and 0.862 respectively, while their average (i.e. balanced accuracy) was equal to 0.881.

Table 54: U-Rhythm diagnostic accuracy sampling data

		Condition	
		Healthy	**Primary Aldosteronism**
Prediction	**Healthy**	106	2
	Primary Aldosteronism	17	18

Appendix 12: ULTRADIAN Questionnaire Regression Analysis Results

Table 55: EQ-5D-5L questionnaire regression analysis results

Question 1 – Mobility				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.65	0.27	-1.04	0.297
Age	1.04	0.02	2.43	0.015
Condition[b]				
Primary aldosteronism	5.38	2.56	3.54	<0.001
Cushing's syndrome	32.93	16.02	7.18	<0.001
Addison's disease	3.14	2.08	1.73	0.083
Number of observations = 331		Likelihood ratio chi^2(5) = 101.78		p-value < 0.001
Pseudo R^2 = 0.245		Log likelihood = -156.609		
Question 2 – Self-Care				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	1.46	0.63	0.88	0.378
Age	1.02	0.02	1.01	0.311
Condition[b]				
Primary aldosteronism	2.23	1.17	1.53	0.127
Cushing's syndrome	8.41	4.44	4.03	<0.001
Addison's disease	0.73	0.79	-0.29	0.774
Number of observations = 332		Likelihood ratio chi^2(5) = 24.14		p-value < 0.001
Pseudo R^2 = 0.090		Log likelihood = -122.519		

Question 3 – Usual Activities				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.61	0.21	-1.44	0.150
Age	1.02	0.01	1.43	0.152
Condition[b]				
Primary aldosteronism	13.57	5.48	6.46	<0.001
Cushing's syndrome	25.88	11.53	7.30	<0.001
Addison's disease	5.46	2.99	3.10	0.002
Number of observations = 332		Likelihood ratio chi^2(5) = 113.38		p-value < 0.001
Pseudo R^2 = 0.222		Log likelihood = -198.239		

Question 4 – Pain/Discomfort				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.74	0.20	-1.12	0.261
Age	1.01	0.01	1.28	0.200
Condition[b]				
Primary aldosteronism	6.58	2.16	5.75	<0.001
Cushing's syndrome	18.95	7.63	7.30	<0.001
Addison's disease	2.51	1.22	1.90	0.057
Number of observations = 332		Likelihood ratio chi^2(5) = 99.12		p-value < 0.001
Pseudo R^2 = 0.158		Log likelihood = -263.377		

Appendix 12: ULTRADIAN Questionnaire Regression Analysis Results

Question 5 – Anxiety/Depression				
Variable	**Odds Ratio**	**Standard Error**	**z**	**p-value**
Gender[a]	0.72	0.19	-1.22	0.221
Age	0.99	0.01	-1.37	0.170
Condition[b]				
Primary aldosteronism	5.23	1.77	4.89	<0.001
Cushing's syndrome	28.50	11.90	8.02	<0.001
Addison's disease	2.41	1.09	1.93	0.053
Number of observations = 331		Likelihood ratio chi^2(5) = 88.21		p-value < 0.001
Pseudo R^2 = 0.135		Log likelihood = -283.493		

*Reference categories: **a** – 'female'; **b** – 'healthy'*

***Only the first time of completion is reported.*

Appendix 13: CPRD Linkage Data Information

Integrated HES Admitted Patient Care Data

The *Hospital Episode Statistics (HES)* contains information on all hospital admissions, outpatient appointments, and Accident and Emergency (A&E) attendances reported by English National Health Service (NHS) healthcare providers, including acute hospital trusts, primary care trusts and mental health trusts. HES data are collected on a monthly basis and are managed by the NHS Digital, formerly known as the *Health and Social Care Information Centre* (248, 284, 288). Although Clinical Practice Research Datalink (CPRD) data have been collected since 1989, CPRD only links data with HES from 1997 and onwards (i.e. the point when the NHS number was introduced – used for data linkage) (248, 284). In HES, diagnostic data are recorded using ICD-10 codes (293), while information on procedures is coded using the *UK Office of Population, Census and Surveys classification (507)*. HES data are available in three forms – a) integrated, b) basic, and c) full HES Admitted Patient Care (APC) data – while outpatient, A&E and diagnostic imaging data can also be provided, when needed (248, 284).

In this study, the integrated HES APC data are used since these are the only HES data that can become available for free when access to CPRD data is provided. These data constitute the lowest level of the available HES data and contain secondary care patient information limited only to key elements such as medical diagnoses, episodes of care and hospitalisations (i.e. discharge date and ICD-10 codes associated with each admission). However, integrated HES data do not include complete hospital episode (i.e. admission dates, primary diagnoses, specialists seen, procedures undertaken), maternity, intensive or high dependency level of care, outpatient, A&E, and diagnostic imaging data, making them limited in the information that they provide (248, 284, 293).

During this analysis, HES data are used, when available, to identify the hospital admissions of the exposed and unexposed patients before and after the index date. Here, it should be mentioned that although the patient CPRD-identifier changes when the patient moves between GPs, the HES-identifier remains the same for the respective patient. This is useful when going back to the CPRD datasets and examining whether there are patients with

duplicate records. Of course, not all CPRD patients are eligible to be linked to HES (e.g. patient or practice region outside England; lack of a valid NHS identifier) (248, 284).

Area-Based Deprivation IMD Data

The *Index of Multiple Deprivation (IMD)* is a measure of deprivation assigned to each small English area (i.e. residential area or neighbourhood) based on its social and economic characteristics (i.e. income; education and skills; employment; living environment; access to services; crime), and housing indicators. IMD is assigned at both patient- and practice-level, and is mainly used to rank areas from least to most deprived (quintile, decile and 'twentile' groupings) rather than as an actual score itself. For the patient, the IMD is derived using their current or most recent postcode of residence, recorded at the GP practice, which is then used to assign a *lower layer super output area (LSOA)* of residence (i.e. predefined, geographical and consistently sized areas designed for the collection and publication of small area statistics). Similarly, the practice's location postcode is mapped to the LSOA boundaries. Afterwards, the LSOA is linked to the quintiles, deciles and 'twentiles' of the English Indices of Multiple Deprivation score, which have been calculated at LSOA level (248, 285).

IMD is provided for all patients and their practices in the CPRD study population when the following criteria are met: a) the practice has consented to participate in the CPRD linkage scheme; b) the patient has not requested against the use of their personal confidential data to the NHS Digital; c) the patient and/or practice have a valid full postcode of residence or location recorded in CPRD GOLD; and d) the postcode can be assigned to an LSOA in England. Here, it is worth noting that the IMDs have been regularly updated (2000, 2004, 2007, 2010 and 2015) due to changing LSOA boundaries and changing deprivation levels in different areas. For ease of comparison across areas and over time, and to prevent the deductive disclosure of a patient's area of residence, generally just one version (quintiles, deciles or 'twentiles') of the year's IMD is provided and used throughout an analysis (248, 285).

In this study, the quintiles of the 2015 area-based deprivation IMD (IMD2015) data were provided. For the analysis, the IMD2015 are used as a proxy of the individuals' socioeconomic status to assess the association between deprivation (e.g. income, living environment, health deprivation, disabilities) and the health services that the patients used.

ONS Mortality Data

In England and Wales, every death must be certified and reported to the *General Register Office (GRO)*[37] *(219)* by providing the relevant individual's information (e.g. by a close relative to the deceased) and a medical certificate with the cause(s) of death, signed by a registered medical practitioner/doctor. Deaths should be registered within five days from the date of death unless there is a serious reason for delay. Regular validation checks, receipt and diagnostic tests are then performed to ensure that the information collected is correct and highlight any inconsistences before the data are finally recorded within the *Office for National Statistics (ONS)* mortality dataset (248, 286, 293).

ONS mortality data are provided if the practice consents to have their data linked and when the following criteria are met. Firstly, the patient must have at least one day of up-to-standard (UTS) follow-up which coincides with the data collection period for the linked dataset. For the most recent linked ONS mortality date, the data collection period relates to the date of death registration (and not the date of death) and runs from January 1998 to April 2017 [**NOTE:** Any analysis using ONS mortality data needs to ensure that person time at risk does not include time before 01 January 1998 (319)]. Secondly, the patient must have died during the study period to be included within the ONS death registration dataset. Thirdly, a set of identifiers must be present in both the CPRD GOLD and death registration record dataset to provide a minimum level of confidence in the linkage algorithm [**NOTE:** Deaths registered between 1996-2000 are coded using ICD-9, with ICD-10 being used thereafter. Moreover, the algorithm for determining the overall cause of death from death certificates changed from 2011 onwards and this can potentially impact specific-cause mortality data] (248, 286, 293).

In this analysis, ONS mortality data are used to cross-validate date of death information contained in the GP records (CPRD), and more accurately identify censoring of patient data when describing and comparing healthcare usage and associated costs.

[37] Official governmental agency responsible for keeping English and Welsh records regarding births, adoptions, marriages, civil partnerships and deaths. GRO normally works in partnership with the Local Registration Service.

Appendix 14: CPRD Study GP Read Codes

Table 56: GP Read Codes used to identify patients with primary aldosteronism

Read Code	Read Term
C151.11	Aldosteronism
C151.13	Conn's syndrome
C151011	Conn's syndrome
C151.00	Hyperaldosteronism
C151000	Primary hyperaldosteronism
C151z00	Hyperaldosteronism NOS
Cyu4700	[X]Other hyperaldosteronism

Appendix 15: CPRD Study Healthcare Resource Use Outcomes

Table 57: Average annual number of primary care consultations (SD, range) for PA patients before and after index date

Time / Consultations	5-10 years before index date (n = 370)	1-5 years before index date (n = 447)	0-1 years before index date (n = 471)	0-1 years after index date (n = 468)	1-5 years after index date (n = 383)	5-10 years after index date (n = 220)
All	6.84 (6.50, 0-47.2)	9.58 (6.93, 0-59.75)	14.90 (10.24, 0-73)	14.40 (12.49, 0-123)	10.55 (8.15, 0-46.25)	10.63 (8.78, 0-50)
Clinic	6.49 (6.23, 0-47.2)	9.03 (6.64, 0-59.5)	13.55 (9.51, 0-68)	12.89 (11.35, 0-116)	9.47 (7.33, 0-43.5)	9.66 (7.84, 0-45.81)
Home	0.01 (0.04, 0-0.6)	0.02 (0.14, 0-1.75)	0.04 (0.31, 0-5)	0.06 (0.47, 0-7)	0.08 (0.51, 0-6.69)	0.15 (0.82, 0-8.6)
Telephone	0.28 (0.73, 0-6.2)	0.47 (0.90, 0-8)	1.19 (2.39, 0-19)	1.30 (2.96, 0-27)	0.89 (1.78, 0-13.86)	0.70 (1.24, 0-7.55)
Out-of-hours	0.07 (0.43, 0-7.8)	0.06 (0.24, 0-3.25)	0.13 (0.51, 0-6)	0.15 (0.74, 0-10.77)	0.11 (0.50, 0-8.11)	0.12 (0.41, 0-4.4)
Administrative tasks	7.79 (9.50, 0-79.2)	16.23 (14.89, 0-96.25)	31.05 (25.75, 0-189)	35.10 (27.31, 0-223.95)	27.96 (21.11, 0-133.80)	10.63 (8.78, 0-50)

Table 58: Average annual number of primary care consultations (SD, range) for non-PA patients before and after index date

Time / Consultations	5-10 years before index date (n = 736)	1-5 years before index date (n = 898)	0-1 years before index date (n = 945)	0-1 years after index date (n = 932)	1-5 years after index date (n = 774)	5-10 years after index date (n = 444)
All	4.43 (4.92, 0-43)	5.39 (5.61, 0-49)	5.76 (6.60, 0-44)	6.17 (7.60, 0-61)	6.39 (6.82, 0-50.5)	7.54 (7.82, 0-65.08)
Clinic	4.20 (4.76, 0-43)	5.05 (5.35, 0-46.25)	5.30 (6.14, 0-44)	5.54 (6.59, 0-57)	5.68 (5.91, 0-47.5)	6.72 (6.85, 0-49.79)
Home	0.00 (0.03, 0-0.6)	0.02 (0.14, 0-2.75)	0.02 (0.33, 0-9)	0.07 (0.79, 0-20.54)	0.10 (0.95, 0-19.09)	0.14 (0.86, 0-11.8)
Telephone	0.21 (0.60, 0-8.6)	0.29 (0.78, 0-9.25)	0.39 (1.22, 0-15)	0.48 (1.54, 0-19.66)	0.54 (1.42, 0-16.45)	0.62 (1.55, 0-16.42)
Out-of-hours	0.02 (0.10, 0-1.77)	0.03 (0.15, 0-1.86)	0.05 (0.28, 0-3.69)	0.08 (0.47, 0-8.11)	0.08 (0.34, 0-4)	0.06 (0.21, 0-2.34)
Administrative tasks	4.64 (7.03, 0-42.8)	8.67 (11.16, 0-83.39)	11.33 (14.15, 0-97)	13.72 (17.27, 0-112)	15.48 (17.07, 0-131.61)	7.54 (7.82, 0-65.08)

Table 59: Average annual number of primary care testing (SD, range) for PA patients before and after index date

Time / Tests	5-10 years before index date (n = 370)	1-5 years before index date (n = 447)	0-1 years before index date (n = 471)	0-1 years after index date (n = 468)	1-5 years after index date (n = 383)	5-10 years after index date (n = 220)
All	2.14 (2.83, 0-15.21)	4.84 (5.07, 0-32.5)	9.28 (9.31, 0-72)	8.17 (9.55, 0-86)	6.96 (6.95, 0-42)	7.72 (6.58, 0-31.98)
Biochemistry	1.04 (1.46, 0-10.93)	2.39 (2.44, 0-16.5)	5.12 (5.11, 0-31)	4.50 (5.34, 0-56)	3.39 (3.17, 0-24.75)	3.44 (2.86, 0-14.70)
Diagnostic imaging	0.15 (0.27, 0-1.8)	0.28 (0.46, 0-3.20)	0.49 (0.88, 0-6)	0.42 (0.82, 0-4)	0.31 (0.54, 0-4)	0.43 (0.68, 0-4.23)
Other diagnostic	0.12 (0.30, 0-2.6)	0.25 (0.47, 0-4.5)	0.46 (1.09, 0-11)	0.23 (0.70, 0-6.89)	0.27 (0.65, 0-5.36)	0.33 (0.66, 0-4.95)
Haematology	0.29 (0.53, 0-4.8)	0.69 (1.22, 0-13.75)	1.11 (2.23, 0-21)	1.12 (2.70, 0-26)	1.29 (2.69, 0-22.5)	1.49 (2.60, 0-16.8)
Microbiology	0.17 (0.34, 0-2)	0.33 (0.67, 0-7.75)	0.45 (1.09, 0-12)	0.42 (1.15, 0-12)	0.42 (0.77, 0-9.37)	0.53 (0.92, 0-5.92)

Appendix 15: CPRD Study Healthcare Resource Use Outcomes

Time / Tests	5-10 years before index date (n = 370)	1-5 years before index date (n = 447)	0-1 years before index date (n = 471)	0-1 years after index date (n = 468)	1-5 years after index date (n = 383)	5-10 years after index date (n = 220)
Other pathological	0.18 (0.54, 0-6.2)	0.26 (0.79, 0-13)	0.51 (1.39, 0-14)	0.43 (1.35, 0-17)	0.32 (0.71, 0-6.36)	0.30 (0.62, 0-3.74)
Serology & immunology	0.06 (0.17, 0-1.2)	0.16 (0.41, 0-4.5)	0.26 (0.80, 0-8)	0.19 (0.64, 0-6)	0.24 (0.61, 0-6.82)	0.26 (0.46, 0-3.82)
Other	0.14 (0.44, 0-4.42)	0.48 (1.08, 0-12)	0.87 (2.12, 0-26)	0.86 (2.33, 0-27)	0.74 (1.55, 0-13.5)	0.96 (1.53, 0-7.88)

Table 60: Average annual number of primary care testing (SD, range) for non-PA patients before and after index date

Time / Tests	5-10 years before index date (n = 736)	1-5 years before index date (n = 898)	0-1 years before index date (n = 945)	0-1 years after index date (n = 932)	1-5 years after index date (n = 774)	5-10 years after index date (n = 444)
All	1.41 (2.62, 0-24.8)	2.33 (3.61, 0-25.25)	3.00 (5.28, 0-60)	3.30 (5.60, 0-43)	3.76 (5.05, 0-44.42)	5.03 (6.81, 0-73.45)
Biochemistry	0.60 (1.33, 0-15.2)	1.01 (1.63, 0-14.25)	1.25 (2.14, 0-15)	1.35 (2.44, 0-20)	1.50 (1.98, 0-11.5)	1.99 (2.28, 0-15.28)
Diagnostic imaging	0.12 (0.28, 0-2.80)	0.18 (0.37, 0-3.25)	0.20 (0.55, 0-4)	0.27 (0.69, 0-6)	0.26 (0.47, 0-4.66)	0.31 (0.49, 0-3.14)
Other diagnostic	0.06 (0.22, 0-2.4)	0.11 (0.38, 0-5.75)	0.15 (0.61, 0-6)	0.15 (0.67, 0-9)	0.20 (0.55, 0-5)	0.32 (0.83, 0-7.69)
Haematology	0.20 (0.60, 0-8.6)	0.40 (1.11, 0-11.25)	0.54 (1.79, 0-28)	0.62 (1.95, 0-27)	0.80 (2.27, 0-27.30)	1.02 (2.80, 0-30.27)
Microbiology	0.14 (0.43, 0-7)	0.19 (0.46, 0-4.75)	0.25 (0.73, 0-9)	0.30 (0.84, 0-8)	0.29 (0.62, 0-6)	0.40 (1.01, 0-16.81)

Appendix 15: CPRD Study Healthcare Resource Use Outcomes

Time / Tests	5-10 years before index date (n = 736)	1-5 years before index date (n = 898)	0-1 years before index date (n = 945)	0-1 years after index date (n = 932)	1-5 years after index date (n = 774)	5-10 years after index date (n = 444)
Other pathological	0.11 (0.45, 0-7)	0.14 (0.58, 0-9.25)	0.20 (1.26, 0-29)	0.21 (1.09, 0-20)	0.20 (0.55, 0-7.5)	0.20 (0.55, 0-4.88)
Serology & immunology	0.05 (0.16, 0-1.87)	0.09 (0.26, 0-2.69)	0.12 (0.53, 0-10)	0.13 (0.53, 0-7.02)	0.16 (0.59, 0-11)	0.23 (0.87, 0-12.52)
Other	0.14 (0.55, 0-9)	0.22 (0.58, 0-6.25)	0.29 (0.86, 0-10)	0.26 (0.90, 0-11)	0.35 (0.77, 0-7.5)	0.57 (1.50, 0-23.32)

Table 61: Average annual number of important for PA primary care testing (SD, range) for PA patients before and after index date

Time / Tests	5-10 years before index date (n = 370)	1-5 years before index date (n = 447)	0-1 years before index date (n = 471)	0-1 years after index date (n = 468)	1-5 years after index date (n = 383)	5-10 years after index date (n = 220)
Blood (relevant to PA)[38]	0.51 (0.76, 0-4.6)	1.23 (1.32, 0-10.45)	3.02 (3.55, 0-26)	2.94 (4.27, 0-54)	1.78 (1.88, 0-12.75)	1.66 (1.52, 0-9.96)
Creatinine blood	0.42 (0.72, 0-4.75)	1.09 (1.29, 0-10.45)	2.69 (3.37, 0-24)	2.68 (4.03, 0-54)	1.67 (1.84, 0-12.5)	1.59 (1.48, 0-9.96)
Serum aldosterone & serum renin	0.00 (0.03, 0-0.48)	0.02 (0.11, 0-1.19)	0.19 (0.51, 0-3)	0.03 (0.18, 0-2)	0.01 (0.06, 0-0.94)	0.01 (0.05, 0-0.61)
Urine	0.04 (0.18, 0-1.6)	0.14 (0.38, 0-3.58)	0.29 (0.66, 0-4)	0.22 (0.62, 0-6)	0.25 (0.47, 0-2.11)	0.35 (0.55, 0-3.27)
Chest X-ray	0.03 (0.10, 0-0.60)	0.05 (0.14, 0-1)	0.08 (0.31, 0-3)	0.06 (0.27, 0-2.90)	0.05 (0.16, 0-1.28)	0.07 (0.19, 0-1.24)
Blood pressure	0.02 (0.08, 0-0.8)	0.03 (0.13, 0-1.02)	0.08 (0.42, 0-4)	0.02 (0.19, 0-3)	0.01 (0.08, 0-0.94)	0.01 (0.05, 0-0.45)

[38] PA-relevant blood tests include: electrolytes, glucose, lipids, potassium and urea

Appendix 15: CPRD Study Healthcare Resource Use Outcomes

Time / Tests	5-10 years before index date (n = 370)	1-5 years before index date (n = 447)	0-1 years before index date (n = 471)	0-1 years after index date (n = 468)	1-5 years after index date (n = 383)	5-10 years after index date (n = 220)
Electrocardiogram	0.05 (0.14, 0-1.37)	0.10 (0.21, 0-1.25)	0.17 (0.46, 0-4)	0.07 (0.30, 0-2.30)	0.06 (0.16, 0-1.10)	0.09 (0.21, 0-1.60)
Other	1.90 (2.59, 0-14)	4.34 (4.81, 0-32.25)	7.84 (8.54, 0-68)	7.15 (9.01, 0-86)	6.49 (6.72, 0-42)	7.32 (6.36, 0-30.15)

Table 62: Average annual number of important for PA primary care testing (SD, range) for non-PA patients before and after index date

Time / Tests	5-10 years before index date (n = 736)	1-5 years before index date (n = 898)	0-1 years before index date (n = 945)	0-1 years after index date (n = 932)	1-5 years after index date (n = 774)	5-10 years after index date (n = 444)
Blood (relevant to PA)[39]	0.23 (0.62, 0-9.2)	0.42 (0.77, 0-7)	0.51 (1.04, 0-10)	0.60 (1.23, 0-13)	0.66 (0.98, 0-10)	0.89 (1.24, 0-12.52)
Creatinine blood	0.17 (0.50, 0-7)	0.34 (0.68, 0-7)	0.44 (0.95, 0-10)	0.53 (1.15, 0-13)	0.58 (0.91, 0-10.5)	0.82 (1.18, 0-12.52)
Serum aldosterone & serum renin	-	-	0.00 (0.03, 0-1)	-	-	-
Urine	0.02 (0.13, 0-1.55)	0.05 (0.22, 0-2.5)	0.06 (0.39, 0-8)	0.07 (0.32, 0-3)	0.09 (0.27, 0-2.25)	0.17 (0.41, 0-2.65)
Chest X-ray	0.02 (0.10, 0-0.81)	0.03 (0.13, 1.5)	0.05 (0.25, 0-3)	0.03 (0.20, 0-2.92)	0.06 (0.22, 0-3.48)	0.07 (0.19, 0-1.79)
Blood pressure	0.00 (0.02, 0-0.4)	0.00 (0.05, 0-1.25)	0.00 (0.08, 0-2)	0.00 (0.08, 0-2)	0.00 (0.05, 0-0.96)	0.01 (0.07, 0-1)

[39] PA-relevant blood tests include: electrolytes, glucose, lipids, potassium and urea

Appendix 15: CPRD Study Healthcare Resource Use Outcomes

Time / Tests	5-10 years before index date (n = 736)	1-5 years before index date (n = 898)	0-1 years before index date (n = 945)	0-1 years after index date (n = 932)	1-5 years after index date (n = 774)	5-10 years after index date (n = 444)
Electrocardiogram	0.03 (0.10, 0-0.94)	0.03 (0.11, 0-1.5)	0.05 (0.22, 0-2)	0.04 (0.25, 0-2.25)	0.05 (0.15, 0-1.25)	0.08 (0.23, 0-2.45)
Other	1.32 (2.49, 0-24.8)	2.19 (3.44, 0-25.25)	2.83 (5.12, 0-60)	3.13 (5.39, 0-43)	3.58 (4.90, 0-43.35)	4.77 (6.58, 0-71.46)

Table 63: Average annual number of primary care drug prescriptions (SD, range) for PA patients before and after index date

Time / Medications	5-10 years before index date (n = 370)	1-5 years before index date (n = 447)	0-1 years before index date (n = 471)	0-1 years after index date (n = 468)	1-5 years after index date (n = 383)	5-10 years after index date (n = 220)
All	9.18 (10.90, 0-68.4)	12.55 (10.90, 0-64.25)	18.67 (14.55, 0-86)	18.68 (16.08, 0-104.44)	16.76 (15.34, 0-86.25)	18.48 (19.35, 0-139.22)
Mineralocorticoid receptor antagonists	0.34 (1.37, 0-11.6)	0.43 (1.25, 0-14.25)	1.14 (2.09, 0-18)	2.26 (2.53, 0-14.04)	1.39 (1.79, 0-13.73)	1.30 (1.53, 0-6.95)
Other diuretics	0.43 (1.08, 0-9)	0.53 (1.11, 0-8.75)	0.65 (1.47, 0-12)	0.44 (1.23, 0-10)	0.44 (1.14, 0-13.62)	0.50 (1.04, 0-6.29)
Other antihypertensive	3.24 (4.63, 0-33.8)	4.90 (5.22, 0-37.75)	7.48 (6.39, 0-35)	6.39 (5.98, 0-31)	5.43 (5.35, 0-33.60)	5.73 (6.75, 0-70.26)
Other	5.16 (8.20, 0-63)	6.70 (7.63, 0-59)	9.41 (11.27, 0-66)	9.60 (12.43, 0-93.34)	9.50 (11.65, 0-70.75)	10.94 (15.24, 0-133.51)

Table 64: Average annual number of primary care drug prescriptions (SD, range) for non-PA patients before and after index date

Time / Medications	5-10 years before index date (n = 736)	1-5 years before index date (n = 898)	0-1 years before index date (n = 945)	0-1 years after index date (n = 932)	1-5 years after index date (n = 774)	5-10 years after index date (n = 444)
All	5.02 (9.82, 0-126.8)	6.53 (9.98, 0-81)	7.44 (11.67, 0-93)	7.93 (12.25, 0-102.95)	8.73 (12.86, 0-126.47)	11.17 (13.37, 0-102.56)
Mineralocorticoid receptor antagonists	0.05 (0.59, 0-11.6)	0.05 (0.47, 0-8.64)	0.05 (0.64, 0-13)	0.06 (0.63, 0-12)	0.04 (0.38, 0-8.5)	0.07 (5.56, 0-9.43)
Other diuretics	0.12 (0.79, 0-11.6)	0.15 (0.69, 0-9.5)	0.17 (0.74, 0-10)	0.20 (0.84, 0-9)	0.22 (0.76, 0-7.74)	0.26 (0.71, 0-7.10)
Other antihypertensive	1.04 (3.94, 0-56.4)	1.41 (3.62, 0-31.25)	1.62 (3.82, 0-29)	1.58 (3.82, 0-48)	1.86 (3.43, 0-22)	2.46 (3.65, 0-23.23)
Other	3.82 (6.85, 0-75)	4.93 (7.61, 0-70.5)	5.61 (9.19, 0-80)	6.09 (10.12, 0-102.95)	6.62 (10.99, 0-126.47)	8.38 (11.34, 0-99.61)

Table 65: Average annual number of secondary care referrals (SD, range) for PA patients before and after index date

Time / Referrals	5-10 years before index date (n = 370)	1-5 years before index date (n = 447)	0-1 years before index date (n = 471)	0-1 years after index date (n = 468)	1-5 years after index date (n = 383)	5-10 years after index date (n = 220)
All	0.47 (0.64, 0-4)	0.76 (1.02, 0-8.25)	1.30 (1.98, 0-27)	0.89 (1.71, 0-20)	0.71 (0.92, 0-5.25)	0.76 (0.99, 0-5.81)
Cardiologist	0.02 (0.07, 0-0.66)	0.04 (0.13, 0-1.40)	0.09 (0.36, 0-3)	0.02 (0.15, 0-1)	0.02 (0.08, 0-0.75)	0.03 (0.11, 0-0.87)
Endocrinologist	0.00 (0.02, 0-0.2)	0.01 (0.06, 0-0.75)	0.08 (0.29, 0-2)	0.04 (0.22, 0-2.89)	0.01 (0.06, 0-0.77)	0.01 (0.07, 0-0.50)
General physician	0.04 (0.11, 0-0.8)	0.06 (0.18, 0-1.45)	0.16 (0.43, 0-3)	0.08 (0.33, 0-4.35)	0.04 (0.14, 0-1.42)	0.05 (0.23, 0-2.86)
Nephrologist	0.00 (0.01, 0-0.2)	0.01 (0.05, 0-0.5)	0.03 (0.23, 0-3)	0.00 (0.07, 0-1)	0.01 (0.05, 0-0.54)	0.01 (0.06, 0-0.83)
Urologist	0.01 (0.04, 0-0.4)	0.03 (0.10, 0-0.75)	0.04 (0.20, 0-2)	0.02 (0.14, 0-1.14)	0.02 (0.09, 0-0.91)	0.02 (0.10, 0-1.09)

Time / Referrals	5-10 years before index date (n = 370)	1-5 years before index date (n = 447)	0-1 years before index date (n = 471)	0-1 years after index date (n = 468)	1-5 years after index date (n = 383)	5-10 years after index date (n = 220)
General pathologist	0.12 (0.38, 0-3.2)	0.15 (0.68, 0-7.5)	0.18 (0.88, 0-8)	0.14 (0.82, 0-11)	0.08 (0.38, 0-4.5)	0.02 (0.18, 0-1.8)
Radiologist	0.04 (0.14, 0-1.08)	0.04 (0.14, 0-1)	0.04 (0.22, 0-2)	0.03 (0.19, 0-2)	0.04 (0.13, 0-1)	0.03 (0.12, 0-0.87)
Other specialty[40]	0.25 (0.39, 0-3.69)	0.43 (0.59, 0-5.25)	0.67 (1.44, 0-22)	0.56 (1.29, 0-19)	0.50 (0.68, 0-4)	0.58 (0.78, 0-4.98)

[40] Other specialty includes referrals to all other secondary care physicians, such as dermatologist, ophthalmologist, surgeon etc.

Table 66: Average annual number of secondary care referrals (SD, range) for non-PA patients before and after index date

Time / Referrals	5-10 years before index date (n = 736)	1-5 years before index date (n = 898)	0-1 years before index date (n = 945)	0-1 years after index date (n = 932)	1-5 years after index date (n = 774)	5-10 years after index date (n = 444)
All	0.38 (0.70, 0-6.6)	0.41 (0.66, 0-6.75)	0.44 (0.99, 0-11)	0.46 (1.08, 0-12.03)	0.48 (0.85, 0-13.15)	0.51 (0.70, 0-3.92)
Cardiologist	0.01 (0.05, 0-0.94)	0.01 (0.04, 0-0.67)	0.01 (0.11, 0-1)	0.01 (0.11, 0-2)	0.02 (0.07, 0-0.60)	0.01 (0.08, 0-0.86)
Endocrinologist	0.00 (0.02, 0-0.4)	0.00 (0.01, 0-0.25)	-	0.00 (0.05, 0-1)	0.00 (0.02, 0-0.25)	0.00 (0.02, 0-0.26)
General physician	0.02 (0.07, 0-0.61)	0.02 (0.08, 0-1.25)	0.02 (0.14, 0-2)	0.02 (0.18, 0-2.85)	0.02 (0.09, 0-0.75)	0.04 (0.19, 0-2.08)
Nephrologist	-	0.00 (0.03, 0-0.91)	-	0.00 (0.05, 0-1)	0.00 (0.02, 0-0.25)	0.00 (0.08, 0-1.71)
Urologist	0.01 (0.05, 0-0.56)	0.01 (0.06, 0-0.56)	0.01 (0.08, 0-1)	0.01 (0.12, 0-1.43)	0.02 (0.08, 0-0.98)	0.01 (0.07, 0-0.79)

Time / Referrals	5-10 years before index date (n = 736)	1-5 years before index date (n = 898)	0-1 years before index date (n = 945)	0-1 years after index date (n = 932)	1-5 years after index date (n = 774)	5-10 years after index date (n = 444)
General pathologist	0.09 (0.37, 0-6.2)	0.06 (0.30, 0-5.25)	0.07 (0.53, 0-10)	0.05 (0.45, 0-7)	0.06 (0.49, 0-11.51)	0.02 (0.14, 0-2.13)
Radiologist	0.03 (0.11, 0-0.8)	0.02 (0.10, 0-1.52)	0.03 (0.21, 0-2)	0.03 (0.23, 0-4.28)	0.04 (0.14, 0-1.5)	0.03 (0.09, 0-0.81)
Other specialty[41]	0.23 (0.51, 0-6.52)	0.29 (0.50, 0-4.25)	0.30 (0.68, 0-5)	0.33 (0.78, 0-8.20)	0.33 (0.56, 0-7.20)	0.39 (0.59, 0-3.55)

[41] Other specialty includes referrals to all other secondary care physicians, such as dermatologist, ophthalmologist, surgeon etc.

Table 67: Average annual number of hospital admissions (SD, range) for PA patients before and after index date

Time	5-10 years before first diagnosis (n = 228)	1-5 years before first diagnosis (n = 276)	0-1 years before first diagnosis (n = 292)	0-1 years after first diagnosis (n = 288)	1-5 years after first diagnosis (n = 235)	5-10 years after first diagnosis (n = 139)
Admissions	0.13 (0.28, 0-1.6)	0.28 (0.64, 0-8)	0.89 (1.30, 0-8)	1.10 (2.41, 0-22)	0.68 (2.15, 0-30.75)	1.68 (11.59, 0-133.25)
Causes:						
Cardiovascular	0.03 (0.12, 0-1)	0.13 (0.28, 0-2.25)	0.64 (1.09, 0-8)	0.75 (1.81, 0-21)	0.32 (0.68, 0-5.93)	0.36 (0.82, 0-5.79)
Endocrine	0.01 (0.05, 0-0.4)	0.06 (0.16, 0-1.25)	0.38 (0.72, 0-4)	0.65 (1.97, 0-22)	0.26 (0.66, 0-5.93)	0.32 (1.78, 0-20.63)
Renal	0.02 (0.07, 0-0.6)	0.03 (0.12, 0-1)	0.09 (0.39, 0-4)	0.06 (0.34, 0-4.35)	0.20 (2.02, 0-30.75)	1.11 (11.19, 0-131.60)
Other	0.12 (0.27, 0-1.6)	0.25 (0.62, 0-8)	0.69 (1.18, 0-8)	0.85 (2.11, 0-22)	0.62 (2.14, 0-30.75)	1.64 (11.59, 0-133.25)

Table 68: Average annual number of hospital admissions (SD, range) for non-PA patients before and after index date

Time	5-10 years before first diagnosis (n = 455)	1-5 years before first diagnosis (n = 558)	0-1 years before first diagnosis (n = 584)	0-1 years after first diagnosis (n = 575)	1-5 years after first diagnosis (n = 474)	5-10 years after first diagnosis (n = 267)
Admissions	0.10 (0.26, 0-2.2)	0.16 (0.47, 0-7.32)	0.20 (0.67, 0-6)	0.36 (1.29, 0-14.22)	0.35 (1.00, 0-11.25)	0.73 (3.41, 0-52.2)
Causes:						
Cardiovascular	0.03 (0.13, 0-1.4)	0.04 (0.19, 0-2.69)	0.06 (0.30, 0-4)	0.14 (0.77, 0-11.06)	0.13 (0.38, 0-3.26)	0.40 (2.92, 0-46.8)
Endocrine	0.01 (0.06, 0-0.6)	0.02 (0.16, 0-3.25)	0.03 (0.26, 0-4)	0.09 (0.54, 0-6)	0.06 (0.25, 0-2.72)	0.10 (0.36, 0-3.53)
Renal	0.01 (0.05, 0-0.41)	0.01 (0.12, 0-2.22)	0.03 (0.25, 0-4)	0.04 (0.29, 0-5.84)	0.04 (0.14, 0-1.83)	0.25 (2.87, 0-46.6)
Other	0.09 (0.24, 0-2.2)	0.14 (0.45, 0-7.32)	0.18 (0.64, 0-6)	0.33 (1.28, 0-14.22)	0.33 (1.00, 0-11.25)	0.71 (3.41, 0-52.2)

Appendix 16: CPRD Study Healthcare Use Prediction Additional Results

Table 69: Prediction of total primary care consultations using negative binomial regression

Variable	IRR	Standard Error	z	p-value
Event time				
1-5 years before	1.226	0.042	5.99	<0.001
1 year before & after	1.326	0.055	6.84	<0.001
1-5 years after	1.446	0.064	8.28	<0.001
5-10 years after	1.703	0.079	11.49	<0.001
PA	1.475	0.084	6.82	<0.001
Event time # PA				
1-5 years before	1.213	0.062	3.77	<0.001
1 year before & after	1.761	0.108	9.23	<0.001
1-5 years after	1.162	0.075	2.33	0.020
5-10 years after	0.980	0.069	-0.29	0.774
Age at index date	1.016	0.001	11.77	<0.001
Gender	0.719	0.028	-8.51	<0.001
Comorbidities				
One	1.594	0.078	9.55	<0.001
Two	1.678	0.175	4.96	<0.001
Three	2.149	0.434	3.79	<0.001
Deprivation				
Less deprived	1.007	0.070	0.09	0.926
Average	0.997	0.055	-0.06	0.951
Deprived	0.984	0.079	-0.21	0.838
Most deprived	1.096	0.082	1.22	0.223
Constant	1.876	0.182	6.49	<0.001
Observation time	1 (exposure)			
Number of observations = 5,691		Number of groups = 1,419		
Wald $chi^2(18)$ = 1,236.94		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Table 70: Prediction of total diagnostic testing using negative binomial regression

Variable	IRR	Standard Error	z	p-value
Event time[a]				
1-5 years before	1.693	0.093	9.55	<0.001
1 year before & after	2.202	0.146	11.94	<0.001
1-5 years after	2.831	0.194	15.23	<0.001
5-10 years after	4.186	0.289	20.72	<0.001
PA[b]	1.480	0.138	4.19	<0.001
Event time # PA[c]				
1-5 years before	1.427	0.113	4.50	<0.001
1 year before & after	2.026	0.182	7.85	<0.001
1-5 years after	1.273	0.120	2.55	0.011
5-10 years after	1.143	0.108	1.42	0.157
Age at index date	1.024	0.002	12.06	<0.001
Gender[d]	0.653	0.037	-7.63	<0.001
Comorbidities[e]				
One	1.884	0.130	9.21	<0.001
Two	2.247	0.396	4.59	<0.001
Three	2.083	0.579	2.64	0.008
Deprivation[f]				
Less deprived	1.054	0.105	0.53	0.598
Average	0.956	0.073	-0.59	0.554
Deprived	0.858	0.090	-1.46	0.143
Most deprived	0.912	0.094	-0.89	0.372
Constant	0.394	0.058	-6.35	<0.001
Observation time[g]	1 (exposure)			
Number of observations = 5,691		Number of groups = 1,419		
Wald $chi^2(18)$ = 1,805.04		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Reference categories:* ***a, c *– '5-10 years before index date';* ***b*** *– 'non-PA';* ***d*** *– 'female';* ***e*** *– 'no comorbidities';* ***f*** *– 'least deprived';* ***g*** *– consideration of patient's GP registration time*

Table 71: Prediction of total drug prescriptions using negative binomial regression

Variable	IRR	Standard Error	z	p-value
Event time[a]				
1-5 years before	1.325	0.078	4.76	<0.001
1 year before & after	1.563	0.104	6.70	<0.001
1-5 years after	1.815	0.145	7.44	<0.001
5-10 years after	2.340	0.183	10.88	<0.001
PA[b]	1.803	0.157	6.77	<0.001
Event time # PA[c]				
1-5 years before	1.125	0.082	1.61	0.106
1 year before & after	1.477	0.125	4.59	<0.001
1-5 years after	1.108	0.110	1.03	0.302
5-10 years after	0.916	0.095	-0.84	0.400
Age at index date	1.026	0.002	12.28	<0.001
Gender[d]	0.761	0.047	-4.41	<0.001
Comorbidities[e]				
One	1.741	0.101	9.53	<0.001
Two	1.728	0.242	3.90	<0.001
Three	1.968	0.420	3.17	0.002
Deprivation[f]				
Less deprived	1.160	0.122	1.41	0.159
Average	1.132	0.095	1.48	0.138
Deprived	1.171	0.144	1.29	0.198
Most deprived	1.470	0.177	3.19	0.001
Constant	0.942	0.150	-0.37	0.709
Observation time[g]	1 (exposure)			
Number of observations = 5,691		Number of groups = 1,419		
Wald $chi^2(18)$ = 1,054.64		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Reference categories:* ***a, c *– '5-10 years before index date';* ***b*** *– 'non-PA';* ***d*** *– 'female';* ***e*** *– 'no comorbidities';* ***f*** *– 'least deprived';* ***g*** *– consideration of patient's GP registration time*

Table 72: Prediction of total secondary care referrals using negative binomial regression

Variable	IRR	Standard Error	z	p-value
Event time[a]				
1-5 years before	1.076	0.075	1.04	0.297
1 year before & after	1.163	0.101	1.75	0.080
1-5 years after	1.288	0.110	2.96	0.003
5-10 years after	1.365	0.117	3.62	<0.001
PA[b]	1.219	0.119	2.02	0.043
Event time # PA[c]				
1-5 years before	1.510	0.161	3.86	<0.001
1 year before & after	2.051	0.258	5.70	<0.001
1-5 years after	1.208	0.145	1.57	0.116
5-10 years after	1.212	0.160	1.46	0.145
Age at index date	1.015	0.002	6.28	<0.001
Gender[d]	0.763	0.042	-4.95	<0.001
Comorbidities[e]				
One	1.355	0.096	4.28	<0.001
Two	1.383	0.270	1.66	0.097
Three	1.764	0.734	1.36	0.173
Deprivation[f]				
Less deprived	1.007	0.091	0.08	0.939
Average	0.908	0.068	-1.28	0.200
Deprived	0.764	0.074	-2.77	0.006
Most deprived	1.015	0.127	0.12	0.904
Constant	0.190	0.033	-9.55	<0.001
Observation time[g]	1 (exposure)			
Number of observations = 5,691		Number of groups = 1,419		
Wald $chi^2(18)$ = 298.08		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Reference categories:* ***a, c *– '5-10 years before index date';* ***b*** *– 'non-PA';* ***d*** *– 'female';* ***e*** *– 'no comorbidities';* ***f*** *– 'least deprived';* ***g*** *– consideration of patient's GP registration time*

Table 73: Prediction of total hospital admissions using negative binomial regression

Variable	IRR	Standard Error	z	p-value
Event time[a]				
1-5 years before	1.477	0.195	2.95	0.003
1 year before & after	2.580	0.433	5.64	<0.001
1-5 years after	3.371	0.636	6.44	<0.001
5-10 years after	7.021	1.890	7.24	<0.001
PA[b]	1.241	0.247	1.08	0.278
Event time # PA[c]				
1-5 years before	1.452	0.284	1.90	0.057
1 year before & after	3.163	0.715	5.09	<0.001
1-5 years after	1.578	0.491	1.47	0.143
5-10 years after	1.444	0.844	0.63	0.529
Age at index date	1.018	0.009	2.06	0.039
Gender[d]	0.954	0.130	-0.35	0.730
Comorbidities[e]				
One	1.736	0.341	2.81	0.005
Two	1.769	0.402	2.51	0.012
Three	2.385	1.469	1.41	0.158
Deprivation[f]				
Less deprived	1.049	0.172	0.29	0.773
Average	1.006	0.171	0.04	0.970
Deprived	1.195	0.347	0.61	0.539
Most deprived	1.374	0.256	1.71	0.088
Constant	0.031	0.019	-5.76	<0.001
Observation time[g]	1 (exposure)			
Number of observations = 3,508		Number of groups = 876		
Wald $chi^2(18)$ = 497.19		p-value < 0.001		

**Abbreviations:* *IRR: incident rate ratio; PA: primary aldosteronism*

Reference categories:* ***a, c *– '5-10 years before index date';* ***b*** *– 'non-PA';* ***d*** *– 'female';* ***e*** *– 'no comorbidities';* ***f*** *– 'least deprived';* ***g*** *– consideration of patient's GP registration time*

Appendix 17: CPRD Study Cost Prediction Analysis Additional Results

Table 74: Prediction of total healthcare costs (without hospital admissions) using GLM with log link function and gamma distribution

Variable	dy/dx	Standard Error	z	p-value
PA[a]	10,825.49	1,047.15	10.34	<0.001
Age at index date	442.08	41.69	10.60	<0.001
Gender[b]	-5,012.54	1,072.71	-4.67	<0.001
Comorbidities[c]				
One	9,858.12	1,331.87	7.40	<0.001
Two	7,952.42	2,924.30	2.72	0.007
Three	9,430.76	5,878.57	1.60	0.109
Years of GP registration	217.99	33.68	6.47	<0.001
Deprivation[d]				
Less deprived	905.89	1,518.70	0.60	0.551
Average	779.90	1,252.10	0.62	0.533
Deprived	1,299.26	1,812.59	0.72	0.473
Most deprived	4,662.70	2,074.16	2.25	0.025

Number of observations = 1,419	Residual df = 1,407
Scale parameter = 1.28	
Deviance = 1,429.17	(1/df) Deviance = 1.02
Pearson = 1,796.19	(1/df) Pearson = 1.28
AIC = 20.94	BIC = -8,782.42
Log pseudolikelihood = -14,846.48	

**Abbreviations:* *AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism*

Reference categories:* ***a *– 'non-PA';* ***b*** *– 'female';* ***c*** *– 'No comorbidities';* ***d*** *– 'least deprived'*

Table 75: Prediction of total healthcare costs (with hospital admissions) using GLM with log link function and gamma distribution

Variable	dy/dx	Standard Error	z	p-value
PA[a]	21,650.93	4,134.20	5.24	<0.001
Age at index date	1,072.89	195.59	5.49	<0.001
Gender[b]	-9,099.34	4,377.36	-2.08	0.038
Comorbidities[c]				
One	24,462.23	9,420.49	2.60	0.009
Two	16,417.57	7,386.13	2.22	0.026
Three	8,972.02	18,319.36	0.49	0.624
Years of GP registration	226.15	114.75	1.97	0.049
Deprivation[d]				
Less deprived	3,570.76	6,988.98	0.51	0.609
Average	901.69	5,377.22	0.17	0.867
Deprived	1,915.49	5,216.74	0.37	0.713
Most deprived	7,284.01	5,535.76	1.32	0.188
Number of observations = 876		Residual df = 864		
Scale parameter = 3.90				
Deviance = 1,215.17		(1/df) Deviance = 1.41		
Pearson = 3,369.15		(1/df) Pearson = 3.90		
AIC = 22.41		BIC = -4,638.75		
Log pseudolikelihood = -9,801.32				

Abbreviations: AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism

***Reference categories: **a** – 'non-PA'; **b** – 'female'; **c** – 'No comorbidities'; **d** – 'least deprived'*

Appendix 18: CPRD Study Sensitivity Analysis I Results

Table 76: Patient characteristics before index date (excluding one-year 'ghost' patients)

Patient Characteristics	PA (n = 470)	Non-PA (n = 858)	Difference between Cohorts (p-value)
Gender; n (%)			
Female	186 (39.57)	360 (41.96)	$chi^2(1) = 0.713$
Male	284 (60.43)	498 (58.04)	p = 0.399
Age at index date (years)[42]; mean (SD, range)	55.18 (13.94, 0-93)	55.79 (14.06, 0-93)	t(1326) = 0.767 p = 0.443
Ethnicity[43]; n (%)			
Black	24 (8.28)	4 (0.75)	chi^2 (3) = 56.250 p < 0.001
White	222 (76.55)	374 (70.43)	
Other	17 (5.86)	22 (4.14)	
Unknown	27 (9.31)	131 (24.67)	
Marital status; n (%)			
Co-habiting/Married	85 (18.09)	161 (18.76)	
Divorced/Separated/Widowed	11 (2.34)	9 (1.05)	$chi^2(4) = 4.320$
Single	18 (3.83)	30 (3.50)	p = 0.364
Data not entered	87 (18.51)	145 (16.90)	
Unknown	269 (57.23)	513 (59.79)	
Socioeconomic status based on residence area (quintiles)[44]; n (%)			
<20% (least deprived)	72 (24.74)	131 (24.67)	
20% - 40%	59 (20.27)	127 (23.92)	$chi^2(4) = 3.457$
40% - 60%	63 (21.65)	112 (21.09)	p = 0.484
60% - 80%	57 (19.59)	81 (15.25)	
>80% (most deprived)	40 (13.75)	80 (15.07)	

[42] Two PA and four non-PA patients were <12 months old. Age '0' means <12 months old at index date.

[43] Only for patients with HES data (290 PA, 531 non-PA).

[44] Only for patients with IMD data (290 PA, 531 non-PA).

Patient Characteristics	PA (n = 470)	Non-PA (n = 858)	Difference between Cohorts (p-value)
Number of comorbidities (based on Charlson's categories)[45]; n (%)			
0	346 (73.62)	713 (83.10)	$chi^2(3) = 17.448$
1	104 (22.13)	126 (14.69)	$p = 0.001$
2	17 (3.62)	16 (1.86)	
3	3 (0.64)	3 (0.35)	
Types of comorbidities (based on Charlson's categories); n (%)			
AIDS	-	1 (0.11)	
Cancer	9 (1.83)	19 (2.16)	
Cerebrovascular disease	14 (2.84)	6 (0.68)	
Chronic pulmonary disease	26 (5.27)	42 (4.77)	
Congestive heart disease	10 (2.03)	4 (0.45)	
Dementia	1 (0.20)	1 (0.11)	
Diabetes	47 (9.53)	51 (5.80)	
Diabetes with complications	7 (1.42)	7 (0.80)	$chi^2(14) = 46.980$
Hemiplegia	-	-	$p < 0.001$
Metastatic tumour	1 (0.20)	1 (0.11)	
Mild liver disease	-	2 (0.23)	
Moderate liver disease	-	-	
Myocardial infarction	6 (1.22)	2 (0.23)	
Peptic ulcer disease	-	-	
Peripheral vascular disease	3 (0.61)	8 (0.91)	
Renal disease	21 (4.26)	15 (1.70)	
Rheumatological disease	2 (0.41)	8 (0.91)	
None	346 (70.18)	713 (81.02)	

**Abbreviations:* *Non-PA: without primary aldosteronism; PA: primary aldosteronism; SD: standard deviation*

[45] The comorbidities that were present within two years before index date are demonstrated.

Table 77: Patient characteristics after index date (excluding one-year 'ghost' patients)

Patient Characteristics	PA (n = 470)	Non-PA (n = 858)	Difference between Cohorts (p-value)
Number of cases transferred out of the GP; n (%)	141 (30.00)	286 (33.33)	$chi^2(1) = 1.547$ p = 0.214
Number of deaths[46]; n (%)	70 (14.89)	130 (15.15)	$chi^2(1) = 0.016$ p = 0.900
Age at death (years); mean (SD, range)	76.54 (10.46, 44-99)	74.70 (11.52, 41-96)	t(198) = -1.114 p = 0.267
Underlying cause of death (ONS data only); n (%)			
Circulatory, endocrine or renal	20 (43.48)	27 (30.34)	$chi^2(1) = 2.308$
Other disease	26 (56.52)	62 (69.66)	p = 0.129
Other recorded cause of death (ONS data only); n (%)			
Circulatory, endocrine or renal	29 (63.04)	51 (57.30)	$chi^2(1) = 0.414$
Other disease	17 (36.96)	38 (42.70)	p = 0.520
Total time registered with GP (years); median (IQR, range)	20.89 (19.29, 1.95-85.34)	21.77 (18.93, 1.25-85.81)	z = 0.107 p = 0.915
Time registered with GP since index date (years); median (IQR, range)	5.31 (6.97, 0.03-25.90)	5.31 (6.56, 0.02-25.65)	z = -0.211 p = 0.833

**Abbreviations: GP: general practice; IQR: interquartile range; Non-PA: without primary aldosteronism; ONS: Office for National Statistics; PA: primary aldosteronism; SD: standard deviation*

[46] CPRD: 44 (9.36%) PA and 101 (11.77%) non-PA deaths
ONS: 46 (9.79%) PA and 89 (10.37%) non-PA deaths

Table 78: Average annual healthcare resource use (SD, range) for PA patients (excluding one-year 'ghost' patients)

Time / Resource Category	5-10 years before index date (n = 367)	1-5 years before index date (n = 444)	0-1 years before index date (n = 468)	0-1 years after index date (n = 465)	1-5 years after index date (n = 380)	5-10 years after index date (n = 217)
Primary care consultations	6.89 (6.51, 0-47.2)	9.62 (6.93, 0-59.75)	15.00 (10.21, 0-73)	14.47 (12.50, 0-123)	10.59 (8.16, 0-46.25)	10.69 (8.81, 0-50)
Diagnostic testing	2.15 (2.83, 0-15.21)	4.88 (5.07, 0-32.5)	9.34 (9.31, 0-72)	8.21 (9.56, 0-86)	7.00 (6.96, 0-42)	7.77 (6.60, 0-31.98)
Drug prescriptions	9.25 (10.91, 0-68.4)	12.64 (10.89, 0-64.25)	18.79 (14.52, 0-86)	18.78 (16.08, 0-104.44)	16.87 (15.35, 0-86.25)	18.69 (19.41, 0-139.22)
Secondary care referrals	0.48 (0.64, 0-4)	0.77 (1.02, 0-8.25)	1.30 (1.98, 0-27)	0.89 (1.71, 0-20)	0.71 (0.92, 0-5.25)	0.77 (1.00, 0-5.81)
Hospital admissions (HES patients only)	**(n = 226)** 0.13 (0.28, 0-1.6)	**(n = 274)** 0.28 (0.64, 0-8)	**(n = 290)** 0.89 (1.31, 0-8)	**(n = 286)** 1.11 (2.42, 0-22)	**(n = 233)** 0.69 (2.16, 0-30.75)	**(n = 137)** 1.70 (11.67, 0-133.25)

**Abbreviations: HES: Hospital Episode Statistics*

Table 79: Average annual healthcare resource use (SD, range) for non-PA patients (excluding one-year 'ghost' patients)

Time / Resource Category	5-10 years before index date (n = 669)	1-5 years before index date (n = 813)	0-1 years before index date (n = 857)	0-1 years after index date (n = 844)	1-5 years after index date (n = 697)	5-10 years after index date (n = 397)
Primary care consultations	4.74 (5.01, 0-43)	5.84 (5.69, 0-49)	6.35 (6.66, 0-44)	6.71 (7.76, 0-61)	6.95 (6.93, 0-50.5)	8.17 (7.96, 0-65.08)
Diagnostic testing	1.53 (2.71, 0-24.8)	2.55 (3.72, 0-25.25)	3.30 (5.45, 0-60)	3.60 (5.79, 0-43)	4.11 (5.19, 0-44.42)	5.49 (7.03, 0-73.45)
Drug prescriptions	5.47 (10.18, 0-126.8)	7.16 (10.28, 0-81)	8.21 (11.99, 0-93)	8.71 (12.61, 0-102.95)	9.62 (13.25, 0-126.47)	12.26 (13.69, 0-102.56)
Secondary care referrals	0.41 (0.73, 0-6.6)	0.44 (0.68, 0-6.75)	0.48 (1.03, 0-11)	0.50 (1.12, 0-12.03)	0.53 (0.88, 0-13.15)	0.55 (0.71, 0-3.92)
Hospital admissions (HES patients only)	**(n = 414)** 0.11 (0.27, 0-2.2)	**(n = 507)** 0.17 (0.49, 0-7.32)	**(n = 531)** 0.22 (0.70, 0-6)	**(n = 522)** 0.36 (1.21, 0-14.22)	**(n = 427)** 0.38 (1.05, 0-11.25)	**(n = 238)** 0.80 (3.61, 0-52.2)

**Abbreviations: HES: Hospital Episode Statistics*

Table 80: Average costs (SD; range) per healthcare resource category (excluding one-year 'ghost' patients)

Healthcare Resource Category	PA	Non-PA	Difference between Cohorts (p-value)
Primary care consultations	£3,982.71 (3,252.88; 0-25,278.61)	£2,355.44 (2,653.40; 0-41,740.34)	t(1326) = -9.847 p < 0.001
Diagnostic testing	£1,137.05 (1,240.29; 0-7,878.94)	£769.05 (1,026.54; 0-8,730.16)	t(1326) = -5.794 p < 0.001
Drug prescriptions	£15,754.01 (16,005.97; 0-112,841.60)	£9,134.54 (12,858.08; 0-103,273.70)	t(326) = -8.209 p < 0.001
Secondary care referrals	£1,588.04 (1,645.48; 0-12,013.75)	£1,003.79 (1,201.89; 0-11,236.24)	t(1326) = -7.403 p < 0.001
Hospital admissions (HES patients only)	£21,996.33 (60,145.92; 0-708,984.50)	£15,307.81 (75,845.08; 0-1,439,457.00)	t(819) = -1.335 p = 0.182
Total costs (all patients) (without hospital admissions)	**£22,461.80 (19,733.99; 695.99-135,365.40)**	**£13,262.82 (15,784.85; 0-123,878.30)**	**t(1326) = -9.274 p < 0.001**
Total costs (HES patients only)	**£45,480.20 (66,651.39; 2,381.61-755,263.70)**	**£28,179.02 (77,114.24; 0-1,480,879.00)**	**t(819) = -3.220 p = 0.001**

**Abbreviations:* HES: Hospital Episode Statistics; Non-PA: without primary aldosteronism; PA: primary aldosteronism*

Table 81: Median costs (IQR; range) per healthcare resource category (excluding one-year 'ghost' patients)

Healthcare Resource Category	PA	Non-PA	Difference between Cohorts (p-value)
Primary care consultations	£3,034.44 (3,546.21; 0-25,278.61)	£1,617.97 (2,511.70; 0-41,740.34)	z = -11.775 p < 0.001
Diagnostic testing	£751.76 (1,140.75; 0-7,878.94)	£418.57 (885.69; 0-8,730.16)	z = -7.803 p < 0.001
Drug prescriptions	£10.765.60 (15,004.25; 0-112,841.60)	£4,368.34 (10,585.37; 0-103,273.70)	z = -11.630 p < 0.001
Secondary care referrals	£1,136.77 (1,543.85; 0-12,013.75)	£624.44 (1,247.47; 0-11,236.24)	z = -8.859 p < 0.001
Hospital admissions (HES patients only)	£11,023.62 (17,434.39; 0-708,984.50)	£3,132.93 (11,889.07; 0-1,439,457.00)	z = -8.711 p < 0.001
Total costs (all patients) (without hospital admissions)	**£16,144.86 (19,601.50; 695.99-135,365.40)**	**£7,828.38 (14,974.83; 0-123,878.30)**	**z = -11.961 p < 0.001**
Total costs (HES patients only)	**£28,404.92 (35,756.48; 2,381.61-755,263.70)**	**£11,968.63 (25,448.28; 0-1,480,879.00)**	**z = -10.356 p < 0.001**

**Abbreviations:* HES: Hospital Episode Statistics; Non-PA: without primary aldosteronism; PA: primary aldosteronism*

Table 82: Prediction of total healthcare costs (without hospital admissions) using GLM with square root link function and gamma distribution (excluding one-year 'ghost' patients)

Variable	dy/dx	Standard Error	z	p-value
PA[a]	8,268.05	852.42	9.70	<0.001
Age at index date	303.12	23.51	12.89	<0.001
Gender[b]	-3,611.30	975.27	-3.70	<0.001
Comorbidities[c]				
One	8,752.28	1,334.94	6.56	<0.001
Two	8,704.41	2,740.69	3.18	0.001
Three	10,750.67	6,133.97	1.75	0.080
Years of GP registration	234.33	39.61	5.92	<0.001
Deprivation[d]				
Less deprived	527.03	1,324.75	0.40	0.691
Average	933.95	1,166.19	0.80	0.423
Deprived	792.92	1,714.57	0.46	0.644
Most deprived	4,004.11	1,841.16	2.17	0.030
Number of observations = 1,328		Residual df = 1,316		
Scale parameter = 1.11				
Deviance = 1,196.87		(1/df) Deviance = 0.91		
Pearson = 1,465.99		(1/df) Pearson = 1.11		
AIC = 21.11		BIC = -8,267.05		
Log pseudolikelihood = -14,004.53				

Abbreviations: AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism

Reference categories:* ***a *– 'non-PA';* ***b*** *– 'female';* ***c*** *– 'No comorbidities';* ***d*** *– 'least deprived'*

Table 83: Prediction of total healthcare costs (with hospital admissions) using GLM with square root link function and gamma distribution (excluding one-year 'ghost' patients)

Variable	dy/dx	Standard Error	z	p-value
PA[a]	17,403.03	3,282.07	5.30	<0.001
Age at index date	840.66	121.37	6.93	<0.001
Gender[b]	-8,089.98	3,281.47	-2.47	0.014
Comorbidities[c]				
One	21,991.93	10,414.64	2.11	0.035
Two	15,643.55	7,022.37	2.23	0.026
Three	11,247.96	19,966.35	0.56	0.573
Years of GP registration	262.14	126.55	2.07	0.038
Deprivation[d]				
Less deprived	4,539.06	6,360.59	0.71	0.475
Average	-771.75	3,789.45	-0.20	0.839
Deprived	2,662.81	4,730.69	0.56	0.574
Most deprived	8,098.14	4,951.42	1.64	0.102
Number of observations = 821		Residual df = 809		
Scale parameter = 3.76				
Deviance = 1,055.07		(1/df) Deviance = 1.30		
Pearson = 3,044.47		(1/df) Pearson = 3.76		
AIC = 22.56		BIC = -4,373.74		
Log pseudolikelihood = -9,247.83				

Abbreviations: AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism

***Reference categories: **a** – 'non-PA'; **b** – 'female'; **c** – 'No comorbidities'; **d** – 'least deprived'*

Table 84: Patient characteristics before index date (excluding five-year 'ghost' patients)

Patient Characteristics	PA (n = 472)	Non-PA (n = 929)	Difference between Cohorts (p-value)
Gender; n (%)			
Female	187 (39.62)	374 (40.26)	$chi^2(1) = 0.053$
Male	285 (60.38)	555 (59.74)	p = 0.817
Age at index date (years)[47]; mean (SD, range)	55.17 (13.94, 0-93)	55.18 (13.99, 0-93)	t(1399) = 0.008 p = 0.994
Ethnicity[48]; n (%)			
Black	24 (8.25)	5 (0.87)	$chi^2(3) = 63.746$ p < 0.001
White	223 (76.63)	393 (68.35)	
Other	17 (5.84)	22 (3.83)	
Unknown	27 (9.28)	155 (26.96)	
Marital status; n (%)			
Co-habiting/Married	86 (18.22)	168 (18.08)	
Divorced/Separated/Widowed	11 (2.33)	9 (0.97)	$chi^2(4) = 4.937$
Single	18 (3.81)	33 (3.55)	p = 0.294
Data not entered	87 (18.43)	159 (17.12)	
Unknown	270 (57.20)	560 (60.28)	
Socioeconomic status based on residence area (quintiles)[49]; n (%)			
<20% (least deprived)	72 (24.66)	142 (24.70)	
20% - 40%	60 (20.55)	134 (23.30)	$chi^2(4) = 3.243$
40% - 60%	63 (21.58)	124 (21.57)	p = 0.518
60% - 80%	57 (19.52)	87 (15.13)	
>80% (most deprived)	40 (13.70)	88 (15.30)	

[47] Two PA and four non-PA patients were <12 months old. Age '0' means <12 months old at index date.

[48] Only for patients with HES data (291 PA, 575 non-PA).

[49] Only for patients with IMD data (291 PA, 575 non-PA).

Patient Characteristics	PA (n = 472)	Non-PA (n = 929)	Difference between Cohorts (p-value)
Number of comorbidities (based on Charlson's categories)[50]; n (%)			
0	348 (73.73)	784 (84.39)	$chi^2(3) = 23.492$
1	104 (22.03)	126 (13.56)	$p < 0.001$
2	17 (3.60)	16 (1.72)	
3	3 (0.64)	3 (0.32)	
Type of comorbidities (based on Charlson's categories); n (%)			
AIDS	-	1 (0.11)	
Cancer	9 (1.82)	19 (2.00)	
Cerebrovascular disease	14 (2.83)	6 (0.63)	
Chronic pulmonary disease	26 (5.25)	42 (4.42)	
Congestive heart disease	10 (2.02)	4 (0.42)	
Dementia	1 (0.20)	1 (0.11)	
Diabetes	47 (9.49)	51 (5.36)	
Diabetes with complications	7 (1.41)	7 (0.74)	$chi^2(14) = 54.713$
Hemiplegia	-	-	$p < 0.001$
Metastatic tumour	1 (0.20)	1 (0.11)	
Mild liver disease	-	2 (0.21)	
Moderate liver disease	-	-	
Myocardial infarction	6 (1.21)	2 (0.21)	
Peptic ulcer disease	-	-	
Peripheral vascular disease	3 (0.61)	8 (0.84)	
Renal disease	21 (4.24)	15 (1.58)	
Rheumatological disease	2 (0.40)	8 (0.84)	
None	348 (70.30)	784 (82.44)	

**Abbreviations:* Non-PA: without primary aldosteronism; PA: primary aldosteronisms; SD: standard deviation*

[50] The comorbidities that were present within two years before index date are demonstrated.

Table 85: Patient characteristics after index date (excluding five-year 'ghost' patients)

Patient Characteristics	PA (n = 472)	Non-PA (n = 929)	Difference between Cohorts (p-value)
Number of cases transferred out of the GP; n (%)	142 (30.08)	308 (33.15)	$chi^2(1) = 1.352$ p = 0.245
Number of deaths[51]; n (%)	70 (14.83)	132 (14.21)	$chi^2(1) = 0.098$ p = 0.754
Age at death (years); mean (SD, range)	76.54 (10.46, 44-99)	74.37 (11.80, 41-96)	t(200) = -1.294 p = 0.197
Underlying cause of death (ONS data only); n (%)			
Circulatory, endocrine or renal	20 (43.48)	27 (29.67)	$chi^2(1) = 2.585$
Other disease	26 (56.52)	64 (70.33)	p = 0.108
Other recorded cause of death (ONS data only); n (%)			
Circulatory, endocrine or renal	29 (63.04)	51 (56.04)	$chi^2(1) = 0.616$
Other disease	17 (36.96)	40 (43.96)	p = 0.432
Total time registered with GP (years); median (IQR, range)	20.94 (19.24, 1.95-85.34)	21.76 (18.80, 1.25-85.81)	z = -0.087 p = 0.931
Time registered with GP since index date (years); median (IQR, range)	5.31 (6.99, 0.03-25.90)	5.41 (6.52, 0.02-25.90)	z = -0.009 p = 0.993

**Abbreviations: GP: general practice; IQR: interquartile range; Non-PA: without primary aldosteronism; ONS: Office for National Statistics; PA: primary aldosteronism; SD: standard deviation*

[51] CPRD: 44 (9.32%) PA and 102 (10.98%) non-PA deaths
ONS: 46 (9.75%) PA and 91 (9.80%) non-PA deaths

Table 86: Average annual healthcare resource use (SD, range) for PA patients (excluding five-year 'ghost' patients)

Time / Resource Category	5-10 years before index date (n = 369)	1-5 years before index date (n = 446)	0-1 years before index date (n = 470)	0-1 years after index date (n = 467)	1-5 years after index date (n = 382)	5-10 years after index date (n = 219)
Primary care consultations	6.86 (6.50, 0-47.2)	9.61 (6.93, 0-59.75)	14.93 (10.23, 0-73)	14.42 (12.50, 0-123)	10.55 (8.16, 0-46.25)	10.62 (8.80, 0-50)
Diagnostic testing	2.14 (2.83, 0-15.21)	4.86 (5.07, 0-32.5)	9.30 (9.31, 0-72)	8.19 (9.55, 0-86)	6.98 (6.95, 0-42)	7.73 (6.59, 0-31.98)
Drug prescriptions	9.20 (10.90, 0-68.4)	12.58 (10.90, 0-64.25)	18.71 (14.54, 0-86)	18.71 (16.09, 0-104.44)	16.79 (15.35, 0-86.25)	18.54 (19.38, 0-139.22)
Secondary care referrals	0.47 (0.64, 0-4)	0.76 (1.02, 0-8.25)	1.30 (1.98, 0-27)	0.89 (1.71, 0-20)	0.71 (0.92, 0-5.25)	0.76 (0.99, 0-5.81)
Hospital admissions (HES patients only)	**(n = 227)** 0.13 (0.28, 0-1.6)	**(n = 275)** 0.28 (0.64, 0-8)	**(n = 291)** 0.89 (1.31, 0-8)	**(n = 287)** 1.11 (2.42, 0-22)	**(n = 234)** 0.68 (2.16, 0-30.75)	**(n = 138)** 1.69 (11.63, 0-133.25)

**Abbreviations: HES: Hospital Episode Statistics*

Table 87: Average annual healthcare resource use (SD, range) for non-PA patients (excluding five-year 'ghost' patients)

Time / Resource Category	5-10 years before index date (n = 722)	1-5 years before index date (n = 881)	0-1 years before index date (n = 928)	0-1 years after index date (n = 915)	1-5 years after index date (n = 759)	5-10 years after index date (n = 437)
Primary care consultations	4.51 (4.93, 0-43)	5.49 (5.62, 0-49)	5.87 (6.62, 0-44)	6.28 (7.63, 0-61)	6.50 (6.83, 0-50.5)	7.62 (7.83, 0-65.08)
Diagnostic testing	1.44 (2.64, 0-24.8)	2.37 (3.63, 0-25.25)	3.05 (5.31, 0-60)	3.36 (5.63, 0-43)	3.82 (5.08, 0-44.42)	5.09 (6.85, 0-73.45)
Drug prescriptions	5.12 (9.89, 0-126.8)	6.66 (10.03, 0-81)	7.58 (11.73, 0-93)	8.07 (12.32, 0-102.95)	8.89 (12.93, 0-126.47)	11.32 (13.42, 0-102.56)
Secondary care referrals	0.39 (0.70, 0-6.6)	0.41 (0.67, 0-6.75)	0.44 (0.99, 0-11)	0.47 (1.08, 0-12.03)	0.49 (0.85, 0-13.15)	0.51 (0.70, 0-3.92)
Hospital admissions (HES patients only)	**(n = 447)** 0.10 (0.26, 0-2.2)	**(n = 549)** 0.16 (0.48, 0-7.32)	**(n = 575)** 0.21 (0.68, 0-6)	**(n = 566)** 0.37 (1.30, 0-14.22)	**(n = 467)** 0.35 (1.01, 0-11.25)	**(n = 264)** 0.72 (3.43, 0-52.2)

**Abbreviations: HES: Hospital Episode Statistics*

Table 88: Average costs (SD; range) per healthcare resource category (excluding five-year 'ghost' patients)

Healthcare Resource Category	PA	Non-PA	Difference between Cohorts (p-value)
Primary care consultations	£3,970.22 (3,252.28; 0-25,278.61)	£2,216.44 (2,600.58; 0-41,740.34)	t(1399) = -10.937 p < 0.001
Diagnostic testing	£1,133.84 (1,238.88; 0-7,878.94)	£721.17 (1,004.09; 0-8,730.16)	t(1399) = -6.705 p < 0.001
Drug prescriptions	£15,694.36 (15,998.36; 0-112,841.60)	£8,503.19 (12,555.86; 0-103,273.70)	t(1399) = -9.212 p < 0.001
Secondary care referrals	£1,585.78 (1,643.06; 0-12,013.75)	£945.33 (1,178.86; 0-11,236.24)	t(1399) = -8.374 p < 0.001
Hospital admissions (HES patients only)	£21,941.17 (60,049.50; 0-708,984.50)	£14,473.57 (70,197.35; 0-1,439,457.00)	t(864) = -1.550 p = 0.122
Total costs (all patients) (without hospital admissions)	**£22,384.20 (19,729.22; 695.99-135,365.40)**	**£12,386.12 (15,485.45; 0-123,878.30)**	**t(1399) = -10.385 p < 0.001**
Total costs (HES patients only)	**£45,370.10 (66,562.88; 2,381.61-755,263.70)**	**£26,489.28 (74,471.50; 0-1,480,879.00)**	**t(864) = -3.650 p < 0.001**

**Abbreviations:* HES: Hospital Episode Statistics; Non-PA: without primary aldosteronism; PA: primary aldosteronism*

Table 89: Median costs (IQR; range) per healthcare resource category (excluding five-year 'ghost' patients)

Healthcare Resource Category	PA	Non-PA	Difference between Cohorts (p-value)
Primary care consultations	£3,014.12 (3,554.01; 0-25,278.61)	£1,472.72 (2,417.13; 0-41,740.34)	z = -13.022 p < 0.001
Diagnostic testing	£751.76 (1,140.80; 0-7,878.94)	£360.95 (844.91; 0-8,730.16)	z = -9.120 p < 0.001
Drug prescriptions	£10,667.23 (15,068.17; 0-112,841.60)	£3,723.93 (9,883.74; 0-103,273.70)	z = -12.999 p < 0.001
Secondary care referrals	£1,136.77 (1,543.30; 0-12,013.75)	£582.70 (1,149.80; 0-11,236.24)	z = -10.139 p < 0.001
Hospital admissions (HES patients only)	£11,015.26 (17,434.39; 0-708,984.50)	£2,804.56 (10,672.22; 0-1,439,457.00)	z = -9.484 p < 0.001
Total costs (all patients) (without hospital admissions)	**£16,072.91 (19,685.09; 695.99-135,365.40)**	**£6,724.04 (14,394.98; 0-123,878.30)**	**z = -13.332 p < 0.001**
Total costs (HES patients only)	**£28,364.02 (35,766.60; 2,381.61-755,263.70)**	**£10,879.59 (24,236.40; 0-1,480,879.00)**	**z = -11.326 p < 0.001**

**Abbreviations:* HES: Hospital Episode Statistics; Non-PA: without primary aldosteronism; PA: primary aldosteronism*

Table 90: Prediction of total healthcare costs (without hospital admissions) using GLM with square root link function and gamma distribution (excluding five-year 'ghost' patients)

Variable	dy/dx	Standard Error	z	p-value
PA[a]	8,865.41	837.89	10.58	<0.001
Age at index date	296.83	22.54	13.17	<0.001
Gender[b]	-4,200.75	949.97	-4.42	<0.001
Comorbidities[c]				
One	9,270.56	1,318.43	7.03	<0.001
Two	9,268.58	2,761.67	3.36	0.001
Three	10,827.63	6,070.00	1.78	0.074
Years of GP registration	216.58	37.95	5.71	<0.001
Deprivation[d]				
Less deprived	375.41	1,274.68	0.29	0.768
Average	593.69	1,146.82	0.52	0.605
Deprived	426.80	1,627.98	0.26	0.793
Most deprived	3,286.39	1,782.12	1.84	0.065
Number of observations = 1,401		Residual df = 1,389		
Scale parameter = 1.26				
Deviance = 1,348.86		(1/df) Deviance = 0.97		
Pearson = 1,743.07		(1/df) Pearson = 1.26		
AIC = 20.96		BIC = -8,714.37		
Log pseudolikelihood = -14,673.53				

Abbreviations: AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism

Reference categories:* ***a *– 'non-PA';* ***b*** *– 'female';* ***c*** *– 'No comorbidities';* ***d*** *– 'least deprived'*

Table 91: Prediction of total healthcare costs (with hospital admissions) using GLM with square root link function and gamma distribution (excluding five-year 'ghost' patients)

Variable	dy/dx	Standard Error	z	p-value
PA[a]	18,441.27	3,106.65	5.94	<0.001
Age at index date	842.58	110.43	7.63	<0.001
Gender[b]	-9,505.02	2,944.81	-3.23	0.001
Comorbidities[c]				
One	22,980.01	10,750.04	2.14	0.033
Two	16,516.39	7,314.40	2.26	0.024
Three	10,918.27	19,444.20	0.56	0.574
Years of GP registration	225.81	117.33	1.92	0.054
Deprivation[d]				
Less deprived	4,091.62	5,868.90	0.70	0.486
Average	-1,680.45	3,363.79	-0.50	0.617
Deprived	1,364.92	4,426.06	0.31	0.758
Most deprived	6,289.56	4,607.21	1.37	0.172

Number of observations = 866	Residual df = 854
Scale parameter = 3.94	
Deviance = 1,174.07	(1/df) Deviance = 1.38
Pearson = 3,365.51	(1/df) Pearson = 3.94
AIC = 22.42	BIC = -4,602.29
Log pseudolikelihood = -9,694.77	

**Abbreviations:* *AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism*

Reference categories:* ***a *– 'non-PA';* ***b*** *– 'female';* ***c*** *– 'No comorbidities';* ***d*** *– 'least deprived'*

Appendix 19: CPRD Study Sensitivity Analysis II Results

Table 92: Patient characteristics before index date (excluding outliers)

Patient Characteristics	PA (n = 441)	Non-PA (n = 855)	Difference between Cohorts (p-value)
Gender; n (%)			
Female	170 (38.55)	330 (38.60)	$chi^2(1) < 0.001$
Male	271 (61.45)	525 (61.40)	p = 0.987
Age at index date (years)[52]; mean (SD, range)	54.61 (13.82, 0-93)	54.44 (13.91, 0-93)	t(1294) = -0.209 p = 0.834
Ethnicity[53]; n (%)			
Black	22 (8.03)	5 (0.94)	chi^2 (3) = 58.301 p < 0.001
White	211 (77.01)	358 (67.29)	
Other	14 (5.11)	19 (3.57)	
Unknown	27 (9.85)	150 (28.20)	
Marital status; n (%)			
Co-habiting/Married	80 (18.14)	152 (17.78)	
Divorced/Separated/Widowed	11 (2.49)	8 (0.94)	$chi^2(4) = 5.490$
Single	17 (3.85)	31 (3.63)	p = 0.241
Data not entered	83 (18.82)	153 (17.89)	
Unknown	250 (56.69)	511 (59.77)	
Socioeconomic status based on residence area (quintiles)[54]; n (%)			
<20% (least deprived)	71 (25.82)	135 (25.38)	
20% - 40%	58 (21.09)	128 (24.06)	$chi^2(4) = 2.530$
40% - 60%	60 (21.82)	112 (21.05)	p = 0.639
60% - 80%	51 (18.55)	80 (15.04)	
>80% (most deprived)	35 (12.73)	77 (14.47)	

[52] Two PA and four non-PA patients were <12 months old. Age '0' means <12 months old at index date.

[53] Only for patients with HES data (274 PA, 532 non-PA).

[54] Only for patients with IMD data (274 PA, 532 non-PA).

Patient Characteristics	PA (n = 441)	Non-PA (n = 855)	Difference between Cohorts (p-value)
Number of comorbidities (based on Charlson's categories)[55]; n (%)			
0	336 (76.19)	730 (85.38)	$chi^2(3) = 19.010$ $p < 0.001$
1	88 (19.95)	111 (12.98)	
2	14 (3.17)	13 (1.52)	
3	3 (0.68)	1 (0.12)	
Types of comorbidities (based on Charlson's categories); n (%)			
AIDS	-	1 (0.11)	
Cancer	7 (1.52)	16 (1.84)	
Cerebrovascular disease	13 (2.82)	5 (0.57)	
Chronic pulmonary disease	19 (4.12)	40 (4.60)	
Congestive heart disease	10 (2.17)	4 (0.46)	
Dementia	1 (0.22)	1 (0.11)	
Diabetes	41 (8.89)	42 (4.83)	
Diabetes with complications	5 (1.08)	7 (0.80)	$chi^2(14) = 50.137$ $p < 0.001$
Hemiplegia	-	-	
Metastatic tumour	1 (0.22)	1 (0.11)	
Mild liver disease	-	1 (0.11)	
Moderate liver disease	-	-	
Myocardial infarction	6 (1.30)	2 (0.23)	
Peptic ulcer disease	-	-	
Peripheral vascular disease	3 (0.65)	5 (0.57)	
Renal disease	18 (3.90)	11 (1.26)	
Rheumatological disease	1 (0.22)	4 (0.46)	
None	336 (72.89)	730 (83.91)	

**Abbreviations:* Non-PA: without primary aldosteronism; PA: primary aldosteronism; SD: standard deviation*

[55] The comorbidities that were present within two years before index date are demonstrated.

Table 93: Patient characteristics after index date (excluding outliers)

Patient Characteristics	PA (n = 441)	Non-PA (n = 855)	Difference between Cohorts (p-value)
Number of cases transferred out of the GP; n (%)	126 (28.57)	278 (32.51)	$chi^2(1) = 2.108$ p = 0.146
Number of deaths[56]; n (%)	59 (13.38)	108 (12.63)	$chi^2(1) = 1.145$ p = 0.704
Age at death (years); mean (SD, range)	76.49 (11.12, 44-99)	74.39 (12.17, 41-96)	t(198) = -1.100 p = 0.273
Underlying cause of death (ONS data only); n (%)			
Circulatory, endocrine or renal	18 (45.00)	21 (28.00)	$chi^2(1) = 3.364$
Other disease	22 (55.00)	54 (72.00)	p = 0.067
Other recorded cause of death (ONS data only); n (%)			
Circulatory, endocrine or renal	25 (62.50)	38 (50.67)	$chi^2(1) = 1.475$
Other disease	15 (37.50)	37 (49.33)	p = 0.225
Total time registered with GP (years); median (IQR, range)	20.86 (18.55, 1.95-85.34)	21.87 (18.81, 1.25-85.81)	z = -0.174 p = 0.862
Time registered with GP since index date (years); median (IQR, range)	5.31 (7.00, 0.11-25.90)	5.31 (6.46, 0.02-25.90)	z = -0.548 p = 0.583

**Abbreviations:* *GP: general practice; IQR: interquartile range; Non-PA: without primary aldosteronism; ONS: Office for National Statistics; PA: primary aldosteronism; SD: standard deviation*

[56] CPRD: 36 (8.16%) PA and 82 (9.59%) non-PA deaths
ONS: 40 (9.07%) PA and 75 (8.77%) non-PA deaths

Table 94: Average annual healthcare resource use (SD, range) for PA patients (excluding outliers)

Time / Resource Category	5-10 years before index date (n = 346)	1-5 years before index date (n = 419)	0-1 years before index date (n = 439)	0-1 years after index date (n = 437)	1-5 years after index date (n = 360)	5-10 years after index date (n = 207)
Primary care consultations	6.14 (5.21, 0-34.2)	9.02 (5.96, 0-33.25)	14.01 (9.29, 0-60)	13.31 (11.44, 0-123)	9.67 (7.06, 0-38.5)	9.64 (7.50, 0-46.82)
Diagnostic testing	1.85 (2.27, 0-11.2)	4.44 (4.38, 0-25.5)	8.41 (7.33, 0-38)	7.23 (7.70, 0-64)	6.15 (5.74, 0-35.78)	7.04 (5.77, 0-28.54)
Drug prescriptions	7.97 (8.99, 0-54.6)	11.51 (9.63, 0-57.75)	17.09 (12.78, 0-86)	17.28 (14.42, 0-104.44)	15.18 (13.11, 0-84.25)	15.46 (12.38, 0-73)
Secondary care referrals	0.42 (0.55, 0-3.2)	0.70 (0.88, 0-5.5)	1.15 (1.47, 0-9)	0.75 (1.26, 0-9)	0.63 (0.79, 0-4.75)	0.68 (0.87, 0-5.81)
Hospital admissions (HES patients only)	**(n = 214)** 0.12 (0.27, 0-1.6)	**(n = 260)** 0.28 (0.65, 0-8)	**(n = 274)** 0.84 (1.23, 0-8)	**(n = 271)** 1.00 (2.19, 0-22)	**(n = 222)** 0.53 (0.87, 0-5.93)	**(n = 129)** 0.53 (2.20, 0-24.60)

**Abbreviations: HES: Hospital Episode Statistics*

Table 95: Average annual healthcare resource use (SD, range) for non-PA patients (excluding outliers)

Time / Resource Category	5-10 years before index date (n = 670)	1-5 years before index date (n = 815)	0-1 years before index date (n = 854)	0-1 years after index date (n = 841)	1-5 years after index date (n = 695)	5-10 years after index date (n = 396)
Primary care consultations	4.23 (4.59, 0-26.77)	5.15 (5.35, 0-36.25)	5.43 (6.32, 0-44)	5.82 (7.17, 0-55.30)	5.95 (6.37, 0-50.5)	6.94 (6.76, 0-49.13)
Diagnostic testing	1.27 (2.30, 0-16)	2.10 (3.17, 0-22.25)	2.57 (4.15, 0-36)	2.87 (4.78, 0-40)	3.37 (4.38, 0-37.03)	4.65 (6.41, 0-73.45)
Drug prescriptions	4.41 (7.20, 0-54)	6.12 (9.77, 0-81)	6.89 (11.15, 0-93)	7.32 (11.62, 0-102.95)	7.96 (11.48, 0-126.47)	10.12 (11.68, 0-59.68)
Secondary care referrals	0.34 (0.60, 0-6.52)	0.36 (0.57, 0-4.5)	0.38 (0.81, 0-5)	0.39 (0.84, 0-5.86)	0.43 (0.68, 0-7.20)	0.45 (0.63, 0-3.55)
Hospital admissions (HES patients only)	**(n = 415)** 0.10 (0.26, 0-2.2)	**(n = 508)** 0.14 (0.35, 0-3.45)	**(n = 532)** 0.18 (0.62, 0-6)	**(n = 523)** 0.33 (1.29, 0-14.22)	**(n = 426)** 0.30 (0.85, 0-11.08)	**(n = 238)** 0.46 (1.01, 0-8.79)

**Abbreviations: HES: Hospital Episode Statistics*

Table 96: Average costs (SD; range) per healthcare resource category (excluding outliers)

Healthcare Resource Category	PA	Non-PA	Difference between Cohorts (p-value)
Primary care consultations	£3,626.63 (2,718.20; 0-21,697.10)	£2,072.14 (2,530.17; 0-41,740.34)	t(1294) = -10.215 p < 0.001
Diagnostic testing	£1,008.59 (1,000.68; 0-6,459.98)	£641.55 (919.96; 0-8,730.16)	t(1294) = -6.603 p < 0.001
Drug prescriptions	£13,698.52 (12,401.29; 0-68,253.87)	£7,457.15 (10,828.61; 0-85,532.54)	t(1294) = -9.349 p < 0.001
Secondary care referrals	£1,411.37 (1,231.76; 0-7,245.87)	£822.29 (945.22; 0-5,752.02)	t(1294) = -9.556 p < 0.001
Hospital admissions (HES patients only)	£17,088.94 (42,144.07; 0-650,376.60)	£10,990.76 (36,898.43; 0-490,309.90)	t(804) = -2.116 p = 0.035
Total costs (all patients) (without hospital admissions)	**£19,745.11 (15,061.62; 695.99-81,940.14)**	**£10,993.13 (13,481.35; 0-89,438.97)**	**t(1294) = -10.634 p < 0.001**
Total costs (HES patients only)	**£38,218.69 (47,616.37; 2,381.61-670,005.10)**	**£21,788.67 (41,468.66; 0-519,507.80)**	**t(804) = -5.062 p < 0.001**

**Abbreviations:* *HES: Hospital Episode Statistics; Non-PA: without primary aldosteronism; PA: primary aldosteronism*

Table 97: Median costs (IQR; range) per healthcare resource category (excluding outliers)

Healthcare Resource Category	PA	Non-PA	Difference between Cohorts (p-value)
Primary care consultations	£2,868.59 (3,298.20; 0-21,697.10)	£1,333.89 (2,281.47; 0-41,740.34)	z = -12.893 p < 0.001
Diagnostic testing	£712.23 (992.65; 0-6,459.98)	£308.88 (752.29; 0-8,730.16)	z = -9.649 p < 0.001
Drug prescriptions	£10,207.03 (13,659.70; 0-68,253.87)	£3,264.76 (8,865.46; 0-85,532.54)	z = -13.173 p < 0.001
Secondary care referrals	£1,088.96 (1,369.43; 0-7,245.87)	£502.20 (1,064.18; 0-5,752.02)	z = -10.656 p < 0.001
Hospital admissions (HES patients only)	£9,366.04 (15,731.04; 0-650,376.60)	£2,338.95 (9,781.87; 0-490,309.90)	z = -9.305 p < 0.001
Total costs (all patients) (without hospital admissions)	**£15,637.40 (17,764.16; 695.99-81,940.14)**	**£5,903.29 (12,569.19; 0-89,438.97)**	**z = -13.528 p < 0.001**
Total costs (HES patients only)	**£26,880.15 (31,027.58; 2,381.61-670,005.10)**	**£10,115.88 (22,590.92; 0-519,507.80)**	**z = -11.321 p < 0.001**

**Abbreviations:* HES: Hospital Episode Statistics; Non-PA: without primary aldosteronism; PA: primary aldosteronism*

Table 98: Prediction of total healthcare costs (without hospital admissions) using GLM with square root link function and gamma distribution (excluding outliers)

Variable	dy/dx	Standard Error	z	p-value
PA[a]	8,581.05	744.04	11.53	<0.001
Age at index date	268.61	20.83	12.89	<0.001
Gender[b]	-4,299.24	791.67	-5.43	<0.001
Comorbidities[c]				
One	7,762.03	1,142.21	6.80	<0.001
Two	5,438.50	2,421.19	2.25	0.025
Three	5,488.38	7,019.70	0.78	0.434
Years of GP registration	205.87	33.78	6.09	<0.001
Deprivation[d]				
Less deprived	48.15	1,229.83	0.04	0.969
Average	16.24	1,054.45	0.02	0.988
Deprived	-528.65	1,470.08	-0.36	0.719
Most deprived	2,063.94	1,818.72	1.13	0.256

Number of observations = 1,296	Residual df = 1,284
Scale parameter = 1.11	
Deviance = 1,245.18	(1/df) Deviance = 0.97
Pearson = 1,425.58	(1/df) Pearson = 1.11
AIC = 20.73	BIC = -7957.30
Log pseudolikelihood = -13,423.77	

**Abbreviations:* *AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism*

Reference categories:* ***a *– 'non-PA';* ***b*** *– 'female';* ***c*** *– 'No comorbidities';* ***d*** *– 'least deprived'*

Table 99: Prediction of total healthcare costs (with hospital admissions) using GLM with square root link function and gamma distribution (excluding outliers)

Variable	dy/dx	Standard Error	z	p-value
PA[a]	17,165.50	2,561.73	6.70	<0.001
Age at index date	696.43	87.61	7.95	<0.001
Gender[b]	-9,376.42	2,674.21	-3.51	<0.001
Comorbidities[c]				
One	8,876.37	3,668.86	2.42	0.016
Two	14,760.29	7,734.70	1.91	0.056
Three	-10,850.10	5,587.99	-1.94	0.052
Years of GP registration	317.35	110.25	2.88	0.004
Deprivation[d]				
Less deprived	3,540.47	5,943.28	0.60	0.551
Average	-2,957.88	2,872.81	-1.03	0.303
Deprived	-1,160.38	3,741.68	-0.31	0.756
Most deprived	476.13	3,610.32	0.13	0.895
Number of observations = 806		Residual df = 794		
Scale parameter = 3.23				
Deviance = 1,016.17		(1/df) Deviance = 1.28		
Pearson = 2,564.92		(1/df) Pearson = 3.23		
AIC = 22.11		BIC = -4,297.35		
Log pseudolikelihood = -8,898.41				

**Abbreviations:* AIC: Akaike information criterion; BIC: Bayesian information criterion; df: degrees of freedom; GP: general practice; PA: primary aldosteronism*

***Reference categories:* **a** – 'non-PA'; **b** – 'female'; **c** – 'No comorbidities'; **d** – 'least deprived'*

Appendix 20: Diagnostic Tests for Primary Aldosteronism

Table 100: Description of screening, diagnostic and subtype classification tests for primary aldosteronism

Test/Procedure	Description
Screening Tests	
Plasma aldosterone-renin ratio (ARR)	*Purpose:* To detect whether there is a normal physiological relationship between aldosterone and the hormone that controls its secretion (renin) or whether this has been pathologically disrupted by autonomous aldosterone secretion. The test should be performed in all patients with suspected PA and should be repeated if initial results are inconclusive or difficult to interpret. *Procedure:* Blood is collected from a vein in the arm using a needle. Blood sample is collected mid-morning after the patient has been sitting, standing or walking for ≥2 hours and seated for 5-15 minutes; Sample is sent to laboratory for analysis; A positive test (i.e. high aldosterone levels and low renin levels) means that additional testing should be performed to confirm/exclude the presence of PA. ARR cut-off values depend on the assay that is used. *Precautions, reactions and adverse effects:* Medications that may affect the ARR (e.g. spironolactone, potassium-wasting diuretics) should be withdrawn ≥4 weeks before the test. If results are inconclusive, other medications that may affect the ARR (e.g. beta-blockers) should be withdrawn as well; Venepuncture can be a stressful procedure for patients. Stress leads to ACTH release which might affect results; Blood samples should be maintained at room temperature during delivery to laboratory.

Test/Procedure	Description
Confirmatory Tests	
Captopril challenge or suppression test (CCT)	*Purpose:* Captopril (ACE inhibitor; used for the treatment of hypertension) is administered to examine whether plasma aldosterone levels will be suppressed (no PA) or remain elevated (PA). The test investigates whether aldosterone levels are independent of control by renin. CCT should be performed after a positive ARR test, alone or in combination with other confirmatory tests, to definitively confirm/exclude the presence of PA. *Procedure:* Blood samples are initially collected to measure baseline plasma aldosterone, renin and cortisol levels; Patient sits/stands for ≥1 hour and then receives 25-50mg of oral captopril; Blood samples are collected at time zero, 1 and 2 hours from captopril administration to measure the new plasma aldosterone, renin and cortisol levels. Patient should remain seated during sampling; If plasma aldosterone levels are not suppressed by >30% and the plasma renin's concentration remains low, PA is confirmed.

Test/Procedure	Description
Fludrocortisone suppression test (FST)	*Purpose:* Fludrocortisone (corticosteroid; similar action to aldosterone in regulating salt and fluid balance in the body) is administered to investigate whether plasma aldosterone levels will be suppressed (no PA) or remain elevated (PA). The test examines whether aldosterone levels are independent of control by renin. FST should be performed after a positive ARR test, alone or in combination with other confirmatory tests, to definitively confirm/exclude the presence of PA. *Procedure:* Patient is normally admitted to hospital; On day 1, blood is collected in the morning to measure baseline plasma aldosterone, renin and cortisol levels; Patient receives 0.1mg oral fludrocortisone acetate and slow-release potassium chloride supplements (8mmol tablets) every 6 hours for 4 days. Potassium blood concentration is measured 2-4 times per day to ensure that it remains close to 4 mmol/L; Patient does also receive a high-sodium diet together with sodium chloride tablets (3x10mmol every 8 hours) for the same period to keep a urinary sodium excretion rate of ≥3 mmol/kg body weight; At 7 and 10 am on the last day, cortisol blood levels are measured, while at 10 am, plasma aldosterone and renin levels are also measured; Patient should remain in a seated/upright position for ≥30 minutes before taking initial and final measurements; On day 4, at 10 am, if plasma aldosterone is >6 ng/dL, plasma renin is <1 ng/mL/hour and cortisol's concentration is lower than that measured at 7 am, PA is confirmed. *Precautions, reactions and adverse effects:* Elderly patients and patients with uncontrolled hypertension, congestive heart failure or arrhythmias must not receive the test; Hypokalaemia is a common side effect during the test; Patients with renal insufficiency may show false negative results for PA.

Test/Procedure	Description
Oral sodium (salt) loading test (OSLT)	*Purpose:* To examine the response in urinary aldosterone and sodium levels after increasing the amount of salt in the diet (high aldosterone levels indicate that PA might be present). OSLT should be performed after a positive ARR test, alone or in combination with other confirmatory tests, to definitively confirm/exclude the presence of PA. *Procedure:* Patient receives a high-sodium diet and adequate oral sodium chloride supplements (total daily salt intake: >200mmol or >6g) for 3-4 days; Patient also receives adequate slow-release potassium chloride supplements to keep plasma potassium in normal levels; On the last day, a 24-hour urine collection is performed to measure aldosterone and sodium levels; PA is highly likely to be present if urinary aldosterone excretion is elevated (>12 μg or 33 nmol/d for 24 hours). *Precautions, reactions and adverse effects:* Patients with severe uncontrolled hypertension, renal insufficiency, congestive heart failure or cardiac arrhythmia, or severe hypokalaemia should not receive this test; Patients with hypokalaemia should have their potassium levels measured every day; 24-hour urine collection may be inconvenient for some patients.

Test/Procedure	Description
Saline infusion test (SIT)	*Purpose:* To measure the response in plasma aldosterone levels after administering salt and water intravenously (high aldosterone levels indicate that PA might be present). SIT should be performed after a positive ARR test, alone or in combination with other confirmatory tests, to definitively confirm/exclude the presence of PA. *Procedure:* Patient is normally admitted to hospital; Blood sample is collected to take the initial measurements of aldosterone, renin, urea, electrolytes and cortisol; 1.5-2L normal (0.9%) saline solution (sodium mixed with water) is infused into the bloodstream between 8-9:30 am to stimulate renin levels, and then (after 4 hours from the initial measurement), blood is collected from adrenal veins to measure again the levels of aldosterone, renin, urea, electrolytes and cortisol; The test lasts ≥6 hours and the patient should be recumbent 1 hour before and during the test; After the test, the patient can continue with their daily routine and take usual medication; PA is highly likely to be present if post-infusion plasma aldosterone levels are >5 ng/dL (>140 pmol/L) and renin levels remain low. *Precautions, reactions and adverse effects:* Beta blockers, diuretics, calcium channel blockers, ACE-inhibitors and angiotensin II blockers must be temporarily stopped 4 weeks before the test to get accurate results; During the test, blood pressure and pulse should be checked and recorded; The test should not be performed in elderly patients or people with low potassium levels, renal disease or kidney problems, severe uncontrolled hypertension, heart failure or arrhythmias, and/or severe hypokalaemia.

Test/Procedure	Description
Subtype Classification Tests	
Adrenal computerised tomography (CT)	*Purpose:* To diagnose pathologic conditions and anatomic abnormalities (e.g. adenomas, tumours, hyperplasia), and to examine their size and shape in the adrenal gland. The test is used as an initial subtype classification test to exclude the presence of adrenocortical carcinoma and assist clinicians during surgery. CT is performed after a positive ARR and one or more positive confirmatory tests for the disease. *Procedure:* Patient is placed on the CT scan table inside an encircling body scanner (*'gantry'*), that takes multiple cross-sectional pictures of the adrenals using X-rays of different angles. The final image is produced because of the variations in organ/tissue density and penetration; The procedure lasts >30 minutes or double this time if it is performed using a contrast dye (administered intravenously or orally) to enhance the image (in this case, CT scan is conducted with and without the dye). Images are interpreted by a radiologist; Abnormal findings in the adrenal glands (e.g. hypodense nodules of <2cm in diameter; nodules of >4cm in diameter) show the presence of adrenal tumours (i.e. adenoma or cancer) or hyperplasia, which may cause their overactivity. *Precautions, reactions and adverse effects:* Patient should remain still and stop breathing during the X-ray exposure since any motion can lead to artifacts on the images.

Test/Procedure	Description
(Bilateral) Adrenal vein sampling (AVS)	*Purpose:* To lateralise the source of the excessive aldosterone secretion (i.e. distinguish unilateral from bilateral PA) by measuring plasma aldosterone levels in both adrenal veins. AVS is performed after a positive ARR test, one or more positive confirmatory tests for the disease, and after finding abnormalities in the adrenal glands when CT is used. *Procedure:* Patient is normally admitted to hospital; A tube is inserted in the groin and placed in the right and left adrenal veins to collect one blood sample from each side; Plasma aldosterone levels are measured and compared between the two samples to detect any differences. A significant difference in aldosterone levels between the two blood samples suggests that unilateral PA is present on the side with the higher levels, while similar levels on both sides indicate overactivity in both adrenal glands (i.e. bilateral PA); There are several ways that adrenal venous sampling can be performed: 1) unstimulated sequential/simultaneous bilateral AVS; 2) unstimulated sequential/simultaneous bilateral AVS followed by bolus cosyntropin-stimulated sequential/simultaneous bilateral AVS; and 3) continuous cosyntropin infusion with sequential bilateral AVS. *Precautions, reactions and adverse effects:* AVS is a difficult procedure and should be conducted by an experienced angiographer to ensure that it will be successful; There is a risk of bleeding or developing a blood clot in the vein due to the insertion of the tube in the adrenal veins.

[**Sources:** (46, 330-332, 335, 336, 348, 351, 508)]

**Abbreviations: ACE: angiotensin-converting enzyme; ACTH: adrenocorticotropic hormone; ARR: aldosterone-renin ratio; AVS: adrenal venous sampling; CCT: captopril challenge/suppression test; CT: computerised tomography; FST: fludrocortisone suppression test; OSLT: oral sodium loading test; PA: primary aldosteronism; SIT: saline infusion test*

Appendix 21: Decision Tree Probabilistic Sensitivity Analysis Results

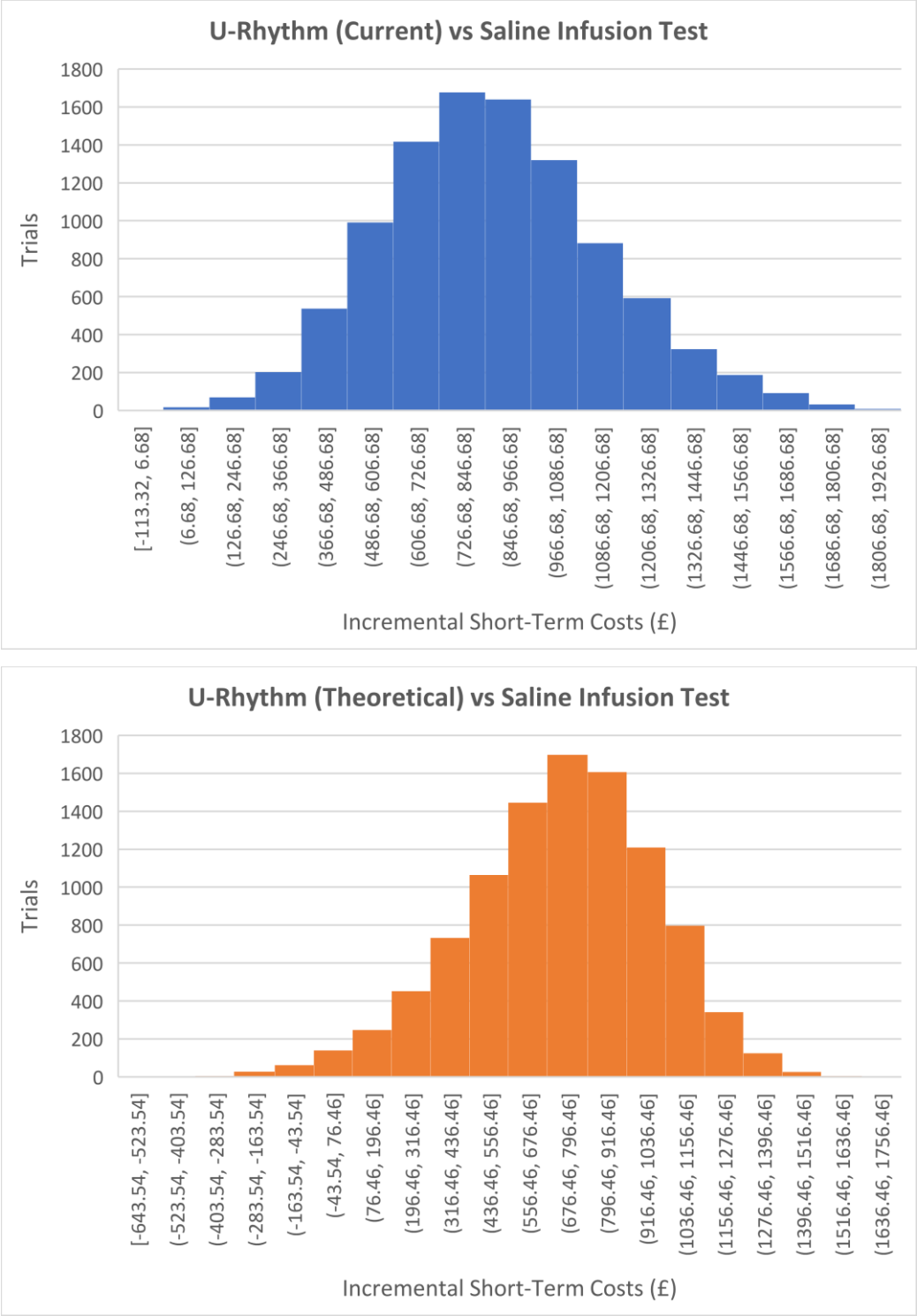
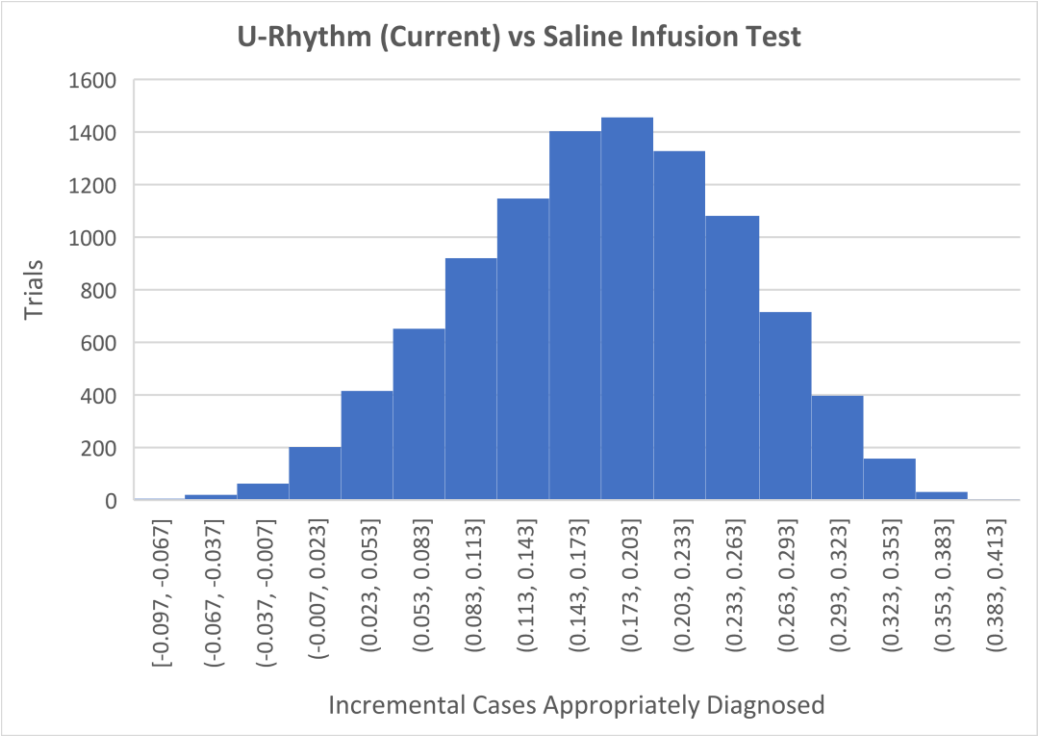


Figure 32: Distribution of incremental short-term costs



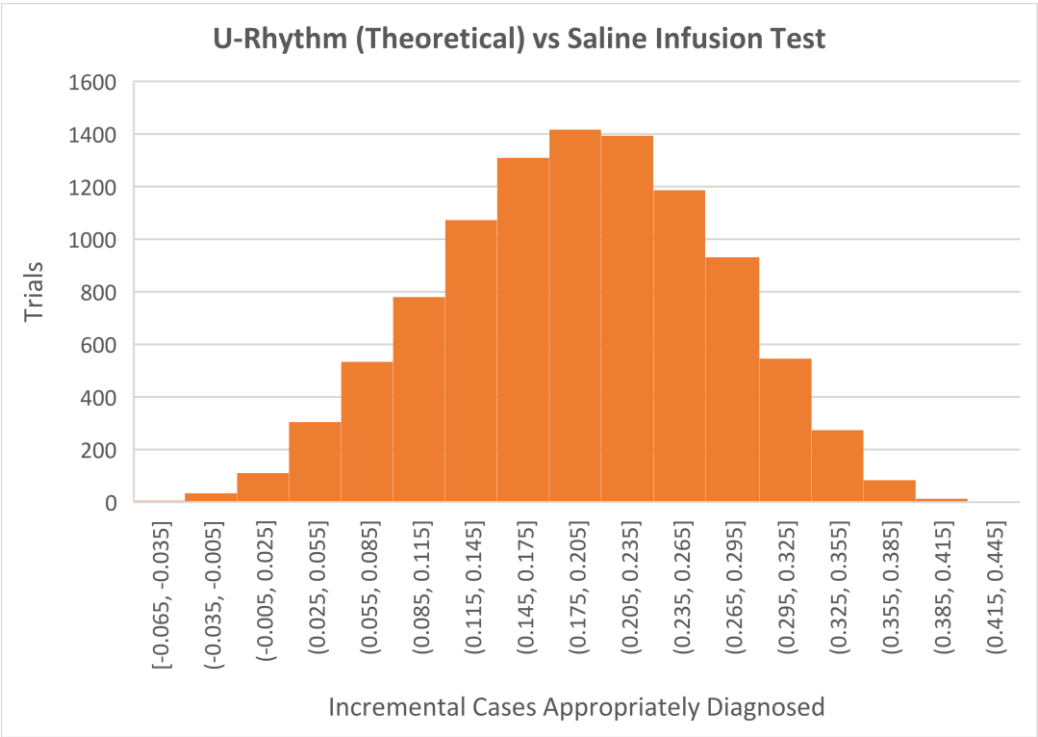


Figure 33: Distribution of incremental cases appropriately diagnosed

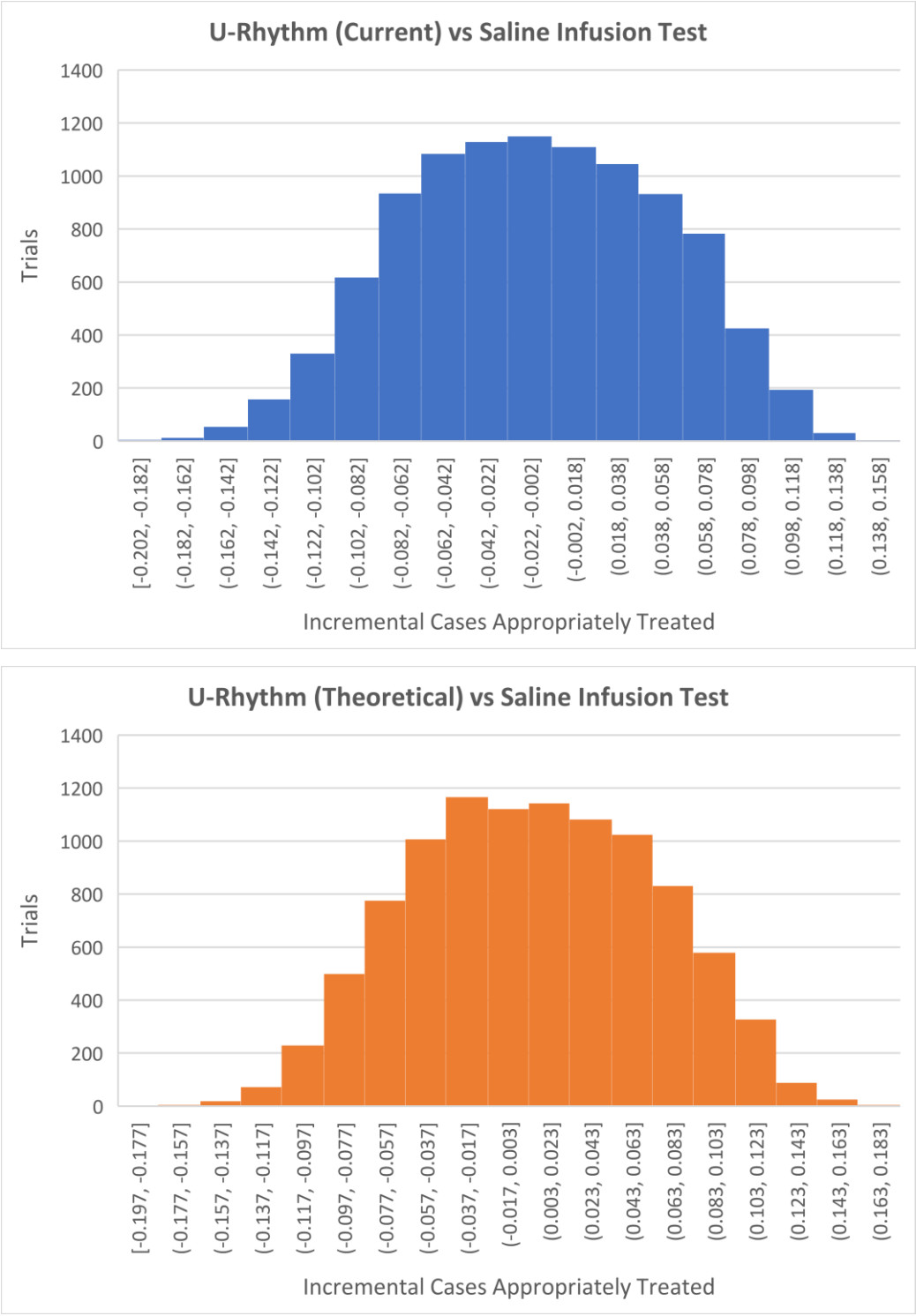
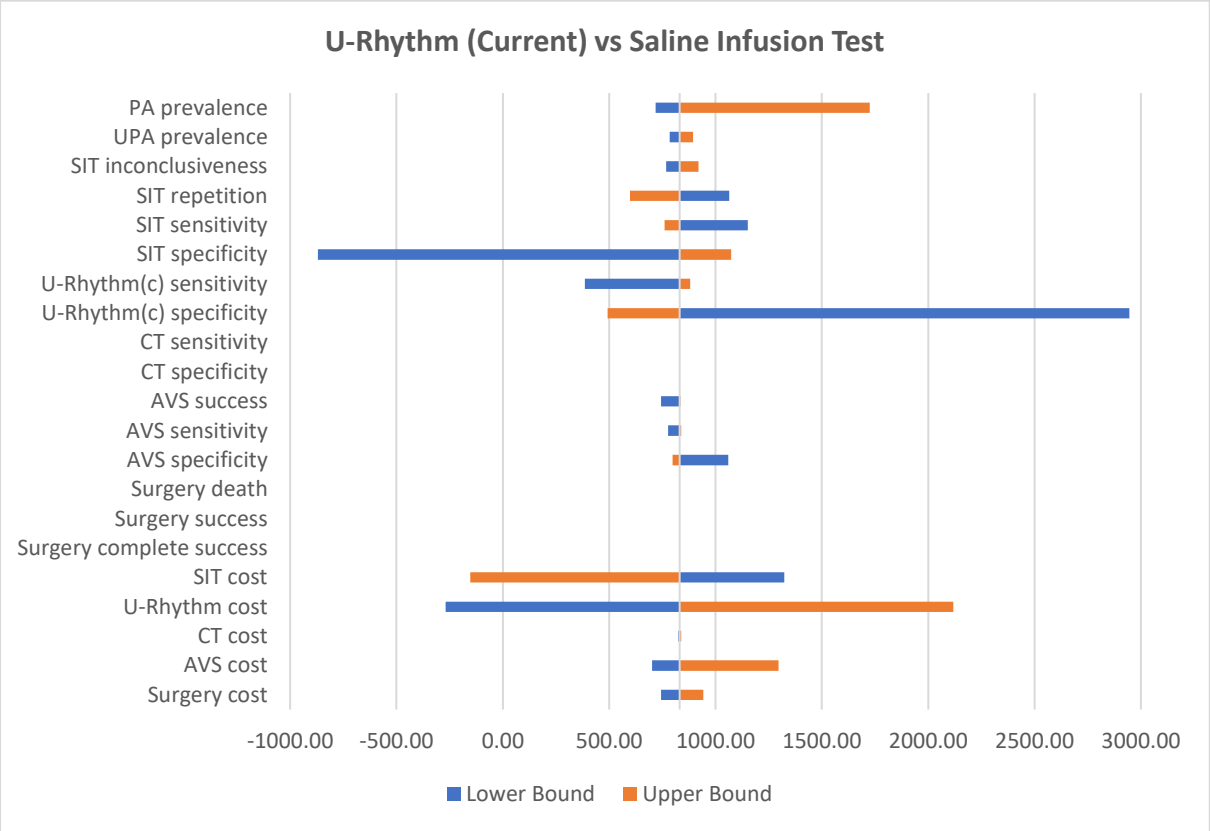


Figure 34: Distribution of incremental cases appropriately treated

Appendix 22: Decision-Analytic Model Internal Validation



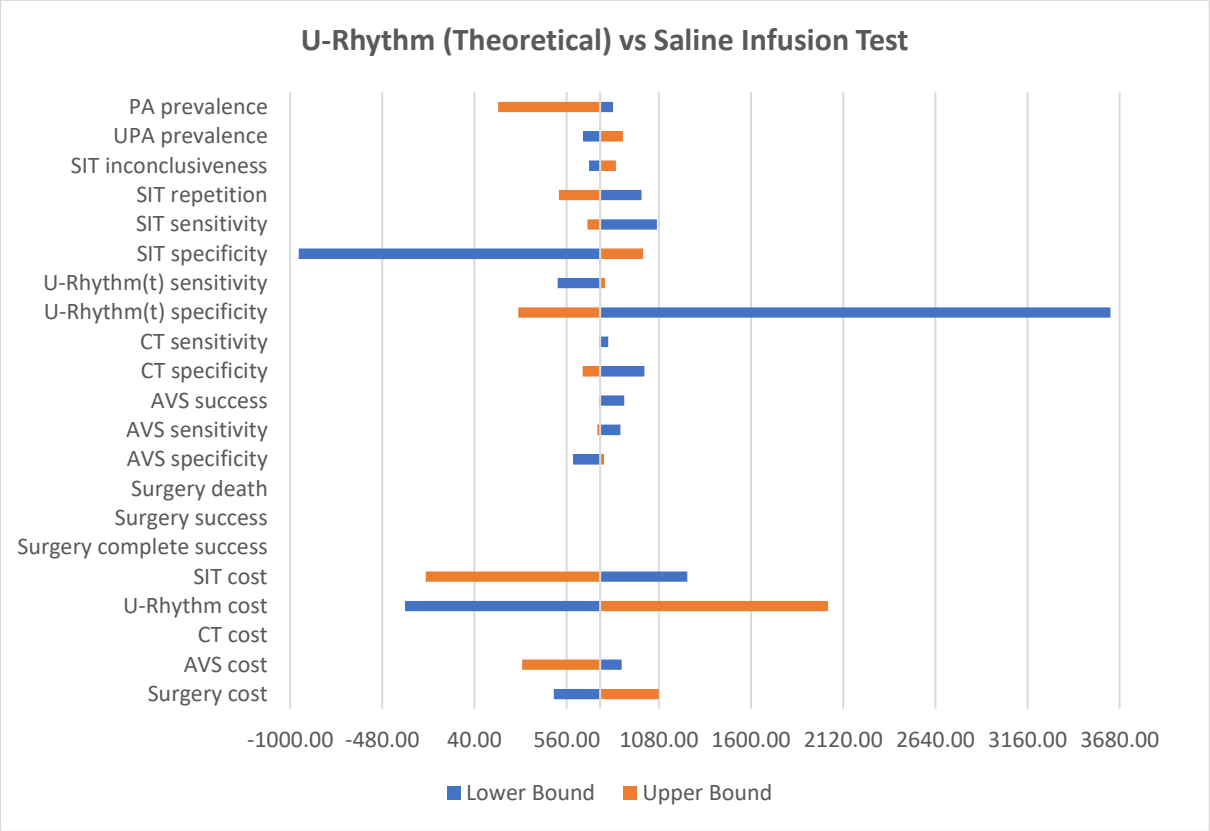
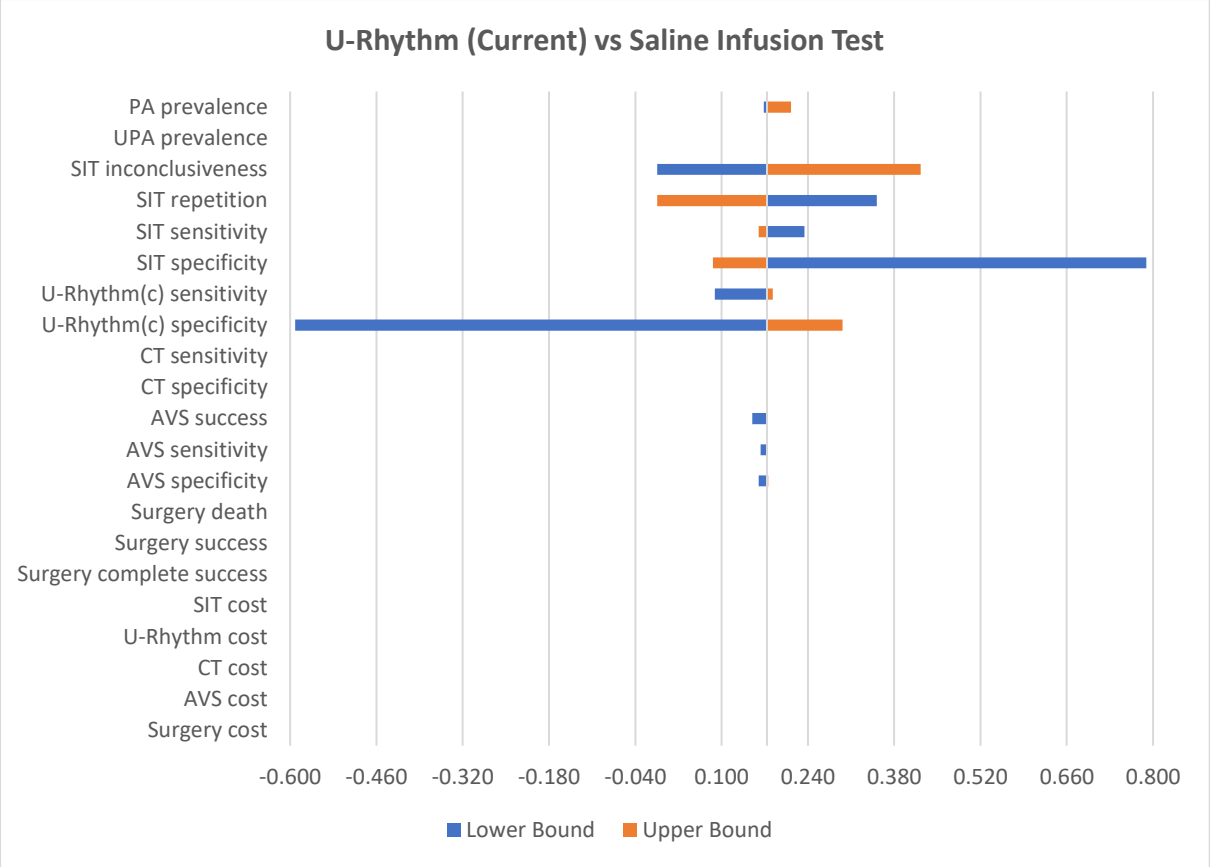


Figure 35: Impact of extreme values on incremental short-term costs



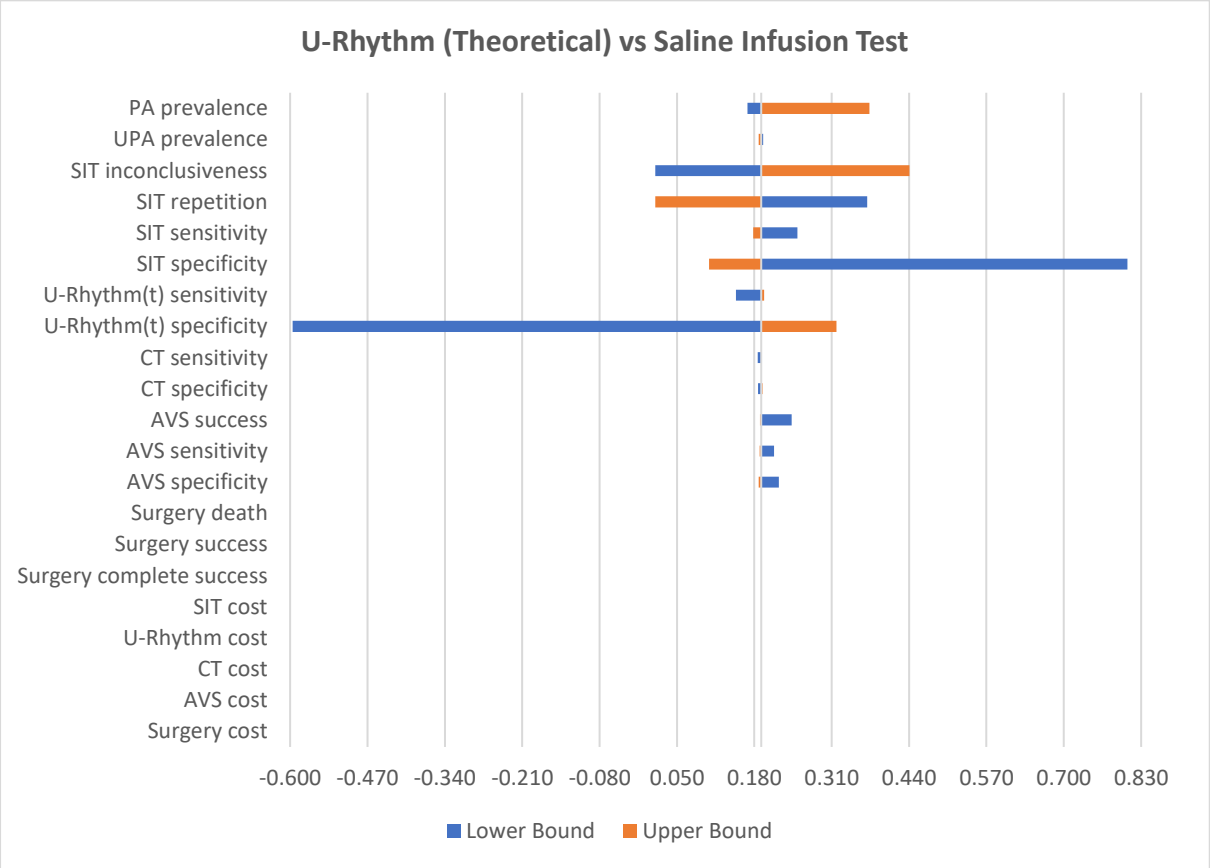
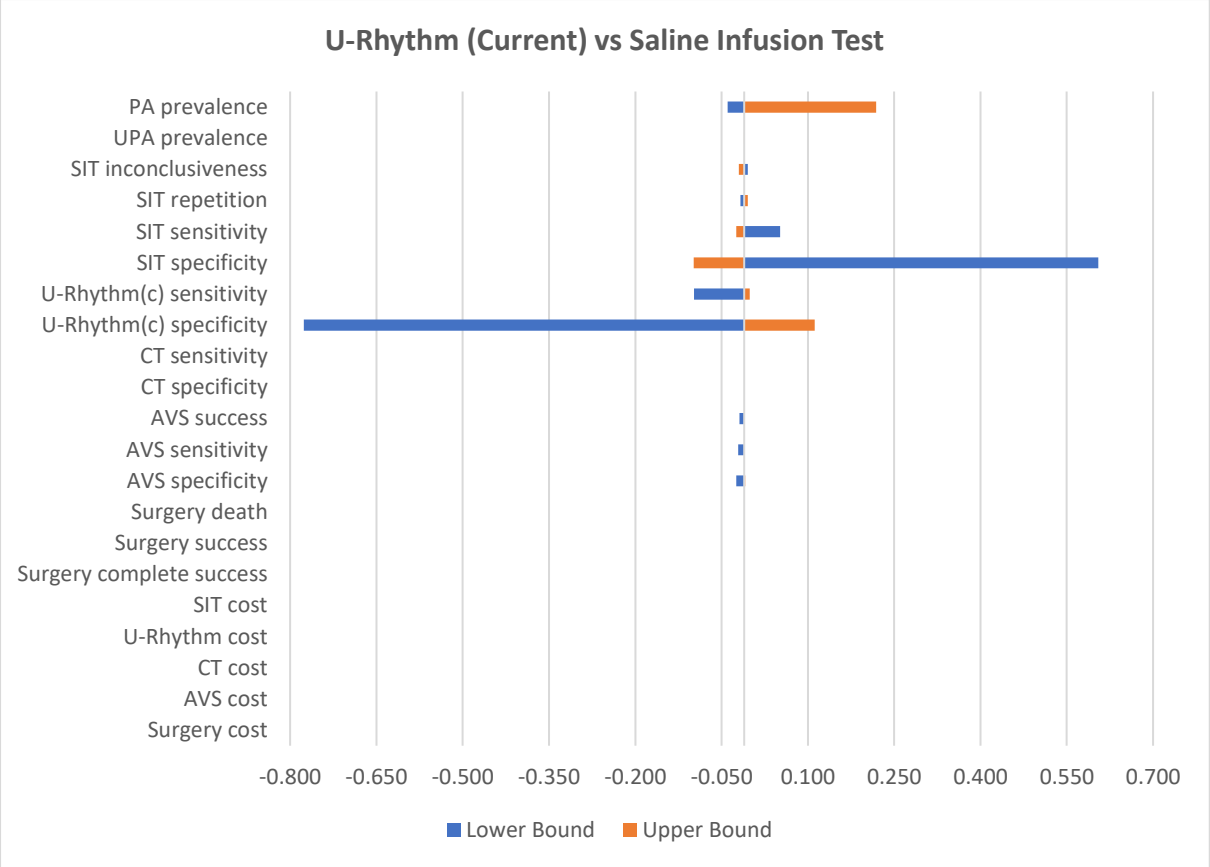


Figure 36: Impact of extreme values on incremental cases appropriately diagnosed



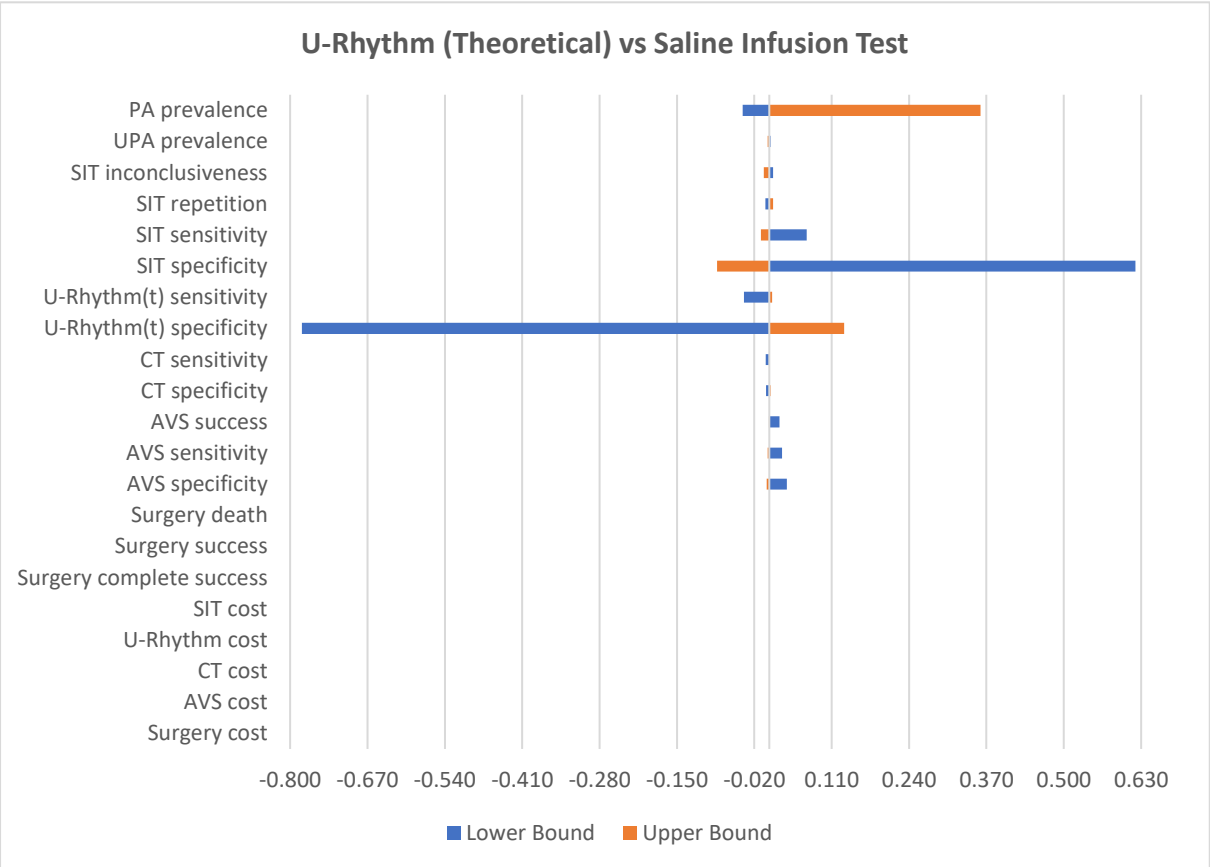


Figure 37: Impact of extreme values on incremental cases appropriately treated