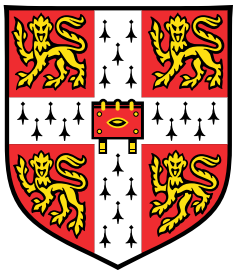
Triage-driven diagnosis for early detection of oesophageal cancer



Marcel Gehrung

Supervisor: Dr. Florian Markowetz

Advisor: Prof. Rebecca Fitzgerald

Cancer Research UK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Homerton College

September 2020

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Marcel Gehrung
September 2020

Abstract

Triage-driven diagnosis for early detection of oesophageal cancer

Marcel Gehrung

In this thesis I present my work to advance the early detection of oesophageal adenocarcinoma by investigating translational aspects of a minimally invasive oesophageal cell sampling technology for the detection of Barrett oesophagus.

Most oesophageal adenocarcinoma patients present with advanced disease, requiring treatment with chemotherapy with or without radiotherapy, followed by surgery to remove the oesophagus, and even then the overall five-year survival is less than 20%. However, if the cancer can be diagnosed at an early, superficial stage then treatment can be performed endoscopically and over 80% of patients survive beyond 5 years. The disease has a clear pre-cancer stage called Barrett oesophagus, making early detection feasible. A novel test called Cytosponge for diagnosing Barrett by cell collection coupled with an immunohistochemical test (Trefoil factor 3 / TFF3) has been developed.

I have investigated two distinct topics, which are key to implement the Cytosponge-TFF3 test in primary and secondary care. First, I analysed and interpreted data of a pragmatic, prospective, multicentre, randomised controlled trial (BEST3) in order to evaluate the use of Cytosponge in primary care. The study aim was to investigate whether offering this test to patients on medication for gastro-oesophageal reflux disease (GERD) would increase the detection of Barrett oesophagus compared with usual care. We were able to show that in patients with GERD the offer of Cytosponge-TFF3 testing results in improved detection (in excess of 10-fold) of Barrett oesophagus.

Second, I devised and implemented a machine learning framework applied to Cytosponge samples with the objective to reduce the pathologists' screening time. I trained and independently validated the framework on data from two clinical trials, analysing a combined total of 4,662 pathology slides from 2,331 patients. The approach exploits screening patterns of expert gastrointestinal pathologists and established decision pathways to define eight triage classes of varying priority for manual expert review. By substitution of manual review with automated review in low-priority classes, I can reduce pathologist workload by 57% while matching the diagnostic performance of expert pathologists.

Acknowledgements

I would like to thank the following people who have helped me undertake this work: My supervisor Florian, for his enthusiasm, encouragement, and patience; Rebecca, for her guidance, opportunities, and mentorship; Mireia, for her support, friendship, and optimism; Adam, for his dedication, humour, and fellowship. The Cancer Research UK Cambridge Institute at the University of Cambridge, for being an amazing community throughout this PhD. The Markowetz/Fitzgerald labs & others for their help, support and input along the way: Maria, Ruben, Michael, Paula, Jo, Evis, Andy, Neus, Adrienne. The Cyted team for all the drive and enthusiasm. Katrin – for being the way you are. And to my late grandfather, who set me off on the road to this PhD a long time ago - before either of us knew.

I would like to dedicate this thesis to my loving parents, Robert and Andrea,
my sister, Sarina, and my late grandparents, Karl and Anneliese.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AE Adverse Event

AI Artificial Intelligence

AUC Area Under The Curve

BE Barrett Oesophagus

BEST1 Barrett Oesophagus Screening Trial 1

BEST2 Barrett Oesophagus Screening Trial 2

BEST3 Barrett Oesophagus Screening Trial 3

BMI Body Mass Index

CAM Class Activation Mapping

CI Confidence Interval

CNN Convolutional Neural Network

COVID-19 Coronavirus Disease 2019

DNA Deoxyribonucleic Acid

GERD Gastro-oesophageal Reflux Disease

GI Gastrointestinal

GP General Practice

Nomenclature

Grad-CAM Gradient-weigthed Class Activation Mapping

H&E Hematoxylin and Eosin

HPV Human Papillomavirus

IM Intestinal Metaplasia

IQR Interquartile Range

ITT Intention-To-Treat

MHRA Medicines and Healthcare products Regulatory Agency

NPV Negative Predictive Value

OAC Oesophageal Adenocarcinoma

OGD Oesophago-Gastric Duodenoscopy

OSCC Oesophageal Squamous Cell Carcinoma

Pap Papanicolaou

PPI Proton Pump Inhibitor

CAM Positive Predictive Value

RFA Radiofrequency Ablation

RNA Ribonucleic Acid

ROC Receiver Operating Characteristic

ROI Region-Of-Interest

RR Rate Ratio

SAE Serious Adverse Event

TFF3 Trefoil Factor 3

TM Trademark

VAS Visual Analogue Scale

Chapter 1

Introduction

Attribution

Parts of sections 1.2 to 1.4 in this chapter were previously included in my first-year report.

Parts of section 1.4 in this chapter were adapted from:
A guide to deep learning on whole slide images
Authors: Berman A, Gehrung M, Markowetz F. *manuscript in preparation*

Author contributions

Figures in this chapter were prepared by me unless otherwise stated. The first-year report was written by me, with feedback from Florian Markowetz, Mireia Crispin-Ortuzar, and Rebecca Fitzgerald. It was initally written in December 2018 and was part of the first-year evaluation process.

A small number of paragraphs in section 1.4 were jointly prepared by myself and Adam Berman as part of the manuscript preparation for a review.

1.1 Early detection of cancer

Cancer is the second most common cause of death globally and is expected to become the leading cause of death in the coming decades [1]. As a disease, it has severe impacts on quality and quantity of life and is a major inhibiting factor for increasing life expectancy [2]. Although some of this increase may be attributed to increased case notification, exposure to key risk factors as well as aging and growth of the population, particularly in developing countries, are driving the growth of cancer-related incidence and mortality worldwide [3].

A main characteristic for cancer is abnormal proliferation by any type of cell present in the body [4]. As a consequence, there are manifold types and sub-types of cancers and it is possible to develop cancers in any organ of the human body. Tumours, a mass or lump of tissue, are characterised as either benign or malignant. The differences between the two types is of significant importance in cancer pathology due to their implications for health and clinical care: Benign tumours, such as warts, are non-invasive and cannot spread to distant sites from their site of origin [5]. Malignant tumours can invade their surrounding tissue and by means of metastasis spread throughout the body. A general rule is that the nomenclature of *cancer* usually refers to malignant tumours which can spread within the body and often complicate curative treatment. Cancers are classified according to the kind of cell from which they arise and the site of origin within the body. The three primary groups based on the kind of cells are carcinomas, leukemias/lymphomas, and sarcomas: Carcinomas arise from epithelial cells. Leukemias or lymphomas arise from immune or blood-forming cells. Sarcomas are solid tumours of connective tissues such as cartilage, muscle or fibrous tissue [4].

The development of cancer usually occurs as a multi-step process in which the initial progenitor cell does not suddenly acquire all features of a cancer cell [6]. Therefore, an initially acquired characteristic does not necessarily initiate a tumour but is potentially the beginning of a process comprising a series of alterations. Eventually, if a cell undergoes a number of alterations that promote abnormal proliferation and forms a malignant tumour which might cause symptoms, a cancer forms that has the potential to invade surrounding tissue and form metastases [7]. The main consequence for this concept of progress is that cancers mostly develop later in life.

The physiological mechanisms driving cancer development are complex and include a range of driving factors such as deregulated cellular metabolism [8], immune evasion [9], genome instability and mutation [10], tumour-promoting inflammation [11], and mechanical

tissue stress [12]. Similarly there are factors constraining cancer development which include immune destruction [4], damage repair mechanisms [13], metabolism-promoting homeostasis [14], and structural integrity of tissue [15]. These factors are summarised in fig. 1.1 where they are presented together with different stages of cancer development. Given the stage-wise progression, there is a potential in detecting cancer at an earlier stage where there is limited invasiveness as well as lack of metastasis.

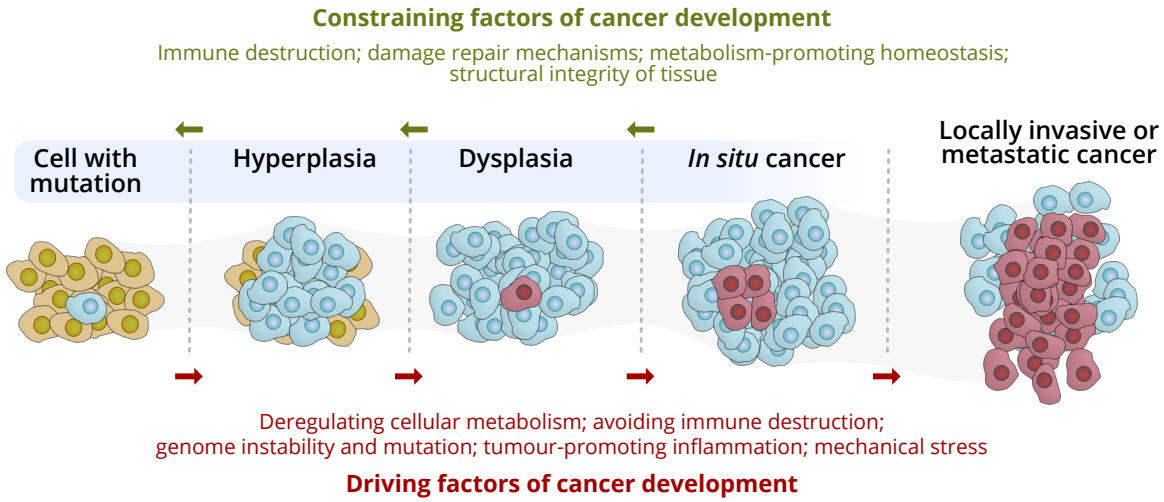


Fig. 1.1 **Factors constraining and driving cancer development.** Tumours usually start with a mutation in a single cell and can continue to develop through stages of hyperplasia, dysplasia, and *in situ* cancer to locally invasive or metastic cancer. If a lesion is detected early in the development (blue shaded area) there is possibility for treatment of the precursor in order to avoid progression.

Early detection of cancer is one of the three key areas in cancer management and is positioned between cancer prevention and cancer treatment. In recent years, early cancer detection has become a priority in policy as well as the general public [16] with a resulting desire to develop novel biomarkers to support diagnosis in a range of healthcare settings. With respect to the patient cohorts in these settings, the concept of early detection is applicable to both healthy and high-risk populations where it may lead to a decrease in morbidity and improved long-term survival [17]. Furthermore, the applicability of early detection biomarkers ideally extends to a stage prior to onset of symptoms, however, for certain cancer types this is not always achievable. In addition, treatment of precursor lesions or early-stage cancers can often be performed in a surgical and potentially minimally invasive way, without the need for radio- or chemotherapy [18] as often required in late-stage cancers. Implications of treatment at an early vs late stage are reduced harm and/or side effects for the patient

and associated health economics (i.e. treatment cost and duration with a resulting gain of quality-adjusted life years).

From an epidemiological perspective, *early detection* can be sub-divided into three different fields: Primary prevention aims to prevent disease (i.e. cancer) before it ever occurs [19]. Secondary prevention, also called screening, is the use of a test among a population with a higher risk of developing cancer in order to detect it sooner [19]. Tertiary prevention can be used to prevent complications in patients who already have been diagnosed with cancer with the aim of reducing morbidity and disability in these patients [20].

Primary and secondary prevention have become a topic of public debate and a variety of diagnostic approaches have been implemented in healthcare systems [21]: for cervical cancer, the Papanicolaou (Pap) test can detect abnormal cellular changes which might develop into cancer [22]. Human papillomavirus (HPV) testing further enables the detection of a viral infection that can cause these cellular changes [23]. The fecal immunochemical test can be used to detect occult blood in the stool which might be indicative for colorecal cancer [24]. Colonoscopy for image-based polyp detection is also recommended for certain risk groups [25]. Additional examples for image-based detection of early tissue changes are mammography for breast cancer [26] and low-dose computed tomography for lung cancer [27], particularly in populations with a history of heavy smoking. The tests listed above have been shown to reduce deaths from those cancers in numerous studies. Other cancer types with limited or developing evidence for early detection are pancreatic, ovarian, oesophageal, prostate, testicular, thyroid, bladder, skin, and oral cancer.

A new, emerging class of early detection tests are blood-based liquid biopsies [28]. These tests are still under development and have not been implemented at scale yet. In brief, blood is sampled from patients and the extracted cell-free, circulating DNA, RNA or proteins can be used to screen for mutations, changes in methylation and other molecular characteristics. This type of test is particularly interesting as it enables simultaneous screening for different cancers and tumour types, including tumours where the primary site is not easily accessible for sampling or imaging. Notable mentions for liquid biopsy tests under development and clinical evaluation are: Galleri (GRAIL) [29], CancerSEEK (Thrive) [30], and Guardant360 (Guardant Health) [31]. While some of these tests might enable the determination of the tissue-of-origin for certain cancers, they will most likely develop into pre-screening, complementary tools for any organ-specific targeted sampling test.

All diagnostic (or screening) methods for early cancer detection need to fulfil a number of conditions in order for the pathway change to be considered efficacious [32]:

- It needs to be a safe and acceptable test with evidence of its ability to detect early-stage disease.
- The cancer should have a recognisable latent (or early) asymptomatic stage.
- The natural history of the cancer and its potential pre-malignant condition needs to be well understood.
- If no early intervention is applied, most cases will progress from a preclinical to a clinical phase.
- Safe and effective treatment for early-stage disease must be available.
- The test should provide health economic benefit.

Aside from the considerations above, awareness for the three key biases which are encountered in screening is also required: First, one of the most common biases is lead-time bias in which a condition might be diagnosed earlier but there is no effect on the date of the patient's death [33]. Second, length-biased sampling which states that cancer screening tests are more effective at identifying slower-growing lesions than fast-growing ones. This causes that screening tests may select for cancers with a potentially favourable outcome [33]. Third, overdiagnosis occurs when indolent lesions are diagnosed that would never cause a health problem for the patient in the future. This third bias is in violation of one of the considerations mentioned above that screening tests will only be considered efficacious if the cancer they target progresses from a pre-clinical to a clinical phase [33].

The combination of these conditions and biases therefore demand a clear understanding of the cancer-of-interest in order to develop an appropriate screening test. One of the deadliest cancers with a need of earlier intervention which also has a targetable pre-malignant lesion is oesophageal cancer.

1.2 Oesophageal cancer

Oesophageal cancer is a cancer arising from the tissue within the oesophagus. On a global scale, it is the sixth most common cause for cancer related deaths with over 570,000 new cases and 510,000 resulting deaths in 2018 [1]. Depending on the location and staging of the cancer, patients presenting with this disease often show symptoms such as dysphagia and weight loss. In the majority of patients, oesophageal cancer is diagnosed at a late stage

and a 5-year overall survival rate of 13% is observed [34]. Oesophageal cancer can be divided into two distinct pathological sub-types: oesophageal adenocarcinoma (OAC) and oesophageal squamous cell carcinoma (OSCC). Both of these sub-types have a divergent profile with respect to epidemiology and risk factors, with OAC being the predomininant type in the western (Europe, US), and OSCC the predominant type in the eastern world (Africa, Asia) [35]. Risk factors for OSCC include tobacco smoking and chewing [36, 37], alcohol consumption [38, 37], low fruit/vegetable intake [39], recurrent thermal injury [40], and HPV infection [41]. Risk factors for OAC include gasto-oesophageal reflux disease (GERD), central visceral obesity, tobacco smoking, male sex, red meat intake, and low fruit/vegetable intake [42]. Epidemiologically, the global prevalence is skewed towards OSCC, which has an eight times higher global incidence rate when compared with OAC.

However, while OSCC has shown a slight decrease in incidence over the last decades, OAC incidence has persistently increased over the past four decades in the western world [43], particularly for white males [44]. Both main sub-types can arise from dysplastic precursor lesions which can be detected using endoscopy.

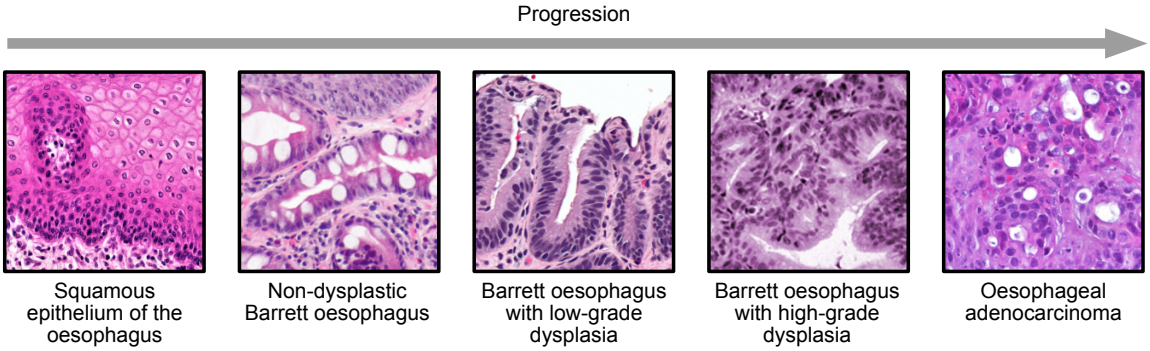


Fig. 1.2 **Progression through several (non-)dysplastic precursor lesions until the eventual manifestation of an adenocarcinoma.** All images are licensed under CC BY-SA 4.0.

If these precursor lesions undergo local endoscopic treatment, patient prognosis dramatically improves resulting in better long-term outcomes. For OAC in particular, this precursor lesion is called Barrett oesophagus (BE) as shown in fig. 1.2. Histologically, BE is a change containing intestinal metaplasia (IM) in which the stratified squamous epithelial lining localised in the lower oesophagus is replaced with columnar epithelium containing goblet cells. In the context of cancer development, BE progresses to OAC through a number of different mechanisms: The exposure of the oesophageal mucosa to stomach acid or bile (GERD), due to reactive oxygen species and nitric oxide, results in DNA damage and a

mutational profile of A-to-C transversions [45–47]. These profiles are commonly observed in BE and OAC and, together with the persistent oesophageal exposure to damaging agents, hypothesised to be a primary driver in pathogenesis of BE and dysplastic lesions [48, 49]. Furthermore, as a result of the mucosal injury, BE often contains genetic, somatic mutations that predispose the lesion to cancer development [50]. Genomic characterisations of BE have shed light on several mechanisms that play a key role in progression to OAC [51]: First, step-wise loss of tumour supressor genes (*CDKN2A* and *TP53*) paired with mutations in *SMAD4* and the disruption of chromatin-modifying events but in the observed absence of whole-genome duplications. Second, significant chromosomal instability in association with aneuploidy. This is the primary consequence of the loss of p53 regulation [52], caused by loss of heterozygosity of chromosome arm 17p [53, 54]. Third, chromothrypsis and kataegis may cause additional chromosal instability [55]. Fourth, greater clonal diversity has been shown to be associated with increased risk of progression. Lastly, epigenetic alterations and specifically hypermethylation of CDKN2A is often observed and further contributes to risk of progression [56].

If a BE lesion has progressed and shows signs of dysplasia, patients usually undergo endoscopic treatment involving musocal resection and epithelial ablation (e.g. radiofrequency ablation / RFA) according to respective guidelines [57, 58]. To detect BE in the first instance, diagnosis usually occurs during an invasive, endoscopic procedure of the upper gastrointestinal tract (oesophago-gastro-duodenoscopy / OGD). However, OGD is only performed on few patients as there is no endoscopic routine screening programme for patients with heartburn symptoms (10 to 15% of the adult population [59]). Given the economic burden and patient discomfort of such a procedure, this indicates a clear need for minimally invasive approaches for the diagnosis of BE and therefore secondary prevention of OAC.

BE can be diagnosed using various approaches which are either in practice, research or under evaluation [60]: endoscopy (including capsule [61], transnasal [62], narrow band imaging, and chromo- and confocal laser endoscopy [63]), electronic nose [64], oral microbiome [65], targeted minimally invasive sampling (e.g. balloon-based [66], sponge-on-string [67]) and liquid biopsy [68]. Endoscopic procedures tend to be more expensive and associated with a risk of complications which reduces the clinical utility for first-time diagnosis, particularly in patient populations with mild to moderate symptoms. Approaches such as the electronic nose or oral microbiome [64, 65] have demonstrated promising data, however, the technologies are early stage and factors like accuracy, feasibility and health economics

are yet to be determined. Liquid biopsy based on circulating microRNAs [68] has shown encouraging results but current markers have insufficient sensitivity and specificity to be considered for clinical use. Balloon-based sampling paired with methylation biomarkers [66] also yielded promising results and was considered as well tolerated by patients. One additional promising technology for the targeted minimally invasive detection of BE which has emerged over the last decade is the Cytosponge technology [69].

1.3 Cytosponge™

The Cytosponge™ is a non-endoscopic diagnostic modality for BE. It is a cell collection device, consisting of a mesh sphere on a string inside a gelatine capsule (fig. 1.3), coupled with an immunohistochemical biomarker called Trefoil Factor 3 (TFF3) to screen for IM. TFF3 is overexpressed in mucin-producing goblet cells and is thought to function as a protector of the mucosa from insults, stabilizer of the mucus layer and promoter for healing of the epithelium [70].

Fig. 1.3 **Cytosponge™ mesh sphere in a gelatine capsule (left) and expanded (right).** Source: University of Cambridge.

The capsule is swallowed by the patient, and the gelatine dissolves in the stomach allowing the mesh sphere to expand to a diameter of 3 cm. After 5 to $7\frac{1}{2}$ minutes, the sponge is withdrawn from the stomach by the attached string, sampling superficial epithelial cells from the top of the stomach, the oesophagus, and the oropharynx. The removed device is placed in a container with preservative solution (such as BD SurePath [71] or CytoRich Red [72]) and processed in a laboratory for (immuno)histochemical staining with TFF3 and Hematoxylin & Eosin (H&E) (fig. 1.4). The stained pathology slides are screened by a pathologist for IM and other potential diagnoses such as eosinophilic oesophagitis,

candida infections, squamous atypia, herpes, and ulcers. A diagnosis of IM is indicated if a columnar-shaped cell which secretes a component of mucus is present in the sample. Usually, these goblet cells are only present in other epithelia (intestines, respiratory tract). However, the presence of columnar epithelium of intestinal type (with goblet cells) in the squamous oesophagus is abnormal and a strong indication for BE or possibly indicative of IM of the gastric cardia, which is also considered a pre-cancerous change. Patients with IM-indicating findings can then be referred for an upper gastrointestinal endoscopy to confirm the diagnosis and receive potential treatment.

A number of previous studies [74, 75] have shown a consistent sensitivity (73.3 % and 79.9 %) and specificity (93.8 % and 92.4 %) for the diagnosis of BE using the Cytosponge coupled with the TFF3 biomarker. A systematic review analysing a number of different Cytosponge studies reported a pooled sensitivity and specificity of 81% and 91%, respectively, for the diagnosis of BE [76]. In the major case-control study for the technology, sensitivity improved with longer BE segments [75]. In all of these studies, the gold standard was OGD with biopsies with an asssumed sensitivity and specificity of 100%.

Another publication performed a patient-level review of five studies with a focus on safety and acceptability of the Cytosponge test [69]. While three studies were focused on patients with GERD and the detection of BE as a primary endpoint, the BEST-RFA study (unpublished) had the objective of detecting BE after RFA treatment and one study focused on the detection of eosinophilic oesophagitis [77]. An overview of the included studies is presented in table 1.1. The review did not intend to assess efficacy of the technology, but rather safety and acceptability of the test as all studies included prospective measurements of these variables. A visual analogue scale (VAS) from 0 to 10 was used to assess acceptability (higher score = more acceptable). Safety was captured by the number of attempts or failures of swallowing and number as well as type of adverse events. Key findings for acceptability from the review [69] included: 134 (5.5%) out of 2,418 patients were either unable to swallow the device or were withdrawn from the study by the clinician. The remaining 2,284 patients completed the VAS, however, only 1,221 had a follow-up endoscopy with an impact on comparative scoring. Cytosponge was significantly more acceptable ($p<0.001$ for each comparison) to patients undergoing endoscopy without sedation (Cytosponge: median VAS of 6 (IQR 5 to 8), endoscopy with sedation: 8 (IQR 5 to 9), endoscopy without sedation: 5 (IQR 3 to 7)). Men provided higher median VAS scores than women for Cytosponge administration (men: 7 (IQR 5 to 8), women: 6 (IQR 5 to 8), $p=0.003$). Primary care patients gave higher median VAS scores compared with secondary care patients (primary care: 7

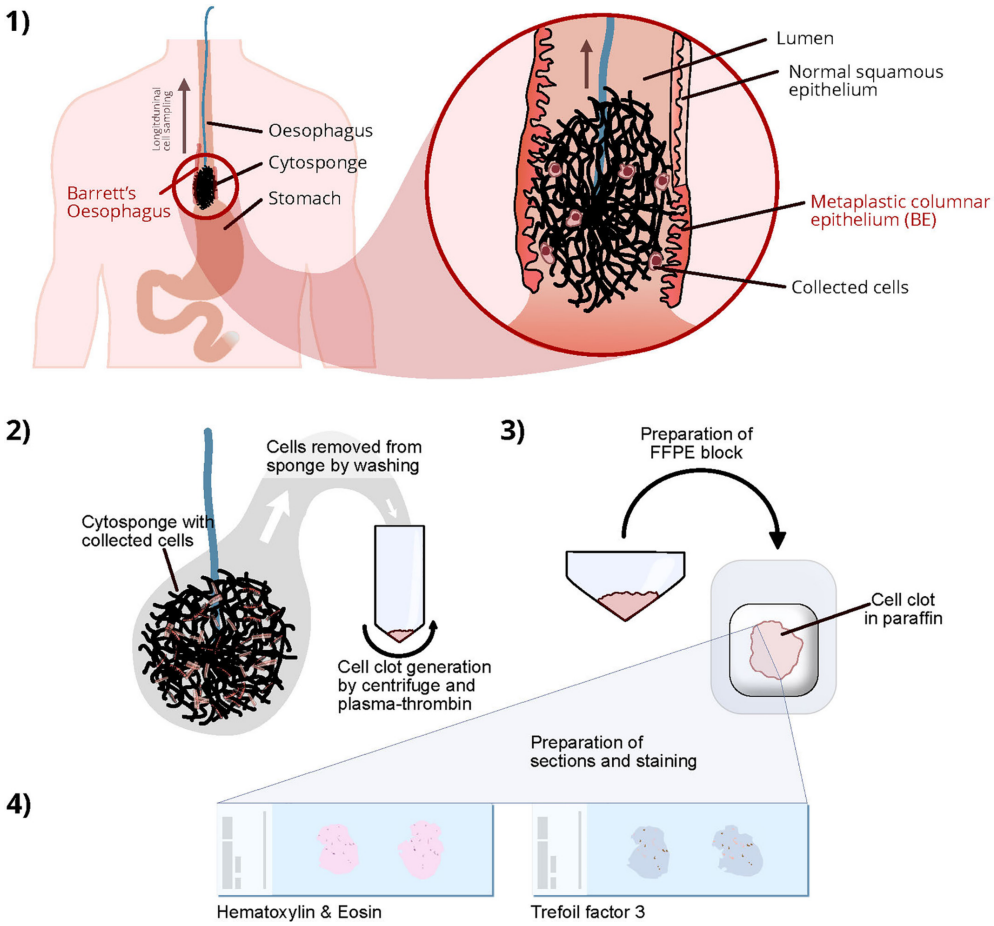


Fig. 1.4 **Schematic outlining the steps of the Cytosponge-TFF3 assay. 1)** The sponge is swallowed, the coating dissolves in the stomach and the sponge deploys. The Cytosponge is then withdrawn using the attached string collecting cells from the gastroesophageal junction and the length of the oesophagus. **2)** The sampled cells are retrieved from the sponge by a series of washing steps, then a cell pellet is made by centrifugation. Plasma-thrombin is added to the cell pellet, making a cell clot. **3)** The cell clot is fixed in formalin and processed into a paraffin block using standard laboratory protocols. **4)** Sections are cut from the paraffin block and stained for assessment by a pathologist. Figure created by me with adapted caption from source: [73].

(IQR 5 to 8), secondary care: 6 (IQR 5 to 8), p<0.001). Patients in secondary care were more likely to fail to swallow the Cytosponge (odds ratio: 5.13, 95% CI 1.48 to 17.79, p<0.01) and patients with BE (5.7%) were more than twice as common to show failure of swallow vs GERD patients (2.1%). 12 serious adverse events (AEs) were recorded in the studies while only two of them were due to the Cytosponge device. The two related events were one detachment of the sponge, which was recovered through endoscopy and one minor pharyngeal, that resolved spontaneously. Other AEs were related to the OGD procedure after Cytosponge administration.

In summary, the Cytosponge device technology is considered as safe with a low risk of adverse events and favourable acceptability ratings when compared to endoscopy without sedation.

In addition to a quantitative assessment of acceptability of the Cytosponge test, a qualitative thematic analysis was prepared in a sample of UK residents living with GERD [78]. 33 participants (17 men and 16 women, median age of 57, range 50 to 69) either had a one-to-one interview (n=10) or participated in one of four focus groups (n=23). 45% had an endoscopy before and none of the participants had a Cytosponge test in the past. There were several concerns highlighted by the participants, based on the anticipated experience: Swallowing of the string, possibility of the Cytosponge getting stuck, vomitting and gagging during the swallow procedure. Participants with prior experience of an endoscopy suggested that Cytosponge will be preferable for practical and economical reasons. It was also noted that the test could be performed at their local primary care practice and did not require sedation.

Additionally, as mentioned in section 1.1, in order to assess whether the technology is beneficial for the patient and payer in the respective healthcare system, health economic evidence needs to be assessed.

Two studies have previously been published that utilise microsimulation models to assess the cost effectiveness of Cytosponge when compared to endoscopy [79, 80]. The first study by Benaglia et al. [79], which was designed for the UK/NHS healthcare market was based on a cohort of men at the age of 50 with a history of GERD. Modelling for every individual was performed for 50 years or until death. The approach assessed three different strategies: Cytosponge test followed by endoscopy in TFF3-positive patients, endoscopy screening, and no screening.

	BEST1	BEST2	BEST (Australia)	BEST2-RFA	EoE study
Country:	UK	UK	Australia	USA	USA
Disease:	GERD	GERD and BE	GERD	BE after RFA treatment	EoE
No. of patients (%):	518 (21.4%)	1,498 (62.0%)	224 (9.3%)	76 (3.1%)	102 (4.2%)
No. of Cytosponge® procedures (%)	518 (19.4%)	1,752 (65.6%)	224 (8.4%)	76 (2.8%)	102 (3.8%)
Time of recruitment:	May 2008 to December 2009	July 2011 to December 2013	May 2010 to August 2014	October 2014 to 2020	December 2012 to 2017
Setting:	Primary care (12 general practices)	Secondary care (11 hospitals)	Secondary care (1 hospital)	Secondary care (1 hospital)	Secondary care (2 hospitals)

Table 1.1 **Previous Cytosponge studies including key characteristics.** Conducted studies with target cohort, number of patients, number of Cytosponge procedures, study start/end date and healthcare setting.

Key model parameters were the prevalence of BE (8%), presence of non-dysplastic disease in BE patients (86%), low-grade dysplasia in BE patients (10%), high-grade dysplasia in BE patients (2%), and intramucosal cancer in BE patients (2%). Other parameters such as costs, utilities and transition rates were derived from NICE guidelines and can be found in the full manuscript [79]. The microsimulation results demonstrated that a Cytosponge test followed by endoscopy in TFF3-positive patients or endoscopy screening alone are cost-effective when compared to no screening at all. The Cytosponge screening would cost less than endoscopy screening.

The second study by Heberle et al. [80] was based on US data and was calibrated for United States Surveillance, Epidemiology and End Results data. The model was based on a birth cohort (1950) of US males starting from age 20 and followed the cohort until the age of 100 or death. One difference when compared to the study above was that at the age of 60 the population was restricted to patients displaying GERD symptoms which have not yet been diagnosed with OAC. The approach assessed the same three screening strategies as Benaglia et al. [79]. Key model parameters were derived from the EACMo from the Massachusets General Hospital or the Microsimulation Screening Analysis model from Erasmus University Medical Center and University of Washington. Cost parameters were based on Medicare reimbursement catalogues and manufacturer discussions. Similar to the previous results [79], this study found that no screening resulted in the poorest outcomes. Endoscopy screening offered the largest health economic benefit with the highest costs. Cytosponge-TFF3 screening with endoscopy in TFF3-positive patients fell between the two other strategies.

Both studies concluded that the use of Cytosponge-TFF3 for screening of patients with GERD would most likely provide more health economic benefit (i.e. more QALYs) as well as a higher rate of cancer detection when compared to no screening, and a lower cost when compared to endoscopic screening.

This summary on safety, efficacy, acceptability, and health economics provides a robust foundation for the implementation of Cytosponge as a targeted screening tool in primary and secondary care. Over the recent years, other novel technologies for pan-oesophageal cell sampling have also emerged, with varying levels of evidence: EsoCheck paired with EsoGuard (Lucid Diagnostics) is a swalloable balloon-based device with subsequent analysis of the samples by using DNA methylation markers [66]. The technology requires a clinician for administration, is well-tolerated with promising sensitivity as well as specificity. Another minimally invasive sponge-on-string device is EsophaCap (Capnostics) which was also paired

with DNA methylation markers [67]. There is limited data available but the biomarker panel has shown its potential to discriminate BE with high accuracy. The panel also has been tested on whole oesophageal brushings with good accuracies for detecting BE (AUC 0.84 to 0.94). DNA methylation markers can also be applied to the Cytosponge technology as an alternative to TFF3 [81]. Last, the WATS3d brush is a wide-area, transepithelial, tissue sampler which can be used to obtain large tissue area samples during endoscopy. In a number of studies, this technology has demonstrated its clinical effiacy and superiority to the conventional OGD biopsy protocol [82–84].

In the context of the Cytosponge technology, the remaining shortcoming with respect to clinical evidence is the lack of randomised, controlled trial evidence of the Cytosponge intervention in comparison to standard of care. Another consideration for widespread Cytosponge adoption is the scalability of the technology which is essential for providing access of the test to the relevant patient populations. In particular, the screening of Cytosponge pathology slides is a laborious process. It initially comprises several repetitive tasks such as checking the amount of sampled material (sample adequacy) and the presence of columnar epithelium of gastric type to confirm that the capsule reached the stomach. Cells sampled on withdrawal of the sponge are mainly squamous cells, gastric columnar epithelium, and respiratory epithelium (and sometimes there is a minor inflammatory component such as inflammatory tissue) (fig. 1.5). After successful quality control, the pathologist performs a diagnostic screen for columnar epithelium with goblet cells by using a combination of the H&E and TFF3 stains. Both of these steps, quality control and diagnosis, are crucial for screening throughput and heavily time consuming.

An idealised clinical setting would entail allocation of as much time as possible for the pathologist to investigate cases which require more detailed screening. These are most likely cases with few columnar epithelium and/or ambiguous presence of goblet cells with uncertainty of epithelial type (gastric / respiratory) or atypical, potentially dysplastic cells. This gives rise to a need for an automated stratification system of samples to support and accelerate clinical decision-making by the pathologist.

1.4 Computational pathology & machine learning

Computational pathology incorporates different sources of raw data (e.g. images, '-omics' and patient demographics), building mathematical models, and inferring diagnostic information based on these data [85]. Systems utilising computational pathology rely on digitised

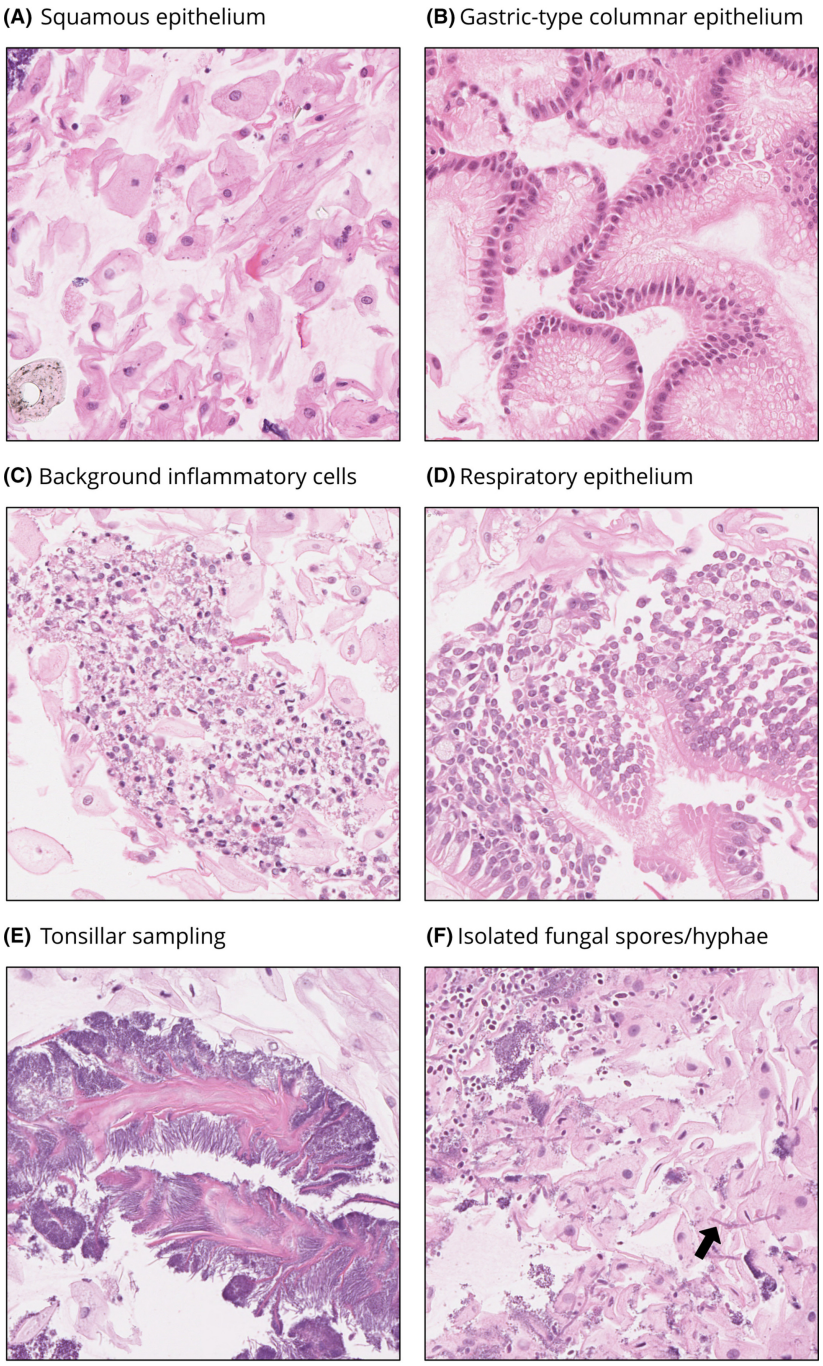


Fig. 1.5 **Normal cellular components of a Cytosponge sample.** All images at 10× magnification. **A**, Squamous epithelial cells from the oesophagus and oropharynx. **B**, Gastric-type columnar epithelium from the stomach and/or a hiatus hernia. **C**, Mixed inflammatory cells, which, provided that they are separate from epithelial cell groups, are not considered to indicate clinically significant oesophagitis. **D**, Strips of respiratory epithelium recognised by their terminal bars and cilia. **E**, Tonsillar sampling characterised by keratinous material and actinomyces organisms. **F**, Fungal spores and occasional hyphae (arrow) with appearances consistent with Candida. As there is no associated significant acute inflammation they are considered to represent commensal organisms rather than being suggestive of Candida oesophagitis. Figure created by me from source: [73].

images of cytological or histological specimens on microscopy slides. They represent a cross-section of relevant biological material which yields information about spatial arrangement in tissue architecture as well as morphological information on a cellular level. This is usually combined with (immuno)histochemical stains which enable screening for morphological or functional information in these samples.

Traditionally, pathologists assess each case individually and subjectively. Therefore, consistency of screening results is not always guaranteed, resulting in suboptimal inter-observer agreement [86, 87], particulary for difficult cases. This is corroborated by the high workload, necessitating rapid screening of individual cases which introduces potential diagnostic errors. [88].

To overcome these issues and allow pathologists to dedicate more time to difficult cases, a framework to assist pathologists in screening is required. Such a framework must integrate the different data available and enable the pathologist to perform processes such as automated quality control or aggregation of diagnostically relevant information.

The excellent pattern recognition ability of the human brain enables rapid classification of labelled and unlabelled images. For pathologists, this classification is learned during specialist pathology training, refined by experience, journal reading, update courses, and peer review. Pathologists generalise well with few images and can apply this knowledge to new situations, even if confounding elements such as stain variation affect the image assessment. When trying to use machine learning to replicate a simple task performed by a pathologist, it is important to consider the choice of training data as well as the type of model. As labelled data are usually available, but limited by the time of pathologists to provide these labels, learning from such examples is achieved by employing supervised learning methods. In brief, a pathologist or expert provides labels for certain images of different tissue types; these examples are then fed into a classifier which is subsequently optimised to produce robust inferences for future images with unknown tissue types. The discrimination of different type-defining characteristics relies on appropriate feature extraction and subsequent classification of these features [89]. Approaches to build such processes have evolved over the last decade, ranging from conventional parametric texture extraction and classification to end-to-end learning using deep convolutional neural networks (CNN) [90]. The latter, a kind of deep learning architecture, have shown outstanding performance when compared to gold standard methods [91] (fig. 1.6a).

Deep learning is a field within machine learning that has grown to prominence in the last several years with the increasing availability of computational capability. In machine learning,

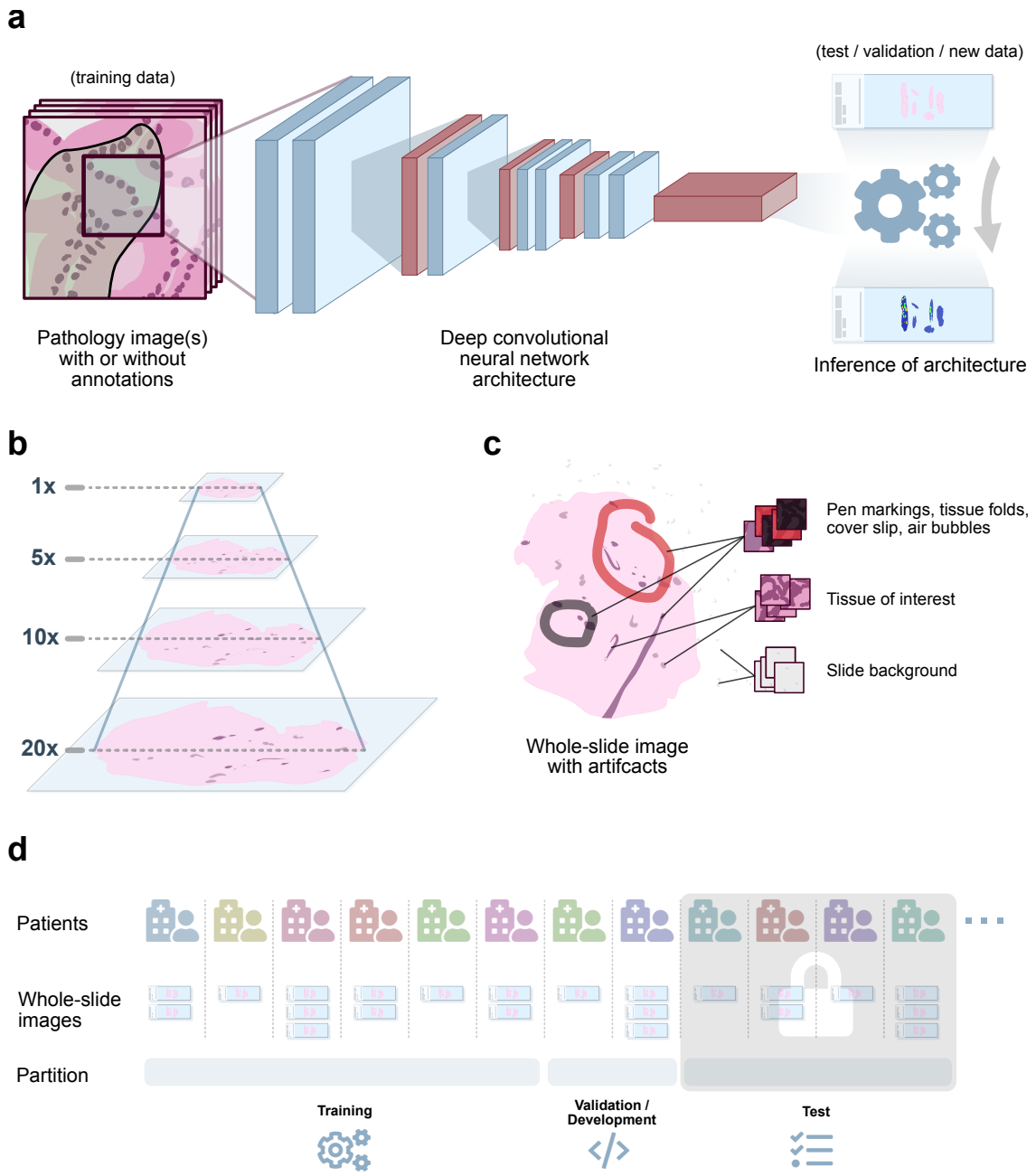


Fig. 1.6 **Overview of general concepts for computational pathology a** A deep convolutional neural network can learn patterns in pathology images and subsequently apply learned classification heuristics on new data. **b** Whole-slide images consist of multiple magnification levels stacked in a pyramidal arrangement. **c** Whole-slide images often have artefacts which need to be accounted for during training. This ranges from pen markings, tissue folds, and slide backgroundd to cover slip boundaries as well as trapped air bubbles. **d** Most approaches in computational pathology rely on distinct partitions of a dataset for model development (e.g. training, validation/development, test). A test set, either internal or external is usually set aside for evaluation of the model to test generalisability and reliability.

input data are transformed into an output not according to deterministic, human-defined rules, but instead, the algorithm learns a mapping between input and output from a collection of ground-truth examples provided to it [92].

In traditional machine learning, prior to training, input data are simplified into a set of features which have been hand-selected by a human in order to reduce the data representation in a way that the learning algorithm will be able to discover patterns to complete the input-output mapping [93]. In deep learning, however, the raw data themselves are fed into the algorithm directly, without any feature-wise simplification [92]. The deep learning algorithm then learns a representation of the data suitable to completing the mapping internally, without any human intervention between raw data and output. This is possible because deep learning algorithms include successive layers of weights with a non-linearity function between them such that increasingly abstract representations of the input data are learned, until a suitable decision boundary can be found to divide the data into the output space. Learning is completed by computing a loss comparing the algorithm's output for a training example with the ground truth of that example, and then performing an algorithm called backpropagation to adjust the weights throughout the network according to that loss [94]. Typically, at least several passes through the full training dataset (called epochs) are needed before the algorithm converges on an optimal set of internal weights for the training set [95].

Due to their large number of layers and weights, deep learning models are able to learn complex relationships between input and output. Deep learning has proven to outperform traditional machine learning and even human performance across many important tasks [96]. However, deep learning models tend to rely on very large training datasets to be successful, so that the model does not overfit to the training data. If model generation is conducted in a proper way and sufficient data are available, the major advantage of not relying on a human to select relevant features for training can significantly improve predictive performance. Hence, deep learning has grown in relevance and popularity enormously across various fields [92].

There are a number of underlying decisions to be made for the generation of a deep learning model, which are known as hyperparameters. Selecting the most appropriate hyperparameters for a training task is a major part of the success of a deep learning system [89, 97]. These include the choice of the receptive field, how to sample from imbalanced classes during training, and selection of the proper performance metric. For example in the case of tissue images, the receptive field is particularly important as it defines the scale at which features are extracted (fig. 1.6b). Furthermore, hyperparameters associated with data augmentation may help to build a robust model applicable on extended datasets. For images,

this particularly includes different spatial and colour distortions which can be introduced during training.

Deep neural networks adapted to image data have proven to match or surpass human performance in many histopathological tasks that previously required an expert pathologist [85, 98]. This includes a range of tasks such as tissue segmentation [99], sample classification [100], survival prediction [101], and genomic/transcriptomic data prediction [102, 103]. Despite major advances in the field in recent years, many approaches still lack considerations for practical implementation in clinical and/or research environments. In a standard clinical workflow, these computational pathology systems need to fulfil three criteria: First, they must provide an interface enabling the pathologist to interact with the data. Second, diagnostic inferences made by the system need to be transparent to allow interpretability. And third, incorporated algorithms and models need to be validated and robust in order to ensure reliability on both tissue and patient level (fig. 1.6c, fig. 1.6d). It remains an open question how clinical decision support systems based on digital pathology can be integrated into the clinical pathway for binary diagnostic tests while fulfilling the three key criteria above.

1.5 Thesis aims

A clear need for early detection of oesophageal cancer has been identified and demands for a novel approach to allow targeted screening of at-risk populations for its precursor lesion Barrett Oesophagus. One of the minimally invasive approaches to detect Barrett, and therefore allow treatment of potentially dysplastic lesions, is the Cytosponge technology in combination with the immunohistochemical marker TFF3. The evidence base for the technology has so far been restricted to a cohort study (BEST1 [74]) and a case-control study (BEST2 [75]). In order to enable adoption and encourage consideration by major public bodies for the development of implementation strategies, randomised controlled evidence is required.

Furthermore, the implication of scale in the context of the Cytosponge technology raises questions around required pathologist capacity for the analysis of samples. Primarily, a significant amount of time is spent by the pathologist to screen clearly negative cases. An automated or semi-automated approach with a minimal false negative rate when compared to a human pathologist would add significant value to enable triaging of pathology slides with the sampled oesophageal cells.

The work presented in this thesis intends to address both of these limitations through:

1. The analysis and evaluation of a pragmatic, randomised controlled trial for Cytosponge-TFF3 in primary care (BEST3) (chapter 2).

2. The development and validation of a deep learning-based tool to enable high-throughput analysis of oesophageal cell samples collected with the Cytosponge technology (chapter 3).

Chapter 2

A pragmatic, randomised controlled trial for Cytosponge-TFF3

Attribution

The text in this chapter was derived from the following publication with changes:
Cytosponge-trefoil factor 3 versus usual care to identify Barrett's oesophagus in a primary care setting: a multicentre, pragmatic, randomised controlled trial
Authors: Fitzgerald RC, Di Pietro M, O'Donovan M, Maroni R, Muldrew B, Debiram-Beecham I, Gehrung M, [...], Sasieni P. *The Lancet 2020*

Personal contributions

This chapter covers the BEST3 study for the evaluation of Cytosponge-TFF3 in primary care. The study was a highly collaborative effort with a large number of individuals involved. My specific contributions to key aspects were:

- Definition and implications for methodology of primary endpoint coding and associated statistical analysis.
- Definition, clinical considerations, and implementation of secondary endpoint assessments.
- Cleaning of data on the basis of raw case report form inputs.

- Statistical analysis of data for assessment of primary endpoint as well as secondary endpoints. (Source code: https://github.com/9xg/phd-thesis-chap2)
- Interpretation of statistical results and development of an understanding regarding clinical implications.
- Aggregation and calculation of statistics and preparation of visual elements for presentation in figures and tables. (figs. 2.1 and 2.2 and tables 2.1, 2.3, 2.4 and 2.8).

A number of these points involved significant discussions between various contributors and multiple iterations. In order to conform to the strict requirements of the Clinical Trials Unit (Director Peter Sasieni) and the Trial Regulations, the statistical analysis was performed by me and Roberta Maroni (Lead Trial Statistician) with supervision from Peter Sasieni. Detailed discussions resolved any discrepancies. An independent trial monitoring committee signed off on the data prior to publication. The Statistical Analysis Plan has been published with the main manuscript [104].

Abstract

The aim of this chapter was the analysis and evaluation of a pragmatic, randomised controlled trial for Cytosponge-TFF3 in primary care. I particularly focused on investigating a diagnostic strategy for the detection of BE, dysplasia, and early cancer. Treatment of dysplastic BE or early cancer has been shown to prevent progression to oesophageal adenocarcinoma.

We conducted a multi-site randomised controlled trial in primary care to evaluate whether offering a Cytosponge-TFF3 test to patients (age $\geq$50 years) on acid-suppressant medication for reflux symptoms increases the rate of BE diagnosis and results in earlier cancer detection. Individuals were randomised 1:1 to receive the Cytosponge-TFF3 test or standard management of reflux. TFF3 positive cases underwent endoscopy. Endpoint data were from coded diagnoses to ensure equity across the arms.

13,657 patients were randomised from socio-demographically diverse GP practices in England. Of 6,832 patients in the intervention arm, 2,679 (39%) expressed interest and 1,750 attended for a Cytosponge examination. 1,654 (99%) patients successfully swallowed the device with a male to female ratio of 48:52 and a median age of 69 years (range 50-96). There were 140 Barrett diagnoses in the Intervention arm (ITT) compared to 13 in usual care giving a rate ratio of 10.2 (95% CI 5.8-18.1), and 10.6 (95% CI 6.0-18.8) when adjusted for cluster randomisation ($p<0.0001$). There were 9 individuals with dysplastic Barrett or stage I oesophago-gastric cancer in the intervention arm and none in the control arm. Overall, 8% of those who undertook a Cytosponge exam had BE and 59% of those endoscoped for a positive Cytosponge-TFF3 result had BE, dysplasia or early cancer.

We were able to show that in patients with reflux the offer of Cytosponge-TFF3 testing results in improved detection of BE and earlier stage cancer compared with usual care.

2.1 Introduction

Heartburn symptoms caused by gastro-oesophageal reflux disease are common, affecting up to 20% of adults in northwest Europe, North America, Australia, and New Zealand and leading to enormous annual healthcare costs [105, 106]. Most of these individuals do not have a diagnosis and are treated over many years with acidsuppressant medication therapy. Symptoms of heartburn are important when one considers the link between heartburn and oesophago-gastric cancer [107]. It is estimated that 3–6% of individuals with gastro-oesophageal refluxpredominant symptoms could have the precursor lesion to oesophageal adenocarcinoma, known as Barrett oesophagus. However, only around 20% of patients with Barrett oesophagus are diagnosed. Therefore, most cases of oesophageal adenocarcinoma are diagnosed *de novo*, without the opportunity to prevent progression [108–110].

The incidence of oesophageal adenocarcinoma is six times higher than it was in the 1990s. Oesophageal adenocarcinoma also has a dismal prognosis due to late presentation, with an overall 5-year survival of less than 20%, despite advances in neoadjuvant therapy and surgery [111, 34]. Clinical guidelines recommend urgent referral for an endoscopy in patients with warning symptoms, such as dysphagia and weight loss, and routine referral for an endoscopy in those with symptoms of gastro-oesophageal reflux that persist despite recommended lifestyle and pharmacological management strategies, and those with multiple additional risk factors for the disease [112, 113, 57, 114]. Nevertheless, the proportion of patients referred for an endoscopy from general practice clinics varies widely, and the referral rates per practice correlate with the stage at diagnosis [115].

A modelling study [116] using data from the USA estimated that the burden of oesophageal adenocarcinoma could be reduced by up to 50% through implementing strategies for the systematic screening and early diagnosis of individuals with gastro-oesophageal reflux, who would otherwise not have been referred for an endoscopy. Early detection needs to be combined with effective interventions to be clinically beneficial. There have been important advances in outpatient-based endoscopic therapies, which are now recommended for low-grade and high-grade dysplasia in Barrett oesophagus, with low rates of recurrence [117–119]. Patients with intramucosal stage I cancers have a survival of more than 90% and can be treated endoscopically, thus mitigating the risks of and side-effects from systemic therapy and an oesophagectomy, which is often required for more advanced disease [120, 121]. In view of the scale of gastro-oesophageal reflux disease, and the costs (both psychological and financial) of investigation, any new clinical strategy needs to be carefully evaluated.

We have developed a test for Barrett oesophagus that is suitable for use in the primary care setting. The test comprises a non-endoscopic cell collection device coupled with an *in vitro* test for the specific biomarker, TFF3, that identifies intestinal metaplasia (the histopathological hallmark of Barrett oesophagus [73]; fig. 2.1). So far, two clinical studies [74, 75] of this new clinical strategy, termed the Cytosponge-TFF3 procedure, have been done in over 2000 patients, with promising data on safety, acceptability, accuracy, and cost-effectiveness [79, 80, 69].

We did this pragmatic, randomised controlled trial, involving patients with recurrent symptoms of gastrooesophageal reflux who had been taking acid-suppressant medication prescribed by their general practitioner, to investigate whether the Cytosponge-TFF3 test, administered in the community setting, leads to the identification of more patients with Barrett oesophagus than does usual clinical practice for endoscopy referral in England. The findings of this trial will lay the foundation for adoption of the Cytosponge-TFF3 test in order to develop real-world implementation strategies. The generated results will also be essential for future health economic analyses assessing the cost effectiveness of the test in primary care in line with requirements for diagnostic or screening methods (chapter 1).

2.2 Methods

2.2.1 Study design and participants

This multi-centre, pragmatic, randomised controlled trial took place in 109 sociodemographically diverse general practice clinics in England.

Patients were eligible for inclusion if they were aged 50 years or older and had records of having been prescribed acid-suppressant therapy (proton-pump inhibitor or histamine-2 receptor antagonists) for at least 6 months in the previous year. Patients with records of having been prescribed non-steroidal anti-inflammatory drugs together with acid-suppressant therapy, suggesting that their reflux symptoms were not the primary basis for the proton-pump inhibitor prescription, and patients who had undergone an endoscopy in the previous 5 years or with a previous diagnosis of Barrett oesophagus, were excluded from the study. All potential participants received an introductory letter to the study and were given 14 days to opt out, after which point they were enrolled in the trial.

The study protocol, which was approved by a central ethics committee, has been made publicly available [122]. Aggregated data were collected from participating primary care

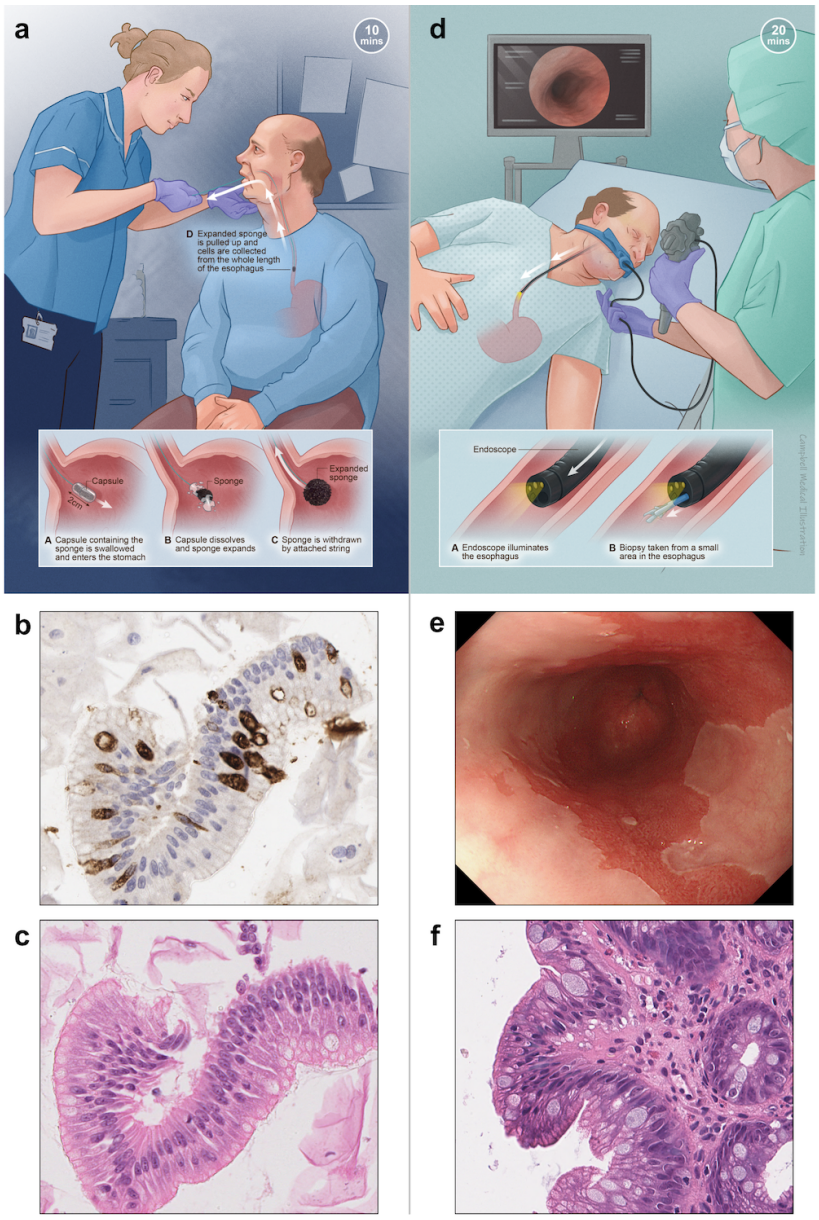


Fig. 2.1 **Comparison of the Cytosponge-TFF3 procedure with the endoscopy procedure.** **a)** Administration and passage of the Cytosponge-TFF3 device to obtain a sample of oesophageal epithelial cells. **d)** Administration and passage of an endoscope to visualise the oesophagus. The Cytosponge-TFF3 sample is processed to a paraffin block and stained with an antibody against **b)** TFF3 and with **c)** haematoxylin and eosin (magnification ×200). **e)** Endoscopic white light view of Barrett oesophagus in deep red compared with surrounding light pink squamous epithelium. **f)** One or more endoscopic biopsies are taken and the tissues are stained with haematoxylin and eosin for histopathological assessment (magnification ×200). **a)** and **d)** were drawn by Campbell Medical Illustration (Glasgow, Scotland). TFF3=trefoil factor 3.

clinics for all potential participants who did not opt out. Written informed consent was obtained before collecting any individual-level patient data and before any clinical procedure was done.

2.2.2 Randomisation and masking

Initially, general practice clinics (ie, clusters) were randomly assigned (1:1) to either the usual care group, in which eligible patients with gastro-oesophageal reflux under the care of these clinics received standard management of their symptoms and were only referred for an endoscopy if required, or the intervention group, in which eligible patients received standard management and were offered the Cytosponge-TFF3 procedure, with a subsequent endoscopy if the procedure identified TFF3-positive cells.

Approximately two-thirds of the way through recruitment, the trial switched to an individual patient-level randomisation approach, which was approved by an independent trial steering committee, the research ethics committee, and the Medicines and Healthcare products Regulatory Agency (MHRA). Cluster randomisation was initially chosen in order to remove selection bias by general practitioners; however, in the trial, all patients were selected by use of the prescribing database, so selection bias was not an issue. After recommendation by the trial steering committee, we switched to individual randomisation during the study, which substantially increased the statistical power and also satisfied patient and clinician demand for the Cytosponge procedure to be available in all clinics. Data from both the cluster and individual randomisations were analysed separately before they were combined, having established that their results were consistent, as required by an independent data monitoring committee and the MHRA.

The trial statistician did the cluster randomisation of general practice clinics by randomly sorting practices within strata (using computer-generated random number sequences) and then allocating alternately. Clinics were not randomly assigned until they had agreed to participate. Individual patient-level randomisation was done by the general practice clinics directly using the BEST3 app, which used a computer-generated random number sequence. Potential participants in both the clinic-level and the patient-level randomisations were informed about the research and given the option to opt out of participation (including data collection) before knowing which group they would be assigned to. All patients who were randomly assigned were followed passively for a weighted overall average of approximately 12 months (range 8–18 months).

The chief investigator (Rebecca Fitzgerald) and the lead statistician (Peter Sasieni) were masked to the aggregated results by group until follow-up was complete. Pathologists analysing endoscopic biopsies for Barrett oesophagus did not know whether the patient had undergone a Cytosponge-TFF3 procedure.

2.2.3 Procedures

Participants randomly assigned to the usual care group received standard care, in which they received prescriptions for acid-suppressant medication and their general practitioner might have provided lifestyle advice or referral for an endoscopy, depending on the severity of their symptoms. Participants randomly assigned to the intervention group received a letter inviting them to undergo a Cytosponge-TFF3 test and, if they expressed interest, were subsequently screened by a nurse via a telephone interview. Sometimes patients were not contactable by telephone or they changed their mind in the intervening period. The telephone screening interview included a symptom screen to ascertain whether participants were taking acid-suppressant therapy for heartburn-predominant symptoms and to exclude any participants who were not deemed to be suitable for the Cytosponge-TFF3 procedure.

Participants were not offered a Cytosponge-TFF3 test if they had dysphagia (as the capsule might not reach the distal oesophagus) or if they were at an increased risk of bleeding because of known cirrhosis or varices, or if they were unable to stop taking anticoagulants. Such participants were still included in the intention-to-treat analysis.

The Cytosponge device was administered by a general practice clinic nurse or a Clinical Research Network nurse, following a training seminar and one-to-one training with an experienced practitioner (Irene Debiram-Beecham), until they were signed off as competent. Samples collected from the Cytosponge procedure were processed centrally and assessed for the presence of Barrett oesophagus by use of haematoxylin and eosin staining and immunohistochemical staining for TFF3 (Ventana Medical Systems, Tuscon, AZ, USA), as described previously [73].

TFF3 staining was evaluated by experienced upper gastrointestinal pathologists, and consensus agreement from two or three pathologists was used in any cases of uncertainty. A sample in which no glandular cells were present was deemed to provide a low-confidence result, as the device might not have reached the stomach and a diagnosis of distal Barrett oesophagus might have therefore been missed. Any sample with glandular groups of cells (indicating that the device had reached the stomach), and that did not have equivocal TFF3

staining, was considered a high-confidence result. Patients with low-confidence or equivocal results, and any with processing failure, were offered a repeat Cytosponge-TFF3 test. All patients with a positive TFF3 test result were offered an endoscopy to confirm the diagnosis of Barrett oesophagus and inform treatment.

After completion of trial follow-up, a random sample of participants from each study group were invited to undergo a research endoscopy procedure. The results of these research endoscopies will be presented elsewhere. All endoscopy samples (both in the usual care group and in the intervention group) were analysed by the local pathologist. Participants with Barrett oesophagus diagnosed by use of the Cytosponge-TFF3 test also had their endoscopic biopsies centrally reviewed to confirm that intestinal metaplasia was present and to identify any dysplasia or cancer (by H&E staining).

A census date 8–18 months after randomisation was set for each general practice clinic. Passive follow-up of all participants, irrespective of study group or whether they had undergone a Cytosponge-TFF3 procedure, was done up to the census date. Census dates were chosen independently of the randomisation, so as to have a weighted average follow-up of 12 months.

The endpoint data collected were coded diagnoses of Barrett metaplasia, Barrett dysplasia, or oesophago-gastric adenocarcinoma, ascertained by at least one of the following three methods: (1) an electronic search of general practice clinic records for new diagnoses of Barrett oesophagus or oesophageal adenocarcinoma, new referrals to gastroenterology departments, or new referrals for oesophagogastroduodenoscopy procedures within the study period, followed by a manual search of the clinical records for those patients with a potential diagnosis of Barrett oesophagus or oesophagogastric adenocarcinoma identified by the electronic search; (2) a full manual search of the general practice clinic records for all participants registered with that clinic; and (3) secure anonymous record linkage between participating general practice clinics and participating endoscopy units to identify individuals who were both study participants and who had been diagnosed with Barrett oesophagus or oesophago-gastric adenocarcinoma during the study period.

2.2.4 Outcomes

The primary outcome was the diagnosis of Barrett oesophagus at 12 months after enrolment, expressed as rate per 1000 person-years, in all participants in the intervention group (regardless of whether they had accepted the offer of the Cytosponge-TFF3 procedure) compared

with all participants in the usual care group. The secondary outcomes were as follows: uptake of the Cytosponge-TFF3 procedure; the number of cases of Barrett oesophagus with dysplasia and intestinal metaplasia-associated cancer, by stage at diagnosis; the positive predictive value of the Cytosponge-TFF3 test, measured in the subset of patients who had a subsequent endoscopy after testing positive for TFF3; and the acceptability and safety of the Cytosponge-TFF3 test.

2.2.5 Statistical analysis

By use of a series of key assumptions (see Statistical Analysis Plan [104]) about the prevalence of Barrett oesophagus, the proportion of endoscopy referrals, and the sensitivity and uptake of the Cytosponge-TFF3 procedure, the expected proportions of Barrett oesophagus diagnoses over 12 months were calculated as 1.38% in the intervention group and 0.60% in the usual care group. On the basis of these assumptions, we calculated that a sample size of 6764 patients was required to achieve a power of 90% and a significance level of 5% if randomisation was done at the individual patient level. To account for the cluster-randomisation design, a variance inflation factor was estimated by strata (defined by number of patients from each clinic who were randomly assigned; 48–65, 66–90, 91–125, 126–175 or 176–198 patients) for the cluster-randomised group, assuming that the intra-class correlation coefficient of the proportion of patients with Barrett oesophagus was 0.025. The actual numbers of participants in each strata were divided by the variance inflation factor to yield the equivalent numbers of individually-randomised patients.

Throughout the trial, we ensured that the projected sum of the equivalent size of the cluster-randomised group and the size of the individual patient-level randomised group was at least 6764 participants. The primary endpoint of Barrett oesophagus diagnoses (excluding cancer diagnoses) in both groups at 12 months after enrolment, was analysed by use of a random-effects log-linear model. The number of Barrett oesophagus diagnoses was the Poisson-distributed outcome, with a fixed effect for the strata, a random effect for each clinic, and an offset for the number of person-years of follow-up. We assumed two different treatment effects (fixed rate ratios / RRs) for the intervention group (one in the first 4 months and the second thereafter) that were eventually combined at a weight ratio of 1:2. In the usual care group, the treatment effect was assumed to be constant over time.

The analysis was first done for the cluster-randomised group, then for the individual patient-randomised group (with no cluster effect), and finally for the whole dataset. When

analysing the whole dataset, the individual patient-randomised group was assigned to a separate stratum. This method was approved by the MHRA.

As only aggregated data about age and sex were available, and we only had access to individual-level data on age, sex, and medication history for patients who successfully swallowed the Cytosponge, no adjustment was made for these factors in the analysis of the primary outcome. Statistical significance was based on a two-sided test with an alpha-value of 5%. The uptake of the Cytosponge-TFF3 procedure was assessed as the number of patients who successfully swallowed the capsule, expressed as a proportion of the patients who were offered the procedure. The number of patients with Barrett oesophagus, Barrett oesophagus with dysplasia, or Barrett oesophagus and cancer is reported by study group and also by the number of participants who underwent the Cytosponge-TFF3 procedure in the intervention group. The positive predictive value of the Cytosponge-TFF3 procedure was calculated from the proportion of patients who underwent the procedure, in whom the subsequent endoscopy and pathological assessment confirmed the diagnosis of Barrett oesophagus, Barrett dysplasia, or cancer (gold standard).

The acceptability of the Cytosponge-TFF3 procedure was estimated from a questionnaire, in which participants rated the procedure using an 11-point visual analogue scale (from 0 to 10); the median and IQR are reported, together with the proportion of participants who scored 5 or more (indicating that the test was somewhat acceptable). The safety of the Cytosponge-TFF3 procedure was assessed by recording any adverse events and serious adverse events that had occurred within 7–14 days of undergoing the procedure. This assessment was done proactively by a nurse via a telephone call with patients. The proportion of patients who had an adverse event, and the type and severity of adverse event, is reported. The adverse events were only collected for participants undergoing the Cytosponge-TFF3 procedure. Since endoscopy is standard of care, no adverse event data was collected in relation to this procedure.

Statistical analyses were done in Stata version 15 (StataCorp LLC, College Station, TX, USA). Pseudorandom numbers for all randomisations were generated in R (R Core Team [2019]). An independent data monitoring committee and a trial steering committee, which included two lay members who provided a patient's perspective, oversaw the trial. The trial is registered with the ISRCTN registry, number ISRCTN68382401.

2.3 Results

2.3.1 Study enrolment, randomisation, demographics, and exclusion of patients

Between March 20, 2017, and March 21, 2019, 113 general practice clinics located in socio-demographically diverse regions in England were enrolled, but four clinics dropped out shortly after being randomly assigned (three in the usual care group and one in the intervention group), leaving 109 clinics, comprising 13,657 patients. These patients were sent an introductory letter and given 14 days to opt out of the study. 143 of these patients opted out before individual patient-level randomisation, leaving 13 514 patients to be randomly assigned. After randomisation, 136 patients in the usual care group and 122 patients in the intervention group withdrew. Additionally, 17 patients (ten in the intervention group and seven in the usual care group) were excluded because they had either died or had been diagnosed with Barrett oesophagus before randomisation, and 17 patients (all in the intervention group) were excluded because their contact details were absent. Of the remaining 13,222 enrolled patients, 7,839 patients from 75 clinics were cluster-randomised, and 5,383 patients from 34 clinics were individually randomised. Overall, 6,388 participants were randomly assigned to the usual care group and 6,834 participants to the intervention group (fig. 2.2).

The demographics of the 13,222 participants included in the final analysis are summarised (table 2.1). The age distribution of participants who successfully swallowed the Cytosponge was similar to that of all participants. The randomly assigned clinics represented all ten deciles of the Index of Multiple Deprivation (data not shown). The median decile of deprivation among participants was seven (with one being the most deprived and ten the least deprived) and the lower quartile was four.

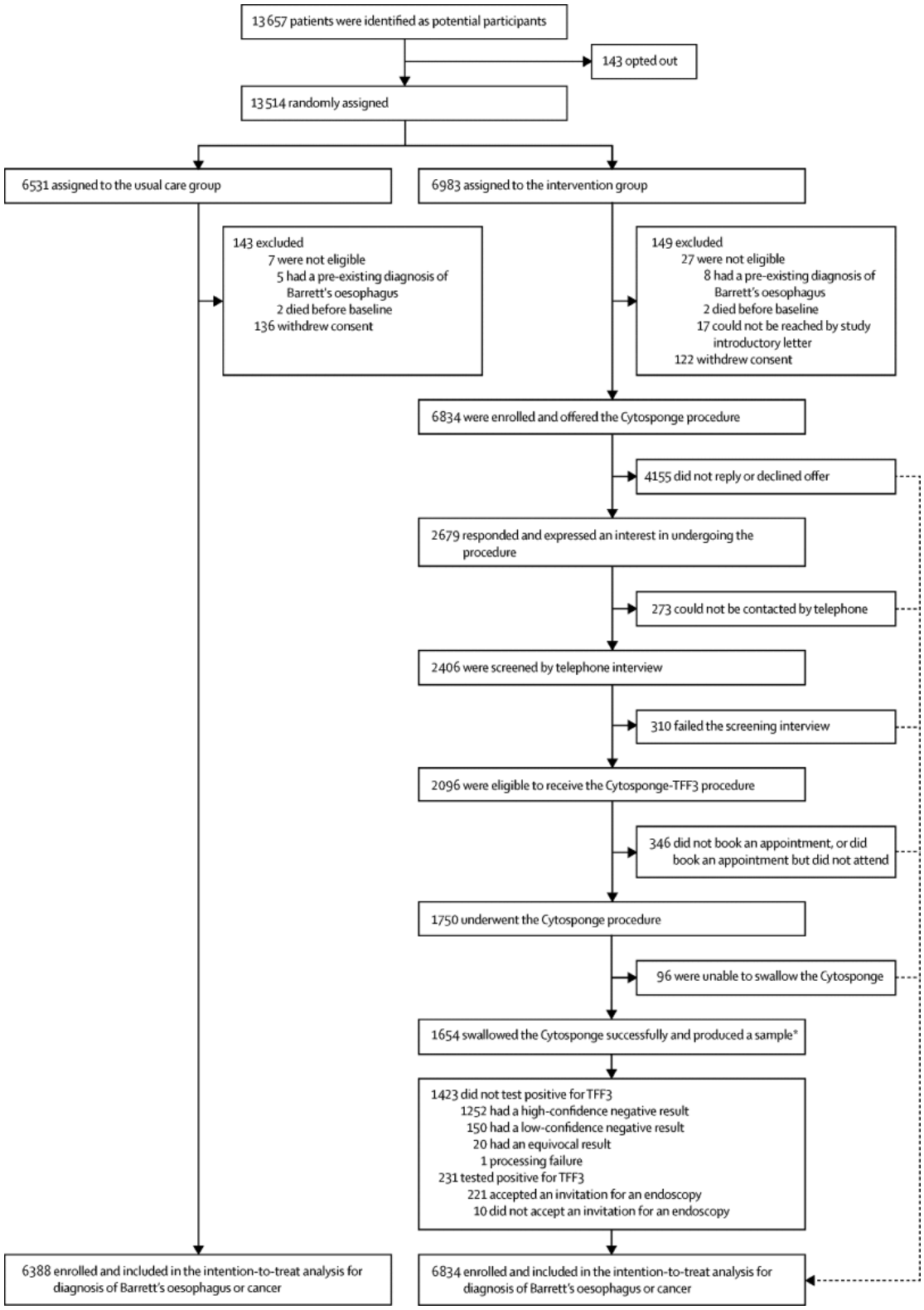


Fig. 2.2 **Trial profile.** *202 (12%) of these 1,654 participants had a repeat Cytosponge test, as the first sample yielded a low-confidence result (defined as the absence of glandular cells in the sample) and a diagnosis of Barrett oesophagus could have therefore been missed; patients with equivocal results, or technical or processing failures, were also invited for a repeat test.

	All participants (n=13,222)	Usual care group (n=6388)	Intervention group	
			All participants (n=6834)	Participants who successfully swallowed the Cytosponge (n=1654)
Sex				
Male	6030 (46%)*	..	..	796 (48%)
Female	7155 (54%)*	..	..	858 (52%)
Age distribution, years				
50–59	3171 (24%)*	..	..	326 (20%)
60–69	4001 (30%)*	..	..	562 (34%)
70–79	4172 (32%)*	..	..	615 (37%)
80–89	1642 (12%)*	..	..	140 (8%)
90–99	199 (2%)*	..	..	11 (1%)
Size of general practice surgery				
Small (48–90 patients)	2083 (16%)	1038 (16%)	1045 (15%)	..
Medium (91–160 patients)	6746 (51%)	3071 (48%)	3675 (54%)	..
Large (161–231 patients)	4393 (33%)	2279 (36%)	2114 (31%)	..
Medication use				
Proton-pump inhibitor only	11818 (92%) †	..	..	1434 (87%) ‡
Histamine-2 receptor antagonists only	613 (5%) †	..	..	170 (10%) ‡
Proton-pump inhibitor plus histamine-2 receptor antagonist	413 (3%) †	..	..	43 (3%) ‡
Socioeconomic factors				
Median Index of Multiple Deprivation decile §	7 (4–9)	6 (4–9)	7 (5–9)	..

Table 2.1 **Baseline characteristics of all randomly assigned participants.** Data are n (%) or median (IQR). Most data were aggregated by site; therefore, there are no data for some fields. * Baseline data were available in an aggregated form; data for age and sex are missing from one site. † Baseline data were available in an aggregated form; data for medication are missing from six sites. ‡ Data for seven patients are not shown in the table, as they had records of over-the-counter acid-suppressant or other medications. § The Index of Multiple Deprivation (with a score of 1 indicating most deprived and a score of 10 indicating least deprived) scores were not available for individual participants and were calculated by assigning the score for each general practice clinic to each patient.

2.3.2 Interest, uptake and results of Cytosponge-TFF3 test

Following a written invitation, 2679 (39%) of 6834 patients in the intervention group responded and expressed an interest in taking part in the Cytosponge-TFF3 procedure. Of these, 2,096 (78%) participants were eligible following the telephone assessment, and 1,750 (65%) provided consent and underwent the procedure. 1654 (95%) of these participants (and 24% of all 6,834 participants in the intervention group) successfully swallowed the device, including 796 men (48%) and 858 (52%) women, with a median age of 69 years (IQR 61–74; table 2.1). 311 (19%) of the 1654 participants who had successfully swallowed the device had a low-confidence negative or equivocal test result, and depending on local capacity, were invited for a repeat Cytosponge-TFF3 test. 202 (65%) of these participants attended the repeat appointment, 190 (94%) of whom successfully swallowed the device, leading to a further 140 patients producing a high-confidence (positive or negative) result. Overall, after the repeat test, 150 (9%) of the 1,654 patients who successfully swallowed the Cytosponge-TFF3 still produced a low-confidence negative result (fig. 2.2). Apart from the eight participants who were found, on review, to have pre-existing Barrett oesophagus, all participants who were invited for the Cytosponge-TFF3 procedure were included in the final intention-to-treat analysis, regardless of whether or not they accepted the invitation.

2.3.3 Offer of Cytosponge-TFF3 results in increased number of Barrett diagnoses when compared to usual care

Barrett oesophagus diagnoses in both groups had to be identified from records of clinical coded diagnoses at all general practice clinics included in the study, the electronic records of local referral hospitals, or both, to ensure equity across the usual care and intervention groups (otherwise, diagnoses from the intervention group would have been more easily ascertained). One diagnosis of Barrett oesophagus in a patient who had a positive Cytosponge-TFF3 test result was omitted from the results, as a coded diagnosis was not identified by any of these data collection methods.

We identified 140 Barrett oesophagus diagnoses in the intervention group (127 in patients who underwent the Cytosponge-TFF3 procedure, and 13 in patients who did not undergo the Cytosponge-TFF3 procedure) compared with 13 diagnoses in the usual care group (table 2.2, table 2.3; see table 2.5 and table 2.7 for the corresponding tables for randomisation groups and a breakdown of the length of Barrett oesophagus detected). 87 (69%) of the 127 participants who were diagnosed with Barrett oesophagus from the Cytosponge-TFF3 procedure were

male. As the results of the cluster-level randomisation and patient-level randomisation both favoured the intervention group, an overall RR was calculated (table 2.2). The estimated cumulative rate of Barrett oesophagus at 12 months was 20.2 per 1,000 person-years in the intervention group and 2.0 per 1,000 person-years in the usual care group (RR adjusted for cluster randomisation 10.6 [95% CI 6.0–18.8], p<0.0001; table 2.2).

	Usual care group (n=6388)	Intervention group (n=6834)	Absolute difference in rates per 1000 person-years (95% CI)	Overall rate ratio (95% CI); p value	Adjusted rate ratios (95% CI); p value		
					Cluster randomised group	Patient-level randomised group	Overall*
Number of participants diagnosed with Barrett oesophagus	13 (<1%)	140 (2%) †	..	..	..	..	..
Follow-up, person-years	6579	6952	..	..	..	..	..
Rate of Barrett oesophagus, per 1000 person-years	2.0	20.2 ‡	18.3 (14.8–21.8)	10.2 (5.8–18.1)	10.0 (5.0–20.0); p<0.0001	12.0 (4.3–33.2); p<0.0001	10.6 (6.0–18.8); p<0.0001

Table 2.2 **Barrett oesophagus diagnoses in the usual care group compared with the intervention group.** Data are n (%), unless otherwise specified. * Overall adjusted rate ratio is a combined rate ratio of the two randomisation groups (cluster randomisation and individual patient-level randomisation) and accounts for clustering. † Number of participants diagnosed with Barrett oesophagus in the intervention group includes all participants who were offered the Cytosponge procedure. ‡ The rate of Barrett oesophagus in the intervention group was calculated as the weighted average of the rate in the first 4 months of follow-up and the rate in the following months, with a weight ratio of 1:2.

2.3.4 Cytosponge-TFF3 can detect Barrett oesophagus with a high positive predictive value

Of 1,654 participants in the intervention group who successfully swallowed the Cytosponge device, 221 (13%) with a positive TFF3 result had a subsequent confirmatory endoscopy. 127 (57%) of these participants were diagnosed with Barrett oesophagus (one of whom had low-grade dysplasia, and three of whom had high-grade dysplasia), and four (2%) participants were diagnosed with stage I oesophago-gastric cancer. Therefore, the Cytosponge-TFF3 procedure had a positive predictive value of 59% (131 of 221 confirmatory endoscopies in patients with a positive Cytosponge-TFF3 result) for Barrett oesophagus, dysplasia, or oesophago-gastric cancer (table 2.2, table 2.3). Of those 90 participants who received a confirmatory endoscopy that did not result in a diagnosis of Barrett oesophagus, dysplasia, or cancer, a further 33 (37%) participants had intestinal metaplasia, identified from a single biopsy taken from the cardia or at the gastrooesophageal junction. Using the available data, we calculated the empirical intraclass correlation coefficient of the proportion of patients with Barrett oesophagus, and found that this value was similar to the expected empirical intraclass correlation coefficient (approximately 0.025). For the secondary endpoints, we compared the number of endoscopic diagnoses of dysplasia and cancer in participants who were offered the Cytosponge-TFF3 procedure with the number of these diagnoses in participants in the usual care group (intention-to-treat analysis).

2.3.5 Cytosponge-TFF3 can detect dysplasia and oesophago-gastric cancer

Nine (<1%) of 6,834 participants were diagnosed with dysplastic Barrett oesophagus (n=4) or stage I oesophago-gastric cancer (n=5) in the intervention group, whereas no participants were diagnosed with dysplastic Barrett oesophagus or stage I oesophagogastric cancer in the usual care group (table 2.3). Of these nine participants in the intervention group, eight were detected as a result of a positive Cytosponge-TFF3 test and a subsequent endoscopy and have all undergone a curative intervention (seven participants underwent endoscopic therapies, and one participant underwent an oesophagectomy for a stage IB cancer involving the first layer of the submucosa; table 2.8). Among those who were offered the Cytosponge-TFF3 procedure but did not have the test (n=5,084), one participant, who initially expressed interest in the procedure, but was referred for an endoscopy before it could be done, was diagnosed with early-stage cancer. Of all 6,388 participants in the usual care group included in the

final analysis, three participants were diagnosed with cancer, of whom two participants were palliative at presentation and died during the study period (table 2.8). In the intervention group, two participants who did not undergo the Cytosponge-TFF3 test were diagnosed with stage IV oesophago-gastric cancer.

	Usual care group (n=6388)	Intervention group		
		Underwent the Cytosponge procedure (n=1750)	Did not undergo the Cytosponge procedure (n=5084)	Overall (n=6834)
Grade of dysplastic Barrett oesophagus				
No dysplasia	13	116	13	129
Indefinite	0	7	0	7
Low-grade	0	1	0	1
High-grade	0	3	0	3
Total	13	127	13	140
Oesophago-gastric cancer stage				
I	0	4	1	5
II	1	0	0	0
III	1	0	0	0
IV	1	0	2	2
Total number of participants with Barrett oesophagus, cancer, or both	16	131	16	147

Table 2.3 **Number of individuals with Barrett oesophagus in the usual care group and intervention group with or without cancer, by grade of dysplasia and cancer stage.**

2.3.6 Acceptability of the Cytosponge-TFF3 is consistent high combined with a small number of adverse events

In the intervention group, an acceptability score for the Cytosponge-TFF3 procedure was provided by 1,464 (89%) of 1,654 participants approximately 1 week after they underwent the procedure. The median acceptability score was 9 (IQR 8–10), with 10 being completely acceptable, and 1,427 (97%) of 1,464 participants gave a score of 5 or higher (table 2.9).

In the intervention group, one serious adverse event associated with the Cytosponge-TFF3 procedure was reported (detachment of the sponge from the thread requiring endoscopy to retrieve the expanded sponge with no adverse sequelae), and three serious adverse events unrelated to the procedure were reported (table 2.4). Of 1,654 participants who successfully swallowed the Cytosponge device, 142 (9%) participants reported an adverse event, including 63 (4%) participants who had a sore throat that required medication or that interfered with eating (table 2.4).

2.4 Discussion

Summary

The results in this chapter directly address the identified need to generate randomised clinical evidence for the Cytosponge-TFF3 test as elaborated in chapter 1. In this multicentre, pragmatic, randomised controlled trial we found that an invitation to have a Cytosponge-TFF3 test led to increased diagnosis of Barrett oesophagus when compared with usual care by general practitioners. This comparison was made in patients identified as being high-risk for this condition, on the basis of a systematic search of electronic patient records for anti-gastro-oesophageal reflux medication. With regard to the secondary endpoint of the proportion of cancer diagnoses, although the numbers were small, we found that all participants in the intervention group who had dysplasia and cancer identified by the Cytosponge-TFF3 procedure were suitable for curative therapy, whereas the cancers detected in the usual care group, and among participants who did not undergo a Cytosponge-TFF3 procedure, had more advanced disease (four of six participants had stage III and IV cancer) and two of these were palliative at presentation and died during the study period. For a device to be suitable for use in general practice clinics, its uptake, safety, and acceptability are key.

	Adverse event severity (n=142)			Total (n=142)
	Low (n=112)	Moderate (n=23)	High (n=7)	
Adverse event				
Sore throat	57 (51%)	5 (22%)	1 (14%)	63 (44%)
Dyspepsia indigestion reflux	16 (14%)	3 (13%)	0	19 (13%)
Oesophageal or gastric pain	11 (10%)	2 (9%)	2 (29%)	15 (11%)
Feeling non-specifically unwell	6 (5%)	3 (13%)	0	9 (6%)
Nausea or vomiting	5 (4%)	3 (13%)	0	8 (6%)
Voice disturbance	3 (3%)	1 (4%)	0	4 (3%)
Diarrhoea or an upset stomach	4 (4%)	1 (4%)	0	5 (4%)
Chest pain or discomfort	2 (2%)	0	0	2 (1%)
Allergic reaction	1 (1%)	0	0	1 (1%)
Anxiety	1 (1%)	0	0	1 (1%)
Bad taste	1 (1%)	0	0	1 (1%)
Paroxysmal positional vertigo	1 (1%)	0	0	1 (1%)
Blood clot excretion	1 (1%)	0	0	1 (1%)
Vasovagal attack	1 (1%)	0	0	1 (1%)
Nosebleed	1 (1%)	0	0	1 (1%)
Headache	1 (1%)	1 (4%)	0	2 (1%)
Bloodshot eye	0	1 (4%)	0	1 (1%)
Chest infection	0	1 (4%)	0	1 (1%)
Abrasion	0	1 (4%)	0	1 (1%)
Fall	0	1 (4%)	0	1 (1%)
Serious adverse event				
Unconscious after minor accident	0	0	1 (14%)	1 (1%)
Detachment of the sponge on day of the procedure	0	0	1 (14%)	1 (1%)
Hernia*	0	0	1 (14%)	1 (1%)
Myocardial infarction †	0	0	1 (14%)	1 (1%)

Table 2.4 **Adverse events in participants who underwent the Cytosponge-TFF3 procedure.** Data are n (%). All percentages calculated with the total number of participants in that column who had an adverse event as the denominator. The severity of adverse events was classified as low, moderate, or high by the nurse during the proactive follow-up telephone call with the patient. Serious adverse events were those classified according to the regulatory requirement for a device trial. * Hernia was repaired 5 days after the procedure. † Occurred 3 days after the procedure.

Acceptability and safety of Cytosponge-TFF3

The acceptability data obtained in our study are encouraging, with a median acceptability score of 9 out of 10, consistent with our previous trials [74, 75]. In this pragmatic trial done across a wide range of demographic areas across England, the proportion of all participants in the intervention group (n=6,834) who expressed an interest in the Cytosponge-TFF3 procedure was 39% (n=2,679), and 24% (n=1,654) of participants had the procedure and successfully swallowed the device, after accounting for inclusion and exclusion criteria and scheduling limitations.

Since we anticipate the Cytosponge-TFF3 procedure being offered by a patient's general practice clinician during a consultation for symptoms of gastro-oesophageal reflux or for a repeat prescription of acid-suppressant medication, as opposed to an unexpected written invitation, and since we will now be able to provide information regarding the efficacy of this procedure from this trial, we predict that the uptake of the Cytosponge-TFF3 procedure will increase substantially compared with that observed in this trial. This hypothesis will require further evaluation in future studies or in clinical implementation research.

The safety of the Cytosponge-TFF3 device has been evaluated previously in a systematic review [69] of 2,672 procedures done across four different studies in the UK, the USA, and Australia. In this review [69], 2,334 (97%) of 2,418 patients swallowed the device successfully and there were two adverse events associated with the device; one was a detachment and one was a self-limiting pharyngeal bleed. These results are similar to those of our trial. The proactive telephone call to patients 7–14 days after they underwent the procedure also allowed us to collect data on side-effects. We found that 63 (4%) of 1,654 participants had a sore throat after the procedure, indicating that patients should be told that they might experience this adverse event after the procedure.

Implications of Cytosponge-TFF3 findings in the context of clinical guidelines

The prevalence of Barrett oesophagus or cancer in the 221 participants who received an endoscopy after testing positive for TFF3 was 59% (n=131). We also identified intestinal metaplasia of the gastro-oesophageal junction and gastric cardia, which was extensive throughout the stomach in some cases, in 33 (15%) of 221 patients. These findings were not included in the primary endpoint, as intestinal metaplasia without visible columnar epithelium is not Barrett oesophagus.

The guidelines for gastric intestinal metaplasia including the cardia were recently reviewed (2019), and UK and US societies suggest that, although the evidence is more scarce than it is for Barrett oesophagus, surveillance endoscopy should be considered when the gastric intestinal metaplasia is extensive or when there are other factors indicating an increased risk of gastric cancer, such as a family history [123, 124].

Overdiagnosis of cancer is a matter of much debate in the screening community, together with whether short segments (1 cm or less) of Barrett oesophagus should be considered as such. The TFF3 test is sensitive and detects some short segments of Barrett oesophagus.

Additionally, since this was a pragmatic trial that relied on a coded diagnosis of Barrett oesophagus, we also identified patients in the usual care group who had short segments of Barrett oesophagus (1 cm or less in length) and were diagnosed as having the condition, reflecting the variable practice in UK hospitals. We expect that these patients can be reassured and probably do not require surveillance. This expectation is consistent with the clinical guidelines, which suggest that patients with over 1 cm of salmon-coloured epithelium containing intestinal metaplasia should be monitored [57, 125].

With regard to the primary endpoint analysis, if we use a stringent criterion to diagnose the most clinically significant cases of Barrett oesophagus (i.e. those 3 cm or more in length; table 2.7), four (<1%) of 6,388 participants would be diagnosed with Barrett oesophagus in the usual care group and 46 (1%) of 6,834 participants would be diagnosed with Barrett oesophagus in the intervention group. This result would still show a positive effect of introducing the Cytosponge-TFF3 procedure into clinical care, with a RR of 10.6 (95% CI 6.0-18.8), after accounting for clustering.

Further guidance will be required to tailor the follow-up of patients diagnosed via the Cytosponge-TFF3 procedure, depending on their degree of risk of progressing to dysplasia or cancer according to the clinical surveillance guidelines.

Future biomarkers and target population of Cytosponge-TFF3 testing

In the future, we expect that additional biomarkers will distinguish indolent Barrett oesophagus from Barrett oesophagus at high risk of progression, so that many patients can be followed up with the Cytosponge-TFF3 procedure, and endoscopy can be reserved for those at a high risk who are likely to require intervention. Identification of risk stratification biomarkers is an ongoing area of research [126].

In this trial, patients were offered the Cytosponge-TFF3 procedure if they required medication for heartburn symptoms. In many health-care systems, a one-off endoscopy would be considered for these patients given that many require long-term medication (i.e., for 3 years or more). The sensitivity and specificity of the Cytosponge-TFF3 procedure have been evaluated previously [75], and our trial was not designed to re-evaluate these aspects. However, based on the number who had an endoscopy following a Cytosponge-TFF3 procedure but did not have Barrett oesophagus or cancer (n=90), and on the number who successfully swallowed the Cytosponge-TFF3 but did not have Barrett oesophagus or cancer (n=1,523), we estimated the specificity of the Cytosponge-TFF3 procedure to detect Barrett oesophagus, dysplasia or cancer to be 94%.

Setting of this trial and capturing of endpoint data

In the future, consideration should be given to the ideal enrichment criteria, which might include a different age cutoff for men compared with women because of the difference in incidence (ie. 87 [69%] of 127 Barrett oesophagus diagnoses in patients who successfully swallowed the Cytosponge were male), and also the inclusion of other risk factors, such as bodymass index.

Among the strengths of our trial is the real-world implementation of the Cytosponge-TFF3 procedure, including the administration of the device by a nurse in the community setting. The TFF3 test was done in a clinically certified laboratory, and the results were communicated in real time. The use of coding to ascertain diagnoses of Barrett oesophagus, dysplasia, and cancer ensured equity across both study groups. Since informed consent from individual patients was obtained only for those who underwent the Cytosponge-TFF3 procedure, the use of coding was the only way to ascertain the diagnoses for participants in the usual care group and those in the intervention group who declined the invitation to undergo the Cytosponge-TFF3 procedure.

Limitations of trial methodology

This trial has some limitations. First, those participants who agreed to undergo the Cytosponge-TFF3 procedure might have had more problematic symptoms than those who did not accept the offer of the procedure. We eliminated this bias by analysing the data of the whole trial as an intention-to-treat analysis. Second, 150 (9%) of 1,654 participants still had a low-confidence result after the offer of a repeat test. Work is ongoing to find out how to

reduce this outcome. Third, there were slightly more women than men agreeing to undergo the Cytosponge-TFF3 procedure, even though Barrett oesophagus is more prevalent in men than in women. In future, strategies to encourage men to attend the procedure, and whether to alter the threshold for testing men versus women, should be considered. Finally, variation in the quality of endoscopies was apparent across the 24 hospitals that took part in the study [127].

Massimiliano di Pietro did a central review of video images and liaised with hospitals to ensure consistency in reporting. Currently, the TFF3 test requires manual reading by a pathologist trained in analysing these specimens, which are much larger and more cytological in nature than endoscopic biopsies. In chapter 3 I am presenting an extensive framework to (semi-)automate quality control and diagnosis of Cytosponge-TFF3 samples and assist pathologists in screening of specimens.

Conclusion

For patients with heartburn-predominant symptoms requiring acid-suppressant therapy for at least 6 months, the Cytosponge-TFF3 procedure is a feasible, safe, and generally acceptable test to administer in the general practice clinic setting. This procedure results in improved detection of Barrett oesophagus, thus enabling a more proactive approach for the identification and minimally invasive treatment of dysplasia and early cancer. An economic evaluation will establish the effect of this strategy, taking into account the additional number of endoscopies required as a result of the Cytosponge-TFF3 procedure. In order to enable implementation of the technology at scale, a need for improved sample analysis workflows has been identified.

Author contributions

RCF and PS are responsible for the integrity of the data and the accuracy of the data analyses. RCF, PS, RM, JO, GR, SGS, FMW, BM, ID-B, and BA conceptualised and designed the study. MDP, MO'D, RM, BM, ID-B, BA, and MT acquired, analysed, and interpreted the data. I drafted the manuscript together with RCF, PS, RM, GR, BM, and JO. I statistically analysed the data together with RM and PS. RCF sought funding for the study and takes overall responsibility for the conduct of the trial and all the reported data.

Competing interests

RCF and MO'D are named on patents related to the Cytosponge-trefoil factor 3 test. Covidien GI Solutions (now Medtronic) licensed the Cytosponge from the Medical Research Council, and the device has now received the CE mark and is cleared by the US Food and Drug Administration. RCF, MO'D, and myself are shareholders in Cyted, a company working on early detection technology. MDP reports personal fees from Medtronic, outside of the submitted work. MT reports personal fees from Medtronic, outside of the submitted work. PS reports fees paid to his organisation from GRAIL, outside of the submitted work. BM is an employee of Cyted. The remaining authors have no conflicts of interest to declare.

2.5 Supplementary tables

	Usual care group (n = 3687)	Intervention group (n = 4152)	Absolute difference (95% CI)	Overall rate ratio (95% CI)	Overall adjusted rate ratio* (95% CI); p-value
Number of participants diagnosed with Barrett oesophagus	9 (0.2%)	92 (2.2%)†	-	-	-
Follow-up, person-years	4,006	4,421	-	-	-
Incidence of Barrett oesophagus, per 1000 person-years	2·2	21·2‡	18·9 (16·8-21·0)	9·4 (4·8-18·7)	10·0 (5·0-20·0);

Table 2.5 **Barrett oesophagus diagnoses in the usual care group compared with the intervention group, cluster-randomised group only.** Data are n (%), unless otherwise specified. *Overall adjusted rate ratio accounts for cluster-level randomisation †Number of participants diagnosed with Barrett oesophagus in the intervention group includes all participants who were offered the Cytosponge procedure. ‡The incidence of Barrett oesophagus in the intervention group was calculated as the weighted average of the incidence in the first 4 months of follow-up and the incidence in the following months, with a weight ratio of 1:2

	Usual care group (n = 2701)	Intervention group (n = 2682)	Absolute difference (95% CI)	Overall rate ratio (95% CI); p-value
Number of participants diagnosed with Barrett oesophagus	4 (0.1%)	48 (1.8%)†	-	-
Follow-up, person-years	2,573	2,531	-	-
Incidence of Barrett oesophagus, per 1000 person-years	1·6	18·6‡	17·1 (11·5-22·6)	12·0 (4·3-33·2); p <0·0001

Table 2.6 **Barrett oesophagus diagnoses in the usual care group compared with the intervention group, individual randomised group only.** Data are n (%), unless otherwise specified. †Number of participants diagnosed with Barrett oesophagus in the intervention group includes all participants who were offered the Cytosponge procedure. ‡The incidence of Barrett oesophagus in the intervention group was calculated as the weighted average of the incidence in the first 4 months of follow-up and the incidence in the following months, with a weight ratio of 1:2

Usual care group (n = 6388)	Intervention group		
	Underwent the Cytosponge procedure (n = 1750)	Did not undergo the Cytosponge procedure (n = 5084)	Overall (n = 6834)
<1cm	1 (8%)	2	0	2 (1%)
1 to <2 cm	3 (23%)	41	3	44 (31%)
2 to <3 cm	3 (23%)	21	4	25 (18%)
3 to <4 cm	1 (8%)	14	1	15 (11%)
4 to <5 cm	0	8	1	9 (6%)
5 to <6 cm	1 (8%)	9	1	10 (7%)
6 to <7 cm	1 (8%)	2	0	2 (1%)
7 to <8 cm	0	3	0	3 (2%)
8+ cm	1 (8%)	6	1	7 (5%)
missing	2 (15%)	21	2	23 (16%)
Total number of participants with Barrett oesophagus	**13 (100%)**	**127**	**13**	**140 (100%)**

Table 2.7 **Length of Barrett oesophagus in cm (Maximal length (M) from Prague CM Classification) across the study arms.** Data are n (%). Only coded Barrett oesophagus diagnoses, i.e. used for the intention-to-treat primary endpoint analysis, are shown.

Case by case	TNM stage	Overall stage	Treatment
Usual care group (n = 6388)	T3N0MX	Stage IIB	Robotic-assisted oesophagectomy
	T3N2M0	Stage IIIB	Palliative radiotherapy + stent / RIP
	T3N3M1	Stage IVB	Best supportive care + stent / RIP 1 month post diagnosis
Intervention group – Underwent the Cytosponge procedure (n = 1750)	LGD	Dysplasia	APC
	LGD-HGD	Dysplasia	RFA
	LGD-HGD	Dysplasia	RFA
	HGD	Dysplasia	EMR
	T1N0MX	Stage I	EMR
	T1bN0M0	Stage I	Oesophagectomy
	SM1 OAC	Stage I	EMR
	T1N0M0 (Gastric on background extensive IM)	Stage I	ESD
Intervention group – Did not undergo the Cytosponge procedure (n = 5084)	T1N0M0	Stage I	EMR, RFA and APC (Patient initially expressed interest in receiving the Cytosponge)
	T3N2M0	Stage IVA	Palliative chemotherapy
	T3N2M1b	Stage IVB	Best supportive care / RIP 3 months post diagnosis

Table 2.8 **Stage and treatment for dysplasia and cancer cases across all study arms.** RIP = Rest in peace, APC = Argon plasma coagulation, RFA = Radiofrequency ablation, EMR = Endoscopic mucosal resection, ESD = Endoscopic submucosal dissection.

Acceptability score*	**Participants who successfully swallowed the Cytosponge (n = 1654)**
0	1 (<0·1%)
1	2 (0·1%)
2	5 (0·3%)
3	13 (0·9%)
4	16 (1·1%)
5	92 (6·2%)
6	63 (4·3%)
7	103 (7·0%)
8	247 (16·9%)
9	317 (21·7%)
10	605 (41·3%)
Total number of patients filling in the questionnaire	**1464 (100·0%)**

Table 2.9 **Cytosponge-TFF3 acceptability scores.** Data are n (%). *11-point visual analogue scale: 0 = unacceptable, 10 = completely acceptable.

Chapter 3

Triage-driven diagnosis of Barrett Oesophagus using deep learning

Attribution

The text in this chapter was derived from the following publication for which I am the first author and main person responsible for the manuscript:
Triage-driven diagnosis for early detection of oesophageal cancer using deep learning
Authors: Gehrung M, Crispin-Ortuzar M, Berman AG, O'Donovan M, Fitzgerald RC, Markowetz F. *In revision at Nature Medicine / available on BioRxiv*

Personal contributions

The work in this chapter was carried out by me in its entirety. All code was written by me. Adam Berman has contributed to some aspects of the code development for the Grad-CAM saliency map visualisation.

Abstract

The aim of this chapter was the design, implementation, and validation of a deep learning method to reduce the pathologists' workload for the analysis of Cytosponge-TFF3 specimens. I particularly focused on investigating an approach that would not replace the pathologist but instead leverage deep learning models to stratify equivocal and unequivocal patient samples.

Deep learning methods have been shown to achieve excellent performance on diagnostic tasks, but it is still an open challenge how to optimally combine them with expert knowledge and existing clinical decision pathways. This question is particularly important for the early detection of cancer, where high volume workflows might potentially benefit substantially from automated analysis.

Here, I present a deep learning framework to analyse samples of the Cytosponge-TFF3 test, a minimally invasive alternative to endoscopy, for detecting Barrett oesophagus, the main precursor of oesophageal cancer. I trained and independently validated the framework on data from two clinical trials, analysing a combined total of 4,662 pathology slides from 2,331 patients. My approach exploits screening patterns of expert gastrointestinal pathologists and established decision pathways to define eight triage classes of varying priority for manual expert review. By substituting manual review with automated review in low-priority classes, I can reduce pathologist workload by 57% while matching the diagnostic performance of expert pathologists. These results lay the foundation for tailored, semi-automated decision support systems embedded in clinical workflows.

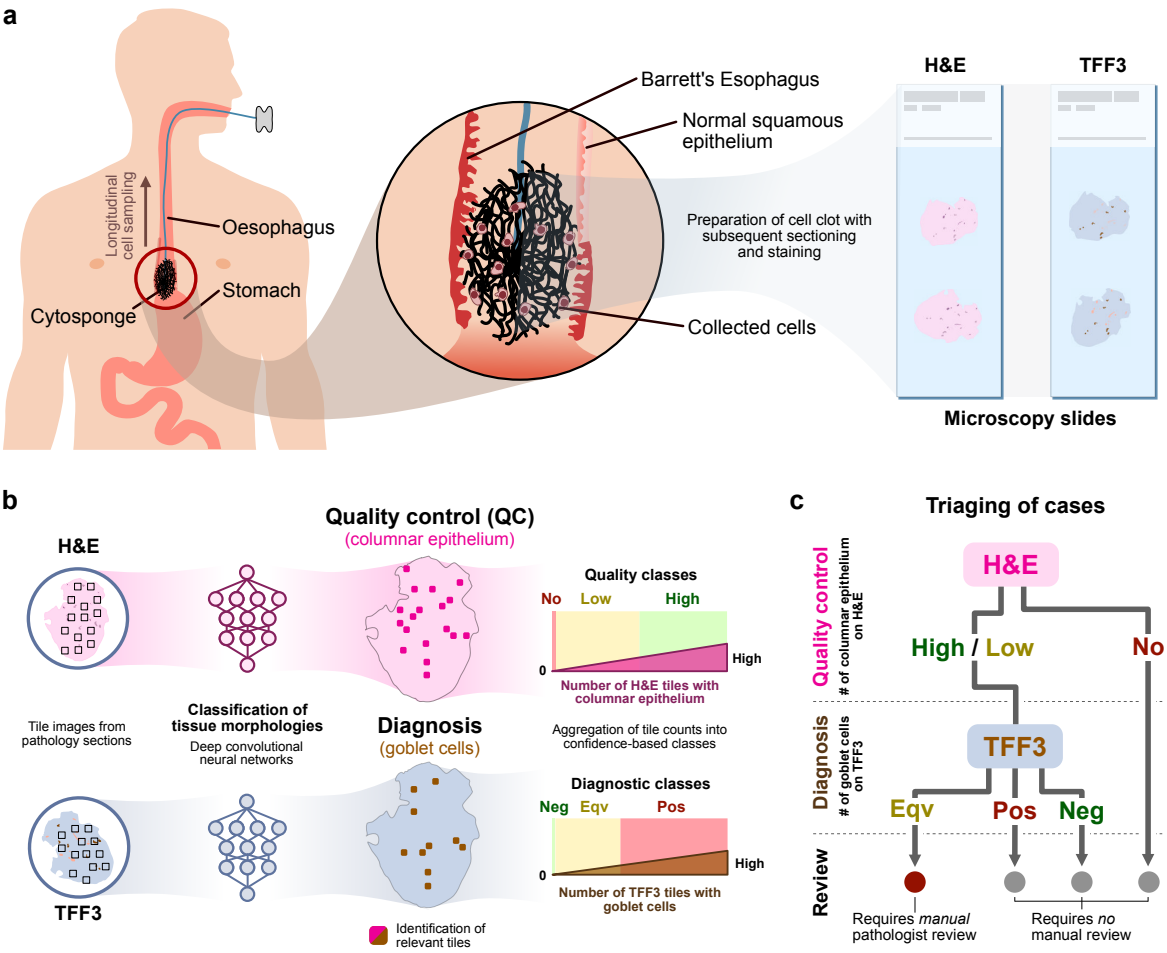
3.1 Introduction

Early detection of cancer often leads to better survival [128], because pre-malignant lesions and early stage tumours can be more effectively treated [17]. Most pre-malignant lesions amenable to early detection rely on targeted sampling and show only minor tissue changes on pathology assessment [22, 129, 130]. In addition, pathology procedures often involve laborious and time-consuming steps which can lead to errors and adversely affect patient care [131]. Recent developments in Artificial Intelligence (AI) have achieved excellent performance on diagnostic tasks [95, 91, 96]. However, understanding how these techniques can be integrated into clinical workflows most efficiently and to assess the actual benefits they bring remains a challenge. The design of a clinical decision support system needs to balance its performance against workload reduction and potential economic impact. Replacing

pathologists entirely could lead to substantial workload reduction, but such an approach would only be viable if performance remains comparable to that of human experts. Between a fully automated approach and the *status quo* of fully manual review lies a semi-automated approach, which uses computational methods to triage patients and only presents pathologists with difficult cases. A semi-automated approach will not reduce workload as much as a fully automated approach, but its performance benefits from existing expert knowledge and heuristics. Here I present such a semi-automated triage system using deep learning for the early detection of oesophageal cancer.

Oesophageal cancer is the sixth most common cause for cancer related deaths [1]. Patients usually present at an advanced stage with dysphagia and weight loss, and the 5-year overall survival of oesophageal adenocarcinoma (OAC), one of two pathological subtypes, is 13% [34]. OAC can arise from a precursor lesion called Barrett oesophagus (BE) [51, 35], providing an effective starting point for early detection. BE occurs in patients with GERD, a digestive disorder where acid and bile from the stomach return into the oesophagus leading to heartburn symptoms. In Western countries, 10 to 15% of the adult population are affected by GERD [59] and, therefore, at an increased risk of having BE. The pathognomonic feature of BE is intestinal metaplasia (IM), a process whereby the stratified squamous epithelial lining localized in the lower oesophagus is replaced with columnar epithelium containing goblet cells [125, 132]. The conventional diagnosis of BE requires an invasive endoscopic procedure of the upper gastrointestinal tract. However, there is no routine endoscopic screening of the GERD population and thus the vast majority of BE patients are undiagnosed [59].

Cytosponge-TFF3 is a non-endoscopic, minimally invasive diagnostic test for BE [74, 75, 78]. It is a cell collection device consisting of a compressed sponge on a string inside a gelatin capsule. The capsule is swallowed by the patient and the gelatin dissolves in the stomach after a few minutes, allowing the sponge to expand. The sponge is then withdrawn from the stomach by the attached string, sampling superficial epithelial cells from the top of the stomach, the oesophagus, and the oropharynx (fig. 3.1a). Therefore, the cellular composition of the sample is dominated by squamous cells, gastric columnar epithelium, and respiratory epithelium as well as any Barrett cells, if present. Following removal, the device is placed in a container with preservative solution and the sampled cells are processed, embedded in paraffin and stained with H&E as well as immunohistochemically stained with TFF3 [73]. H&E stains allow the identification and quantification of cellular phenotypes, which is critical for quality control. TFF3 is over-expressed in mucin-producing goblet cells which are a key feature of BE. TFF3 also functions as a protector of the mucosa from insults,



d

Dataset	Study	Patients	Controls (NE)	Cases (BE)	Slides
Training/Development	BEST2	100	51	49	200
Calibration	BEST2	187	83	104	374
Internal Validation	BEST2	525	262	263	1050
External Validation	BEST3	1519	1400	119	3038
Total:		2331	1796	535	4662

Fig. 3.1 **Cytosponge procedure with conceptual patient triage scheme and data summary. a** During withdrawal the sponge samples superficial epithelial cells from the top of the stomach and the oesophagus. These cells are processed into a cell block, then sectioned and stained with H&E and TFF3. **b** Convolutional neural networks, trained on an independent training dataset, are used for inference of H&E and TFF3 stains. The resulting tile maps are analysed for relevant regions (columnar epithelium on H&E and goblet cells on TFF3 stain) and aggregated into quality control and diagnostic classes based on tile detections. **c** Quality and diagnostic classes are mapped to a conceptualised pathway for sample stratification. The review layer (bottom) describes to what extent a human pathologist has to review the microscopy slides. (Pos = Positive, Neg = Negative) **d** Overview of data used in this study.

stabilizes the mucus layer, and promotes healing of the epithelium [70]. TFF3 stains allow the identification and quantification of goblet cells, which are indicative of IM. Therefore, TFF3 is the key diagnostic biomarker for BE [73].

The Cytosponge-TFF3 approach has profound and well-tested clinical significance. It offers, with substantial clinical trial data underpinning its efficacy, a long-awaited diagnostic alternative to endoscopy (BEST1 [74], BEST2 [75], BEST3 [104]). The BEST3 study, covered in chapter 2, found that the Cytosponge-TFF3 test had in excess of a 10-fold increase in detection of Barrett compared to usual clinical care in which patients with heartburn receive medication and an endoscopy if deemed necessary. This performance makes the Cytosponge a major advance in patient management. The BEST3 study also concluded that the pathology assessment is a major bottleneck for scaling the test to large patient populations. Since the analysis of Cytosponge-TFF3 pathology slides is a very laborious process due to the large amount of sampled cellular material. It comprises several time-consuming tasks such as assessing the amount of sampled material and checking the presence of gastric-type columnar epithelium to confirm that the capsule reached the stomach, followed by assessment for the presence of goblet cells indicative of BE. Though effective, the laboriousness of this process gives rise to a major opportunity for a clinical decision support system to improve analysis and scalability of the Cytosponge-TFF3 test.

Here, I use a deep learning approach for quality control and diagnosis of pathology slides for the Cytosponge-TFF3 test (fig. 3.1b). I propose a triage-driven approach, which retains diagnostic accuracy by leveraging the decision-making rules of expert gastrointestinal pathologists (fig. 3.1c). I train, calibrate, and internally validate my approach on data of the BEST2 multi-centre clinical trial [75] and externally validate it in an independent cohort from the recent BEST3 multi-centre trial [104] (fig. 3.1d). Additionally, I explore in a simulation study how well my results generalise to more general populations.

3.2 Methods

3.2.1 Study design and dataset

The multicentre Barrett Oesophagus Screening Trial 2 (BEST2) [75] case-control study (study registration: ISRCTN12730505) investigates the automated analysis of Cytosponge-TFF3 samples as a secondary objective. Ethics approval was obtained from the East of England - Cambridge Central Research Ethics Committee (number 10/H0308/71) and registered in

the UK Clinical Research Network Study Portfolio (9461). Patients enrolled underwent a Cytosponge procedure followed by an endoscopy with biopsies where required. The objective of this work was the comparison of: fully manual review of Cytosponge-TFF3 pathology slides by human experts, fully automated review of Cytosponge-TFF3 pathology slides by a deep learning-based method, and triage-driven, semi-automated review of Cytosponge-TFF3 pathology by a hybrid method relying on deep learning methods and human experts.

812 patients were randomly selected from the entire BEST2 cohort (from 11 hospitals in the UK) for digitisation of their respective H&E and TFF3 pathology slides (1624 in total) on an Aperio AT2 digital whole-slide scanner (Leica Biosystems Nussloch GmbH, Germany) at 40x magnification.

BEST2 patients were randomly partitioned into three distinct subsets: 100 patients for training/development (labels unblinded for training purposes), 187 patients for calibration (labels unblinded for calibration), and 525 patients as an internal validation set (labels unblinded after validation). The distribution of patients with or without Barrett oesophagus (BE) for each partition is shown in fig. 3.1d.

For independent external validation I used data from the Barrett Oesophagus Screening Trial 3 (BEST3) [104] randomised controlled trial (study registration: ISRCTN68382401). Ethics approval was obtained from the East of England - Cambridge Central Research Ethics Committee (number 16/EE/0546). Patients enrolled either were invited to a Cytosponge procedure or received standard of care. Both arms were followed up after 8 to 18 months (weighted overall average of approx. 12 months). Only patients who underwent a Cytosponge procedures or were referred as part of usual care received an endoscopy. A patient was considered as positive for Barrett Oesophagus if they either had a diagnosis at endoscopy or as a result of a coded search in records from the primary care site.

1,519 patients were randomly selected from the entire BEST3 cohort (from 109 primary care sites in the UK) for digitisation of their respective H&E and TFF3 pathology slides (1638 in total) on Hamamatsu S60 and S210 whole-slide scanners (Hamamatsu, Japan) at 40x magnification. For each patient, the repeat test was used if one as performed due to inadquace of the baseline test.

All BEST3 patients were processed using the fully automated and triage-driven, semi-automated approach presented in this work. Labels were unblinded after validation.

Confidence intervals in this work were defined as the 2.5th and 97.5th percentiles on distributions of 500 samples (with replacement) of the respective dataset size.

Cytosponge-TFF3 procedure

The Cytosponge-TFF3 technology has been introduced in chapter 1.

Endoscopy procedure

Esophago-gastroduodenoscopies were carried out by an endoscopist after the Cytosponge test. BE was defined as endoscopically visible columnar-lined oesophagus that measured at least 1 cm circumferentially or at least 3 cm in non-circumferential tongues according to the Prague criteria ($\geq$C1 or $\geq$M3 [133]). An additional criterion for BE was histopathological evidence of intestinal metaplasia (IM) on at least one endoscopy biopsy. For cases with suspected BE, diagnostic biopsies were collected following the recommended Seattle surveillance protocol [134]. When reviewing the biopsy data, all of the pathologists were blinded to the result of the Cytosponge-TFF3 test.

3.2.2 Annotation and pre-processing of whole-slide images

Whole-slide image annotation for training

One H&E- and one TFF3-stained slide for each of the 100 BEST2 patients from the training set were manually annotated and reviewed by an expert pathologist (Maria O'Donovan) using the ASAP software [135]. Regions of interest (ROIs) were selected in the digitised pathology slides at a magnification of 40x. Each of these ROIs was labeled with a class for training. For the H&E-based quality control model, four different classes were identified: gastric-type columnar epithelium, respiratory-type columnar epithelium, intestinal metaplasia, and background (including other cellular material such as squamous cells and slide artefacts). Gastric-type columnar epithelial cells were considered as the marker for quality control, as their presence confirms that the Cytosponge has reached the stomach. For the TFF3-based diagnostic model, three classes were identified: TFF3-positive regions (darkly stained goblet cells), TFF3-equivocal regions (regions of ambiguous staining that may be goblet cells), and background. TFF3-positive cells were considered as the marker for the presence of IM, as they indicate that the patient might have BE. All slides were annotated using the existing patient-level ground truth data for comparison. I aimed for a representative fraction of available material on each slide to be labelled.

Tesselation of whole-slide images for training

Tesselation, or tiling, of whole-slide images was performed in order to prepare data prior to model training. A custom tiling method was developed to optimise the yield and coverage of annotated cellular material in the images. Whereas packing problems of squares in polygons can be neglected for large annotations, optimal coverage for tiles in combination with small annotation sizes is not straightforward and requires a tailored solution. Annotations with an area of $1.5 * \textit{tile area}$ or larger were cropped into tiles by taking the top-left coordinate of the enveloping bounding box and iterating tiles along the x- and y-axis of the image. Tiles with an intersection of less than 0.33 (for H&E) or 0.66 (for TFF3) with their corresponding annotation were rejected. Annotations with an area smaller than $1.5 * \textit{tile area}$ were treated as single examples and a tile was placed in the center-of-mass of the respective annotation. Tiles with sufficient annotation coverage (determined by intersection) were extracted and labelled according to the class of their parent annotation. For this work, a tile size of 400-by-400 pixels (corresponding to 200-by-200 µm at a magnification of 40x) was selected in accordance with sizes of relevant tissue features. Tiles were extracted from whole-slide images as JPEG images with minimal compression.

Model training using deep learning

I implemented two different deep learning frameworks: one for performing quality control on H&E-stained slides, and a second one for performing automated BE diagnosis from the TFF3-stained slide images. Both deep learning frameworks for quality control and diagnosis were created by comparative transfer learning of multiple convolutional neural network architectures: AlexNet [136], DenseNet [137], Inception v3 [138], ResNet-18 [139], SqueezeNet [140], and VGG-16 [141]. All architectures were initialised with the best parameter set that was achieved on the ImageNet competition. Training tile images were resized as required for the individual architectures, resulting in a change of effective magnification from 22x to 30x. I then unfroze all layers to enable fine-tuning of the entire network. For all models, training continued on two NVIDIA GTX 1080Ti graphics cards for 25 epochs with an architecture-specific batch size (ResNet-18: 128, VGG-16: 48, Inception v3: 48, AlexNet: 64, SqueezeNet: 256, DenseNet: 84) and a learning rate that decayed by a factor of 0.1 every 7 epochs. All models used cross-entropy loss. To account for slight variations in the training data, random vertical/horizontal flip, random rotation, and random color jitter (variation in hue, contrast, brightness, and saturation) were introduced for data augmentation. Differences

in tile class sizes were accounted for by using a modified imbalanced dataset sampler, a function which oversamples from minority classes and undersamples from majority classes. The parameter set of epoch with the highest accuracy on the development subset was selected for further use. All models were trained using the PyTorch deep learning framework [142]. Final model versions used a split of 85:15 patients for training and development subset. I further investigated the effect of increased training set sizes by incrementally increasing the training subset while fixing the development subset size (fig. 3.2).

3.2.3 Evaluation and visualisation of tile-level models

Evaluation of tile-level performance

In order to compare the performance of all six deep learning architectures, I calculated class-specific performance in the quality control and diagnosis frameworks (table 3.1). To obtain these numbers, I selected the epochs with the best weighted accuracy score on the development subset for each training run. I then calculated precision and recall of all four classes in the H&E-based model and all three classes in the TFF3-based model in the selected epoch. For visual comparison, I also created 2D inference maps of samples which where classified as positive or negative by a pathologist for quality control and diagnosis, respectively. Tile-level results were not used to select architectures for the fully automated or semi-automated, triage-driven approach. The best performing architectures according to relevant class precision and recall on tile level for quality control and diagnosis were selected for saliency map generation.

Generation of saliency maps using Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) class localisation maps are created by visualising the gradients flowing into the final convolutional layer of the network, just before the fully-connected layers [143]. Since convolutional layers contain class-specific spatial information from the input image which is lost in the fully connected layers, this is the optimal point for map generation. Unlike conventional class-activation maps (CAMs), Grad-CAM has the benefit of not requiring any modifications to the existing model architecture, nor does it require any retraining of the model [143]. In order to create the class-specific Grad-CAM localisation map for class c, $L^c_{\text{Grad-CAM}}$, it is first necessary to compute the gradient $\frac{\partial y^c}{\partial A^k}$ of the score y^c for class c with respect to the feature map A^k of the final convolutional layer [143]. Once $\frac{\partial y^c}{\partial A^k}$ has been computed for each feature map k, these backward-flowing

gradients are global-average-pooled across the width and height of the network (indexed by i and j) to yield α_k^c, the weights of neuron importance for each of the feature maps k [143]:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

α_k^c, the neuron-importance weights for each feature map k, therefore estimate the salience of each feature map to the prediction of class c [143]. Note that Z corresponds to the number of pixels in the respective feature map. Finally, to get class c-specific Grad-CAM localisation map $L^c_{\text{Grad-CAM}}$, I take the $ReLU$ of the weighted sum of the feature maps A^k, where each feature map k's weight is α_k^c [143]:

$$L^c_{\text{Grad-CAM}} = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{3.1}$$

Note that the $ReLU$ operation is used to retain only the features which have a positive influence on the prediction of class c, and that the resulting localisation map will be the same size as the feature maps of the last convolutional layer [143].

I generated saliency maps for both models trained on H&E and TFF3, respectively. The target layer from the VGG-16 architecture was the last feature layer (no. 30) before several stacked fully connected layers. Tiles were randomly selected from the development subset. For qualitative comparison between saliency maps and manual landmarks, I asked one expert pathologist (Maria O'Donovan) to highlight important areas. Areas highlighted by the pathologist provide a representation of features which a human observer uses for classification of tile images. To investigate qualitative agreement of landmarks by the pathologist with generated saliency maps, a side-by-side comparison of tile images and respective saliency maps was prepared (fig. 3.4).

3.2.4 Calibration and evaluation of fully-automated, patient-level models for BE detection

Model inference on calibration and validation cohort

All six deep learning architectures trained separately for quality control and diagnosis tasks were applied to pathology slides randomised to calibration and validation cohort. Whole-slide images were tesselated on the fly as described above. Detection of tissue was achieved by luminance thresholding of tile values in the LAB colour space. Tiles were forward-passed

through the trained deep learning architectures and softmax probabilities were aggregated for each tile position.

Aggregation of classifications on tile level to the patient level

I explored two different aggregation approaches based on propagation of the individual tile-level classifications to patient-level classifications for quality control and diagnosis: a fully automated approach which operates on the basis of a single operating point, and a semi-automated, triage-driven approach which leverages two operating points. For the former approach, performance was assessed using sensitivity and specificity; for the latter, performance was assessed using an incremental substitution scheme with simultaneous analysis of sensitivity and specificity. For both approaches, tile-level probabilities had to be thresholded to obtain the number of positive tiles per slide for quality control and diagnosis. In the following section, I describe how tile-level probabilities were thresholded and how the operating points on the resulting numbers of positive tiles (quality control and diagnosis) were then calibrated and evaluated as part of each approach.

Determination of tile-level probability thresholds

In order to generalise the tile-level probabilities to the number of positive tiles per patient, I determined thresholds for each model and endpoint (quality control and diagnosis). The probability threshold of individual tiles for quality control and diagnosis had to be determined, then, the resulting number of positive tiles per threshold was assessed against the best ROC-AUC on the calibration cohort (fig. 3.6, table 3.2).

To achieve the best-performing threshold for individual tile probabilities and subsequent aggregation, I iterated over a range of tile thresholds on a fine grid from 0 to 1 (in 0.005 steps and inclusive of 0.999, 0.9999, and 0.99999). For the quality control model on H&E, the relevant class was gastric-type columnar epithelium. For the diagnosis model on TFF3, the relevant class was TFF3-positive goblet cells.

In order to determine the resulting number of positive tiles per threshold, probability thresholds for quality control were compared (ROC-AUC) to the pathologist ground truth of H&E slide analysis. Probability thresholds for diagnosis were compared (ROC-AUC) to endoscopy (confirmation of BE presence by endoscopist and IM on endoscopy biopsy by pathologist) ground truth. This step was required to determine the optimal threshold for

individual tile classification. This threshold was then used in the calibration and validation of the fully automated and semi-automated, triage-driven model as described in the next section.

Calibration of fully automated model

All six deep learning architectures trained for quality control and diagnosis were applied to the whole-slide images from the calibration cohort (see Model inference). The number of positive tiles per sample for quality control and diagnosis was determined as described above. To determine an adequate operating point for the fully automated patient-level model, ROC analysis was performed on the number of detected tiles (quality control and diagnosis) per patient. On the same set of patients, I calculated the performance by an expert pathologist. In order to determine the ideal cut-off for number of detected tiles, I fixed the specificity of each model to the performance of an expert pathologist on the calibration cohort. The resulting operating point was then chosen for validation of the fully automated model in the validation cohort. Tile-level thresholds which yielded the best sensitivity on the calibration cohort were used for evaluating all approaches on the validation cohort. The best-performing architecture (assessed by sensitivity) on the calibration cohort was considered the representative model for application on the validation cohort. However, due to the simplicity of operating point determination, the performance of all other architectures on the validation cohort was also investigated.

Evaluation of fully automated model using ROC analysis

All six deep learning architectures trained for quality control and diagnosis were applied to the whole-slide images from the validation cohort (see Model inference). The number of positive tiles per sample for quality control and diagnosis was determined as described above. Subsequently, the previously determined operating point (calibration) for each of the deep learning architectures was applied. The binary results were then compared against ground truth of the quality control and diagnosis models. For quality control on H&E, the results were compared to the ground truth of the pathologist who was reading the H&E slide of the Cytosponge test. For diagnosis on TFF3, the results were compared with endoscopy ground truth (with confirmation of BE presence by endoscopist and IM on endoscopy biopsy by pathologist). Sensitivities and specificities on the validation cohort were calculated for all models with an additional presentation of ROCs for visualisation (table 3.3, fig. 3.7). For

comparison with other approaches, performance metrics of the architecture selected during calibration of the fully automated model were used.

3.2.5 Calibration and evaluation of semi-automated, patient-level models for BE detection

Calibration of triage-driven, semi-automated model

All six deep learning architectures trained for quality control and diagnosis were applied to the whole-slide images from the calibration cohort (see Model inference). For calibration, only the best model (according to ROC-AUC) was presented to two expert observers to determine operating points. The number of positive tiles per sample for quality control and diagnosis was determined as described above (fig. 3.9). The objective of this approach was a more granular classification of patients into three classes for quality control and diagnosis and subsequent stratification by different class combinations. Therefore, two operating points were determined for each model, instead of one.

Both observers were presented with the number of detected tiles and relevant ground truth (Cytosponge pathology and endoscopy) for quality control and diagnosis models. They were instructed to choose two operating points for each task: First, an operating point which optimises sensitivity with a low number of false positives. Second, an operating point which separates the intermediate region of the first and second operating point from samples with optimised specificity and a low number of false negatives. The resulting operating points were then chosen for validation of the semi-automated, triage-driven model in the validation cohort (fig. 3.8).

The two operating points for quality control and diagnosis resulted in three tiers per framework and were labelled as follows: for quality control, samples above the first operating point were to be considered as high confidence, samples between the first and second operating point as low confidence, and samples below the second operating point as no confidence. For diagnosis, samples above the first operating point were to be considered as high confidence positive, samples between the first and second operating points as low confidence equivocal, and samples below the second operating point as high confidence negative. Eight triage classes (number 1 to 8) were composed by all possible combinations of quality control and diagnosis classes. The combination (no confidence in quality and high confidence in diagnosis) is likely artifactual and was therefore merged (with no confidence in quality and equivocal in diagnosis) to form triage class 4. Two expert observers then

ranked all eight classes from lowest to highest likelihood for patients having BE. They further assigned a qualitative rank for priority of manual review based on the subjective difficulty to review samples that are part of specific triage classes.

Evaluation of triage-driven model on internal validation cohort

The triage-driven, semi-automated model was evaluated by applying a cumulative substitution scheme on the internal validation cohort. The base scenario for all cumulative substitutions was the performance of the pathologists on the entire validation cohort. At every substitution, the pathologists' Cytosponge-TFF3 results were substituted with automated review in the respective triage classes. Then, sensitivity, specificity, and proportion of patients substituted with automated review were calculated and compared against the previous substitution steps. The substitution scheme was applied starting from both ends of the triage class list. First, class 1 was substituted with automated review, then classes 1 and 2, then classes 1, 2, and 3, and so on. Second, class 8 was substituted with automated review, then classes 8 and 7, then classes 8, 7, and 6, and so on. I then analysed the sensitivity and specificity curves for deviations from their previous values for each step in both applications of the scheme. Classes which caused a drop in sensitivity or specificity on substitution were considered as 'difficult' and I retained manual review by a pathologist for associated samples. For each of the difficult classes I then summed up the number of patients that fell into these classes and divided by the total number of patient in the validation cohort. This ratio was to be considered as the potential workload reduction which this substitution scheme could achieve without notable loss in performance.

Simulation of cohort variation and impact on workload reduction

In order to assess workload reduction in cohorts with different compositions, I simulated the distribution of patients within triage classes with varying BE prevalences and sample confidences. Let P be a set of all patients with two subsets: $Q \subseteq P$ contains all patients with BE and its complement $R = P \setminus Q$ contains all patients without BE. I count the proportions of patients in each triage class in each of the sets P, Q, R as vectors $\mathbf{c}^P$, $\mathbf{c}^Q$ and $\mathbf{c}^R$, respectively. My simulation consists in re-weighting these vectors to reflect different BE prevalences and sample confidences. For each element of a range of BE prevalences ($\mathbf{s}_{\text{prev}} = \{0.00, 0.01, ..., 0.55\}$) I multiply $\mathbf{c}^Q$ by $s \in \mathbf{s}_{\text{prev}}$ and $\mathbf{c}_R$ by $1 - s$. At the same time, for each element of a range of relative sample confidences ($\mathbf{t}_{\text{conf}} = \{-0.25, -0.24, ..., 0.25\}$)

I shift proportions of $\mathbf{c}^P$ between triage classes $\{1,3,4,5,6,7\}$ and $\{2,8\}$ by adding $t \in \mathbf{t}_{\text{conf}}$ to one set of classes and subtracting it from the other. Reduction of workload (W) at every simulation step was defined as $\mathbf{c}^P$ for classes 4, 5, and 6 over classes 1, 2, 3, 7, and 8:

$$W = \frac{\mathbf{c}_4^P + \mathbf{c}_5^P + \mathbf{c}_6^P}{\mathbf{c}_1^P + \mathbf{c}_2^P + \mathbf{c}_3^P + \mathbf{c}_7^P + \mathbf{c}_8^P}$$

Evaluation of triage-driven model on external validation cohort

The triage-driven, semi-automated model was further evaluated applying it with frozen model parameters on the external validation cohort. Processing of images was performed as described on the internal validation cohort above. The trial from the data originates was investigating real-world implementation of the Cytosponge device technology. Therefore, endoscopy data endoscopy data was only available for positive Cytosponge patients and those who had Barrett diagnosed at follow-up as a result of standard of care. This resulted in a difference of available data as the study was designed for PPV instead of sensitivity and specificity. The NPV was also calculated by using aggregated findings from the primary trial endpoint. An analysis according to the presented substitution scheme was additionally performed (fig. 3.12)

3.3 Results

3.3.1 Deep learning models achieve high performance for tile-level classifications

The first step of my approach is based on the tile-level detection of different classes of cells relevant for quality control and diagnosis of BE. For model development and internal validation, I used 812 Cytosponge-TFF3 patient samples with paired pathology and endoscopy data from the BEST2 clinical case-control study [75]. Samples were randomly divided into training/development (n=100), calibration (n=187) and internal validation (n=525) sets (fig. 3.1d). An additional independent dataset (n=1,519) from the BEST3 study was used for external validation of the developed approach.

Training sets of larger size did not improve tile-level accuracy (fig. 3.2). Training, calibration, and validation sets were kept separate. Endoscopic as well as Cytosponge pathology diagnoses were only unblinded after tile-wise tissue classification models were

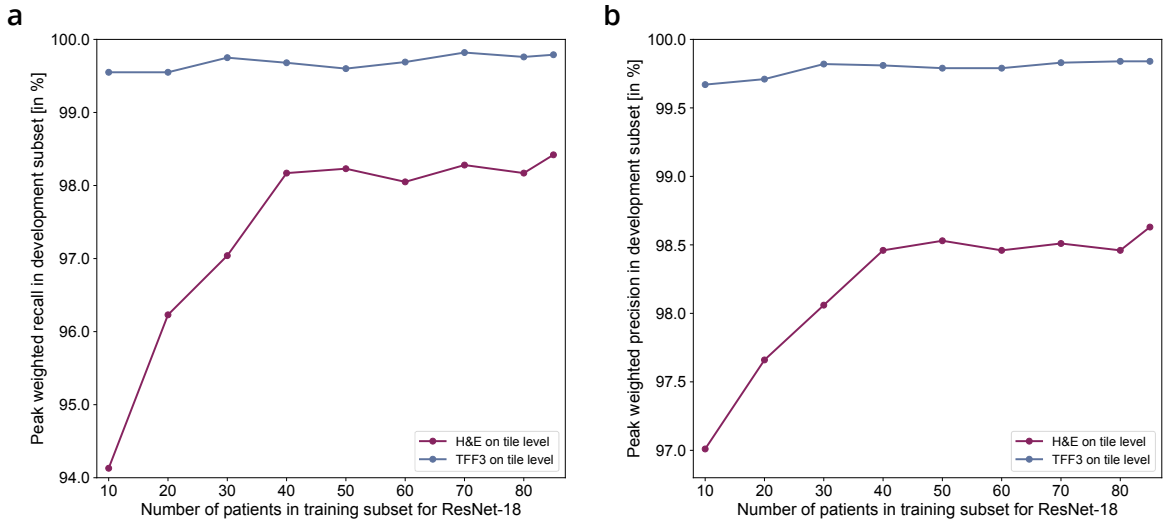


Fig. 3.2 **Differential increase of training partition size for ResNet-18.** Training subset refers to the relative proportion of the training partition used in the model training phase. Development subset refers to the relative proportion of the training partition used in the model development phase. The peak development weighted recall (a) and precision (b) correspond to the best performing cohort for each training run. The size of the development set was fixed at 15 patients. For each patient, an average of 3,500 tiles was used. For both H&E and TFF3 no substantial increase in performance metrics could be observed after a training subset size of 50 patients. H&E benefited more from an increased number of patients than the TFF3 model. This difference is associated with the increased complexity of detecting different tissue morphologies on H&E vs. brown goblet cells on TFF3.

calibrated and validated, respectively. All training slides were tessellated prior to training: For H&E I derived 193,734 tiles from 100 slides and for TFF3 I derived 235,932 tiles from 100 slides (based on the size of annotated areas, see Methods). All tiles were 200-by-200 µm and all labels were taken from expert slide annotations.

For both quality control (H&E) and diagnostic (TFF3) tasks, I trained several state-of-the-art networks (AlexNet [136], DenseNet [137], Inception v3 [138], ResNet-18 [139], SqueezeNet [140], and VGG-16 [141]) and evaluated their performance on the development datatset. Using individual tiles, I compared tile-level precision and recall for classifying columnar epithelium using the presence of gastric-type cells (on H&E) and positive goblet cells (on TFF3) (table 3.1, description in Methods): For gastric-type columnar epithelium, VGG-16, DenseNet and Inception v3 achieved the highest recalls (0.950, 0.947, 0.940, respectively) with consistent precisions (0.843, 0.865, 0.857). For goblet cells, VGG-16, Inception v3, and ResNet-18 achieved the highest recalls (0.919, 0.919, 0.912) with consistent precisions (0.856, 0.856, 0.827). Examples for whole slide images classified positive and negative for quality control and diagnosis are shown in section 3.3.1a.

3.3.2 Saliency maps agree with pathologist criteria for classification of tissue tiles

To understand which characteristics of the tile images were relevant to my models' classifications, I generated saliency maps using Gradient-weighted Class Activation Mapping (Grad-CAM) [143]. These maps highlight the local regions of an image most relevant to a model's identification of a particular class. I generated saliency maps for classes in one H&E-based model (VGG-16) and one TFF3-based model (VGG-16) (section 3.3.1b). For the gastric-type columnar epithelium class of the H&E-based model, the saliency maps highlight gastric cells by both the linear organisation of their nuclei as well as the presence of a straight border between the cells and the lumen. For the positive class of the TFF3-based model, I found that the saliency maps highlighted the mucin-containing goblet cells that characterise IM with high precision. In addition to the three representative examples in section 3.3.1b, I compared landmarks selected by an expert pathologist with tile images and respective saliency maps (fig. 3.4). The saliency maps confirm that the models learned features are similar to those used by pathologists to identify different tissue classes.

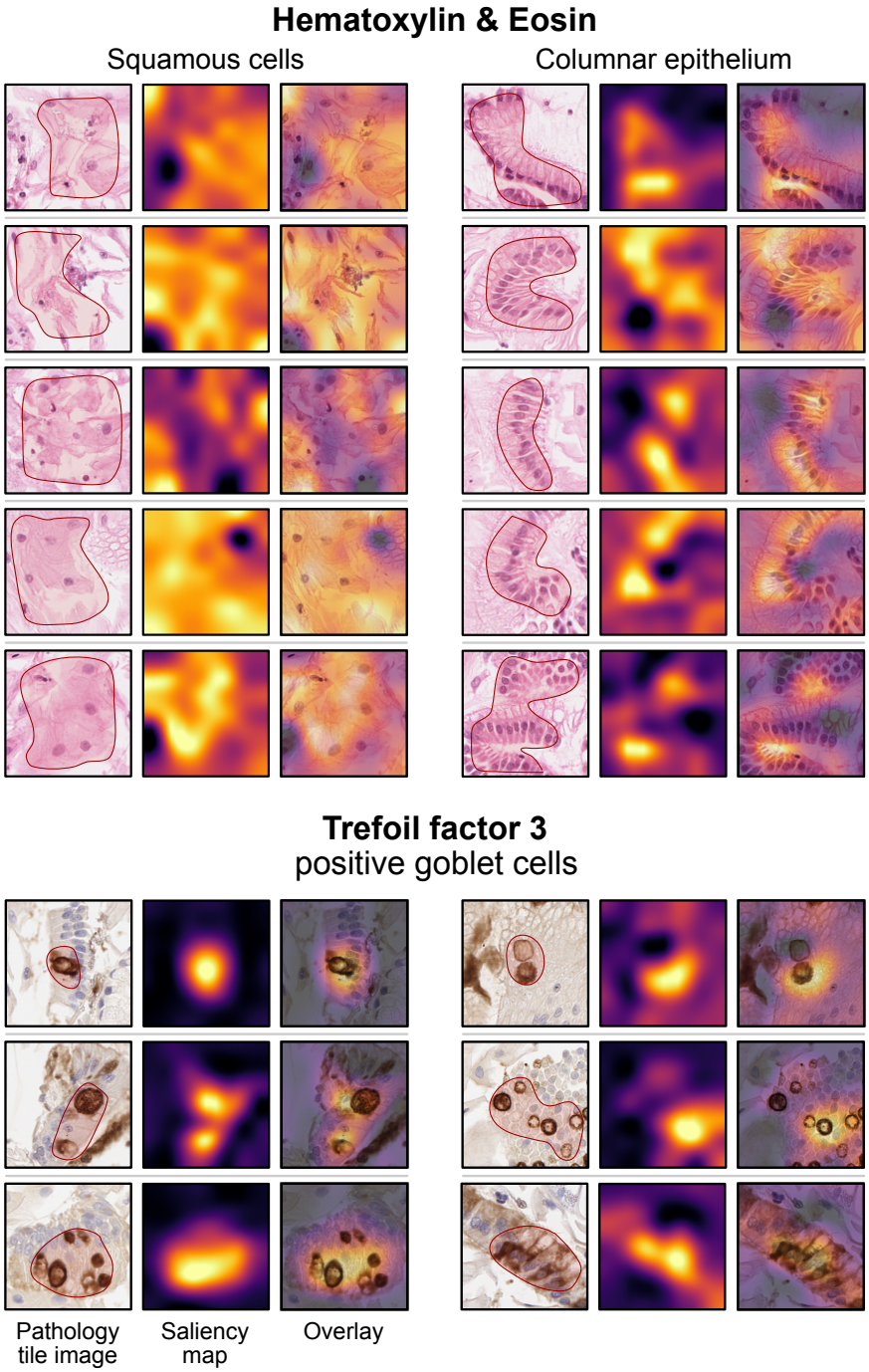


Fig. 3.4 **Comparison of pathologist landmarks with saliency maps extracted from VGG-16 architectures.** Additional examples of saliency maps for Hematoxylin & Eosin stain (squamous cells and columnar epithelium) and Trefoil factor 3 (positive goblet cells). Landmarks selected by an expert pathologist are shown as overlays with red borders on pathology tile images. For all classes, there was visual agreement between highlighted areas by the pathologist and saliency map activations.

3.3.3 Fully automated approach shows suboptimal performance

Tile-level classifications were aggregated into patient-level classifications using tile counts above thresholds determined by the specificity of expert pathologists on the calibration cohort (Methods, table 3.2, fig. 3.5).

I then performed ROC analysis with matched Cytosponge pathology and endoscopy ground truth on the internal validation cohort (section 3.3.1c-e).

First, the patient-level scores were compared against the binary Cytosponge-TFF3 ground truth by the pathologist on the internal validation set. For quality control, VGG-16 ranked highest for detecting columnar epithelium in H&E stains (ROC-AUC: 0.99 (CI 95%: 0.98 - 0.99)). SqueezeNet, the least complex architecture I trained, ranked lowest (ROC-AUC: 0.97 (CI 95%: 0.95 - 0.98), section 3.3.1c). For diagnosis, VGG-16 ranked highest for detecting goblet cells in TFF3 stains (ROC-AUC: 0.97 (CI 95%: 0.96 - 0.99), section 3.3.1d). Again, SqueezeNet ranked lowest (ROC-AUC: 0.94 (CI 95%: 0.92 - 0.96)). Confidence intervals were derived by bootstrapping (Methods). Results for all architectures are presented in table 3.3, and fig. 3.7a/b.

In summary, for both quality control and diagnosis in comparison to Cytosponge-TFF3 pathology ground truth, VGG-16 provided the highest performance, and SqueezeNet the lowest.

Next, patient-level counts were compared to endoscopy ground truth for detecting BE on the internal validation set (Methods). This ground truth was defined according to the Prague criteria (Methods) with confirmed IM on endoscopy biopsies [57]. To calculate sensitivity and specificity for the fully automated method on the internal validation cohort, I used operating points determined on the calibration cohort (table 3.2). VGG-16 ranked highest for detecting patients with BE from TFF3 stains (ROC-AUC: 0.88 (CI 95%: 0.85 - 0.91), Sensitivity: 72.62% (CI: 67.42% - 78.21%), Specificity: (93.13% (CI: 90.04% - 96.13%)), section 3.3.1e). SqueezeNet ranked lowest for detecting patients with BE from TFF3 stains (ROC-AUC: 0.85 (CI 95%: 0.81 - 0.88), Sensitivity: 69.58% (CI: 63.92% - 75.52%), Specificity: 92.37% (88.47% - 95.52%), section 3.3.1e). For comparison, the pathologists achieve a sensitivity of 81.7% (CI 95%: 77.4% - 86.5%) and a specificity of 92.7% (CI 95%: 89.6% - 95.6%). Performances of all architectures are presented in table 3.3, and fig. 3.7c. In summary, results for the fully automated approach on the internal validation cohort showed a loss of sensitivity of 9.1% for BE detection when compared to an expert pathologist.

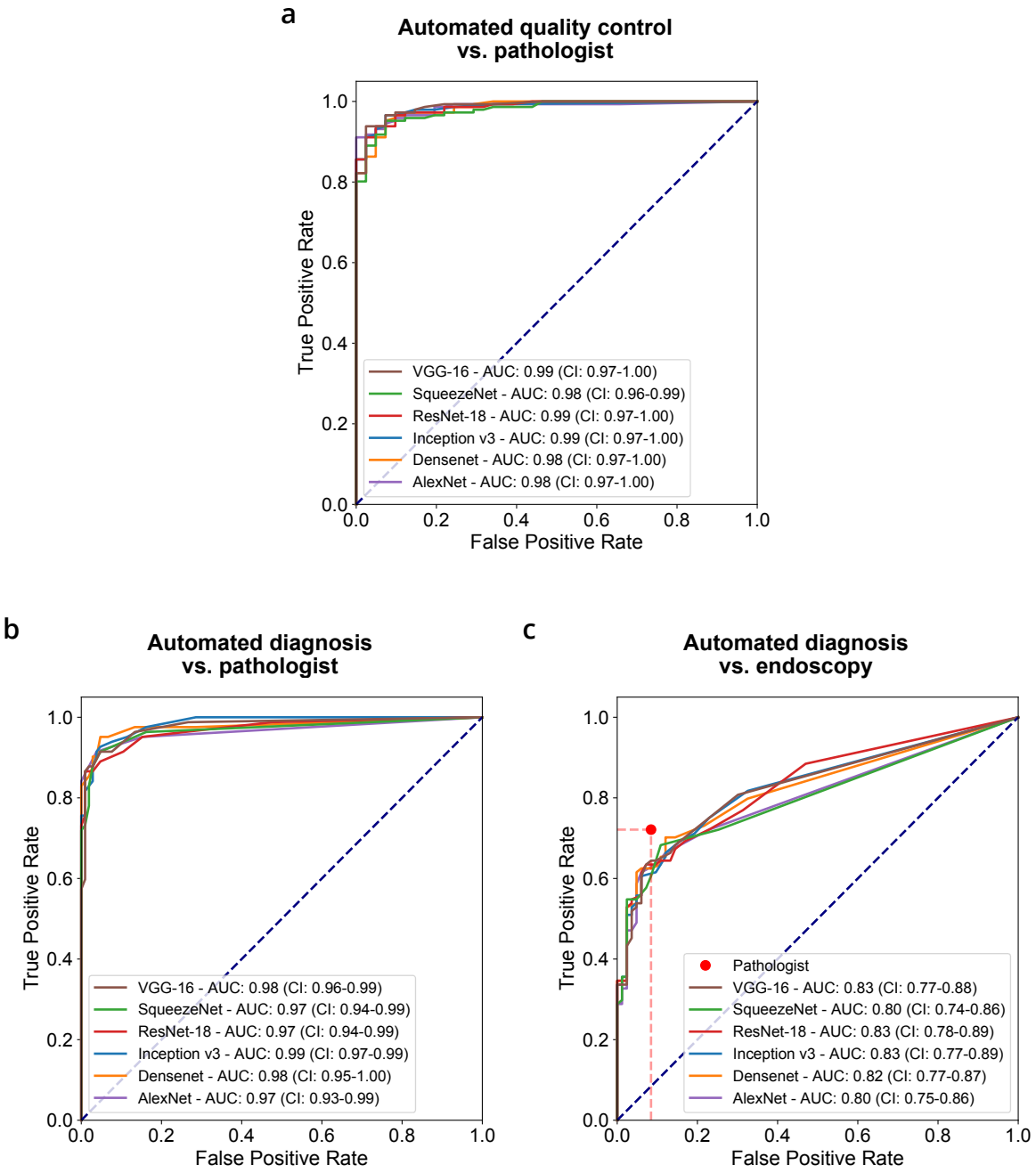


Fig. 3.5 **Performance of all deep learning architectures on the calibration cohort.** (a) ROC analysis of number of tiles containing columnnar epithelium on H&E compared with pathologist ground truth from Cytosponge (b) ROC analysis of number of tiles containing positive goblet cells on TFF3 compared with pathologist ground truth from Cytosponge (c) ROC analysis of number of tiles containing positive goblet cells on TFF3 compared with endoscopy (with confirmed IM) ground truth. A weak AUC dependency on architecture complexity can be observed.

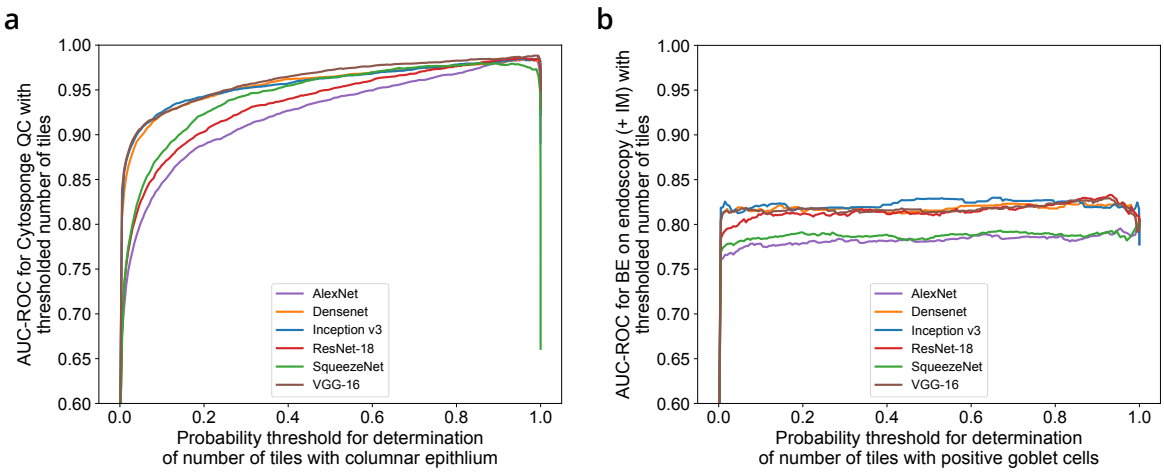


Fig. 3.6 **Determination of probability thresholds in order to obtain number of tiles.** Both plots show the AUC-ROC for individual probability thresholds (after softmax) which are used to decide whether a tile falls into the relevant class. (a) AUC-ROC for quality control (QC) ground truth determined by the pathologist compared with number of tiles containing columnar epithelium at individual probability thresholds. (b) AUC-ROC for diagnosis ground truth determined by the endoscopy (with confirmed IM on pathology) compared with number of tiles containing positive goblet cells at individual probability thresholds.

3.3.4 Triage-driven approach selects patients for manual review

I then explored whether a different modelling approach based on established decision pathways could boost performance. I developed a triage-driven, semi-automated approach as an alternative to the fully automated approach described above. Both approaches use the same patient-level aggregations as input, but their outputs are different: the fully automated approach tries to directly mimic pathology assessment by classifying patients as positive or negative for BE. In contrast, the triage approach defines different quality and diagnostic confidence classes to select challenging patient samples for manual review. Although it cannot reduce workload as much as a fully automated approach, a triage approach keeps sample stratification more interpretable and transparent.

I first selected deep learning architectures and defined cut-offs for different quality and diagnostic confidence classes based on thresholds determined by two expert observers on the calibration cohort (fig. 3.9, Methods).

For quality confidence classes, pathologists conclude that the sponge reached the stomach if they observe columnar epithelial groups [73, 75]. I encoded these subjective metrics in a quantitative scheme where the number of tiles detected with gastric-type columnar epithelium on H&E were classified as no confidence, low confidence, or high confidence (fig. 3.9a,

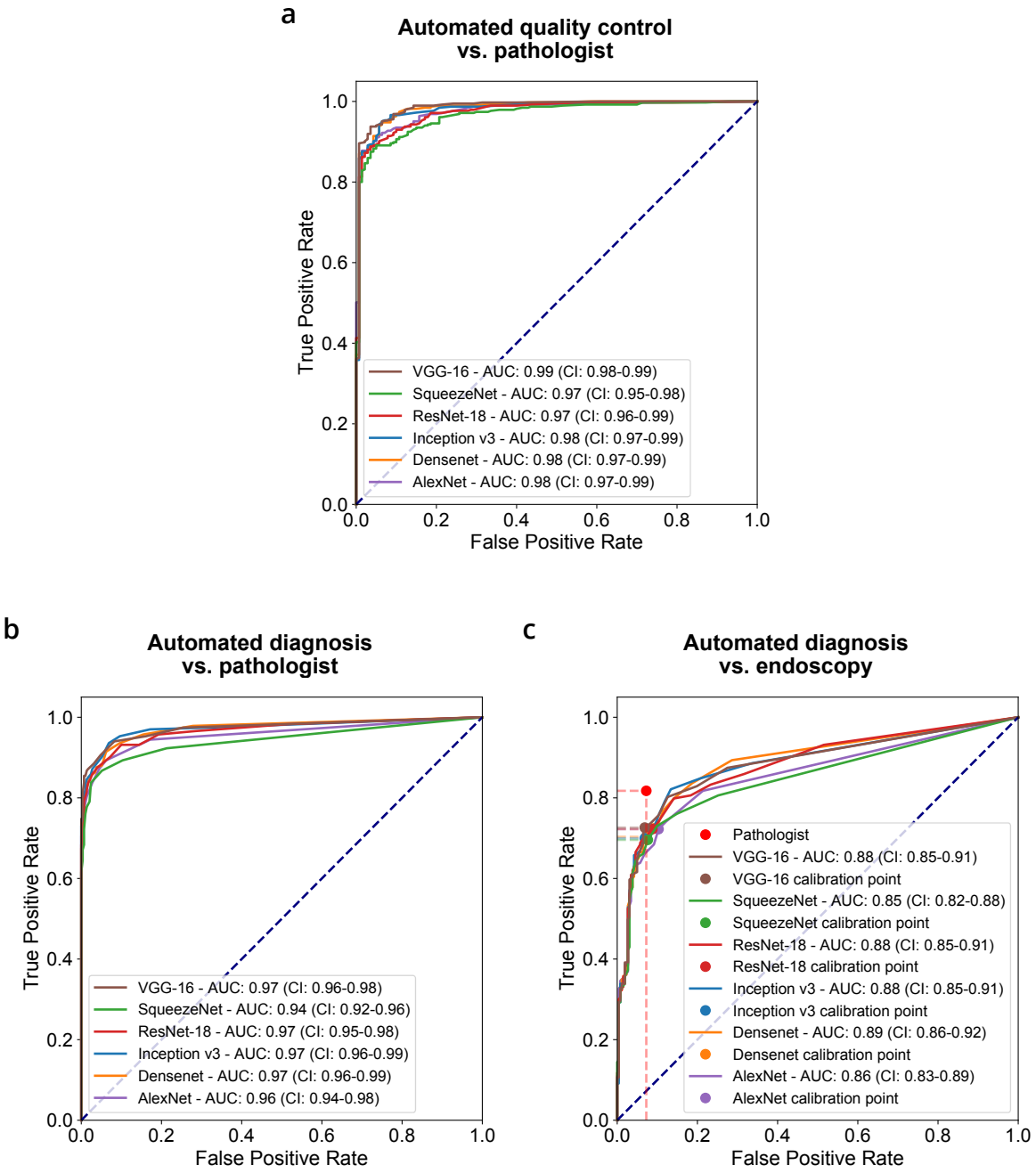


Fig. 3.7 **Performance of all deep learning architectures on the validation cohort.** (a) ROC analysis of number of tiles containing columnnar epithelium on H&E compared with pathologist ground truth from Cytosponge (b) ROC analysis of number of tiles containing positive goblet cells on TFF3 compared with pathologist ground truth from Cytosponge (c) ROC analysis of number of tiles containing positive goblet cells on TFF3 compared with endoscopy (with confirmed IM) ground truth. As in the calibration cohort, a weak AUC dependency on architecture complexity can be observed.

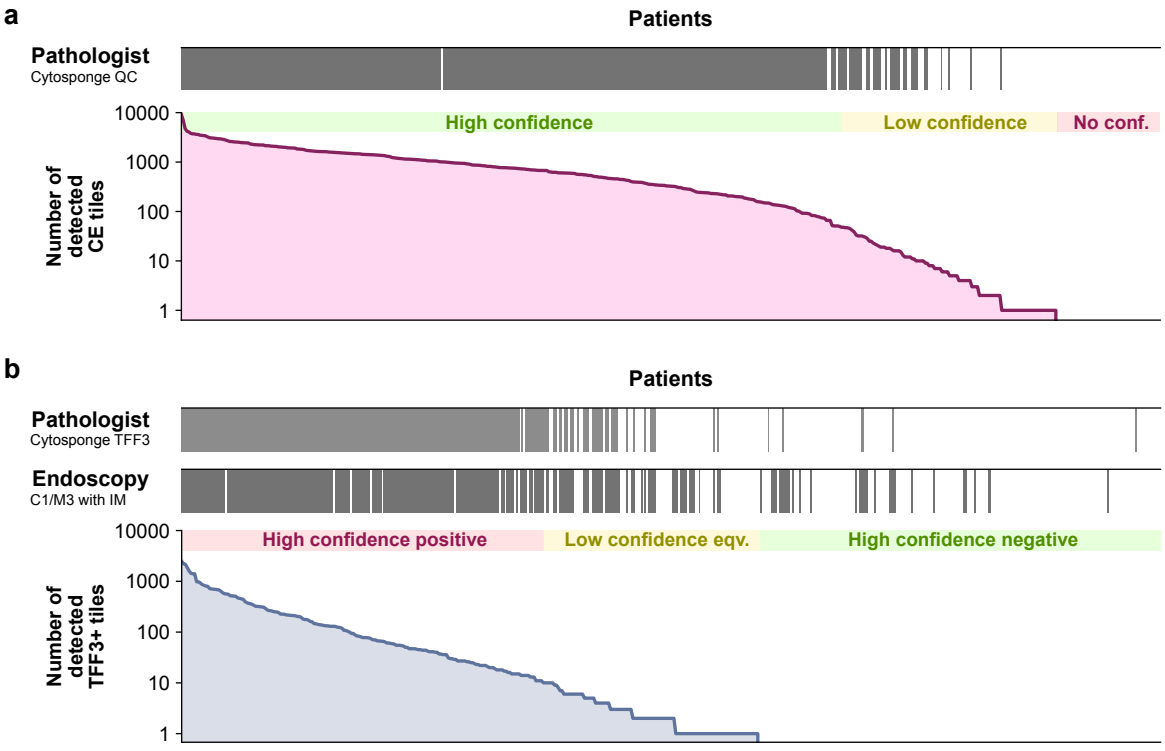


Fig. 3.8 **Application of quality control and diagnostic confidence class scheme to internal validation cohort.** **a** Quality ground truth by pathologist from Cytosponge (top) compared with number of detected columnar epithelium (CE) tiles on H&E detected by VGG-16 (bottom). **b** Diagnosis ground truth by pathologist from Cytosponge (top), Endoscopy (with confirmed IM on biopsy) ground truth (middle) compared with number of detected TFF3-positive tiles on TFF3 detected by ResNet-18 (bottom) / eqv. = equivocal.

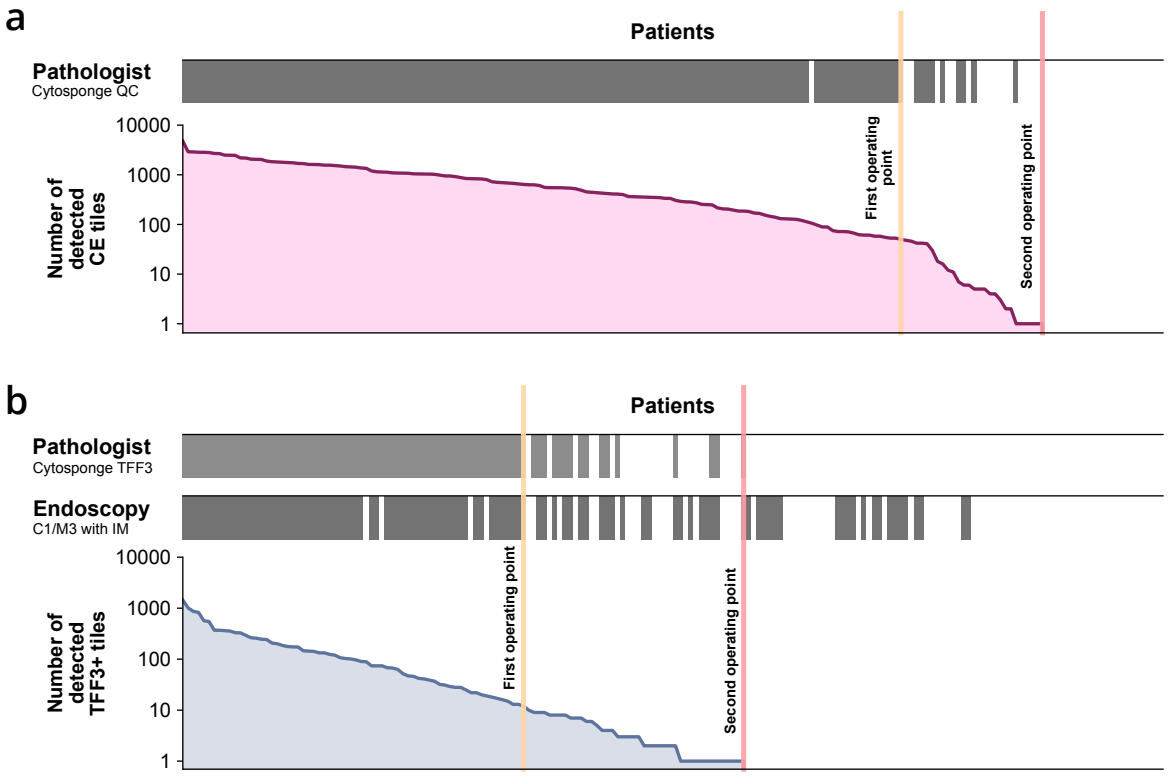


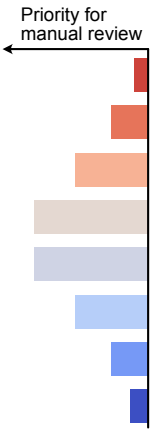
Fig. 3.9 **Application of quality control and diagnostic confidence class scheme to calibration cohort. a** Quality ground truth by pathologist from Cytosponge (top) compared with number of detected columnar epithelium (CE) tiles on H&E detected by VGG-16 (bottom). **b** Diagnosis ground truth by pathologist from Cytosponge (top), Endoscopy (with confirmed IM on biopsy) ground truth (middle) compared with number of detected TFF3-positive tiles on TFF3 detected by ResNet-18 (bottom) / eqv. = equivocal.

table 3.4). For diagnostic confidence classes, the number of tiles detected with TFF3-positive goblet cells were classified as high confidence negative, low confidence equivocal, or high confidence positive (fig. 3.9b, table 3.4). On the internal validation cohort, I observed a visual agreement between these confidence classes and pathology and endoscopy ground truths (fig. 3.8, table 3.5).

I then combined the quality and diagnostic classes into eight triage classes of varying priority for manual review (fig. 3.10a). The relative priority of each class was determined by expert pathologists: Cases with low confidence in sample quality (none or few columnar epithelium detected on H&E) or low confidence in diagnosis (few goblet cells detected on TFF3) should be prioritised for human expert assessment over cases with high-confidence positive or negative evidence. In my internal validation cohort, I find that only 13.0% of patients fall into the triage classes with high priority (4 and 5), while 87.0% fall into the other six classes (fig. 3.10a).

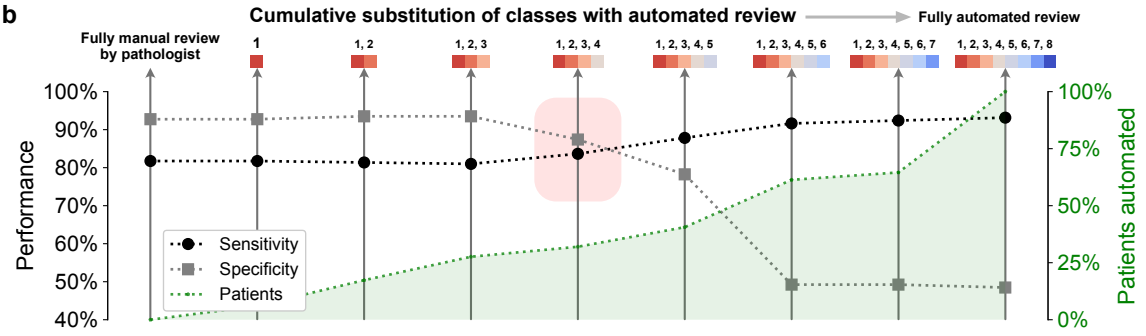
I next asked which classes can be substituted by automated review while retaining the accuracy of full manual review by a human pathologist (sensitivity: 81.7%; specificity: 92.7%). I applied a cumulative substitution scheme and started by substituting class 1 with automated review, then classes 1 and 2, then classes 1, 2, and 3, and so on. In the validation cohort, I found that sensitivity and specificity remain stable if classes 1, 2, and 3 are substituted, but decrease with the substitution of class 4, 5, and 6 (fig. 3.10b). Repeating this procedure starting with class 8 shows that sensitivity and specificity are stable if classes 8 or 7 are substituted, but decrease with the substitution of classes 6, 5, and 4 (fig. 3.10c). These results show that five of the eight classes (1, 2, 3, 7, 8) can be substituted by automated review while three classes (4, 5, 6) should be manually reviewed by a pathologist. This substitution scheme would result in similar performance (sensitivity: 82.5% (CI 95%: 77.3% - 87.2%); specificity: 92.7% (CI 95%: 89.6% - 95.9%)) as fully manual review by a pathologist. These classes cover the majority of patients (66.3% (CI 95%: 62.7% - 70.1%) in validation cohort) and triage-driven, semi automated review would thus save 66% of the pathologists' workload (Methods) by enabling them to focus on difficult cases while leaving easy cases for automated review.

a



No.	Quality control	Diagnosis	Automated result	Validation
1	No confidence	High confidence negative	Negative	33 (6.1%)
2	High confidence	High confidence negative	Negative	59 (11.2%)
3	Low confidence	High confidence negative	Negative	54 (10.3%)
4	No confidence	Low confidence equivocal*	Positive	23 (4.4%)
5	Low confidence	Low confidence equivocal	Positive	45 (8.6%)
6	High confidence	Low confidence equivocal	Positive	109 (20.8%)
7	Low confidence	High confidence positive	Positive	17 (3.2%)
8	High confidence	High confidence positive	Positive	186 (35.4%)
				525 in total

b



c

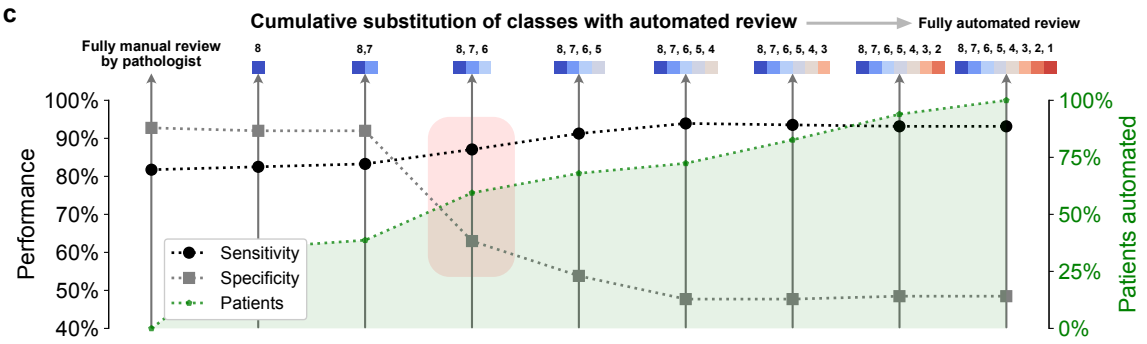


Fig. 3.10 **Triage-driven approach with incremental triage class substitution scheme on internal validation set.** **a** Table of quality control and diagnosis classes. Each class has been assigned a qualitative priority for manual review. Column 'Automated result' refers to the label a sample would be assigned if all samples of this class were automatically reviewed. Asterisk (*): includes combination of no confidence (quality control) and high confidence positive (diagnosis) despite minimal likelihood of occurrence. **b** Cumulative substitution scheme starting with fully manual review, followed by substitution with automated review of class no. 1, then 1 and 2, etc. Red rectangle indicates a drop of performance at substitution stage. **c** Cumulative substitution scheme starting with fully manual review, followed by substitution with automated review of class no. 8, then 8 and 7, etc. Red rectangle indicates a drop of performance at substitution stage.

3.3.5 Simulation of varying cohort composition corroborates reduction in expected workload

My case-control cohort is not representative of a real-world population eligible for Cytosponge-TFF3 testing. In my internal validation set I observed a disease prevalence of 50.0%, while the prevalence expected in a real-world population with GERD symptoms ranges from 3.0% to 7.5% [74, 144–146]. Additionally, the allocation of samples to triage classes depends directly on the amount of sampled cellular material and the resulting sample confidence, which can vary widely and might improve with future refinements of the device administration procedure.

To understand how my results generalize, I devised a simulation approach to vary how many samples have BE and how many samples are allocated to high/low confidence triage classes (Methods). To simulate the change in workload over a range of possible prevalences of BE, I first determined the proportion of patients with and without BE in each triage class and then weighted each vector of proportions by a new prevalence ranging from 0 to 55%. To simulate the effect that relative changes in overall sample confidence have on the workload, I first determined the proportion of patients in triage classes with highest sample confidence (determined by quality control and diagnostic class: 2 and 8) and lower sample confidence (1, 3, 4, 5, 6, and 7). I then modified the proportion of high confidence samples and inversely adapted the proportion of lower confidence samples within a range from -25% to 25%.

Over a fine grid of varying disease prevalence and changes in sample confidence, I observed a negative impact of decreasing cohort BE prevalence and a positive impact of sample confidence on the potential workload reduction (fig. 3.11a). According to this simulation, in a realistic cohort with a BE prevalence of 7%, I would still be able to reduce the pathology workload by 57%. In order to retain the same workload reduction I observed in the validation cohort, the proportion of samples with high confidence in a realistic cohort would need to be increased by 15%.

3.3.6 External validation of triage-driven approach

Finally, I tested the validity of my results and the extrapolation in the simulation study in an independent test set of 3,038 slides from 1,519 patients from from 109 primary care sites in the UK (BEST3 trial) [104]. All slides were processed in the same way and with the same model parameters as the BEST2 validation cohort (fig. 3.12, table 3.6). Following the method described in the previous section, I used manual pathologist reviews for samples that fell into

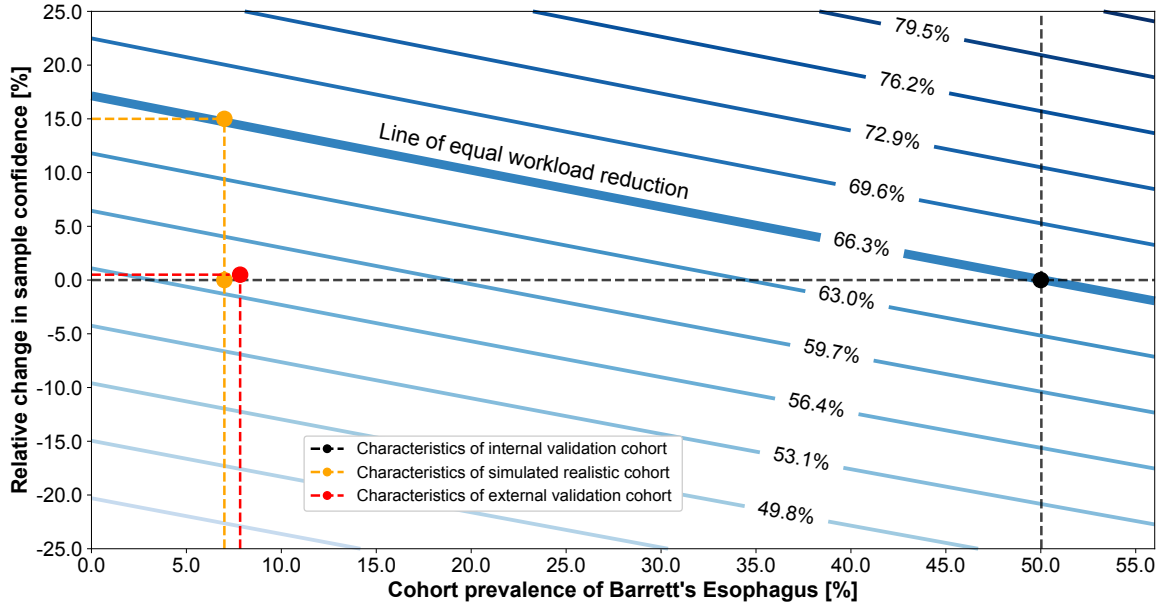


Fig. 3.11 **Triage model applied to external validation cohort and simulation of cohort variation.** Simulation of changes in cohort prevalence of BE and sample confidence with impact on workload reduction. Every contour line (blue) represents the same level of workload reduction as indicated by the percentages. Solid black lines indicate the workload reduction of the validation cohort. The dotted yellow line illustrates the workload reduction of a realistic primary care referral cohort (with 7% prevalence) with no change in sample confidence classes (lower yellow marker) and the confidence change required to match the same amount of workload reduction as in the validation cohort (upper yellow marker). The results from the external validation cohort are shown in red.

triage classes 4, 5 and 6. In the BEST3 trial, endoscopy data was only available for positive Cytosponge patients and those who had Barrett diagnosed at follow-up as a result of standard of care. In addition, the trial was not designed to investigate sensitivity or specificity but positive predictive value (PPV) instead. I also calculated the negative predictive value (NPV) based on findings aggregated through the primary endpoint analysis (coded BE diagnosis in patient records). For this external validation cohort, fully manual review by pathologists resulted in a PPV of 56.08% and NPV of 99.02%. After application of the triage-driven, semi-automated approach the PPV of the overall cohort was 53.37% and the NPV 99.39% (fig. 3.12).

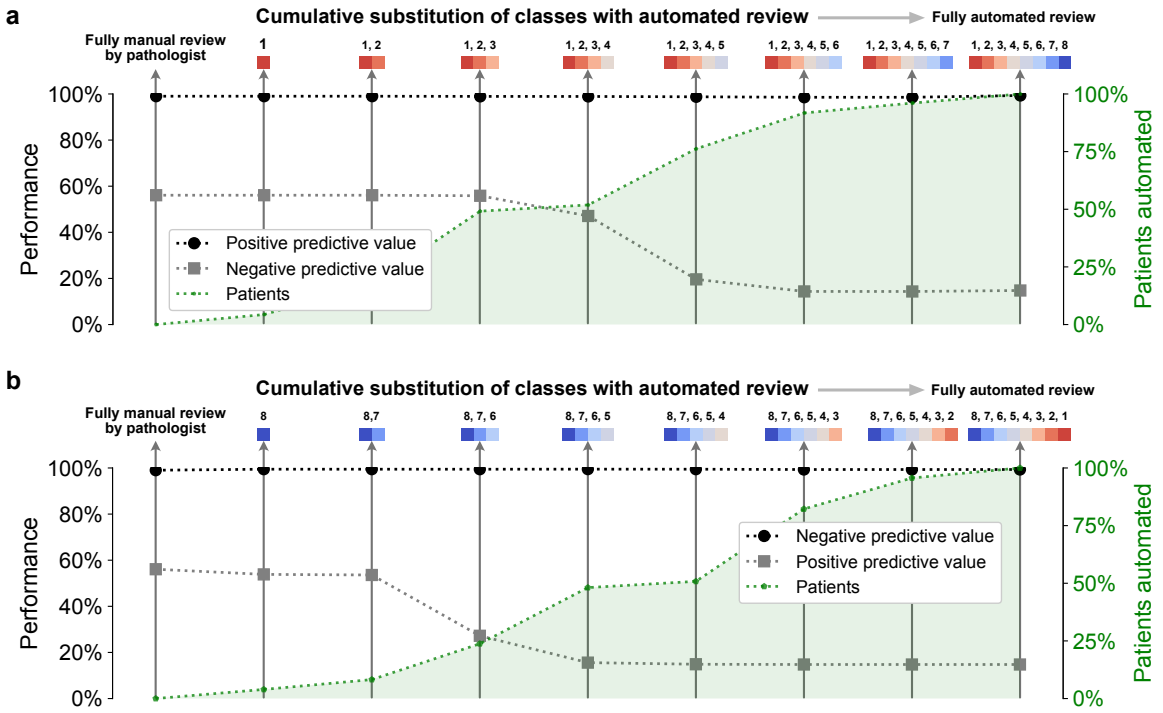


Fig. 3.12 **Performance of semi-automated, triage-driven model on external validation cohort. a** Cumulative substitution scheme starting with fully manual review, followed by substitution with automated review of class no. 1, then 1 and 2, etc. **b** Cumulative substitution scheme starting with fully manual review, followed by substitution with automated review of class no. 8, then 8 and 7, etc.

Using this approach in a realistic primary care setting would have resulted in the following key results: In total 872 patients out of 1,519 patients (57.41%) would have been reviewed automatically while 42.59% would have had to be reviewed manually. This agrees with the simulated, expected value (fig. 3.11) of workload reduction given the prevalence (7.8%) of BE in this external validation cohort. Six additional patients would have been diagnosed

with BE while being missed by the pathologist at the cost of 19 additional endoscopies when compared to fully manual review. One patient would have received an automated negative diagnosis even though the pathologist scored it as positive with BE finding at endoscopy.

3.4 Discussion

Summary

I have presented a triage-driven approach that analyses samples of the Cytosponge-TFF3 test using deep learning for the early detection of oesophageal cancer. My approach combines quality control and diagnostic metrics of pathology slides to stratify patients into 8 triage classes which determine whether a patient sample requires manual or if automated review would suffice.

Benefits of the semi-automated, triage-driven approach

For the analysis of Cytosponge-TFF3 samples, my triaging approach has several benefits: I am able to substantially reduce workload and match the sensitivity and specificity of expert pathologists. In my internal validation cohort, fully manual review by a pathologist achieves 81.7% sensitivity and 92.7% specificity. In a fully automated approach, I observed a sensitivity of 72.6% and a specificity of 93.1%. With my triage-driven approach, I demonstrate that up to 66% of cases can be reviewed automatically while achieving a sensitivity of 82.5% and specificity of 92.7%, a performance marginally superior to fully manual review by pathologists. Further, in an external validation cohort from a large randomised controlled trial I observed a PPV of 53.37% and NPV of 99.39%. For comparison, pathologist review resulted in very similar values with a PPV of 56.08% and NPV of 99.02%. While a small number of additional endoscopies would have been triggered, they would have also yielded more positive diagnoses. In this more realistic cohort, 57.41% workload for the pathologists would have been reduced. These results (fig. 3.11) have several implications: First, a fully automated review would reduce sensitivity (at fixed specificity) and therefore suffer from a loss of clinical utility. Second, while a triage-driven approach is not able to reduce workload as much as a fully automated approach, the described triage classes provide a logical way for stage-wise clinical adoption and performance testing in routine practice.

Another benefit of my approach is that I was able to directly adopt heuristics applied by pathologists familiar with Cytosponge-TFF3 samples in my algorithmic design process. As a result, my approach demonstrates traceability and interpretability [91]: First, I mimicked the screening process of samples observed by expert pathologists by replicating their decision-making scheme (fig. 3.1c). Second, the saliency maps I generated from deep learning models

to visualize learned features in the pathology images show strong agreement with manual landmarks placed by pathologists (fig. 3.4).

As a further benefit, my triage approach achieves strong performance from only 287 samples for training and calibration by incorporating informative prior knowledge about biological and clinical test characteristics, followed by rigorous testing in independent cohorts. This performance compares favorably to previous fully automated approaches reporting expert-level performance that relied on very large datasets with training set sizes ranging from 10,000 to more than 100,000 examples [100, 147] - dataset sizes that cannot be expected for most applications.

Finally, a quantitative analysis of workload reduction across varying disease prevalences and sample confidences shows that my approach is expected to generalize well to a real-world population. A more general population would have a lower disease prevalence than a case-control study, which would cause a larger workload due to the distribution of BE/non-BE patients within the individual triage classes. I was further able to confirm this simulation with an external validation cohort. These findings provide realistic expectations of how clinical decision-making systems are affected by bias in cohort composition.

Limitations of the methodology and technical considerations

My approach has several limitations: First, while samples used in this work were generated at multiple centres they were processed at only a single site (Addenbrookes Hospital, Cambridge, UK). Thus, my data might not fully reflect the variation in histology sectioning and staining across different laboratories [148]. I compensated for this limitation through data augmentation by spatial and color profile distortion. Additionally, my data are not too far from future real-world applications, because for large-scale rollout of the Cytosponge test a centralised laboratory is envisaged to ensure processing as well as analysis with proper quality assurance. In future work, I plan to test whether the superiority of the triage-driven approach over fully manual pathologist review will generalize by incorporating multi-centre data from ongoing and future Cytosponge-TFF3 studies to evaluate this effect more extensively.

Second, the underlying machine learning model could be further optimized. For example, instead of using a transfer learning model based on pre-training with a primary dataset, one could train a model from scratch, which has proven to improve results in some CNN applications [149]. In addition, the tile size needs further investigation because it determines the receptive field in which the CNNs build feature representations of images. My tile size

was chosen by expert pathologists to capture relevant structures like columnar epithelium and goblet cells. Although good performance was observed, a refined multi-scale classification with several magnifications might be necessary to achieve better classification of tissue types. Further improvements might be realised from using attention-based models to reduce the laborious annotation steps required for expanding the training data [101] or aggregating tiles to patient level with more sophisticated approaches based on sequence models [100].

Third, a major determinant of workload reduction is the quality and therefore diagnostic confidence attributed to a sample. However, what determines the amount of columnar material sampled is unknown. One hypothesis is that the strength of oesophageal peristalsis, which can be influenced by variations in device ingestion, may be associated with the likelihood of the Cytosponge reaching the stomach. We plan to investigate determinants of sample quality by comparing the data generated by the trained deep learning models with patient and device administrator profiles.

Conclusion

In summary, my triage approach differs from previous applications of deep learning to medical images [95, 100] which used fully automated approaches on extremely large datasets. I show that for a modest dataset size, leveraging existing heuristics of pathologist decision-making in a triage-based approach is a powerful alternative to fully automated classification models, which generalises well to an independent validation cohort. These results lay the foundation for tailored, semi-automated decision support systems embedded in clinical workflows.

Author contributions

I conceived and led the analysis; MCO and AB contributed to the analysis; I wrote the code for analysis with contributions by AB; MO and RCF were involved in collection and labelling of the data; RCF conceived the study; RCF and FM directed the project; I wrote the manuscript with FM's support and the assistance as well as feedback of all other co-authors.

Competing interests

The Cytosponge device technology and the associated TFF3 biomarker have been licensed to Covidien GI solutions (now owned by Medtronic) by the Medical Research Council. MCO, FM and myself are named inventors on a patent pertaining to technology applied in this work. RCF and MO are named inventors on patents pertaining to the Cytosponge and associated technology. MO, RCF, and myself are shareholders of Cyted Ltd, a company working on early detection technology.

3.5 Supplementary tables

	AlexNet	DenseNet	Inception	ResNet	SqueezeNet	VGG
H&E						
Overall accuracy	0.977	**0.990**	0.989	0.984	0.959	0.988
Precision						
Background	0.999	0.999	0.999	0.999	0.999	0.999
CE (gastric type)	0.791	**0.865**	0.857	0.807	0.763	0.843
CE (respiratory type)	0.389	0.750	**0.895**	0.667	0.241	0.741
Intestinal Metaplasia	0.393	**0.688**	0.609	0.518	0.215	0.640
Recall						
Background	0.984	0.995	**0.996**	0.991	0.963	0.995
CE (gastric type)	0.893	0.947	0.940	0.921	0.935	**0.950**
CE (respiratory type)	0.802	0.779	0.588	0.794	**0.832**	0.634
Intestinal Metaplasia	0.606	0.610	0.629	0.606	**0.643**	0.568
TFF3						
Overall accuracy	0.996	**0.999**	0.998	0.998	**0.999**	0.998
Precision						
Positive	0.752	**0.903**	0.856	0.827	0.589	0.856
Equivocal	0.233	0.513	**0.533**	0.385	0.133	0.404
Negative	1.000	1.000	1.000	1.000	1.000	1.000
Recall						
Positive	0.912	0.890	**0.919**	0.912	0.897	**0.919**
Equivocal	0.465	0.465	0.372	0.465	**0.767**	0.442
Negative	0.997	**1.000**	**1.000**	0.999	0.991	0.999

Table 3.1 **Tile-level precision and recall for all classes from H&E and TFF3 models.** This data is derived from the tiles in the development set. (DenseNet = DenseNet-121, Inception = Inception v3, ResNet = ResNet-18, VGG = VGG-16). The highest value(s) per row is/are highlighted in bold.

	AlexNet	DenseNet	Inception	ResNet	SqueezeNet	VGG
Quality control						
Probability threshold	0.97	0.96	0.995	0.96	0.85	0.99
AUC	0.985	0.984	0.986	0.986	0.980	**0.988**
Diagnosis						
Probability threshold	0.9999	0.87	0.655	0.93	0.99999	0.93
AUC	0.80	0.82	0.83	0.83	0.80	0.83
Sensitivity at fixed specificity (91.57%)	63.4%	62.5%	61.5%	63.5%	60.6%	**64.4%**
Tile number threshold	3	8	10	9	4	6

Table 3.2 **Individual probability threshold calibration with associated performance based on differential ROC analysis for quality control and diagnosis.** The AUC for quality control relates to the performance on the calibration cohort at the given probability threshold for individual tiles containing columnar epithelium on H&E. The AUC for diagnosis relates to the performance on the calibration cohort at the given probability threshold for individual tiles containing positive goblet cells on TFF3. Sensitivity is based on a fixed value of specificity derived from the pathologist performance on the calibration cohort. The tile number threshold is the resulting cutoff from the fixed specificity.

	AUC (CI 95%) vs. pathologist	AUC (CI 95%) vs. endoscopy	Sensitivity (CI 95%)	Specificity (CI 95%)
Quality control				
AlexNet	0.98 (0.97-0.99)	n/a	n/a	n/a
DenseNet	0.98 (0.97-0.99)	n/a	n/a	n/a
Inception v3	0.98 (0.97-0.99)	n/a	n/a	n/a
ResNet-18	0.97 (0.96-0.99)	n/a	n/a	n/a
SqueezeNet	0.97 (0.95-0.98)	n/a	n/a	n/a
VGG-16	0.99 (0.98-0.99)	n/a	n/a	n/a
Diagnosis				
Pathologist	n/a	n/a	81.75% (76.67%-85.92%)	92.75% (89.37%-95.51%)
AlexNet	0.96 (0.94-0.98)	0.86 (0.83-0.89)	72.24% (66.98%-77.37%)	89.70% (85.80%-92.97%)
DenseNet	0.97 (0.96-0.99)	0.89 (0.86-0.91)	70.34% (64.84%-76.24%)	92.75% (89.84%-95.85%)
Inception v3	0.97 (0.96-0.99)	0.88 (0.85-0.91)	69.96% (64.71%-75.65%)	93.13% (89.74%-96.03%)
ResNet-18	0.97 (0.95-0.98)	0.88 (0.85-0.91)	72.24% (66.67%-77.18%)	91.22% (87.72%-94.64%)
SqueezeNet	0.94 (0.92-0.96)	0.85 (0.82-0.88)	69.58% (63.59%-74.54%)	92.37% (88.85%-95.42%)
VGG-16	**0.97 (0.96-0.99)**	**0.88 (0.85-0.91)**	**72.62% (66.72%-77.64%)**	**93.13% (89.75%-96.05%)**

Table 3.3 **Performance of all architectures after application on the validation cohort**. Quality control models relied on pathologist calls on sample quality. Sensitivities or specificities were not determined due to irrelevance in the fully automated model approach. Diagnosis models relied on thresholds determined on the calibration cohort.

Quality classes	No confidence	Low confidence	High confidence
No. of patients	22	27	138
Proportion	11.8%	14.4%	73.8%
QC positive (path)	0	9	137
QC negative (path)	22	18	1
Diagnostic classes	High conf. negative	Low conf. equivocal	High conf. positive
No. of patients	56	59	72
Proportion	30.0%	31.5%	38.5%
TFF3 positive (path)	1	10	71
TFF3 negative (path)	55	49	1
Barrett oesophagus	12	26	66
No Barrett oesophagus	44	33	6

Table 3.4 **Characteristics of patients in quality control and diagnosis classes from calibration cohort.** For each of the three quality control and diagnosis classes, the number of patients within the class and the paired ground truth is shown.

Quality classes	No confidence	Low confidence	High confidence
No. of patients	55	116	354
Proportion	10.5%	22.1%	67.4
QC positive (path)	0	35	350
QC negative (path)	55	81	4
Diagnostic classes	High conf. negative	Low conf. equivocal	High conf. positive
No. of patients	145	177	203
Proportion	27.6%	33.7%	38.7%
TFF3 positive (path)	4	33	197
TFF3 negative (path)	141	144	6
Barrett oesophagus	18	61	184
No Barrett oesophagus	127	116	19

Table 3.5 **Characteristics of patients in quality control and diagnosis classes from validation cohort.** For each of the three quality control and diagnosis classes, the number of patients within the class and the paired ground truth is shown.

Quality classes	No confidence	Low confidence	High confidence
No. of patients	107	912	500
Proportion	7.1%	60.0%	32.9
QC positive (path)	38	733	350
QC negative (path)	69	179	4
Diagnostic classes	High conf. negative	Low conf. equivocal	High conf. positive
No. of patients	747	646	126
Proportion	49.2%	42.5%	8.3%
TFF3 positive (path)	1	83	105
TFF3 negative (path)	746	563	21
Barrett oesophagus	5	38	76
No Barrett oesophagus	742	608	50

Table 3.6 **Characteristics of patients in quality control and diagnosis classes from external validation cohort.** For each of the three quality control and diagnosis classes, the number of patients within the class and the paired ground truth is shown.

Chapter 4

Discussion

The two main objectives of my PhD were to investigate clinical and technical aspects of the Cytosponge-TFF3 test with a focus on requirements for translating the technology into clinical practice. The Cytosponge-TFF3 technology provides a tool for the early detection of Barrett oesophagus, a precursor lesion for oesophageal adenocarcinoma. To better understand the efficacy of the diagnostic test I investigated its performance compared to usual care in a large randomised controlled trial in primary care (chapter 2). Furthermore, I identified, conceptualised, and developed an approach to support the scalability of the technology by using machine learning (chapter 3). This triage-driven approach enables equivocal samples to be presented to a pathologist for review while unequivocal samples only need to undergo automated review.

4.1 Approaches for early detection of oesophageal cancer

Early detection of cancer has developed into an important field in oncology over the last decades [150]. These developments particularly apply to healthy and high-risk patient populations where earlier detection of cancer or pre-malignant lesions comes with an opportunity for treatment with curative intent (section 1.1). A common theme amongst cancers with implemented screening programmes is that they are usually characterised by a high incidence and/or prevalence. They also have a particular tumour biology that enables early intervention due to the presence of pre-malignant or gradually progressing lesions. A significant implication of these early interventions is the health economic benefit which comes from reducing the burden on the healthcare system by lowering the prevalence of late-stage disease.

In recent decades, a number of technologies were developed which are geared towards the earlier detection of cancer (section 1.1).

For oesophageal cancer in particular, minimally invasive sampling gained significant traction with the development of balloon-based sampling and sponge-on-string devices (sections 1.2 and 1.3). The sponge-on-string device with the most advanced level of evidence is Cytosponge-TFF3. Notable studies include BEST1 (cohort study) and BEST2 (case-control study). Both of these studies have a particular focus on diagnostic performance, safety, and acceptability with a direct comparison to OGD as gold standard. A key question raised by promising results of these two studies was whether, in a real-world setting, the detection of BE can be increased by offering a Cytosponge test to eligible patients with GERD in primary care. The direct comparator would be usual care which, for patients with GERD, refers to the prescription of proton pump inhibitors (PPI), lifestyle monitoring and potential endoscopy referral. This was assessed as a primary endpoint in BEST3, a multicentre, pragmatic, randomised controlled trial (chapter 2).

4.2 Clinical evidence base for Cytosponge-TFF3

The BEST3 study found that an offer of Cytosponge-TFF3 to patients on anti-GERD medication can increase the rate of detection in excess of 10-fold when compared to usual care. We further observed a stage shift in which multiple patients in the intervention arm were diagnosed with dysplasia or cancer and were treatable with curative intent. In the usual care arm, detected cancers were more advanced with a prognosis generally involving palliative care. Out of 6,834 patients in the intervention arm of the trial, 2,679 (39%) expressed interest, and 1,654 (24%) underwent the procedure. The low uptake is hypothesised to be related to how patients were invited into the study. Patients received invitation by letter which could have potentially discouraged them from participating when compared to a direct, personal invitation by their GP. This uptake-determining factor is now being investigated as part of the DELTA project [151], a publicly funded project with a consortium of academic and industry partners which conducts real-world implementation pilot studies for Cytosponge-TFF3.

Side effects of the Cytosponge-TFF3 procedure were limited to a small number of patients (63 out of 1,654, 4%) indicating they had a sore throat post procedure. Other side effects included dyspepsia and oesophageal/gastric pain and were in line with findings in previous studies [69]. Patients rated the Cytosponge test as acceptable with a median score of 9 out of 10.

The PPV of patients being diagnosed with Barrett oesophagus (of any length) after a positive Cytosponge-TFF3 test was 59% (121 out of 221 patients). This is a remarkable improvement in performance when compared to a simulated PPV of 24.3% on the basis of 3% BE prevalence from a previous study [75]. The difference with respect to the simulated PPV is the slightly higher BE prevalence in the BEST3 patient population but also an increased sensitivity due to the BEST2 study relying on an ITT analysis where repeat tests were not included. Additionally, we diagnosed intestinal metaplasia of the gastro-oesophageal junction and/or gastric cardia in 33 (15%) out of the 221 patients. Recently reviewed guidelines suggest that there might be clinical significance of gastric IM for development of gastric cancer, however, further evidence is needed to conclude whether patients wit limited extent of IM should undergo regular surveillance. A more extensive discussion of this matter can be found in chapter 2, section 2.4.

Comparability to other minimally invasive sampling technologies on the basis of the BEST3 results is limited due to the type of studies conducted for these technologies so far. Relevant trials which will provide diagnostic accuracy in larger patient populations are ongoing and likely to conclude in 2022 (EsoGuard, ClinicalTrials.gov Identifier: NCT04293458) and 2025 (EsophaCap, ClinicalTrials.gov Identifier: NCT04214119). Despite the capability of Cytosponge-TFF3 to pick up short and long BE segments, a resulting question is how these patients should be followed up if diagnosed. Risk stratification biomarkers for BE have been explored to identify patients at higher risk of progression to dysplasia or cancer [126, 152, 153]. These biomarkers will aid to identify BE lesions with a higher risk of progression so these patients can be prioritised for subsequent endoscopy.

At time of submission of this thesis, this is of particular relevance as the COVID-19 pandemic has put severe pressure on endoscopy services with an over 30% reduction in diagnostic endoscopies for the first quarter of 2020 in England, UK [154]. As a response to this crisis, a refined biomarker panel based on atypia screening and p53 immunohistochemistry similar to Ross-Innes et al. [126] was used by di Pietro et al. [154] to prioritise patients on waiting lists for upper GI endoscopies. The results are encouraging and are currently being validated in a nested case-control study to generate further evidence. However, the patient population for the implementation of such approaches needs to be strictly monitored in order to avoid application of the technology outside its current evidence base. This particularly applies to patients who present with moderate to severe symptoms of dysphagia where an urgent endoscopy should be prioritised over a Cytosponge-TFF3 test.

In the BEST3 study, the selection of the ideal target population for Cytosponge-TFF3 testing was limited to the following characteristics: age of 50 years or older, on acid-suppressant for more than 6 months, and no endoscopy procedure within the past 5 years. For future refinements of this population, it should be considered to include other risk factors such as gender, BMI, extended prescription history, and other potential risk factors. Various approaches to enriching the population for testing are currently explored in collaboration with the University of Oxford as part of the DELTA project [151].

An important limitation of the Cytosponge test is that in a small fraction of patients the Cytosponge capsule fails to reach the stomach and therefore the sample contains few or no gastric columnar epithelium. The consequence is a low-confidence result which was the case in 150 (9%) out of 1,654 patients. A correlation between the instructions for the device administrator and the sample quality has been observed with multiple reports indicating that the way in which the capsule is swallowed (larger vs smaller sips of water) impacts the oesophageal peristalsis of the patient and therefore the passage into the stomach. A re-test in these patients results in additional cost and work is underway to investigate means to reduce the observed low-confidence rate and reduce potential distress for the patient.

In summary, I have identified important performance characteristics of the Cytosponge-TFF3 technology for clinical implementation and aspects for improvements of the technology's *status quo* which are now being investigated as part of the DELTA real-world implementation study [151].

4.3 Pathology assessment of oesophageal cells samples

In a population with a low prevalence of BE, most patients are expected to have a TFF3-negative test result. Therefore, most samples will have no significant clinical finding while they will consume a substantial amount of time for pathologist review whereas a number of patients will produce a positive test which will require confirmation via endoscopy. In addition, a smaller number of patients will have an equivoical test results for which the pathologist requires additional time to inspect the H&E and TFF3 sections.

The BEST3 study, just like previous studies, was based on centralised testing of the oesophageal cell samples which enabled pragmatic quality assurance of the laboratory process and pathologist screening. Unpublished evidence from commercial studies (CASE1 and CASE2) have shown that lack of consistent sample preparation and interpretation of the Cytosponge-TFF3 test can lead to highly discrepant results with a significant effect on

diagnostic performance. The primary cause for these results is that both of these studies relied on testing in individual or multiple laboratories without adequate quality assurance procedures in place. Given the laboratory procedure for clot generation and the pathologist review which requires training and regular quality assurance, a centralised approach for testing is envisaged in order to ensure maximal clinical utility for patients undergoing the Cytosponge test.

At present, the Cytosponge-TFF3 test relies on review by a pathologist of microscopy slides with H&E and TFF3 stains. While the diagnostic concordance was shown to be high in a previous study [75] and double reporting will not be required for most samples, the anticipated adoption of the test technology demands for more scalable approaches due to the limited number and capacity of available GI pathologists. This is further exacerbated by the fact that, in comparison to endoscopic biopsies, Cytosponge sections are significantly larger and therefore require more screening time. Additional aspects of the screening procedure is the careful analysis for the presence of gastric-type columnar epithelium as referred to in section 4.2. While the presence of gastric-type columnar epithelium indicates that the Cytosponge has reached the stomach, respiratory-type columnar epithelium can also be present in the cell samples. Especially during screening of TFF3-positive samples, it has to be confirmed that the TFF3 overexpression does not come from respiratory-type columnar epithelium as occasionally observed in normal and inflamed airways. If TFF3 overexpression is observed, it is recommended to carefully inspect the adjacent H&E section in order to classify the morphology of the columnar epithelium as gastric or respiratory type [73]. Overcalling of Cytosponge-TFF3 samples is important to monitor as it would severely impair the specificity of the test which is essential for its use as a targeted screening tool.

The pattern in the distribution of diagnostic results forms the basis of a key test characteristic which can be exploited for accelerating the screening process. As introduced in chapter 1, section 1.3, I have identified a clear need for high-throughput approaches which can support the screening of Cytosponge-TFF3 by leveraging the associated screening heuristic.

4.4 High-throughput approaches for Cytosponge-TFF3

In order to tackle the need of a pathology assessment tool for workload reduction, I developed a comprehensive approach to (semi-)automate the analysis of Cytosponge-TFF3 samples. In this work, two methods have been presented: First, an automated approach which substitutes the pathologist by generating automated results for all samples. Second, a semi-automated

approach which combines quality control and diagnostic metrics of individual samples into a triage class which determines whether a sample should be reviewed by a pathologist or can be scored in an automated way. This approach enables pathologists to focus only on equivocal cases, substantially reducing their overall workload.

It is important to highlight that a loss of sensitivity, as demonstrated by the fully-automated model, would have a substantial impact on the clinical utility of the diagnostic test. This particularly applies to a targeted screening test such as Cytosponge-TFF3 which will likely be applied to large patient populations. A fully automated approach in this setting would result in reduced health economic benefit. The shortcomings of the fully-automated model can be resolved by the semi-automated model which avoids the automated scoring of equivocal patient samples and flags them for pathologist review. Due to the internal structure of the triage-driven, semi-automated model it would also be possible to pursue clinical implementation by initially only substituting a lower number of triage classes with automated review. A key advantage of such an approach would be the opportunity to closely audit the technical behaviour and clinical consequences of the model in an applied setting.

The key finding of chapter 3 was the semi-automated, triage-driven model as it resulted in superior sensitivity and specificity than the fully-automated model. The fully automated model showed a reduced sensitivity of 9.1% (at fixed specificity) on the internal validation cohort whereas the semi-automated model demonstrated a marginal increase in sensitivity (0.8%) when compared to pathologists while reducing their workload by 66%. The triage-driven model was applied to an external validation cohort (BEST3) with a small loss of 2.71% in PPV and 0.37% in NPV. Due to the variation in cohort composition (i.e. difference in BE prevalence), the workload for the pathologist would have been reduced by 57%.

The triage-driven, semi-automated model was developed with direct involvement of experienced cytopathologists. This is essential for the development of clinical decision support systems as the lack of human involvement often causes issues in addressing the appropriate clinical questions which results in implementation failures [155]. Key characteristics of my system were determined by close observation of pathologists carrying out the screening of Cytosponge samples. This process enabled the identification of the distinct quality control and diagnostic processes which were then developed into a metric on which the triage classes are based. However, as a consequence of the close pathologist involvement, it was clear that the developed models have to be interpretable to build confidence with prospective implementation stakeholders. By using Grad-CAM, an established and well-tested method for visualisation of model focus on tissue morphology or architecture, it was possible to enable a

better understanding of the internal decision processes of the trained deep learning models. As most machine learning models for digital pathology will involve a human pathologist for the foreseeable future, transparency and interpretability has been recognised as key feature of deep learning models for medical applications [156–158].

The method presented in this work was based on two large datasets (BEST2 [75] and BEST3 [104]; chapter 2) which were partitioned in order to enable training, calibration, and validation of the models. It is important to highlight that the deep learning model itself relied on tile images of only 100 patients for training, with a further 187 patients for calibration of various operating points. The resulting performance on the validation data sets is particularly impressive as the underlying origin of the external validation set was different and another antibody for TFF3 staining was used. A consequence of this observation might be that the training and calibration dataset were heterogenous enough to train the models to generalise and anticipate potential domain shifts with respect to age of pathology slides, fading of stains etc. Another finding in this work is that the recognition of IM-positive staining pattern becomes difficult when focusing on a particular region rather than the entire slide. When compared to analysis by a pathologist, the automated scoring will have the benefit to objectively score these regions one after the other without any biases involved. This obviously excludes biases which might have been introduced by the initial training data. Given the results from the validation set, it can be concluded that the training data introduced a minimal bias into the model which can be accounted for to some extent by the ability to triage equivocal cases for pathologist review. The triage-driven, semi-automated approach has been demonstrated to be an effective tool to mitigate issues which could be introduced by fully-automated models or other underlying biases by avoiding automated classification of equivocal patient samples.

Deep learning models for pathology applications are particularly susceptible to failure after a domain shift. Such a shift occurs, for example, due to experimental variations such as (immuno)histochemical stain intensity, or when applying the model on data from a different study, or on images acquired using different scanners. Data augmentation, where different spatial and colour distortions are introduced during training, may help to build a robust model applicable to extended datasets with respect to relevant clinical endpoints. For the analysis of Cytosponge-TFF3 samples I observed limited variability, which is confounded as all samples have been processed in the tissue bank at Cambridge University Hospital. The models were made robust to these variables by augmenting the training data. Ongoing development and implementation studies in the DELTA project [151] will rely on centralised

processing in a new laboratory and performance will be assessed and reported once these data are available. Further technical optimisation of the developed deep learning models is currently being performed with a focus on two main areas: Attention-based models [101] as they could potentially eliminate the laborious annotation step required for expanding the training data and might increase test specificity further. Aggregation of tiles to patient level by means of a more sophisticated approach [100] which is relying on sequence models to potentially reduce false positive rate of TFF3 images.

I also devised and implemented a simulation study based on a simple prevalence and quality model to critically assess the model performances by investigating the cohort composition bias of the validation data sets. The visualisation of the simulation enables a quick estimate of how effective the semi-automated model might perform depending on the identified key characteristics (*i.e.* BE prevalence and sample quality). Whereas BE prevalence is an intuitive variable in a certain patient population and depends on the selection criteria, sample quality and its influences are currently under investigation (chapter 4, section 4.2) The simulation is an approach which can easily be extended to other variables and, particularly for dichotomous diagnostic problems, may prove useful to understand how decision support systems impact clinical pathways and identify confounding factors that influence performance.

4.5 Real-world implementation of Cytosponge-TFF3

The combined outcome of both areas investigated in chapter 2 and chapter 3 provides additional relevant evidence for the Cytosponge-TFF3 technology, further enabling the implementation of a diagnostic test in clinical care. Pilot implementations to investigate clinical utility of Cytosponge-TFF3 in the National Health Service and abroad are underway. Furthermore, projects backed by commercial and public funding have been initiated that expand on the concept of the triage-driven computational pathology approach [151]. Another important outcome of the work presented in this thesis is the spin-out company Cyted which I co-founded and am leading as the Chief Executive Officer. Since late 2019, we raised £8.7M in equity and grant funding and the team size has recently exceeded 20 employees. At Cyted, we are working on various commercial projects with a key focus on implementation of Cytosponge in clinical care and the development of digital diagnostic infrastructure for histo/cytopathology as well as molecular diagnostic tests. The ongoing real-world implementation pilots which are supported by Cyted can be divided into four distinct pathways:

1. Patients presenting to their GP with reflux symptoms and especially those with increased risk for Barrett and oesophageal cancer (Male, > 50 years, BMI > 30). The GP may be unsure whether to refer for endoscopy and whether they need long term PPI;

2. Patients on repeat prescriptions for reflux disease who have not had an endoscopy in the last 5 years;

3. Routine referrals to secondary care with heartburn or reflux predominant symptoms and no alarm symptoms (e.g. no anaemia, dysphagia, weight loss etc). This may be especially relevant during the COVID-19 pandemic when endoscopy services are resctricted and waiting lists are long. The BEST2 control arm [75] was representative of patients in this group with no Barrett diagnosed. Further data will be audited for this patient group.

4. Routine endoscopy, including for surveillance, is severely curtailed during the COVID-19 pandemic. Therefore, Cytosponge can also be considered for patients with known Barrett who would usually have endoscopic surveillance. For these individuals information on their risk will be provided through use of additional biomarkers (p53 staining and pathological assessment for atypia, as in Ross-Innes et al. [126]). This information can be used to help prioritise patients for endoscopy.

The success of sustainable implementation in healthcare systems of the individual pathways will be monitored and reported in the future.

4.6 Future outlook

Early detection of oesophageal cancer will undergo a radical change in the next decade. Areas of ongoing development and innovation can be divided into three distinct groups:

First, non-endoscopic detection of oesophageal diseases. The Cytosponge-TFF3 technology is one of several technologies which has been developed in recent years. Other sponge-on-string devices (e.g. EsophaCap) have been developed in conjunction with various types of biomarkers including cytopathology and epigenetic testing [159, 160]. Whether cytopathology or molecular tests will prove to be more accurate is, based on previous studies across different devices, debatable and there remain open questions with respect to feasibility and scalability. Balloon-based devices (e.g. EsoCheck) might enable more targeted sampling with promising results based on methylation testing [66]. Clinical adoption of these

technologies will depend on the clinical evidence and performance in randomised trials and real-world settings which are underway for some of the above.

Second, advanced endoscopic interventions for detection and sampling of (pre-)malignant tissue. Endoscopy relying on new imaging technologies such as hyperspectral imaging [161] and deep learning [162] will improve detection of dysplastic and cancerous lesions. These imaging-based technologies will potentially be used after patients have been identified to undergo endoscopy by a prior non-endoscopic testing procedure. Advanced approaches for tissue sampling will remove the need of unnecessary biopsies by enabling brush-based tissue collection [84] for laboratory analysis.

Last, biomarker development for more advanced patient stratification. The determination of progression risk has become an area of importance in the recent years. Multiple research studies relying on histopathological or genomic tests have shown that it is possible to stratify patients into those at risk of progressing from BE to dysplasia or cancer and those who will likely have indolent lesions [126, 163–165]. The integration of such tests with non-endoscopic sampling techniques has been previously demonstrated [126] but gives rise to a need for additional advanced clinical evidence. Enriched tissue-of-interest sampling by using techniques like biopsies or brushes will likely benefit first of any new biomarkers that rely on high signal-to-background ratios. Application to non-endoscopic tissue/cell sampling devices will follow in due course as and when cellular enrichment techniques will be established.

Combined, these advancements in the early detection of oesophageal cancer will substantially increase quality and quantity of life for affected patients. Challenges of implementing early detection technologies at scale have to be considered for the transformation of research methods into clinical services. These considerations include the importance of diagnostics in precision medicine to ensure minimisation of overdiagnosis and overtreatment. This thesis addresses a number of implementation challenges for a specific technology, considers characteristics of the associated clinical pathway, and highlights translational technological aspects that will enable earlier detection of oesophageal cancer.

Chapter 5

Publications

5.1 Manuscripts

This is a list of all publications as a result of my PhD between April 2018 and September 2020.

Cytosponge-trefoil factor 3 versus usual care to identify Barrett's oesophagus in a primary care setting: a multicentre, pragmatic, randomised controlled trial
Authors: Fitzgerald RC, Di Pietro M, O'Donovan M, Maroni R , Muldrew B, Debiram-Beecham I, Gehrung M, [...], Sasieni P. *The Lancet 2020*

Triage-driven diagnosis for early detection of esophageal cancer using deep learning
Authors: Gehrung M, Crispin-Ortuzar M, Berman AG, O'Donovan M, Fitzgerald RC, Markowetz F. *In revision at Nature Medicine*

Role of TFF3 as an adjunct in the diagnosis of Barrett's esophagus using a minimally invasive esophageal sampling device — The Cytosponge™
Authors: Paterson AL, Gehrung M, Fitzgerald RC, O'Donovan M. *Diagnostic Cytopathology 2020*

A guide to deep learning on whole slide images
Authors: Berman A, Gehrung M, Markowetz F. *manuscript in preparation*

Three-Dimensional Printed Molds for Image-Guided Surgical Biopsies: An Open Source Computational Platform
Authors: Crispin-Ortuzar M*, Gehrung M*, Ursprung S*, [...] Steward GD, Sala E, Markowetz F. *JCO Clinical Cancer Informatics 2020* / * equal contribution

Data integration for biomarker validation using miRNA and computational pathology from non-endoscopic oesophageal cell samples
Authors: Masque-Soler N, Gehrung M, Kosmidou C, Markowetz F, Fitzgerald R. *validation experiments ongoing / manuscript in preparation*

5.2 Patents

Patent application GB2009208.6: **Automated Assessment of Pathology Samples.** Application filed in 2020 by Cancer Research Technologies Ltd.

5.3 Software packages

PathML (under development) - https://github.com/9xg/pathml

References

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[2] Anne Shrestha, Charlene Martin, Maria Burton, Stephen Walters, Karen Collins, and Lynda Wyld. Quality of life versus length of life considerations in cancer patients: a systematic literature review. *Psycho-oncology*, 28(7):1367–1380, 2019.

[3] Geoffrey M Cooper and Robert E Hausman. *The cell: Molecular approach*. Medicinska naklada, 2004.

[4] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[5] Halliday A Idikio. Human cancer classification: a systems biology-based model integrating morphology, cancer stem cells, proteomics, and genomics. *Journal of Cancer*, 2:107, 2011.

[6] Sibaji Sarkar, Garrick Horn, Kimberly Moulton, Anuja Oza, Shannon Byler, Shannon Kokolus, and McKenna Longacre. Cancer development, progression, and therapy: an epigenetic overview. *International journal of molecular sciences*, 14(10):21087–21113, 2013.

[7] Nadezhda V Krakhmal, MV Zavyalova, EV Denisov, SV Vtorushin, and VM Perelmuter. Cancer invasion: patterns and mechanisms. *Acta Naturae ( )*, 7(2 (25)), 2015.

[8] Ralph J DeBerardinis and Navdeep S Chandel. Fundamentals of cancer metabolism. *Science advances*, 2(5):e1600200, 2016.

[9] Dass S Vinay, Elizabeth P Ryan, Graham Pawelec, Wamidh H Talib, John Stagg, Eyad Elkord, Terry Lichtor, William K Decker, Richard L Whelan, HMC Shantha Kumara, et al. Immune evasion in cancer: Mechanistic basis and therapeutic strategies. In *Seminars in cancer biology*, volume 35, pages S185–S198. Elsevier, 2015.

[10] Penny A Jeggo, Laurence H Pearl, and Antony M Carr. Dna repair, genome stability and cancer: a historical perspective. *Nature Reviews Cancer*, 16(1):35, 2016.

[11] Sergei I Grivennikov, Florian R Greten, and Michael Karin. Immunity, inflammation, and cancer. *Cell*, 140(6):883–899, 2010.

References

[12] Josette M Northcott, Ivory S Dean, Janna K Mouw, and Valerie M Weaver. Feeling stress: The mechanics of cancer progression and aggression. *Frontiers in cell and developmental biology*, 6:17, 2018.

[13] Marc-Olivier Turgeon, Nicholas JS Perry, and George Poulogiannis. Dna damage, repair, and cancer metabolism. *Frontiers in oncology*, 8:15, 2018.

[14] Yugang Wang, Yan Xia, and Zhimin Lu. Metabolic features of cancer cells, 2018.

[15] Fabian Spill, Daniel S Reynolds, Roger D Kamm, and Muhammad H Zaman. Impact of the physical microenvironment on tumor progression and metastasis. *Current opinion in biotechnology*, 40:41–48, 2016.

[16] André M Ilbawi and Benjamin O Anderson. Cancer in global health: How do prevention and early detection strategies relate? *Science translational medicine*, 7(278):278cm1–278cm1, 2015.

[17] Joshua D Schiffman, Paul G Fisher, and Peter Gibbs. Early detection of cancer: past, present, and future. *American Society of Clinical Oncology Educational Book*, 35(1):57–65, 2015.

[18] Tobore Onojighofia Tobore. On the need for the development of a cancer early detection, diagnostic, prognosis, and treatment response system. *Future Science OA*, 6(2):FSO439, 2019.

[19] John S Spratt. The primary and secondary prevention of cancer. *Journal of surgical oncology*, 18(3):219–230, 1981.

[20] Paul B Jacobsen and Michael A Andrykowski. Tertiary prevention in cancer care: Understanding and addressing the psychological dimensions of cancer during the active treatment period. *American Psychologist*, 70(2):134, 2015.

[21] Robert A Smith, Kimberly S Andrews, Durado Brooks, Stacey A Fedewa, Deana Manassaram-Baptiste, Debbie Saslow, and Richard C Wender. Cancer screening in the united states, 2019: A review of current american cancer society guidelines and current issues in cancer screening. *CA: a cancer journal for clinicians*, 69(3):184–210, 2019.

[22] Kavita Nanda, Douglas C McCrory, Evan R Myers, Lori A Bastian, Vic Hasselblad, Jason D Hickey, and David B Matchar. Accuracy of the papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review. *Annals of internal medicine*, 132(10):810–819, 2000.

[23] Matejka Rebolj, Janet Rimmer, Karin Denton, John Tidy, Christopher Mathews, Kay Ellis, John Smith, Chris Evans, Thomas Giles, Viki Frew, et al. Primary cervical screening with high risk human papillomavirus testing: observational study. *bmj*, 364, 2019.

[24] Jeffrey K Lee, Elizabeth G Liles, Stephen Bent, Theodore R Levin, and Douglas A Corley. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. *Annals of internal medicine*, 160(3):171–181, 2014.

[25] Perry J Pickhardt, Cesare Hassan, Steve Halligan, and Riccardo Marmo. Colorectal cancer: Ct colonography and colonoscopy for detection—systematic review and meta-analysis. *Radiology*, 259(2):393–405, 2011.

[26] US Preventive Services Task Force et al. Screening for breast cancer: Us preventive services task force recommendation statement. *Annals of internal medicine*, 151(10):716, 2009.

[27] Peter B Bach, Joshua N Mirkin, Thomas K Oliver, Christopher G Azzoli, Donald A Berry, Otis W Brawley, Tim Byers, Graham A Colditz, Michael K Gould, James R Jett, et al. Benefits and harms of ct screening for lung cancer: a systematic review. *Jama*, 307(22):2418–2429, 2012.

[28] Jonathan CM Wan, Charles Massie, Javier Garcia-Corbacho, Florent Mouliere, James D Brenton, Carlos Caldas, Simon Pacey, Richard Baird, and Nitzan Rosenfeld. Liquid biopsies come of age: towards implementation of circulating tumour dna. *Nature Reviews Cancer*, 17(4):223, 2017.

[29] MC Liu, GR Oxnard, EA Klein, C Swanton, MV Seiden, Minetta C Liu, Geoffrey R Oxnard, Eric A Klein, David Smith, Donald Richards, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free dna. *Annals of Oncology*, 2020.

[30] Joshua D Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359(6378):926–930, 2018.

[31] Richard B Lanman, Stefanie A Mortimer, Oliver A Zill, Dragan Sebisanovic, Rene Lopez, Sibel Blau, Eric A Collisson, Stephen G Divers, Dave SB Hoon, E Scott Kopetz, et al. Analytical and clinical validation of a digital sequencing panel for quantitative, highly accurate evaluation of cell-free circulating tumor dna. *PloS one*, 10(10):e0140712, 2015.

[32] Jennifer M Croswell, David F Ransohoff, and Barnett S Kramer. Principles of cancer screening: lessons from history and study design issues. In *Seminars in oncology*, volume 37, pages 202–215. Elsevier, 2010.

[33] James F Holland and Raphael E Pollock. *Holland-Frei cancer medicine 8*, volume 8. PMPH-USA, 2010.

[34] Heiko Pohl, Brenda Sirovich, and H Gilbert Welch. Esophageal adenocarcinoma incidence: are we reaching the peak? *Cancer Epidemiology and Prevention Biomarkers*, 19(6):1468–1470, 2010.

[35] Yonne Peters, Ali Al-Kaabi, Nicholas J. Shaheen, Amitabh Chak, Andrew Blum, Rhonda F. Souza, Massimiliano Di Pietro, Prasad G. Iyer, Oliver Pech, Rebecca C. Fitzgerald, and Peter D. Siersema. Barrett oesophagus. *Nature Reviews Disease Primers*, 5(1), May 2019.

[36] Neal D Freedman, Christian C Abnet, Michael F Leitzmann, Traci Mouw, Amy F Subar, Albert R Hollenbeck, and Arthur Schatzkin. A prospective study of tobacco, alcohol, and the risk of esophageal and gastric cancer subtypes. *American journal of epidemiology*, 165(12):1424–1433, 2007.

[37] Anoop Prabhu, Kenneth O Obi, and Joel H Rubenstein. The synergistic effects of alcohol and tobacco consumption on the risk of esophageal squamous cell carcinoma: a meta-analysis. *American Journal of Gastroenterology*, 109(6):822–827, 2014.

[38] C Yang, Keitaro Matsuo, Hidemi Ito, Kaoru Hirose, Kenji Wakai, Toshiko Saito, Masayuki Shinoda, Shunzo Hatooka, Kazuko Mizutani, and Kazuo Tajima. Esophageal cancer risk by aldh2 and adh2 polymorphisms and alcohol consumption: exploration of gene-environment and gene-gene interactions. *Asian Pacific Journal of Cancer Prevention*, 6(3):256, 2005.

[39] Neal D Freedman, Yikyung Park, Amy F Subar, Albert R Hollenbeck, Michael F Leitzmann, Arthur Schatzkin, and Christian C Abnet. Fruit and vegetable intake and esophageal cancer in a large prospective cohort study. *International journal of cancer*, 121(12):2753–2760, 2007.

[40] Farhad Islami, Paolo Boffetta, Jian-Song Ren, Leah Pedoeim, Dara Khatib, and Farin Kamangar. High-temperature beverages and foods and esophageal cancer risk—a systematic review. *International journal of cancer*, 125(3):491–524, 2009.

[41] Ethan B Ludmir, Sarah J Stephens, Manisha Palta, Christopher G Willett, and Brian G Czito. Human papillomavirus tumor infection in esophageal squamous cell carcinoma. *Journal of gastrointestinal oncology*, 6(3):287, 2015.

[42] Lesley A Anderson, RG Peter Watson, Seamus J Murphy, Brian T Johnston, Harry Comber, Jim Mc Guigan, John V Reynolds, and Liam J Murray. Risk factors for barrett's oesophagus and oesophageal adenocarcinoma: results from the finbar study. *World journal of gastroenterology: WJG*, 13(10):1585, 2007.

[43] Melina Arnold, Isabelle Soerjomataram, Jacques Ferlay, and David Forman. Global incidence of oesophageal cancer by histological subtype in 2012. *Gut*, 64(3):381–387, 2015.

[44] Jesper Lagergren and Pernilla Lagergren. Oesophageal cancer. *Bmj*, 341:c6280, 2010.

[45] Katerina Dvorak, Aaron Goldman, Jianping Kong, John P Lynch, Lloyd Hutchinson, Jean Marie Houghton, Hao Chen, Xiaoxin Chen, Kausilia K Krishnadath, and Wytske M Westra. Molecular mechanisms of barrett's esophagus and adenocarcinoma. *Annals of the New York Academy of Sciences*, 1232(1):381–391, 2011.

[46] Nadine M Vaninetti, Laurette Geldenhuys, Geoffrey A Porter, Harvey Risch, Pierre Hainaut, Duane L Guernsey, and Alan G Casson. Inducible nitric oxide synthase, nitrotyrosine and p53 mutations in the molecular pathogenesis of barrett's esophagus and esophageal adenocarcinoma. *Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center*, 47(4):275–285, 2008.

[47] Austin M Dulak, Petar Stojanov, Shouyong Peng, Michael S Lawrence, Cameron Fox, Chip Stewart, Santhoshi Bandla, Yu Imamura, Steven E Schumacher, Erica Shefler, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature genetics*, 45(5):478–486, 2013.

[48] Michael Quante, Govind Bhagat, Julian A Abrams, Frederic Marache, Pamela Good, Michele D Lee, Yoomi Lee, Richard Friedman, Samuel Asfaha, Zinaida Dubeykovskaya, et al. Bile acid and inflammation activate gastric cardia stem cells in a mouse model of barrett-like metaplasia. *Cancer cell*, 21(1):36–51, 2012.

[49] Xia Wang, Hong Ouyang, Yusuke Yamamoto, Pooja Ashok Kumar, Tay Seok Wei, Rania Dagher, Matthew Vincent, Xin Lu, Andrew M Bellizzi, Khek Yu Ho, et al. Residual embryonic cells as precursors of a barrett's-like metaplasia. *Cell*, 145(7):1023–1035, 2011.

[50] Gianmarco Contino, Thomas L Vaughan, David Whiteman, and Rebecca C Fitzgerald. The evolving genomic landscape of barrett's esophagus and esophageal adenocarcinoma. *Gastroenterology*, 153(3):657–673, 2017.

[51] Elizabeth C Smyth, Jesper Lagergren, Rebecca C Fitzgerald, Florian Lordick, Manish A Shah, Pernilla Lagergren, and David Cunningham. Oesophageal cancer. *Nature reviews Disease primers*, 3:17048, 2017.

[52] Matthew D Stachler, Amaro Taylor-Weiner, Shouyong Peng, Aaron McKenna, Agoston T Agoston, Robert D Odze, Jon M Davison, Katie S Nason, Massimo Loda, Ignaty Leshchiner, et al. Paired exome analysis of barrett's esophagus and adenocarcinoma. *Nature genetics*, 47(9):1047–1055, 2015.

[53] Brian J Reid, Laura J Prevo, Patricia C Galipeau, Carissa A Sanchez, Gary Longton, Douglas S Levine, Patricia L Blount, and Peter S Rabinovitch. Predictors of progression in barrett's esophagus ii: baseline 17p (p53) loss of heterozygosity identifies a patient subset at increased risk for neoplastic progression. *The American journal of gastroenterology*, 96(10):2839–2848, 2001.

[54] Patricia C Galipeau, Laura J Prevo, Carissa A Sanchez, Gary M Longton, and Brian J Reid. Clonal expansion and loss of heterozygosity at chromosomes 9p and 17p in premalignant esophageal (barrett's) tissue. *Journal of the National Cancer Institute*, 91(24):2087–2095, 1999.

[55] Caryn S Ross-Innes, Jennifer Becq, Andrew Warren, R Keira Cheetham, Helen Northen, Maria O'Donovan, Shalini Malhotra, Massimiliano Di Pietro, Sergii Ivakhno, Miao He, et al. Whole-genome sequencing provides new insights into the clonal architecture of barrett's esophagus and esophageal adenocarcinoma. *Nature genetics*, 47(9):1038–1046, 2015.

[56] Bodo Klump, Chih-Jen Hsieh, Karlheinz Holzmann, Michael Gregor, and Rainer Porschen. Hypermethylation of the cdkn2/p16 promoter during neoplastic progression in barrett's esophagus. *Gastroenterology*, 115(6):1381–1386, 1998.

[57] Rebecca C Fitzgerald, Massimiliano di Pietro, Krish Ragunath, Yeng Ang, Jin-Yong Kang, Peter Watson, Nigel Trudgill, Praful Patel, Philip V Kaye, Scott Sanders, et al. British society of gastroenterology guidelines on the diagnosis and management of barrett's oesophagus. *Gut*, 63(1):7–42, 2014.

[58] Massimiliano di Pietro and Rebecca C Fitzgerald. Revised british society of gastroenterology recommendation on the diagnosis and management of barrett's oesophagus with low-grade dysplasia. *Gut*, 67(2):392–393, 2018.

[59] Hashem B El-Serag, Stephen Sweet, Christopher C Winchester, and John Dent. Update on the epidemiology of gastro-oesophageal reflux disease: a systematic review. *Gut*, 63(6):871–880, 2014.

[60] Vedha Sanghi and Prashanthi N Thota. Barrett's esophagus: novel strategies for screening and surveillance. *Therapeutic advances in chronic disease*, 10:2040622319837851, 2019.

[61] Miguel Muñoz-Navas. Capsule endoscopy. *World Journal of Gastroenterology: WJG*, 15(13):1584, 2009.

[62] Clare Parker, Estratios Alexandridis, John Plevris, James O'Hara, and Simon Panter. Transnasal endoscopy: no gagging no panic! *Frontline gastroenterology*, 7(4):246–256, 2016.

[63] James East, Jasper L Vleugels, Philip Roelandt, Pradeep Bhandari, Raf Bisschops, Evelien Dekker, Cesare Hassan, Gareth Horgan, Ralf Kiesslich, Gaius Longcroft-Wheaton, et al. Advanced endoscopic imaging: european society of gastrointestinal endoscopy (esge) technology review. *Endoscopy*, 2016.

[64] Yonne Peters, Ruud WM Schrauwen, Adriaan C Tan, Sanne K Bogers, Bart de Jong, and Peter D Siersema. Detection of barrett's oesophagus through exhaled breath using an electronic nose device. *Gut*, 69(7):1169–1172, 2020.

[65] Erik J Snider, Griselda Compres, Daniel E Freedberg, Marla J Giddins, Hossein Khiabanian, Charles J Lightdale, Yael R Nobel, Nora C Toussaint, Anne-Catrin Uhlemann, and Julian A Abrams. Barrett's esophagus is associated with a distinct oral microbiome. *Clinical and translational gastroenterology*, 9(3), 2018.

[66] Helen R Moinova, Thomas LaFramboise, James D Lutterbaugh, Apoorva Krishna Chandar, John Dumot, Ashley Faulx, Wendy Brock, Omar De la Cruz Cabrera, Kishore Guda, Jill S Barnholtz-Sloan, et al. Identifying dna methylation biomarkers for non-endoscopic detection of barrett's esophagus. *Science translational medicine*, 10(424), 2018.

[67] Prasad G Iyer, William R Taylor, Michele L Johnson, Ramona L Lansing, Kristyn A Maixner, Tracy C Yab, Julie A Simonson, Mary E Devens, Seth W Slettedahl, Douglas W Mahoney, et al. Highly discriminant methylated dna markers for the non-endoscopic detection of barrett's esophagus. *American Journal of Gastroenterology*, 113(8):1156–1166, 2018.

[68] Pauline Bus, Christine Kestens, Fiebo Jan Willem Ten Kate, Wilbert Peters, Joost Paulus Hubertus Drenth, Jeanine Merel Leonoor Roodhart, Peter Derk Siersema, and Jantine Wilhelmina Paula Maria van Baal. Profiling of circulating microrna s in patients with barrett's esophagus and esophageal adenocarcinoma. *Journal of gastroenterology*, 51(6):560–570, 2016.

[69] Wladyslaw Januszewicz, Wei Keith Tan, Katie Lehovsky, Irene Debiram-Beecham, Tara Nuckcheddy, Susan Moist, Sudarshan Kadri, Massimiliano di Pietro, Alex Boussioutas, Nicholas J Shaheen, et al. Safety and acceptability of esophageal cytosponge cell collection device in a pooled analysis of data from individual patients. *Clinical Gastroenterology and Hepatology*, 17(4):647–656, 2019.

[70] Pierre Lao-Sirieix, Alex Boussioutas, Sudarshan R Kadri, Maria O'Donovan, Irene Debiram, Lakshmi Harihar, Rebecca C Fitzgerald, et al. Non-endoscopic screening biomarkers for barrett's oesophagus: from microarray analysis to the clinic. *Gut*, 2009.

[71] Keith D Tardif, Michael T Pyne, Elisabeth Malmberg, Tatum C Lunt, and Robert Schlaberg. Cervical cytology specimen stability in surepath preservative and analytical sensitivity for hpv testing with the cobas and hybrid capture 2 tests. *PloS one*, 11(2):e0149611, 2016.

[72] James Weidmann, Abhijeet Chaubal, and Marluce Bibbo. Cellular fixation. a study of cytorich red and cytospin collection fluid. *Acta cytologica*, 41(1):182, 1997.

[73] Anna L. Paterson, Marcel Gehrung, Rebecca C. Fitzgerald, and Maria O'Donovan. Role of tff3 as an adjunct in the diagnosis of barrett's esophagus using a minimally invasive esophageal sampling device—the cytospongetm. *Diagnostic Cytopathology*, 2019.

[74] Sudarshan R Kadri, Pierre Lao-Sirieix, Maria O'Donovan, Irene Debiram, Madhumita Das, Jane M Blazeby, Jon Emery, Alex Boussioutas, Helen Morris, Fiona M Walter, et al. Acceptability and accuracy of a non-endoscopic screening test for barrett's oesophagus in primary care: cohort study. *Bmj*, 341:c4372, 2010.

[75] Caryn S Ross-Innes, Irene Debiram-Beecham, Maria O'Donovan, Elaine Walker, Sibu Varghese, Pierre Lao-Sirieix, Laurence Lovat, Michael Griffin, Krish Ragunath, Rehan Haidry, et al. Evaluation of a minimally invasive cell sampling device coupled with assessment of trefoil factor 3 expression for diagnosing barrett's esophagus: a multi-center case–control study. *PLoS medicine*, 12(1):e1001780, 2015.

[76] Umair Iqbal, Osama Siddique, Anais Ovalle, Hafsa Anwar, and Steven F Moss. Safety and efficacy of a minimally invasive cell sampling device ('cytosponge') in the diagnosis of esophageal pathology: a systematic review. *European Journal of Gastroenterology & Hepatology*, 30(11):1261–1269, 2018.

[77] David A Katzka, Thomas C Smyrk, Jeffrey A Alexander, Debra M Geno, RoseMary A Beitia, Audrey O Chang, Nicholas J Shaheen, Rebecca C Fitzgerald, and Evan S Dellon. Accuracy and safety of the cytosponge for assessing histologic activity in eosinophilic esophagitis: a two-center study. *The American Journal of Gastroenterology*, 112(10):1538, 2017.

[78] Madeleine Freeman, Judith Offman, Fiona M Walter, Peter Sasieni, and Samuel G Smith. Acceptability of the cytosponge procedure for detecting barrett's oesophagus: a qualitative study. *BMJ open*, 7(3):e013901, 2017.

[79] Tatiana Benaglia, Linda D Sharples, Rebecca C Fitzgerald, and Georgios Lyratzopoulos. Health benefits and cost effectiveness of endoscopic and nonendoscopic cytosponge screening for barrett's esophagus. *Gastroenterology*, 144(1):62–73, 2013.

[80] Curtis R Heberle, Amir-Houshang Omidvari, Ayman Ali, Sonja Kroep, Chung Yin Kong, John M Inadomi, Joel H Rubenstein, Angela C Tramontano, Emily C Dowling, William D Hazelton, et al. Cost effectiveness of screening patients with gastroesophageal reflux disease for barrett's esophagus with a minimally invasive cell sampling device. *Clinical Gastroenterology and Hepatology*, 15(9):1397–1404, 2017.

[81] Hamza Chettouh, Oliver Mowforth, Núria Galeano-Dalmau, Navya Bezawada, Caryn Ross-Innes, Shona MacRae, Irene Debiram-Beecham, Maria O'donovan, and Rebecca C Fitzgerald. Methylation panel is a diagnostic biomarker for barrett's oesophagus in endoscopic biopsies and non-endoscopic cytology specimens. *Gut*, 67(11):1942–1949, 2018.

[82] JF Johanson, J Frakes, D Eisen, et al. Computer-assisted analysis of abrasive transepithelial brush biopsies increases the effectiveness of esophageal screening: a multicenter prospective clinical trial by the endocdx collaborative group. *Digestive diseases and sciences*, 56(3):767–772, 2011.

[83] Seth A Gross, Michael S Smith, Vivek Kaul, and US Collaborative WATS3D Study Group. Increased detection of barrett's esophagus and esophageal dysplasia with adjunctive use of wide-area transepithelial sample with three-dimensional computer-assisted analysis (wats). *United European gastroenterology journal*, 6(4):529–535, 2018.

[84] Vivek Kaul, Seth Gross, F Scott Corbett, Zubair Malik, Michael S Smith, Christina Tofani, and Anthony Infantolino. Clinical utility of wide-area transepithelial sampling with three-dimensional computer-assisted analysis (wats3d) in identifying barrett's esophagus and associated neoplasia. *Diseases of the Esophagus*, 2020.

[85] David N Louis, Georg K Gerber, Jason M Baron, Lyn Bry, Anand S Dighe, Gad Getz, John M Higgins, Frank C Kuo, William J Lane, James S Michaelson, et al. Computational pathology: an emerging definition. *Archives of pathology & laboratory medicine*, 138(9):1133–1138, 2014.

[86] Elizabeth Montgomery. Is there a way for pathologists to decrease interobserver variability in the diagnosis of dysplasia? *Archives of pathology & laboratory medicine*, 129(2):174–176, 2005.

[87] Myrtle J van der Wel, Helen G Coleman, Jacques JGHM Bergman, Marnix Jansen, and Sybren L Meijer. Histopathologist features predictive of diagnostic concordance at expert level among a large international sample of pathologists diagnosing barrett's dysplasia using digital pathology. *Gut*, 69(5):811–822, 2020.

[88] Stephen S Raab and Dana M Grzybicki. Anatomic pathology workload and error, 2006.

[89] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[90] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[91] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[92] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[93] Sotiris B Kotsiantis. Supervised machine learning: A review of classification techniques. 2007.

[94] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[95] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24, 2019.

[96] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

[97] Anamika Dhillon and Gyanendra K Verma. Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, 9(2):85–112, 2020.

[98] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: An overview. *Frontiers in Medicine*, 6, 2019.

[99] Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan, and Guanghua Xiao. Pathology image analysis using segmentation deep learning algorithms. *The American journal of pathology*, 189(9):1686–1698, 2019.

[100] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[101] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019.

[102] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, pages 1–11, 2020.

[103] Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, pages 1–11, 2020.

[104] Rebecca Fitzgerald, M Di Pietro, M O'Donovan, M Maroni, Beth Muldrew, Irene Debiram-Beecham, M Gehrung, J Offman, M Tripathi, SG Smith, et al. Cytosponge-trefoil factor 3 versus usual care to identify barrett's oesophagus in a primary care setting: a prospective, multicentre, pragmatic, randomised controlled trial. *The Lancet*, 2020.

[105] James E Everhart and Constance E Ruhl. Burden of digestive diseases in the united states part i: overall and upper gastrointestinal diseases. *Gastroenterology*, 136(2):376–386, 2009.

[106] GR Locke 3rd, Nicholas J Talley, Sara L Fett, Alan R Zinsmeister, and LJ Melton 3rd. Prevalence and clinical spectrum of gastroesophageal reflux: a population-based study in olmsted county, minnesota. *Gastroenterology*, 112(5):1448–1456, 1997.

[107] Jesper Lagergren, Reinhold Bergström, Anders Lindgren, and Olof Nyrén. Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma. *New England journal of medicine*, 340(11):825–831, 1999.

[108] Gareth S Dulai, Sushovan Guha, Katherine L Kahn, Jeffrey Gornbein, and Wilfred M Weinstein. Preoperative prevalence of barrett's esophagus in esophageal adenocarcinoma: a systematic review. *Gastroenterology*, 122(1):26–33, 2002.

[109] Hashem B El-Serag, Aanand D Naik, Zhigang Duan, Mohammad Shakhatreh, Ashley Helm, Amita Pathak, Marilyn Hinojosa-Lindsey, Jason Hou, Theresa Nguyen, John Chen, et al. Surveillance endoscopy is associated with improved outcomes of oesophageal adenocarcinoma detected in patients with barrett's oesophagus. *Gut*, 65(8):1252–1260, 2016.

[110] K Visrodia, S Singh, R Krishnamoorthi, DA Ahlquist, KK Wang, Prasad G Iyer, and David A Katzka. Systematic review with meta-analysis: prevalent vs. incident oesophageal adenocarcinoma and high-grade dysplasia in barrett's oesophagus. *Alimentary pharmacology & therapeutics*, 44(8):775–784, 2016.

[111] Melina Arnold, Mark J Rutherford, Aude Bardot, Jacques Ferlay, Therese ML Andersson, Tor Åge Myklebust, Hanna Tervonen, Vicky Thursfield, David Ransom, Lorraine Shack, et al. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (icbp survmark-2): a population-based study. *The Lancet Oncology*, 20(11):1493–1505, 2019.

[112] Nicholas J Shaheen, Gary W Falk, Prasad G Iyer, and Lauren B Gerson. Acg clinical guideline: diagnosis and management of barrett's esophagus. *American Journal of Gastroenterology*, 111(1):30–50, 2016.

[113] NICE. Dyspepsia and gastro-oesophageal reflux disease: investigation and management of dyspepsia, symptoms suggestive of gastro-oesophageal reflux disease, or both. 2014.

[114] American Gastroenterological Association et al. American gastroenterological association medical position statement on the management of barrett's esophagus. *Gastroenterology*, 140(3):1084–1091, 2011.

[115] Mustafa Shawihdi, Elizabeth Thompson, Neil Kapoor, Geraint Powell, Richard P Sturgess, Nick Stern, Michael Roughton, Michael G Pearson, and Keith Bodger. Variation in gastroscopy rate in english general practice and outcome for oesophagogastric cancer: retrospective analysis of hospital episode statistics. *Gut*, 63(2):250–261, 2014.

[116] Thomas L Vaughan and Rebecca C Fitzgerald. Precision prevention of oesophageal adenocarcinoma. *Nature Reviews Gastroenterology & Hepatology*, 12(4):243–248, 2015.

[117] K Nadine Phoa, Frederike GI Van Vilsteren, Bas LAM Weusten, Raf Bisschops, Erik J Schoon, Krish Ragunath, Grant Fullarton, Massimiliano Di Pietro, Narayanasamy Ravi, Mike Visser, et al. Radiofrequency ablation vs endoscopic surveillance for patients with barrett esophagus and low-grade dysplasia: a randomized clinical trial. *Jama*, 311(12):1209–1217, 2014.

[118] Nicholas J Shaheen, Prateek Sharma, Bergein F Overholt, Herbert C Wolfsen, Richard E Sampliner, Kenneth K Wang, Joseph A Galanko, Mary P Bronner, John R Goldblum, Ana E Bennett, et al. Radiofrequency ablation in barrett's esophagus with dysplasia. *New England Journal of Medicine*, 360(22):2277–2288, 2009.

[119] Prateek Sharma, Nicholas J Shaheen, David Katzka, and Jacques JGHM Bergman. Aga clinical practice update on endoscopic treatment of barrett's esophagus with dysplasia and/or early cancer: expert review. *Gastroenterology*, 158(3):760–769, 2020.

[120] Sachin Wani, Dayna Early, Steve Edmundowicz, and Prateek Sharma. Management of high-grade dysplasia and intramucosal adenocarcinoma in barrett's esophagus. *Clinical Gastroenterology and Hepatology*, 10(7):704–711, 2012.

[121] Oliver Pech, Andrea May, Hendrik Manner, Angelika Behrens, Jürgen Pohl, Maren Weferling, Urs Hartmann, Nicola Manner, Josephus Huijsmans, Liebwin Gossner, et al. Long-term efficacy and safety of endoscopic resection for patients with mucosal adenocarcinoma of the esophagus. *Gastroenterology*, 146(3):652–660, 2014.

[122] Judith Offman, Beth Muldrew, Maria O'Donovan, Irene Debiram-Beecham, Francesca Pesola, Irene Kaimi, Samuel G Smith, Ashley Wilson, Zohrah Khan, Pierre Lao-Sirieix, et al. Barrett's oesophagus trial 3 (best3): study protocol for a randomised controlled trial comparing the cytosponge-tff3 test with usual care to facilitate the diagnosis of oesophageal pre-cancer in primary care patients with chronic acid reflux. *BMC cancer*, 18(1):784, 2018.

[123] Samir Gupta, Dan Li, Hashem B El Serag, Perica Davitkov, Osama Altayar, Shahnaz Sultan, Yngve Falck-Ytter, and Reem A Mustafa. Aga clinical practice guidelines on management of gastric intestinal metaplasia. *Gastroenterology*, 158(3):693–702, 2020.

[124] Matthew Banks, David Graham, Marnix Jansen, Takuji Gotoda, Sergio Coda, Massimiliano Di Pietro, Noriya Uedo, Pradeep Bhandari, D Mark Pritchard, Ernst J Kuipers, et al. British society of gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. *Gut*, 68(9):1545–1575, 2019.

[125] Stuart J Spechler and Rhonda F Souza. Barrett's esophagus. *The New England journal of medicine*, 371(9):836–845, 2014.

[126] Caryn S Ross-Innes, Hamza Chettouh, Achilleas Achilleos, Nuria Galeano-Dalmau, Irene Debiram-Beecham, Shona MacRae, Petros Fessas, Elaine Walker, Sibu Varghese, Theodore Evan, et al. Risk stratification of barrett's oesophagus using a non-endoscopic sampling method coupled with a biomarker panel: a cohort study. *The Lancet Gastroenterology & Hepatology*, 2(1):23–31, 2017.

[127] Prateek Sharma, Sravanthi Parasa, and Nicholas Shaheen. Developing quality metrics for upper endoscopy. *Gastroenterology*, 158(1):9–13, 2020.

[128] Nigel Hawkes. Cancer survival data emphasise importance of early diagnosis, 2019.

[129] Peggy R CyR. Atypical moles. *American family physician*, 78(6), 2008.

[130] Ian Talbot, Ashley Price, and Manuel Salto-Tellez. *Biopsy pathology in colorectal disease*. CRC Press, 2006.

[131] Raymond Maung. Pathologists' workload and patient safety. *Diagnostic Histopathology*, 22(8):283–287, 2016.

[132] Robert Odze. Histology of barrett's metaplasia: Do goblet cells matter? *Digestive diseases and sciences*, 63(8):2042–2051, 2018.

[133] Prateek Sharma, John Dent, David Armstrong, Jacques JGHM Bergman, Liebwin Gossner, Yoshio Hoshihara, Janusz A Jankowski, Ola Junghard, Lars Lundell, Guido NJ Tytgat, et al. The development and validation of an endoscopic grading system for barrett's esophagus: the prague c & m criteria. *Gastroenterology*, 131(5):1392–1399, 2006.

[134] Douglas S Levine, Rodger C Haggitt, Patricia L Blount, Peter S Rabinovitch, Valerie W Rusch, and Brian J Reid. An endoscopic biopsy protocol can differentiate high-grade dysplasia from early adenocarcinoma in barrett's esophagus. *Gastroenterology*, 105(1):40–50, 1993.

[135] Computation Pathology Group, part of the Diagnostic Image Analysis Group, at the Radboud University Medical Center. Asap.

[136] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[137] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[138] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[139] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[140] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[141] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[142] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[143] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[144] Xinqing Fan and Ned Snyder. Prevalence of barrett's esophagus in patients with or without gerd symptoms: role of race, age, and gender. *Digestive diseases and sciences*, 54(3):572–577, 2009.

[145] Douglas K Rex, Oscar W Cummings, Michael Shaw, Mark D Cumings, Roy KH Wong, Raj S Vasudeva, Donal Dunne, Emad Y Rahmani, and Debra J Helper. Screening for barrett's esophagus in colonoscopy patients with and without heartburn. *Gastroenterology*, 125(6):1670–1677, 2003.

[146] JL Herrera Elizondo, R Monreal Robles, D García Compean, EI González Moreno, OD Borjas Almaguer, HJ Maldonado Garza, and JA González González. Prevalence of barrett's esophagus: An observational study from a gastroenterology clinic. *Revista de Gastroenterología de México (English Edition)*, 82(4):296–300, 2017.

[147] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[148] Babak Ehteshami Bejnordi, Nadya Timofeeva, Irene Otte-Höller, Nico Karssemeijer, and Jeroen AWM van der Laak. Quantitative analysis of stain variability in histology slides and an algorithm for standardization. In *Medical Imaging 2014: Digital Pathology*, volume 9041, page 904108. International Society for Optics and Photonics, 2014.

[149] Brady Kieffer, Morteza Babaie, Shivam Kalra, and Hamid R Tizhoosh. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2017.

[150] Nora Pashayan and Paul DP Pharoah. The challenge of early detection in cancer. *Science*, 368(6491):589–590, 2020.

[151] DELTA project. https://www.deltaproject.org. Accessed: 2020-09-13.

[152] Lucas C Duits, Pierre Lao-Sirieix, W Asher Wolf, Maria O'Donovan, Nuria Galeano-Dalmau, Sybren L Meijer, G Johan A Offerhaus, James Redman, Jason Crawte, Sebastian Zeki, et al. A biomarker panel predicts progression of barrett's esophagus to esophageal adenocarcinoma. *Diseases of the Esophagus*, 32(1):doy102, 2019.

[153] Sarah Killcoyne, Eleanor Gregson, David C. Wedge, Dan J. Woodcock, Matthew Eldridge, Rachel de la Rue, Ahmad Miremadi, Sujath Abbas, Adrienn Blasko, Wladyslaw Januszewicz, Aikaterini Varanou Jenkins, Moritz Gerstung, and Rebecca C. Fitzgerald. Genomic copy number predicts oesophageal cancer years before transformation. *bioRxiv*, 2020.

[154] Massimiliano di Pietro, Ines Modolell, Maria O'Donovan, Catherine Price, Nastazja D Pilonis, Irene Debiram-Beecham, and Rebecca C Fitzgerald. Use of cytosponge as a triaging tool to upper gastrointestinal endoscopy during the covid-19 pandemic. *The Lancet Gastroenterology & Hepatology*, 5(9):805–806, 2020.

[155] Tim Bezemer, Mark CH De Groot, Enja Blasse, Maarten J Ten Berg, Teus H Kappen, Annelien L Bredenoord, Wouter W Van Solinge, Imo E Hoefer, and Saskia Haitjema. A human (e) factor in clinical decision support systems. *Journal of medical Internet research*, 21(3):e11732, 2019.

[156] Zizhao Zhang, Pingjun Chen, Mason McGough, Fuyong Xing, Chunbao Wang, Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui, Jasreman Dhillon, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245, 2019.

[157] Yahui Jiang, Meng Yang, Shuhao Wang, Xiangchun Li, and Yan Sun. Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Communications*, 40(4):154–166, 2020.

[158] Michael Wainberg, Daniele Merico, Andrew Delong, and Brendan J Frey. Deep learning in biomedicine. *Nature biotechnology*, 36(9):829–838, 2018.

[159] Zhongren Zhou, Irina Kalatskaya, Donna Russell, Norman Marcon, Maria Cirocco, Paul M Krzyzanowski, Cathy Streutker, Hua Liang, Virginia R Litle, Tony E Godfrey, et al. Combined esophacap cytology and muc2 immunohistochemistry for screening of intestinal metaplasia, dysplasia and carcinoma. *Clinical and experimental gastroenterology*, 12:219, 2019.

[160] Zhixiong Wang, Swetha Kambhampati, Yulan Cheng, Ke Ma, Cem Simsek, Alan H Tieu, John M Abraham, Xi Liu, Vishnu Prasath, Mark Duncan, et al. Methylation biomarker panel performance in esophacap cytology samples for diagnosing barrett's esophagus: a prospective validation study. *Clinical Cancer Research*, 25(7):2127–2135, 2019.

[161] Jonghee Yoon, James Joseph, Dale J Waterhouse, A Siri Luthman, George SD Gordon, Massimiliano Di Pietro, Wladyslaw Januszewicz, Rebecca C Fitzgerald, and Sarah E Bohndiek. A clinically translatable hyperspectral endoscopy (hyse) system for imaging the gastrointestinal tract. *Nature communications*, 10(1):1–13, 2019.

[162] Alanna Ebigbo, Robert Mendel, Andreas Probst, Johannes Manzeneder, Friederike Prinz, Luis A de Souza Jr, Joao Papa, Christoph Palm, and Helmut Messmann. Real-time use of artificial intelligence in the evaluation of cancer in barrett's oesophagus. *Gut*, 69(4):615–616, 2020.

[163] Sarah Killcoyne, Eleanor Gregson, David C Wedge, Dan J Woodcock, Matthew Eldridge, Rachel de la Rue, Ahmad Miremadi, Sujath Abbas, Adrienn Blasko, Wladyslaw Januszewicz, et al. Genomic copy number predicts oesophageal cancer years before transformation. *Nature Medicine*, 2020.

[164] Ross J Porter, Graeme I Murray, Daniel P Brice, Russell D Petty, and Mairi H McLean. Novel biomarkers for risk stratification of barrett's oesophagus associated neoplastic progression–epithelial hmgb1 expression and stromal lymphocytic phenotype. *British Journal of Cancer*, 122(4):545–554, 2020.

[165] Mamoun Younes, Keith Brown, Gregory Y Lauwers, Gulchin Ergun, Frank Meriano, A Carl Schmulen, Alberto Barroso, and Atilla Ertan. p53 protein accumulation predicts malignant progression in barrett's metaplasia: a prospective study of 275 patients. *Histopathology*, 71(1):27–33, 2017.