Investigating the potential of real-world data to improve outcomes for patients with lung cancer

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Biology, Medicine and Health

2023

Isabella M Fornacon-Wood
School of Medical Sciences

Contents

Word count 46,318

List of Tables

List of Figures

List of Abbreviations

AUC	Area Under the Curve
CI	Concordance Index
CBCT	Cone Beam Computed Tomography
CT	Computed Tomography
CTV	Clinical Target Volume
DFS	Disease Free Survival
DM	Distant Metastasis
ECOG	Eastern Cooperative Oncology Group
EHR	Electronic Health Record
FDA	Food and Drug Administration
GRD	Gross Residual Disease
GTV	Gross Tumour Volume
H&N	Head and Neck
HR	Hazard Ratio
IBSI	Imaging Biomarker Standardization Initiative
ICC	Intraclass Correlation Coefficient
IGRT	Image-Guided Radiotherapy
iGTV	Internal Gross Tumour Volume
IMRT	Intensity Modulated Radiotherapy
LHS	Learning Healthcare System
LINAC	Linear Accelerator
LR	Local Relapse
LRR	Local Regional Recurrence
LR-RFS	Loco-Regional Recurrence-Free Survival

MIP	Maximum Intensity Projection
MLC	Multileaf Collimator
MRI	Magnetic Resonance Imaging
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NSCLC	Non-small Cell Lung Cancer
OR	Odds Ratio
OS	Overall Survival
PCI	Prophylactic Cranial Irradiation
pCR	Pathological Complete Response
pCT	Radiotherapy Planning CT scan
PET	Positron Emission Tomography
PFS	Recurrence Free Survival
PTV	Planning Target Volume
RCT	Randomized Control Trial
RFS	Recurrence Free Survival
ROI	Region Of Interest
RQS	Radiomics Quality Score
SCLC	Small Cell Lung Cancer
TNM	TNM Classification of Malignant Tumours
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
VMAT	Volumetric Modulated Arc Therapy
3D-CRT	3D Conformal Radiotherapy
95% CI	95% Confidence Interval

Abstract

There is increasing interest in using real-world data, routinely collected data relating to the health or healthcare delivery of patients, to generate evidence that has the potential to alter clinical decision making. Such real-world evidence could help to fill gaps in clinical knowledge, particularly for patients under-represented in clinical trials and for changes in radiotherapy workflows which occur as technology and techniques advance, often without clinical evidence to support the potential benefits. This thesis investigates the potential of real-world data to improve outcomes for patients with lung cancer through the analysis of routinely collected clinical and imaging data.

First, the radiomics literature was reviewed to assess whether radiomics had the potential to personalise lung cancer treatment. The reviewed literature suffered from significant limitations, and no single radiomics biomarker or methodological approach was used widely, suggesting substantial barriers to clinical translation remain.

Next, the reliability of radiomic features was assessed across four feature extraction platforms. It was found that choice of feature extraction platform, Imaging Biomarker Standardisation Initiative (IBSI) compliance, parameter settings and platform version affected feature reliability. This highlights the difficulty in trusting radiomics biomarkers, and the importance of using the latest version of an IBSI compliant software to ensure reproducibility of radiomics, a key requirement for clinical translation.

The potential of real-world clinical data was then evaluated in the context of various retrospective changes to practice. First, the introduction of Intensity-modulated radiotherapy (IMRT) at The Christie NHS Foundation Trust was investigated, finding that the proportion of patients treated with curative-intent radiotherapy had increased and patient survival had improved following the introduction of IMRT. Second, the impact of the COVID-19 pandemic on outcomes for patients with lung cancer was evaluated, finding that patients who had a change to their radiotherapy or chemotherapy treatment did not have significantly worse survival or relapse rates compared to patients whose treatments were not changed; however, patients who had a change to their radiotherapy did have increased odds of $\geq$ grade 3 acute toxicity.

Finally, the potential of Bayesian methodology for assessing changes to clinical practice was investigated. A Bayesian analysis of a change to image-guided radiotherapy protocol found a reduced hazard of death for patients who had residual set-up errors towards the heart post-protocol change. This suggests the potential for Bayesian methodology to evaluate prospective incremental changes to practice.

Together, these results demonstrate the potential real-world datasets have to monitor and improve outcomes for patients with lung cancer.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, the University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations) and in the University's policy on Presentation of Theses.

Acknowledgements

I would like to thank my brilliant supervisors, Gareth Price, Corinne Faivre-Finn and James O'Connor, for supporting and guiding me through my PhD. Thank you for encouraging me to pursue my own areas of interest and providing plenty of opportunities to get involved with interesting research projects. I have learnt so much from all of you and look forward to continuing to do meaningful work in my new role as a postdoc.

Thanks to the entire RRR department for keeping work fun and social, even through the many lockdowns and working from home. To my office friends, thank you for the fun nights out and endless conversations about data issues with a cup of Yorkshire tea and M&S biscuits.

I would like to thank the MCRC for funding my PhD.

A special thank you to my whole family, especially my parents, for your love and encouragement every day. Your support through the pandemic and having a baby helped me to finish this PhD; I couldn't have done it without you.

Finally, a special thank you to Jack, my rock, for your endless love and support. And to Heidi, you put a smile on my face every day. This thesis is dedicated to you, a reminder you can do anything you set your mind to.

The Author

- BSc Physics, Universität Leipzig, 2017.
- MRes Translational Medicine, The University of Manchester, 2018.

Rationale

This thesis is presented in alternative format due to having published, or intending to publish, the studies in peer-reviewed journals. This thesis has been constructed by including the studies chronologically, each investigating different aspects of a Learning Healthcare System. This starts with generating evidence for clinical decision making, to retrospectively investigating changes to clinical practice. The thesis therefore tells a coherent story, and critically analyses the work and methods used, as required. My contribution to each of the studies is included in a page before each chapter.

Chapter 1

Introduction

1.1 Radiotherapy

Soon after Röntgen's discovery of x-rays in 1895 and Curie's discovery of radium in 1898, methods were developed to use ionising radiation in the treatment of cancer [1]. Initially, x-rays were used to treat skin malignancies, while radium was inserted directly into tumours in the first use of brachytherapy. A key breakthrough in using radiation for cancer was discovered in 1939 by Henri Coutard, who found that treating patients with lower doses over a longer period of time, rather than a high dose all at once, decreased side effects while improving control of the cancer [2]. He noted there was a fine line between the energy that would cure versus harm a patient. Technological advances throughout the 20th century have led to sophisticated machinery capable of delivering high energy x-ray beams to solid tumours, known as radiotherapy, while progress in radiobiology has helped to increase knowledge of the effects of radiation on tumours and healthy tissue [3].

Radiotherapy plays an essential role in the treatment of over 50% of patients with cancer [4, 5]. According to a report published by the Department of Health Cancer Policy Team, out of all patients in the UK who are cured of their cancer, 40% will have had radiotherapy as a part of their treatment and 16% will have been cured through radiotherapy alone [6]. In North America, it is estimated that 29% of cancer survivors have been treated with radiotherapy [7]. Radiotherapy works by using ionising radiation to

kill cancerous cells through DNA damage [8]. Radiation damage does not discriminate between healthy and cancerous tissue, so great care must be taken to avoid healthy tissue and organs near the tumour, so called 'organs at risk' (OARs). Fractionation, splitting the total radiotherapy dose into fractions and delivering these over multiple days or weeks, allows maximum destruction of cancerous cells whilst minimising destruction to the surrounding tissue. This happens through DNA damage; healthy cells are more able to repair from DNA damage than malignant cells, so over the course of a treatment the irradiated healthy tissue has time to repair, while more malignant cells are destroyed each time [9].

Radiotherapy begins with a tailored treatment plan. A radiotherapy planning scan, usually a Computed Tomography (CT) scan, is taken in the position the patient would be in for the radiotherapy treatment. The visible tumour is then manually delineated by a radiation oncologist, along with nearby organs at risk. Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI) scans can give complementary information to the delineating clinician on disease context and spread. The delineated tumour is known as the Gross Tumour Volume (GTV). The GTV margins are then expanded to create the Clinical Target Volume (CTV), which is based on the probability of malignant cells being outside the GTV and hopes to incorporate areas containing microscopic disease. The Planning Target Volume (PTV) expands the margins further to take into account anatomical motion, uncertainties and margin errors to ensure the prescribed dose is being delivered to the CTV [10]. Each contour is presented in Figure 1.1. In the case of tumours that are affected by motion from breathing, for example lung tumours, 4D-CT scans are taken which take multiple 3D scans over the course of a breathing cycle, allowing tumour motion to be incorporated into the radiotherapy planning process [11]. The Internal Gross Tumour Volume (iGTV) compromises of the GTV and the extent of its motion over the breathing cycle. This is contoured by either combining the individual GTV contours on each image of the breathing cycle, or contouring directly on the Maximum Intensity Projection (MIP) image, an image created from the maximum intensity values from each image.

The radiotherapy delivery is optimised using beams from different angels to ensure maximum dose is given to the target volume and minimum dose is received by the

surrounding organs at risk. The radiotherapy beam is generated using a linear accelerator (LINAC) and is shaped using a multileaf collimator (MLC) in the head of the LINAC. The MLC consists of tungsten leaves that can move to block parts of the beam to generate any beam shape desired [12]. Radiotherapy delivered in this way is known as 3D conformal radiotherapy.

Figure 1.1: The GTV in red encloses the visible tumour, the CTV in green extends the GTV to include microscopic spread of the tumour and the PTV in yellow ensures the maximum dose is delivered to the CTV.

1.1.1 Intensity-modulated radiotherapy

A key advancement in the field of radiotherapy is the development of intensity-modulated radiotherapy (IMRT). With this treatment a radiotherapy dose is prescribed to the tumour and a maximum dose to surrounding organs is defined, then beam delivery is optimised around these parameters (called inverse planning) by dynamically moving the MLC leaves during delivery to generate intensity-modulated fields allowing a sharper dose fall-off and less dose to surrounding organs at risk [13, 14]. Volumetric modulated arc radiotherapy (VMAT) is a newer form of IMRT where the machine rotates around the patient in an arc shape and continuously changes the shape of the beams to conform around the tumour from all directions, significantly decreasing treatment time. IMRT techniques allow higher radiation doses to be delivered to the tumour while sparing organs at risk to give a greater chance of treatment response with minimal toxicity.

1.1.2 Image-guided radiotherapy

Imaging machines can be integrated into LINACs to allow patients to be scanned before treatment. This is called image-guided radiotherapy (IGRT) and has become

standard of care in the UK [15]. Immediately prior to treatment, a cone beam CT (CBCT) scan is taken and the tumour position in the image is matched to the planning CT scan [16]. CBCT scans use diverging x-rays to provide a volumetric image of the patient's tumour, using a lower radiation dose than a traditional CT scan. 4D CBCT scans acquire respiratory motion data during the imaging process, allowing tumour motion to be taken into account. The CBCT or 4D CBCT image is then aligned with the planning CT scan, whether 3D or 4D, and the difference in the patient's position determined. If the patient's position has moved significantly from the planning position, the distance between them termed a set-up error, the treatment couch can be adjusted to ensure the radiotherapy beams are reaching the tumour. If large anatomical changes are detected, reactive adaptive re-planning is required to ensure the tumour is receiving the correct dose. IGRT reduces radiation damage to healthy tissue caused by set-up errors, and ensures, as much as possible, the full dose is actually being delivered to the PTV as planned.

1.1.3 Changes to radiotherapy workflows

IMRT and IGRT are examples of advances in radiotherapy that were implemented without evidence of clinical benefit from Randomised Control Trials (RCTs). Changes to radiotherapy workflows happen often as technology and techniques advance, often without formal evaluation as there is an assumed benefit to technological advancements based on biological or physical characteristics. This in itself makes it difficult to evaluate technological changes in a RCT as it would be unethical to randomise patients to a potentially lesser treatment when there is biological or dosimetric evidence suggesting the newer technique is superior [17]. For example, IGRT was implemented without RCT evidence as it is difficult to argue that there is a clinical equipoise between patients who are imaged before treatment to ensure the tumour is in the planning position, and patients who are not imaged. In the case of IMRT, dosimetric studies revealed it can achieve better dose conformality [18, 19], so it was implemented with an assumed clinical benefit rather than RCT evidence. Furthermore, RCTs are time consuming; it takes years to develop, recruit and follow-up patients to finally determine any benefit of technical changes and during that time newer advancements may be available making the original reason for the trial obsolete. These complexities

in evaluating technical changes to radiotherapy practice mean there is a distinct lack of evidence on the impact of such changes on patient outcomes [20].

1.2 Lung cancer

Lung cancer is the leading cause of cancer deaths worldwide, accounting for 18% of all cancer deaths [21]. Over half of patients die within one year of diagnosis and the 5-year survival is only 8% for men and 12% for women [22]. There had been little improvement in survival for patients with lung cancer since the 1970s [22]; however, recent advances in immunotherapy for patients without actionable mutations and tyrosine kinase inhibitors for patients with actionable mutations have led to improved survival rates [23–25]. Many patients are diagnosed with advanced disease which is more difficult to treat [26]. This is in part due to the symptoms of lung cancer generally presenting when the disease is in the later stages and even when symptoms present earlier, people delay seeing their GP as symptoms are non-specific, such as coughing and chest pain, or smoking-related side effects [27]. Lung cancer screening initiatives have been rolled out in the UK for those at high risk of the disease, after pilot initiatives led to many lung cancers being diagnosed at early stages [28]. There is a high association between lung cancer mortality rates and social deprivation in the UK [29], likely due to the high prevalence of smoking in deprived areas [30], unhealthy lifestyles and a lack of symptom awareness.

There are two types of lung cancer, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). 85% of all lung cancer cases present as NSCLC. There are three main types of NSCLC: adenocarcinoma, squamous cell carcinoma and large cell carcinoma. Early stage NSCLC can be treated with surgery or, if surgery is deemed too dangerous for the patient, radical radiotherapy. Stereotactic ablative radiotherapy (SABR) is hypofractionated, larger doses given per fraction, and prescribed for small tumours (up to 5 cm) that are localised and can be precisely targeted by the radiation whilst avoiding damage to the surrounding tissue. It can achieve a local control rate similar to that of surgery [31]. If the patient is not eligible for SABR or surgery, for example if the patient is frail and suffers from many comorbidities, but the disease is early stage, conventional radiotherapy can be prescribed (i.e. 30-33 Gy in 60-66

fractions), or hypofractionated radiotherapy (i.e. 60 Gy in 15 fractions). Most cases of NSCLC present at a locally advanced stage and treatment depends on various criteria including comorbidities, tumour volume, dose to healthy tissue and the fitness of the patient. Patient fitness is characterised by the performance status, such as the Eastern Cooperative Oncology Group (ECOG) performance status, which scores a patient from 0 to 5, 0 meaning the patient is fully active and needs no help in daily activities, 1 meaning the patient cannot perform physically strenuous activities but is still able to carry out light work, to 5 meaning the patient is dead. The fittest patients with a performance score of 0/1 are prescribed the gold standard treatment of concurrent chemo-radiotherapy if their cancer is inoperable, as they are deemed fit enough to withstand the toxicities associated with receiving radiotherapy and chemotherapy concurrently [32]. This is then followed by consolidation immunotherapy if they meet the response criteria. Lung and kidney function are assessed before treatment to ensure the patient can tolerate the toxicities [33]. If a patient's tumour is particularly large or they are not fit enough to withstand a concurrent regimen, sequential chemo-radiotherapy is preferred. If the patient suffers from poor renal function or is too ill for chemotherapy, radiotherapy alone is prescribed. For metastatic NSCLC, radiotherapy, chemotherapy, immunotherapy and targeted agents can be prescribed for palliative treatment.

Small cell lung cancer is the rarer, more aggressive form of lung cancer that typically presents in the advanced stages and has a median survival of 7 months with treatment [34]. Incidence of SCLC has decreased over the years, likely due to an increase in smoking cessation, although in developing countries incidence is the same or increasing [35]. SCLC has historically been classified as either limited or extensive in stage, where extensive means the cancer has spread outside the thorax. Standard of care for limited stage SCLC is concurrent chemo-radiotherapy if the patient is fit, if not then sequential chemo-radiotherapy or radiotherapy alone. For extensive stage SCLC, chemotherapy and immunotherapy is considered for fit patients, otherwise chemotherapy alone. Chemotherapy is prescribed for limited as well as extensive stage SCLC due to the high risk of micro-metastasis. SCLC responds well to initial treatment of

chemo-radiotherapy; however, it has a very high relapse risk and risk of brain metastasis. Prophylactic cranial irradiation (PCI) is given as standard of care for patients with all stages of SCLC as it reduces risk of brain metastasis and relapse [36]. However, recent evidence is challenging the use of PCI, particularly in the extensive stage setting [37].

1.3 Real-world data

Patients with lung cancer tend to be old, frail, and have a high comorbidity burden. Such patients are typically under-represented in RCTs [38, 39]. It is vital that clear evidence is generated for these patients to ensure they are getting the best treatments.

Real-world data are defined by the Food and Drug Administration (FDA) as routinely collected data relating to the health or healthcare delivery of patients from various sources [40]. Real-world data have the potential to provide evidence for patients typically excluded or under-represented in RCTs, as data are captured for every patient. Real-world data are often though of as the clinical data captured in patient's electronic health records, but also includes the large volumes of imaging data patients accrue during their cancer care.

1.3.1 Real-world clinical data

Clinical data are captured in patient's Electronic Health Records (EHRs), documenting their healthcare over their lifetime. Data are captured on patient demographics, diagnoses, treatments and side effects, laboratory tests, imaging reports, hospital episodes, outcomes and more. EHRs are used worldwide to document and store healthcare data with the aim of supporting continuity of care [41]. In the UK, the implementation of EHRs has been driven by governmental initiatives and financial investments in the last two decades [42], and the current NHS Long Term Plan aims for all hospitals to become fully digitised so clinicians can interact with care records where ever they are [43]. There is increasing interest in using the large, rich datasets derived from EHR in research, for example to determine incidence/prevalence of a disease, find potential

risk factors or improve the quality of services [44]. As data are captured for every patient, there is the potential to develop representative clinical models that describe and benefit the entire population. Real-world evidence generated from real-world data can complement evidence from RCTs.

1.3.2 Real-world imaging data

Imaging is used routinely in lung cancer management, helping to determine diagnosis, prognosis and predict the optimal treatment for each patient. Imaging allows visualisation of morphological characteristics of the tumour that can qualitatively help to determine the stage of the cancer, by identifying and quantifying the extent to which the tumour has spread to lymph nodes or other organs. Imaging is non-invasive and is performed throughout the treatment pathway to monitor response to treatment. Imaging biomarkers, biological features detected from medical images, can help inform treatment decisions and are indispensable in oncology [45]. For example, TNM Classification of Malignant Tumours (TNM) staging is used worldwide to classify solid tumours by taking into account tumour size, invasion of nearby tissue, involved lymph nodes and presence of metastasis. These factors can all be determined through various imaging techniques and in combination become a prognostic factor that helps the clinician to tailor a treatment specific to the patient. Finding imaging biomarkers beyond TNM that help stratify patient treatment options is crucial in this era of personalised medicine, particularly in lung cancer where images are taken routinely throughout the care pathway and improvements in survival have been limited to small patient groups where immunotherapy has been shown to be beneficial.

Radiomics

Imaging biomarkers have been used in health care for decades, however over the last 10 years or so the concept and work flow of extracting image-based features from medical images has blossomed. Radiomics is the extraction of numerous quantitative features from medical images to quantify tumour phenotypes [46]. In contrast to imaging biomarker studies where features of interest are chosen *a priori*, radiomics is a data-driven approach where statistical methods are used to find the features most correlated to the measure of interest. These image-based features can be combined with

patient characteristics and other biomarkers, such as genomics and pathology, to create individualised predictive models of outcome. The rationale behind creating a radiographic tumour phenotype is that medical images harbour information on underlying pathology which is not revealed by qualitative assessment [47]. If no particular imaging biomarker is of interest to study, then radiomics can be a hypothesis-generating approach to finding features that are correlated to a measure of interest.

Studies involving radiomics have increased exponentially in the last few years. The workflow for radiomics analyses is described in detail in Chapter 2 and shown in Figure 1.2. Briefly, images of tumours are first acquired and reconstructed. The visible tumour is then contoured and pre-processing steps can be performed on the image such as discretisation, binning voxel intensities to reduce the total number of intensity values in the image, or filtering. Features are then extracted; this can include thousands of features describing the tumour's shape, texture and intensity. The final steps involve data analysis and statistical modelling techniques to correlate the features to clinical endpoints such as survival or treatment response, or biological endpoints such as genetic mutational status or histology.

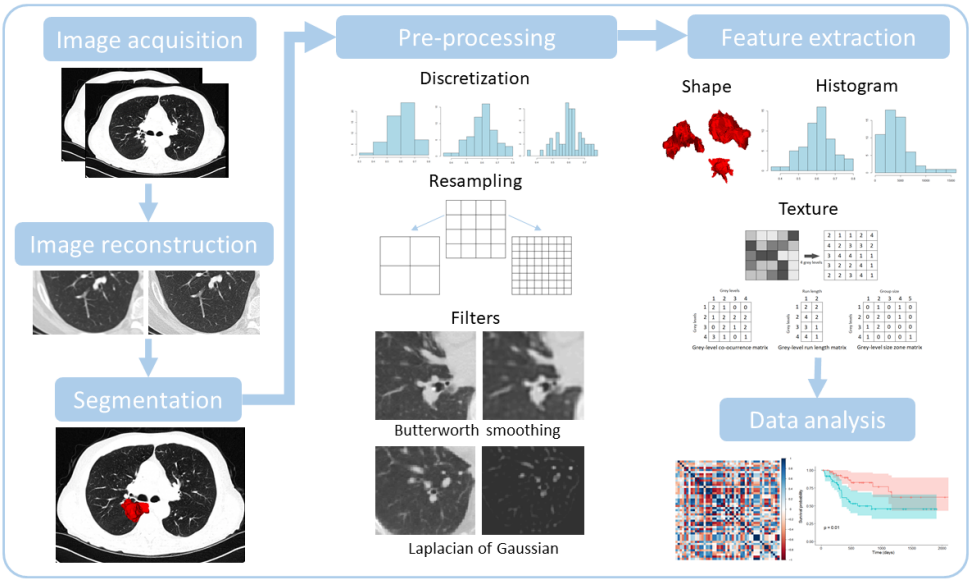

Figure 1.2: Visualisation of the steps in the radiomics workflow. First, images are acquired and reconstructed. The region of interest is then segmented, from which features will be extracted. Next, pre-processing steps are performed to modify the images before feature extraction. Shape, first order (or histogram) and texture features are then extracted from the region of interest. Finally, data analysis steps attempt to find correlations between features and the specified outcome.

Many limitations are inherent at each step of the radiomics workflow, discussed in detail in Chapters 2 and 3. Key issues include the lack of standardisation with regards to image acquisition, image pre-processing and feature extraction which has been found to affect feature repeatability and reproducibility [48]. For example, features extracted from images of the same phantom acquired on different CT scanners are not always reproducible [49, 50]; in one study the variability between feature values was comparable to the inter-patient variability [50]. The Imaging Biomarker Standardisation Initiative (IBSI) have done extensive work to standardise the radiomics feature extraction process by developing a guideline for feature nomenclature and definitions, and publishing a digital phantom and benchmark datasets so software developers can benchmark their feature values to be IBSI compliant, promoting reproducibility [51, 52].

There are vast amounts of imaging data available in the field of oncology, and there is potential to harness the information within using radiomics. Individualised statistical models created using real-world imaging and clinical data could help to identify potential risk factors, and ultimately lead to changes being made that may improve outcomes for patients with cancer.

1.3.3 Why are we interested in real-world data?

RCTs are the gold standard in evidence-based medicine for evaluating the effectiveness of treatments. Their design ensures results are not affected by confounding, unmeasured variables that influence both cause and effect, as patients are randomised between treatment options on entry into the trial. This allows a reliable cause and effect relationship between the intervention and outcome to be established.

While RCTs are a powerful tool for providing evidence of treatment effectiveness, there are cases where RCT evidence does not exist. For example, RCT evidence generally does not exist in cases where there is a lack of clinical equipoise, i.e. genuine uncertainty as to whether one intervention is better than another. As discussed in Section 1.1.3, technical changes to radiotherapy treatments are often implemented without RCT evidence, as technological advances are assumed to lead to superior treatments. While there may be dosimetric or mathematical evidence suggesting the

advancement is superior, ensuring patient outcomes are also improving, or at least not getting worse due to a perhaps unforeseen complication, is also important. Real-world data are vast, and the large, rich datasets can be used to investigate such clinical questions and generate real-world evidence to fill the evidence-gap when RCT data do not exist. Real-world data, therefore, have the potential to ensure new advances are actually effective, and are leading to better outcomes for patients [53].

RCTs are highly selective and tend to exclude, or at least under-represent, patients who are old, frail and have multiple comorbidities [38, 39]. One study found that 60% of RCT study eligibility criteria relate to comorbidity or performance status [54]. It is therefore questionable whether results from RCTs have external validity, i.e. that results are generalizable to the rest of the patient population [55]. A key advantage of real-world data is that data exist for all patients, including those who are old, frail and have multiple comorbidities. Real-world datasets can be used to check for real-world effectiveness of interventions, particularly for patients under-represented in RCTs. Furthermore, RCTs tend to have short follow-up periods, so long-term effects of interventions may not be captured. As real-world data derived from EHRs are longitudinal in nature, analysis of long-term follow-up data could reveal rare adverse events that otherwise may not be captured. This of course depends on high quality, long-term data being collected, which can be difficult in practice.

Many treatment decisions made in radiation oncology are done so without RCT evidence to back them [56]. Real-world evidence generated from real-world data could help to provide evidence where RCT evidence does not, and will not, exist.

1.3.4 The challenges of real-world data

The main challenge of using real-world data to generate evidence is the fact data are collected for clinical use and not necessarily for research or analytical purposes. This means the quality of the data is less standardised and potentially inferior to that from RCTs, where data entry is well defined and controlled. Real-world data are often incomplete, inaccurate, inconsistent and care records can be dis-jointed making it difficult to merge databases from different sources. Laboratory and imaging results may not be comparable across patients due to different protocols and techniques used.

Retrospective analysis of real-world data suffer with these issues the most, as prospective data collection can mitigate these issues by clearly defining how data should be collected.

Developing evidence of both treatment benefit and risk in the real-world setting requires high quality data on outcomes such as survival, local and distant relapse, as well as long term toxicity and quality of life metrics, which are not always available. The implementation of electronic Patient-Reported Outcome Measures (ePROMs) for patients with cancer treated at the Christie NHS Foundation Trust enables collection of data on patient symptoms and quality of life, an important aspect of treatments not captured in most EHRs [57]. As well as capturing high quality data, data must be captured in a structured way so that they can be easily analysed. Unstructured data are captured as free-text in clinical notes and reports, or imaging data for example. It is estimated that 80% of healthcare data are unstructured [58] and are therefore unusable in analyses without manual curation or the development of tools to be able to convert the data to be amenable for computer processing.

Missing data are a further issue associated with real-world data, and when missing data are informative, i.e. there is a reason the data are missing and they are not missing at random, this can introduce bias into an analysis. For example, missing data can be informative if a particular test is only done on patients with severe disease. If you included the data from this particular test in your analysis, your results would be biased as only patients with severe disease would have data for that test. Methods to overcome missing data include only using complete cases, imputing missing data, and carrying observations forward in the case of longitudinal data [59]. When missing data are informative, only using complete cases can introduce bias and reduce statistical power by reducing the size of the available dataset.

Real-world data are observational by nature, i.e. non-interventional, and as such real-world evidence can suffer from different types of bias. Selection bias occurs when selection of participants into a study is not representative of the wider population of interest. This is a particular issue for retrospective observational studies as patient inclusion into an analysis is heavily dependent on clinical decisions made in practice

which could lead to a biased result. For example, a study including patients with lung cancer who received immunotherapy would have to deal the bias that may arise from the fact that only patients who responded well to initial chemo-radiotherapy and had a good performance status would have been offered adjuvant immunotherapy, and therefore any results from that study would not be applicable to patients who did not respond to chemo-radiotherapy or had poor PS. Selection bias compromises external validity. Confounding occurs when there are systematic differences between baseline variables of the comparison groups that affect outcome. RCTs mitigate this bias by randomising patients on entry to the study, but observational data do not have this advantage and as such choice of a particular treatment or intervention could be correlated to the outcome of interest. For example, comparing the survival of patients with lung cancer who received concurrent versus sequential chemo-radiotherapy would be confounded by the fact that concurrent chemo-radiotherapy is only offered to the fittest patients. The concurrent group would therefore be expected to live longer in any case due to better underlying health, rather than purely due to the concurrent treatment. Methods to mitigate confounding include adjusting for known confounders in multivariable analyses or using propensity score matching [60]. However, there is always the limitation that unknown confounders could be influencing the results and compromising internal validity. It is therefore important to include multidisciplinary teams when designing and analysing an observational study with real-world data to ensure all confounders are being taken into account and there is in-depth knowledge of how treatment decisions are made.

1.3.5 The potential of real-world data

RCTs and real-world evidence are not mutually exclusive. The randomisation from RCTs is vital to ensure validity of results; however, studies using real-world data can provide complimentary evidence. Moreover, randomised trials can be done using real-world data to generate real-world evidence in what is known as a pragmatic trial. Pragmatic trials use randomisation in a real-world setting to evaluate interventions, as opposed to RCTs which use optimal settings [61]. They allow the effectiveness of treatments to be tested in the real-world population. The importance and potential of

real-world EHR data was highlighted during the COVID-19 pandemic. The RECOVERY trial, a pragmatic trial, found the first effective treatment against COVID-19, an inexpensive steroid called dexamethasone which reduced deaths by up to a third, saving countless lives [62]. It had a simple design with minimal data entry requirements, making it different to standard RCTs, which was necessary at the time to ensure minimal burden to an already overworked and understaffed workforce. Staff were only required to submit essential information, and then routinely collected data available in EHRs were linked to the trial data to complete the database. This allowed rapid and reliable results to be discovered. The RECOVERY trial highlights the potential real-world data have to generate real-world evidence in pragmatic trials and allow the discovery of safe, effective treatments that will improve the lives of patients.

The potential of real-world data is being recognised. In 2022, the Department of Health and Social Care commissioned an independent review into improving the safety and security of healthcare data used in research, and how these data can be harnessed to improve the lives of patients [63]. The findings from the review, named the Goldacre report, have shaped the Health and Social Care Data Strategy published in June 2022, 'Data saves lives: reshaping health and social care with data' [64], which describes a vision and a roadmap to make better use of NHS data to save lives. The NHS long term plan also sets out an aim to drive digital transformation within the NHS, including making clinical data available for research and ensuring clinical records are all digitised [43]. The National Institute for Health and Care Excellence (NICE) have ambitions to include real-world data in the development of their evidence-based guidelines [65].

1.4 Learning Healthcare Systems

Real-world data have the potential to fill current gaps in clinical knowledge, particularly in the case of technical changes to radiotherapy workflows. A Learning Healthcare System (LHS) is a potential environment within which technical changes to practice could be evaluated [66]. In the LHS concept, routine, real-world data collated in patient's EHRs are analysed to monitor outcomes and identify areas for improvement.

Changes to practice are implemented and learning cycles monitor the changes in order to improve healthcare delivery [67]. The learning cycle is pictured in Figure 1.3. Embedding evidence generation and learning cycles into clinical practice could allow accelerated clinical translation of findings, as well as generate evidence in cases where traditional trials are not practical or possible. This iterative approach to improving healthcare delivery is also known as rapid learning.

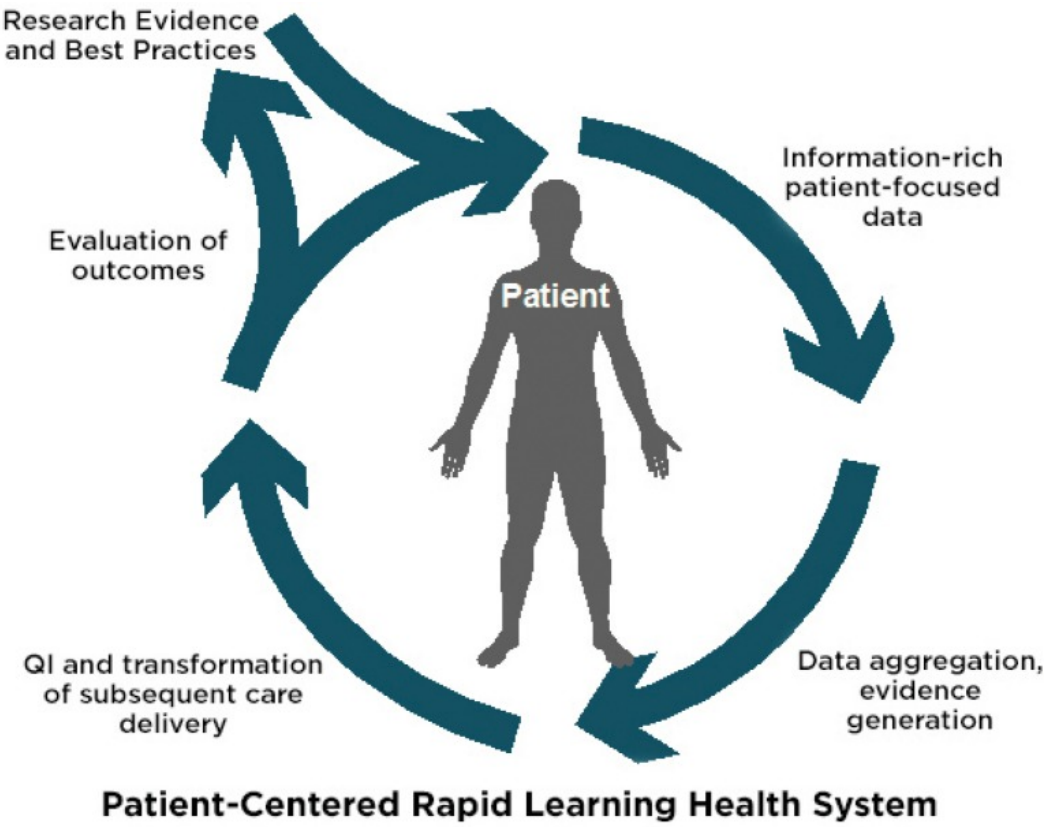

Figure 1.3: The Learning Healthcare System [68].

1.4.1 Analyses within a Learning Healthcare System

A LHS encompasses several steps, and each step requires a different analytical approach. The first step in a LHS is to analyse data to generate clinical insight or evidence that answers a clinical question, shown as the Research Evidence and Best Practices in Figure 1.3. This step can include evidence generated from observational clinical and imaging data, or large scale clinical audits for example. Hypothesis generating analyses, such as large scale radiomics studies, can help identify potential

biomarkers or variables that can be taken forward and tested in larger validation studies, or prospectively in a LHS. Such retrospective research evidence plays a key part in a LHS by generating evidence that, if convincing enough, could subsequently influence clinical practice. Examples of changes to radiotherapy workflows that might come from retrospective analyses include reducing radiotherapy margins or an OAR dose limit.

Once a change to practice has been decided on, implementation within a LHS ensures rapid feedback on the impact of the change on patient outcomes as more patients are treated. As data would generally have been collected as a part of routine care in patient's EHRs, this does not impose extra time or effort for the healthcare provider. Outcomes of interest could include treatment-related side effects, patient reported outcomes, quality of life metrics or increased costs. The real-world evidence generated from the results could influence the next learning cycle, i.e. make further adjustments to practice in light of the results, or influence guidelines for other centres to implement the change if it proves to be successful.

Analysing patient's data prospectively as soon as outcomes are available allows the investigator to rapidly find out if the changes are beneficial or otherwise. A vital part of the LHS is the experimental methodology used to analyse the data to evaluate the impact of the change to practice. The real-world evidence generated needs to be of high quality if it is to influence guidelines and ensure patients are receiving the best quality of care. Experimental or quasi-experimental designs can be used for rapid learning to evaluate the impact of changes to practice. An experimental design includes point-of-care randomisation, ensuring high internal validity as with a RCT. Since the intervention is being tested on real-world patients, the experimental design also has improved external validity compared to a RCT. Quasi-experimental designs do not randomise and will therefore suffer from lower internal validity due to unknown confounding, but will have high external validity due to reduced selection bias [69]. Both methods attempt to establish a cause-and-effect relationship between the intervention or change to practice, and the outcomes of interest.

A non-randomised quasi-experimental design may be easier to implement clinically in

radiotherapy, considering the lack of clinical equipoise means technical changes tend to be implemented for all patients, rather than randomising patients between protocols. A recent review of the LHS/rapid learning literature in radiation oncology found that out of 16 studies, 15 used a quasi-experimental design as opposed to randomising patients [70], 7 of which used a pre/post design i.e. comparing the outcomes of patients before and after a change to practice without using a control group. Such designs do not account for secular trends and may suffer from confounding as patient cohorts may be systematically different before and after the intervention [71]. This can be mitigated against with multivariable analyses adjusting for known confounders, although only one study in the review did this [70]. Moreover, no studies adjusted for multiple comparisons when performing multiple significance tests on patient outcomes or variable distributions [70]. The review concludes that there is little consensus as to the best experimental methods to evaluate changes to practice using real-world data in a LHS [70]. There is, therefore, an unmet need to develop statistical frameworks that help investigators understand, as easily and quickly as possible, whether a change to practice has impacted clinical outcomes or not.

1.5 Aims

Real-world data offer an opportunity to improve outcomes for patients with lung cancer. In particular, real-world data could help to provide evidence that changes made to radiotherapy practice are beneficial, or at the least not detrimental. Real-world clinical and imaging data can help to achieve this by providing large, representative and inclusive datasets that can be analysed to find relationships between variables of interest and outcomes. Retrospective studies done on real-world datasets form part of the LHS concept by generating insight which can inform changes to practice, ensuring patients are receiving the best standard of care.

This thesis is presented in alternative format and includes numerous individual studies which together investigate different aspects of a LHS using concrete case studies with retrospective, real-world data. First, clinical insight is generated from retrospective analysis of both imaging and clinical real-world data. This is followed by investigating the potential of using real-world data to assess the impact of changes to practice. The

overall aims of this thesis are to:

1. Investigate the potential of routine, real-world radiomics imaging biomarkers to generate clinical insight and improve patient outcomes through supported decision making.

2. Develop approaches for using real-world data to assess whether changes to clinical practice affect patient outcomes.

The thesis starts with a thorough review in Chapter 2, investigating whether radiomics, using routinely collected imaging data, has the potential to personalise lung cancer treatment and improve clinical outcomes, addressing aim 1. Reported methodological concerns with CT-based NSCLC radiomics are summarised along with potential solutions. The published literature that use radiomics to predict patient outcomes or aspects of tumour biology is then critically appraised with respect to the methodological concerns identified in this review. Different scoring systems that appraise radiomics and prediction modelling studies are applied to each study and compared to each other. This work has been published in Lung Cancer [72] and is reproduced here, subject to formatting for consistency throughout the thesis.

The results from Chapter 2 led to the work in Chapter 3, having identified a gap in the literature. Chapter 3 addresses aim 1 by investigating whether choice of radiomic feature extraction software influences the statistical reliability of features and the ability to predict clinical outcome. Four software platforms are compared across three clinical datasets, and the impact of IBSI compliance, feature calculation settings and software version are investigated. This work has been published in European Radiology [73] and is reproduced here, subject to formatting for consistency throughout the thesis.

Chapter 4 addresses aim 2 by using routinely collected data to evaluate whether the introduction of IMRT at The Christie NHS Foundation Trust had an effect on the proportion of patients treated with curative-intent radiotherapy and whether patient survival was affected. This work has been published in Frontiers in Oncology [74] and is reproduced here, subject to formatting for consistency throughout the thesis.

Chapter 5 also addresses aim 2, by analysing real-world data collected prospectively for Lung Radiotherapy during the COVID-19 Pandemic (COVID-RT Lung) to assess the impact of changes to treatments for patients with lung cancer during the first wave of the COVID-19 pandemic. This work has been published in Clinical Oncology [75] and is reproduced here, subject to formatting for consistency throughout the thesis. This paper is a follow-up paper to one published describing the changes made to treatments of patients enrolled in COVID-RT Lung [76] which is included in the Appendix.

Chapters 6 and 7 both address aim 2, by assessing the potential benefits of using Bayesian methodology to evaluate changes to practice with real-world data. Chapter 6 takes the form of a teaching article, explaining the differences between frequentist and Bayesian statistical methodologies using a simulated dataset based on the dataset used in Chapter 7. This work has been published in the International Journal of Radiation Oncology-Biology-Physics [77] and is reproduced here, subject to formatting for consistency throughout the thesis. A Letter to the Editor in response to this paper [78] and our reply [79] is included in the Appendix. Chapter 7 then uses the Bayesian methodology on a real-world dataset to investigate whether a change in IGRT patient set-up protocol at The Christie NHS Foundation Trust reduced the risk of death associated with having residual set-up errors towards the heart. This work has been published in Radiotherapy and Oncology [80] and is reproduced here, subject to formatting for consistency throughout the thesis.

Chapter 2

Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype

This chapter has been published in Lung Cancer 2020 Volume 146 p197-208. The section numbering, references and formatting have been modified to be consistent with the rest of the thesis.

Authors

Isabella Fornacon-Wood[1], Corinne Faivre-Finn[1,2], James P B O'Connor[1,3] and Gareth J Price[1]

Affiliations

[1] Division of Cancer Sciences, University of Manchester, Manchester, UK.

[2] Department of Radiation Oncology, The Christie NHS Foundation Trust, Manchester, UK.

[3] Department of Radiology, The Christie NHS Foundation Trust, Manchester, UK.

Author contributions

I performed the systematic literature search and tabulated the results, scoring each paper according to the different scoring systems. I performed the statistical analysis in R. I wrote the manuscript, which was reviewed by all co-authors.

Abstract

Radiomics has become a popular image analysis method in the last few years. Its key hypothesis is that medical images harbor biological, prognostic and predictive information that is not revealed upon visual inspection. In contrast to previous work with a priori defined imaging biomarkers, radiomics instead calculates image features at scale and uses statistical methods to identify those most strongly associated to outcome. This builds on years of research into computer aided diagnosis and pattern recognition. While the potential of radiomics to aid personalized medicine is widely recognized, several technical limitations exist which hinder biomarker translation. Aspects of the radiomic workflow lack repeatability or reproducibility under particular circumstances, which is a key requirement for the translation of imaging biomarkers into clinical practice. One of the most commonly studied uses of radiomics is for personalized medicine applications in Non-Small Cell Lung Cancer (NSCLC). In this review, we summarize reported methodological limitations in CT based radiomic analyses together with suggested solutions. We then evaluate the current NSCLC radiomics literature to assess the risk associated with accepting the published conclusions with respect to these limitations. We review different complementary scoring systems and initiatives that can be used to critically appraise data from radiomics studies. Wider awareness should improve the quality of ongoing and future radiomics studies and advance their potential as clinically relevant biomarkers for personalized medicine in patients with NSCLC.

2.1 Introduction

Lung cancer remains the leading cause of cancer-related mortality worldwide [81]. The 5 year survival for patients with non-small cell lung cancer (NSCLC), the most common form of the disease, is 10-20% [22, 82]. Despite advances in treatment options in recent years, survival rates have changed little [22, 83]. Given the patient variability and tumor heterogeneity of this cancer, personalizing treatment is key to improving survival beyond the current poor prognosis [84]. One requirement for successful delivery of personalized medicine is the identification and validation of biomarkers that can predict which patients will benefit from a given therapy. There is an unmet need for such

biomarkers in lung cancer [85].

Medical imaging plays a key role in the diagnosis and treatment of lung cancer, making the use of image-based biomarkers to guide clinical decision-making attractive. Over the last several decades, a number of biomarkers derived from CT, PET and MRI that measure tumor size, shape and texture, or quantify aspects of the tumor microenvironment have been used in lung cancer studies for diagnosis, prediction, prognostication and response monitoring [85–87].

There is currently substantial interest in using computer algorithms to extend this approach to extract tens to thousands of image 'features' in an analysis pipeline strategy termed 'radiomics'. Such methods test the hypothesis that medical images harbor data that will provide biomarkers for personalized medicine, but that the optimum biomarkers are not readily determined a priori [88]. Imaging biomarker studies postulate that medical images contain biological, prognostic and predictive information that is not apparent when clinicians view scans [47]. In radiomics, this information is extracted from digital images using computer algorithms to form 'radiomic signatures', a type of quantitative imaging biomarker formed by combining the radiomics features that have the strongest association to the measured outcome. The radiomics workflow consists of a series of steps [89] summarized in Figure 2.1. Proponents of radiomics hypothesize that these data-driven approaches will select the most statistically significant signature that relates to an outcome measure of interest. This approach is extremely popular, but to date the resultant imaging biomarkers have not been validated as useful tools for personalized medicine [90].

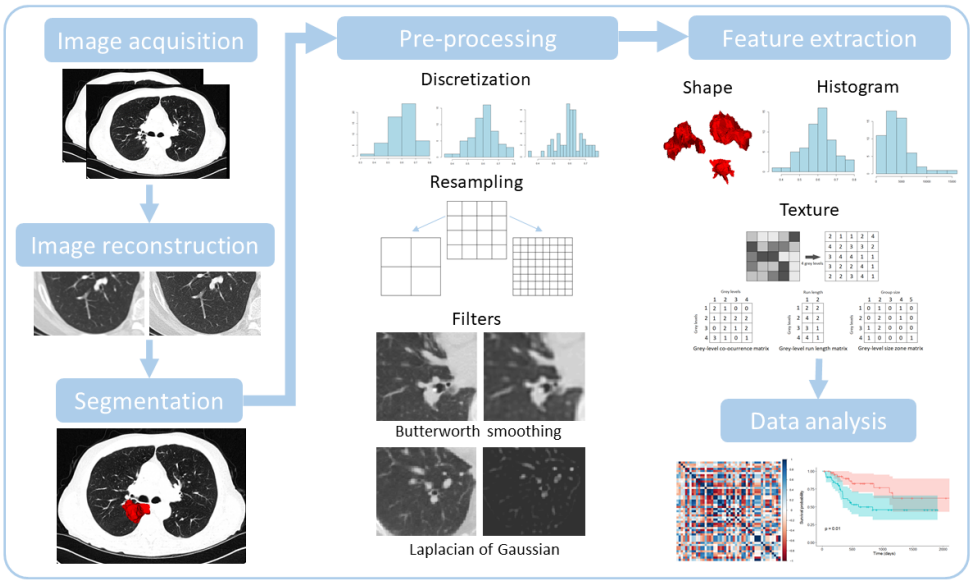

Figure 2.1: Visualization of the steps in the radiomics workflow. First, images are acquired and reconstructed. The region of interest is then segmented, from which features will be extracted. Next, pre-processing steps are performed to modify the images before feature extraction. Shape, first order (or histogram) and texture features are then extracted from the region of interest. Finally, data analysis steps attempt to find correlations between features and the specified outcome.

CT is the most commonly used modality worldwide for diagnosis, treatment planning, and follow-up in all stages of lung cancer, meaning that informative imaging biomarkers discovered from these data could be translated rapidly into clinical practice. In this review, we summarize the literature supporting use of CT radiomic biomarkers to guide decision-making in patients with NSCLC. We appraise the published reports of CT radiomics biomarkers as predictive, prognostic or biologically informative tools and review literature highlighting methodological limitations. Our aims are to evaluate how robust the conclusions of these studies are and to assess how well the current standardization and reporting tools inform readers of the potential limitations when interpreting their results.

2.2 The potential of radiomics for personalized decision-making in NSCLC

A review of the literature found 43 CT image based studies that evaluated the prognostic or predictive role of radiomic signatures in patients with NSCLC (Table 2.1).

Three of these studies, together with a further 21 we separately identified, evaluated the role of radiomic signatures in appraising aspects of tumor biology including genomic or pathologic biomarkers, signalling pathways, and disease classification in NSCLC (Table 2.2).

In addition, 42 studies reported on radiomics methodological limitations, potential problems, and possible solutions in CT based studies using data from NSCLC patients or imaging phantoms. The frequency of publications, for all types of NSCLC radiomics study, has markedly increased over the last six years (Figure 2.2). Our search strategies are described in detail in supplementary materials.

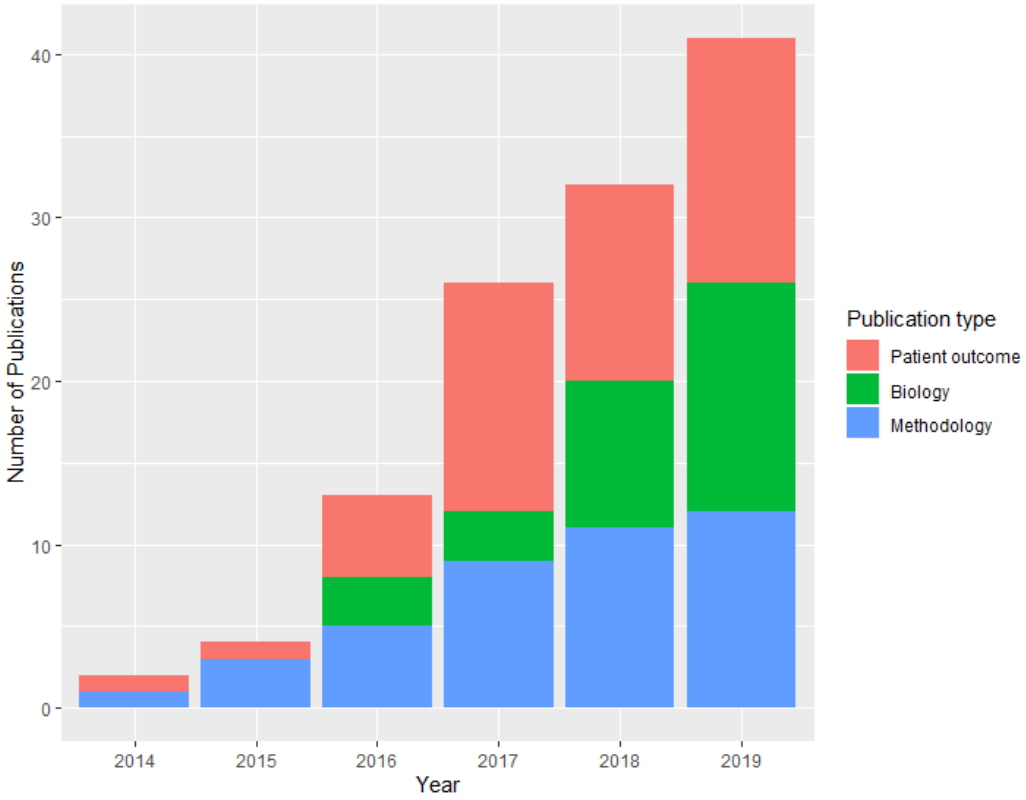

Figure 2.2: Frequency of CT NSCLC radiomics studies published from 2014 to 2019. Publications are categorized as those investigating radiomics methodological concerns, those evaluating radiomic signatures as prognostic or predictive biomarkers of patient outcome, and those evaluating radiomic signatures as biomarkers of tumor biology.

Table 2.1: Radiomics studies in NSCLC, categorized into sections based on their investigated endpoint. The Data column specifies the total number of patients involved in the study, in brackets split by training and validation cohorts if applicable and specifying other cancer types of cohorts if applicable. Note: Studies marked with * are validation studies and their RQS score components refer to methodology based on the previous published data. This table has been simplified to clarify presentation – more details for each study are available in Supplementary Table 2.6.

Reference	Stage	Data (training + validation)	Radiomic features in final model	Result
Overall survival				
Aerts et al. 2014	1-3b	647 pCT (422 + 225)	Shape, first order and texture	CI=0.65
Van Timmeren et al. 2017*	1-4	252 pCT and CBCT (102 + 56 + 94)	Shape, first order and texture	CI=0.69, 0.61, 0.59 (pCT) CI=0.66,0.63,0.59 (CBCT)
Grossman et al. 2017*	1-3	351 diagnostic CT (262 + 89)	Shape, first order and texture	CI=0.60
Grossman et al. 2017	1-3	351 diagnostic CT (262 + 89)	Not specified	CI=0.61
Yu et al. 2017	1	442 diagnostic CT (147 + 295)	First order and texture	CI=0.64
Chaddad et al. 2017	1-3b	315 pCT	Shape and texture	Average AUC=0.70-0.76
Fave et al. 2017	3	107 4DCT end of exhale, planning and CBCT	Shape and texture	CI=0.672
Li et al. 2017	1-2a	59 follow up CT	Texture	AUC=0.81

Table 2.1 continued from previous page

Reference	Stage	Data (training + validation)	Radiomic features in final model	Result
Li et al. 2017	1-2a	92 4DCT Average-CT or 50% phase-CT images were used for analysis	Shape and first order	AUC=0.728
Tang et al. 2018	1-3	290 staging CT (114 + 176)	Shape, first order and texture	CI=0.72
Bianconi et al. 2018	1-3	203 pCT	Shape and texture	HR=1.06-1.48
De Jong et al. 2018*	4	195 diagnostic CT	Shape, first order and texture	CI=0.576
Lee et al. 2018	1-3	339 CT (type not defined, just pre-operative within 2 weeks before surgery)	Shape, first order and texture	CI=0.772
He et al. 2018	1-3	186 CT (298 after oversampling (223 + 75)) type not defined	Not specified	AUC=0.9296
Starkov et al. 2018	1	116 pCT	Texture	High risk vs low risk median p-values=0.04–0.07
Yang et al. 2018	1-4	371 CT (239 + 132)	First order and texture	CI=0.702
Wang et al. 2019	3	70 pre-treatment and 97 post treatment CT from 118 patients	Texture	CI=0.743

Table 2.1 continued from previous page

Reference	Stage	Data (training + validation)	Radiomic features in final model	Result
Shi et al. 2019	3	11 CBCT from 23 patients	First order	HR=0.21
Van Timmeren et al. 2019	1-4	337 pCT and 2154 CBCTs from 337 patients (141 + 94 + 61 + 41)	First order and texture	CI=0.59, 0.54, 0.57
Huang et al. 2019	1-4	371 CT (254 + 63 + 54)	Shape, first order and texture	CI=0.621, 0.649
Franceschini et al. 2019	1-2	102 4DCT (start of inspiration) (70 + 32)	Shape and texture	AUC=0.85
Local or metastatic recurrence				
Coroller et al. 2015	2-3	182 pCT (98 + 84)	First order and texture	CI=0.6
Mattonen et al. 2016	1	45 follow-up CT	First order and texture	AUC=0.85
Huynh et al. 2016	1-2	113 CT (free breathing)	First order and texture	Median CI=0.67
Huynh et al. 2017	1-2a	112 free breathing CT and AIP CT	Shape, first order and texture	AIP radiomics CI=0.667 FB radiomics CI=0.601
Fave et al. 2017	3	107 4DCT end of exhale, planning and CBCT	Shape and texture	CI=0.632, 0.558 (DM, LRR)
Li et al. 2017	1-2a	59 follow up CT	Texture	AUC=0.80, 0.80 (RFS, LR-RFS)

Table 2.1 continued from previous page

Reference	Stage	Data (training + validation)	Radiomic features in final model	Result
Li et al. 2017	1-2a	92 4DCT Average-CT or 50% phase-CT images were used for analysis	Shape	AUC=0.747, 0.690 (RFS, LL-RFS)
Dou et al. 2018	2-3	200 pCT (100 + 100)	Texture	CI=0.65
Ferreira Junior et al. 2018	1-4	68 CT (52 + 16)	Shape and texture	AUC=0.75, 0.71 (lymph node metastasis, DM)
Yang et al. 2018	1-3	159 CT (106 + 53)	Shape, first order and texture	AUC=0.856
Zhong et al. 2018	1-2	492 CT	First order and texture	AUC=0.972
Lafata et al. 2019	1	70 CT	Texture	Maximum AUC=0.72, 0.83, 0.60 (recurrence, LR, non-LR)
Akinci D'Antonoli et al. 2019	1-2b	124 CT	Shape, first order and texture	AUC 0.731, 0.750 (LR, DM)
He et al. 2019	Not specified	717 CT (423 + 294)	First order and texture	CI=0.734
Xu et al. 2019	3-4	132 CT (106 + 26)	Texture	AUC=0.642

Table 2.1 continued from previous page

Reference	Stage	Data (training + validation)	Radiomic features in final model	Result
Franceschini et al. 2019	1-2	102 4DCT (start of inspiration) (70 + 32)	Shape, first order and texture	AUC=0.73
Ferreira-Junior et al. 2019	1-4	85 CT	Shape, first order and texture	AUC=0.92, 0.84 (DM, nodal metastasis)
Cong et al. 2019	1a	649 venous phase CT (455 + 194)	Shape, first order and texture	AUC=0.851
Treatment response, disease-free or progression-free survival				
Coroller et al. 2016	2-3	127 pCT	Shape, first order and texture	Median AUC=0.65, 0.61 (GRD, pCR)
Huang et al. 2016	1-2	282 CT (141 + 141)	First order and texture	HR=2.09
Song et al. 2016	1-4	152 CT (80 + 72)	Texture	HR= 2.35, 2.75
Coroller et al. 2017	2-3	85 pCT	Shape, first order and texture	Median AUC=0.68, 0.71 (pCR, GRD)
Tunali et al. 2019	3b-4	228 CT	Texture	AUC=0.804
Franceschini et al. 2019	1-2	102 4DCT (start of inspiration) (70 + 32)	Texture	AUC=0.88
Lung toxicity				
Moran et al. 2017	1	14 diagnostic CT	First order and texture	AUC=0.689-0.750

Table 2.1 continued from previous page

Reference	Stage	Data (training + validation)	Radiomic features in final model	Result
Krafft et al. 2018	Not specified	192 50% 4DCT phase	First order and texture	Average AUC=0.68
Staging				
Yuan et al. 2018	1	327 CT	First order and texture	AUC=0.938
Yang et al. 2019	1-3	256 CT	First order and texture	AUC= 0.93

Abbreviations: AUC, area under the curve; CBCT, cone-beam CT; CI, concordance index; DFS, disease free survival; DM, distant metastasis; GRD, gross residual disease; H&N, head and neck; HR, hazard ratio; LR, local relapse; LRR, local regional recurrence; LR-RFS, loco-regional recurrence-free survival; OS, overall survival; pCR, pathological complete response; pCT, radiotherapy planning CT scan; PFS, progression free survival; RFS, recurrence free survival.

Table 2.2: Radiomics studies in NSCLC with an aspect of biology as the endpoint. The column labeled 'Data' specifies the total number of patients involved in the study, in brackets split by training and validation cohorts if applicable and specifying other cancer types of cohorts if applicable. This table has been simplified to clarify presentation – more details for each study are available in Supplementary Table 2.7

Reference	Stage	Endpoint	Data (training + validation)	Radiomic features in final model	Result
Genomics					
Aerts et al. 2016	Early stage	EGFR	47 diagnostic CT and follow-up	Shape and texture	AUC=0.74-0.91
Rios Velazquez et al. 2017	1-4	EGFR, KRAS	705 diagnostic CT (353 + 352)	Shape, first order and texture	AUC=0.69-0.80
Mei et al. 2018	Not specified	EGFR	296 CT	Texture	AUC=0.664
Digumarthy et al. 2019	Not specified	EGFR	93 CT	First order	AUC=0.713
Jia et al. 2019	1-4	EGFR	504 CT (345 + 158)	Shape, first order and texture	AUC=0.802
Li et al. 2019	1-4	EGFR subtypes (19Del and L858R)	312 CT (236 + 76)	Shape and first order	AUC= 0.775-0.793
Tu et al. 2019	1-4	EGFR	404 CT (243 + 161)	First order and texture	AUC=0.775
Yang et al. 2019	Not specified	EGFR	467 CT (306 + 161)	Shape, first order and texture	AUC=0.789

Table 2.2 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Radiomic features in final model	Result
Wang et al. 2019	1-2	EGFR, TP53	61 CT (41 + 20)	First order and texture	AUC=0.604, 0.586
Wang et al. 2019	1-2	Tumor mutation burden	61 CT (41 + 20)	Texture	AUC=0.606
Signaling pathways					
Grossman et al. 2017	1-3	Various	351 CT (262 + 89)	Shape, first order and texture	AUC=0.62-0.72
Bak et al. 2018	1-4	Various	57 CT	First order and texture	OR=0.08-23.94
Histopathology					
Patil et al. 2016	Not specified	ADC, LCC, SCC, NOS	317 pCT	Shape, first order and texture	88% accuracy
Wu et al. 2016	1-4	ADC, SCC	350 pCT (198 + 152)	First order and texture	AUC=0.72
Ferreira Junior et al. 2018	1-4	ADC, SCC	68 CT (52 + 16)	Not specified	AUC=0.81
Zhu et al. 2018	Not specified	ADC, SCC	129 CT (81 + 48)	First order and texture	AUC=0.893
Digumarthy et al. 2019	Not specified	ADC, SCC	93 CT	First order	AUC=0.744
E et al. 2019	Not specified	ADC, SCC, SCLC	229 CT	Shape, first order and texture	AUC=0.657-0.875

Table 2.2 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Radiomic features in final model	Result
Ferreira-Junior et al. 2019	1-4	ADC, SCC	85 CT	Shape, first order, texture	AUC=0.88
Liu et al. 2019	Not specified	ADC, LCC, SCC, NOS	349 CT (278 + 71)	Not specified	AUC=0.86
Zhou et al. 2018	1-4	Ki-67	110 CT	Shape and texture	AUC=0.61-0.77
Gu et al. 2019	Not specified	Ki-67	245 CT	First order and texture	AUC=0.776
Song et al. 2017	1-3	Micropapillary pattern	339 CT	First order	AUC=0.751
Chen et al. 2018	Not specified	Degree of differentiation	487 CT (303 + 184)	First order and texture	AUC=0.782
She et al. 2018	Not specified	Invasive vs non-invasive adenocarcinoma	402 CT (207 + 195)	Shape, first order and texture	AUC=0.89
Yang et al. 2019	Not specified	Invasive vs non-invasive adenocarcinoma	192 CT (116 + 76)	First order and texture	AUC=0.77

Abbreviations: ADC, adenocarcinoma; AUC, area under the curve; CI, concordance index; EGFR, epidermal growth factor receptor; KRAS, Kirsten rat sarcoma viral oncogene homolog; LCC, large cell carcinoma; NOS, not otherwise specified; OR, odds ratio; SCC, squamous cell carcinoma.

The initial studies labelled as 'radiomics' were published in 2014 and 2015. Aerts and colleagues showed that a radiomic signature based on shape and texture metrics was associated with overall survival, validating the signature in patients with NSCLC and patients with head and neck cancers [91]. The study also found positive associations between the radiomic signature and gene expression. Coroller and colleagues showed that a different set of texture metrics were associated with the subsequent development of distant metastases [92]. The hypothesized mechanism was that tumor heterogeneity, identified by the radiomics analyses, drives worse outcomes. Both studies were performed using radiotherapy planning CT data.

Over the next four years (2015–2019), 41 CT studies were published that linked radiomics to lung cancer patient outcome. In general, studies sought to evaluate whether or not radiomic signatures could outperform existing methods for patient risk stratification. 20 studies related radiomics to overall survival [91, 93–111], 18 to the likelihood of local or metastatic recurrence [92, 104, 108–110, 112–124], 6 to response, disease-free or progression-free survival [104, 125–129], and 2 to staging [130, 131]. Two further studies focused on the association of radiomics signatures to lung toxicity [132, 133], Four studies investigated multiple endpoints.

The majority of studies derived radiomics signatures in radiotherapy planning or diagnostic images acquired prior to therapy. Nearly all studies evaluated patients undergoing treatment with cytotoxic chemo-radiotherapy. More recently, a number of studies have evaluated the potential of radiomics to improve patient stratification for targeted therapies and immunotherapy agents [111, 129, 134]. For example, Tang and colleagues linked radiomic features to a tumor immune phenotype in patients with stage I-III NSCLC, finding patients with heterogeneous tumors, which correlated with low PD-L1 and high CD3 cell count, had better prognosis [111].

There are 24 CT studies evaluating how radiomic signatures of NSCLC relate to genomics [134–142], signalling pathways [105, 143] and histopathology [112, 119, 137, 144–154]. For example, Rios Velazquez and colleagues found distinct imaging phenotypes for EGFR and KRAS mutations from CT images of patients with NSCLC [135]. Some of the studies that relate radiomics to patient outcome also relate their radiomic

signature to genomics [91] or biological markers [99].

Collectively, these 64 studies present a positive view of the potential for radiomics signatures to deliver personalized medicine. However, two important limitations are readily apparent. Firstly, while nearly all studies report at least one positive association between CT radiomic signature and either outcome (OS, PFS, recurrence or toxicity) or tumor biology (genomic or pathology biomarkers and signalling pathways), the particular radiomic signature derived varies substantially between studies. Consequently, few study signatures are directly comparable with one another, and so the literature does not identify specific candidate radiomic signatures for further large multicenter evaluation.

Secondly, it has become clear that studies can suffer from significant technical limitations. Studies of these limitations have also increased over the last five years, although at a slower pace than the patient outcome studies Figure 2.2.

2.3 Reported methodological limitations of CT based radiomics studies

All biomarkers, including radiomic signatures, must undergo technical and biological validation to become robust tools used to guide clinical decision-making. These validation steps take a biomarker from discovery to research assay where the biomarker can be used with confidence to determine an outcome in a research setting (termed 'crossing translational gap 1'). The regulatory approval process (through e.g. the FDA or EMA) then takes the biomarker from research assay to clinically approved assay for use in decision-making in patients (termed 'crossing translational gap 2') [90].

To date, very few radiomics signatures have crossed either of these translational gaps. The first radiology product with radiomics capabilities to receive such approvals was QuantX for detection of breast abnormalities based on MRI, receiving FDA approval in 2017 [155]. Soon afterwards, Feedback Medical received CE approval for TexRAD Lung, a quantitative image texture analysis technology [156].

In this section, we evaluate the methodological limitations preventing CT based radiomics signatures from crossing these translational gaps. We review the potential problems and proffered solutions identified in 42 studies of imaging phantoms or patients with NSCLC (summarized in Table 2.3 and expanded in Supplementary Table 2.5).

Table 2.3: Potential problems at each step of the radiomics workflow along with possible solutions offered by the literature. Each workflow step with potential problems and solutions identified by the literature is labelled with a letter A-H to reference in-text. Note: Modelling does not have a letter associated with since there is no consensus on the best statistical modelling strategies.

Problem area		Potential problems	Potential solutions
Image acquisition	A	Different scanners and acquisition protocols affect feature reproducibility [49, 50, 157–167]	Image phantoms on different scanners to provide baseline [49], establish credibility of scanners and protocols [50], catalogue reproducible features [159, 166], model a correction algorithm [158], harmonize data [160].
	B	Patient motion affects feature reproducibility [161, 168, 169]	Set motion tolerances, reduce ROI boundaries [161], use single phase from 4D images [168], find robust features using 4DCT data [169].
Image acquisition and reconstruction	C	Image resolution parameters (voxel size, slice thickness) affect feature values [49, 157, 170–174], model performance [175].	Control resolution [49] parameters in prospective studies, resample to common resolution and voxel depth [170–172, 174], apply smoothing image filters [171], apply deep learning methods [176].
Image reconstruction	D	Image reconstruction algorithm and reconstruction parameters (kernel) affects features [173, 177, 178]	Pre-processing image correction [177] and harmonization of acquisition techniques [173, 178].
Segmentation	E	Delineation variability [159, 179–183] affects features and is time consuming [182, 183]. Results from one disease site are not necessarily transferrable to another [184].	Expert ROI definition [179], multiple observers [179, 180, 184], identification of stable features with respect to delineation [159, 180, 181], automated segmentation [182, 183], image filtering [184]

Table 2.3 continued from previous page

Problem area		Potential problems	Potential solutions
Pre-processing	F	Number of grey levels used to discretize histogram and texture features affects feature values [172, 174, 185], as does bin width [170].	Texture features can be normalized to reduce dependency on the number of grey levels [174], number of grey levels used for discretization should be recorded with feature formula. 128 grey levels may be optimal for texture features, along with thresholding [185]
Feature extraction		No studies found in the literature search.	
Feature correlation	G	Strong correlations between tumor volume and radiomic features exist [174, 186–188]	Normalization of features to volume [174], bit depth resampling [186], feature redesign [186], more robust statistics to check added value of radiomics signatures [187].
Test re-test	H	Radiomic features may not be repeatable over multiple measurements [189–191], repeatable features are not generalizable to other disease sites [192].	Test-retest data acquisition [189, 192], use of multiple 4D phases [189, 191], use of simulated retest by image perturbation [190].
Modelling clinical outcome		Different modelling strategies affect model performance [193–196]	Sample sizes above 50 give better predictive performance [194], as does normalizing features [193]. No consensus on best modelling strategies to use.

2.3.1 Image acquisition

Many radiomics studies are retrospective evaluations of CT images, often with data acquired at multiple different institutions and on different CT scanner vendor platforms. Consequently, nearly all studies contend with variations in image acquisition and reconstruction protocols.

Studies assessing the impact of different CT scanners and protocols on radiomic features have shown some features have poor reproducibility [49, 50, 161, 165–167]. Performing phantom studies on different scanners as a quality assurance step may ensure a level of feature consistency [50]. Indeed, one study showed that using a controlled protocol across different CT scanners reduced feature variability by over 50% compared to using local protocols [49]. Other studies used post-extraction deep learning [176] or correction factors [158] to reduce feature variability.

Restricting study data to one scanner make and model along with one set of acquisition parameters, to reduce variability in image capture acquisition, is seldom feasible for a multicentre research study. Therefore, many of these issues still remain when setting up a well powered prospective clinical trial with radiomic signatures as exploratory endpoints.

2.3.2 Image reconstruction

Retrospective data analyses are constrained by image reconstruction parameters determined by clinical department protocols, chosen to optimize image anatomical quality. While variations in image reconstruction, slice thickness and in plane pixel dimensions may have negligible effect for clinical interpretation, they can induce variability in radiomic feature values, since many features correlate to these parameters [49, 170–174].

Resampling the image to an equal voxel size has reduced feature dependency on acquisition in some studies [170, 172] but not others [49, 171]. Smoothing filters have also been suggested as a method for reducing voxel size dependency [171], as has limiting inclusion criteria to particular resolution ranges. For example, Lu et al. found that features calculated from images with 1.25mm and 2.5mm thick slices were comparable

to each other but that both differed from those calculated on 5mm slice thickness images [173].

Reconstruction techniques also influence feature values with studies demonstrating differences between features calculated on images reconstructed with soft or sharp kernels [173, 178]. Potential solutions include the application of correction factors based on the image noise power spectrum [177]. Solutions that balance feature robustness with the need to make image inclusion criteria as permissive as possible are vital given the small cohorts size issues that blight many studies.

2.3.3 Segmentation

The ROI definition for feature extraction is known to be a particularly sensitive step in the radiomics pipeline [179–183]. Radiomics studies are popular in radiotherapy given the ready availability of pre-defined ROIs on treatment planning scans, typically using the clinically defined Gross Tumor Volume (GTV). The subjectivity of GTV definition can depend on the operator, as expert delineations may generate features with better predictive power than those from a non-specialist [179].

Frequently suggested solutions include the inclusion of multiple observers or the use of semi-automated delineation tools [182, 183]. However, few studies have adopted these solutions, most likely due to the difficulty of getting clinically qualified staff to delineate ROIs. In studies not using radiotherapy planning CT scans, the ROIs must be drawn specifically for the purpose of the radiomics analysis and will suffer from all of the same issues discussed above.

2.3.4 Pre-processing

The preparation of images for feature extraction has a marked effect on feature value. Reducing the number of image grey-levels (voxel depth re-binning) is a commonly used method to supress image noise. However, studies have shown that radiomic features are not comparable when computed with a differing intensity bin sizes [170, 172, 174]. This has led to the proposed use of standardized bin resolution [174].

2.3.5 Feature extraction

Radiomics features span a range of calculation classes. Shape features contain information about the ROI morphology (such as volume and measures of sphericity). First-order image intensity features assess properties of the intensity histogram of voxels within the ROI (e.g. the mean intensity and other statistical moments of the histogram). Texture features summarize different measures of the way in which voxel intensities change across the ROI (e.g. voxel variation coarseness and homogeneity). These features may be calculated on the original image or derived after various filters have been applied that modify particular aspects of it, for example to enhance the edges where image intensity changes [89].

Many different software platforms exist for performing the feature extraction step, including free open-source software, commercial software, and software developed in-house by individual institutions. The Image Biomarker Standardization Initiative (IBSI) is an international collaboration between research groups with the aim of standardizing image biomarker extraction [51]. To date only one study has investigated whether feature extraction software influences radiomic features from CT scans of patients with NSCLC [73], which shows, consistent with data from other cancer types [197, 198], that this can have substantial impact on feature values.

2.3.6 Feature correlation

Since many tens to thousands of features are calculated from images in radiomics, it is unsurprising that many features often correlate with one another. However, the fact that features often correlate strongly with tumor volume and clinical factors [174, 186, 187] is not well appreciated. While it has been suggested that radiomic feature calculations formulae should be modified to be account for tumor volume [174], it is crucial that studies also include transparent and robust feature reduction steps to account for other clinical prognostic and predictive factors. Robust feature reduction is also crucial in limiting the risk of model overfitting.

2.3.7 Test-retest

As highlighted by several studies, [189, 192] and by consensus statements on imaging biomarkers [90], radiomics studies usually lack an assessment of the signatures' single centre repeatability or multicentre reproducibility. The use of test-retest datasets in which multiple images of the same subjects or phantom have been acquired in quick succession have been proposed as a means to assess repeatability [189, 192]. Alternative options include the use of multiple 4D image phases [189] and the simulation of retest data by image perturbation [190] where test-retest data are not available. Few radiomic studies incorporate any of these approaches.

2.3.8 Modelling clinical outcome

Typically, studies derive between tens to a few thousand image features in development datasets [45]. Dimensionality reduction to remove highly correlated and unstable radiomic features is often employed before finding the most informative features for a specific outcome, such as overall survival, treatment-related toxicities or cancer recurrence in a test dataset. Many different statistical options exist for deriving a model based on radiomic features. The choice of model and statistical methods can influence results [194–196].

Random forests have been found by some authors to give higher performance compared to other methods for classification tasks using radiomics features [194, 196], with Naïve Bayes and Support Vector Machines also reported to perform well [194]. For radiomic feature based time-to-event analyses, one study found cox regression with gradient boost performed better than traditional cox regression (0.614 versus 0.660 concordance index) [195]. In terms of feature selection, there is no consensus on the best method to use. Optimal performance of feature selection techniques depend on the outcome of interest [194]. A contemporary non-radiomics study of classifier performance in radiotherapy datasets found that random forest and elastic net logistic regression performed best, but that classification accuracy depended on the specific dataset [199]. To summarize, there is limited consensus as to the best machine learning methods to employ for radiomics studies, and that the optimum choice may depend on the specific dataset used in the study.

Regardless of feature selection and modelling methodology, the resulting model (often termed a 'radiomic signature') should be robustly validated in line with the TRIPOD guidelines to ascertain if it is reproducible across different clinical datasets. This tests if the observed signature relates to the desired outcome in a different patient group, and aims to reduce the risk of overfitting in the training cohort [45].

Lastly, whatever approach is taken it is vital that investigators test whether incorporating radiomic features into a clinical model adds any benefit to well-known clinical prognostic factors such as tumor stage and performance status. Radiomic features will only have clinical utility if they provide more predictive information than is currently available in the clinic.

2.4 Assessing the quality of radiomics studies in NSCLC

We evaluated the quality of the 43 radiomics studies we identified that report a relationship between a CT defined radiomic signature and clinical outcome in patients with NSCLC (Supplementary Table 2.6) using both established assessment tools and the results of our review of methodological limitations reported above. We then applied the same tools to the 24 studies that evaluated the relationship between CT radiomic signatures and genomic, protein expression, and pathology biomarkers in patients with NSCLC (Supplementary Table 2.7). Some studies investigated multiple endpoints, so in total we evaluated 75 outcomes. The four tools we use to interpret the technical validation of these studies are:

1. The strength of the validation in each study, assessed by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines [200]. TRIPOD provides an ordinal score (1-4, with 4 being the most robust). These guidelines are not specific to radiomics studies, but provide insight into the level of validation in a study (details in Supplementary Table 2.8).

2. The Radiomics Quality Score (RQS) developed by Lambin and colleagues [201]. RQS provides a checklist to evaluate aspects of study design, by assessing various technical and statistical aspects of the radiomics pipeline. It consists of 16

components, each of which award or penalize points, to provide the RQS. The total number of points available range from -8 to 36 (the more points the better) and are often presented as a percentage (Supplementary Table 2.9).

3. Qualitative assessment of radiomics methodological limitations resulting from our literature review and labelled as A–H and listed in Table 2.3.

4. The reported evidence for added value of the radiomics signature to a clinical model of outcome tested in the study (for the patient outcome studies only). This provides an assessment of clinical utility.

2.5 Interpreting the quality of radiomics studies in NSCLC

Studies linking CT radiomics signatures to clinical outcome and tumor biology were found to have a high incidence of methodological limitations (summarized in Table 2.4). Overall, half of studies had a TRIPOD type of either 1a or 1b (meaning the results were not validated or validated within the same dataset). Only 13/75 studies had TRIPOD type of 3 or 4 (meaning the results were validated in an external dataset). The median RQS was 6 (range of -8 to 36). Details on RQS and TRIPOD are found in supplementary material. We found that 70% of studies (52 of 75) had six or more methodological limitations, and no study had less than three methodological limitations. Finally, over half of studies relating radiomics to patient outcome did test the added benefit of the radiomic signature to a clinical model.

Table 2.4: Summary of the 4 assessment criteria - TRIPOD score, RQS, number of methodological limitations and testing the added value of radiomics to a clinical model. The added value of radiomics to a clinical model was only tested for the patient outcome studies (N=50).

	N=75
TRIPOD type (n (%))	
1a – no validation	10 (13)
1b – internal validation	27 (36)
2a – dataset randomly split for validation	18 (24)
2b – dataset non-randomly split for validation	7 (9)
3 – external validation	10 (13)
4 – validation only	3 (4)
RQS (median, [IQR])	6 [2-12.25]
Number of methodological limitations (n (%))	
0-2	0 (0)
3	4 (5)
4	4 (5)
5	15 (20)
6	21 (28)
7	23 (31)
8	8 (11)
	N=50
Added value of radiomics to clinical model tested? (n (%))	
Yes	32 (64)
No	18 (36)

Our analysis suggests that the four assessment tools provide useful and complimentary critiques. Figure 2.3A shows that the TRIPOD ordinal score focusing on validation and the RQS score focusing on study reporting are correlated (Pearson correlation coefficient 0.70). This reflects the importance the RQS places on study validation. However, both the TRIPOD score and RQS score were relatively independent of our assessment of study methodological limitations (Figure 2.3B-C, Pearson correlation coefficients -0.12 and 0.13). Indeed, some studies with high TRIPOD and RQS scores had several technical limitations listed. For example, two studies with a TRIPOD score of 4 and the highest reported RQS scores (16 and 18 respectively) [103, 105] had five and six identified methodological limitations respectively. In contrast, one study with a low TRIPOD score of 1b and a moderate RQS score (of 7) had just three

pipeline technical limitations [108].

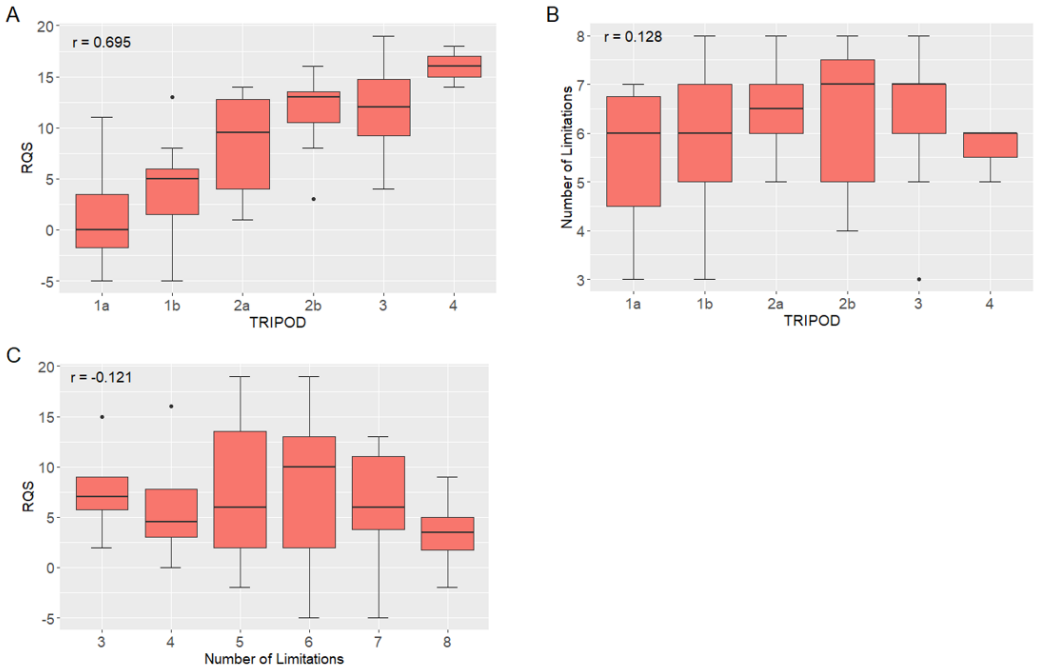

Figure 2.3: The assessment of the literature plotted against each other as boxplots. (A) RQS versus TRIPOD, (B) RQS versus the number of methodological limitations found in this review and (C) TRIPOD versus the number of methodological limitations found in this review.

An illustrative example is given by three studies [94, 103, 105] that externally validated the landmark radiomic signature developed by Aerts and colleagues in 2014 [91]. However, subsequent work [187, 202] has suggested that the prognostic value of the signature reflected the correlation of the signature with tumor volume, rather than reflecting underlying tumor heterogeneity. An important workflow step our review identified is the assessment of feature correlations and potential confounders (G). While the RQS recommends performing multivariable analysis and testing the benefit of the radiomics signature to a gold standard, it does not explicitly recommend testing for feature correlations or confounders.

Study quality depends not only on quality of reporting, but also on ensuring that features used are robust against potential problems. There is a raised recognition of the methodological issues that limit the potential utility of the radiomics concept, as shown by the increase in studies in this area Figure 2.2. However, we find that only 39% of the patient outcome studies and 50% of the biology studies we identified cite

methodology papers. This suggests that there is still limited appreciation of the need to employ more rigorous radiomics workflows. The IBSI guidelines and RQS are aimed at addressing these issues. For example the IBSI reference manual gives recommendations for image processing techniques as well as suggesting standardized feature definitions, nomenclature, and guidelines for reporting [51]. The RQS rewards the use of test-retest approaches, multiple segmentation analyses, and the use of phantoms to resolve inter-scanner differences.

However, our review of limitations highlights further concerns, such as differing slice thickness or voxel size (C) and the specification of grey-level binning size (F). These are not included in RQS (only 58% of studies in Table 2.1, Table 2.2 specified the grey-level binning method or size). The IBSI guidelines, the RQS and TRIPOD assessment schemes are important steps that should improve the technical quality of radiomics studies. However, they are not sufficient alone and review of the literature suggests a need to either update them to include more granular limitations or to use them alongside other assessment tools.

One result of the increase prevalence of studies investigating methodological limitations that would accelerate clinical translation would be the identification of a subset of robust features that should be used in outcome studies. Unfortunately, comparing results across studies is difficult. In addition to the risks to reliability listed in Table 2.3, the software used for feature extraction often uses different nomenclature (one of issues the IBSI addresses) and can calculate ostensibly similar features in different ways and with different parameter settings so that they are not comparable [73]. Software use varied greatly across all studies included in this review. Of the patient outcome and biology studies, 15% did not specify the software used, 48% used in-house developed software and just 37% used free or commercial options. These numbers are similar for the methodology studies; 14% did not specify the software used, 40% used in-house developed software and 47% used free or commercial options. Four of the patient outcome and biology studies did not specify the features in the final radiomic signature at all. The result is that there is no consensus on which particular features or feature signatures should be used for clinical studies. However, there are now increasing numbers of studies that employ the techniques used to determine which features are

reliable. Table 2.3 and Supplementary Tables 2.6 and 2.7 list the remaining limitations for each clinical and biological study - 42% of the assessed studies applied at least one of the suggested solutions to methodological limitations to increase feature robustness. Of these studies, 46% used a test re-test dataset, 58% used multiple segmentations and 4% tested CT model dependence.

A further important step in the radiomics workflow where community consensus would increase the comparability of studies is that of the optimal machine learning techniques that should be used to develop the resulting statistical models. We found that the top feature reduction technique used in all studies was univariable analysis (53%) followed by LASSO (27%). The most common modelling technique was logistic regression (39%) followed by cox regression (34%). 16% of studies used random forest and 11% SVM, both of which were highlighted as high performing by the methodology studies [194, 196]. The techniques used in each study are listed in Supplementary Tables 2.6 and 2.7. Four outcome studies used multiple modelling techniques to determine which one performed best on their data; a recommended method as model performance is dataset-dependent [199]. Out of these four studies, the best performing classifiers were random forest [144] and Naïve Bayes [112, 150]. One study did not reveal the best performing model [152].

The lack of consensus in how to address limitations to the reliability of radiomics features, or of a preferred way to conduct the subsequent statistical modelling, means there is still significant variability in approach, with each finely tuned to its own particular dataset. Progress along the imaging biomarker translation roadmap [90] is dependent on the development of reliable measures that can be used to test clinical hypotheses. These findings agree with those of previous authors [51, 201] and show there is still an unmet need to move away from the current heterogeneous landscape to one that is more standardized. The validation of existing signatures in different datasets [94, 103, 105] discussed above is a vital part of this effort.

Lastly, in addition to the assessment of technical quality, radiomic signatures need to be evaluated for clinical relevance. It is important to test whether incorporating radiomic features into a clinical model improves performance over known prognostic

or predictive factors. This need is well-recognized with 64% of the studies in Table 2.1 making its assessment. Future studies will be most impactful if they explicitly evaluate the clinical utility of a radiomic signature as part of data reporting.

In summary, use of the four different assessment tools allows us to draw three conclusions. Firstly, there is a high prevalence of methodological limitations among CT radiomics studies exploring the potential of the approach to guide personalized medicine. Secondly, there remains considerable variability in the approach to addressing these limitations, and that modelling approaches are likely tuned to specific datasets. Thirdly, different assessment tools provided complementary information, which taken together provided the greatest insight into how study data could be improved.

2.6 Future directions

Personalized medicine is of great potential benefit to patients, but this vision is dependent on the identification of stratification and predictive biomarkers [84]. Imaging biomarkers, derived from routinely acquired patient images, have enormous translational potential given the ubiquity of imaging in clinical workflows. Evaluation of the radiomics literature in NSCLC reveals the exponential rate of publication of new radiomics studies, which, in their conclusions, present a very positive view of the potential for radiomics to deliver this goal.

This review puts these findings in context for NSCLC, but the messages are likely to be generic to all cancer types. All published studies are at risk of translational hurdles due to technical and methodological issues. Importantly, some of these limitations are well recognized, well investigated and have solutions proposed that are beginning to be applied to clinical studies. In distinction, other limitations are poorly understood or researched, and so substantial barriers to translation remain. In addition, wider concerns surrounding over-fitting data and biological validation persist. Lastly, no single radiomic signature or methodological approach is used widely, so further work is required to identify candidates to take forward in larger multicenter studies.

The fact that all the radiomics studies identified in the NSCLC literature have some

limitations should not infer that the published data and conclusions are incorrect; rather that risk exists in interpreting their findings at face value. Standardization issues, variability in methodology and a general lack of reporting hinders comparison of results across studies. Identifying limitations, by employing recognized assessment methodology tools, can help inform and educate design of future radiomics studies in NSCLC and beyond. This will improve study quality and expedite the translation of radiomic biomarkers as tools in personalized medicine.

2.7 Acknowledgements

This work was supported by Cancer Research UK through the Cancer Research UK Manchester Centre: [C147/A18083] and [C147/A25254]; and an Advanced Clinician Scientist Fellowship to Professor James P B O'Connor [C19221/A22746]. Professor Corinne Faivre-Finn and Professor James P B O'Connor are supported by the NIHR Manchester Biomedical Research Centre.

2.8 Supplementary materials

Supplementary

Literature search

Search strategy

Publications that report radiomics analyses on NSCLC data with the aim of predicting patient outcome were identified by searching the PubMed database using the key words "radiomics" and "lung cancer" or "NSCLC". The search was conducted on the 08/01/2020 and no start date limit was used.

A second search was undertaken to find studies that addressed a methodological concern of radiomics. The PubMed database was searched using a combination of the following keys words (a) "radiomics" or "radiomics" and (b) "cancer" and (c) "standardization" or "reliable" or "impact of" or "improvement" or "repeatable" or "reproducible" or "repeatability" or "reproducibility" or "test–retest" or "variability" or "limitation" or

"limitations" or "vulnerability" or "vulnerabilities" or "stability" or "stable" or "robustness" or "robust" or "quality" or "agreement" or "effect of". The search was conducted on the 13/01/2020 and no start date limit was used.

Search outcomes

The results of the search for radiomics studies in lung cancer were screened by the title and abstract to find studies whose primary aim was either to create predictive radiomics models of clinical outcome or link radiomics to biology for NSCLC patients from CT images. Inclusion criteria were publications assessing outcomes of overall survival, metastases, treatment-induced toxicities or finding biological correlations. Studies using a modality other than CT, where the primary cancer was not NSCLC and review articles were not included in this step. 282 publications were found and after screening titles and abstracts based on the inclusion criteria, 116 publications remained. Exclusion criteria included CT studies not from planning CT, CBCT or diagnostic CT, if access to the article could not be gained, if the article was in a language other than English, if the study included deep learning as opposed to the traditional radiomics workflow discussed in this review, and studies predicting nodule malignancy. Studies of analysis reproducibility or methodology limitations were also excluded from this search, as they were included in the second evaluation. In all, 64 publications remained for analysis (Supplementary Figure 2.4). Included studies are summarized in Tables 2.1 and 2.2 and expanded in Supplementary Tables 2.6 and 2.7.

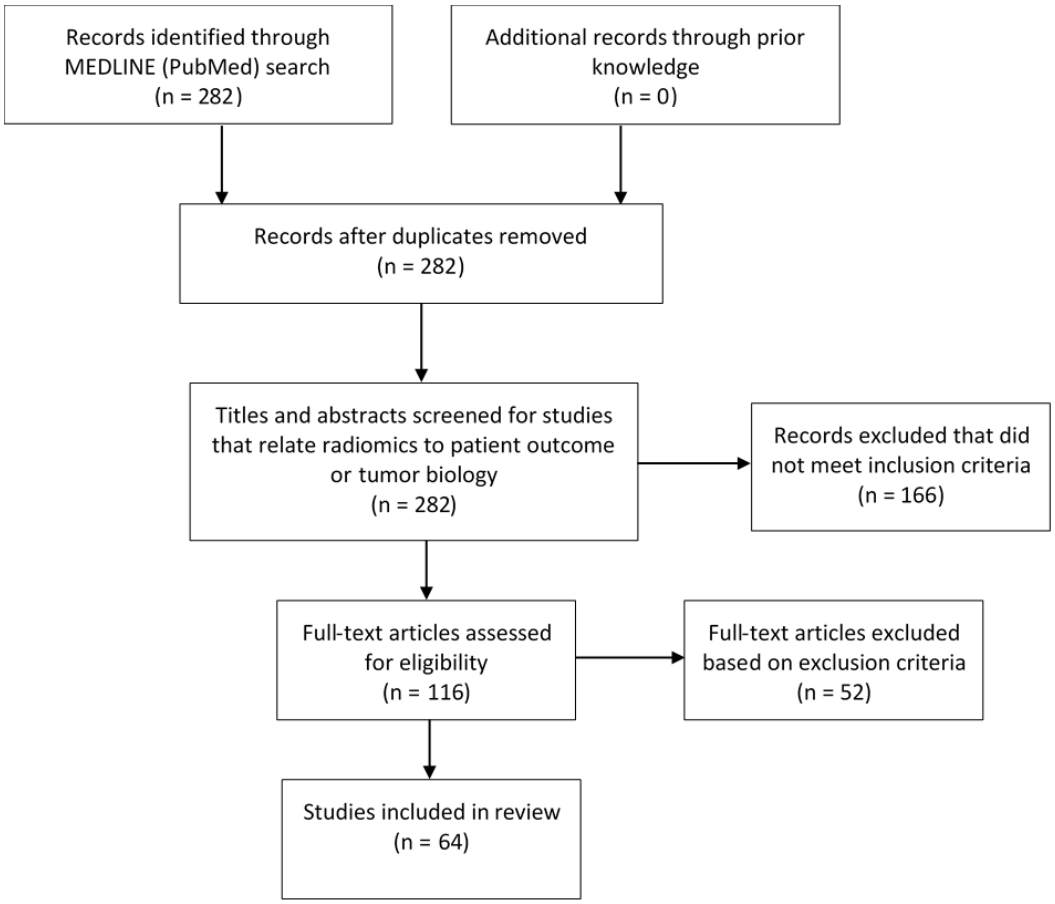

Figure 2.4: Flow diagram for the patient outcome and biology radiomics studies in lung cancer search outcomes.

The results of the search for studies of radiomics limitations were screened by titles and abstract to identify studies whose primary aim was to address a radiomics-based methodological concern using human or phantom scans. 489 publications were found and a further 3 studies previously known to the authors but not returned by our search were added to the results. After applying screening by inclusion criteria 132 studies remained. The following exclusion criteria were then applied: the study data was not CT-based, the CT data was from a cancer other than NSCLC or not clearly specified, the study investigated variability in deep learning models rather than the traditional radiomics workflow, and the article was a review or report of a published public dataset, rather than original research. This approach left 42 studies for inclusion in this review (Supplementary Figure 2.5). Included studies are presented in Supplementary Table 2.5 and summarized in Table 2.3.

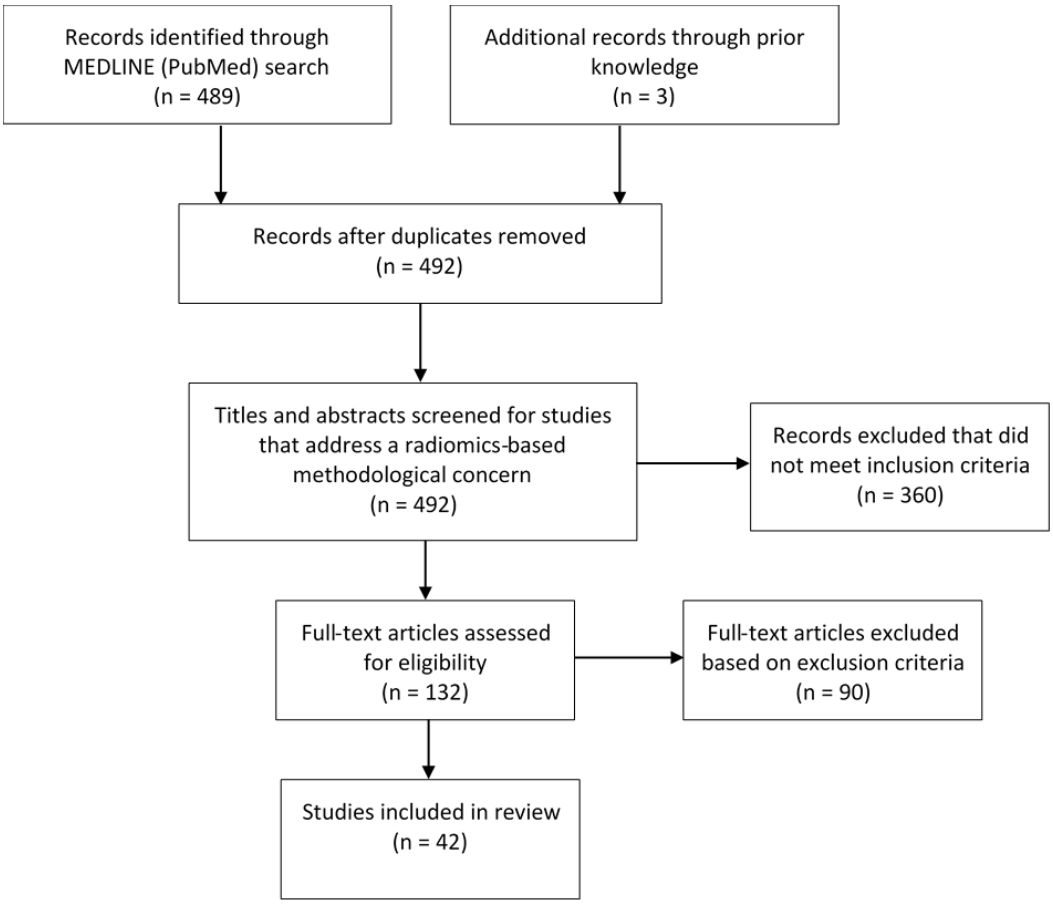

Figure 2.5: Flow diagram for the methodological radiomics studies in lung cancer search outcomes.

Search constraints and limitations

Conference abstracts were not included in the search. While this would have increased the number of studies included, abstracts had insufficient detail for our critical appraisal. Publication bias is a potential limitation of this review, since negative results are less likely to be published. This review included publications that investigated methodological concerns in the radiomics workflow for NSCLC and phantom CT scans and as such concerns that had been addressed in another cancer type or imaging modality were excluded from this analysis.

Table 2.5: Radiomics methodological studies selected for inclusion.

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Ger et al. 2018	Image acquisition. Pre-processing.	Updated CCR phantom. 20 NSCLC and 30 HNSCC CT scans.	IBEX.	First order and texture.	CT scanner, protocol and slice thickness affect feature values. In general, resampling does not change feature values or remove feature correlations with slice thickness.	Correct for the CT manufacturer and model by scanning a phantom on each scanner. Control or limit the range of slice thickness in studies.
Fave et al. 2015	Image acquisition. Test-retest. Volume dependence.	CCR phantom. 10 NSCLC CBCT test-retest scans.	IBEX.	First order and texture.	Features are more likely to be reproducible when the same CBCT scanner manufacturer and protocol are used. Increased motion reduces feature reproducibility.	A motion threshold of at most 10mm, preferably 5mm, increases feature reproducibility, as does excluding edges of the ROI. Texture features should not be compared across images acquired using different imaging protocols and CBCT manufacturers.
Lafata et al. 2018	Image acquisition.	Dynamic digital phantom simulation. 31 NSCLC free breathing CT scans, AIP and end of exhale 4DCT scans.	In-house Matlab.	Shape, first order and texture.	Image noise and motion affects feature reproducibility.	The end of exhale phase of a 4DCT is least affected by motion.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Du et al. 2019	Image acquisition.	20 NSCLC 4DCT scans. 140 NSCLC 4DCT scans.	3D Slicer.	Shape, first order and texture.	Motion affects feature reproducibility.	Assessing feature stability across all 4DCT phases can find features robust to motion which can improve the predictive performance of radiomic models.
Larue et al. 2017	Image acquisition. Pre-processing.	CCR phantom.	In-house.	First order and texture.	Feature values are different between CT scanners, even with similar acquisition protocols. Most radiomic features are affected by slice thickness. Choice of bin width influences feature values. The resampling method can induce variability in texture features.	Grey-level discretization could be optimized to improve prognostic value of features. Resampling decreases variability and reduces feature correlations with slice thickness. Cubic or linear interpolation induce less feature variability than nearest neighbour interpolation when resampling to 1x1x3mm3 voxels.
Mackin et al. 2015	Image acquisition.	CCR phantom. 20 NSCLC CT scans.	IBEX.	First order and texture.	Feature values are different between CT scanners.	Credentialing CT scanners could reduce variability across features measured on different scanners. CT scanners could be corrected for during data analysis.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Mackin et al. 2017	Image acquisition. Pre-processing.	CCR phantom. 8 NSCLC unconstructed CT scans.	IBEX.	Shape, first order and texture.	Variation in pixel size causes the intra-patient variability to be large relative to the inter-patient variability.	Feature variability due to differences in pixel size can be reduced through resampling and Butterworth low-pass filtering.
MacKin et al. 2018	Image acquisition.	CCR phantom.	IBEX.	First order and texture.	The impact of noise, i.e. tube current values, is more apparent for homogeneous materials than for textured.	Features are not substantially affected by variations in x-ray tube current. Tube current would not need to be harmonized across patients in a study.
Mahmood et al. 2017	Image acquisition. Image reconstruction.	Anthropomorphic phantom.	IBEX.	First order and texture.	Texture features are not reproducible across different CT scanners, even whilst using almost identical scanning parameters.	Robust correction factors need to be developed to reduce feature variability across CT scanners.
Midya et al. 2018	Image acquisition. Image reconstruction.	A uniform water phantom and an anthropomorphic phantom. A single abdominal CT scan.	In-house Matlab.	First order and texture.	CT scanner tube current, noise index and reconstruction technique influence feature reproducibility.	Only use features robust to changes in tube current, noise index and reconstruction technique.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Shafiq-ul-Hassan et al. 2017	Image acquisition. Image reconstruction.	CCR phantom.	In-house.	Shape, first order and texture.	Most texture features are reconstruction kernel dependent.	The variability in features due to different reconstruction kernels can be reduced by applying the Noise Power Spectrum peak frequency and ROI maximum intensity as correction factors.
Shafiq-ul-Hassan et al. 2017	Image acquisition. Pre-processing.	CCR phantom.	In-house.	Shape, first order and texture.	Features are dependent on voxel size and number of grey levels used for discretization.	Feature definitions can be normalized by voxel size and number of grey levels to reduce their dependency.
Yasaka et al. 2017	Image acquisition. Pre-processing.	CCR phantom.	TexRad.	First order.	Unfiltered and filtered features are variable across different CT scanners.	Feature variability due to different CT scanners needs to be taken into consideration.
Lu et al. 2016	Image acquisition. Image reconstruction.	32 NSCLC CT scans.	Not specified.	Shape, first order and texture.	Changing reconstruction algorithm and slice thickness affects feature values.	Image acquisition techniques needs to be standardized.
Zhao et al. 2016	Image acquisition. Test re-test. Image reconstruction.	31 NSCLC CT test-retest scans.	In-house.	Shape, first order and texture.	Features depend on the reconstruction algorithm used.	Features derived from images reconstructed with sharp and smooth algorithms should not be compared.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Li et al. 2018	Image acquisition. Image reconstruction.	51 NSCLC CT scans.	In-house Matlab.	Shape and texture.	Images with 1mm slice thickness give models with greater predictive performance than those with 5mm slice thickness.	Use thinner slice thickness to increase predictive performance of models.
Park et al. 2019	Image acquisition.	100 NSCLC CT scans.	Not specified.	First order and texture.	Radiomic features are not reproducible across different slice thicknesses.	Reproducibility can be improved by converting images to 1mm slice thickness using a convolutional neural network-based super-resolution algorithm.
Kim et al. 2019	Image acquisition. Image reconstruction.	Thoracic phantom.	Not specified.	First order and texture.	CT slice thickness, exposure setting and reconstruction algorithm affect radiomic features.	Image acquisition and reconstruction parameters need to be standardized to avoid variability.
Zhovannik et al. 2019	Image acquisition.	Phantom. 221 NSCLC pCT scans.	PyRadiomics.	First order and texture.	Radiomic features depend on scanner signal-to-noise ratio (exposure setting).	A correction algorithm can be modelled to make radiomic features reproducible across different signal-to-noise ratios.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Tunali et al. 2019	Image acquisition. Segmentation.	40 NSCLC CT scans. 32 NSCLC non-contrast enhanced CT scans. 212 NSCLC CT scans.	In-house C++.	First order and texture.	Radiomic features are not reproducible across multiple segmentations or different image acquisitions.	Use stable and reproducible radiomic features as a feature selection tool in the radiomics workflow.
Kakino et al. 2019	Image acquisition.	269 NSCLC diagnostic delayed phase CT scans.	PyRadiomics.	First order and texture.	Not all features are reproducible across contrast enhanced and non-contrast enhanced CT images. This also depends on patient characteristics.	Do not combine contrast enhanced and non-contrast enhanced images in radiomics analysis.
Hepp et al. 2020	Image acquisition.	69 NSCLC CT with simulated dose reduction.	PyRadiomics.	First order and texture.	Radiomic feature values differed when CT dose level was changed.	Differences in CT dose levels should be taken into account in radiomics studies.
Mahon et al. 2019	Image acquisition.	Gammex CT electron density phantom and Qasar body phantom. 135 NSCLC CT.	PyRadiomics.	Shape, first order and texture.	Variability in imaging protocols can induce variability in extracted radiomic features.	ComBat harmonization can harmonize radiomic features extracted from CT images using different imaging protocols.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Haga et al. 2018	Segmentation.	40 NSCLC maximum exhale 4DCT scans.	In-house Matlab.	Shape, first order and texture.	Differences in tumor segmentation can cause different features to become significant in the feature selection stage.	Multiple segmentation analysis can reveal features that are robust to delineation uncertainties. For good predictions ROIs need to be contoured by a specialist, such as a radiation oncologist.
Huang et al. 2017	Segmentation.	46 NSCLC CT scans.	In-house.	Shape, first order, texture and delta features.	Differences in tumor segmentation can lead to differences in a feature's predictive power.	Multiple segmentation analysis should be done to find robust features.
Kalpathy-Cramer et al. 2016	Segmentation.	40 NSCLC CT scans. Thoracic phantom.	Various.	Shape, first order and texture.	Some features are not robust to multiple segmentations.	Features need to be assessed for their robustness to segmentation and their usefulness in a predictive models.
Owens et al. 2018	Segmentation.	10 NSCLC CT scans.	IBEX.	Shape, first order and texture.	Segmentation is time consuming and subject to inter-observer variability.	Semi-automatic segmentations performed by non-specialists can give segmentations comparable to those from clinicians.
Parmar et al. 2014	Segmentation.	20 NSCLC CT scans.	In-house Matlab.	Shape, first order and texture.	Segmentation is time consuming and subject to inter-observer variability.	Semi-automatic segmentation led to a smaller range of feature values across observers than manual segmentation.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Pavic et al. 2018	Segmentation.	11 NSCLC, 11 HNSCC, 11 MPM CT scans.	In-house Python.	Shape, first order and texture.	Features affected by delineation uncertainty are different between cancer types. Shape features are most affected across all tumor types.	Filtering the image increases the number of stable features. Multiple segmentation analysis should be performed to find robust features. Averaging texture matrices rather than merging results in more stable features with respect to segmentation.
Shafiq-ul-Hassan et al. 2018	Pre-processing. Volume dependence.	CCR phantom. 18 NSCLC CT scans.	Not specified.	First order and texture.	Some texture features are not stable across a different number of grey levels used for discretization or voxels in the ROI.	Normalization by the number of grey levels or number of voxels In the ROI made some features more reproducible.
Fave et al. 2016	Pre-processing. Volume dependence.	107 NSCLC 4DCT end of exhale scans.	IBEX.	First order and texture.	Some features are entirely ROI volume dependent. Features tended to be more correlated with ROI volume after Butterworth smoothing.	Feature formulas can be corrected to remove ROI volume dependence.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Wang et al. 2019	Pre-processing. Volume dependence.	50 NSCLC CT scans.	Matlab open-source toolkit [185].	First order and texture.	Texture features may be sensitive to the number of grey levels used for discretization, the method of discretization and the use of an intensity threshold.	Discretization of 128 grey levels provides a set of reproducible texture features, regardless of discretization method. Thresholding the ROIs before feature extraction also improves reproducibility.
Welch et al. 2019	Volume dependence.	421 NSCLC CT scans.	PyRadiomics.	Shape, first order and texture.	Features from a previously published radiomic signature were found to be correlated with tumor volume.	Features should be tested for multicollinearity using statistical analysis or by data perturbation.
Choi et al. 2018	Volume dependence.	14 NSCLC free breathing pCT scans.	Not specified.	First order and texture.	Features may not be robust to variations in tumor size and may be correlated with the normal lung volume surrounding the tumor.	Simulations involving tumors of different sizes can reveal features robust to changes in tumor volume.
Larue et al. 2017	Test-retest.	26 NSCLC CT scans test-retest. 20 NSCLC and 20 oesophageal 4DCT scans. 120 oesophageal CT scans.	In-house.	Shape, first order and texture.	Test-retest is not always available for the phenotype of interest.	A 4DCT dataset can be used to find robust features across phases, as an alternative to a test-retest dataset.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Van Timmeren et al. 2016	Test-retest.	40 rectal cancer CT scans. 27 NSCLC CT scans.	Not specified.	Shape, first order and texture.	Test-retest results are not generalizable across different cancer types.	Test-retest analysis should be performed for each cancer type with controlled CT scanners and imaging protocols.
Zwanenburg et al. 2019	Test-retest.	31 NSCLC and 19 HNSCC CT scans.	In-house Python.	Shape, first order and texture.	Test-retest is not always available for the phenotype of interest.	Image perturbation could be an alternative to test-retest, giving multiple images to compare features across.
Tanaka et al. 2019	Test-retest.	14 NSCLC 4DCT scans. 14 NSCLC CT scans test-retest.	IBEX.	Shape, first order and texture.	Results from test-retest may not be generalizable to different CT protocols.	4DCT could be used as an alternative to test-retest imaging. Finding robust features across phases around the end-of-exhale phase rather than all 10 phases prevents excessive dimension reduction.
Parmar et al. 2015	Modelling.	464 NSCLC pCT scans (spiral thoracic CT with or without contrast).	In-house Matlab.	Shape, first order and texture.	Choice of classification method causes variation in a model's predictive performance.	Particular combinations of feature selection and classification methods give classification models with high predictive performance.
Sun et al. 2018	Modelling.	283 NSCLC pCT scans (spiral thoracic CT with or without contrast).	In-house Matlab.	Shape, first order and texture.	Statistical methods to predict overall survival differ in their predictive performance.	Particular combinations of feature selection and machine learning methods give survival models with high predictive performance.

Table 2.5 continued from previous page

Reference	Workflow stage(s)	Study data	Software	Features	Identified problems	Solution
Zhang et al. 2017	Modelling.	112 NSCLC CT scans.	In-house Matlab.	First order and texture.	Endpoints, feature selection and classification methods affect predictive performance.	Sample sizes above 50 give better predictive performance. Subsampling data to add to the minority class increases predictive performance.
Haga et al. 2019	Modelling.	40 NSCLC maximum exhale 4DCT scans. 29 NSCLC CT scans.	In-house Matlab.	Shape, first order and texture.	Feature normalization can affect predictive performance.	Performance of classification models can be improved by normalizing features, particularly z-score normalization.

Abbreviations: AUC, area under the curve; CBCT, cone-beam CT; CI, concordance index; DFS, disease free survival; DM, distant metastasis; GRD, gross residual disease; H&N, head and neck; HR, hazard ratio; LR, local relapse; LRR, local regional recurrence; LR-RFS, loco-regional recurrence-free survival; OS, overall survival; pCR, pathological complete response; pCT, radiotherapy planning CT scan; PFS, progression free survival; RFS, recurrence free survival.

Table 2.6: Radiomics studies in NSCLC, split into sections based on their investigated endpoint. The Data column specifies the total number of patients involved in the study, in brackets split by training and validation cohorts if applicable and specifying other cancer types of cohorts if applicable. Note: Studies marked with * are validation studies and their RQS score components refer to methodology based on the previous published data. The 'Added value?' column shows whether the added value of radiomics to a clinical model was tested.

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Overall survival											
Aerts et al. 2014 [91]	1-3b	647 pCT (422 + 225)	In-house MATLAB	Shape, first order and texture	CI=0.65	Test re-test, Multiple segmentations, Univariable analysis	Cox regression	3	19	A, B, C, D, G	Yes
Van Timmeren et al. 2017*	1-4	252 pCT and CBCT (102 + 56 + 94)	In-house MATLAB	Shape, first order and texture	CI=0.69, 0.61, 0.59 (pCT) CI=0.66, 0.63, 0.59 (CBCT)	Validation of Aerts et al. 2014 [91]	Validation of Aerts et al. 2014 [91]	4	16	A, B, C, D, F, G	No
Grossman et al. 2017*	1-3	351 diagnostic CT (262 + 89)	Not specified	Shape, first order and texture	CI=0.60	Validation of Aerts et al. 2014 [91]	Validation of Aerts et al. 2014 [91]	4	18	A, B, C, D, G	Yes

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Grossman et al. 2017	1-3	351 diagnostic CT (262 + 89)	Not specified	Not specified	CI=0.61	mRMR, Stepwise selection	Cox regression	3	9	A, B, C, D, E, G, H	Yes
Yu et al. 2017	1	442 diagnostic CT (147 + 295)	IBEX	First order and texture	CI=0.64	Multiple segmentations, Random survival forests, Correlation analysis, Correlation to tumor size, Univariable analysis	Cox regression	3	15	A, B, C	Yes
Chaddad et al. 2017	1-3b	315 pCT	In-house MATLAB	Shape and texture	Average AUC=0.70-0.76	None performed	Random forest	1b	6	A, B, C, D, E, G, H	Yes

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Fave et al. 2017	3	107 4DCT end of exhale, planning and CBCT	IBEX	Shape and texture	CI=0.672	CT model dependence, Correlation to tumor volume, Stepwise selection	Cox regression	1b	7	D, E, H	Yes
Li et al. 2017	1-2a	59 follow up CT	Definiens Developer	Texture	AUC=0.81	Correlation analysis, PCA, Univariable analysis, Stepwise selection or backward stepwise selection	Cox regression	1b	6	A, B, C, D, E, F, G, H	Yes

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Li et al. 2017	1-2a	92 4DCT Average-CT or 50% phase-CT images were used for analysis	Definiens Developer	Shape and first order	AUC=0.728	Correlation analysis, Stepwise selection or backward stepwise selection	Cox regression	1b	6	B, C, D, E, F, G, H	Yes
Tang et al. 2018	1-3	290 staging CT (114 + 176)	IBEX	Shape, first order and texture	CI=0.72	Multiple segmentations, Clustering, Univariable analysis	Cox regression	3	10	A, B, C, D, F, G, H	No
Bianconi et al. 2018	1-3	203 pCT	Not specified	Shape and texture	HR=1.06-1.48	Univariable analysis	Kaplan-Meier	1a	1	A, B, C, D, E, H	No
De Jong et al. 2018*	4	195 diagnostic CT	In-house MATLAB and CERR	Shape, first order and texture	CI=0.576	Validation of Aerts et al. 2014 [91]	Validation of Aerts et al. 2014 [91]	4	14	A, B, C, D, F, G	Yes

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRI-POD	RQS (max 36)	Methodological limitations	Added value?
Lee et al. 2018	1-3	339 CT (type not defined, just preoperative within 2 weeks before surgery)	In-house MATLAB	Shape, first order and texture	CI=0.772	Univariable analysis, Stepwise selection, LASSO	Cox regression	1b	5	A, B, C, D, E, G, H	Yes
He et al. 2018	1-3	186 CT (298 after oversampling (223 + 75)) type not defined	PyRadiomics	Not specified	AUC=0.9296	None performed	Random forest	2a	1	A, B, C, D, E, F, G, H	No
Starkov et al. 2018	1	116 pCT	MATLAB Generalized Riesz-Wavelet Toolbox v 1.0	Texture	High risk vs low risk median p-values=0.04–0.07	LASSO	Kaplan-Meier	1b	-5	A, B, C, D, E, F, H	No

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Yang et al. 2018	1-4	371 CT (239 + 132)	In-house MATLAB	First order and texture	CI=0.702	Multiple segmentations, LASSO	Cox regression	3	11	A, B, C, D, F, G, H	Yes
Wang et al. 2019	3	70 pretreatment and 97 post treatment CT from 118 patients	Not specified	Texture	CI=0.743	Multiple segmentations, Clustering, Random survival forest, Backward stepwise selection, Correlation analysis	Cox regression	1b	6	B, C, D, F, H	No
Shi et al. 2019	3	11 CBCT from 23 patients	IBEX	First order	HR=0.21	Test re-test, Multiple segmentation, Correlation analysis	Kaplan-Meier	1a	4	A, B, C, D	Yes

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Van Timmeren et al. 2019	1-4	337 pCT and 2154 CBCTs from 337 patients (141 + 94 + 61 + 41)	In-house MATLAB	First order and texture	CI=0.59, 0.54, 0.57	Correlation analysis, LASSO	Cox regression	3	19	B, C, D, E, G, H	Yes
Huang et al. 2019	1-4	371 CT (254 + 63 + 54)	In-house MATLAB	Shape, first order and texture	CI=0.621, 0.649	Test re-test, LASSO	Cox regression	2a	4	A, B, C, D, E, F, G	No
Franceschini et al. 2019	1-2	102 4DCT (start of inspiration) (70 + 32)	LIFEx	Shape and texture	AUC=0.85	Univariable analysis, Elastic net Backward stepwise selection	Cox regression	2a	2	C, D, E, G, H	No

Local or metastatic recurrence

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Coroller et al. 2015	2-3	182 pCT (98 + 84)	In-house MATLAB and CERR	First order and texture	CI=0.6	mRMR, Univariable analysis, Stepwise selection	Cox regression	2b	13	A, B, C, D, E, G, H	Yes
Mattonen et al. 2016	1	45 follow-up CT	In-house MATLAB	First order and texture	AUC=0.85	Stepwise selection	SVM	1b	-2	B, C, D, E, G, H	Yes
Huynh et al. 2016	1-2	113 CT (free breathing)	In-house MATLAB and 3D Slicer	First order and texture	Median CI=0.67	Test re-test, PCA, Univariable analysis	Cox regression	1b	6	A, B, C, D, E, G	Yes
Huynh et al. 2017	1-2a	112 free breathing CT and AIP CT	In-house MATLAB and 3D Slicer	Shape, first order and texture	AIP radiomics CI=0.667 FB radiomics CI=0.601	Test re-test, PCA, Univariable analysis, LASSO	Cox regression	1b	4	A, B, C, E, G	Yes

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Fave et al. 2017	3	107 4DCT end of exhale, planning and CBCT	IBEX	Shape and texture	CI=0.632, 0.558 (DM, LRR)	Stepwise selection	Cox regression	1b	7	D, E, H	Yes
Li et al. 2017	1-2a	59 follow up CT	Definiens Developer	Texture	AUC=0.80, 0.80 (RFS, LR-RFS)	Correlation analysis, PCA, Univariable analysis, Stepwise selection or backward stepwise selection	Cox regression	1b	6	A, B, C, D, E, F, H	Yes

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Li et al. 2017	1-2a	92 4DCT Average-CT or 50% phase-CT images were used for analysis	Definiens Developer	Shape	AUC=0.747, 0.690 (RFS, LL-RFS)	Correlation analysis, Univariable analysis, Stepwise selection or backward stepwise selection	Cox regression	1b	6	B, C, D, E, F, H	Yes
Dou et al. 2018	1-3	200 pCT (100 + 100)	PyRadiomics	Texture	CI=0.65	Test re-test, mRMR, Stepwise selection	Cox regression	2b	16	A, C, E, G	Yes
Ferreira Junior et al. 2018	1-4	68 CT (52 + 16)	IBEX	Shape and texture	AUC=0.75, 0.71 (lymph node metastasis, DM)	ReliefF	Naive Bayes, k-nearest neighbors and neural network	2a	9	A, B, C, D, E, F, G, H	No

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Yang et al. 2018	1-3	159 CT (106 + 53)	PyRadiomics	Shape, first order and texture	AUC=0.856	LASSO, Backward stepwise selection	Logistic regression	2b	13	B, E, F, G, H	Yes
Zhong et al. 2018	1-2	492 CT	MaZda	First order and texture	AUC=0.972	Multiple segmentations, ReliefF, PCA	SVM	1b	3	A, B, C, D, F, H, G	Yes
Lafata et al. 2019	1	70 CT	In-house MATLAB	Texture	Maximum AUC=0.72, 0.83, 0.60 (recurrence, LR, non-LR)	Univariable analysis, Truncated singular value decomposition, LASSO	Logistic regression	1b	2	A, B, C, D, E, F, G, H	No
Akinci D'Antonoli et al. 2019	1-2b	124 CT	Moddicom	Shape, first order and texture	AUC 0.731, 0.750 (LR, DM)	Univariable analysis, Stepwise selection	Cox regression	1b	13	A, B, E, F, G, H	Yes

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
He et al. 2019	Not specified	717 CT (423 + 294)	In-house MATLAB	First order and texture	CI=0.734	Multiple segmentations, Correlation analysis, Univariable analysis, LASSO, Backward stepwise selection	Logistic regression	2b	13	A, B, C, D, F, G, H	No
Xu et al. 2019	3-4	132 CT (106 + 26)	In-house MATLAB	Texture	AUC=0.642	Test re-test, LASSO	Cox regression	2a	12	B, C, D, E, G, H	No
Franceschini et al. 2019	1-2	102 4DCT (start of inspiration) (70 + 32)	LIFEx	Shape, first order and texture	AUC=0.73	Backward stepwise selection	Logistic regression	2a	2	C, D, E, G, H	No

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRI-POD	RQS (max 36)	Methodological limitations	Added value?
Ferreira-Junior et al. 2019	1-4	85 CT	IBEX	Shape, first order and texture	AUC=0.92, 0.84 (DM, nodal metastasis)	Univariable analysis, ReliefF	Neural network	1b	-2	A, B, C, G, H	No
Cong et al. 2019	1a	649 venous phase CT (455 + 194)	Artificial Intelligence Kit	Shape, first order and texture	AUC=0.851	Multiple segmentations, Univariable analysis, LASSO	Random forest	2a	14	B, C, D, G, H	Yes
Treatment response, disease-free or progression-free survival											
Coroller et al. 2016	2-3	127 pCT	In-house MATLAB and 3D Slicer	Shape, first order and texture	Median AUC=0.65, 0.61 (GRD, pCR)	Test re-test, PCA, Univariable analysis	Logistic regression	1b	7	A, B, C, D, E, G	Yes
Huang et al. 2016	1-2	282 CT (141 + 141)	In-house MATLAB	First order and texture	HR=2.09	Multiple segmentations, LASSO	Cox regression	2a	13	A, B, C, D, F, G, H	Yes

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Song et al. 2016	1-4	152 CT (80 + 72)	Not specified	Texture	HR= 2.35, 2.75	Univariable analysis	Cox regression	2a	4	A, B, C, D, E, F, G, H	No
Coroller et al. 2017	2-3	85 pCT	In-house MATLAB and 3D Slicer	Shape, first order and texture	Median AUC=0.68, =0.71 (pCR, GRD)	Test re-test, PCA, Univariable analysis	Random forest	1b	3	A, B, C, D, E, G	Yes
Tunali et al. 2019	3b-4	228 CT	In-house MATLAB and C++	Texture	AUC=0.804	Test re-test, Univariable analysis, Correlation to tumor volume, Backwards stepwise selection	Logistic regression	1a	5	A, B, D, E	Yes
Franceschini et al. 2019	1-2	102 4DCT (start of inspiration) (70 + 32)	LIFEx	Texture	AUC=0.88	Univariable analysis, Elastic net, Backward stepwise selection	Cox regression	2a	2	C, D, E, G, H	No

Table 2.6 continued from previous page

Reference	Stage	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations	Added value?
Lung toxicity											
Moran et al. 2017	1	14 diagnostic CT	Not specified	First order and texture	AUC=0.689-0.750	Univariable analysis	Logistic regression	1a	-2	A, B, C, E, G, H	Yes
Krafft et al. 2018	Not specified	192 50% 4DCT phase	In-house MATLAB	First order and texture	Average AUC=0.68	LASSO	Logistic regression	1b	0	A, E, G, H	Yes
Staging											
Yuan et al. 2018	1	327 CT	Artificial intelligence kit	First order and texture	AUC=0.938	Recursive feature elimination	SVM	1b	2	A, B, C, F, G, H	No
Yang et al. 2019	1-3	256 CT	PyRadiomics	First order and texture	AUC= 0.93	LASSO	Logistic regression	1b	-3	B, C, E, F, G, H	No

Abbreviations: AUC, area under the curve; CBCT, cone-beam CT; CI, concordance index; DFS, disease free survival; DM, distant metastasis; GRD, gross residual disease; H&N, head and neck; HR, hazard ratio; LR, local relapse; LRR, local regional recurrence; LR-RFS, loco-regional recurrence-free survival; OS, overall survival; pCR, pathological complete response; pCT, radiotherapy planning CT scan; PFS, progression free survival; RFS, recurrence free survival.

Table 2.7: Radiomics studies in NSCLC with an aspect of biology as the endpoint. The Data column specifies the total number of patients involved in the study, in brackets split by training and validation cohorts if applicable and specifying other cancer types of cohorts if applicable.

Reference	Stage	Endpoint	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations
Genomics											
Aerts et al. 2016	Early stage	EGFR	47 diagnostic CT and follow-up	Not specified	Shape and texture	AUC=0.74-0.91	Coefficient of variation Correlation analysis Univariable analysis	Logistic regression	1a	2	B, E, F
Rios Velazquez et al. 2017	1-4	EGFR, KRAS	705 diagnostic CT (353 + 352)	In-house plug in for 3D Slicer	Shape, first order and texture	AUC=0.69-0.80	Test retest, PCA, mRMR	Random forest	3	14	A, B, C, D, E, F
Mei et al. 2018	Not specified	EGFR	296 CT	PyRadiomics	Texture	AUC=0.664	Univariable analysis	Logistic regression	1a	-2	A, B, C, E, F, G, H
Digumarthy et al. 2019	Not specified	EGFR	93 CT	TexRAD	First order	AUC=0.713	Univariable analysis	Logistic regression	1a	-1	A, B, C, E, F, G, H

Table 2.7 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRI-POD	RQS (max 36)	Methodological limitations
Jia et al. 2019	1-4	EGFR	504 CT (345 + 158)	Not specified	Shape, first order and texture	AUC=0.802	Univariable analysis	Random forest	2a	5	A, B, D, E, F, G, H
Li et al. 2019	1-4	EGFR subtypes (19Del and L858R)	312 CT (236 + 76)	In-house C++	Shape and first order	AUC= 0.775-0.793	Multiple segmentations, Univariable analysis, Stepwise selection	Logistic regression	2b	14	B, C, F, G, H
Tu et al. 2019	1-4	EGFR	404 CT (243 + 161)	In-house MATLAB	First order and texture	AUC=0.775	Multiple segmentations, Univariable analysis, Clustering, Backwards stepwise selection	Logistic regression	2a	13	A, B, C, D, F, G, H

Table 2.7 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations
Yang et al. 2019	Not specified	EGFR	467 CT (306 + 161)	PyRadiomics	Shape, first order and texture	AUC=0.789	Mean decrease impurity importance from random forest	Random forest	2a	13	B, D, E, F, G, H
Wang et al. 2019	1-2	EGFR, TP53	61 CT (41 + 20)	PyRadiomics	First order and texture	AUC=0.604, 0.586	LASSO	SVM	2a	10	B, C, E, F, G, H
Wang et al. 2019	1-2	Tumor mutation burden	61 CT (41 + 20)	PyRadiomics	Texture	AUC=0.606	LASSO	SVM	2a	10	B, C, E, F, G, H
Signaling pathways											
Grossman et al. 2017	1-3	Various	351 CT (262 + 89)	Not specified	Shape, first order and texture	AUC=0.62-0.72	Clustering	Logistic regression	3	9	A, B, C, D, E, G, H
Bak et al. 2018	1-4	Various	57 CT	In-house MATLAB	First order and texture	OR=0.08-23.94	Univariable analysis	Logistic regression	1a	-5	B, C, E, F, G, H

Table 2.7 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations
Histopathology											
Patil et al. 2016	Not specified	ADC, LCC, SCC, NOS	317 pCT	In-house MATLAB	Shape, first order and texture	88% accuracy	None	SVM	1b	3	A, B, C, D, E, G, H
Wu et al. 2016	1-4	ADC, SCC	350 pCT (198 + 152)	In-house MATLAB	First order and texture	AUC=0.72	Correlation analysis, Univariable analysis	Random forest, naive Bayes, and k-nearest neighbors	3	13	A, B, C, D, E, G, H
Ferreira Junior et al. 2018	1-4	ADC, SCC	68 CT (52 + 16)	IBEX	Not specified	AUC=0.81	ReliefF	Naive Bayes and k-nearest neighbors and neural network	2a	6	A, B, C, D, E, G, H

Table 2.7 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations
Zhu et al. 2018	Not specified	ADC, SCC	129 CT (81 + 48)	In-house MATLAB	First order and texture	AUC=0.893	Multiple segmentations, LASSO	Logistic regression	2a	11	A, B, C, D, F, G, H
Digumarthy et al. 2019	Not specified	ADC, SCC	93 CT	TexRAD	First order	AUC=0.744	Univariable analysis	Logistic regression	1a	-1	A, B, C, E, F, G, H
E et al. 2019	Not specified	ADC, SCC, SCLC	229 CT	In-house MATLAB	Shape, first order and texture	AUC=0.657-0.875	Test re-test, Clustering, mRMR, Incremental forward search	Naive Bayes, logistic regression and random forest	1b	5	B, C, E, F, G
Ferreira-Junior et al. 2019	1-4	ADC, SCC	85 CT	IBEX	Shape, first order, texture	AUC=0.88	Univariable analysis, ReliefF	Neural network	1b	-2	A, B, C, G, H

Table 2.7 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations
Liu et al. 2019	Not specified	ADC, LCC, SCC, NOS	349 CT (278 + 71)	Not specified	Not specified	AUC=0.86	l2,1-norm minimization	SVM	3	4	A, B, C, D, E, F, G, H
Zhou et al. 2018	1-4	Ki-67	110 CT	3D Slicer	Shape and texture	AUC=0.61-0.77	Univariable analysis, Backwards stepwise selection	Logistic regression	1a	11	B, D, E, F, G, H

Table 2.7 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations
Gu et al. 2019	Not specified	Ki-67	245 CT	MaZda	First order and texture	AUC=0.776	Feature selection algorithm based on random forest	Logistic regression, linear discriminant analysis, classification tree and regression tree, k-neighbour clustering, SVM and random forest	1b	-2	A, B, C, D, E, F, G, H

Table 2.7 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations
Song et al. 2017	1-3	Micropapillary pattern	339 CT	Not specified	First order	AUC=0.751	Multiple segmentation, Univariable analysis, Stepwise selection	Logistic regression	1b	8	B, C, D, F, G, H
Chen et al. 2018	Not specified	Degree of differentiation	487 CT (303 + 184)	In-house MATLAB	First order and texture	AUC=0.782	Univariable analysis, mRMR, Backwards stepwise selection	Logistic regression	2b	3	A, B, C, D, E, F, G, H
She et al. 2018	Not specified	Invasive vs non-invasive ADC	402 CT (207 + 195)	In-house Python	Shape, first order and texture	AUC=0.89	LASSO	Logistic regression	2b	8	A, B, C, D, E, F, G, H

Table 2.7 continued from previous page

Reference	Stage	Endpoint	Data (training + validation)	Software	Radiomic features in final model	Result	Feature selection	Model building	TRIPOD	RQS (max 36)	Methodological limitations
Yang et al. 2019	Not specified	Invasive vs non-invasive ADC	192 CT (116 + 76)	Artificial intelligence kit	First order and texture	AUC=0.77	Multiple segmentations, Correlation analysis, LASSO	Logistic regression	2a	13	B, C, D, F, G, H

Abbreviations: ADC, adenocarcinoma; AUC, area under the curve; CI, concordance index; EGFR, epidermal growth factor receptor; KRAS, Kirsten rat sarcoma viral oncogene homolog; LCC, large cell carcinoma; NOS, not otherwise specified; OR, odds ratio; SCC, squamous cell carcinoma.

Table 2.8: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) analysis types [200].

Analysis type	**Description**
Type 1a	Development of a prediction model where predictive performance is then directly evaluated using exactly the same data (apparent performance).
Type 1b	Development of a prediction model using the entire data set, but then using resampling (e.g. bootstrapping or cross-validation) techniques to evaluate the performance and optimism of the developed model.
Type 2a	The data are randomly split into two groups: one to develop the prediction model, and one to evaluate its predictive performance.
Type 2b	The data are non-randomly split (e.g. by location or time) into two groups: one to develop the prediction model and one to evaluate its predictive performance.
Type 3	Development of a prediction model using one data set and an evaluation of its performance on separate data (e.g. from a different study).
Type 4	The evaluation of the predictive performance of an existing (published) prediction model on separate data.

Table 2.9: The radiomics quality score (RQS) scoring criteria developed by Lambin et al. [201].

Criteria	Points
Image protocol quality – well-documented image protocols (e.g., contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/ replicability	+1 (if protocols are well-documented) +1 (if public protocol is used)
Multiple segmentations – possible actions are: segmentation by different physicians/ algorithms/-software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyze feature robustness to segmentation variabilities	+1
Phantom study on all scanners – detect inter-scanner differences and vendor-dependent features. Analyze feature robustness to these sources of variability	+1
Imaging at multiple time points – collect individuals' images at additional time points. Analyze feature robustness to temporal variabilities (e.g., organ movement, organ expansion/shrinkage).	+1
Feature reduction or adjustment for multiple testing – decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features	+3 (if neither measure is implemented) +3 (if either measure is implemented)
Multivariable analysis with non radiomic features (e.g., EGFR mutation) – is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non radiomics features	+1
Detect and discuss biological correlates – demonstration of phenotypic differences (possibly associated with underlying gene–protein expression patterns) deepens understanding of radiomics and biology	+1
Cut-off analyses – determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results	+1
Discrimination statistics – report discrimination statistics (e.g., C-statistic, ROC curve, AUC) and their statistical significance (e.g., p-values, confidence intervals). One can also apply resampling method (e.g., bootstrapping, cross-validation)	+1 (if a discrimination statistic and its statistical significance are reported) +1 (if also an resampling method technique is applied)

Table 2.9 continued from previous page

Criteria	Points
Calibration statistics – report calibration statistics (e.g., Calibration-in-the-large/slope, calibration plots) and their statistical significance (e.g., p-values, confidence intervals). One can also apply resampling method (e.g., bootstrapping, cross-validation)	+1 (if a calibration statistic and its statistical significance are reported) +1 (if also an resampling method technique is applied)
Prospective study registered in a trial database – provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker	+7 (for prospective validation of a radiomics signature in an appropriate trial)
Validation – the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance	-5 (if validation is missing) +2 (if validation is based on a dataset from the same institute) +3 (if validation is based on a dataset from another institute) +4 (if validation is based on two datasets from two distinct institutes) +4 (if the study validates a previously published signature) +5 (if validation is based on three or more datasets from distinct institutes) *Datasets should be of comparable size and should have at least 10 events per model feature.
Comparison to 'gold standard' – assess the extent to which the model agrees with/is superior to the current 'gold standard' method (e.g., TNM-staging for survival prediction). This comparison shows the added value of radiomics	+2
Potential clinical utility – report on the current and potential application of the model in a clinical setting (e.g., decision curve analysis)	+2

Table 2.9 continued from previous page

Criteria	Points
Cost-effectiveness analysis – report on the cost-effectiveness of the clinical application (e.g., quality adjusted life years generated)	+1
Open science and data – make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study	+1 (if scans are open source) +1 (if region of interest segmentations are open source) +1 (if code is open source) +1 (if radiomics features are calculated on a set of representative ROIs and the calculated features + representative ROIs are open source
	Total points (36 = 100%)

Chapter 3

Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform

This chapter has been published in European Radiology 2020 Volume 30 Issue 11 p6241-6250. The section numbering, references and formatting have been modified to be consistent with the rest of the thesis.

Authors

Isabella Fornacon-Wood[1], Hitesh Mistry[1], Christoph J Ackermann[2], Fiona Blackhall[1,3], Andrew McPartlin[4], Corinne Faivre-Finn[1,4], Gareth J Price[1] and James P B O'Connor[1,5]

Affiliations

[1] Division of Cancer Sciences, University of Manchester, Manchester, UK.

[2] Department of Medical Oncology, Spital STS AG, Thun, Switzerland.

[3] Department of Medical Oncology, The Christie NHS Foundation Trust, Manchester,

UK.

[4] Department of Radiation Oncology, The Christie NHS Foundation Trust, Manchester, UK.

[5] Department of Radiology, The Christie NHS Foundation Trust, Manchester, UK.

Author contributions

I collated the CT images for each cohort and extracted the gross tumour volume from the radiotherapy structure set for the H&N and SCLC cohorts. I extracted the radiomic features using the four different software platforms and wrote the R script to compare the features across platforms. I wrote the PostgreSQL script to extract survival data for the H&N cohort, and the R script to perform survival analysis. I wrote the manuscript, which was reviewed by all co-authors.

Abstract

Objective

To investigate the effects of Image Biomarker Standardisation Initiative (IBSI) compliance, harmonisation of calculation settings and platform version on the statistical reliability of radiomic features and their corresponding ability to predict clinical outcome.

Methods

The statistical reliability of radiomic features was assessed retrospectively in three clinical datasets (patient numbers: 108 head and neck cancer, 37 small-cell lung cancer, 47 non-small-cell lung cancer). Features were calculated using four platforms (PyRadiomics, LIFEx, CERR and IBEX). PyRadiomics, LIFEx and CERR are IBSI-compliant, whereas IBEX is not. The effects of IBSI compliance, user-defined calculation settings and platform version were assessed by calculating intraclass correlation coefficients and confidence intervals. The influence of platform choice on the relationship between radiomic biomarkers and survival was evaluated using univariable cox regression in the largest dataset.

Results

The reliability of radiomic features calculated by the different software platforms was only excellent (ICC > 0.9) for $4/17$ radiomic features when comparing all four platforms. Reliability improved to ICC > 0.9 for $15/17$ radiomic features when analysis was restricted to the three IBSI-compliant platforms. Failure to harmonise calculation settings resulted in poor reliability, even across the IBSI-compliant platforms. Software platform version also had a marked effect on feature reliability in CERR and LIFEx. Features identified as having significant relationship to survival varied between platforms, as did the direction of hazard ratios.

Conclusion

IBSI compliance, user-defined calculation settings and choice of platform version all influence the statistical reliability and corresponding performance of prognostic models in radiomics.

3.1 Introduction

There is considerable current interest in calculating features from medical images using high-throughput methods and then relating these features to clinical endpoints [88, 201]. This approach has been termed 'radiomics'. The principal hypothesis is that medical images contain information beyond that identified readily by traditional radiological examination, and that this information can be extracted through advanced image analysis. Since imaging plays a key role in cancer diagnosis, treatment and follow-up, radiomics provides potential non-invasive and inexpensive methods for developing biomarkers for prognosis and/or prediction in oncology.

The potential value of radiomic biomarkers has been well documented [88, 89], but recent literature have highlighted potential barriers to the translation of radiomics into useful decision-making tools [48, 90]. For example, studies have demonstrated that radiomic features can be heavily influenced by scanner acquisition and reconstruction parameters [166, 203], or inter-observer variability in defining target lesions [184], both of which influence model performance [179, 204].

One critical aspect of the radiomics workflow that remains relatively unexamined is the implementation of the software platforms used to calculate radiomic features. Many radiomic software platforms are reported in the literature, ranging from in-house developments [91], to open-source [205–207], freeware [208] and commercial offerings [209]. With in-house and commercial products, the source code for calculating features is not always publicly available. This can prevent comparison of results between studies in the literature. This is contrary to current moves towards an open-science approach in 'big data' analyses and in artificial intelligence, where open-source and freeware developers publish feature definitions alongside software code, including the values chosen for any calculation settings, and the user-defined free parameters that are required for the calculation of some features [210].

Several studies have previously demonstrated that features can vary when calculated in different software platforms [197, 211, 212]. The Image Biomarker Standardisation Initiative (IBSI) is an international collaboration developed to help standardise

radiomic feature calculation and has provided a framework to deliver practical solutions to this problem [51]. The IBSI has made recommendations concerning feature calculation, standardised feature definition and nomenclature. It has also provided a digital phantom with benchmark values to validate feature calculation platforms (to become IBSI-compliant) [213]. However, IBSI does not address calculation settings or evaluate versions of software.

In this article, we expand on this work by looking in three clinical datasets. We aimed to investigate the effects of IBSI compliance, harmonisation of calculation settings and choice of platform version on the statistical reliability of radiomic features and their corresponding ability to predict clinical outcome.

3.2 Methods and materials

In this study, we evaluated three different clinical datasets using four different radiomic feature calculation platforms.

3.2.1 Patient data

Data analysis was performed following institutional board approval and was compliant with UK research governance (ref. 17/NW/0060). We examined three datasets:

1. One hundred eight radiotherapy planning contrast-enhanced CT scans from patients with oropharyngeal head and neck (H&N) cancer treated with either chemo-radiotherapy or radiotherapy alone at The Christie NHS Foundation Trust, Manchester, UK.

2. Thirty-seven radiotherapy planning contrast-enhanced CT scans from a cohort of patients with small-cell lung cancer (SCLC) who had been enrolled in the CONVERT trial [214], acquired in nine different institutions (supplementary material A).

3. Forty-seven diagnostic contrast-enhanced CT scans from a cohort of patients with stage 4 non-small-cell lung cancer (NSCLC) cancer treated with first-line immunotherapy at The Christie NHS Foundation Trust, Manchester, UK.

The gross tumour volume, the extent of the visible tumour on the CT scan, was extracted from the radiotherapy structure set for both the H&N and SCLC cohorts. Original contours were drawn by the treating physician using the Pinnacle3 Treatment Planning system (versions 8.0, 9.0, 9.8 or 16.0, Philips Healthcare) and used as the analysis region of interest (ROI). Twelve H&N and 10 SCLC patients did not have contrast due to poor renal function or IV access. For the NSCLC dataset, ROIs were drawn by a thoracic oncologist (C.A.; 5 years' experience) using the same Pinnacle software (version 9.8). ROIs were checked by a board-certified radiologist J.O.C.: 14 years' experience). Full details of patient cohorts, image acquisition and reconstruction are detailed in Supplementary Tables 3.4 and 3.5.

3.2.2 Radiomic software platform selection

To our knowledge, 14 different radiomics software platforms are reported in the literature (Table 3.1) [205–208, 215–220]. Four of these software platforms are freely available, used widely in the literature and have mathematical equations documented to sufficient detail to understand the basis for their analysis.

Table 3.1: Details of various software packages available for radiomic feature calculation. The listed number of citations are those that cite the initial publication introducing the platform according to PubMed (search on 30/01/2020)

Software	**Year of publication**	**Citations**	**IBSI-compliant?**	**Free?**	**Open source?**	**Feature sets calculated**	**Mathematical equations documented?**
MaZda [216]	2009	366	✗	✓	✗	Shape, intensity and texture	✗
Chang-Gung Image Texture Analysis (CGITA) [217]	2014	65	✗	✓	✓	Intensity and texture	✗
IBEX [206]	2015	134	✗	✓	✓	Shape, intensity and texture	✓
Moddicom [218]	2015	13	✗	✓	✓	Shape, intensity and texture	✗
PyRadiomics [207]	2017	324	✓	✓	✓	Shape, intensity and texture	✓
LIFEx [208]	2018	84	✓	✓	✗	Shape, intensity and texture	✓
Quantitative Image Feature Engine (QIFE) [219]	2018	13	✗	✓	✓	Shape, intensity and texture	✗
CERR [205]	2018	25	✓	✓	✓	Shape, intensity and texture	✓

Table 3.1 continued from previous page

Software	Year of publication	Citations	IBSI-compliant?	Free?	Open source?	Feature sets calculated	Mathematical equations documented?
MITK Phenotyping [220]	2019	6	✓	✓	✓	Shape, intensity and texture	✓
RaCat [215]	2019	4	✓	✓	✓	Shape, intensity and texture	✗
PORTS v.1.1 matlab software (www.ncihub.org/resources/1663)	Not published	Not published	✗	✓	✓	Intensity and texture	✓
MatLab package (www.github.com/mvallieres/radiomics)	Not published	Not published	✓	✓	✓	Shape, intensity and texture	✓
TexRad	Not published	Not published	Unknown	✗	✗	Unknown	Unknown
Oncoradiomics	Not published	Not published	Unknown	✗	✗	Unknown	Unknown

For all of the study, we used the latest version of the following platforms: LIFEx v5.47 [208], IBEX v1.0 beta [206], PyRadiomics v2.2.0 [207] and the Computational Environment for Radiological Research (CERR) commit a1c8181 (05/09/2019) available at https://github.com/cerr/CERR [205]. Notably, LIFEx, PyRadiomics and CERR claim compatibility with the IBSI standard, whereas IBEX does not (Table 3.1).

For the comparison between software versions, we used LIFEx v5.1, CERR commit 50530f7 (29/08/2019) and PyRadiomics v2.1.2. IBEX has only released one version.

3.2.3 Feature calculation

We analysed radiomic features common to the four software platforms. These 17 features included three shape parameters, four intensity feature, one histogram feature, six 3D grey level co-occurrence matrix (GLCM) features and three 3D neighbourhood grey tone difference matrix (NGTDM) features measuring ROI heterogeneity (Table 3.2; example of the shape feature 'sphericity' shown in Figure 3.1). Since naming conventions for these features are not consistent across software (see Table 3.2), we used the feature names most closely in keeping with IBSI nomenclature, but simplified where appropriate. No image pre-processing was performed.

Table 3.2: Differences in naming conventions defined by the IBSI across the radiomic software.

Feature	**IBSI terminology**	**LIFEx**	**IBEX**	**PyRadiomics**	**CERR**
Volume	Volume (mesh) and volume (voxel counting)	Volume	Volume	Mesh volume and voxel volume	Volume
Sphericity	Sphericity	Sphericity	Sphericity	Sphericity	Sphericity
Area	Surface area (mesh)	Surface area	Surface area	Surface area	Surface area
Skewness	Discretised intensity skewness	Histogram skewness	Intensity histogram skewness	First-order skewness	Skewness
GLCM correlation	GLCM correlation	GLCM correlation	GLCM correlation	GLCM correlation	GLCM correlation
GLCM contrast	GLCM contrast	GLCM contrast = variance	GLCM contrast	GLCM contrast	GLCM contrast
GLCM angular Second moment	GLCM angular Second moment	GLCM energy = angular second moment	GLCM energy	GLCM joint energy	GLCM joint energy
GLCM joint entropy	GLCM joint entropy	GLCM entropy Log2 = joint entropy	GLCM entropy	GLCM joint entropy	GLCM joint entropy
GLCM difference average	GLCM difference average	GLCM dissimilarly	GLCM dissimilarly	GLCM difference average	Dissimilarity (difference average)

Table 3.2 continued from previous page

Feature	IBSI terminology	LIFEx	IBEX	PyRadiomics	CERR
GLCM inverse difference	GLCM inverse difference	GLCM homogeneity = inverse difference	GLCM homogeneity	GLCM ID	GLCM inverse difference
NGTDM busyness	NGTDM busyness	NGLDM busyness	Neighbour intensity difference busyness	NGTDM busyness	NGTDM busyness
NGTDM coarseness	NGTDM coarseness	NGLDM coarseness	Neighbour intensity difference coarseness	NGTDM coarseness	NGTDM coarseness
NGTDM contrast	NGTDM contrast	NGLDM contrast	Neighbour intensity difference contrast	NGTDM contrast	NGTDM contrast
Minimum	Minimum intensity	Conventional HU minimum	Global Minimum	First-order minimum	Minimum
Maximum	Maximum intensity	Conventional HU maximum	Global maximum	First-order maximum	Maximum
Mean	Mean intensity	Conventional HU mean	Global mean	First-order mean	Mean

Table 3.2 continued from previous page

Feature	**IBSI terminology**	**LIFEx**	**IBEX**	**PyRadiomics**	**CERR**
Standard deviation	Not defined (variance is defined)	Conventional HU standard deviation	Global standard deviation	First-order standard deviation	Standard deviation

Abbreviations: *ID*, inverse difference; *GLCM*, grey-level co-occurrence matrix; *HU*, Hounsfield Unit; *NGLDM*, neighborhood grey-level different matrix; *NGTDM*, neighboring grey tone difference matrix.

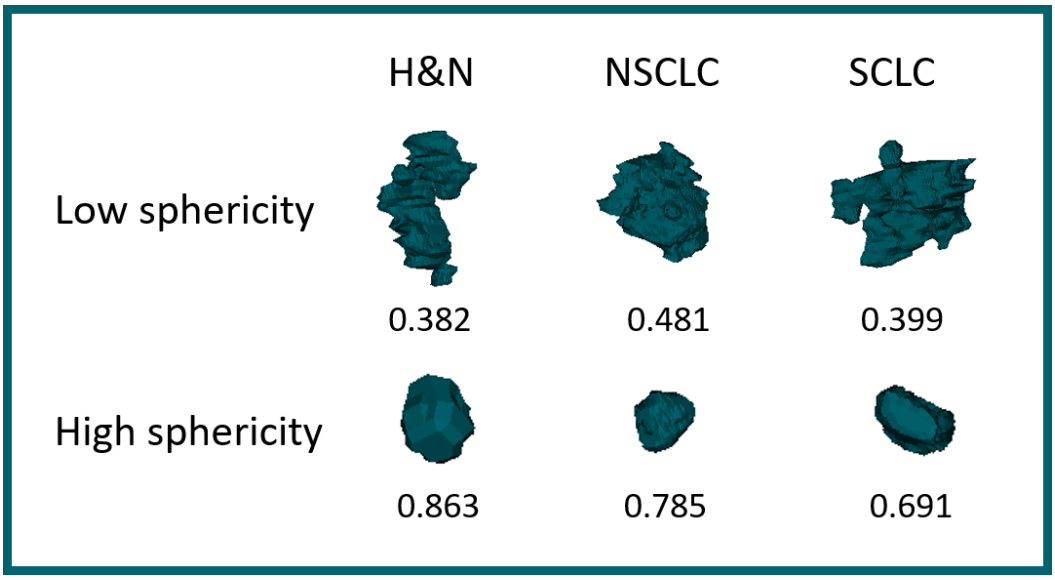

Figure 3.1: Example tumours and corresponding values for the feature 'sphericity' from each dataset.

The absolute numerical value of some radiomic features depend heavily on choice of default or user-defined settings. For example, the number of bins used to discretise image intensities do not have consistent default values across the platforms (see Table 3.3). Therefore, as well as performing inter-platform comparison of the results from different platforms, we also investigated the effect harmonising these parameters to common values. The harmonised calculation settings are presented in Table 3.3. Differences between platforms are detailed in supplementary material B.

Table 3.3: Default calculation settings for each software platform along with the harmonised settings used in this study.

Calculation settings	**LIFEx**	**IBEX**	**PyRadiomics**	**CERR**	**Harmonised settings (this study)**
Histogram					
Number of grey levels	400	256	Bin width 25	Bin width 25	64
Lower bound	- 1000	0	Minimum	0	Minimum
Upper bound	3000	4096	Maximum	500	Maximum
GLCM					
Number of grey levels	400	100	Bin width 25	Bin width 25	64
Lower bound	- 1000	0	Minimum	0	Minimum
Upper bound	3000	2100	Maximum	500	Maximum
Directions	13	13	13	4	13
Offset	1	1, 4 and 7	1	1	1
Symmetric	Yes	Yes	Yes	Yes	Yes
NGTDM					
Number of grey levels	400	256	Bin width 25	Bin width 25	64
Lower bound	- 1000	0	Minimum	0	Minimum
Upper bound	3000	4096	Maximum	500	Maximum
Distance	1	2	1	1	1

Abbreviations: *GLCM*, grey-level co-occurrence matrix; *NGTDM*, neighboring grey tone difference matrix.

3.2.4 Statistical analysis

To assess the effect of software platform variation on the reliability of radiomic biomarkers, we calculated two-way mixed effect intraclass correlation coefficients (ICC) and their 95% confidence intervals (CIs) for each feature. The ICC quantifies the absolute agreement between features computed by each platform. The ICC estimates and CI were stratified to indicate poor (ICC CI < 0.5), moderate ($0.5 <$ ICC CI < 0.75), good ($0.75 <$ ICC CI < 0.9) and excellent (ICC CI > 0.9) reliability [221]. Negative ICC estimates and CI were truncated at zero.

To assess the effect of software platform variation on the relationship of radiomic biomarkers to clinical outcome, we applied univariable cox regression against overall survival in the H&N dataset for each feature in Table 3.2. We repeated this analysis for each software platform using both their default calculation settings and the harmonised settings. Feature values were normalised to uniform scale (mean 0, standard deviation 1) to permit relative comparison of effect sizes.

All statistical analyses were performed in R 3.5.2 [222] with packages irr v0.84 [223] and survival v2.44.1.1 [224].

3.3 Results

3.3.1 Poor radiomic biomarker reliability across software platforms is improved by IBSI standardisation

We assessed the statistical reliability between radiomic features calculated from four software platforms using harmonised calculation settings in three clinical datasets. The distribution of feature values across all platforms and cohorts is available in the supplementary data. In each case, ICC and confidence intervals were derived (Figure 3.2A). Reliability between all four software was excellent (ICC CI > 0.9) in all datasets for only 4/17 features (volume, skewness, mean and maximum intensity). Reliability between software was poor (ICC CI < 0.5) in all datasets for 6/17 features (sphericity, some GLCM features and all NGTDM features). The other features had moderate or good reliability. Overall, the level of reliability for each individual feature

was highly consistent across the three clinical datasets.

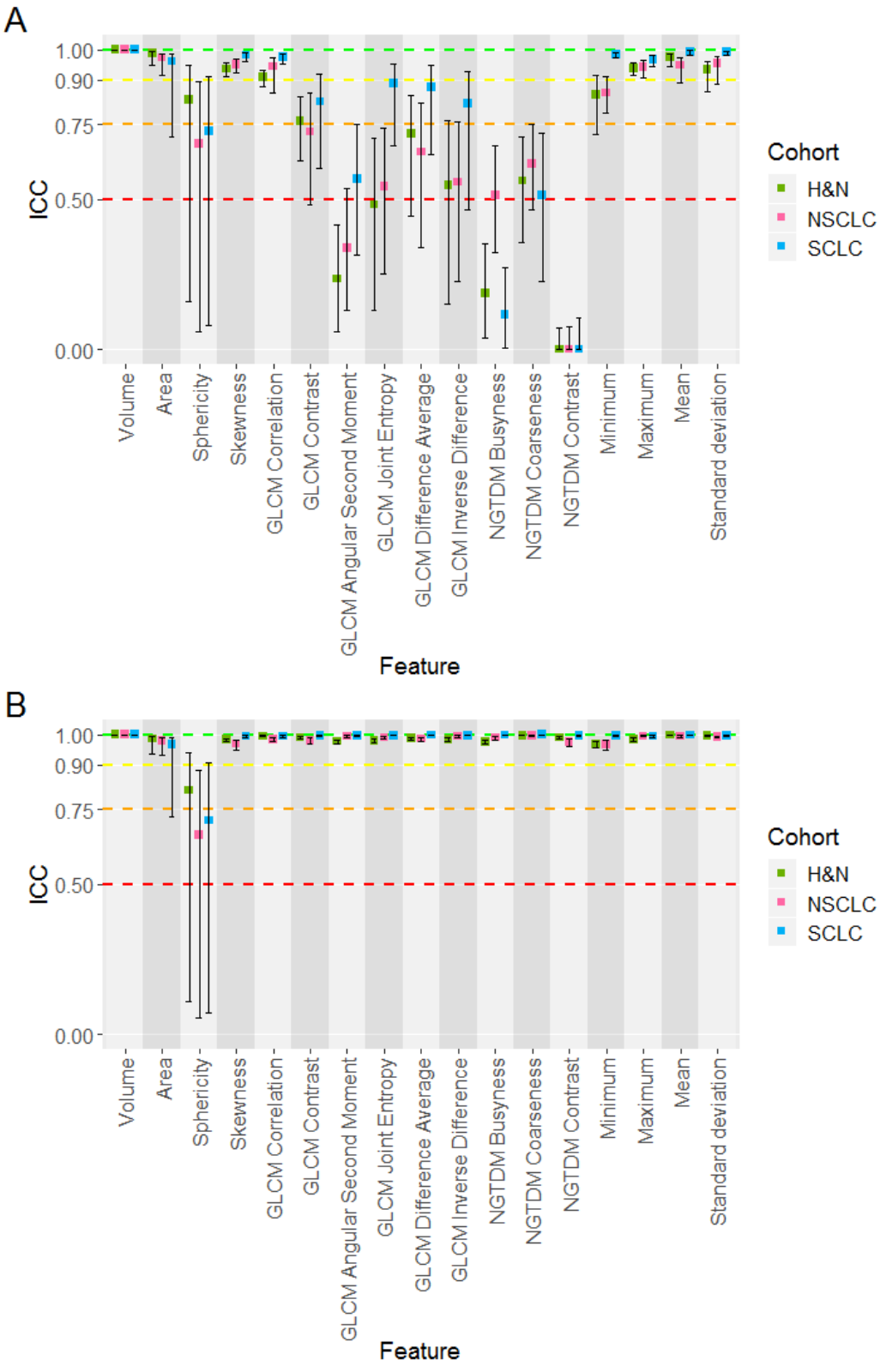

Figure 3.2: Boxplots of ICC estimates and CI for each cohort (H&N in green, NSCLC in pink, SCLC in blue) for all 17 features, showing the statistical reliability between the different software platforms. **A** ICC estimates and CI for all four software with harmonised calculation settings. **B** ICC estimates and CI for the three IBSI-compliant software with harmonised calculation settings (i.e. with IBEX excluded from analysis).

We repeated the analysis for only the IBSI-compliant software platforms, by removing IBEX data (Figure 3.2B). This had a marked effect, with 15/17 features now showing excellent reliability across all datasets. Overall, these data show that the level of reliability across different radiomic biomarkers can vary substantially between different software platforms in the absence of IBSI-compliant standardisation. Once standardisation is adopted, this divergence is reduced substantially for most radiomic biomarkers.

3.3.2 IBSI standardisation is only effective when calculation settings are harmonised

IBSI guidelines provide clear instructions and definitions for the process of image biomarker calculation. However, no recommendations are given for calculation settings. We evaluated the influence of using the default calculation settings versus harmonising them across software platforms using the three IBSI-compliant software platforms (Figure 3.3A). Reliability was excellent for only 6/17 features (volume, skewness, standard deviation and mean, minimum, maximum intensity) when default calculation settings were used, despite all software being IBSI-compliant. In distinction, 10/17 features (sphericity, all six GLCM-based features and all three NGTDM-based features) had poor reliability across all three datasets.

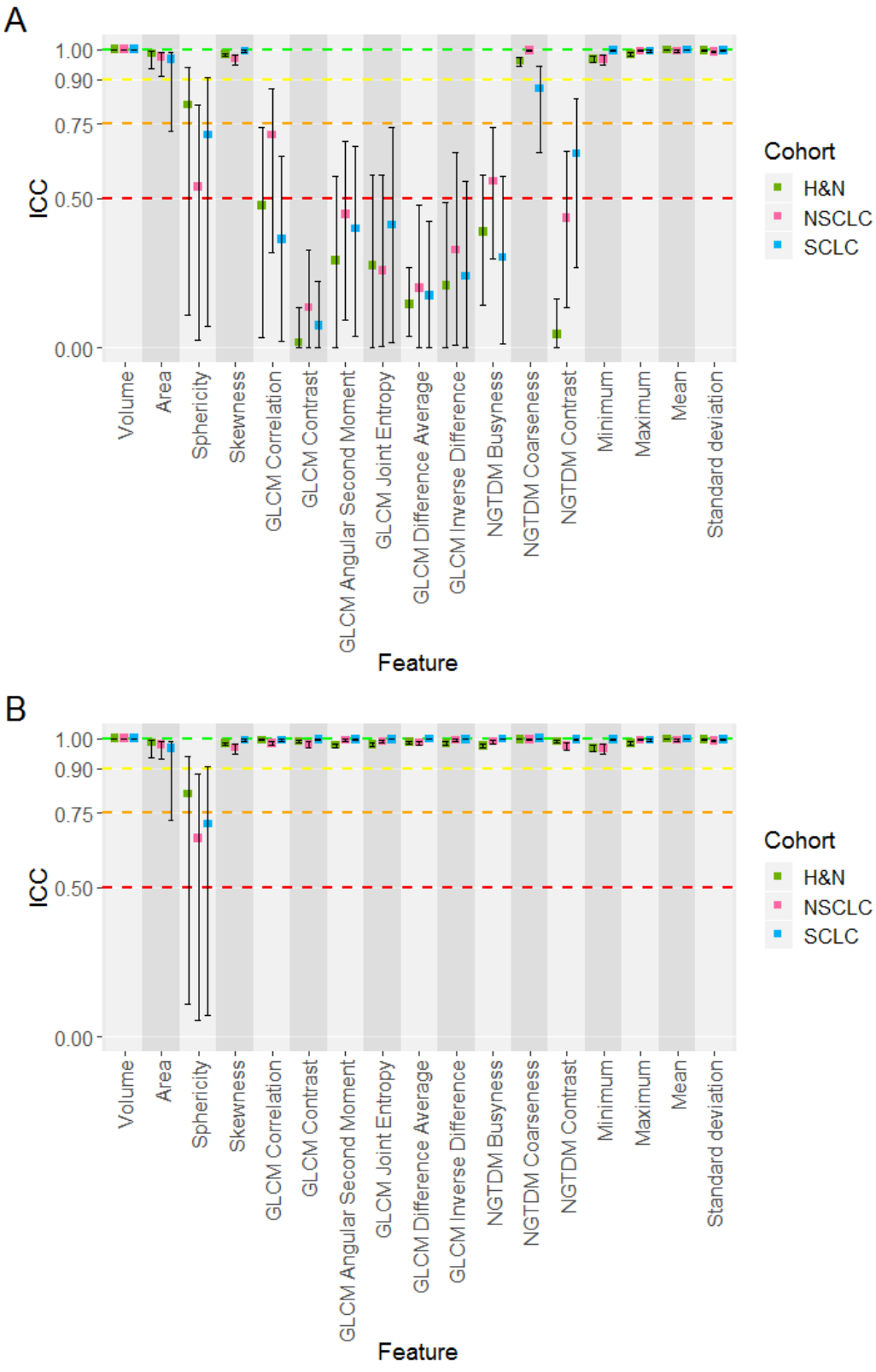

Figure 3.3: Boxplots of ICC estimates and CI for each cohort (H&N in green, NSCLC in pink, SCLC in blue) across all 17 features, showing the statistical reliability between the different software platforms. **A** ICC estimates and CI for the three IBSI-compliant software with default calculation settings (i.e. with IBEX excluded from analysis). **B** ICC estimates and CI for the three IBSI-compliant software with harmonised calculation settings (i.e. with IBEX excluded from analysis).

Once calculation settings were harmonised, the reliability reverted to that seen for IBSI-compliant software (Figure 3.3B). These data reveal the importance of these user-defined free parameters to the calculation of radiomic features. Without harmonisation of calculation settings, even IBSI-compliant platforms generate unreliable features, with the effect remarkably consistent across the three different tumour types and two different types of CT data (diagnostic and radiotherapy planning scans).

3.3.3 Different versions of each software platform influence the statistical reliability of radiomic biomarkers

Software platforms undergo frequent updates. We evaluated the effect of changing between software versions for all three IBSI-compliant platforms by calculating the ICC between the newer and older versions. PyRadiomics had excellent reliability for all features (Figure 3.4A). CERR had a discretisation error in an older version (commit 50530f7 (29/08/2019) available at https://github.com/cerr/CERR) which affected texture features calculation (GLCM and NGTDM) (Figure 3.4B). We identified this difference and, after making the developers aware, the source of error issue was discovered and corrected for the newest version, which is used in our full analysis.

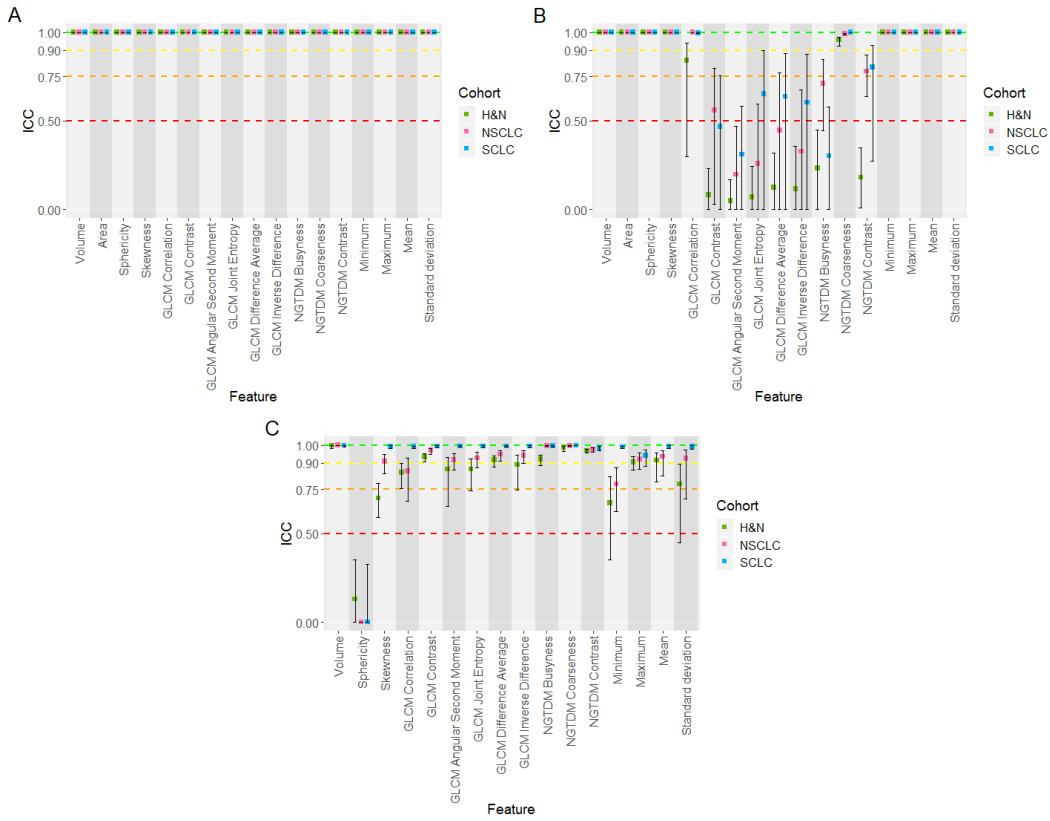

Figure 3.4: Boxplots of ICC estimates and CI for each cohort (H&N in green, NSCLC in pink, SCLC in blue) across all 17 features, showing the reliability between different versions of the same software platform. ICC estimates and CI are presented for **A** PyRadiomics version 2.2.0 versus 2.1.2 with harmonised calculation settings, **B** CERR commit a1c8181 versus 50530f7 with harmonised calculation settings and **C** LIFEx version 5.47 versus 5.1 with harmonised calculation settings (NB: area is not calculated in LIFEx version 5.1 and so does not appear in **C**).

Initial experiments showed that sphericity had poor reliability in all datasets, even when comparison was restricted to IBSI-compliant software platforms (Figure 3.2B). Investigation traced this uncertainty to LIFEx (the sphericity values for CERR and PyRadiomics had ICC estimates with 95% CI of 0.996 to 0.999 (CI 0.992-1) for the three clinical datasets). Comparing the latest LIFEx release (5.1) with the development version used in this study (5.47) shows significant changes in sphericity (Figure 3.4C). The minimum value calculation also changed between these versions with a knock-on effect on dependent features, such as skewness, some GLCM features and standard deviation.

Taken together, these data reveal the importance of study authors reporting which software version was used for data analysis. The data also highlight the difficulty in

comparing studies that initially appear to be similar to one another.

3.3.4 Software platform and calculation settings affect the significance and direction of correlation of radiomic features to overall survival

We assessed how the choice of software platform and calculation settings influences the relationship of radiomic features to patient outcome. These analyses were performed in the largest of our clinical datasets (H&N cancer; N = 108). Overall survival was determined, with 28 patients dying within the follow-up period of 2.2 years. Univariable Cox regression results are presented for all 17 features with harmonised calculation settings and default calculation settings (Figure 3.5).

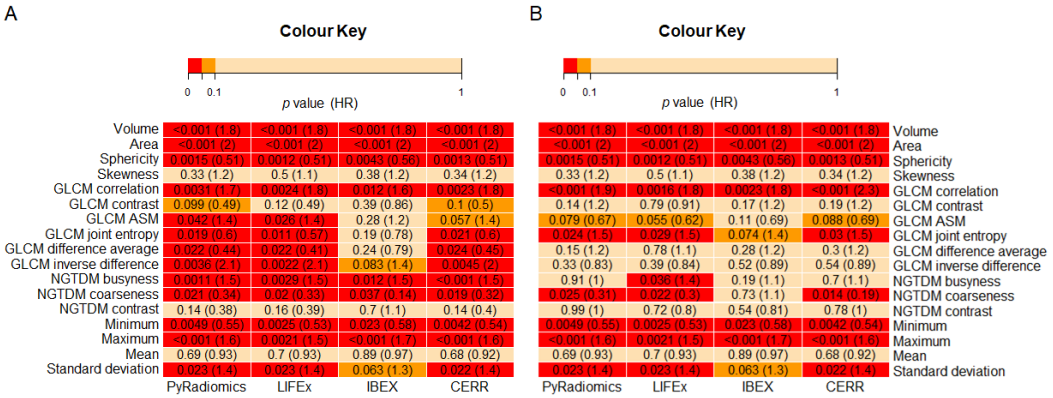

Figure 3.5: Heat-map of the p values (and associated hazard ratios) from univariable Cox regression for each radiomic feature, with harmonised calculation settings on the left **A** and default calculation settings on the right **B**. Cells are colour-coded according to the following p value thresholds: p value < 0.05 (red), $0.05 < p$ value < 0.1 (orange) and p value > 0.1 light orange. *ASM*, angular second moment; *HR*, hazard ratio.

The p values and associated hazard ratios for each feature when using harmonised calculation settings are presented in Figure 3.5A. Eight features (volume, area, sphericity, GLCM correlation, NGTDM busyness, NGTDM coarseness, minimum and maximum) were significant at $p < 0.05$ in all four platforms. A further five features (GLCM angular second moment, GLCM joint entropy, GLCM difference average, GLCM inverse difference and standard deviation) were significant at $p < 0.05$ for the three IBSI-compliant software platforms but not in IBEX. When a given radiomic feature was

deemed significant at the $p < 0.05$ threshold for multiple software platforms, the hazard ratios were generally in close agreement across the software platforms.

The p values and associated hazard ratios for each feature when using default calculation settings are presented in Figure 3.5B. Since shape and most first-order features are not dependent on these parameters, they were unaffected by the changed calculation settings. Texture features, however, are dependent on the user-defined calculation settings and all became no longer significant at the $p < 0.05$ threshold, with the exception of GLCM correlation. Notably, IBEX diverged further from agreement with the three IBSI-compliant software platforms.

Of particular note, the hazard ratio for GLCM joint entropy changed from 0.56–0.59 (i.e. less than 1.0 and significant p value) when harmonised calculation settings were used to 1.5 (i.e. more than 1.0 and significant p value) when default calculation settings were used. Thus, significant correlations were detected that had opposing hazard ratio directions depending on choice of parameter input. This effect is shown clearly in Figure 3.6, where the direction of the hazard ratio changed from protective to harmful. These data reveal that both IBSI compliance and calculation settings can affect the significance and direction of relationships between radiomic features and clinical outcome.

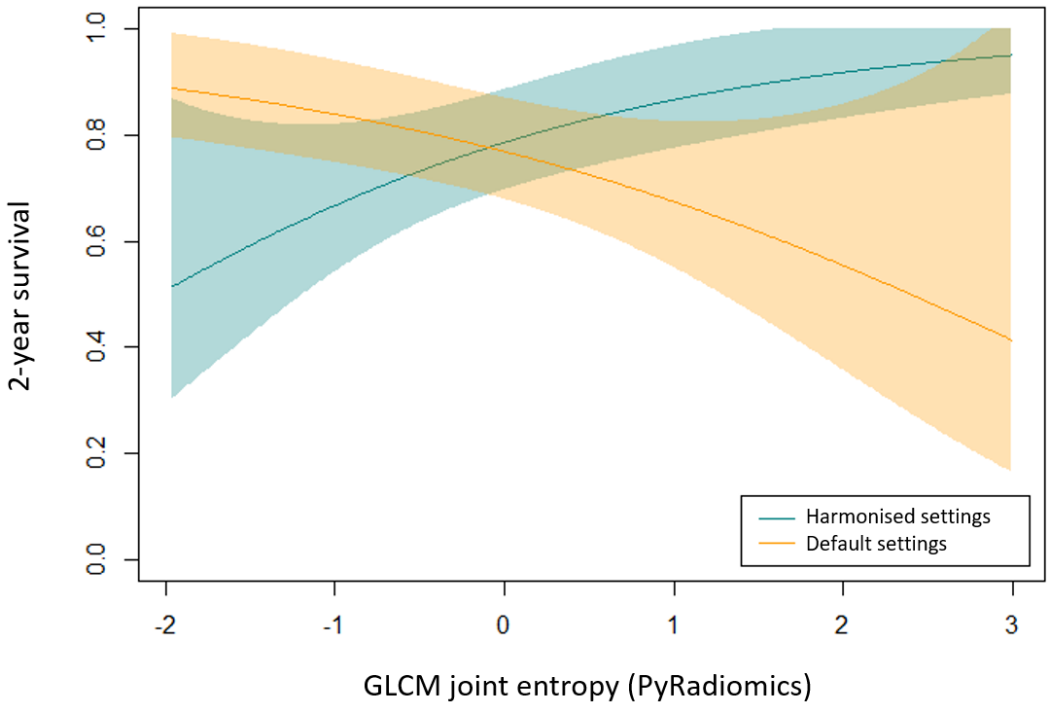

Figure 3.6: GLCM joint entropy (here calculated in PyRadiomics) against 2-year survival for patients with H&N cancer when calculated with harmonised settings (blue) and default settings (orange).

3.4 Discussion

Radiomics has great potential to produce independent predictive biomarkers for personalised healthcare, particularly in the management of patients with cancer [201]. Many studies have been published describing prognostic and predictive radiomic signatures, but significant methodological limitations have hindered clinical translation of these techniques [225].

In this study, we investigated the importance of IBSI compliance, harmonising calculation settings and choice of platform version when using different radiomics calculation platforms. We tested how these factors affect the statistical reliability of features and showed how these factors also influence the relationship between radiomic biomarkers and clinical outcome (in this case, the overall survival).

Radiomic feature calculation is an important part of the radiomics workflow. Studies can use a variety of commercial or freely available software platforms to achieve this [209] or use in-house developed software. A study by Foy et al compared two in-house

developed software to IBEX and found that for head and neck CT scans, histogram features had excellent reliability but GLCM features varied between poor and excellent reliability [211]. The software packages in that study were not IBSI-compliant.

Our study demonstrates the benefits of standardising feature calculation platforms according to the IBSI. Features calculated in IBSI-compliant software had greater statistical reliability than features calculated in non-compliant platforms, but only when calculation settings were also harmonised. The method of grey level discretisation has been shown to affect feature reproducibility within the same software platform [170, 174]. Our results both confirm these findings and extend the principle to all those user-defined parameters listed in Table 3.3, emphasising the need to harmonise calculation settings even when an IBSI-compliant platform is used. Results were highly consistent across three clinical datasets.

Our data has also highlighted the importance of inter-software comparison. By doing so, we identified potential errors in both the CERR and LIFEx code bases, leading to subsequent corrections and improved reliability. It is vital that investigators document the version and date of the software platform used in their study to ensure results are reproducible between institutions. Our data also highlight the benefits of open-source tools and the importance of the relevant scientific communities actively working with their developers to improve them.

Univariable survival analysis revealed substantial differences in prognostic power between supposedly similar features derived from different software platforms. We make three observations. Firstly, some features had significant association with H&N cancer overall survival in the IBSI-compliant software but not in IBEX. These findings concur with Liang et al who investigated two platforms and found differences in downstream clustering of known prognostic factors in patients with nasopharyngeal carcinoma [212]. Similar conclusions were drawn by Bogowicz et al who investigated this in PET scans of patients with H&N cancer [197]. Secondly, when only evaluating IBSI-compliant software, there was a divergence of feature to survival correlation between software platforms when calculation settings varied.

Thirdly, our study demonstrates that when different calculation settings are used, the

relationship of significant features to survival can remain significant but the direction of that relationship (hazard ratio) can invert from protective to harmful. This effect may reflect that for some features, altering calculation settings radically alters the biophysical property being measured. In this study, there is no ground truth against which the 'true' direction of a feature can be established, but the data demonstrates the important role calculation settings play in selecting features for radiomic signatures.

There are several limitations to this study. Our inclusion criteria for feature calculation platforms that they are freely available, widely cited, and sufficiently well documented for analysis limited the number of assessed platforms to four, only one of which was not IBSI-compliant. There are also more features available in each of the software platforms that were not included in this study, as only features that were available across all four software platforms were analysed. The clinical datasets used were sufficiently large to evaluate ICC with CIs but the number of events only permitted univariable survival analysis of outcome. Lastly, LIFEx is a closed-source project, which precluded thorough investigation of the observed difference in sphericity calculation compared to other IBSI-compliant software.

In conclusion, this study has shown that use of IBSI-compliant radiomic feature calculation platforms appears to increase the statistical reliability of features. However, even IBSI-compliant platforms are affected strongly by user-defined calculation settings and changes between software versions. Future radiomics studies should be aware of potential differences between software platforms and ensure platforms used for radiomics studies are IBSI-compliant. Studies should ensure software version and user-defined parameters are clearly reported. Furthermore, the radiomics community should consider working towards a recommended set of harmonised calculation settings. Locking imaging biomarkers down in this way will improve the technical quality of data from subsequent studies, a vital step towards their translation into clinical decision-making tools [45].

3.5 Acknowledgements

This work was supported by CRUK via the funding to Cancer Research UK Manchester Centre: (C147/A18083) and (C147/A25254) and to Professor James P B O'Connor (C19221/A22746). Professor Fiona Blackhall, Professor Corinne Faivre-Finn and Professor James P B O'Connor are supported by the NIHR Manchester Biomedical Research Centre.

3.6 Supplementary materials

Table 3.4: Patient characteristics for the H&N, NSCLC and SCLC cohorts.

	H&N (n=108)	**NSCLC (n=47)**	**SCLC (n=37)**
Sex (n (%))			
Female	26 (24.1)	20 (42.6)	15 (40.5)
Male	82 (75.9)	27 (57.4)	22 (59.5)
Age at start of treatment (median [IQR])	61.50 [56.00, 67.00]	68.50 [63.00, 73.00]	60.00 [54.00, 66.00]
T stage (n (%))			
X	0 (0)	3 (6.4)	0 (0)
1	18 (16.7)	5 (10.6)	0 (0)
2	41 (38.0)	12 (25.5)	9 (24.3)
3	20 (18.5)	7 (14.9)	14 (37.8)
4	26 (24.1)	20 (42.6)	12 (32.4)
Unknown	3 (2.8)	0 (0)	2 (5.4)
N stage (n (%))			
X	0 (0)	2 (4.2)	0 (0)
0	40 (37.0)	6 (12.8)	9 (24.3)
1	14 (13.0)	2 (4.3)	4 (10.8)
2	46 (42.6)	22 (46.8)	18 (48.6)
3	6 (5.6)	15 (31.9)	2 (5.4)
Unknown	2 (1.9)	0 (0)	4 (10.8)
Performance status (n (%))			
0	56 (51.9)	5 (10.6)	13 (35.1)
1	40 (37.0)	38 (80.9)	24 (64.9)
2	9 (8.3)	4 (8.5)	0 (0)
3	3 (2.8)	0 (0)	0 (0)
HPV status (n (%))			
Negative	14 (13.0)	NA	NA
Positive	51 (47.2)	NA	NA
Unknown	43 (39.8)	NA	NA
Chemotherapy (n (%))			
No	50 (46.3)	47 (100)	0 (0)
Yes	58 (53.7)	0 (0)	37 (100)
Radiotherapy prescribed dose (median [IQR])	65.40 [59.03, 66.00]	NA	55.00 [45.00, 66.00]

Table 3.5: Image acquisition and reconstruction parameters for the SCLC, NSCLC and H&N CT datasets.

	H&N (n=108)	NSCLC (n=47)	SCLC (n=37)
Manufacturer, model	Philips, Brilliance Big Bore (n=94, 87.0%) SIEMENS, SOMATOM Definition AS (n=14, 13.0%)	GE MEDICAL SYSTEMS, LightSpeed VCT (n=4, 8.5%) GE MEDICAL SYSTEMS, Optima CT660 (n=3, 6.4%) Philips, Ingenuity CT (n=1, 2.1%) SIEMENS, SOMATOM Definition AS (n=38, 80.9%) SIEMENS, SOMATOM Definition AS+ (n=1, 2.1%)	GE MEDICAL SYSTEMS, HiSpeed CT/I (n=9, 24.3%) GE MEDICAL SYSTEMS, LightSpeed RT16 (n=4, 10.8%) Philips, Brilliance Big Bore (n=3, 8.1%) Philips, Gemini (n=1, 2.7%) SIEMENS, Definition AS (n=16, 43.2%) SIEMENS, Sensation Open (n=2, 5.4%) SIEMENS, Spirit (n=2, 5.4%)
Slice thickness (mm) (median [IQR])	3 [3.00, 3.00]	3 [3.00, 3.00]	3.00 [3.00, 5.00]
Pixel spacing (mm) (median [IQR])	1.17 x 1.17 [1.17, 1.18]	0.68 x 0.68 [0.63, 0.79]	0.98 x 0.98 [0.94, 0.98]
Tube voltage (kVp)	120	120	120
Tube current (mAs) (median [IQR])	177.00 [99.00, 296.75]	313.00 [267.00, 413.00]	167.00 [140.00, 285.00]

Table 3.5 continued from previous page

	H&N (n=108)	NSCLC (n=47)	SCLC (n=37)
Convolution kernel	B (n=94, 87.0%) B31f (n=14, 13.0%)	['I70f', '2'] (n=20, 42.6%) ['I70f', '3'] (n=2, 4.3%) B70f (n=1, 2.1%) B80f (n=16, 34.0%) C (n=1, 2.1%) LUNG (n=7, 14.9%)	B (n=4, 10.8%) B31f (n=16, 43.2%) B31s (n=2, 5.4%) B41s (n=2, 5.4%) SOFT (n=9, 24.3%) STANDARD (n=4, 10.8%)

Supplementary material A

The 37 radiotherapy planning contrast-enhanced CT scans from a cohort of patients with small cell lung cancer (SCLC) were acquired in nine different institutions, namely the Christie Hospital (Manchester, UK, N=28), Beatson Cancer Centre (Glasgow, UK, N=1), Bristol Haematology & Oncology Centre (Bristol, UK, N=1), Freeman Hospital (Newcastle-upon-Tyne, UK, N=1), Royal Marsden Hospital (London, UK, N=1), Institut Ste Catherine, Avignon, France (N=1), Centre Hospitalier Universitaire de Clermont-Ferrand (Clermont-Ferrand, France, N=2), Universiteit Gent (Gent, Belgium, N=1), Medical University of Gdansk (Gdansk, Poland, N=1).

Supplementary material B

In LIFEx, features are calculated on the largest cluster of continuous voxels within the ROI only. To be able to compare results from LIFEx to results from PyRadiomics, IBEX and CERR, which use the whole ROI regardless of whether the voxels are continuous or not, only ROI's with one cluster of voxels according to LIFEx were analyzed. This left 37 ROIs for comparison in the SCLC dataset, 108 ROIs in the H&N dataset and 47 ROIs in the NSCLC dataset.

In IBEX, the Hounsfield Units (HU) of the CT scan have 1000 added to them to ensure non-negative values, despite the fact that the lowest HU for a CT scan is -1014. Negative HU after this transformation are truncated at 0. To adjust for this in the minimum, maximum and mean comparison in Table 3.3, 1000 HU were taken from the IBEX values.

The IBSI define two methods for calculating the volume of a region of interest (ROI). The first is a mesh-based approach, where the surface of the ROI is represented as a mesh of triangles. The second method simply multiplies the volume of one voxel by the total number of voxels in the ROI. The voxel counting method does not handle partial volume effects at the ROI edge, which is particularly important for smaller volumes, and therefore the mesh-based approach is preferred [51]. PyRadiomics provide both options for volume calculation. In LIFEx, IBEX and CERR, volume is calculated using a voxel-counting approach.

The neighborhood grey tone difference matrix (NGTDM) as defined by the IBSI, PyRadiomics and CERR varies in its nomenclature. In IBEX it is known as the neighbor intensity difference matrix and in LIFEx as the neighborhood grey-level different matrix (NGLDM). The original definition of the NGTDM was developed by Amadasun and King [226]. The IBSI define the NGLDM as a different matrix entirely, originally developed by Sun and Wee [227], however the NGLDM in LIFEx is the same as the NGTDM as defined by the IBSI. Other than this, LIFEx correct their feature names to comply with the IBSI, for example "GLCM Homogeneity=Inverse Difference" since inverse difference is the IBSI-compliant feature definition.

Chapter 4

Impact of Introducing Intensity Modulated Radiotherapy on Curative Intent Radiotherapy and Survival for Lung Cancer

This chapter has been published in Frontiers in Oncology 2022 Volume 12 p835844. The section numbering, references and formatting have been modified to be consistent with the rest of the thesis.

Authors

Isabella Fornacon-Wood[1†], Clara Chan[2†], Neil Bayman[2], Kathryn Banfill[1,2], Joanna Coote[2], Alex Garbett[2], Margaret Harris[2], Andrew Hudson[2], Jason Kennedy[3], Laura Pemberton[2], Ahmed Salem[1,2], Hamid Sheikh[2], Philip Whitehurst[4], David Woolf[2], Gareth Price[1,4‡] and Corinne Faivre-Finn[1,2‡]

[†] These authors share first authorship. [‡] These authors share last authorship.

Affiliations

[1] Division of Cancer Sciences, University of Manchester, Manchester, UK.

[2] Department of Clinical Oncology, The Christie NHS Foundation Trust, Manchester, UK.

[3] Radiotherapy Related Research, The Christie NHS Foundation Trust, Manchester, UK.

[4] Department of Medical Physics, The Christie NHS Foundation Trust, Manchester, UK.

Author contributions

I wrote the PostgreSQL scripts to collate the GTV and PTV data, joining them to the database provided by J.K. and cleaned by C.C.. I wrote the R script to perform the statistical analysis. I co-wrote the manuscript with C.C., which was reviewed by all co-authors.

Abstract

Background

Lung cancer survival remains poor. The introduction of Intensity-Modulated Radiotherapy (IMRT) allows treatment of more complex tumours as it improves conformity around the tumour and greater normal tissue sparing. However, there is limited evidence assessing the clinical impact of IMRT. In this study, we evaluated whether the introduction of IMRT had an influence on the proportion of patients treated with curative-intent radiotherapy over time, and whether this had an effect on patient survival.

Materials and Methods

Patients treated with thoracic radiotherapy at our institute between 2005 and 2020 were retrospectively identified and grouped into three time periods: A) 2005-2008 (pre-IMRT), B) 2009-2012 (selective use of IMRT), and C) 2013-2020 (full access to IMRT). Data on performance status (PS), stage, age, gross tumour volume (GTV), planning target volume (PTV) and survival were collected. The proportion of patients treated with a curative dose between these periods was compared. Multivariable survival models were fitted to evaluate the hazard for patients treated in each time period, adjusting for PS, stage, age and tumour volume.

Results

12,499 patients were included in the analysis (n=2675 (A), n=3127 (B), and n=6697 (C)). The proportion of patients treated with curative-intent radiotherapy increased between the 3 time periods, from 38.1% to 50.2% to 65.6% ($p<0.001$). When stage IV patients were excluded, this increased to 40.1% to 58.1% to 82.9% ($p<0.001$). This trend was seen across all PS and stages. The GTV size increased across the time periods and PTV size decreased. Patients treated with curative-intent during period C had a survival improvement compared to time period A when adjusting for clinical variables (HR=0.725 (0.632-0.831), $p<0.001$).

Conclusion

IMRT was associated with more patients receiving curative-intent radiotherapy. In addition, it facilitated the treatment of larger tumours that historically would have been

treated palliatively. Despite treating larger, more complex tumours with curative-intent, a survival benefit was seen for patients treated when full access to IMRT was available (2013-2020). This study highlights the impact of IMRT on thoracic oncology practice, accepting that improved survival may also be attributed to a number of other contributing factors, including improvements in staging, other technological radiotherapy advances and changes to systemic treatment.

4.1 Introduction

Lung cancer is the third most common cancer and the leading cause of cancer death in the UK [81]. For some time it has been recognised that better treatments are urgently required to improve lung cancer survival. Over the last two decades, increasing knowledge regarding the biology of lung cancer has led to the development of new systemic agents such as tyrosine kinase inhibitors and immunotherapy, leading to improvements in survival in locally advanced and metastatic non-small cell lung cancer. However outcome of lung cancer patients remains poor compared to the majority of other cancer types [22, 83].

Radiotherapy (RT) plays an important role in the management of lung cancer with over 50% patients receiving this modality at some point during their cancer journey [6]. Radiotherapy can either be given with palliative intent to control symptoms, or radically with curative intent – in patients with early and locally advanced disease.

Radiotherapy treatment planning is a careful balancing act between optimal tumour control and limitation of damage to normal tissue. In order to avoid undue toxicity, dose constraints are placed on the normal tissues such as the lungs, heart, oesophagus and spinal cord to minimise functional damage. The radiotherapy dose delivered to the tumour is therefore often limited by the dose that can be safely delivered to the normal tissues. This is particularly challenging in patients who have large volume disease and/or disease close to critical normal structures, such as the spinal cord. In some situations this can lead to patients being treated with a safer, lower, but ultimately palliative dose. As local control correlates with improved survival [32, 228], these patients naturally have a poorer outcome.

Over the last two decades, great advancements have been made in radiotherapy technology [1, 229]. Prior to the 1980's radical lung patients were planned with fluoroscopy, however the introduction of computed tomography (CT) allowed improved tumour localisation and conformal planning. In addition, the advent of the multi-leaf collimator (MLC) enabled fields to be shaped around a target volume. This three-dimensional conformal radiotherapy (3DCRT) has been the gold standard for radical RT to the lung since the 1980's. Subsequently, 4D planning was introduced which incorporates tumour motion into the radiotherapy planning process, allowing more bespoke plans based on tumour motion and a reduction in margins. In addition there have been improved methods of image guidance, allowing the verification of the tumour position during the treatment course with increasing accuracy. This again has allowed a reduction in tumour margins and therefore dose delivered to normal tissue [230]. Despite these improvements in technology, there are still a significant proportion of lung cancer patients, in particular those with locally advanced disease, who are treated with a palliative approach either due to the treatment volume or its proximity to a critical structure [231].

Intensity-modulated radiotherapy (IMRT) is an advanced form of 3DCRT that modifies the intensity of the radiation across each beam, sculpting the high-dose volume around the site of disease and thereby sparing adjacent organs at risk. This technology has been available since the early 2000s, however the routine implementation of IMRT in the setting of lung cancer treatment has been slow, due partly to the increased planning and quality assurance time required by this techniques, and a perceived lack of evidence for using it [232]. To date there are a handful of large retrospective studies evaluating 3DCRT against IMRT in lung cancer, and only one publication in a randomised, prospective setting which addresses this issue [233]. There is a lack of data on the impact of modern RT technology on patient management and outcome, particularly for patients that are typically excluded from clinical trials [20].

We have been treating lung cancer patients in our institution routinely with IMRT for over a decade. This study aims to evaluate whether the introduction of IMRT has had an influence on the proportion of patients we are able to treat with curative intent over time, and whether this has had any impact on patient survival.

4.2 Methods and Materials

A retrospective review of patients in our institution treated with thoracic RT for lung cancer between 2005-2020 was carried out. Approval was granted to collect and analyse this patient data by the UK Computer Aided Theragnostics (ukCAT) Research Database Management Committee (REC reference: 17/NW/0060).

Patients between 2005-2012 were identified by ICD-10 codes on MOSAIQ and patients between 2013-2020 were identified via the Christie web portal (CWP – an in house e-record system designed to collect structured data on patients, tumour characteristics and outcome data). For all patients, data on age, sex, ECOG performance status (PS), stage, gross tumour volume (GTV), planning target volume (PTV) and survival were collected. For patients planned using 4D-CT imaging, GTV data was synthesized from the internal gross tumour volume (iGTV) using a previously published method [234].

Patients were grouped into 3 time periods, determined by the year the first radiotherapy fraction was delivered: A (2005-2008, pre IMRT), B (2009-2012, some availability IMRT) and C (2013-2020, full access IMRT). SABR was introduced in 2011 in our institution. Any patient who received an absolute physical dose of greater than 40 Gy was classed as having 'curative-intent' thoracic RT. This dose was chosen to cover patients receiving radical doses such as 45 Gy/30 fractions twice-daily (EQD2 43.1 Gy) or 40 Gy/15 fractions daily (EQD2 42.2 Gy) for limited stage small cell lung cancer (SCLC). For patients receiving palliative radiotherapy, records were manually checked to ensure these patients received palliative radiotherapy to the lung (and not a site of metastatic disease). Those that had not were excluded from this study.

The proportion of patients treated with curative-intent RT was compared between the 3 time periods and the Chi-squared test was used to compare differences between the groups. We performed 2 analyses, one including all stages and the other including only patients with stage I-III. The proportion of patients treated with curative-intent RT was also compared across all PS groupings and stages of disease. For curative-intent patients, the trend of tumour volume treated over time was reviewed and the Mann–Whitney U test used to compare GTV and PTV across time periods. Survival

curves were generated using the Kaplan-Meier method and compared using the log-rank test. Univariable and multivariable cox survival models were fitted to evaluate the hazard of being treated in one of the 3 time periods, adjusting for baseline PS, stage at diagnosis, age at the start of treatment and GTV. These analyses were then repeated excluding patients who had received stereotactic radiotherapy (SABR). All statistical analyses were performed in R 4.0.0 [222] with package survival v3.1-12 [224].

4.3 Results

In total, 12499 patients were identified as having received radiotherapy to the lung between 2005 and 2020; 2675 in group A (2005-2008, pre IMRT), 3127 in group B (2009-2012, some availability IMRT) and 6697 in group C (2013-2020, full access IMRT). Patients in time period B receiving IMRT were planned with this technique only if 3D conformal radiotherapy was unable to achieve a dosimetrically acceptable radical plan.

Baseline characteristics are presented in Table 4.1. Median age was 70 (63-77), 71 (64-78) and 72 (65-78) in each group respectively. 985 patients received SABR, 0 in group A, 33 in group B and 952 in group C.

Table 4.1: Baseline characteristics.

	A: 2005-2008 n=2675	**B: 2009-2012 n=3127**	**C: 2013-2020 n=6697**
Age at start of treatment (median [IQR])	70.00 [63.00, 77.00]	71.00 [64.00, 78.00]	72.00 [65.00, 78.00]
Sex (n (%))			
Male	1527 (57.8)	1729 (56.0)	3435 (52.3)
Female	1117 (42.2)	1358 (44.0)	3139 (47.7)
Treatment intent (n (%))			
Curative	1018 (38.1)	1570 (50.2)	4391 (65.6)
Palliative	1657 (61.9)	1557 (49.8)	2306 (34.4)
SABR (n (%))	0 (0.0)	33 (1.1)	952 (14.2)
ECOG performance status (n (%))			
0	284 (10.6)	281 (9.0)	588 (8.8)
1	852 (31.9)	1071 (34.3)	2301 (34.4)
2	474 (17.7)	762 (24.4)	2012 (30.0)
3	167 (6.2)	348 (11.1)	813 (12.1)
4	3 (0.1)	5 (0.2)	17 (0.3)
Missing	895 (33.5)	660 (21.1)	966 (14.4)
Stage (n (%))			
I	321 (12.0)	443 (14.2)	1490 (22.2)
II	158 (5.9)	243 (7.8)	628 (9.4)
III	552 (20.6)	810 (25.9)	1875 (28.0)
IV	142 (5.3)	512 (16.4)	1706 (25.5)
Missing	1502 (56.1)	1119 (35.8)	998 (14.9)

Abbreviations: *SABR*, stereotactic ablative radiotherapy; *ECOG*, Eastern Cooperative Oncology Group.

There was a progressive increase in the proportion of patients receiving curative-intent radiotherapy year on year since 2005, with a step wise change occurring from 2011 as shown in Figure 4.1. This increase in the proportion of patients receiving a curative dose was highlighted further when patients were grouped into the 3 previously specified time periods (Figure 4.2). Patients receiving curative-intent RT increased between groups A (2005-2008) and B (2009-2013) (38.1% to 50.2%, $p<0.0001$), and B and C (2014-2020) (50.2% to 65.6%, $p<0.0001$). Results were similar when the patients treated with SABR were removed from the analysis (Supplementary Figures 4.8

and 4.9). These percentages increased when only stage I-III patients were examined (Figure 4.3) with patients receiving curative-intent RT increasing from 40.1% to 58% to 82.9% in A, B and C respectively.

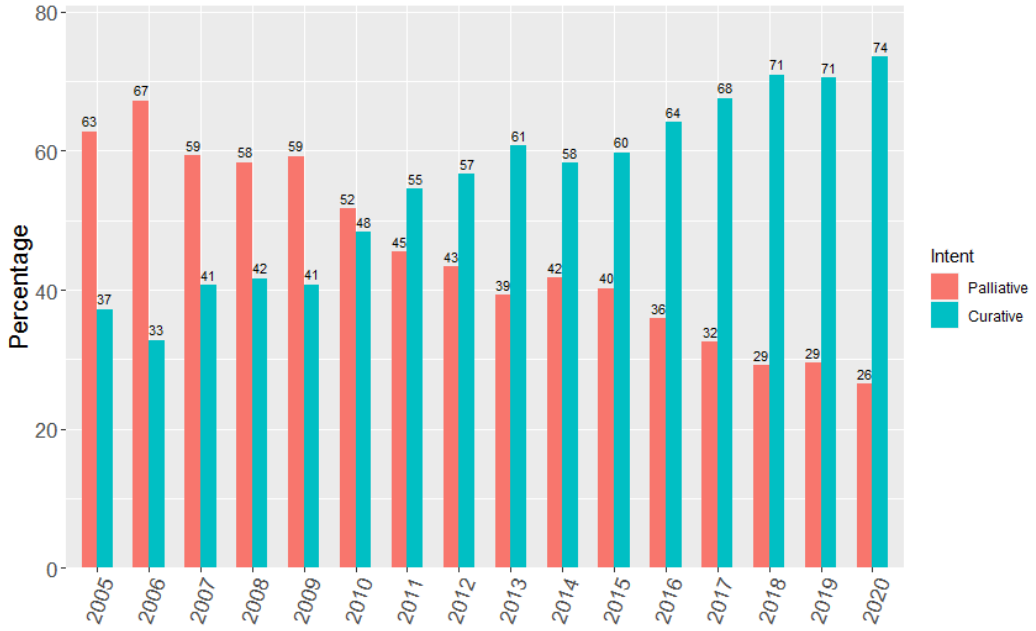

Figure 4.1: Yearly percentage of patients treated with curative versus palliative intent radiotherapy from 2005 to 2020.

Figure 4.2: Percentage of patients treated with curative versus palliative intent radiotherapy (whole population) in each of the pre-specified time periods.

Figure 4.3: Percentage of patients treated with curative versus palliative intent radiotherapy (stages I-III) in each of the pre-specified time periods.

Further sub-classification according to PS and stage are presented in Tables 2, 3 respectively. The proportion of patients treated with curative-intent radiotherapy increased between the three time periods, regardless of PS and stage of disease. Stage IV patients have been included to reflect the increasing use of 'radical' radiotherapy to achieve optimal local disease control, typically in the setting of oligometastatic disease. Results were similar when patients treated with SABR were removed from the analysis (Supplementary Tables 4.5 and 4.6). Table 4 presents sub-classification according to PS for stage III patients only, showing that the proportion of curative-intent patients has increased across all PS for these patients.

Table 4.2: Proportion of patients treated with curative-intent radiotherapy across each PS and time period.

PS	A: 2005-2008 % curative-intent (n curative-intent/n total)	B: 2009-2012 % curative-intent (n curative-intent/n total)	C: 2013-2020 % curative-intent (n curative-intent/n total)
0 (n=1153)	52.1 (148/284)	65.5 (184/281)	71.3 (419/588)
1 (n=4224)	43.9 (374/852)	60.5 (648/1071)	70.7 (1627/2301)
2 (n=3248)	34.8 (165/474)	51.3 (391/762)	67.8 (1365/2012)
3 (n=1328)	15.6 (26/167)	21.8 (76/348)	47.8 (389/813)

Table 4.3: Proportion of patients treated with curative-intent radiotherapy across each stage and time period.

Stage	A: 2005-2008 % curative-intent (n curative-intent/n total)	B: 2009-2012 % curative-intent (n curative-intent/n total)	C: 2013-2020 % curative-intent (n curative-intent/n total)
I (n=2254)	76.9 (247/321)	91.4 (405/443)	97.5 (1453/1490)
II (n=1029)	70.3 (111/158)	84.8 (206/243)	91.6 (575/628)
III (n=3237)	40.4 (223/552)	66.4 (538/810)	75.9 (1424/1875)
IV* (n=2360)	2.11 (3/142)	9.96 (51/512)	14.9 (255/1706)

* Patients with oligometastatic disease treated with curative intent.

Table 4.4: Proportion of patients treated with curative-intent radiotherapy across each PS and time period for stage III patients only.

PS	A: 2005-2008 % curative-intent (n curative-intent/n total)	B: 2009-2012 % curative-intent (n curative-intent/n total)	C: 2013-2020 % curative-intent (n curative-intent/n total)
0 (n=451)	66.7 (48/72)	79.4 (77/97)	87.2 (246/282)
1 (n=1430)	46.0 (116/252)	77.9 (306/393)	85.2 (669/785)
2 (n=819)	28.7 (31/108)	57.6 (110/191)	72.1 (375/520)
3 (n=296)	10.5 (4/38)	32.9 (27/82)	34.1 (60/176)

GTV data was available for 4306 patients treated with curative-intent. The distribution of GTVs in each time period is presented in Figure 4.4A, showing larger GTVs have been treated in group C compared to A and B. Median GTV was 35.5 cm^3 [16.8, 60.1], 39.2 cm^3 [15.1, 82.9] and 32.5 cm^3 [9.9, 91.8] for groups A, B and C respectively. There was a significant decrease in median GTV between time periods B and C (p=0.00597). However, when patients treated with SABR (n=546) were removed from the analysis (violin plot in Figure Figure 4.4B), median GTV was 35.5 cm^3 [16.8, 60.1], 41.7 cm^3 [16.3, 85.8] and 47.6 cm^3 [17.6, 112.1] for groups A, B and C respectively, showing a significant increase in GTV size in each time period in non-SABR patients (A to B, p=0.00383; B to C, p=0.00136). The maximum treated GTV also increased across each time period, from 254.0 cm^3 to 534.4 cm^3 to 916.3 cm^3.

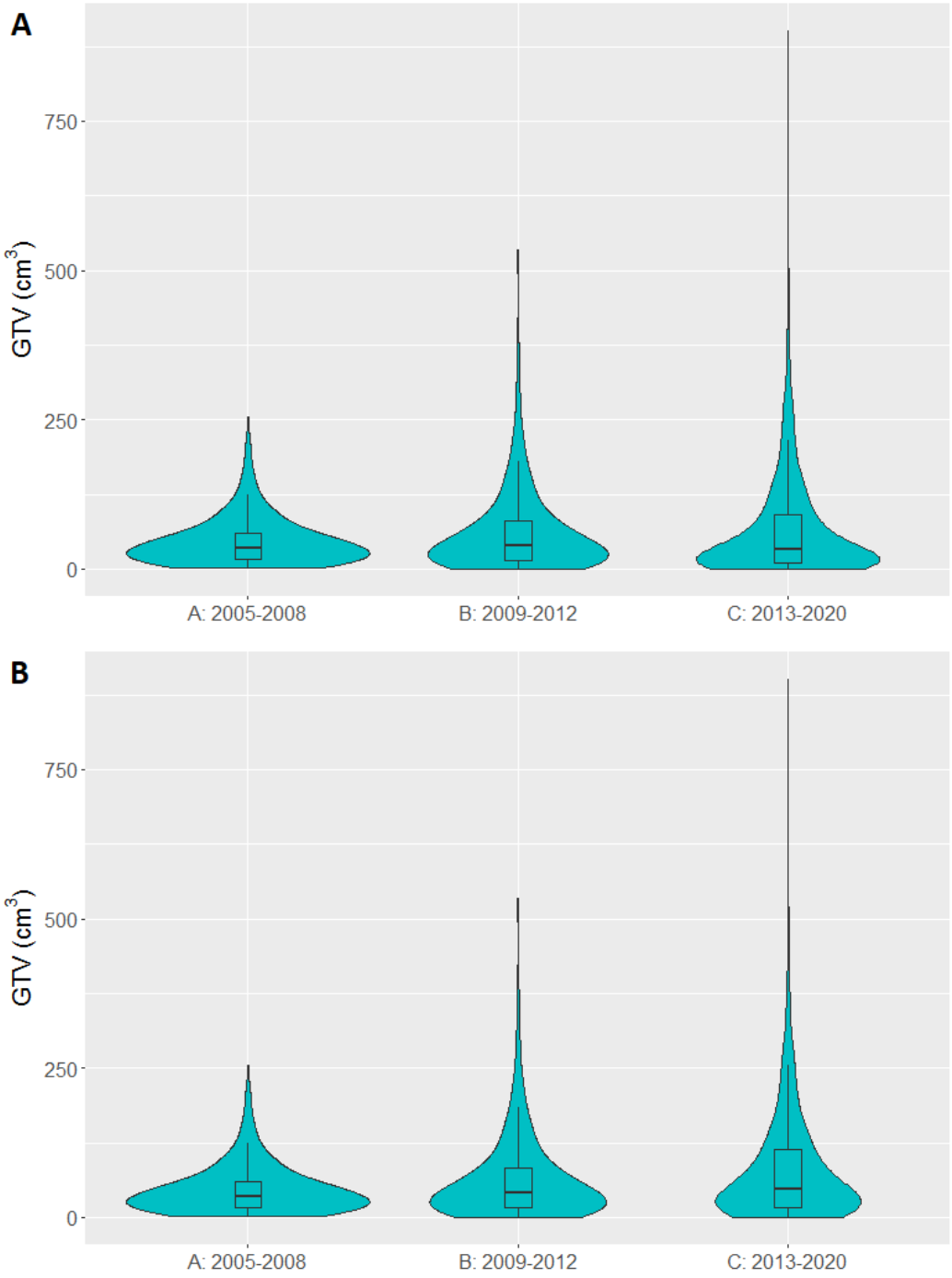

Figure 4.4: Violin plot presenting the distribution of GTVs in patients treated with curative-intent radiotherapy in each time period. (A) SABR patients included (B) SABR patients excluded.

PTV data was available for 4915 curative-intent patients. The distribution of PTVs in each time period is presented in Figure 4.5. Median PTV was 319.2 cm^3 [225.8, 433.2], 326.3 cm^3 [202.3, 502.2] and 235.9 cm^3 [97.8, 401.7] for groups A, B and C respectively. There was a significant decrease in PTV between time periods B and

C ($p<0.0001$). When patients treated with SABR were removed from the analysis, median PTV was 319.2 cm^3 [225.8, 433.2], 334.1 cm^3 [211.8, 506.7] and 282.2 cm^3 [169.7, 438.9] for groups A, B and C respectively, again showing a significant decrease in PTV between time periods B and C ($p<0.0001$).

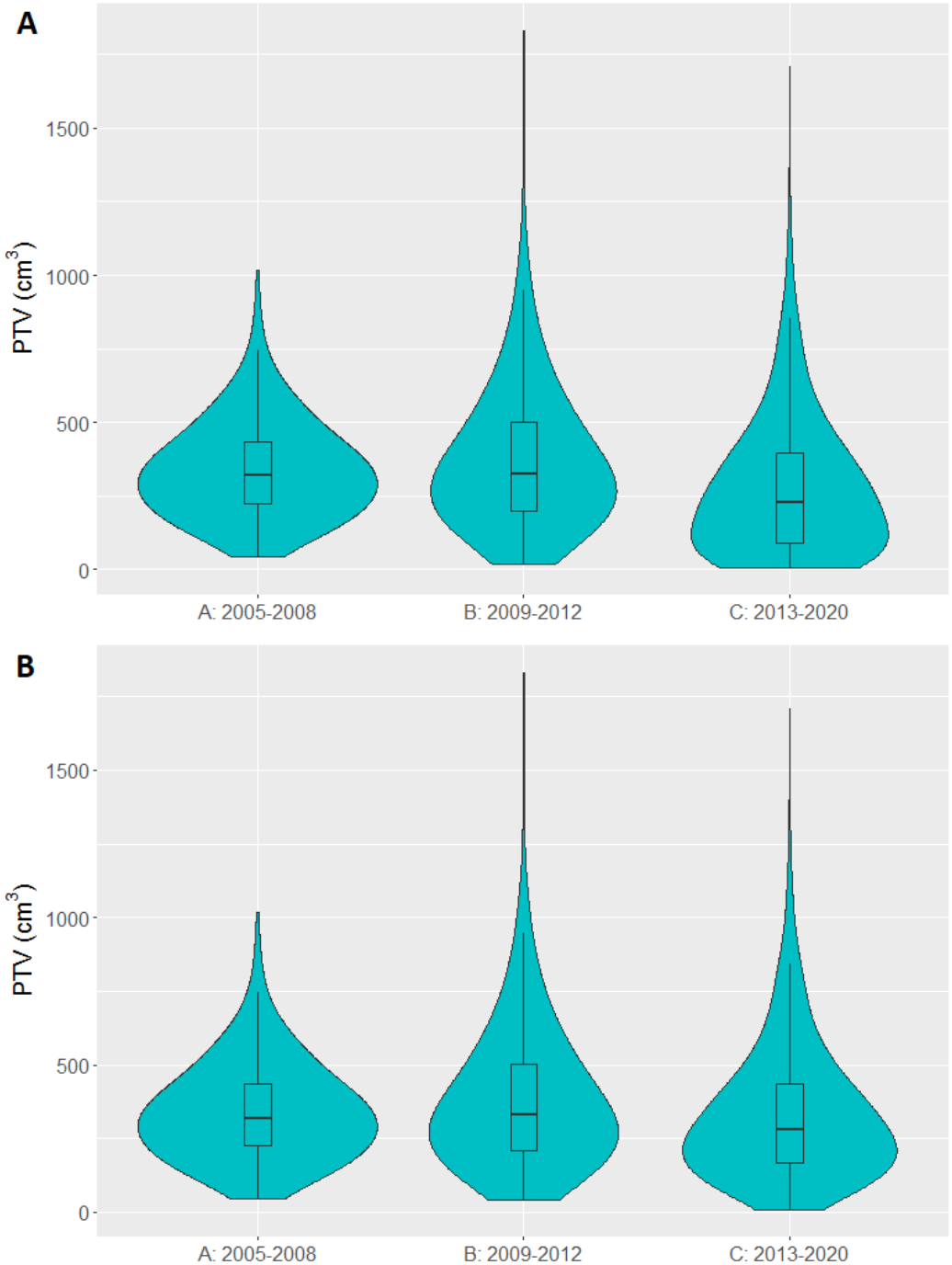

Figure 4.5: Violin plot presenting the distribution of PTVs from patients treated with curative-intent radiotherapy in each time period. (A) SABR patients included (B) SABR patients excluded.

Univariable survival analysis showed that the survival of patients treated with curative-intent radiotherapy has significantly improved in time period C compared to A (HR=0.847 (0.786-0.913), p<0.001). When patients treated with SABR were removed from the analysis, there was only a survival benefit for patients in time period B compared to A (HR=1.09 (1.00-1.18), p=0.0486), not for time period C compared to A (HR=0.949 (0.879-1.02), p=0.180). Kaplan-Meier curves are presented in Figure 4.6 for all curative-intent patients and curative-intent without SABR. Multivariable survival analysis, however, showed a survival benefit for patients treated in time period C compared to A for all curative-intent patients (HR=0.725 (0.632-0.831), p<0.001) as well as when patients treated with SABR were removed from the analysis (HR=0.757 (0.658-0.870), p<0.001). Full results are presented in Supplementary Tables 4.7 and 4.8.

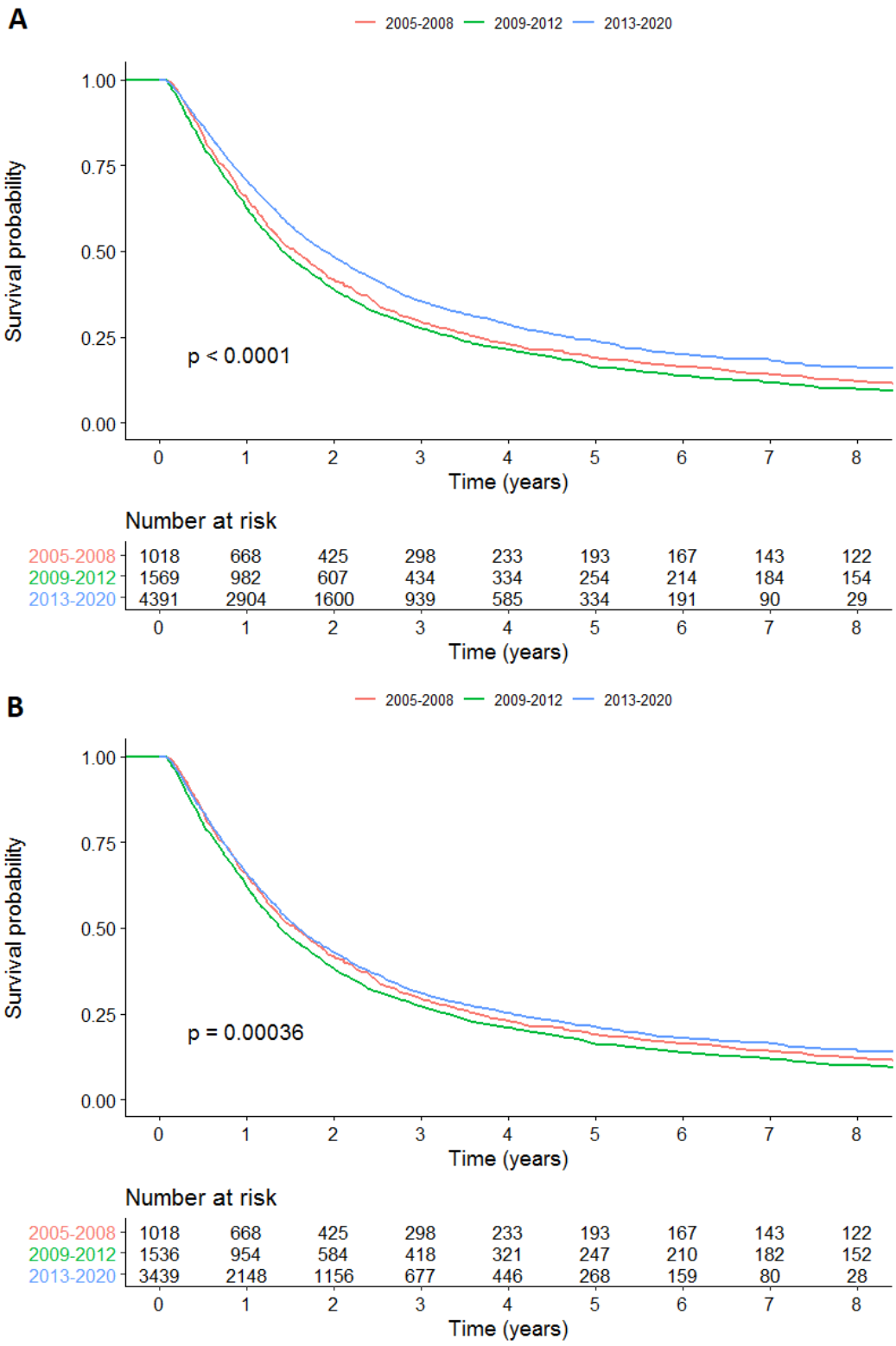

Figure 4.6: Kaplan-Meier survival curves for each time period for all patients treated with curative-intent radiotherapy (A) and curative-intent without SABR (B).

We conducted an analysis in patients with stage III disease. Kaplan-Meier curve is

presented in Figure 4.7 for patients with stage III treated with curative-intent. Univariable survival analysis showed no significant improvement or worsening of survival for time period C compared to A (HR=0.969 (0.832, 1.13), p=0.683). Multivariable survival analysis however, showed a survival benefit for patients treated in time period C compared to A for patients with stage III disease treated with curative-intent (HR=0.740 (0.600-0.913), p=0.00489). Full results are presented in Supplementary Table 4.9.

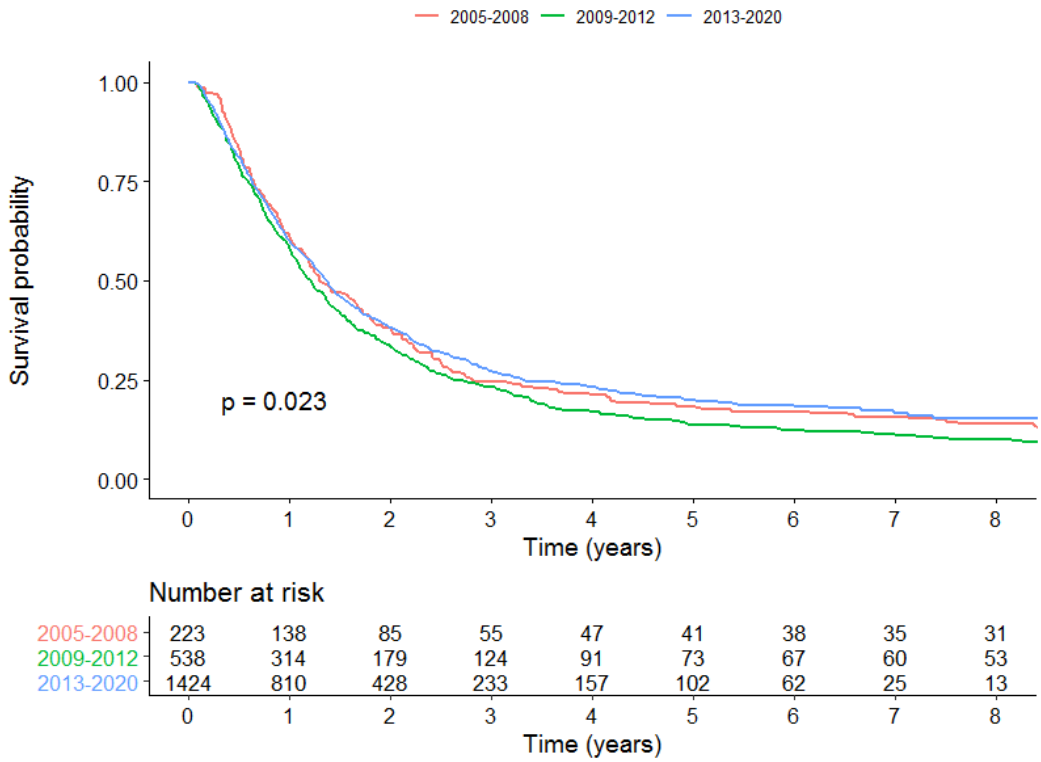

Figure 4.7: Kaplan-Meier survival curves for each time period for patients with stage III disease curative-intent radiotherapy.

4.4 Discussion

In this big data analysis, there has been a steady increase in the proportion of patients treated with curative-intent radiotherapy, across all PS groups and stages of disease. In addition, survival improved in the era when there was full access to IMRT (2013-2020) compared to no access to IMRT (2005-2008) when clinical variables were adjusted for. The introduction of IMRT has allowed the delivery of curative-intent doses to patients with tumours previously considered to be unsuitable for such an approach

due to large volume or proximity to critical organs at risk. In addition, the normal tissue sparing that IMRT facilitates enabled the treatment of patients with poorer performance status due to better tolerance of the treatment.

Our analysis showed that the proportion of patients with stage III lung cancer receiving curative-intent treatment has increased over the time periods, across all PS. This change has been partly facilitated by IMRT which allows the treatment of large and complex volumes. Other factors may have played a role, such as 4D CT planning (introduced 2011) facilitating more individualised treatment volumes, and also availability of radiotherapy at new satellite centres from 2010, allowing more patients (particularly the elderly and patients with poorer PS) to be treated nearer home. Accepting this, we still feel that as GTV volumes increased over the time periods studied, the introduction of IMRT is likely to have contributed greatly towards the proportion of patients able to receive curative-intent treatment. Baseline PET imaging has been standard at our institution since 2001 and so should not account for differences observed between the groups.

The survival benefit demonstrated on multivariable analysis was not seen in the unadjusted analysis, reflecting that patients with poorer performance status and larger tumours are being treated in the latest time period. As lung cancer outcome is associated with tumour volume [235, 236], it was expected that the survival in this group might have been worse in comparison to earlier time frames. However, survival improved for patients in the latest time period despite larger gross tumour volumes and an increased number of patients with poorer PS suggesting that planning with IMRT leads to at least non-inferior survival. In particular, when patients treated with SABR were removed from the analysis we showed that despite a significant increase in GTV in patients treated with curative-intent, the survival benefit in the latest time period remained.

Whilst this survival gain could be partly attributed to IMRT it is important to recognise that other changes in lung cancer management have occurred in the intervening time period we examined, and so we cannot claim that IMRT has directly led to an improvement in survival. Technological advances such as SABR, 4D radiotherapy and

image guidance radiotherapy have allowed reduced radiotherapy planning margins, leading to reduced normal tissue doses. The doses we used for curative intent stayed the same throughout the study. In our series, median PTV volumes were lower in the later timeframe (C) compared to either of the earlier timeframes, even when patients treated with SABR were excluded. This is likely to reflect a change in our CTV-PTV expansion margins which were introduced in the later time period following a move to daily image verification. It is unlikely that this reduction in PTV volume is responsible for the increased survival seen in the later timeframe, as although the difference was found to be statistically significant, in clinical terms the differences in PTV volume seen between group C and groups A and B is small. Also, GTV volume is known to be an independent prognostic factor for lung cancer survival, and we have previously demonstrated that this parameter increased between the three time periods.

Non-radiotherapy factors such as improved diagnostic imaging techniques, endobronchial ultrasound (EBUS) and associated stage migration, a change in staging classification, improvements in systemic therapy and supportive care may all have led to better outcomes. With regards to systemic therapy, the last 15 years have seen better integration of radiotherapy and systemic treatment, as well as the development of more targeted agents and immunotherapy that can be used on progression. Unfortunately due to the fact that this study started in 2005 data on chemotherapy were not as complete in the first time period due to lack of availability of electronic records for systemic therapy at the time. It was therefore not possible to guarantee a full, accurate and therefore meaningful collection of data on systemic treatment the patient may have received at the time of radiotherapy, or subsequently on progression. It is worth noting however that systemic treatment in the context of concurrent chemoradiotherapy had not changed significantly until the introduction of adjuvant Durvalumab, which has only been in routine use in the U.K. since 2019 (the latter part of our latest time period).

There are other limitations to this study including its retrospective design, and as is always the case when performing big data analyses, there is a significant amount of missing data within the clinical variables, including the lack of data on systemic therapy. This was more evident in the earlier time frames which were prior to our

in house electronic e-record being created, which facilitated the prospective collection of key data on outcome forms. We feel the large number of patients included in this analysis in part mitigates the issue of missing data [59]. Furthermore, this study reports on a unique dataset that evaluates real-world data from patients that are typically excluded from clinical trials. It is also worth noting that we have purposefully included a heterogeneous population of lung cancer patients with differing histologies into this analysis as we were interested in evaluating the impact of IMRT on curative-intent treatment. Admittedly the dose threshold for curative intent of greater than 40 Gy may also have included patients with NSCLC who did not fully complete their treatment, but in the context of such a large study, the numbers of patients whom this applies to are expected to be low.

These results are of particular importance in the UK, following publication of the most recent national lung cancer audit [231]. This highlighted that the majority of stage III NSCLC patients are receiving best supportive care or palliative treatment, even when patients have a PS of 0/1. In addition, there was a large regional variation in the percentage of patients receiving curative intent treatment from 8-80% [231]. It has been suggested that the centres offering a greater proportion of patients curative intent treatment may have better access to optimal radiotherapy planning techniques and image guided treatment [237]. Indeed, in the Royal College of Radiologists (RCR) published consensus statements for radiotherapy for lung cancer, it is recommended that patients receiving radical radiotherapy are planned with advanced techniques such as IMRT or VMAT [238].

The implementation of IMRT for the curative-intent treatment of lung cancer has lagged behind that of other disease sites such as head & neck cancers. This may stem from a perceived lack of high level evidence for using the technique. To date, there has only been one prospective study looking at the impact of IMRT on treatment toxicity and survival [233]. Chun et al. compared the outcome of patients treated with IMRT to 3D-CRT within the RTOG 0617 trial, reporting that despite larger planning target volumes in the IMRT group, patients had lower rates of grade 3+ pneumonitis and lower cardiac doses, however no difference in survival between the groups was observed [233]. A retrospective study by Yom et al. showed that patients treated with IMRT had

larger GTVs compared to matched patients treated with 3D-CRT. Similarly to Chun et al., they reported lower rates of grade 3+ pneumonitis in the IMRT group [239]. On the other hand, due to the complexity and cost of delivering IMRT, it has been suggested that 3D-CRT is still an equally sound option for locally advanced NSCLC, particularly for less experienced centres [240]. A meta-analysis of studies comparing IMRT to 3D-CRT reported survival to be similar between the two techniques, however there were reduced incidence of grade 2 pneumonitis and increased grade 3 oesophagitis in the IMRT group [241]. Overall the available data suggests that IMRT facilitates treatment of larger volumes, does not lead to inferior survival in NSCLC patients and should be employed to reduce dose to organs at risk, particularly to the heart and lung [233, 241].

IMRT and other advanced radiotherapy planning techniques offer the opportunity to achieve more than just treating larger volumes. Due to its ability to sculpt dose around the treatment volume, it may be possible to safely deliver a higher dose to the tumour, without compromising normal tissue toxicity. The hypothesis is that higher dose should equate to improved local control, and subsequently better survival. The RTOG 0617 study results however suggested that dose escalating with conventional fractionation does not seem to offer a benefit. It should be noted that only 47% patients in this study were planned with IMRT, dose to the heart was not prioritised in radiotherapy planning and further analysis has shown that higher cardiac dose in this trial is associated with worse survival [233]. Since the publication of RTOG 0617, further studies have demonstrated that excess radiation dose to the heart is associated with a decrease in survival [242]. A number of studies are have addressed the question of isotoxic dose escalation and dose painting based on FDG PETCT which are facilitated by the use of IMRT [243].

Looking forward, it may be possible in the future to perform causal inference analyses, which would help establish whether the increased proportion of patients treated with curative intent, and their improved survival, is indeed attributable to the introduction on IMRT. The data could also be enhanced by including treatment related toxicity, something that can now be achieved through the use of patient reported outcomes and proactive, prospective clinician reported toxicity, which we are now documenting at

our centre on an eform at each outpatient visit [57].

In summary, this big data analysis has demonstrated that the introduction of IMRT was associated with an increasing proportion of patients with lung cancer receiving curative-intent radiotherapy, across all PS and stages of disease. Despite treating larger, more complex tumours with curative-intent, and more patients with poor performance status, a survival benefit was seen for patients treated when full access to IMRT was available. This study highlights the impact IMRT has had on our practice, acknowledging that other contributing factors such as improvement in staging, technical radiotherapy and systemic therapy may have also contributed to the improved survival. We would recommend that IMRT is available for routine use for lung cancer patients who are being considered for treatment with curative intent. Current evidence suggests that this technique, at the very least, leads to non-inferior outcomes, and may facilitate improved outcomes firstly through the greater number of patients with stage III disease being able to receive a curative-intent dose, and secondly through a reduction of dose to the normal tissues.

4.5 Funding

This work was supported by CRUK via the funding to Cancer Research UK Manchester Centre: [C147/A18083] and [C147/A25254]. CF-F is supported by NIHR Manchester Biomedical Research Centre.

4.6 Supplementary materials

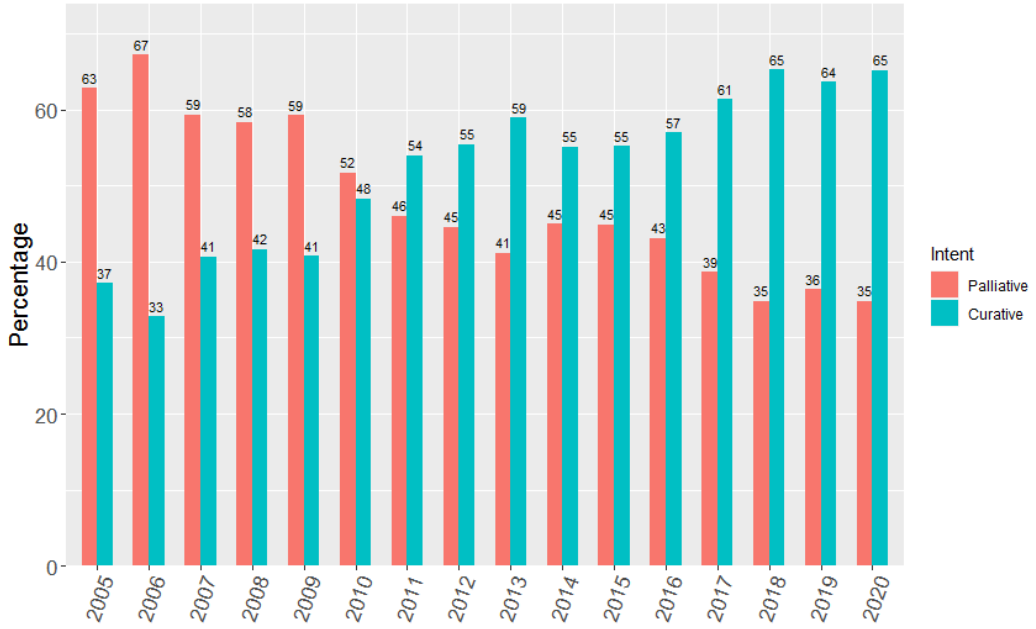

Figure 4.8: Percentage of patients treated with curative versus palliative intent, non-SABR radiotherapy year on year from 2005 to 2020.

Figure 4.9: Percentage of patients treated with curative versus palliative intent, non-SABR radiotherapy in each of the pre-specified time periods.

Table 4.5: Proportion of patients treated with curative-intent, non-SABR radiotherapy across each PS and time period.

PS	**A: 2005-2008** **% curative-intent (n curative-intent/n total)**	**B: 2009-2012** **% curative-intent (n curative-intent/n total)**	**C: 2013-2020** **% curative-intent (n curative-intent/n total)**
0 (n=1119)	52.1 (148/284)	65.1 (181/278)	69.7 (388/557)
1 (n=3938)	43.9 (374/852)	60.1 (636/1059)	66.7 (1353/2027)
2 (n=2863)	34.8 (165/474)	59.7 (381/752)	60.5 (990/1637)
3 (n=1190)	15.6 (26/167)	21.4 (74/346)	37.4 (253/677)

Table 4.6: Proportion of patients treated with curative-intent, non-SABR radiotherapy across each stage and time period.

Stage	**A: 2005-2008** **% curative-intent (n curative-intent/n total)**	**B: 2009-2012** **% curative-intent (n curative-intent/n total)**	**C: 2013-2020** **% curative-intent (n curative-intent/n total)**
I (n=1420)	76.9 (247/321)	90.9 (378/416)	94.6 (646/683)
II (n=1008)	70.3 (111/158)	84.8 (206/243)	91.3 (554/607)
III (n=3231)	40.4 (223/552)	66.3 (536/808)	75.9 (1420/1871)
IV* (n=2351)	2.11 (3/142)	9.96 (51/512)	14.5 (246/1697)

Table 4.7: Survival analysis results from the multivariable analysis of all curative-intent patients. 3188 patients with no missing variables were included.

	HR (95% CI)	***P* value**
Time period (ref A: 2005-2008)		
B: 2009-2012	0.953 (0.822, 1.11)	0.523
C: 2013-2020	0.725 (0.632, 0.831)	<0.001
PS (ref 0)		
1	1.31 (1.15, 1.50)	<0.001
2	1.64 (1.42, 1.89)	<0.001
3	1.73 (1.44, 2.08)	<0.001
4	1.28 (0.318, 5.16)	0.726
Stage (ref I)		
II	1.54 (1.36, 1.74)	<0.001
III	1.67 (1.50, 1.86)	<0.001
IV	2.04 (1.70, 2.43)	<0.001
Age at start of treatment (years)	1.01 (1.01,1.02)	<0.001
GTV (cm^3)	1.00 (1.00, 1.00)	<0.001

Table 4.8: Survival analysis results from the multivariable analysis of curative-intent patients without SABR. 2749 patients with no missing variables were included.

	HR (95% CI)	***P* value**
Time period (ref A: 2005-2008)		
B: 2009-2012	0.966 (0.832, 1.12)	0.646
C: 2013-2020	0.757 (0.658, 0.870)	<0.001
PS (ref 0)		
1	1.29 (1.12, 1.47)	<0.001
2	1.58 (1.37, 1.83)	<0.001
3	1.62 (1.33, 1.97)	<0.001
4	1.04 (0.145, 7.40)	0.972
Stage (ref I)		
II	1.41 (1.24, 1.61)	<0.001
III	1.53 (1.37, 1.72)	<0.001
IV	1.93 (1.59, 2.33)	<0.001
Age at start of treatment (years)	1.01 (1.01,1.02)	<0.001
GTV (cm^3)	1.00 (1.00, 1.00)	<0.001

Table 4.9: Survival analysis results from the multivariable analysis of stage III curative-intent patients. 1370 patients with no missing variables were included.

	HR (95% CI)	***P* value**
Time period (ref A: 2005-2008)		
B: 2009-2012	0.966 (0.771, 1.21)	0.767
C: 2013-2020	0.740 (0.600, 0.913)	0.00489
PS (ref 0)		
1	1.30 (1.09, 1.54)	0.00282
2	1.58 (1.30, 1.91)	<0.001
3	1.86 (1.37, 2.52)	<0.001
Age at start of treatment (years)	1.01 (1.01,1.02)	<0.001
GTV (cm^3)	1.00 (1.00, 1.00)	<0.001

Chapter 5

Impact of the COVID-19 pandemic on outcomes for patients with lung cancer receiving curative-intent radiotherapy in the UK

This chapter has been published in Clinical Oncology 2023 Volume 35 Issue 10 pages e593-e600. The section numbering, references and formatting have been modified to be consistent with the rest of the thesis.

Authors

Isabella Fornacon-Wood[1], Kathryn Banfill[1,2], Shahreen Ahmad[3], Anna Britten[4], Carrie Carson[5], Nicole Dorey[6], Matthew Hatton[7], Crispin Hiley[8], Kamalram Thippu Jayaprakash[9], Apurna Jegannathen[10], Pek Koh[11], Andrew C Kidd[12], Niki Panakis[13], Clive Peedell[14], Adam Peters[15], Anthony Pope[16], Ceri Powell[17], Claire Stilwell[18], Betsan Thomas[19], Elizabeth Toy[20], Kate Wicks[1], Victoria Wood[21], Sundus Yahya[22], Gareth Price[1] and Corinne Faivre-Finn[1,2]

Affiliations

[1] The University of Manchester, Manchester, UK.

[2] The Christie NHS Foundation Trust, Manchester, UK.

[3] Guy's and St Thomas' NHS Foundation Trust, London, UK.

[4] Brighton and Sussex University Hospitals NHS Trust, Brighton, UK.

[5] The Northern Ireland Cancer Centre, Belfast, UK.

[6] Torbay and South Devon NHS Foundation Trust, Torquay, UK.

[7] Weston Park Hospital, Sheffield, UK.

[8] University College London Hospitals, London, UK.

[9] Oncology Centre, Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK.

[10] University Hospitals North Midlands, Stoke-on-Trent, UK.

[11] Royal Wolverhampton NHS Trust, Wolverhampton, UK.

[12] Singleton Hospital, Sketty Lane, Sketty, Swansea, UK.

[13] Oxford Universities NHS Trust, Oxford, UK.

[14] The James Cook University Hospital, Middlesbrough, UK.

[15] Beatson West of Scotland Cancer Centre, UK.

[16] Clatterbridge Cancer Centre, Bebington, UK.

[17] Velindre Cancer Centre, Cardiff, UK.

[18] Aberdeen Royal Infirmary, Aberdeen, UK.

[19] Swansea Bay University Hospital, Swansea, UK.

[20] Royal Devon and Exeter NHS Foundation Trust, Exeter, UK.

[21] University Hospitals Southampton NHS Foundation Trust, Southampton, UK.

[22] University Hospitals Birmingham, Birmingham, UK.

Author contributions

I prepared the database and decided on the statistical analysis plan. I wrote the R script to perform the statistical analysis. I wrote the manuscript, which was reviewed by all authors.

Abstract

Background

Previous work found that during the first wave of the COVID-19 pandemic, 34% of patients with lung cancer treated with curative-intent radiotherapy (RT) in the UK had a change to their centre's usual standard of care treatment (Banfill 2021). We present the impact of these changes on patient outcomes.

Materials and Methods

The COVID-RT Lung database was a prospective multicentre UK cohort study including patients with stage I-III lung cancer referred for and/or treated with radical RT between April and October 2020. Data were collected on patient demographics, RT and systemic treatments, toxicity, relapse, and death. Multivariable cox and logistic regression were used to assess the impact of having a change to RT on survival, distant relapse and $\geq$ grade 3 acute toxicity. The impact of omitting chemotherapy on survival and relapse was assessed using multivariable cox regression.

Results

Patient and follow-up forms were available for 1280 patients. Seven hundred and sixty-five (59.8%) patients were aged over 70 and 603 (47.1%) were female. Median follow-up was 213 days (119, 376). Patients with stage I-II NSCLC who had a change to their RT had no significant increase in distant relapse (p=0.859) or death (p=0.884); however, did have increased odds of $\geq$ grade 3 acute toxicity (p=0.0348). Patients with stage III NSCLC who had a change to their RT had no significant increase in distant relapse (p=0.216) or death (p=0.789); however, did have increased odds of $\geq$ grade 3 acute toxicity (p<0.001). Patients with stage III NSCLC who had their chemotherapy omitted had no significant increase in distant relapse (p=0.0827) or death (p=0.0661).

Conclusion

This study suggests changes to RT and chemotherapy made in response to the COVID-19 pandemic did not significantly affect distant relapse or survival. Changes to RT, namely increased hypofractionation, led to increased odds of $\geq$ grade 3 acute toxicity. These results are important as hypofractionated treatments can help to reduce

hospital attendances in the context of potential future emergency situations.

5.1 Introduction

The COVID-19 pandemic put an unprecedented demand on NHS services which in turn affected cancer treatments, including radiotherapy [244, 245]. The effects of the COVID-19 pandemic on outcomes for patients with cancer is of increasing concern. It is known that patients with cancer have higher rates of severe disease and death with COVID-19 compared to the general population [246], however evidence is still lacking on the indirect impact of COVID-19 on cancer treatments.

Radiotherapy plays a key role in the treatment of lung cancer, with radiotherapy being indicated for 61-82% of patients [247, 248]. Radiotherapy alone or in combination with chemotherapy and/or immunotherapy is an important treatment modality in the curative-intent setting. These patients are particularly vulnerable due to the immunosuppressive nature of treatments and multiple comorbidities [249]. Therefore, guidelines were rapidly produced at the start of the pandemic with the aim to reduce hospital visits without compromising treatment benefit by using reduced-fractionation regimens for patients receiving curative-intent radiotherapy [250].

COVID-RT lung was a UK data collection initiative that aimed to assess the impact of the COVID-19 pandemic in patients with stage I-III lung cancer receiving curative-intent radiotherapy [76]. Previous analysis reported that 34% of patients had a change to their centre's usual standard of care treatment [76], and 17.5% of patients had a different radiotherapy dose and/or fractionation, with an increased use of hypofractionated regimens as was recommended by UK guidelines [250]. In patients with stage III disease considered for chemotherapy, 10.7% of patients had this treatment modality omitted, and 6.7% had a reduced chemotherapy dose. We present in this paper the impact of these changes to treatment on patient outcomes.

5.2 Methods and Materials

COVID-RT Lung is a prospective, multicentre UK cohort study. Data were prospectively collected on all patients with stage I-III lung cancer referred for and/or treated with curative-intent radiotherapy (biologically equivalent dose >50 Gy) between April and October 2020. The data collection procedure has been described previously [76]. For this analysis, the following baseline clinical information was extracted from the COVID-RT Lung database on the 25/07/2022: age at the time of treatment; gender; histology; stage; baseline Eastern Cooperative Oncology Group (ECOG) Performance Status (PS); radiotherapy dose and fractionation; dates of radiotherapy; chemotherapy delivery; and immunotherapy delivery. Data were also collected on whether patients had a change to their centre's standard of care treatment pre-COVID-19, radiotherapy treatment, or chemotherapy treatment. The specific changes made to radiotherapy and chemotherapy treatment were not recorded. A reduction in systemic treatment was defined as a reduced number of planned cycles and/or a reduced dose for any single cycle. Follow-up data were collected 12 months after the end of treatment, or earlier for acute toxicity or if the patient died. Follow-up data were collected on: distant and loco-regional relapse within a year post radiotherapy, death, $\geq$ grade 3 treatment related acute toxicity (toxicity within 3 months of the end of radiotherapy) and late toxicity (toxicity from 3 months after the end of radiotherapy) according to CTCAE v5.0 (oesophageal, pulmonary and cardiac toxicity, lung infection and chest pain).

Baseline characteristics were summarised as counts and percentages, and medians with the lower and upper quartiles. Age was dichotomised at 70 years in line with the UK Government's shielding advice. Dose per fraction was grouped into $<$ 2 Gy/fraction, 2 Gy/fraction, >2-2.9Gy/fraction, 3-5.9 Gy/fraction and $\geq$ 6 Gy/fraction. These groupings were chosen to highlight different fractionation regimens; 2 Gy/fraction represents conventional fractionation, and above that moderate to ultra-hypofractionation. Patients with a radiological diagnosis of cancer were assumed to have NSCLC. Hazard ratios (HR) and 95% confidence intervals were estimated to describe the hazard of death, and distant and loco-regional relapse for patients who had a change to their radiotherapy dose and/or fractionation, in response to the COIVD-19 pandemic, using

multivariable cox regression. Regressions were adjusted for age, gender, PS, whether the patient received chemotherapy, and radiation dose per fraction. Adjusted odds ratios (aOR) and 95% confidence were estimated to describe the risk of developing $\geq$ grade 3 acute toxicity for patients who had a change to their radiotherapy dose and/or fractionation, in response to the COVID-19 pandemic, using multivariable logistic regression, adjusting for age, gender, PS, whether the patient received chemotherapy, and radiation dose per fraction. HRs and 95% confidence intervals were estimated to describe the hazard of death and distant relapse for patients who had their chemotherapy omitted using multivariable cox regression, adjusting for age, gender, PS, whether the patient had a change to their radiotherapy dose/and or fractionation, and radiation dose per fraction. Multivariable analysis was not performed for the SCLC data, or late toxicity, due to insufficient sample size in these cohorts. Multivariable logistic regression was used to determine whether patients who had a change to their chemotherapy regimen in response to the COVID-19 pandemic were also more likely to have a change to their radiotherapy dose and/or fractionation, adjusting for age and PS. Mean dose per fraction was compared between groups using the t-test. Patients who had their radiotherapy delivered in < 15 fractions were removed from the stage III NSCLC and SCLC analyses, as they are palliative or Stereotactic Ablative Body Radiotherapy (SABR) regimens.

All statistical analyses were performed in R 4.0.0 [222] with package survival v3.1-12 [224].

5.3 Results

Completed patient and follow-up forms were available in the COVID-RT Lung database for 1280 patients (out of 1717) treated between April and October 2020. Median follow-up was 213 days (119, 376). Baseline characteristics split by change to local standard of care treatment are presented in Table 5.1. Seven hundred and sixty-five (59.8%) patients were aged over 70 and 603 (47.1%) were female. Two hundred and fifty-nine (61.7%) patients who had a change to treatment were aged over 70 and 116 (27.6%) patients who had a change to treatment had a PS of 2-3. A higher proportion of patients who had their treatment changed received 3-5.9 Gy/fraction compared to

patients who had no change to treatment (16.0% vs 5.6%).

Changes to local standard of care treatment have been presented previously [76]. To briefly summarise, the main change to treatment for patients with stage I-II disease was a change to radiotherapy dose and/or fractionation (16.1%), followed by radiotherapy being given instead of surgery (9.5%) [76]. For patients with stage III disease, the main change was a change to radiotherapy dose and/or fractionation (19.5%), followed by having their chemotherapy omitted (10.7%), or receiving a reduced chemotherapy dose (6.8%) [76].

Table 5.1: Baseline characteristics. The percentages describe the percentage of patients for each clinical variable per column.

	No change to treatment (N=860)	**Change to treatment (N=420)**	**All patients (N=1280)**
Age in years, n (%)			
< 70	351 (40.8)	161 (38.3)	512 (40.0)
≥ 70	506 (58.8)	259 (61.7)	765 (59.8)
Missing	3 (0.3)	0	3 (0.2)
Gender, n (%)			
Female	422 (49.1)	181 (43.1)	603 (47.1)
Male	437 (50.8)	239 (56.9)	676 (52.8)
Missing	1 (0.1)	0	1 (0.1)
PS, n (%)			
0	100 (11.6)	65 (15.5)	165 (12.9)
1	416 (48.4)	239 (56.9)	655 (51.2)
2–3	342 (39.8)	116 (27.6)	458 (35.8)
Missing	2 (0.2)	0	2 (0.2)
Histology, n (%)			
NSCLC	508 (59.1)	276 (65.7)	784 (61.3)
SCLC	62 (7.3)	54 (12.9)	116 (9.1)
Radiological diagnosis	289 (33.6)	215 (51.2)	379 (29.6)
Missing	1 (0.1)	0	1 (0.1)
Stage, n (%)			
I	395 (45.9)	147 (35.0)	542 (42.3)
II	135 (15.7)	56 (13.3)	191 (14.9)
III	327 (38.0)	215 (51.2)	542 (42.3)
Missing	3 (0.3)	2 (0.5)	5 (0.4)
Mean dose per fraction in Gy/fraction (SD)	6.61 (5.12)	6.61 (6.70)	6.61 (5.68)

Table 5.1 continued from previous page

	No change to treatment (N=860)	**Change to treatment (N=420)**	**All patients (N=1280)**
Dose per fraction grouped, n (%)			
$<$ 2 Gy/fraction	36 (4.2)	1 (0.2)	37 (2.9)
2 Gy/fraction	37 (4.3)	9 (2.1)	46 (3.6)
>2-2.9 Gy/fraction	372 (43.3)	211 (50.2)	583 (45.5)
3-5.9 Gy/fraction	48 (5.6)	67 (16.0)	115 (9.0)
$\geq$ 6 Gy/fraction	357 (41.5)	127 (30.2)	484 (37.8)
Missing	10 (1.2)	5 (1.2)	15 (1.2)

Abbreviations: *Gy*, Gray; *NSCLC*, non-small cell lung cancer; *PS*, Performance status; *RT*, Radiotherapy; *SCLC*, small cell lung cancer; *SD*, standard deviation.

5.3.1 Changes to radiotherapy dose and/or fractionation

Stage I-II NSCLC

Seven-hundred and six patients had stage I-II NSCLC, of which 106 (15.0%) had a change to their radiotherapy dose and/or fractionation. Table 5.2 presents toxicity and outcomes data for these patients. Rates of distant and loco-regional relapse and death were similar between stage I-II NSCLC patients who had a change to their radiotherapy and those who did not (6.6% vs 8.7%, 11.3% vs 11.7% and 11.3% vs 13.0%). For patients who had a change to their radiotherapy, 5 (4.7%) had $\geq$ grade 3 acute toxicity and 1 (0.9%) had $\geq$ grade 3 late toxicity. For patients who did not have a change to their radiotherapy, 13 (2.2%) had $\geq$ grade 3 acute toxicity and 8 (1.3%) had $\geq$ grade 3 late toxicity.

Multivariable analysis showed that patients with stage I-II NSCLC who had a change to their radiotherapy dose and/or fractionation had no significant increased hazard of distant relapse (HR=1.09 (0.412, 2.90), p=0.859), loco-regional relapse (HR=1.25 (0.609, 2.58), p=0.541), or death (HR=0.951 (0.480, 1.88), p=0.884). These patients did, however, have increased odds of developing $\geq$ grade 3 acute toxicity (aOR=3.46 (1.01, 10.6), p=0.0348) in this dataset. The full multivariable results can be found in Supplementary Tables 5.5 and 5.6. Patients with stage I-II NSCLC who had a change

to their radiotherapy dose and/or fractionation received a higher dose per fraction (mean 12.1 vs 9.22, p<0.001) compared to patients who had no change.

Table 5.2: Toxicity and disease status for patients with stage I-II NSCLC split by whether they had a change to their radiotherapy dose and/or fractionation or not.

	Change to RT (N=106)	**No change to RT (N=600)**
Acute toxicity ≥ grade 3, n (%)	5 (4.7)	13 (2.2)
Oesophageal	1 (0.9)	4 (0.7)
Pulmonary	2 (1.9)	4 (0.7)
Cardiac	0	2 (0.3)
Lung infection	2 (1.9)	2 (0.3)
Chest pain	0	0
Other	0	1 (0.2)
Missing	6 (5.7)	34 (5.7)
Late toxicity ≥ grade 3, n (%)	1 (0.9)	8 (1.3)
Oesophageal	0	1 (0.2)
Pulmonary	0	2 (0.3)
Cardiac	0	0
Lung infection	1 (0.9)	3 (0.5)
Chest pain	0	2 (0.3)
Other	0	0
Missing	22 (20.8)	121 (20.2)
Disease status, n (%)		
Distant relapse	7 (6.6)	52 (8.7)
Loco-regional relapse	12 (11.3)	70 (11.7)
No evidence recurrence	79 (74.5)	416 (69.3)
Death, n (%)	12 (11.3)	78 (13.0)

Abbreviations: *RT*, Radiotherapy.

Supplementary Table 5.7 presents ≥ grade 3 acute and late toxicity for patients with stage I-II NSCLC who received 5 fraction SABR versus 3 fraction SABR. Rates of ≥ grade 3 acute (2.3% vs 3.5%) and late toxicity (1.5% vs 2.1%) were similar between patients who received 5 fraction SABR versus 3 fraction SABR. Overall toxicity rates were low. There were no cases of ≥ grade 3 chest wall pain for patients who received 3 or 5 fraction SABR in this dataset.

Stage III NSCLC

Four hundred and twenty-five patients had stage III NSCLC, of which 77 (18.1%) had a change to their radiotherapy dose and/or fractionation. Table 5.3 presents toxicity and outcomes data for these patients. Patients who had a change to their radiotherapy had lower rates of distant (10.4% vs 21.3%) and loco-regional (7.8% vs 21.6%) relapse, but death rates were similar (23.4% vs 25.9%). A higher proportion of patients who had a change to their radiotherapy had $\geq$ grade 3 acute and late toxicity compared to patients who did not have a change to their radiotherapy (acute toxicity 19.5% vs 9.2%, late toxicity 5.2% vs 1.7%). For patients who had a change to their radiotherapy, the majority of $\geq$ grade 3 acute (80.0%) and late (75.0%) toxicity was seen in patients who received concurrent chemotherapy; however, for patients who did not have a change to their radiotherapy, the majority of $\geq$ grade 3 acute (46.9%) and late (66.7%) toxicity was seen in patients who had no chemotherapy, although numbers are low (Supplementary Table 5.8).

Multivariable analysis revealed that patients with stage III NSCLC who had a change to their radiotherapy dose and/or fractionation had no significant increased hazard of distant relapse (HR=1.71 (0.731, 4.00), p=0.216), loco-regional relapse (HR=0.880 (0.316, 2.45), p=0.806), or death (HR=1.08 (0.606, 1.94), p=0.789). They did however, have increased odds of developing $\geq$ grade 3 acute toxicity (aOR=4.78 (2.11, 10.7), p<0.001) in this dataset. The full multivariable results can be found in Supplementary Tables 5.9 and 5.10. Patients with stage III NSCLC who had a change to their radiotherapy dose and/or fractionation received a higher dose per fraction (mean 3.01 vs 2.65, p<0.001) compared to patients who had no change to their radiotherapy dose and/or fractionation.

Table 5.3: Toxicity and disease status for patients with stage III NSCLC split by whether they had a change to their radiotherapy dose and/or fractionation or not.

	Change to RT (N=77)	**No change to RT (N=348)**
Acute toxicity ≥ grade 3, n (%)	15 (19.5)	32 (9.2)
Oesophageal	5 (6.5)	15 (4.3)
Pulmonary	6 (7.8)	7 (2.0)
Cardiac	0	0
Lung infection	2 (2.6)	4 (1.1)
Chest pain	0	1 (0.3)
Other	1 (1.3)	5 (1.4)
Missing	11 (14.3)	7 (2.0)
Late toxicity ≥ grade 3, n (%)	4 (5.2)	6 (1.7)
Oesophageal	3 (3.9)	1 (0.3)
Pulmonary	1 (1.3)	4 (1.1)
Cardiac	0	0
Lung infection	0	0
Chest pain	0	0
Other	0	1 (0.3)
Missing	22 (28.6)	63 (18.1)
Disease status, n (%)		
Distant relapse	8 (10.4)	74 (21.3)
Loco-regional relapse	6 (7.8)	75 (21.6)
No evidence recurrence	46 (59.7)	188 (54.0)
Death, n (%)	18 (23.4)	90 (25.9)

Abbreviations: *RT*, Radiotherapy.

Three hundred and twenty-eight (77.2%) patients with stage III NSCLC received 55 Gy in 20 fractions, of which 169 (51.5%) had radiotherapy alone, 79 (24.1%) received concurrent chemotherapy and 80 (24.4%) sequential chemotherapy. Thirty (7.1%) patients with stage III NSCLC received 60-66 Gy in 30-33 fractions, of which 24 (80.0%) received concurrent chemotherapy, 5 (16.7%) received sequential chemotherapy and 1 (3.3%) had radiotherapy alone.

Seventeen (21.5%) patients with stage III NSCLC who received 55 Gy in 20 fractions with concurrent chemotherapy had ≥ grade 3 acute toxicity, 7 (8.9%) of which were pulmonary and 5 (6.3%) oesophageal. 4 (5.1%) patients had ≥ grade 3 late toxicity. Five (20.8%) patients with stage III NSCLC who received 60-66 Gy in 30-33 fractions with concurrent chemotherapy had ≥ grade 3 acute toxicity, 1 (20.0%) pulmonary, 1

(20.0%) oesophageal and 2 (40.0%) lung infection. Four (16.7%) patients had ≥ grade 3 late toxicity.

SCLC

One hundred and nine patients had SCLC, of which 34 (31.2%) had a change to their radiotherapy dose and/or fractionation. Supplementary Table 5.11 presents toxicity and outcomes data for these patients. Patients who had a change to their radiotherapy had higher rates of distant relapse (35.3% vs 26.7%), loco-regional relapse (20.6% vs 14.7%), and death (35.3% vs 26.7%). 2 (5.9%) patients who had a change to their radiotherapy had ≥ grade 3 acute toxicity and 2 (5.9%) had ≥ grade 3 late toxicity. For patients who did not have a change to their radiotherapy, 9 (12.0%) had ≥ grade 3 acute toxicity and none had ≥ grade 3 late toxicity. Patients with SCLC who had a change to their radiotherapy dose and/or fractionation received a higher dose per fraction (mean 2.68 vs 2.34, $p<0.001$) compared to patients who had no change to their radiotherapy dose and/or fractionation.

Fifty-six (51.4%) patients with SCLC received 40 Gy in 15 fractions, of which 36 (64.3%) received sequential chemotherapy, 12 (21.4%) concurrent chemotherapy and 8 (14.3%) radiotherapy alone. Nineteen (17.4%) patients received 55 Gy in 20 fractions, of which 14 (73.7%) received sequential chemotherapy, 3 (15.8%) radiotherapy alone and 2 (10.5%) concurrent chemotherapy. 16 (14.7%) patients received 45 Gy in 30 twice-daily fractions, all of whom received concurrent chemotherapy. Three (2.8%) patients received 60-66 Gy in 30-33 fractions, of which 2 (66.7%) received concurrent chemotherapy and 1 (33.3%) received sequential chemotherapy. No patients with SCLC who had a change to their radiotherapy received ≤ 2 Gy/fraction.

5.3.2 Changes to chemotherapy regimen

Stage III NSCLC

Two hundred and sixty-one (61.4%) patients with stage III NSCLC were considered for chemotherapy as part of their management plan. However, 48 (18.0%) had their chemotherapy omitted and 35 (13.4%) had their chemotherapy dose and/or number of planned cycles reduced (Table 5.4). Patients who had their chemotherapy omitted had

a higher rate of distant relapse compared to those who had no change (31.2% vs 14.6%), and a higher rate of death (35.4% vs 20.2%). Fifty-six (21.5%) patients with stage III NSCLC who were considered for chemotherapy had consolidation immunotherapy. Of the patients with stage III NSCLC who had a change to their chemotherapy regimen, 12 (14.5%) also had a change to their RT dose and/or fractionation. Patients who had a change to their chemotherapy regimen were significantly less likely to have a change to their radiotherapy dose and/or fractionation (aOR=0.479 (0.222, 0.963), p=0.0470).

Multivariable analysis demonstrated no significant increase in distant relapse (HR=1.85 (0.923, 3.71), p=0.0827), loco-regional relapse (HR=1.03 (0.468, 2.27), p=0.940) or death (HR=1.80 (0.961, 3.40) P=0.0661) for patients who had their chemotherapy omitted, suggesting the higher rates of distant relapse and death in this group were not significantly associated with having their chemotherapy omitted. The full multivariable results can be found in Supplementary Table 5.12.

Table 5.4: Disease status for patients with stage III NSCLC split by whether they had their chemotherapy omitted, reduced, or received standard of care chemotherapy i.e. no change to chemotherapy regimen.

	Chemotherapy omitted (N=48)	**Chemotherapy dose/number of cycles reduced (N=35)**	**No change to chemotherapy (N=178)**
Disease status, n (%)			
Distant relapse	15 (31.2)	5 (14.3)	26 (14.6)
Loco-regional relapse	10 (20.8)	2 (5.7)	28 (15.7)
No evidence recurrence	24 (50.0)	23 (65.7)	98 (55.1)
Death, n (%)	17 (35.4)	5 (14.3)	36 (20.2)

SCLC

One hundred and four (95.4%) patients with SCLC were considered for chemotherapy as part of their management plan. However, 7 (6.7%) had their chemotherapy omitted and 14 (13.5%) had their chemotherapy dose and/or number of planned cycles reduced (Supplementary Table 5.13). Patients who had their chemotherapy omitted or reduced had higher rates of distant relapse compared to those who had no change (42.9% vs 50.0% vs 26.5%), and loco-regional relapse (57.1% vs 21.4% vs 10.8%). Rates

of death were similar between patients who had their chemotherapy reduced vs no change (35.7% vs 28.9%), and 1 (14.3%) patient who had their chemotherapy omitted died. Of the patients with SCLC who had a change to their chemotherapy regimen, 10 (47.6%) also had a change to their RT dose and/or fractionation. Patients who had a change to their chemotherapy regimen were not significantly more or less likely to have a change to their radiotherapy dose and/or fractionation (aOR=2.35 (0.851, 6.51), p=0.0965).

5.4 Discussion

The initial analyses of the COVID-RT Lung data found a third of patients had their treatment changed, from what they would usually have received, due to the COVID-19 pandemic. The most common change was receiving a different radiotherapy dose and/or fractionation to the centre's usual standard of care, typically increased use of hypofractionated radiotherapy [76]. This increased used of hypofractionated radiotherapy during the COVID-19 pandemic is in line with UK recommendations to reduce hospital attendances [250]. The key findings in this analysis are that there was no significant impact on distant/loco-regional relapse or mortality for patients with NSCLC who had a change to their radiotherapy dose and/or fractionation, and there was a small increase in $\geq$ grade 3 acute toxicity. Furthermore, for patients with stage III NSCLC who were considered for chemotherapy, omitting or reducing chemotherapy dose and/or number of cycles did not lead to a significant impact on distant/loco-regional relapse or mortality.

The effect of hypofractionated radiotherapy on outcomes is an important consideration, particularly as it has the advantage of fewer hospital visits and reduced overall treatment times. Although we did not have information on the specific radiotherapy regimen changes that took place in this study, patients who had a change to their radiotherapy received a higher dose per fraction, indicating increased use of hypofractionation. A randomised phase III trial with 96 patients with stage II-III NSCLC not fit for concurrent chemotherapy compared 60 Gy in 15 fractions over 3 weeks (hypofractionated arm) to 60 Gy in 30 fractions over 6 weeks (conventional arm), reporting no significant difference in 1-year survival (37.7% in the hypofractionated arm

vs 44.6% in the conventional arm), local and distant relapse, and ≥ grade 3 toxicity, although there was a higher rate of grade 2 toxicity in the hypofractionated group [251]. Our study, in contrast, did find increased ≥ grade 3 toxicity for patients who had a change to their radiotherapy, and therefore increased dose per fraction, however we did not compare specific radiotherapy regimens.

A retrospective analysis of 111 patients with NSCLC compared node-negative patients (a surrogate for patients not eligible for SABR) to node-positive patients (a surrogate for those unfit for chemo-radiotherapy) who received 60 Gy in 15 fractions at one institution [252]. The study found acceptable 1-year survival rates (86.5% node-negative versus 69.1% node-positive), local control and ≥ grade 3 toxicity. The study is limited by selection bias, as patients were treated with 60 Gy in 15 fractions when conventional radiotherapy or SABR were not appropriate. Another retrospective population-based study in the UK, however, reported significantly worse survival for patients with stage I-III NSCLC treated with 55 Gy in 20 fractions compared to 60-66 Gy in 30-33 fractions (25 months v 28 months, $p = 0.02$) [253]. This study was retrospective in nature, and did not distinguish between patients who received concurrent and sequential chemotherapy. The survival differential may therefore be caused by selection bias rather than the hypofractionated regimen, as patients in the UK often receive 55 Gy in 20 fractions following induction chemotherapy or if they are not fit enough for chemotherapy. Our study suggests the use of hypofractionated treatments during the pandemic did not affect survival.

The increased odds of ≥ grade 3 acute toxicity for patients with NSCLC who had a change to their radiotherapy dose and/or fractionation is to be expected due to the higher dose per fraction used in hypofractionated radiotherapy. The rate of COVID-19 infection in the COVID-RT Lung study was low (33 (2.1%) patients) [76]. Outcomes for patients with cancer who are infected with COVID-19 are worse [246], so our results suggest the changes to treatment put in place to reduce COVID-19 exposure were effective and may have prevented vulnerable deaths due to COVID-19, at the expense of a small increase in acute toxicity. Data on long-term toxicity were not collected. For patients with stage III NSCLC who had a change to their radiotherapy, the majority of toxicity was seen in patients who received concurrent chemotherapy;

however, for patients who did not have a change to their radiotherapy, the majority of toxicity was seen in patients who had no chemotherapy. There was a higher rate of patients with PS 2-3 in the no change to treatment group, which may explain this difference in toxicity, however it is important to note that numbers for this analysis were low.

Most patients with stage III NSCLC received the moderately hypofractionated regimen of 55 Gy in 20 fractions in this study, compared with the conventional radiotherapy regimen of 60-66 Gy in 30-33 fractions (77.2% vs 7.1%). Rates of $\geq$ grade 3 acute toxicity were similar between both regimens for patients who also received concurrent chemotherapy. A randomised phase II trial including 130 patients with stage III NSCLC and PS 0-1 receiving either concurrent or sequential chemotherapy with standard 55 Gy in 20 fractions over 4 weeks (SOCCAR) reported that 32% of patients who had concurrent chemotherapy had $\geq$ grade 3 acute toxicity [254]. Our study reported a lower rate of $\geq$ grade 3 acute toxicity (21.5%), which may relate to the use of more advanced radiotherapy techniques, such as IMRT and VMAT, since the completion of SOCCAR recruitment in 2010.

No patients with SCLC who had a change to their radiotherapy received twice-daily radiotherapy, suggesting once-daily radiotherapy was preferred to reduce the time spent in hospital during the pandemic. The CONVERT trial compared once-daily to twice-daily chemo-radiotherapy and found no difference in survival outcomes, although the study was powered to show superiority of once-daily chemo-radiotherapy and not equivalence [214].

This analysis found that patients with stage III NSCLC who had their chemotherapy omitted did not have a significant increase in risk of distant/loco-regional relapse or death in the multivariable analysis. This is in contrast to meta-analyses that show chemotherapy improves survival and tumour control in locally advanced lung cancer [32, 255]. Our results do show, however, that there was a higher rate of distant relapse and death in the chemotherapy omitted group compared to standard of care. This suggests there may be a baseline difference between patients who had their chemotherapy omitted and those who did not, which was taken into account in the multivariable

analysis by adjusting for clinical variables. Also, it is likely, since the confidence interval for the hazard ratio of both distant relapse and death is close to 1 and the p-values close to the significance cut-off of 0.05, that there may be a negative effect of omitting chemotherapy that could not be detected with the sample size or follow-up available in COVID-RT Lung. Unfortunately, a limitation of this study is that we could not collect more data as this is a unique dataset from a fixed time period. The risk of death due to COVID-19 is less now as many vulnerable patients with lung cancer are vaccinated and there are more effective treatments for patients hospitalised with COVID-19 [256]. Therefore, given the evidence that chemotherapy given in addition to radiotherapy improves survival in patients with stage III NSCLC [257], chemotherapy should no longer be omitted due to COVID-19 risk.

The results from this study are encouraging since the National Lung Cancer Audit found decreased 1-year survival for patients with lung cancer from 2019 to 2020, reversing the improvement in survival seen previously [258]. Our analysis did not find an increased risk of death for patients who had a change to radiotherapy or chemotherapy treatments between April and October 2020; however, our data only included patients who were treated with curative-intent radiotherapy and median follow-up is shorter. Previous analysis found that patients who had a change to treatment were more likely to be elderly ($\geq$ 70 years) [76], which is in line with the UK Government's shielding advice. This is important to consider when interpreting the mortality results from this analysis.

This study is subject to limitations, namely the sample size and short median follow-up of 7 months. Effects may, therefore, have been missed due to a lack of statistical power. Due to the unique circumstance of data collection, more data could not be collected during the time period encapsulating the first wave of the COVID-19 pandemic, from April to October 2020. Unfortunately, 437 follow-up forms were not filled out and therefore these patients could not be included in the outcomes analysis. Longer follow-up would be required to verify the findings in this study. Despite this, our study provides valuable information to inform treatments for patients with lung cancer in such exceptional circumstances.

In conclusion, this study showed that changes made to radiotherapy and chemotherapy treatments during the COVID-19 pandemic did not significantly impact distant/loco-regional relapse or survival. Patients who had a change to their radiotherapy treatment, namely increased hypofractionation, had increased odds of $\geq$ grade 3 acute toxicity. These results are important as they can inform practice in the context of potential future emergency situations requiring a need to reduce hospital attendances. Furthermore, hypofractionated treatments are a more convenient and cheaper alternative to conventional fractionation regimens without significant compromise on tumour control or mortality.

5.5 Funding

This work was supported by NIHR Manchester Biomedical Research Centre, grant number BRC1215-20007, and CRUK via the funding to Cancer Research UK Manchester Centre: [C147/A18083] and [C147/ A25254].

5.6 Supplementary materials

5.6.1 Changes to radiotherapy dose and/or fractionation

Stage I-II NSCLC

Table 5.5: Survival, distant relapse and loco-regional relapse results from the multivariable analysis of patients with stage I-II NSCLC, investigating the effect of having a change to radiotherapy dose and/or fractionation.

	Survival		**Distant relapse**		**Loco-regional relapse**	
	HR (95% CI)	**P value**	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
Change to radiotherapy dose and/or fractionation	0.951 (0.480, 1.88)	0.884	1.09 (0.412, 2.90)	0.859	1.25 (0.609, 2.58)	0.541
Gender (reference male)						
Female	0.999 (0.646, 1.54)	0.996	1.69 (0.957, 2.97)	0.0706	1.21 (0.764, 1.93)	0.413
Age (reference <70 years)						
≥ 70 years	1.01 (0.643, 1.58)	0.970	0.838 (0.467, 1.50)	0.552	1.55 (0.892, 2.68)	0.120
PS (reference 0)						
1	1.14 (0.462, 2.81)	0.779	0.783 (0.331, 1.86)	0.579	0.655 (0.315, 1.36)	0.258
2-3	1.83 (0.748, 4.49)	0.185	0.604 (0.245, 1.49)	0.274	0.643 (0.305, 1.36)	0.247
Chemotherapy	2.16 (0.847, 5.52)	0.107	2.39 (0.757, 7.54)	0.138	1.70 (0.395, 7.29)	0.478
Dose per fraction (reference 2 Gy/fraction)						
>2- 2.9 Gy/fraction	1.15 (0.141, 9.31)	0.898	1.40 (0.154, 12.8)	0.764	0.944 (0.0857, 10.4)	0.963
<2 Gy/fraction	0.437 (0.0266, 7.17)	0.562	1.99 (0.104, 38.2)	0.648	1.49 (0.0677, 32.9)	0.800
3-5.9 Gy/fraction	0.732 (0.0820, 6.54)	0.780	0.804 (0.0803, 8.04)	0.852	0.806 (0.0748, 8.68)	0.859
≥ 6 Gy/fraction	0.660 (0.0808, 540)	0.699	0.374 (0.0394, 3.54)	0.391	0.619 (0.0561, 6.82)	0.695

Abbreviations: *Gy*, Gray; *HR*, Hazard Ratio; *PS*, Performance Status.

Table 5.6: Results from the multivariable analysis of ≥ grade 3 acute toxicity in patients with stage I-II NSCLC, investigating the effect of having a change to radiotherapy dose and/or fractionation.

	aOR (95% CI)	**P value**
Change to radiotherapy dose and/or fractionation	3.46 (1.01, 10.6)	0.0348
Gender (reference male)		
Female	1.99 (0.734, 5.81)	0.185
Age (reference <70 years)		
≥ 70 years	1.04 (0.378, 3.22)	0.941
PS (reference 0)		
1	0.816 (0.183, 5.92)	0.810
2-3	1.08 (0.233, 8.29)	0.929
Chemotherapy	2.26 (0.218, 15.8)	0.440
Dose per fraction (reference 2 Gy/fraction)		
>2- 2.9 Gy/fraction	0.500 (0.0448, 12.7)	0.605
<2 Gy/fraction	3.50e-8 (NA, 3.87e+100)	0.997
3-5.9 Gy/fraction	2.18e-8 (NA, 4.50e+31)	0.990
≥ 6 Gy/fraction	0.300 (0.0249, 8.49)	0.393

Abbreviations: *aOR*, adjusted Odds Ratio; *Gy*, Gray; *PS*, Performance Status.

Table 5.7: Toxicity data for patients with stage I-II NSCLC who received 5 fraction SABR versus 3 fraction SABR.

	5 fraction SABR (N=262)	**3 fraction SABR (N=143)**
Acute toxicity ≥ grade 3, n (%)	6 (2.3)	5 (3.5)
Oesophageal	1 (0.4)	1 (0.7)
Pulmonary	4 (1.5)	2 (1.4)
Cardiac	0	2 (1.4)
Lung infection	1 (0.4)	0
Chest pain	0	0
Other	0	0
Missing	15 (5.7)	6 (4.2)
Late toxicity ≥ grade 3, n (%)	4 (1.5)	3 (2.1)
Oesophageal	0	0
Pulmonary	1 (0.4)	1 (0.7)
Cardiac	0	0
Lung infection	3 (1.1)	1 (0.7)
Chest pain	0	1 (0.7)
Other	0	0
Missing	46 (17.6)	43 (30.1)

Abbreviations: *SABR*, Stereotactic Ablative Radiotherapy.

Stage III NSCLC

Table 5.8: Rates of ≥ grade 3 acute and late toxicity for patients with stage III NSCLC who received concurrent, sequential or no chemotherapy, split by whether they also had a change to their radiotherapy.

	Change to RT		**No change to RT**	
	≥ grade 3 acute toxicity (N=15)	**≥ grade 3 late toxicity (N=4)**	**≥ grade 3 acute toxicity (N=32)**	**≥ grade 3 late toxicity (N=6)**
Concurrent chemotherapy	12 (80.0)	3 (75.0)	13 (40.6)	1 (16.7)
Sequential chemotherapy	2 (13.3)	1 (25.0)	3 (9.4)	1 (16.7)
No chemotherapy	1 (6.7)	0	15 (46.9)	4 (66.7)

Abbreviations: *RT*, Radiotherapy.

Table 5.9: Survival, distant relapse and loco-regional relapse results from the multivariable analysis of patients with stage III NSCLC, investigating the effect of having a change to radiotherapy dose and/or fractionation.

	Survival		**Distant relapse**		**Loco-regional relapse**	
	HR (95% CI)	**P value**	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
Change to radiotherapy dose and/or fractionation	1.08 (0.606, 1.94)	0.789	1.71 (0.731, 4.00)	0.216	0.880 (0.316, 2.45)	0.806
Gender (reference male)						
Female	0.601 (0.392, 0.920)	0.019	0.777 (0.492, 1.23)	0.280	0.617 (0.383, 0.995)	0.0478
Age (reference <70 years)						
≥ 70 years	1.23 (0.833, 1.83)	0.294	1.06 (0.680, 1.66)	0.786	0.829 (0.524, 1.31)	0.425
PS (reference 0)						
1	1.29 (0.710, 2.36)	0.400	1.38 (0.734, 2.58)	0.320	2.07 (0.997, 4.28)	0.0509
2-3	1.67 (0.872, 3.19)	0.122	1.08 (0.513, 2.27)	0.843	1.70 (0.739, 3.93)	0.212
Chemotherapy	0.750 (0.482, 1.17)	0.204	0.662 (0.396, 1.11)	0.116	0.666 (0.391, 1.14)	0.135
Dose per fraction (reference 2 Gy/fraction)						
>2- 2.9 Gy/fraction	1.45 (0.514, 4.10)	0.481	2.14 (0.506, 9.03)	0.302	0.867 (0.329, 2.28)	0.771
<2 Gy/fraction	2.08 (0.494, 8.78)	0.318	1.95 (0.258, 14.8)	0.517	2.15 (0.516, 8.93)	0.293
3-5.9 Gy/fraction	1.28 (0.337, 4.89)	0.713	1.80 (0.249, 13.0)	0.559	2.85 (0.661, 12.3)	0.160
≥ 6 Gy/fraction	NA	NA	NA	NA	NA	NA

Abbreviations: *Gy*, Gray; *HR*, Hazard Ratio; *PS*, Performance Status.

Table 5.10: Results from the multivariable analysis of ≥ grade 3 acute toxicity in patients with stage III NSCLC, investigating the effect of having a change to radiotherapy dose and/or fractionation.

	aOR (95% CI)	**P value**
Change to radiotherapy dose and/or fractionation	4.78 (2.11, 10.7)	0.000148
Gender (reference male)		
Female	1.15 (0.589, 2.20)	0.685
Age (reference <70 years)		0.848
≥ 70 years	1.07 (0.549, 1.06)	
PS (reference 0)		
1	1.19 (0.512, 3.06)	0.693
2-3	1.60 (0.561, 4.82)	0.385
Chemotherapy	1.83 (0.871, 3.96)	0.115
Dose per fraction (reference 2 Gy/fraction)		
>2- 2.9 Gy/fraction	0.479 (0.177, 1.45)	0.163
<2 Gy/fraction	2.37 (0.385, 12.9)	0.324
3-5.9 Gy/fraction	0.0873 (0.0101, 0.533)	0.0127
≥ 6 Gy/fraction	NA	NA

Abbreviations: *aOR*, adjusted Odds Ratio; *Gy*, Gray; *PS*, Performance Status.

SCLC

Table 5.11: Follow-up data for patients with SCLC split by whether they had a change to their radiotherapy dose and/or fractionation or not.

	Change to RT (N=34)	**No change to RT (N=75)**
Acute toxicity ≥ grade 3, n (%)	2 (5.9)	9 (12.0)
Oesophageal	2 (5.9)	3 (4.0)
Pulmonary	0	0
Cardiac	0	1 (1.3)
Lung infection	0	0
Chest pain	0	0
Other	0	5 (6.7)
Missing	1 (2.9)	3 (4.0)
Late toxicity ≥ grade 3, n (%)	2 (5.9)	0
Oesophageal	1 (2.9)	0
Pulmonary	1 (2.9)	0
Cardiac	0	0
Lung infection	0	0
Chest pain	0	0
Other	0	0
Missing	4 (11.8)	13 (17.3)
Disease status, n (%)		
Distant relapse	12 (35.3)	20 (26.7)
Loco-regional relapse	7 (20.6)	11 (14.7)
No evidence recurrence	17 (50.0)	38 (50.7)
Death, n (%)	12 (35.3)	20 (26.7)

Abbreviations: *RT*, Radiotherapy.

5.6.2 Changes to chemotherapy regimen

Stage III NSCLC

Table 5.12: Survival, distant relapse and loco-regional relapse results from the multivariable analysis of patients with stage III NSCLC who were considered for chemotherapy, investigating the effect of having chemotherapy omitted or the dose/number of cycles reduced.

	Survival		**Distant relapse**		**Loco-regional relapse**	
	HR (95% CI)	**P value**	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
Change to chemotherapy regimen (reference no change)						
Chemotherapy omitted	1.81 (0.961, 3.40)	0.0661	1.85 (0.923, 3.71)	0.0827	1.03 (0.468, 2.27)	0.940
Chemotherapy dose/number of cycles reduced	0.733 (0.273, 1.97)	0.537	1.26 (0.451, 3.49)	0.664	0.387 (0.0883, 1.70)	0.208
Gender (reference male)						
Female	0.591 (0.334, 1.05)	0.0712	0.985 (0.539, 1.80)	0.960	0.695 (0.366, 1.32)	0.267
Age (reference <70 years)						
≥ 70 years	1.34 (0.771, 2.33)	0.298	0.878 (0.455, 1.69)	0.698	0.717 (0.347, 1.48)	0.367
PS (reference 0)						
1	1.04 (0.497, 2.17)	0.922	1.46 (0.652, 3.25)	0.360	2.98 (1.03, 8.60)	0.0431
2-3	2.11 (0.921, 4.85)	0.0773	2.58 (0.919, 7.26)	0.0718	3.59 (0.951, 13.5)	0.0594
Change to radiotherapy dose and/or fractionation	1.15 (0.524, 2.54)	0.725	1.78 (0.669, 4.71)	0.249	0.794 (0.196, 3.21)	0.747
Dose per fraction (reference 2 Gy/fraction)						
> 2- 2.9 Gy/fraction	1.31 (0.453, 3.80)	0.617	1.81 (0.420, 7.75)	0.427	0.820 (0.305, 2.21)	0.695
<2 Gy/fraction	8.90 (1.51, 52.4)	0.0156	1.33e-5 (0, Inf)	0.998	4.14e-6 (0, Inf)	0.998
3-5.9 Gy/fraction	1.87 (0.333, 10.5)	0.478	1.28 (0.0902, 18.0)	0.857	5.69 (0.720, 45.0)	0.0992
≥ 6 Gy/fraction	NA	NA	NA	NA	NA	NA

Abbreviations: *Gy*, Gray; *HR*, Hazard Ratio; *PS*, Performance Status.

SCLC

Table 5.13: Disease status for patients with SCLC split by whether they had their chemotherapy omitted, reduced, or received standard of care chemotherapy i.e. no change to chemotherapy regimen.

	Chemotherapy omitted (N=7)	**Chemotherapy dose/number of cycles reduced (N=14)**	**No change to chemotherapy (N=83)**
Disease status, n (%)			
Distant relapse	3 (42.9)	7 (50.0)	22 (26.5)
Loco-regional relapse	4 (57.1)	3 (21.4)	9 (10.8)
No evidence recurrence	3 (42.9)	6 (42.9)	43 (51.8)
Death, n (%)	1 (14.3)	5 (35.7)	24 (28.9)

Abbreviations: *RT*, Radiotherapy.

Chapter 6

Understanding the Differences Between Bayesian and Frequentist Statistics

This chapter has been published in the International Journal of Radiation Oncology-Biology-Physics 2022 Volume 112 Issue 5 p1076-1082. The section numbering, references and formatting have been modified to be consistent with the rest of the thesis.

Chapters 6 and 7 were originally one study, and then were split into two for publication. The work in this chapter highlights the differences between frequentist and Bayesian statistics using a simulated dataset and is therefore not novel; however, it has been included to aid clinical interpretation of the methods in Chapter 7.

Authors

Isabella Fornacon-Wood[1], Hitesh Mistry[1], Corinne Johnson-Hart[2], Corinne Faivre-Finn[1,3], James P B O'Connor[1,4] and Gareth J Price[1]

Affiliations

[1] Division of Cancer Sciences, University of Manchester, Manchester, UK.

[2] Department of Medical Physics, The Christie NHS Foundation Trust, Manchester, UK.

[3] Department of Radiation Oncology, The Christie NHS Foundation Trust, Manchester, UK.

[4] Department of Radiology, The Christie NHS Foundation Trust, Manchester, UK.

Author contributions

I modified an R script written by G.J.P to simulate the dataset used in the analysis. I wrote the R code to analyse the data using frequentist and Bayesian methods. I wrote the manuscript, which was reviewed by all co-authors.

Highlights

- There are 2 main approaches to statistical inference, frequentist and Bayesian, differing in their interpretation of uncertainty.
- The frequentist approach deals with long-run probabilities (ie, how probable is this data set given the null hypothesis), whereas the Bayesian approach deals with the probability of a hypothesis given a particular data set.
- Bayesian analysis incorporates prior information into the analysis, whereas a frequentist analysis is purely driven by the data.
- The Bayesian approach can calculate the probability that a particular hypothesis is true, whereas the frequentist approach calculates the probability of obtaining another data set at least as extreme as the one collected (giving the P value).
- Interpretation of results is more intuitive with a Bayesian approach compared with the frequentist approach, which can often be misinterpreted.

6.1 Case Vignette

Changes to radiation therapy workflows happen continuously as technology and techniques are optimized. One may wonder what the clinical outcomes of such changes are, or if there is any effect at all, because formal evaluation through a clinical trial rarely takes place. An example might be the modification of treatment protocols such as those used in image guided radiation therapy (IGRT). Suppose a large cancer center uses an IGRT action threshold during lung cancer patient setup. Patients whose position in their daily cone beam computed tomography (CBCT) is more than a threshold distance from that in their radiation therapy planning CT scan have their position corrected before treatment, whereas those with smaller offsets are treated without correction. The center updates its protocol to reduce the action threshold so that smaller setup errors are corrected. Published results suggest this change may have an impact on patient survival, and the team wants to determine whether this is the case in their center. Statisticians are tasked with identifying the appropriate statistical methodologies for such an evaluation as well as considering whether the approach

could be embedded into routine practice to monitor the impact of future changes in clinical management.

6.2 Introduction

Just like Liverpool versus Manchester United, the Yankees versus the Red Sox, or Coke versus Pepsi, there are 2 main schools of statistics, and you may have heard the proponents of each noisily arguing their respective benefits. The first is the frequentist approach, which dominates the medical literature and consists of null hypothesis significance testing (think P values and confidence intervals). The other is the Bayesian approach, governed by Bayes' theorem. The fundamental difference between these 2 schools is their interpretation of uncertainty and probability [259]: the frequentist approach assigns probabilities to data, not to hypotheses, whereas the Bayesian approach assigns probabilities to hypotheses. Furthermore, Bayesian models incorporate prior knowledge into the analysis, updating hypotheses probabilities as more data become available. The goal of this article is to educate readers about the differences between frequentist and Bayesian inference, discuss potential advantages or disadvantages of each approach, and use the case vignette to highlight how these 2 methods may be implemented in a real-world example in a radiation therapy clinic.

6.3 Introduction to Frequentist Statistics

Frequentist statistics is all about probability in the long run; the data set collected and analyzed is one of many hypothetical data sets addressing the same question, and uncertainty is due to sampling error alone. For example, the probability of getting heads when flipping a coin in the long run is 0.5; if we flip the coin many times, we would expect to see heads 50% of the time, whereas if we had flipped the coin only a few times we could reasonably expect to observe a different distribution (eg, all heads) just by chance.

Frequentist inference begins by assuming a null hypothesis to be true before data are collected (eg, that there is no effect of a particular treatment on survival). Investigators then collect data, analyze them, and ask, "How surprising is my result if there is

actually no effect of the treatment on survival?" The data would be surprising if there was a low probability by chance alone of obtaining another data set at least as extreme (ie, far away from the null hypothesis and unlikely to occur by chance) than that collected (eg, showing a large difference in survival between patients having different treatments when our null hypothesis states there is no difference). If this is the case, then the collected data are considered unlikely under the null hypothesis and we can reject it, inferring that the null hypothesis does not adequately explain our data and that something else (eg, the new treatment) must account for our results.

This probability of obtaining another data set as extreme as the one collected is known as the P value. The P value is often criticized for being misunderstood and misused in the field of medicine [260, 261]. For example, in contrast to popular belief, the P value is not a measure of how correct a hypothesis is, nor is it a measure of the size or importance of an effect [261]. In particular, large P values do not provide evidence of no effect [262]. A P value is simply the probability of obtaining another data set at least as extreme as the one collected by chance alone.

For example, suppose we run an analysis and get a highly significant P value of .001 for our treatment variable—how do we interpret this? Formally, there is a 0.1% chance of collecting data equal to or more extreme than this result if the null hypothesis were true—it would be surprising to collect these data if there is, in fact, no effect of treatment on survival. If we had run the analysis and got a P value of .46, then there is a 46% chance of collecting data equal to or more extreme than this if the null hypothesis is true—it would not be surprising to obtain these results if there is no effect of treatment on survival. This result is not evidence of no effect, however, because the inference began by assuming there would be no effect of treatment. The key here is that probabilistic statements (ie, P values) can only be made about the data, not about hypotheses or parameters (ie, the treatment effect) [263].

Reporting confidence intervals can improve the interpretation of results compared with a P value alone and can give information on the size and direction of an effect [264]. A 95% confidence interval tells us that if we were to repeat the experiment over and over (remember, frequentist statistics are long run), 95% of the computed confidence

intervals would contain the true mean [265]. This is different than saying there is 95% chance the true mean lies within the interval, because frequentist statistics cannot assign probabilities to parameters—the true mean either lies within the interval or it does not [266].

6.4 Introduction to Bayesian Statistics

Bayesian statistics are named after the Reverend Thomas Bayes, whose theorem describes a method to update probabilities based on data and past knowledge. In contrast to the frequentist approach, parameters and hypotheses are seen as probability distributions and the data as fixed. This idea is perhaps more intuitive because generally the data we collect are the only data set we have, so it does not necessarily make sense to perform statistical analysis assuming it is one of many potential data sets. Probability distributions summarize the current state of knowledge about a parameter or hypothesis and can be updated as more data becomes available using Bayes' theorem, presented in Equation (6.1).

$$p(\theta|Data) = \frac{p(Data|\theta) \cdot p(\theta)}{p(Data)} \tag{6.1}$$

The probability distribution that summarizes what is known about an outcome before a test or piece of information is obtained is known as the prior distribution, often just dubbed "the prior." Such an outcome may be the prevalence of disease or a specific diagnosis. The prior probability of the outcome θ (eg, of having the disease) is labeled $P(\theta)$ in Equation (6.1). The prior is one of the key differences between frequentist and Bayesian inference; frequentist analyses base their results only on the data they collect. The prior could be formulated by expert beliefs, historical data, or a combination of the two.

Consider now that one has a positive test for the disease. What is the probability of θ (having the disease), given these new data (ie, that the test was positive)? The notation for this scenario is $P(\theta|test_+)$, where the "|" can be translated as "given" and we set the Data variable in Equation (6.1) to $test_+$. In Bayesian terminology,

this probability is known as the posterior distribution (or post test distribution) and summarizes what is known about the outcome using both the prior information and the new data.

Among all potential patients, regardless of whether they have the disease, the probability of a positive test (ie, true positive plus false positive) is $P(test_+)$. Therefore, the relative probability of having a positive test if a patient is truly positive versus the probability of someone randomly selected from the population (who may or may not have the disease) having a positive test is the ratio $P(test_+|\theta)/P(test_+)$. The posterior probability of having the disease if you have a positive test is then the baseline prevalence of the disease in the population (the prior probability) multiplied by this factor.

We can further illustrate how Bayes' theorem and the use of prior information can help to answer important questions using the example of COVID-19 testing. For this example, let's assume that a test for COVID-19 infection is guaranteed (100% chance) to detect the COVID-19 virus in someone who has the infection ($P[test_+|virus]$=1.0) and has a 99.9% chance of correctly identifying that someone does not have the virus (or a 0.1% chance of a false positive: $P[test_-|no\ virus]$=0.999 and $P[test_+|no\ virus]$=0.001) [267]. That sounds like a high probability, but what we are really interested in is if you have a positive test, how likely is it that you actually have the virus ($P[virus|test_+]$). We can calculate this probability using Bayes' theorem. As discussed, this probability depends on some prior information—how likely you were to have the virus (ie, $P[virus]$) before taking the test.

$$p(virus|test_+) = \frac{p(test_+|virus) \cdot p(virus)}{p(test_+)} \tag{6.2}$$

If we assume the prior probability is the prevalence of the virus in the population at the height of the pandemic, 2%, then $P(virus)$=0.02. Out of 1 million people, 20,000 will have the virus and 980,000 will not. If we test them all, 20,000 will have a true positive test, 979,020 will have a true negative test, and 980 will have a false positive test (980,000 x 0.001). Thus, there would be 20,980 positive tests in total ($P[test_+]$=20,980/1,000,000=0.02098), and the chance of having the virus

given a positive test could be calculated using Equation (6.2): $P(virus|test_+)$=(1.0 x 0.02)/0.02098=0.953, which can be approximated as 95%. In this setting, if the test is positive, one should believe the test.

However, what if the prevalence of COVID-19 was estimated to be much less, say, 0.2% ($P[virus]$=0.002), such as during a period of a governmental stay-at-home order? In this scenario, when we test 1 million people, 2000 will have a true positive test, 997,002 will have a true negative test, and 998 will have a false positive test (998,000 x 0.001), giving 2998 positive test results, and $P(test_+)$=2988/1,000,000=0.002988. The probability of having the virus after a positive test now becomes $P(virus|test_+)$=(1.0 x 0.002)/0.002988=0.669. In this scenario, one may truly question whether a positive test is diagnostic of infection, because the likelihood of a false positive is approximately 1 in 3.

Therefore, knowledge of the prior information (in this case, the prevalence of COVID-19 in a population) can alter the chance of having a virus when a test is positive from 95% to 67% when the sensitivity and specificity of the test remain unaltered. On the other hand, if we were testing hospitalized patients, for example, the prevalence would be expected to be much higher, and there would be a lower chance of obtaining a false positive result. This simple calculation highlights the importance of taking into account prior information, something a frequentist analysis does not do.

Sometimes formulating a prior is not that easy or clear. In such cases, the use of priors can be seen as a drawback to Bayesian inference, particularly when results depend on the chosen prior, and thus the analysis could be manipulated to get a positive result. In such cases, an "uninformative" prior that provides no additional information could be used, or multiple priors (eg, with either optimistic or skeptical assumptions) could be tested to determine the sensitivity of the results to particular priors. As with all analyses, it is vital that researchers are transparent in their methods and assumptions.

The posterior distribution captures our "updated" estimate of the probability of the outcome after incorporating new data, including our uncertainties, and can be analyzed to give various statistics, such as the mean and 95% credible interval. The 95%

credible interval is different from a frequentist 95% confidence interval; it is the parameter range that has a 95% probability of including the true parameter value. The frequentist confidence interval is often misinterpreted in this way; however, one must remember that the 95% confidence interval assumes that the experiment is hypothetically repeated over and over and that 95% of the computed confidence intervals would contain the true mean. Perhaps more importantly, and one of the big advantages of Bayesian inference, is that the posterior distribution can also be used to directly calculate the probability of different hypotheses (eg, that one treatment is superior to another or that survival is improved by at least 3 months).

In this respect, Bayesian inference is more intuitive at its core and in closer alignment with our natural mode of probabilistic reasoning than frequentist inference. For example, we are more interested in the probability that 1 treatment is superior to another (Bayesian probability) than in the probability of obtaining certain data assuming the treatments are equal (frequentist null hypothesis). This advantage in interpretability remains even if our analysis uses an uninformative prior.

6.5 Case Study in Radiation Therapy

Let's look now at the real-world example in radiation therapy from our case vignette and compare the use of a frequentist and Bayesian approach to evaluating the clinical impact of a change in practice. In image guided radiation therapy (IGRT), an action threshold is often used as a decision threshold. At each daily fraction, the patient is set up on the radiation therapy treatment couch. They are then imaged using a CBCT system and their position is compared with the ideal position in the treatment plan. If the offset between the daily CBCT image and the planning CT image is greater than the action threshold, the couch is moved to better align the patient's position to that in the plan. If the offset is less than the threshold, the patient's setup is considered accurate enough and they receive their daily treatment without any shifts, the assumption being that setup errors less than the action threshold will not alter the clinical target volume dose owing to the planning target volume margin or change organ-at-risk doses enough to have a clinical impact. Previous analyses of patients with lung cancer treated with IGRT have shown that the residual setup errors that remain

after the action threshold has been applied (ie, positional errors that are less than the threshold and are considered acceptable enough to treat with) are associated with survival [268]. Patients with average residual setup errors that pushed the radiation therapy dose toward the heart were found to have worse survival than those who had average residual errors moving the dose away from the heart. In this previous study, the action threshold was 5 mm.

Let us assume that after this finding, the department decided to reduce the action threshold from 5 mm to 2 mm in the hope of ameliorating the effect. A year after implementing the change in protocol (post-protocol change), a physician wants to know the impact on clinical outcome. The hypothesis is that the effect of the average residual setup error direction (toward vs away from the heart) on patient survival will be decreased after reducing the action threshold from 5 mm to 2 mm (ie, the hazard ratio [HR] between the 2 directions will have decreased), because only trivial offsets would remain with the new policy.

We have 2 simulated data sets: 1 before protocol change, where residual setup errors range from –5 mm to 5 mm, as in the original study, and 1 after protocol change, with errors from –2 mm to 2 mm. The direction of the average residual setup error is calculated and dichotomized as either closer to or farther from the heart. We want to know if the difference in survival between patients with average residual setup errors toward versus away from the heart observed in the pre-change patients (5 mm action threshold) is reduced with the new 2 mm protocol (ie, the HR of death between toward and away patients is reduced). We fit survival models to the pre- and post-protocol change data separately and look at how the HR changes, adjusting for the clinical variables accounted for in the cited study [268]: performance status, age, prescribed radiation therapy fractions, and tumor volume. The hypothesis is that the HR should be closer to 1 in the post-protocol change data, because a smaller action threshold should lead to smaller residual setup errors, and hence, less difference in the magnitude of heart irradiation and subsequent toxicity.

First, we take a frequentist approach. We fit a survival model to the pre-protocol change data and post-protocol change data separately. The null hypothesis is that

there is no effect of setup error on survival, and our model will indicate how surprising the data we collect are if this is true. Before protocol change, we have an HR of 1.26 (95% confidence interval, 1.05-1.48), and P=0.013; that is, the risk of death is 26% (with a confidence interval of 5%-48%) higher in the group of patients with the average setup error direction toward the heart compared with those with the average setup error direction away from the heart. If there were actually no effect of setup error on survival, it would be surprising to see an HR this extreme (ie, this far from 1.0). After the protocol change, we have an HR of 1.07 (95% confidence interval, 0.91-1.27), and P=0.41. This P value is greater than the widely used 5% threshold, so seeing these results is not particularly unexpected if there were no difference in outcome associated with setup error. However, remember that this is not evidence of no effect, so we cannot conclude that the effect of residual setup error on survival is eliminated. Indeed, if we look at the 95% confidence interval, the HR of the post-protocol change could be as high as 1.27 (ie, a 27% increase in the risk of death). If we had a very tight confidence interval around the null (HR=1), we would be more confident that there was no effect, but with such a wide confidence interval, the result is not informative.

Now we can analyze the data using Bayesian statistics. We again fit a survival model to the post-protocol change data, but this time, we specify a prior. To assess the sensitivity of the results to the prior, we choose 3 priors - an uninformative, a skeptical, and an enthusiastic one. The uninformative prior lets the post-protocol change data alone drive the inference, like in a frequentist setting. The skeptical prior (ie, skeptical that the change reduces the impact of the setup error) uses the data about the impact of the residual setup errors on survival from before the new protocol was changed (the pre-protocol change patient cohort). The use of such historic data is a common scenario as it allows us to include existing observations (eg, from previous clinical trials) in our analysis and, for example, indicate how confident we are in the quality of the evidence (for observational data we might include greater uncertainty in the prior distribution). Where we are investigating a change in practice, it also allows us to be cautious in our assessment of the intervention by starting from the assumption that patients will experience the historically observed outcomes. The enthusiastic prior (ie,

enthusiastic that there will no longer be an impact of setup error) assumes that there will be no survival difference between patients with residual setup errors toward versus away from the heart before the model sees the post-protocol change data. Enthusiastic priors are typically used for sensitivity analyses to evaluate how strongly the choice of the prior influences the analysis result. Given enough data, both the enthusiastic and skeptical priors would eventually evolve to the same posterior distribution (conversely, experiments with few data are unlikely to move the prior by a large amount).

Figure 6.1 presents the probability distributions for the HR before and after protocol change for the different priors, calculated via the Bayes theorem (Equation (6.1)). We know the prior distribution ($P[\theta]$, or the belief about the distribution of the HR of death between patients with average setup errors toward and away from the heart) and obtain the distribution of the likelihood function ($P[Data|\theta]$, the probability of observing our survival data for a range of possible HR values) from our survival model. This information allows us to calculate the numerator of the Bayes equation, but calculating the denominator, $P(Data)$, is often very difficult for nontrivial (ie, real-life) situations, meaning it is often not possible to directly calculate the posterior distribution. Fortunately, we can instead use a computer to numerically sample a close approximation to it, typically using a technique called Markov chain Monte Carlo (MCMC) sampling. Briefly, MCMC iteratively samples different positions (ie, parameter values) in a hypothetical posterior distribution, using a set of rules to decide how to move from 1 position to the next (the rules are the Markov chain). At each new position we calculate the ratio of the new posterior probability to that in the previous position based on the obtained data. As Equation (6.3) shows, taking the ratio of the probability at the 2 positions means that we cancel out of the part of the Bayes equation that we have difficulty calculating, meaning we can easily calculate the ratio for each successive position. If the ratio is greater than 1 (ie, the new position is more probable given the data), we accept the move, and if it is less than 1 (ie, the previous position was more probable given the data), we calculate a random number and accept the move if the ratio is larger than the random number, rejecting it otherwise. By repeating this process many times and keeping a record of the different accepted sample positions (the record is known as the MCMC trace), we obtain an increasingly

accurate estimation of the posterior distribution as the histogram of how often each parameter position "wins" relative to the competing value. Most commonly used analytical software languages contain MCMC sampling libraries to calculate posterior distributions using more sophisticated implementations of this general approach [269, 270].

$$\frac{p(\theta_{new}|Data)}{p(\theta_{previous}|Data)} = \frac{\frac{p(Data|\theta_{new})\cdot p(\theta_{new})}{p(Data)}}{\frac{p(Data|\theta_{previous})\cdot p(\theta_{previous})}{p(Data)}} = \frac{p(Data|\theta_{new})\cdot p(\theta_{new})}{p(Data|\theta_{previous})\cdot p(\theta_{previous})} \quad (6.3)$$

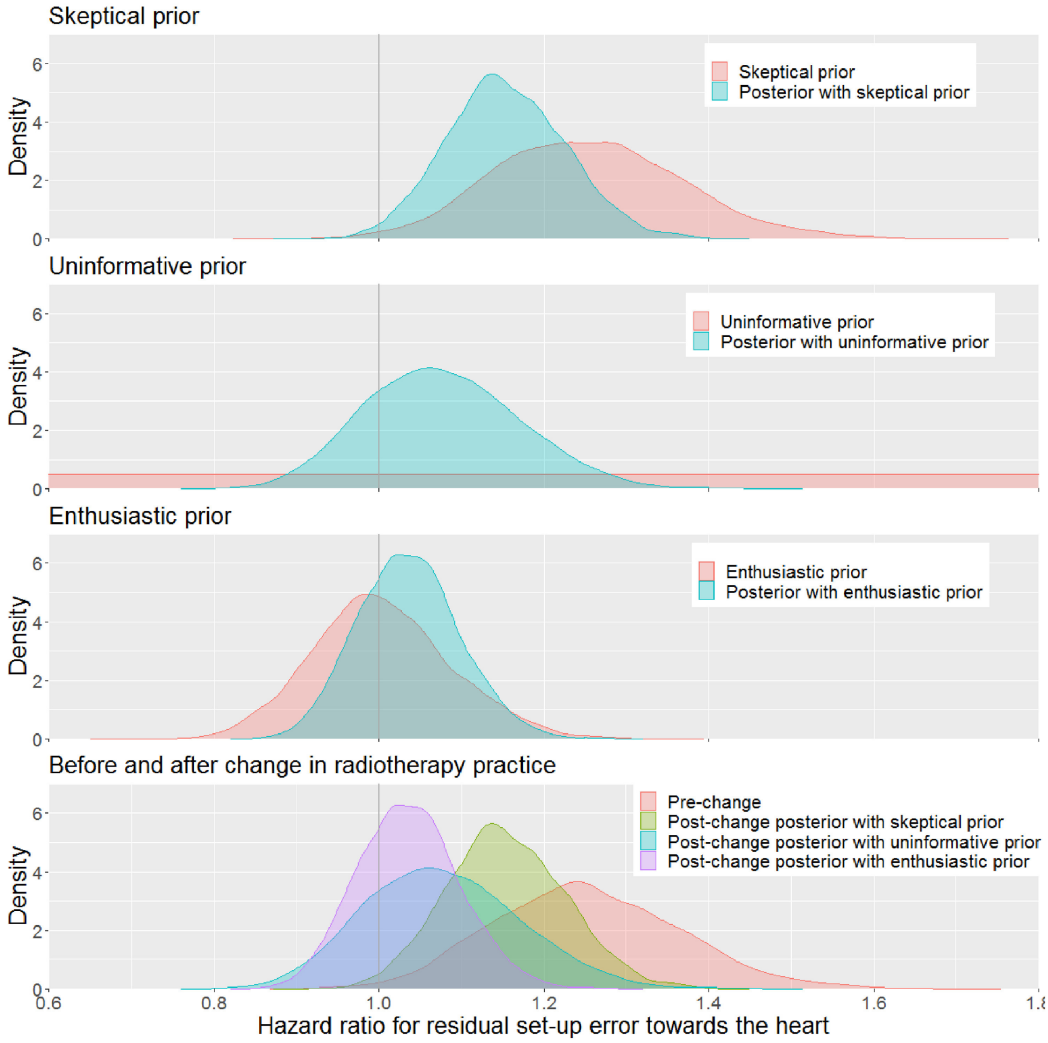

Figure 6.1: Posterior distributions of the hazard ratio (HR) for residual setup error direction toward the heart. The red distribution in the top 3 plots shows the skeptical, uninformative, and enthusiastic priors used to calculate the HR distributions after protocol change, with the corresponding posteriors shown in blue. The bottom plot shows all 3 posterior HR distributions together with the HR distribution from before the protocol change (the pre-change cohort).

We can see that the post-protocol change HRs shown in Figure 6.1 are different depending on which prior was chosen; the posterior calculated with the skeptical prior has the largest HR and is closest to the pre-protocol change HR distribution, the posterior with the enthusiastic prior has the HR distribution closest to the null (HR=1), and the posterior using the uninformative prior has an HR somewhere in the middle. This result shows us how choice of prior influences inference in the Bayesian setting. With the skeptical prior, there is a post-protocol change HR=1.15 (credible interval, 1.02-1.3). With the uninformative prior, it is slightly less (1.08 [0.91-1.27]), and with

the enthusiastic prior, it is even less (1.04 [0.91-1.16]). Another key advantage of the Bayesian approach is that the HR point estimate and 95% credible intervals are not the only information that can be obtained from the posterior distributions; we can also directly quantify the evidence for multiple hypotheses using probabilities, examples of which are summarized in Table 6.1.

Table 6.1: Bayesian probabilities for the skeptical, uninformative, and enthusiastic priors.

	Skeptical prior	**Uninformative prior**	**Enthusiastic prior**
P(HR reduced)	0.760	0.882	0.961
P(HR>1)	0.988	0.786	0.712

Abbreviations: *HR*, hazard ratio; *P(HR reduced)*, probability that the HR for patients with average residual setup error direction toward the heart (compared with those with average residual setup error directions away from the heart) was reduced after the protocol change; *P(HR>1)*, probability that the HR for the average residual setup error direction toward the heart was greater than 1 after protocol change.

Interpretation of the results is different depending on the prior used. For the skeptical prior, there is a high probability that an effect of residual setup error direction toward the heart exists after protocol change, because we have a high probability (0.988) of an HR>1. The probability that the HR is reduced after protocol change is lower, at 0.760. With the uninformative prior, there is a moderate probability both that an HR>1 exists after protocol change and that the HR is reduced. With the enthusiastic prior, there is a high probability that the HR is reduced after protocol change and a lower probability (0.712) that an HR>1 exists.

So which prior and interpretation do we trust? That depends on our beliefs prior to the analysis and can be subjective. Presenting how sensitive results are to the choice of prior is therefore important so readers can fully understand how the prior affects the results and how this in turn affects the study's interpretation. For example, the results in Table 6.1 show that even when using an enthusiastic prior that assumed changing the IGRT protocol would eliminate the increased risk of death in patients whose setup errors move the heart toward the high dose region, there was a reasonably high probability that a residual effect remained. On the other hand, even when using a skeptical prior that assumed there would be no change in the HR after the protocol

change, there was a reasonably high probability that the HR was reduced. As such, we can be confident that this finding (ie, the HR was decreased by the protocol change, but a reduced effect still remained) is a real effect, whereas the frequentist analysis does not allow us to make this inference and would be at risk of researchers reaching the opposite conclusion if only looking at the P value.

Use of an uninformative prior lets the data alone drive the inference and gives results similar to a frequentist analysis, but a Bayesian analysis allows one to directly calculate the probabilities for potential hypotheses. Furthermore, Bayesian posteriors allow us to calculate the probability of many different hypotheses (eg, that the HR is >1, that the HR is >1.1, that the effect is reduced by 20%). If we wanted to do this in the frequentist setting, we would need to separately assess different hypotheses and consider if multiplicity corrections were required. Similarly, the ease with which Bayesian analyses can accommodate prior information also means that it is straightforward to collect more data after an analysis has taken place if results are inconclusive.

Unlike the somewhat rigid orthodoxy that has developed around the interpretation of frequentist analyses (eg, P-value thresholds), Bayesian probabilities need to be placed into context to aid subsequent decision-making. The probability of different (competing) hypotheses can be directly cited as evidence and used to inform clinical decisions (in our example, if we consider that there is a strong probability that a residual setup error effect remains, we might then want to reduce the action threshold further). Formal frameworks for decision-making based on Bayesian probabilities have also been developed that are more akin to the hypothesis-testing approaches used in the frequentist setting. Bayes factors, for example, assess the ratio of the likelihood of 1 hypothesis over an alternative hypothesis given the observed data. The higher the Bayes factor, the more likely that 1 hypothesis (in the numerator) is correct. Standardized scales of this factor have been developed to decide which hypothesis is the most compelling. A more detailed explanation of Bayes factors is given by both Goodman [271] and Schonbrodt and Wagenmakers [272].

6.6 Conclusions

Both the frequentist and Bayesian approaches are useful for data analysis as long as they are interpreted correctly. The strength of the Bayesian approach is the incorporation of prior information and the ability to directly calculate the probability of different hypotheses from the posterior distribution. It is more in line with our natural mode of reasoning; if we are sure in our belief of something (eg, we have a strong prior formed from evidence of a large effect from a phase 3 clinical trial), data will have to be highly convincing to alter this belief, whereas if we are unsure either way (uninformative prior), inference is driven more by the data than by the prior, and our beliefs are more easily overturned. In contrast, the frequentist approach is driven by the data only, so there is no issue of subjectivity owing to the prior. Although on one hand, this means one cannot manipulate priors to get a particular result in a frequentist analysis, the findings can be manipulated in other ways, such as P-value fishing, and the interpretation of the results is less intuitive and can easily lead to misinterpretation. However you choose to analyze your data, it is of utmost importance to be transparent in the methodology and to correctly interpret the results in a manner consistent with the underlying statistical approach and any limitations in how the data were collected (eg, in a prospective randomized trial or from a retrospective observational cohort). Proper inference, and thus the clinical impact of one's analysis, depends critically on these principles.

6.7 Dos and don'ts

- Do define the data available and the clinical question and then select the most appropriate statistical analysis.
- Do investigate the influence of your prior on the result when performing a Bayesian analysis.
- Do ensure your interpretation of either a frequentist or a Bayesian analysis is correct, with particular attention to P values, confidence intervals, and credible intervals.

- Do not assume a Bayesian analysis will solve the problem of insufficient or poor-quality (ie, incorrectly recorded) data; the most important part of an analysis is the quality of the data.

Chapter 7

Bayesian methods provide a practical real-world evidence framework for evaluating the impact of changes in radiotherapy

This chapter has been published in Radiotherapy and Oncology 2022 Volume 176 p53-58. The section numbering, references and formatting have been modified to be consistent with the rest of the thesis.

Authors

Isabella Fornacon-Wood[1], Hitesh Mistry[1], Corinne Johnson-Hart[2], Corinne Faivre-Finn[1,3], James P B O'Connor[1,4] and Gareth J Price[1]

Affiliations

[1] Division of Cancer Sciences, University of Manchester, Manchester, UK.

[2] Department of Medical Physics, The Christie NHS Foundation Trust, Manchester, UK.

[3] Department of Radiation Oncology, The Christie NHS Foundation Trust, Manchester, UK.

[4] Department of Radiology, The Christie NHS Foundation Trust, Manchester, UK.

Author contributions

I adapted the PostgreSQL scripts written by C.J.H. to collate the post-protocol change clinical cohort. I wrote the R script for the Bayesian survival analysis. I wrote the manuscript, which was reviewed by all co-authors.

Abstract

Purpose

Retrospective studies have identified a link between the average set-up error of lung cancer patients treated with image-guided radiotherapy (IGRT) and survival. The IGRT protocol was subsequently changed to reduce the action threshold. In this study, we use a Bayesian approach to evaluate the clinical impact of this change to practice using routine 'real-world' patient data.

Methods and materials

Two cohorts of NSCLC patients treated with IGRT were compared: pre-protocol change (N=780, 5 mm action threshold) and post-protocol change (N=411, 2 mm action threshold). Survival models were fitted to each cohort and changes in the hazard ratios (HR) associated with residual set-up errors was assessed. The influence of using an uninformative and a skeptical prior in the model was investigated.

Results

Following the reduction of the action threshold, the HR for residual set-up error towards the heart was reduced by up to 10%. Median patient survival increased for patients with set-up errors towards the heart, and remained similar for patients with set-up errors away from the heart. Depending on the prior used, a residual hazard ratio may remain.

Conclusion

Our analysis found a reduced hazard of death and increased survival for patients with residual set-up errors towards versus away from the heart post-protocol change. This study demonstrates the value of a Bayesian approach in the assessment of technical changes in radiotherapy practice and supports the consideration of adopting this approach in further prospective evaluations of changes to clinical practice.

7.1 Introduction

Radiotherapy plays a key role in the treatment of cancer. In particular, radiotherapy is indicated as a treatment option in more than 60% of lung cancer patients during

the course of their management [273]. Radiotherapy has a rich history of technological innovation [230], with much of this transformation occurring through rapid successive changes to the techniques and technologies at each workflow stage. Randomized controlled trials (RCTs) are frequently used to evaluate well defined changes in radiotherapy such as large changes in dose and fractionation [274], or the addition of new systemic treatments [275]. However, conventional RCTs are not well suited to the evaluation of incremental technical changes [20, 276]. Not only may such changes evolve further during trial recruitment [277, 278], but also there is often an implicit assumption that advances will be associated with clinical benefit, making it difficult to argue the equipoise needed for randomisation [17]. As a result, incremental changes in technique are often adopted within radiotherapy departments without formal evaluation.

Real-world data can be defined as observational data that is collected electronically as a part of patients' routine care. It offers an opportunity to provide evidence in patient populations well-known to be under-represented in conventional medical research. The potential of such data has been recognized by the UK National Institute for Health and Care Excellence (NICE) [65] and US Food and Drugs Administration (FDA) [40] who have developed real-world evidence frameworks.

Previous work by Johnson-Hart et al. [268] found that the average residual set-up error following Image Guided Radiotherapy (IGRT) position correction was related to survival in cohorts of patients with lung cancer and esophageal cancer. Patients who had an average residual set-up error moving the radiotherapy dose towards the heart had worse survival than those with set-up errors moving the dose away from the heart. The action threshold (the discrepancy allowed between planning position and treatment day position) was subsequently reduced from 5 mm to 2 mm at The Christie NHS Foundation Trust.

The aim of this paper is to use a Bayesian approach [77] with real-world data to evaluate the clinical impact of this change in IGRT protocol and investigate the effect of incorporating prior information into the analysis.

7.2 Methods and materials

Two anonymized cohorts of NSCLC patients treated with each action threshold were retrospectively collected: i) Pre-protocol change: 780 patients treated before November 2016 (action threshold 5 mm); and ii) Post-protocol change: 411 patients treated between November 2016 and March 2020 (action threshold 2 mm). All data analysis was performed following institutional board approval and was compliant with UK research governance (ref. 17/NW/0060).

The data preparation steps for each cohort have been previously described in full by Johnson-Hart et al. [268] and are summarized in Supplementary Materials. Age, performance status, prescribed radiotherapy fractions and gross tumor volume (GTV) were collected for each cohort with missing data imputed using a random forest method (R library randomForestSRC v2.9.2). We investigated three clinical research questions:

1. Did the introduction of a reduced action threshold reduce the HR of death for patients with average residual set-up errors towards versus away from the heart?

2. Does a residual HR of death remain for patients with average set-up error towards versus away from the heart post-protocol change?

3. Was patient survival improved post-protocol change?

The first and second research questions were addressed by assessing how the HR of the pre-existing survival differential between patients with average residual set-up errors towards and away from the heart changed following the introduction of the new IGRT protocol. The third research question evaluated changes in median survival by considering patients who had average set-up errors towards and away from the heart separately.

Johnson-Hart et al. found that the association of residual set-up error with survival is not constant with time, as the Kaplan-Meier survival curves split and then come back together [268]. The excess mortality associated with residual set-up errors is hypothesized to result from radiation induced cardiac toxicity that manifests soon

after completion of radiotherapy in patients with lung cancer. Thereafter, cancer deaths dominate mortality in both cohorts bringing the Kaplan-Meier curves back together [268]. Rather than adding complexity to the analysis by modelling a time-varying hazard ratio in this exemplar analysis, we selected a set of constant hazard ratio multivariable Weibull survival models [279] after right censoring patients at 12, 18, 24 and 30 months. This parametric survival model can provide more power than non-parametric methods as it assumes that an underlying distribution for survival can be characterized by a small number of parameters (2 in the case of the Weibull model: the shape and scale parameters), enabling a better fit to survival data than non– or semi-parametric methods [280].

All survival models included the explanatory variables reported by Johnson-Hart et al: age, performance status, prescribed radiotherapy fractions and logarithm of GTV [268]. The first and second analyses also included the average residual set-up error direction as a binary variable (towards or away from the heart). The HRs of death for each factor were calculated from the model beta coefficients using the formula HR=exp(-beta coefficient*shape). In the third analysis, median patient survival was calculated using the formula median survival=scale*(ln(2))ˆ1/shape. The shape and scale parameters are the Weibull parameters that describe the overall shape and fitting of the survival curve [279, 281].

A brief introduction to Bayesian analysis is provided in Supplementary Materials. Two priors were used during analysis, a skeptical prior (i.e. based on the belief that large effects from an intervention are unlikely) and an uninformative prior (i.e. there is no prior information available). This led to two models: a skeptical model, using a skeptical prior calculated from the pre-protocol change cohort (i.e. incorporating the increase in hazard of death associated with residual set-up errors towards the heart) and an uninformative model, using an uninformative prior that let the post-protocol change cohort drive the model fit. The uninformative prior was default in the R package brms [269]; an improper flat prior over real numbers for the model variables, and a gamma (0.01, 0.01) for the shape parameter. Posterior distributions were calculated using Markov Chain Monte Carlo (MCMC) sampling (4 chains, 10,000 iterations, 5000 warm up) with MCMC chain convergence checked graphically (Supplementary

Figure 7.4). The HR and median survival distributions were plotted and the mean and 95% credible intervals tabulated for each. We calculated the probability for various hypotheses for each analysis: that the post-protocol change HR was less than the pre-protocol change HR; that the post-protocol change HR remained greater than 1; and that median survival had increased or decreased post-protocol change.

All statistical analyses were performed in R 4.0.0 [222] with package brms v2.13.0 [269, 282].

7.3 Results

The clinical variables for each cohort are listed in Table 7.1, demonstrating that patient, tumor and treatment characteristics are well balanced between cohorts.

Table 7.1: Baseline characteristics.

Variable	**Pre-protocol change cohort (N=780)**	**Post-protocol change cohort (N=411)**
Age in years, median (IQR)	71 (64-78)	72 (65-78)
Sex, n (%)		
Male	421 (54.0)	229 (55.7)
Female	359 (46.0)	182 (44.3)
ECOG PS, n (%)		
0	111 (14.2)	35 (8.5)
1	364 (46.7)	190 (46.2)
2	198 (25.4)	127 (30.9)
3	47 (6.0)	32 (7.8)
4	2 (0.3)	1 (0.2)
Missing	58 (7.4)	26 (6.3)
Stage, n (%)		
I	18 (2.3)	22 (5.4)
II	113 (14.5)	68 (16.5)
III	520 (66.7)	235 (57.2)
IV	38 (4.9)	13 (3.2)
Missing	91 (11.7)	73 (17.8)
GTV in cm^3, median (IQR)	45 (20-94)	48 (20-119)
Missing, n (%)	104 (13)	33 (8)
Dose and Fractionation, n (%)		
60-66 Gy in 30-33	159 (20.4)	50 (12.2)
55 Gy in 20	621 (79.6)	361 (87.8)
Residual set-up error direction, n (%)		
Towards	337 (43.2)	220 (53.5)
Away	443 (56.8)	191 (46.5)
Follow-up months, median (IQR)	99.3 (66.0, 118.2)	32.9 (29.0, 39.6)

Abbreviations: *GTV*, gross tumor volume; *IQR*, interquartile range.

Following the reduction in action threshold, the HR for residual set-up error towards the heart reduced (Table 7.2 and Figure 7.1). There was a high probability that the HR reduced by at least 5% for the skeptical model (Probability P>0.78), and 10% for the uninformative model (P>0.9) (Table 7.3). The HR was higher for the skeptical model than for the uninformative model. The HRs for the other variables included in the models remained similar pre- and post-protocol change (Supplementary Table 7.4).

Probabilities of a 15%, 20% and 25% reduction in HR are presented in Supplementary Table 2. Figure 7.2 shows the evolution of the HR posterior distribution as patients were sequentially added to the analysis, for both the uninformative and the skeptical model. The mean HR of the posterior distribution shifted to the left (i.e. towards HR of 1) as more patients were added to the models. The probability the HR had reduced increased from 0.608 to 0.994 for the uninformative model, and from 0.541 to 0.915 for the skeptical model. The skeptical model's HR posterior distribution approached HR=1 more slowly than the uninformative model. Together, these show that reducing the action threshold led to a reduced hazard of death for patients with residual set-up errors towards the heart, and the conclusion was reached more quickly for the uninformative model than for the skeptical model.

Table 7.2: HR for residual set-up error towards the heart for models with different censored follow-up times. Mean posterior HR is presented with 95% credible intervals.

Follow-up months	**Pre-protocol change (N=780)**	**Post-protocol change, uninformative prior (N=411)**	**Post-protocol change, skeptical prior (N=411)**
	Mean HR and 95% credible interval		
12	1.58 (1.24, 2.00)	0.937 (0.659, 1.30)	1.31 (1.07, 1.59)
18	1.43 (1.16, 1.73)	0.908 (0.68, 1.19)	1.19 (1.01, 1.40)
24	1.34 (1.11, 1.60)	0.909 (0.702, 1.16)	1.14 (0.980, 1.31)
30	1.26 (1.05, 1.49)	0.903 (0.702, 1.14)	1.09 (0.950, 1.26)

Abbreviations: *HR*, hazard ratio.

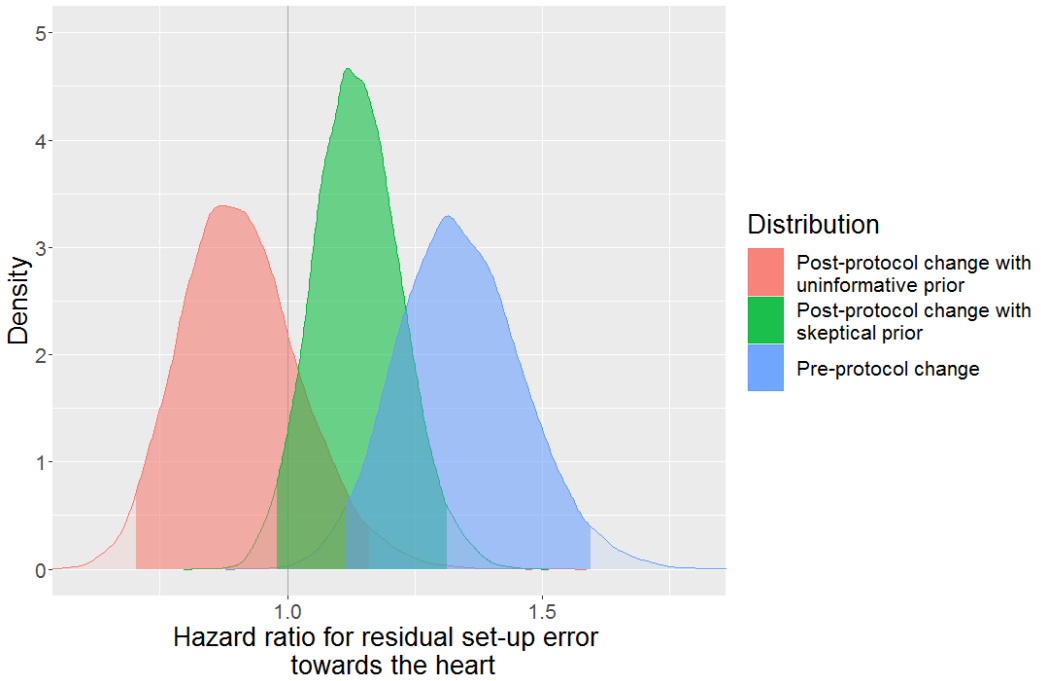

Figure 7.1: Posterior distributions for the HR of residual set-up error towards the heart with 24 months censored follow-up. The blue distribution presents the pre-protocol change data which creates the skeptical prior i.e. the HR in the data pre-protocol change. The red distribution is the posterior of the HR post-protocol change for the uninformative model, and the green distribution the posterior of the HR post-protocol change for the skeptical model.

Table 7.3: Bayesian probabilities calculated directly from the posterior distributions for the HR of residual set-up error towards the heart with different censored follow-up times. P(HR_post < HR_pre) is the probability that the HR is reduced in the post-protocol change cohort compared to the pre-protocol change cohort. P(5%|10% HR reduction) is the probability that the HR is reduced by 5% or 10% in the post-protocol change cohort. P(HR_post > 1) is the probability that the HR in the post-protocol change cohort is greater than 1.

	Uninformative prior				**Skeptical prior**			
Follow-up months	**P(HR_post < HR_pre)**	**P(5% HR reduction)**	**P(10% HR reduction)**	**P(HR_post > 1)**	**P(HR_post < HR_pre)**	**P(5% HR reduction)**	**P(10% HR reduction)**	**P(HR_post > 1)**
12	0.994	0.988	0.977	0.325	0.886	0.809	0.704	0.996
18	0.995	0.991	0.978	0.229	0.911	0.832	0.711	0.981
24	0.994	0.986	0.966	0.209	0.915	0.828	0.685	0.954
30	0.986	0.970	0.934	0.186	0.892	0.786	0.615	0.891

Abbreviations: *HR*, hazard ratio.

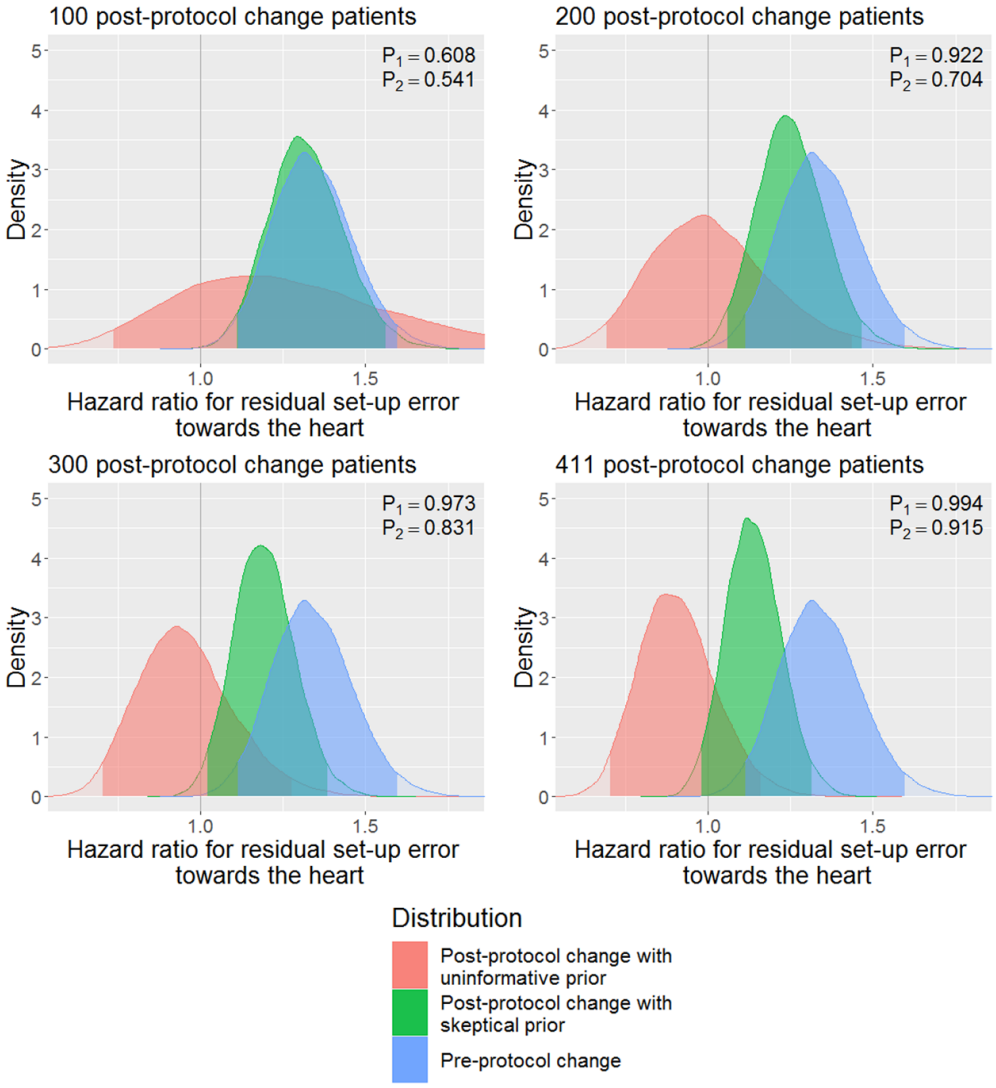

Figure 7.2: Evolving posterior distributions as more post-protocol data is added to the models. P1 and P2 are the probabilities that the HR for residual set-up error towards the heart is less in the post-protocol change data than the pre-protocol change data for the uninformative and skeptical models respectively. The blue distribution presents the pre-protocol change data, which creates the skeptical prior, the red distribution the posterior of the HR post-protocol change for the uninformative model, and the green distribution the posterior of the HR post-protocol change for the skeptical model.

The probability that the post-protocol change HR remained greater than 1 (i.e. that we still observe an effect) was moderately low ($P<0.4$) for the uninformative model, and high ($P>0.8$) for the skeptical model. This suggests a residual HR may remain following the reduction in action threshold, depending on choice of prior. The median survival (with 24 months follow-up) for patients with residual set-up errors towards the heart remained similar pre and post-protocol change for the skeptical model, from

17.1 (15.2–19.3) months to 17.7 (16.0–19.5) months, but increased for the uninformative model to 21.8 (18.2–26.4) months. For patients with residual set-up errors away from the heart, median survival decreased slightly from 20.8 (18.8–23.0) months to 19.6 (18.0–21.2) months (skeptical model) and 20.2 (17.0–24.3) months (uninformative model). Results for all follow-up times are presented in Supplementary Tables 7.6 and 7.7. Posterior distributions for median survival are presented in Figure 7.3. The uninformative model suggests there is a high probability that patients with residual set-up errors towards the heart have increased median survival post-protocol change (P=0.987), likely by at least 1 month (P=0.959). The skeptical model suggests median survival is moderately likely to have increased (P=0.678). For patients with residual set-up errors away from the heart, the uninformative model suggests median survival is moderately likely to have decreased (P=0.620), whereas the skeptical model suggests it is likely to have decreased (P=0.812). These results suggest patient survival increased post-protocol change for patients with residual set-up errors towards the heart, but patient survival may have decreased for patients with residual set-up errors away from the heart, depending on choice of prior.

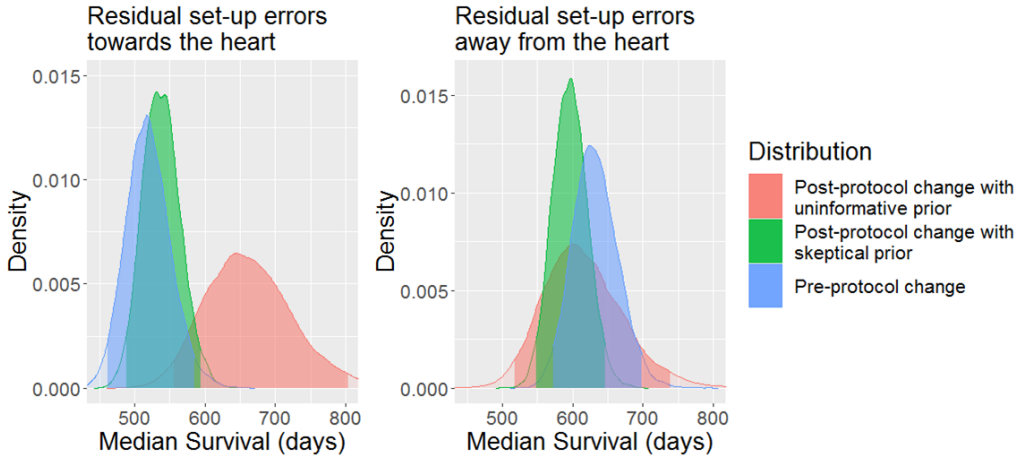

Figure 7.3: Posterior distributions for the median survival of patients with residual set-up errors towards the heart (left) and away from the heart (right) with 24 months censored follow-up. The blue distribution presents the skeptical prior, i.e. the median survival in the data pre-protocol change. The red distribution is the median survival posterior for the uninformative model, and the green distribution the median survival posterior for the skeptical model.

7.4 Discussion

Evaluating changes to clinical practice in radiotherapy is important to ensure technological advances lead to improved patient outcome, or more importantly, to equivalence with no worsening of outcome. Learning from data routinely collected in radiotherapy departments has the potential to allow us to monitor the impact of such changes as part of standard practice.

In this study, we used a Bayesian approach to analyze the impact of a change in IGRT protocol. We found that the difference between patient outcomes associated with patients having average residual set-up errors towards and away from the heart during a course of radiotherapy [268] was reduced when the action threshold was changed from 5 mm to 2 mm. We also found an increase in median survival post-protocol change for patients with average residual set-up errors towards the heart and, more weakly, a decrease in median survival in those patients with residual set-up errors away from the heart. These findings are in keeping with the theory that the survival differential results from cardiac toxicity caused by increased or decreased heart radiation dose when moving it on average towards (a harmful effect) or away from (a protective effect) the high dose treatment region [283].

The main limitation in reaching this conclusion is that it is drawn from a historically controlled retrospective observational dataset in which the outcome is simply compared before and after the intervention (the change in protocol). This experimental design is at risk from secular confounding, and may overestimate the effect of the intervention [284]. However, this approach is the closest fit to how radiotherapy is typically changed in practice, and would thus be the most amenable to adoption with minimal impact on current workflows. Furthermore, in the context of radiotherapy treatment we often have a good understanding of the potential confounders to our analyses. For example, Table 7.1 shows the distribution of confounding variables is very similar in the cohorts treated with the two different protocols, and the multivariable analysis reported in Supplementary Table 7.4 also shows there have been no major changes in the HR associated with each. Together these data give us confidence that the circumstances of the two patient groups are comparable and there is a good chance the effect we

observe is due to the change in IGRT protocol. Brink et al. used a time-dependent Cox model to analyse the effect of residual set-up errors on survival, finding the influence of set-up errors does vary over time [285], as expected from the Kaplan-Meier in Johnson-Hart et al.'s original study [268]. Instead of using a time varying hazard function we analyzed the data with different follow-up periods, finding similarly that the effect of residual set-up errors changes as patients are followed up for longer. We can see from Tables 7.2 and 7.3 and Supplementary Tables 7.5 to 7.7 that although the hazard ratios and probabilities raw values change, the interpretation when comparing the values pre-protocol change and post-protocol change does not.

This study compared a model with an uninformative prior for the HR of residual set-up error towards versus away from the heart, to a skeptical one incorporating the knowledge that there was a previously known harmful effect of having residual set-up errors towards the heart. The results highlight the influence using prior information has on the results. Both models suggest (with different certainties) a reduction in the difference in risk of death due to the protocol change, and differences in median survival following the change, which allows us to be confident in reporting these result. However, the models differ in their conclusion on whether a residual risk remains. The use of a skeptical prior in an analysis is important to ensure results are not over-interpreted and while using multiple priors for a Bayesian analysis is recommended, it is not necessary to interpret each result with equal weight [286]. In this case we cannot conclusively say whether a residual risk remains as the skeptical model incorporates the original effect the 5 mm action threshold had on patient survival, but equally we cannot rule it out. This result suggests further investigation of any residual risk is warranted.

The skeptical and uninformative models gave high and moderate probabilities that median survival decreased for patients with average residual set-up errors away from the heart. While ideally both models would clearly be in agreement, there is actually little difference between the point estimates for the uninformative and skeptical models (19.6 vs 20.2 months). Instead, this difference in probability is driven by the tighter distribution in the posterior when using the skeptical prior (Figure 7.3) that results from the additional information provided by the prior. Dependence on choice of prior

could be seen as a criticism of the Bayesian approach, but can also be viewed as a means to incorporate uncertainty and difference in expert opinion into the analytical process [287]. Meaningful priors can be generated for real world scenarios, such as we describe, by using historical data, as was done in this study, previous clinical trial data, or expert belief if no data is available. Incorporating historical data into analyses in the form of a prior has been shown to improve study power by improving the precision of the estimates [288, 289]. Indeed, Ryan et al. found that the use of informative priors in the analysis of RCTs could lead to fewer patients being enrolled and earlier completion of trials, due to increased study power for hypothesis testing [290]. This can clearly be seen in Figure 7.3 where the post-protocol change HR distribution is consistently tighter with the skeptical model than with the uninformative one.

We chose to analyze this data using a Bayesian approach rather a frequentist one due to the ease of interpretability with Bayesian analyses [77]. Whilst frequentist analyses are ubiquitous in medical sciences, so too is their misinterpretation [261, 262, 291]. Largely, this results from the underlying assumptions of the Null Hypothesis Significance Testing approach, which require careful study to fully appreciate. In contrast, by directly calculating parameter distributions Bayesian analyses allow multiple hypotheses to be tested while incorporating prior information from previous studies or expert belief. An important point is that even when an uninformative prior is used, and thus the results will be otherwise similar to those from a frequentist analysis, the approach to asking questions of the posterior is still Bayesian, including the ability to evaluate multiple hypotheses. Furthermore, in the context of evaluating changes to radiotherapy practice with real world data, Bayesian methodology allows for continuous updating of the posterior distribution as more data becomes available, as shown in Figure 7.2.

To conclude, the data we present provides real-world evidence that reducing the IGRT action threshold improved patients' clinical outcomes. The use of Bayesian methodology permitted us to test our results using both a skeptical and an uninformative prior, allowing us to be confident in some of our conclusions (decreased hazard ratio between patients with residual set-up errors towards and away from the heart, increased median survival in patients with shifts towards the heart) and cautious in others (the presence

of a residual hazard ratio with the 2 mm action threshold). The Bayesian approach is well suited for the assessment of technical changes in radiotherapy practice, and should be considered for the prospective evaluation of such changes.

7.5 Acknowledgements

This work was supported by CRUK via the funding to Cancer Research UK Manchester Centre: [C147/A18083] and [C147/A25254] and to Professor James O'Connor [C19221/A22746]. Professor Corinne Faivre-Finn and Professor James O'Connor are supported by NIHR Manchester Biomedical Research Centre.

7.6 Supplementary materials

7.6.1 Data preparation steps

Briefly, the image registration translations required to match the bony structures on the Cone Beam CT (CBCT) taken at the time of treatment to the planning CT scan were collected for each treatment fraction from CBCT (Elekta XVI) database archives. The action threshold was virtually applied to the recorded daily CBCT to CT image registration offsets, with those with offsets greater than the action threshold set to zero (assuming a perfect correction process). This yielded the residual set-up errors on days where CBCT images were acquired. These data were imputed to non-imaging days through nearest neighbor interpolation. The direction of residual errors relative to the heart were calculated using the center of mass of the heart and PTV delineations in the planning scan.

7.6.2 Bayesian analysis

In a Bayesian analysis, the data is fixed and parameters have probability distributions. This is in contrast to a frequentist analysis where probabilistic statements are made about the data (i.e. a P value tells us how surprising the data is, given the null hypothesis) and not hypotheses. In a Bayesian analysis, probabilistic statements can be made about hypotheses and parameters (i.e. how likely is one treatment to be

superior to another?). Bayesian models can incorporate prior information, known as the prior distribution, which gets updated when the model is introduced to the collected data, using Bayes Rule, to give the posterior distribution. The posterior probability distribution summarizes the current state of knowledge about a parameter, and the mean or median value can be derived to get a point estimate. Credible intervals can also be derived, which give a parameter range that has a particular probability of including the true unknown parameter value i.e. 95% credible interval. This is different to a frequentist confidence interval, which cannot be assigned a probability. Probabilities of hypotheses can also be calculated from the posterior distribution, for example, that one treatment is superior to another, or that survival is improved by at least 3 months.

7.6.3 Supplementary Figures

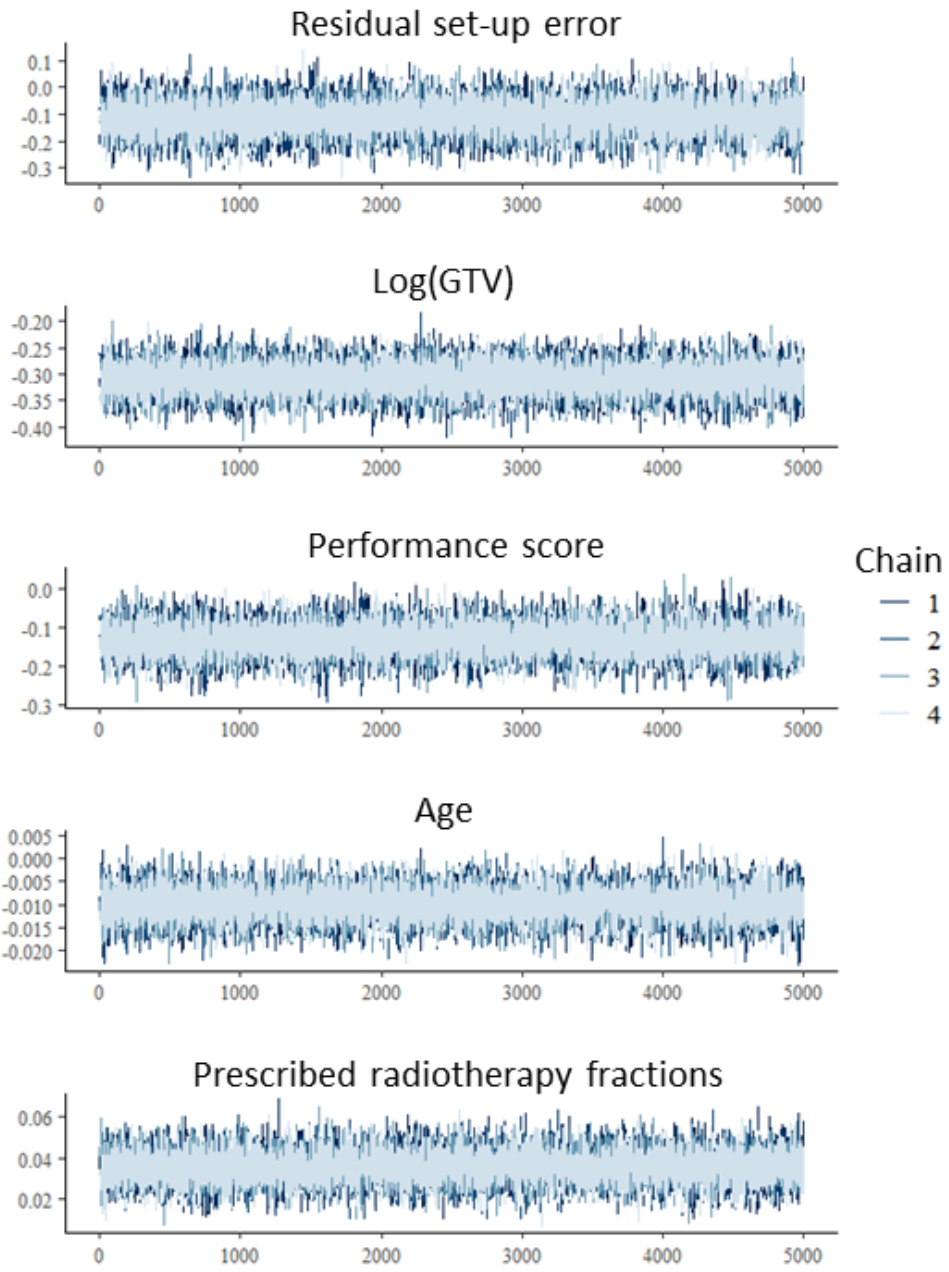

Figure 7.4: Example of chain convergence for the Bayesian survival model.

7.6.4 Supplementary Tables

Table 7.4: HR for all variables in the Bayesian survival model with different censored follow-up times. Mean posterior HR is presented with 95% credible intervals.

Follow-up months	**Variable**	**Pre-protocol change HR (N=780)**	**Post-protocol change, uninformative prior (N=411)**	**Post-protocol change, skeptical prior (N=411)**
12	Residual set-up error towards heart	1.58 (1.24, 2.00)	0.937 (0.659, 1.30)	1.31 (1.07, 1.59)
	PS	1.17 (0.994, 1.37)	1.01 (0.795, 1.27)	1.12 (0.981, 1.28)
	Age	1.00 (0.989, 1.02)	1.01 (0.992, 1.03)	1.01 (0.994, 1.02)
	Prescribed fractions	0.958 (0.930, 0.987)	0.942 (0.893, 0.990)	0.959 (0.935, 0.983)
	Log(GTV)	1.54 (1.37, 1.73)	1.72 (1.46, 2.02)	1.58 (1.44, 1.73)
18	Residual set-up error towards heart	1.43 (1.16, 1.73)	0.908 (0.680, 1.19)	1.19 (1.01, 1.4)
	PS	1.17 (1.02, 1.33)	1.10 (0.907, 1.33)	1.14 (1.02, 1.26)
	Age	1.00 (0.992, 1.01)	1.01 (0.993, 1.03)	1.00 (0.996, 1.01)
	Prescribed fractions	0.958 (0.935, 0.981)	0.960 (0.920, 0.997)	0.963 (0.944, 0.983)
	Log(GTV)	1.62 (1.46, 1.78)	1.54 (1.35, 1.75)	1.57 (1.45, 1.69)
24	Residual set-up error towards heart	1.34 (1.11, 1.60)	0.909 (0.702, 1.16)	1.14 (0.98, 1.31)
	PS	1.15 (1.02, 1.29)	1.18 (0.984, 1.39)	1.15 (1.04, 1.27)
	Age	1.00 (0.995, 1.01)	1.01 (0.999, 1.03)	1.01 (1.00, 1.02)
	Prescribed fractions	0.962 (0.941, 0.982)	0.97 (0.934, 1.00)	0.969 (0.951, 0.986)
	Log(GTV)	1.57 (1.44, 1.72)	1.51 (1.33, 1.69)	1.53 (1.43, 1.64)
30	Residual set-up error towards heart	1.26 (1.05, 1.49)	0.903 (0.702, 1.14)	1.09 (0.950, 1.26)
	PS	1.13 (1.01, 1.26)	1.16 (0.979, 1.36)	1.13 (1.03, 1.24)

Table 7.4 continued from previous page

Follow-up months	Variable	Pre-protocol change HR (N=780)	Post-protocol change, uninformative prior (N=411)	Post-protocol change, skeptical prior (N=411)
	Age	1.01 (0.998, 1.02)	1.02 (1.00, 1.03)	1.01 (1.00, 1.02)
	Prescribed fractions	0.965 (0.945, 0.984)	0.974 (0.941, 1.01)	0.971 (0.955, 0.988)
	Log(GTV)	1.53 (1.41, 1.66)	1.46 (1.30, 1.63)	1.50 (1.40, 1.60)

Abbreviations: *HR*, hazard ratio; *Log(GTV)*, logarithm of gross tumor volume.

Table 7.5: Probabilities calculated directly from the posterior distributions for the HR of residual set-up error towards the heart with different censored follow-up times.

	Uninformative prior			Skeptical prior		
Follow-up months	**P(15% HR reduction)**	**P(20% HR reduction)**	**P(25% HR reduction)**	**P(15% HR reduction)**	**P(20% HR reduction)**	**P(25% HR reduction)**
12	0.960	0.926	0.876	0.571	0.414	0.266
18	0.955	0.908	0.834	0.551	0.369	0.205
24	0.927	0.855	0.745	0.493	0.299	0.141
30	0.874	0.772	0.628	0.412	0.225	0.090

Abbreviations: *HR*, hazard ratio.

Table 7.6: Median survival pre- and post-protocol change for patients with residual set-up errors towards the heart.

Follow-up months	Pre-protocol change	Post-protocol change, skeptical prior	Post-protocol change, uninformative prior
	Mean HR and 95% credible interval		
12	15.9 (13.6-18.9)	16.0 (14.2-18.2)	21.6 (16.4-29.9)
18	16.3 (14.4-18.6)	16.8 (15.2-18.6)	21.5 (17.5-26.8)
24	17.1 (15.2-19.3)	17.7 (16.0-19.5)	21.8 (18.2-26.4)
30	17.9 (15.8-20.2)	18.2 (16.5-20.1)	21.6 (18.3-25.7)

Abbreviations: *HR*, hazard ratio.

Table 7.7: Median survival pre- and post-protocol change for patients with residual set-up errors away from the heart.

Follow-up months	Pre-protocol change	Post-protocol change, skeptical prior	Post-protocol change, uninformative prior
	Mean HR and 95% credible interval		
12	22.8 (18.6-28.7)	20.4 (17.7-23.8)	19.8 (15.1-27.4)
18	20.8 (18.5-23.6)	19.4 (17.7-21.3)	19.7 (16.2-24.5)
24	20.8 (18.8-23.0)	19.6 (18.0-21.2)	20.2 (17.0-24.3)
30	21.1 (19.3-23.2)	19.8 (18.3-21.4)	20.1 (17.1-23.9)

Abbreviations: *HR*, hazard ratio.

Chapter 8

Discussion

The aim of this thesis was to investigate the potential of real-world clinical and imaging data to gain clinical insight from patients with lung cancer. Real-world data derived from EHRs can provide large, representative and inclusive datasets which can be used to investigate important clinical questions and generate real-world evidence. In particular, patients with lung cancer who are old, frail and comorbid tend to be under-represented in RCTs, and could, therefore, benefit from the evidence generated from real-world data. Furthermore, in radiotherapy, changes to workflows are often not assessed in RCTs due to impracticality; they are expensive, difficult to implement and take a long time. There is also an implicit assumption new technological advances are superior to previous, and this can lead to evidence-gaps when the clinical impact of the implemented changes is not assessed. Methods of evidence generation are required to ensure such changes are beneficial, or at the least not detrimental, to patient outcomes. Ensuring decisions made in oncology are backed by high quality evidence is key to improving outcomes for all patients. The ultimate goal would be to prospectively monitor changes to clinical practice in a LHS environment using real-world data to generate high quality, real-world evidence. An essential part of the LHS concept is generating clinical insight through retrospective analyses. This thesis included numerous individual studies of retrospective analyses of real-world clinical and imaging data, which together addressed the following aims:

1. To investigate the potential of routine, real-world radiomics imaging biomarkers

to generate clinical insight and improve patient outcomes through supported decision making.

2. To develop approaches for using real-world data to assess whether changes to clinical practice affect patient outcomes.

The first aim was addressed in Chapters 2 and 3. Chapter 2 systematically reviewed the radiomics literature and found potential methodological limitations along the radiomics workflow that could be hindering clinical translation of radiomics biomarkers. The clinical radiomics literature in lung cancer was assessed, finding that all studies suffered from significant technical limitations, and that no single radiomics biomarker or methodological approach was used widely. Therefore, substantial barriers to clinical translation of radiomics biomarkers remain and further methodological studies are required to overcome these barriers.

Chapter 3 investigated how the platform used to extract radiomics features from real-world imaging data affected feature reliability and ability to predict survival. This work addressed a gap in the literature identified in Chapter 2. Choice of feature extraction platform, IBSI compliance, parameter settings and platform version were all found to affect feature reliability. This work led to the recommendations that radiomics studies use the latest version of an IBSI compliant software with harmonised parameter settings, and to publish parameter settings and software version to ensure reproducibility of the resulting radiomics biomarker, a key requirement for clinical translation.

The second aim was addressed in Chapters 4 to 7, which all investigated changes to clinical practice. Chapter 4 investigated the introduction of IMRT at The Christie NHS Foundation Trust by comparing 3 cohorts of patients treated in different time periods: pre-IMRT, some availability IMRT and full access to IMRT. The study found that the proportion of patients treated with curative-intent increased over the 3 time periods, and a survival benefit was found for patients treated in the latest time period with full access to IMRT compared to patients treated pre-IMRT, despite treating larger tumours and patients with poorer performance status.

Chapter 5 found that patients who had a change to their radiotherapy or chemotherapy treatments due to the COVID-19 pandemic did not have significantly worse overall survival or progression-free survival compared to patients whose treatments were not changed. Patients who had a change to their radiotherapy treatment had increased odds of $\geq$ grade 3 acute toxicity. This study produced real-world evidence that the recommendations put in place to reduce hospital attendances during the pandemic did not negatively affect outcomes.

Chapters 6 and 7 addressed the second aim by investigating whether Bayesian methodology would be suited to assessing changes to clinical practice. Chapter 6 presented the differences between Bayesian and frequentist statistics. Data were simulated to investigate the effect of a protocol change reducing the action threshold (the discrepancy allowed between planning position and treatment day position) in IGRT from 5 mm to 2 mm, using data from previous work that found that patients who had residual-set up errors pushing the heart towards the radiotherapy high dose region had worse survival compared to patients who had residual-set up errors pushing the heart away from the high dose region. The simulated data were analysed using both frequentist and Bayesian methods, demonstrating that Bayesian methods can lead to more informative and intuitive results. It also demonstrated how results are influenced by the choice of prior in a Bayesian analysis.

Chapter 7 then used Bayesian methodology to investigate the effect of a real-world dataset that the simulated dataset from Chapter 6 was based on. Models incorporating a skeptical prior (i.e. the model incorporated the known increased hazard of death for patients with residual-set up errors towards the heart pre-protocol change) and an uninformative prior (i.e. no prior information influenced the model) were compared. A reduced hazard of death was found for patients who had residual set-up errors that moved the heart towards the radiotherapy high dose region post-protocol change, for both the skeptical and the uninformative model. Depending on the choice of prior, a residual hazard of death may remain post-protocol change. There is strong real-world evidence where results agree with different priors, and weaker real-world evidence where results disagree depending on choice of prior. Overall, the results from Chapters 6 and 7 suggest that Bayesian methods would be well suited to evaluating the

impact of changes to practice in a LHS, as the data pre-change could be incorporated into the analysis and the results from different priors compared to decide how confident one is in the results.

8.1 Novelty and comparison to recent studies

Despite huge research interest in radiomics, clinical translation of radiomics biomarkers has been limited [292]. This is likely due to a lack of standardised methodology across radiomics studies, leading to irreproducible results [72]. The results from Chapters 2 and 3 aim to benchmark the methodology of radiomics studies and suggest how shortcomings may be improved. Another review article also evaluated the quality of radiomics studies in oncology according to RQS, a score that summarises the quality of a radiomics study, and TRIPOD, a statement provided to aid in transparent reporting of prediction models, finding the quality and the reporting of the included studies were poor [293], in agreement with the results from Chapter 2. The work in Chapter 2, however, goes beyond this by summarising the methodological limitations with radiomics and providing solutions to common issues along the radiomics workflow. It is also interesting to note that there is a difference in the reporting of results, for example both AUC (area under the curve) and CI (concordance index) are reported in Table 2.1, highlighting that reporting of results is not standardised. A roadmap for the clinical translation of radiomics biomarkers into clinically useful tools was recently published by a group of radiologists, physicists, and statisticians [292], providing 16 criteria to achieve this aim. The criteria include defining the target population for which the radiomics test will be useful, standardising imaging protocols, ensuring reproducibility of the test, as well as ensuring there is sufficient patient benefit resulting from acting on the results of the radiomic test. These criteria are reassuringly similar to the conclusions of the results from Chapters 2 and 3, and the work from Chapter 3 was referenced.

Various studies have investigated methodological issues with radiomics, which were summarised in Chapter 2. The work in Chapter 3 was the first to investigate the impact of IBSI compliance (an initiative set up to standardise feature extraction) and platform version on feature reliability in lung cancer. Mistakes were found in the code

of LIFEx and CERR, 2 widely used platforms in the radiomics community, and reliability was improved once the mistakes were corrected. Differences in feature reliability between platforms had previously been investigated in other cancer sites and imaging modalities, for example Foy et al. compared features extracted from Mammograms and H&N CT scans across two in-house platforms, MaZda v4.6 and IBEX, finding less than excellent reliability for many features, particularly texture features [211]. Default settings were initially used, and then texture feature settings were harmonised, as much as possible, and reliability increased, but not for all features. The study did not include IBSI-compliant platforms, which may have increased reliability further after harmonising parameter settings. Bogowicz et al. compared two in-house developed platforms and extracted features from H&N PET scans, finding 88% of features were not reproducible between platforms [197]. This study implemented a fixed bin size for image intensity discretisation, however did not document whether other parameters were harmonised, such as lower and upper bounds for histogram and texture features which can differ between platforms, as was found in Chapter 3 [73]. Harmonising all parameter settings may have improved feature reproducibility between platforms. Liang et al. compared the IBSI compliant platform PyRadiomics v2.1.2 with the non-IBSI compliant Moddicom v0.51 across CT and MRI images, finding 76.1% of features were highly correlated from the CT extracted features, but only 28.6% of the MRI extracted features were highly correlated [212]. The study also extracted MRI features using CERR, an IBSI-compliant platform, and found shape and intensity features to be reproducible between CERR and PyRadiomics, but not texture. This is in contrast to the study in Chapter 3, which found CERR and PyRadiomics to have excellent reliability for all features; however, the version of CERR that was used in the analysis was not documented, so it is likely an older and different version of CERR was used. It is also unclear whether parameter settings were default or harmonised; if default, then this may explain the difference in texture features. These studies highlight the importance of the IBSI to standardise feature extraction platforms, and suggest the issues of reliability found in Chapter 3 for SCLC planning CTs, NSCLC diagnostic CTs and H&N planning CTs are also applicable to other cancer sites and imaging modalities. Chapter 3 is a particularly important contribution to the radiomics community as it highlights how important using the latest version of a publicly available software

is, as well as harmonising parameters when comparing results from different software. Furthermore, the comparison of IBSI compliant platforms revealed not all features had excellent reliability, so it is important to test platforms to ensure IBSI-compliance rather than assuming it to be true.

A recently published paper compared features extracted from three IBSI platforms across CT and MRI scans and found some features to have poor reliability, particularly when extracted from perfusion MRI-based maps [294]. However, the study did not harmonise parameter settings which may have led to improved reliability as found in Chapter 3 [73]. LIFEx was one of the three platforms that were compared, and a discretisation error was found in LIFEx v7.0 when compared to v6.65. These versions of LIFEx are updates to the ones compared in Chapter 3, suggesting that even using the latest version of a platform may not be reliable. Platforms should therefore be checked before use to ensure IBSI compliance. A highly cited 'how-to' guide for radiomics analyses recommends using an IBSI-compliant feature extraction platform [295], concurring with the results in Chapter 3. However, this guide does not mention the importance of using and testing the latest version of the platform, which is also important when considering some versions may have mistakes in the feature calculation code [73, 294]. Ensuring software are actually IBSI-compliant when they claim to be is important, as studies could become obsolete with software updates if the features calculated were different. It can be difficult to know this without explicitly testing the software against another before conducting the study.

Chapter 4 provided real-world evidence that outcomes had improved since the introduction of IMRT, and more patients were being treated with curative-intent. While this was not a causal analysis, we are confident that IMRT at least contributed to this increase in patients being treated with curative intent and improved survival, due to the finding in Chapter 4 that patients treated when there was full access to IMRT had poorer performance status and larger tumours than patients treated in previous time periods. This evidence is important, as a lung cancer national audit found that the majority of stage III NSCLC patients were receiving palliative treatments, even those with a PS of 0/1 [231]. A secondary analysis of a RCT compared patients treated with IMRT to 3D conformal radiotherapy, finding survival to be similar between the groups

[233]. Patients treated with IMRT had larger planning target volumes [233], which is different to our findings that PTV decreased in the latest time period with full access to IMRT. We did find that gross tumour volume increased in the latest time period, so the discrepancy may be due to differences in PTV margins. The IMRT group had lower rates of $\geq$ grade 3 pneumonitis and received lower cardiac doses, and the the volume of the heart receiving 40 Gy was associated with worse survival [233]. Given the results from Chapter 7 and the wider literature suggesting increased radiation dose to the heart may lead to worse survival [242, 296, 297], IMRT could help improve survival through decreased cardiac toxicity.

The study in Chapter 5 is the first, to our knowledge, to look at differential outcomes for patients who did and did not have their treatment changed due to the COVID-19 pandemic. The findings that patients who had a change to their radiotherapy or chemotherapy treatment did not have worse outcomes will be useful in the context of another pandemic or emergency situation requiring a reduction in hospital visits. The National Lung Cancer Audit found 1-year survival decreased from 2019 to 2020 [258], so the results from Chapter 5 are encouraging, although our study only included patients treated with curative-intent. Our results also suggest outcomes are comparable between conventional and hypofractionated radiotherapy, as patients who had a change to their radiotherapy treatment received a higher dose per fraction compared to patients who did not have a change, which is in line with published guidelines recommending hypofractionation during the COVID-19 pandemic [250]. This result is in contrast with a study by Brada et al., who performed a retrospective analysis of 169,863 patients and found that patients with stage I-III NSCLC who received moderately hypofractionated radiotherapy of 55 Gy in 20 fractions had worse survival than patients who received conventional radiotherapy of 60-66 Gy in 30-33 fractions (25 months vs 28 months, p=0.02) [253]. This study, however, suffers from selection bias as it does not differentiate between patients who received concurrent and sequential chemotherapy, and patients in the UK often receive 55 Gy in 20 fractions if they are not fit enough for chemotherapy, or following induction chemotherapy. Therefore, the survival difference could be due to patients being more unwell in the hypofractionated arm rather than due to hypofractionation itself. The study did adjust for potential

confounders

High quality clinical evidence that hypofractionated radiotherapy is equivalent, or superior, to conventional radiotherapy in locally advanced lung cancer is lacking [298]. A randomised phase III trial investigated whether 60 Gy in 15 fractions over 3 weeks (hypofractionated arm) was superior to 60 Gy in 30 fractions over 6 weeks (conventional arm) for patients with stage II-III NSCLC not fit for concurrent chemotherapy, however the trial was closed early when an interim analysis failed to show a survival benefit [251]. The study was not powered to show equivalence; however, no significant difference in 1-year survival (37.7% in the hypofractionated arm vs 44.6% in the conventional arm), local and distant relapse, and $\geq$ grade 3 toxicity was found, although there was a higher rate of grade 2 toxicity in the hypofractionated group. In contrast, our study in Chapter 5 found increased odds of $\geq$ grade 3 toxicity for patients who had a change to their radiotherapy, however we did not compare specific radiotherapy regimens.

The COVID-19 pandemic highlighted the importance of real-world data and generating rapid, high-quality evidence to inform clinical practice. A LHS could lead to accelerated identification of effective treatments, as well as allow continuous monitoring and improvement of treatments using routinely collected data [299]. A team at Stanford Hospital developed a LHS environment during the COVID-19 pandemic to generate evidence that would inform guidelines to manage a predicted surge in patients infected with COVID-19 [300]. In 2 weeks, 4 clinical questions were answered using EHR data which influenced hospital guidelines. For example, an analysis found that patients who were discharged with home oxygen had similar readmission rates to those who no longer required oxygen, allowing quicker discharge of patients still on oxygen [300]. Had the infrastructure been in place, the work in Chapter 5 could have been evaluated in a LHS environment across multiple hospitals, although this would require data sharing agreements and therefore added complexity. Whilst LHS are often talked about in the literature, their use in the real world is limited [70].

While the COVID-19 pandemic was a specific situation where changes to practice occurred and rapid answers to clinical questions were required, radiotherapy is also

a field where changes to practice often occur without formal evaluation. Chapters 6 and 7 demonstrated how Bayesian methodology could be used to assess the impact of changes to clinical practice in radiotherapy, using the example case of a change in IGRT protocol. Chapter 7 provided real-world evidence that a change in IGRT protocol led to a reduced hazard of death for patients who had residual set-up errors moving the radiotherapy dose towards the heart. Work by McWilliam et al. found that dose to the base of the heart was significantly associated with worse patient survival [296, 297], suggesting the reduced action threshold investigated in Chapter 7 may have reduced heart toxicity and therefore risk of death.

The work in Chapter 6 used a simulated dataset based on the data used in Chapter 7. To simulate the datasets, a patient distribution was first created using different explanatory variables (age, performance status, stage and GTV). Dose to GTV, dose to heart and shift to heart were added, and hazard ratios assigned to all parameters based on the literature. A baseline hazard was simulated using a weibull model with multivariable hazard ratios. Uniform patient accrual was assumed and patients were censored according to the accrual and follow-up period specified. The action threshold could be modified (i.e. to create the 5mm and 2mm datasets), and the hazard ratio of the residual set-up error per mm was included which was derived from the original study by Johnson-Hart et al. [268]. This allowed two datasets to be simulated, one with a 2mm action threshold and one with a 5mm action threshold, both simulating the effect of having a residual-set up error pushing the dose towards versus away from the heart.

The Bayesian analysis of the change in IGRT protocol presented in Chapter 7 used a before/after design to assess the protocol change. Bayesian designs have previously been adopted in the field of radiotherapy to compare treatments using observational data. A study by He et al. used Bayesian statistics to compare patients with oesophageal cancer who received either 3DCRT or IMRT, finding the group treated with IMRT had lower risks of pericardial effusion, pleural effusion and death [301]. The authors calculated probabilities of harmful effects, deciding that probabilities ≥ 0.95 or ≤ 0.05 were significant. This is a frequentist way of thinking, particularly the use of the arbitrary cut-off value 0.05. The study also did not specify what prior distributions

were used. As discussed and demonstrated in Chapters 6 and 7, incorporating prior information into analyses can improve precision, as well as allow sensitivity analyses to be performed to discover how reliable the results are. Ensuring methods and statistical analyses are well documented and reported is vital to allow critical appraisal and comparison of studies, as highlighted in the radiomics review in Chapter 2.

8.2 Limitations and future work

This thesis has demonstrated the potential of using real-world imaging and clinical data to generate clinical insight and real-world evidence. However, there are limitations to the included studies and plenty of opportunities for future work.

The work in Chapters 2 and 3 aim to improve the quality of future radiomics studies. Future radiomics work should be methodological in nature, aiming to ensure results are reproducible and repeatable across different imaging machines, protocols and users. Radiomics may be moving towards a fully machine-learning based approach in the future, such as using deep-learning methods to extract and create features, and perform analyses [292]. Chapter 2 did not include deep-learning studies; however, the use of artificial intelligence to learn from imaging and clinical data is of increasing interest and may help to solve standardisation issues by developing robust imaging features and complex correction algorithms. Further research will be required to ensure results from such models are interpretable, and that models can be tested and re-calibrated as required as time goes on. Indeed, the radiomics quality score has been updated since the work in Chapter 2 to the RQS 2.0, which now includes a specific score for deep learning radiomics studies [302]. Additionally, the TRIPOD statement, initially developed to aid in improving reporting and critical appraisal of prediction/prognostic models, is in the process of being extended for machine learning models with TRIPOD-AI [303]. Most of the studies included in Chapter 2 were done on pre-treatment or planning scans; however, there is the potential to use follow-up scans, for example to use radiomics to replace repeat tissue biopsies or compliment liquid biopsies. It may be more useful to look at endpoints that come from biopsies, such as mutational status, than typical endpoints investigated such as overall survival, which some studies did look at in Chapter 2. We found no feature class to be more prominent amongst

particular endpoints; however we did not look into this in detail. It would be interesting to look at whether shape or texture features are generally favoured for particular outcomes.

A key challenge of real-world data is that it is collected during the course of clinical care and is not intended to be used in analyses. This means the datasets used in real-world analyses are not always of high quality due to poor recording, missing data and variations in terminology for example [304, 305]. The work in Chapter 4 included data collected from as early as 2005 which predated EHRs at The Christie NHS Foundation Trust; therefore, potentially important confounding variables, such as systemic therapies or radiotherapy technique, were not available. Furthermore, GTV and PTV data were not complete due to older radiotherapy planning data being archived and difficult to retrieve. Had the radiotherapy technique used been known for the work in Chapter 4, a quasi-experimental comparative analysis of patients treated with IMRT versus 3D conformal RT could have been performed; for example, in a difference-in-differences analysis which would have compared the difference in outcomes over time between patients treated with IMRT and 3D conformal radiotherapy. Moreover, ideally we would have looked at all patients with lung cancer to ensure we had the optimal denominator when looking at the proportion of patients treated with curative intent, rather than just those treated with radiotherapy. However, we only had data from patients who were referred for radiotherapy. Datasets that include patients within greater Manchester diagnosed with cancer could mitigate this issue. It is also important to note that the number of patients treated with palliative radiotherapy could have decreased due to improved systemic anti-cancer therapies; however, we did not have data on chemotherapy or immunotherapy to investigate this. In Chapter 5, the finding that patients had increased grade 3 toxicity could be confounded by undiagnosed COVID-19 infection, since the results were driven by pulmonary toxicity. Since COVID-19 testing was sporadic during the first wave of the pandemic, we do not have the data to be able to investigate this.

Structured toxicity data have been collected at The Christie NHS Foundation Trust since 2020 in lung cancer, however prior to this it was difficult and time consuming

to derive toxicity data from EHRs. Unstructured notes may contain toxicity information; however, it is often not graded (for example according to Common Terminology Criteria for Adverse Events (CTCAE), it is time consuming to go through patient's notes to find these data and it is not guaranteed to be recorded for all patients. Had structured toxicity data been collected for the patients in the IMRT analysis in Chapter 4, it would have been interesting to look at whether patients who received IMRT had reduced toxicity compared to 3D conformal radiotherapy, as the modulation of the beam intensities helps to avoid nearby organs at risk and therefore reduces normal tissue toxicity. It would have also been interesting to look at the impact of the IGRT protocol change, evaluated in Chapter 7, on cardiac toxicity as well as survival, had these data been collected. Future work could investigate these questions when high quality data on toxicity outcomes become available, for example from electronic patient-reported outcome measures (ePROMS) which have recently been implemented at The Christie NHS Foundation Trust [57].

Further data quality issues associated with real-world data include missing data. A large proportion of data were missing in the IMRT analysis in Chapter 4. The multivariable analysis only included patients who had complete data; however, this reduced the power of the analysis due to reducing the sample size. A larger proportion of missing data occurred in the earliest time period, as this pre-dated EHRs at The Christie NHS Foundation Trust. Within each time period, however, data were assumed to be missing at random, as data entry relied on the treating clinician and not on the patient's prognosis. Therefore, it is unlikely that the worse survival seen in the earliest time period (pre IMRT) compared to the latest time period (full access to IMRT) is biased due to missing data. A key advantage of the COVID-19 study in Chapter 5 was that the data had been collected prospectively, meaning there was very little missing data for the clinical variables. Unfortunately, we cannot know how many patients are missing from that analysis as it was up to the individual centres to fill out the forms for each patient referred for curative-intent radiotherapy. Prospective data collection could increase the quality of real-world datasets, but at the cost of increased time and effort on behalf of healthcare professionals inputting the data. It would also require educating clinicians who input data on the importance of ensuring the inputted data

are accurate. Research is ongoing at The Christie NHS Foundation Trust to quantify the quality of data in EHRs, which could help to appraise results of studies that use the data. Furthermore, a large amount of clinical data are held in an unstructured format, so future work around developing computational tools to convert the unstructured data into easily useable formats is key to increasing the amount and type of data available.

RCTs are the gold standard of evidence in medicine, and their design allows causal relationships to be established. It is much more difficult to establish a causal relationship in an observational, real-world study; however, statistical techniques can help to infer causality. It is important to remember, however, that correlations are not necessarily causal. Adjusting for confounding variables, as done in Chapters 4 to 7, is important to ensure results are not biased and helps towards inferring causality [306]. Cox proportional hazard models were fitted throughout this thesis, and this model assumes proportional hazards. This was checked using Schoenfeld residuals. A framework for inferring causality from observational data in the absence of a RCT has been developed by Hernan et al., called the target trial [307]. The target trial framework works to emulate a RCT and derive effect estimates from observational data that are the same, or very similar, to those that would have been derived from a RCT. Such a framework could be used to investigate comparative studies. In order to emulate randomisation, however, confounding variables must be accounted for. If the data required is not available, then a successful target trial is not possible.

It is vital to have clinical engagement in real-world evidence studies to ensure all confounding variables are taken into account, as much as possible depending on the data available, as well as helping to interpret results soundly. In Chapters 4 to 7, confounding factors were carefully identified through discussions with treating clinicians and accounted for where possible. In Chapter 4, we could not conclude a causal relationship between the survival benefit seen in the time period with full access to IMRT, as other changes to clinical practice had occurred during the time span of the study which were not accounted for due to a lack of available data. On the other hand, we can be more confident that the IGRT protocol change in Chapter 7 did lead to a reduced hazard of death for patients with residual-set up errors towards the heart because of

the adjustment for key confounding variables, the balance of variables between cohorts and the fact the posterior hazard ratios for the clinical variables included in the models did not change significantly between protocols. To increase confidence in the results of observational studies in future work, casual graphs, such as directed acyclic graphs, could be used to determine casual and non-casual pathways between variables. This can help to determine which variables need to be accounted for.

Another limitation of working with retrospective, real-world data can be the sample size. Real-world datasets can be large, such as the IMRT work in Chapter 4 which included 12,499 patients; however, depending on the data collected they can also be small. It can, therefore, be difficult to identify small signals from the data due to insufficient statistical power. In Chapter 5, the amount of data collected could not be increased due to the unique time period they were collected in: the first wave of the COVID-19 pandemic. The association between omitting chemotherapy and survival and distant relapse was not significant, but the confidence intervals were close to 1, suggesting there may be a small effect which could not be detected with the sample size in the study. Of course, we cannot conclusively infer this, as there may well be no significant impact; however, had we had more data this could have been investigated. A Bayesian analysis would have allowed incorporation of prior information, so we could have included the known detrimental effect omitting chemotherapy has for patients with lung cancer [255], as well as allow a decision to be made that more data are required to reach a conclusion without penalty. Future work could attempt to find larger retrospective datasets during that same period; however, it is unlikely that data will have been collected on treatment changes made due to COVID-19. Sample size was also an issue in the Bayesian analysis of the IGRT protocol change in Chapter 7, as some results differed depending on the choice of prior used, and more data would be required to make a solid conclusion that both priors agreed with. Unfortunately, more data could not be collected for that study either, as all patients with lung cancer treated with the 2 mm IGRT protocol at The Christie NHS Foundation Trust were included. Shortly after the implementation of the 2 mm action threshold, the action threshold was eliminated such that all patient set-up errors were corrected. Future work could involve collecting data from other centres who also implemented this protocol change

and repeat the analysis.

The Bayesian methodology used in Chapters 6 and 7 highlights how the choice of prior affects results. In particular, Chapter 6 shows how the uninformative prior does not give exactly the same results as the frequentist analysis. If a prior is truly uninformative, these should be identical. It is important to note that even uninformative priors contain some information, so are not completely uninformative. It can, therefore, be difficult to pick an appropriate prior, both informative or uninformative. Priors are subjective, and frequentist methods may therefore be preferred by some as they allow only the data to drive the inference. While testing multiple priors can help ensure validity of results, it could also lead to exploitation if the investigator only chooses to present results with priors that make their results look better. It is therefore important to check results against an uninformative prior, and have solid reasoning for picking a particular prior. Deciding on the prior in an analysis plan prior to performing the analysis would be best practice.

Bayesian methods could now be applied to other retrospective changes to clinical practice, across different cancer sites. It would be interesting to also investigate different endpoints, if the data are available, for example treatment-related toxicity or ePROMS. Implementing a causal framework into the analysis would also be of interest to improve the quality of the resulting real-world evidence, for example by emulating a target trial [307]. The ultimate goal would be to design high quality, real-world evidence studies to prospectively monitor changes to practice in a LHS environment. Such methods could be resource-intensive, due to data quality checks and updating results in real-time using a Bayesian approach. One such study is the RAPID-RT project [308], which aims to prospectively use rapid learning in the clinic to introduce a heart dose-limit for lung cancer radiotherapy and optimise the limit via learning cycles to give the best clinical outcome. The study also aims to investigate the costs associated with such a LHS environment, and how much resource is required to expand it to other centres. It is an exciting opportunity to demonstrate how real-world data can be used within a LHS to generate real-world evidence to compliment RCTs.

8.3 Impact

The radiomics review presented in Chapter 2 has been cited 57 times (Scopus, 22/06/2023) since its publication in 2020 [72], suggesting the radiomics community is becoming aware of the limitations in the literature and taking steps to improve future studies.

The study presented in Chapter 3 discovered errors in two of the four radiomic feature extraction platforms that were compared [73]. An error in CERR was found when calculating texture features; the minimum and maximum value of the defined ROI was meant to be used but instead the platform used the minimum and maximum of the entire image. An error in LIFEx was found when calculating sphericity, however the platform is closed-source and so the specific issue could not be identified. The developers of CERR fixed the error as soon as they were made aware (https://github.com/cerr/CERR/commit/50530f7). LIFEx took 6 months to fix the error, during which the paper was published. An ESTRO poster presentation was adapted to include the updated and corrected version of the LIFEx code, LIFEx version 6.0, demonstrating excellent reliability with other software for all features (Figure 8.1) [309]. Over 400 papers have been published using LIFEx software; unfortunately, most papers do not specify the software version used in their analysis. This work, therefore, led to effective improvement in two tools used widely by the radiomics community. Furthermore, the paper covering work in Chapter 3 has been cited 85 times (Scopus, 22/06/2023) since publication in 2020, suggesting the radiomics community is taking steps to use IBSI compliant platforms, and in particular the corrected version of LIFEx [73]. This work has also been cited in two key roadmap papers written by leading experts in translational imaging science [292, 310].

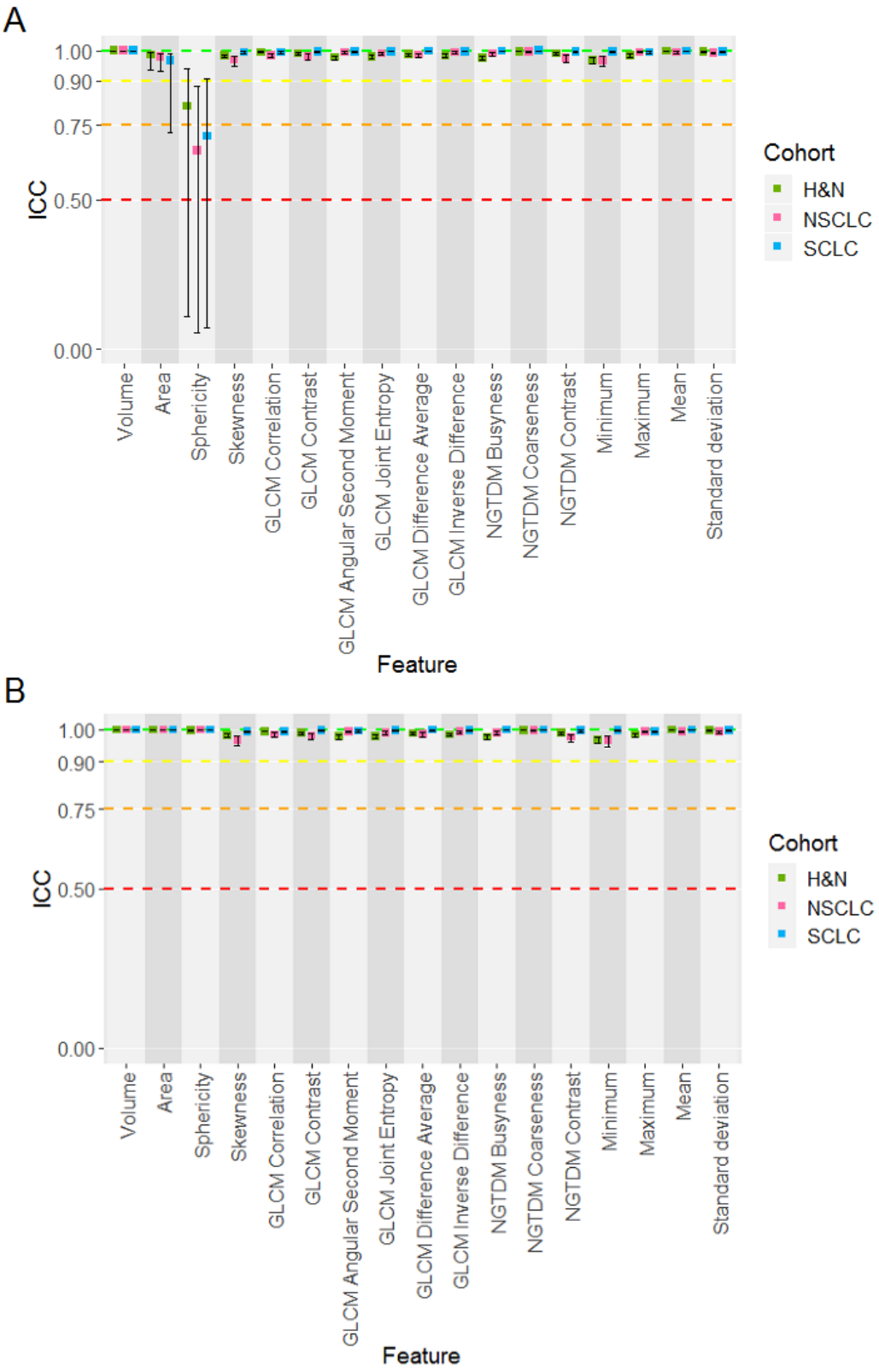

Figure 8.1: Boxplot of ICC estimates and CI for each cohort (H&N in green, NSCLC in pink, SCLC in blue) across comparable features, showing the statistical reliability across three IBSI-compliant software with harmonised calculation parameters. **A** ICC estimates and CI for the three IBSI-compliant software from Chapter 3 (PyRadiomics, CERR and LIFEx v5.47) **B** ICC estimates and CI for the three IBSI-compliant software with the corrected version of LIFEx (LIFEx version 6.0).

The IMRT work presented in Chapter 4 and the COVID-19 work presented in Chapter 5 both demonstrate how real-world data can be used to monitor and assess the impact of changes to practice. In particular, the IMRT study demonstrated how real-world data can be used to ensure new treatments are beneficial to patient populations where clinical trials are not typically done [74]. The COVID-19 study contributed to the body of evidence that the increased use of hypofractionated radiotherapy treatments did not significantly compromise tumour control or mortality, while being more convenient and a cheaper alternative to conventional fractionation regimens. Furthermore, the study suggests the guidelines put in the place at the start of the COVID-19 pandemic were effective, and may have prevented vulnerable deaths due to COVID-19.

The Bayesian versus frequentist teaching article presented in Chapter 6 [77] received a Letter to the Editor that argued in favour of frequentist statistics over Bayesian [78]. We replied to their points, concluding that the study design and statistical assumptions are more important to the results than the analytical approach itself [79]. Both letters can be found in the Appendix. This shows the impact of our original paper by engaging radiation oncologists in such discussions. A podcast was recorded for the International Journal of Radiation Oncology-Biology-Physics discussing the results of the paper (https://www.redjournal.org/audio-do/red-journal-podcast-april-1-2022) and the paper has been used in teaching materials. The paper has also been cited in fields outside of radiation oncology, such as the financial, educational and environmental sectors, promoting the wider use of Bayesian methods.

The Bayesian analysis of the IGRT reduced action threshold presented in Chapter 7 concluded that Bayesian methods would be well suited to monitoring of changes to clinical practice [80]. The methodological research helped to inform the statistical design of the NIHR funded RAPID-RT study, which aims to prospectively assess the impact of a change in protocol aiming to avoid a region of the heart associated with increased risk of death [296, 297].

Chapter 9

Conclusion

This thesis presents multiple studies with the overall aim of investigating the potential of using real-world data to gain clinical insight from patients with lung cancer. The key findings of this thesis are:

1. Clinical translation of radiomics, using real-world imaging data, is presently hindered by various technical validation shortcomings.

2. The platform used for radiomics feature extraction, as well as parameter settings, IBSI-compliance and platform version, affect feature reliability which in turn can affect outcomes analyses.

3. The introduction of IMRT for patients with lung cancer led to an increased number of patients being treated with curative intent and improved patient survival.

4. Changes made to radiotherapy treatments for patients with stage I-III NSCLC due to the COVID-19 pandemic did not significantly affect overall survival or progression-free survival; however, did increase odds of $\geq$ grade 3 acute toxicity.

5. A change in protocol reducing the IGRT action threshold reduced the risk of death associated with residual set-up errors for patients with lung cancer.

6. Bayesian methods are well suited to monitoring changes to clinical practice with real-world data.

Together, these results demonstrate the potential large, real-world datasets have to monitor and improve outcomes for patients with lung cancer. The research setting is a powerful tool for generating evidence and insight from retrospective data. The NHS have key aims to make better use of EHR data to save lives [64], and NICE and the FDA have developed real-world evidence frameworks to make better use of real-world evidence in the development of guidelines and medical products [40, 65]. The promise of real-world data is being recognised. The ultimate goal and next challenge is to embed real-world evidence generation into clinical practice to prospectively monitor changes to practice and the impact on patient outcomes, for example in a LHS environment. In this way, practice could be adapted in real-time in response to study results. Ultimately, this would ensure every patient is receiving the best possible treatment backed by the highest quality evidence.

Chapter 10

Publications and Presentations

10.1 Publications

10.1.1 Journal articles

H-index=4

- Fornacon-Wood I, Banfill K, Ahmad S, et al. Impact of the COVID-19 Pandemic on Outcomes for Patients with Lung Cancer Receiving Curative-Intent Radiotherapy in the UK. Clinical Oncology. 2023;35(10):e593–600. doi.org/10.1016/j.clon.2023.07.005.

- Fornacon-Wood I, Mistry H, Johnson-Hart C, Faivre-Finn C, O'Connor JPB, Price GJ. Bayesian methods provide a practical real-world evidence framework for evaluating the impact of changes in radiotherapy. Radiother Oncol. 2022;176:53-58. doi:10.1016/j.radonc.2022.09.009.

- Fornacon-Wood I, Mistry H, Johnson-Hart C, Faivre-Finn C, O'Connor JPB, Price GJ. Understanding the Differences Between Bayesian and Frequentist Statistics. Int J Radiat Oncol Biol Phys. 2022;112(5):1076-1082. doi:10.1016/j.ijrobp.2021.12.011

- Banfill K, Croxford W, Fornacon-Wood I, et al. Changes in the Management of Patients having Radical Radiotherapy for Lung Cancer during the First Wave of

the COVID-19 Pandemic in the UK. Clin Oncol. 2022;34(1):19-27. doi:10.1016/j.clon.2021.10.009

- Fornacon-Wood I, Chan C, Bayman N, et al. Impact of Introducing Intensity Modulated Radiotherapy on Curative Intent Radiotherapy and Survival for Lung Cancer. Front Oncol. 2022;12:835844. doi:10.3389/fonc.2022.835844

- Fornacon-Wood I, Faivre-Finn C, O'Connor JPB, Price GJ. Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype. Lung Cancer. 2020;146:197-208. doi:10.1016/j.lungcan.2020.05.028

- Fornacon-Wood I, Mistry H, Ackermann CJ, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. Eur Radiol. 2020;30(11):6241-6250. doi:10.1007/s00330-020-06957-9

10.1.2 Conference abstracts

- Fornacon-Wood I, Banfill K, Ahmad S, et al. MO-0467 Impact of the COVID-19 pandemic on patient outcomes after curative-intent radiotherapy in the UK. Radiother Oncol. 2023;182:S370-S371. 10.1016/S0167-8140(23)08389-5

- Chan C, Fornacon-Wood I, Bayman N, et al. OC-0437 Impact of introducing IMRT on curative intent radiotherapy and survival for lung cancer. Radiother Oncol. 2022;170:S380-S381. doi:10.1016/s0167-8140(22)02573-7

- Chan C, Fornacon-Wood I, Bayman N, et al. 104 Impact of introducing intensity modulated radiotherapy (IMRT) on curative intent radiotherapy and survival for lung cancer. Lung Cancer. 2022;165:S49. doi:10.1016/S0169-5002(22)00151-9

- Price J, Fornacon-Wood I, Thomson D, et al. PO-1001 The effect of switching to carboplatin chemo-RT for cycle 2 in cisplatin-ineligible HNSCC patients. Radiother Oncol. 2021;161:S833-S834. doi:10.1016/s0167-8140(21)07452-1

- Fornacon-Wood I, Mistry H, Johnson-Hart C, O'Connor JPB, Faivre-Finn C, Price G. PD-0776 A Bayesian approach to evaluate the impact of change in IGRT protocol using real world data. Radiother Oncol. 2021;161:S608. doi:

10.1016/s0167-8140(21)07055-9

- Croxford W, Banfill K, Fornacon-Wood I, et al. PO-1198 Changes in radical radiotherapy for lung cancer patients in the UK during the COVID-19 pandemic. Radiother Oncol. 2021;161:S994-S995. doi:10.1016/s0167-8140(21)07649-0

- Fornacon-Wood I, O'Connor J, Faivre-Finn C, Price G. PH-0652: Standardization influences repeatability and prognostic value of radiomic features. Radiother Oncol. 2020;152:S362-S363. doi:10.1016/s0167-8140(21)00674-5

- Ackermann C, Fornacon-Wood I, Tay R, et al. P1.04-44 Radiomics for Predicting Response to First-Line Anti-PD1 Therapy in Advanced NSCLC. J Thorac Oncol. 2019;14(10):S457-S458. doi:10.1016/j.jtho.2019.08.947

10.1.3 Non peer-reviewed

- Fornacon-Wood I, Davey A. Statistical resources for medical physics and beyond. European Society for Radiotherapy and Oncology (ESTRO) Physics Newsletter, May 2021.

10.2 Presentations

- ESTRO, May 2023. Oral presentation 'Impact of the COVID-19 pandemic on patient outcomes after curative-intent radiotherapy in the UK'.

- Manchester Cancer Research Centre, March 2023. Oral presentation 'Investigating the potential of real-world data to improve outcomes for patients with lung cancer.'

- ESTRO, August 2021. Poster presentation 'A Bayesian approach to evaluate the impact of change in IGRT protocol using real world data'.

- Manchester Cancer Research Centre, November 2020. Oral presentation 'Improving patient outcomes in radiotherapy using real world data.'

- ESTRO, November 2020. Poster presentation 'Standardization influences repeatability and prognostic value of radiomic features'.

10.3 Awards

- Awarded an Institute of Physics and Engineering in Medicine students and trainee travel grant of £300 to present work at ESTRO 2020.
- Nominated for Best Poster Award in the Physics category at ESTRO 2020.

Bibliography

1. Connell, P. P. & Hellman, S. Advances in radiotherapy and implications for the next century: A historical perspective. *Cancer Research* **69,** 383–392 (2009).
2. Coutard, H. PRINCIPLES OF X RAY THERAPY OF MALIGNANT DISEASES. *The Lancet* **224,** 1–8 (1934).
3. Schaue, D. & Mcbride, W. H. Opportunities and challenges of radiotherapy for treating cancer. *Nature Reviews Clinical Oncology* **12,** 527–540 (June 2015).
4. Borras, J. M. *et al.* The optimal utilization proportion of external beam radiotherapy in European countries: An ESTRO-HERO analysis. *Radiotherapy and Oncology* **116,** 38–44 (2015).
5. Delaney, G., Jacob, S., Featherstone, C. & Barton, M. The role of radiotherapy in cancer treatment: Estimating optimal utilization from a review of evidence-based clinical guidelines. *Cancer* **104,** 1129–1137 (2005).
6. Department of Health Cancer Policy Team. *Radiotherapy Services in England 2012* tech. rep. (2012), 58.
7. Bryant, A. K., Banegas, M. P., Martinez, M. E., Mell, L. K. & Murphy, J. D. Trends in radiation therapy among cancer survivors in the United States, 2000-2030. *Cancer Epidemiology Biomarkers and Prevention* **26,** 963–970 (June 2017).
8. Baskar, R., Dai, J., Wenlong, N., Yeo, R. & Yeoh, K. W. Biological response of cancer cells to radiation treatment. *Frontiers in Molecular Biosciences* **1** (Nov. 2014).
9. Bernier, J., Hall, E. J. & Giaccia, A. Radiation oncology: A century of achievements. *Nature Reviews Cancer* **4,** 737–747 (Sept. 2004).

10. Van Herk, M. Errors and Margins in Radiotherapy. *Seminars in Radiation Oncology* **14,** 52–64 (2004).

11. Keall, P. J. *et al.* The management of respiratory motion in radiation oncology report of AAPM Task Group 76. *Medical Physics* **33,** 3874–3900 (2006).

12. Jeraj, M. & Robar, V. Multilead collimator in radiotherapy. *Radiology and Oncology* **38,** 235–240 (2004).

13. Taylor, A. & Powell, M. E. B. Intensity-modulated radiotherapy - What is it? *Cancer Imaging* **4,** 68–73 (2004).

14. Purdy, J. A. Intensity-modulated radiotherapy: Current status and issues of interest. *International Journal of Radiation Oncology Biology Physics* **51,** 880–914 (Nov. 2001).

15. National Radiotherapy Advisory Group. *Radiotherapy: developing a world class service for England* tech. rep. (2007).

16. Jaffray, D. A. Image-guided radiotherapy: from current concept to future perspectives. *Nature Reviews Clinical Oncology* **9,** 688–699 (2012).

17. Pignol, J. P. & Janus, C. The evaluation of innovation in radiation oncology - What can we do and what should we do? *Acta Oncologica* **54,** 1251–1253 (Oct. 2015).

18. Arbea, L., Ramos, L. I., Martínez-Monge, R., Moreno, M. & Aristu, J. Intensity-modulated radiation therapy (IMRT) vs. 3D conformal radiotherapy (3DCRT) in locally advanced rectal cancer (LARC): Dosimetric comparison and clinical implications. *Radiation Oncology* **5,** 1–9 (Feb. 2010).

19. Bree, I. d., van Hinsberg, M. G. & van Veelen, L. R. High-dose radiotherapy in inoperable nonsmall cell lung cancer: Comparison of volumetric modulated arc therapy, dynamic IMRT and 3D conformal radiotherapy. *Medical Dosimetry* **37,** 353–357 (2012).

20. Van Loon, J., Grutters, J. & Macbeth, F. Evaluation of novel radiotherapy technologies: What evidence is needed to assess their clinical and cost effectiveness, and how should we get it? *The Lancet Oncology* **13,** e169–e177 (Apr. 2012).

21. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **71,** 209–249 (May 2021).

22. Cancer Research UK. *Lung cancer survival* https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer#heading-Two%20https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/survival. *Date accessed: 29/06/2023.*

23. Antonia, S. J. *et al.* Overall Survival with Durvalumab after Chemoradiotherapy in Stage III NSCLC. *New England Journal of Medicine* **379,** 2342–2350 (2018).

24. Gandhi, L. *et al.* Pembrolizumab plus Chemotherapy in Metastatic Non–Small-Cell Lung Cancer. *New England Journal of Medicine* **378,** 2078–2092 (May 2018).

25. Garon, E. B. *et al.* Five-year overall survival for patients with advanced non-small-cell lung cancer treated with pembrolizumab: Results from the phase i KEYNOTE-001 study. *Journal of Clinical Oncology* **37,** 2518–2527 (Oct. 2019).

26. Heuvers, M. E., Hegmans, J. P., Stricker, B. H. & Aerts, J. G. Improving lung cancer survival; time to move on. *BMC Pulmonary Medicine* **12,** 1–4 (Dec. 2012).

27. Corner, J., Hopkinson, J., Fitzsimmons, D., Barclay, S. & Muers, M. Is late diagnosis of lung cancer inevitable? Interview study of patients' recollections of symptoms before diagnosis. *Thorax* **60,** 314–319 (Apr. 2005).

28. NHS England. *Hundreds of people diagnosed with cancer early through life-saving NHS lung checks* https://www.england.nhs.uk/2022/04/hundreds-of-people-diagnosed-with-cancer-early-through-life-saving-nhs-lung-checks/. *Date accessed: 12/01/2023.*

29. Peake, M. D. Deprivation, distance and death in lung cancer. *Thorax* **70,** 108–109 (Feb. 2015).

30. Payne, N. W. *et al.* Socio-economic deprivation and cancer incidence in England: Quantifying the role of smoking. *PloS one* **17,** e0272202 (Sept. 2022).

31. Senan, S., Palma, D. A. & Lagerwaard, F. J. Stereotactic ablative radiotherapy for stage I NSCLC: Recent advances and controversies. *Journal of Thoracic Disease* **3,** 189–196 (2011).

32. Aupérin, A. *et al.* Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non - small-cell lung cancer. *Journal of Clinical Oncology* **28,** 2181–2190 (May 2010).

33. Lim, E. *et al.* Guidelines on the radical management of patients with lung cancer. *Thorax* **65** (2010).

34. Wang, S. *et al.* Survival changes in patients with small cell lung cancer and disparities between different sexes, socioeconomic statuses and ages. *Scientific Reports* **7** (2017).

35. Govindan, R. *et al.* Changing epidemiology of small-cell lung cancer in the United States over the last 30 years: Analysis of the surveillance, epidemiologic, and end results database. *Journal of Clinical Oncology* **24,** 4539–4544 (2006).

36. Aupérin, A. *et al.* Prophylactic Cranial Irradiation for Patients with Small-Cell Lung Cancer in Complete Remission. *New England Journal of Medicine* **341,** 476–484 (1999).

37. Dingemans, A. M. *et al.* Small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* **32,** 839–853 (July 2021).

38. Unger, J. M., Hershman, D. L., Fleury, M. E. & Vaidya, R. Association of Patient Comorbid Conditions with Cancer Clinical Trial Participation. *JAMA Oncology* **5,** 326–333 (Mar. 2019).

39. Hutchins, L. F., Unger, J. M., Crowley, J. J., Coltman, C. A. & Albain, K. S. Underrepresentation of Patients 65 Years of Age or Older in Cancer-Treatment Trials. *New England Journal of Medicine* **341,** 2061–2067 (Dec. 1999).

40. U.S. Food and Drug Administration. *Framework for FDA's Real-World Evidence Program* tech. rep. (2018).

41. Aminpour, F., Sadoughi, F. & Ahamdi, M. Utilization of open source electronic health record around the world: A systematic review. *Journal of Research in Medical Sciences* **19,** 57–64 (2014).

42. Robertson, A. *et al.* Implementation and adoption of nationwide electronic health records in secondary care in England: Qualitative analysis of interim results from a prospective national evaluation. *BMJ (Online)* **341,** 872 (Sept. 2010).

43. NHS England. *The NHS Long Term Plan* tech. rep. (2019).

44. Cowie, M. R. *et al.* Electronic health records to facilitate clinical research. *Clinical Research in Cardiology* **106,** 1 (Jan. 2017).

45. O'Connor, J. P. Rethinking the role of clinical imaging. *eLife* **6,** e30563 (2017).

46. Lambin, P. *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer* **48,** 441–446 (2012).

47. Gillies, R. J., Anderson, A. R., Gatenby, R. A. & Morse, D. L. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clinical Radiology* **65,** 517–521 (2010).

48. Traverso, A., Wee, L., Dekker, A. & Gillies, R. Repeatability and reproducibility of radiomic features: A systematic review. *International Journal of Radiation Oncology*Biology*Physics* **102,** 1143–1158 (Nov. 2018).

49. Ger, R. B. *et al.* Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies. *Scientific Reports* **8,** 13047 (Aug. 2018).

50. Mackin, D. *et al.* Measuring computed tomography scanner variability of radiomics features. *Investigative Radiology* **50,** 757–765 (Nov. 2015).

51. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. *Image biomarker standardisation initiative* http://arxiv.org/abs/1612.07003%0Ahttp://dx.doi.org/10.1148/radiol.2020191145. *arXiv preprint 1612.07003*

52. Zwanenburg, A. *et al.* The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295,** 328–338 (May 2020).

53. De Lusignan, S., Crawford, L. & Munro, N. Creating and using real-world evidence to answer questions about clinical effectiveness. *Journal of Innovation in Health Informatics* **22,** 368–373 (July 2015).

54. Unger, J. M. *et al.* Comparison of survival outcomes among cancer patients treated in and out of clinical trials. *Journal of the National Cancer Institute* **106** (2014).

55. Rothwell, P. M. Factors That Can Affect the External Validity of Randomised Controlled Trials. *PLoS Clinical Trials* **1,** e9 (May 2006).

56. Apisarnthanarax, S. *et al.* Applicability of randomized trials in radiation oncology to standard clinical practice. *Cancer* **119,** 3092–3099 (Aug. 2013).

57. Crockett, C. *et al.* The Routine Clinical Implementation of Electronic Patient-reported Outcome Measures (ePROMs) at The Christie NHS Foundation Trust. *Clinical Oncology* **33,** 761–764 (2021).

58. SyTrue. *Why unstructured data holds the key to intelligent healthcare systems* https://hitconsultant.net/2015/03/31/tapping-unstructured-data-healthcares-biggest-hurdle-realized/. *Date accessed: 12/01/2023.*

59. Kang, H. The prevention and handling of the missing data. *Korean Journal of Anesthesiology* **64,** 402–406 (May 2013).

60. Assimon, M. M. Confounding in Observational Studies Evaluating the Safety and Effectiveness of Medical Treatments. *Kidney360* **2,** 1156–1159 (2021).

61. Patsopoulos, N. A. A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience* **13,** 217–224 (2011).

62. The RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19. *New England Journal of Medicine* **384,** 693–704 (2021).

63. Goldacre, B., Morley, J. & Hamilton, N. *Better, broader, safer: Using health data for research and analysis* tech. rep. (2022).

64. Department of Health and Social Care. *Data saves lives: reshaping health and social care with data* https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data. *Date accessed: 23/10/2022.*

65. National Institute for Health and Care Excellence. *NICE real-world evidence framework* tech. rep. (2022).

66. Lambin, P. *et al.* 'Rapid Learning health care in oncology' - An approach towards decision support systems enabling customised radiotherapy. *Radiotherapy and Oncology* **109,** 159–164 (Oct. 2013).

67. Friedman, C. P. *et al.* Toward a science of learning systems: A research agenda for the high-functioning Learning Health System. *Journal of the American Medical Informatics Association* **22,** 43–50 (Jan. 2015).

68. Wysham, N. G. *et al.* Development and Refinement of a Learning Health Systems Training Program. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* **4,** 23 (Nov. 2016).

69. Bosdriesz, J. R. *et al.* Evidence-based medicine—When observational studies are better than randomized controlled trials. *Nephrology* **25,** 737–743 (Oct. 2020).

70. Price, G. *et al.* Learning healthcare systems and rapid learning in radiation oncology: Where are we and where are we going? *Radiotherapy and Oncology* **164,** 183–195 (2021).

71. Axelrod, D. A. & Hayward, R. *Nonrandomized interventional study designs (quasi-experimental designs) in Clinical Research Methods for Surgeons* 63–76 (Humana Press, 2007).

72. Fornacon-Wood, I., Faivre-Finn, C., O'Connor, J. P. & Price, G. J. Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype. *Lung Cancer* **146,** 197–208 (Aug. 2020).

73. Fornacon-Wood, I. *et al.* Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *European Radiology* **30,** 6241–6250 (Nov. 2020).

74. Fornacon-Wood, I. *et al.* Impact of Introducing Intensity Modulated Radiotherapy on Curative Intent Radiotherapy and Survival for Lung Cancer. *Frontiers in Oncology* **12,** 835844 (May 2022).

75. Fornacon-Wood, I. *et al.* Impact of the COVID-19 Pandemic on Outcomes for Patients with Lung Cancer Receiving Curative-intent Radiotherapy in the UK. *Clinical oncology (Royal College of Radiologists (Great Britain))* **35,** e593–e600 (Oct. 2023).

76. Banfill, K. *et al.* Changes in the Management of Patients having Radical Radiotherapy for Lung Cancer during the First Wave of the COVID-19 Pandemic in the UK. *Clinical Oncology* **34,** 19–27 (Jan. 2022).

77. Fornacon-Wood, I. *et al.* Understanding the Differences Between Bayesian and Frequentist Statistics. *International Journal of Radiation Oncology Biology Physics* **112,** 1076–1082 (Apr. 2022).

78. Chowdhry, A. K. *et al.* In Regard to Fornacon-Wood et al. [Letter to the Editor]. *International Journal of Radiation Oncology*Biology*Physics* **115,** 249–250 (Jan. 2023).

79. Fornacon-Wood, I., Mistry, H., Price, G. J., Faivre-Finn, C. & O'Connor, J. P. In Reply to Chowdhry et al. [Letter to the Editor]. *International Journal of Radiation Oncology*Biology*Physics* **115,** 250–251 (Jan. 2023).

80. Fornacon-Wood, I. *et al.* Bayesian methods provide a practical real-world evidence framework for evaluating the impact of changes in radiotherapy. *Radiotherapy and Oncology* **176,** 53–58 (Nov. 2022).

81. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **68,** 394–424 (Nov. 2018).

82. National Cancer Institute. *SEER Cancer Statistics Review, 1975 - 2011* https://seer.cancer.gov/archive/csr/1975_2011/. *Date accessed: 09/09/2019.*

83. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians* **69,** 7–34 (2019).

84. Schilsky, R. L. Personalized medicine in oncology: The future is now. *Nature Reviews Drug Discovery* **9,** 363–366 (May 2010).

85. Salem, A. *et al.* Targeting hypoxia to improve non–small cell lung cancer outcome. *Journal of the National Cancer Institute* **110,** 14–29 (2018).

86. De Langen, A. J. *et al.* Monitoring Response to Antiangiogenic Therapy in Non-Small Cell Lung Cancer Using Imaging Markers Derived from PET and Dynamic Contrast-Enhanced MRI. *Journal of Nuclear Medicine* **52,** 48–55 (2011).

87. Nishino, M., Hatabu, H., Johnson, B. E. & McLoud, T. C. State of the Art: Response Assessment in Lung Cancer in the Era of Genomic Medicine. *Radiology* **271,** 6–27 (2014).

88. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **278,** 563–577 (2016).

89. Aerts, H. J. The potential of radiomic-based phenotyping in precision medicine: a review. *JAMA Oncology* **2,** 1636–1642 (Dec. 2016).

90. O'Connor, J. P. *et al.* Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology* **14,** 169–186 (2017).

91. Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications* **5,** 4006 (June 2014).

92. Coroller, T. P. *et al.* CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology* **114,** 345–350 (Mar. 2015).

93. Bianconi, F. *et al.* Evaluation of shape and textural features from CT as prognostic biomarkers in non-small cell lung cancer. *Anticancer Research* **38,** 2155–2160 (Apr. 2018).

94. De Jong, E. E. *et al.* Applicability of a prognostic CT-based radiomic signature model trained on stage I-III non-small cell lung cancer in stage IV non-small cell lung cancer. *Lung Cancer* **124,** 6–11 (Oct. 2018).

95. Lee, G. *et al.* Comprehensive Computed Tomography Radiomics Analysis of Lung Adenocarcinoma for Prognostication. *The Oncologist* **23,** 806–813 (July 2018).

96. He, B. *et al.* A biomarker basing on radiomics for the prediction of overall survival in non-small cell lung cancer patients. *Respiratory Research* **19,** 199 (Oct. 2018).

97. Starkov, P. *et al.* The use of texture-based radiomics CT analysis to predict outcomes in early-stage non-small cell lung cancer treated with stereotactic ablative radiotherapy. *British Journal of Radiology* **92,** 20180228 (Feb. 2019).

98. Yang, L. *et al.* Development of a radiomics nomogram based on the 2D and 3D CT features to predict the survival of non-small cell lung cancer patients. *European Radiology* **29,** 2196–2206 (May 2019).

99. Wang, L. *et al.* Integrative nomogram of CT imaging, clinical, and hematological features for survival prediction of patients with locally advanced non-small cell lung cancer. *European Radiology* **29,** 2958–2967 (June 2019).

100. Shi, L. *et al.* Cone-beam computed tomography-based delta-radiomics for early response assessment in radiotherapy for locally advanced lung cancer. *Physics in medicine and biology* **65,** 15009 (Jan. 2019).

101. Van Timmeren, J. E. *et al.* Longitudinal radiomics of cone-beam CT images from non-small cell lung cancer patients: Evaluation of the added prognostic value for overall survival and locoregional recurrence. *Radiotherapy and Oncology* **136,** 78–85 (July 2019).

102. Huang, L. *et al.* Assessment of a Radiomic Signature Developed in a General NSCLC Cohort for Predicting Overall Survival of ALK-Positive Patients With Different Treatment Types. *Clinical Lung Cancer* **20,** e638–e651 (Nov. 2019).

103. Van Timmeren, J. E. *et al.* Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. *Radiotherapy and Oncology* **123,** 363–369 (June 2017).

104. Franceschini, D. *et al.* A radiomic approach to predicting nodal relapse and disease-specific survival in patients treated with stereotactic body radiation therapy for early-stage non-small cell lung cancer. *Strahlentherapie und Onkologie* (Nov. 2019).

105. Grossmann, P. *et al.* Defining the biological basis of radiomic phenotypes in lung cancer. *eLife* **6,** 1–22 (July 2017).

106. Yu, W. *et al.* Development and Validation of a Predictive Radiomics Model for Clinical Outcomes in Stage I Non-small Cell Lung Cancer. *International Journal of Radiation Oncology Biology Physics* **102,** 1090–1097 (Nov. 2017).

107. Chaddad, A., Desrosiers, C., Toews, M. & Abdulkarim, B. Predicting survival time of lung cancer patients using radiomic analysis. *Oncotarget* **8,** 104393–104407 (Nov. 2017).

108. Fave, X. *et al.* Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. *Scientific Reports* **7,** 588 (Apr. 2017).

109. Li, Q. *et al.* CT imaging features associated with recurrence in non-small cell lung cancer patients after stereotactic body radiotherapy. *Radiation Oncology* **12,** 158 (Sept. 2017).

110. Li, Q. *et al.* Imaging features from pretreatment CT scans are associated with clinical outcomes in nonsmall-cell lung cancer patients treated with stereotactic body radiotherapy. *Medical Physics* **44,** 4341–4349 (Aug. 2017).

111. Tang, C. *et al.* Development of an Immune-Pathology Informed Radiomics Model for Non-Small Cell Lung Cancer. *Scientific Reports* **8,** 1922 (Jan. 2018).

112. Ferreira Junior, J. R., Koenigkam-Santos, M., Cipriano, F. E. G., Fabro, A. T. & Azevedo-Marques, P. M. d. Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. *Computer Methods and Programs in Biomedicine* **159,** 23–30 (June 2018).

113. Yang, X. *et al.* A new approach to predict lymph node metastasis in solid lung adenocarcinoma: A radiomics nomogram. *Journal of Thoracic Disease* **10,** S807–S819 (Apr. 2018).

114. Zhong, Y. *et al.* Radiomics approach to prediction of occult mediastinal lymph node metastasis of lung adenocarcinoma. *American Journal of Roentgenology* **211,** 109–113 (July 2018).

115. Lafata, K. J. *et al.* Association of pre-treatment radiomic features with lung cancer recurrence following stereotactic body radiation therapy. *Physics in Medicine and Biology* **64** (2019).

116. Akinci D'Antonoli, T. *et al.* CT Radiomics Signature of Tumor and Peritumoral Lung Parenchyma to Predict Nonsmall Cell Lung Cancer Postsurgical Recurrence Risk. *Academic Radiology* **27,** 497–507 (July 2020).

117. He, L. *et al.* Radiomics-based predictive risk score: A scoring system for preoperatively predicting risk of lymph node metastasis in patients with resectable non-small cell lung cancer. *Chinese Journal of Cancer Research* **31,** 641–652 (Aug. 2019).

118. Xu, X. *et al.* Application of radiomics signature captured from pretreatment thoracic CT to predict brain metastases in stage III/IV ALK-positive non-small cell lung cancer patients. *Journal of thoracic disease* **11,** 4516–4528 (Nov. 2019).

119. Ferreira-Junior, J. R. *et al.* CT-based radiomics for prediction of histologic subtype and metastatic disease in primary malignant lung neoplasms. *International Journal of Computer Assisted Radiology and Surgery* **15,** 163–172 (Jan. 2020).

120. Cong, M. *et al.* Development of a predictive radiomics model for lymph node metastases in pre-surgical CT-based stage IA non-small cell lung cancer. *Lung Cancer* **139,** 73–79 (Jan. 2020).

121. Mattonen, S. A. *et al.* Detection of Local Cancer Recurrence after Stereotactic Ablative Radiation Therapy for Lung Cancer: Physician Performance Versus Radiomic Assessment. *International Journal of Radiation Oncology Biology Physics* **94,** 1121–1128 (Apr. 2016).

122. Huynh, E. *et al.* CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiotherapy and Oncology* **120,** 258–266 (Aug. 2016).

123. Huynh, E. *et al.* Associations of radiomic data extracted from static and respiratory-gated CT scans with disease recurrence in lung cancer patients treated with SBRT. *PLoS ONE* **12,** e0169172 (2017).

124. Dou, T. H., Coroller, T. P., van Griethuysen, J. J., Mak, R. H. & Aerts, H. J. Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC. *PLoS ONE* **13,** e0206108 (2018).

125. Coroller, T. P. *et al.* Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiotherapy and Oncology* **119,** 480–486 (June 2016).

126. Huang, Y. *et al.* Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non—Small Cell Lung Cancer. *Radiology* **281,** 947–957 (Dec. 2016).

127. Song, J. *et al.* Association between tumor heterogeneity and progression-free survival in non-small cell lung cancer patients with EGFR mutations undergoing tyrosine kinase inhibitors therapy. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS,* 1268–1271 (Aug. 2016).

128. Coroller, T. P. *et al.* Radiomic-Based Pathological Response Prediction from Primary Tumors and Lymph Nodes in NSCLC. *Journal of Thoracic Oncology* **12,** 467–476 (Mar. 2017).

129. Tunali, I. *et al.* Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: An early report. *Lung Cancer* **129,** 75–79 (Mar. 2019).

130. Yuan, M. *et al.* Prognostic Impact of the Findings on Thin-Section Computed Tomography in stage 1 lung adenocarcinoma with visceral pleural invasion. *Scientific Reports* **8,** 4743 (Mar. 2018).

131. Yang, M. *et al.* Imaging phenotype using radiomics to predict dry pleural dissemination in non-small cell lung cancer. *Annals of translational medicine* **7,** 259 (June 2019).

132. Moran, A., Daly, M. E., Yip, S. S. & Yamamoto, T. Radiomics-based Assessment of Radiation-induced Lung Injury After Stereotactic Body Radiotherapy. *Clinical Lung Cancer* **18,** e425–e431 (Nov. 2017).

133. Krafft, S. P. *et al.* The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis. *Medical Physics* **45,** 5317–5324 (Nov. 2018).

134. Aerts, H. J. *et al.* Defining a Radiomic Response Phenotype: A Pilot Study using targeted therapy in NSCLC. *Scientific Reports* **6,** 33860 (Sept. 2016).

135. Rios Velazquez, E. *et al.* Somatic mutations drive distinct imaging phenotypes in lung cancer. *Cancer Research* **77,** 3922–3930 (July 2017).

136. Mei, D., Luo, Y., Wang, Y. & Gong, J. CT texture analysis of lung adenocarcinoma: Can Radiomic features be surrogate biomarkers for EGFR mutation statuses. *Cancer Imaging* **18,** 52 (Dec. 2018).

137. Digumarthy, S. R., Padole, A. M., Gullo, R. L., Sequist, L. V. & Kalra, M. K. Can CT radiomic analysis in NSCLC predict histology and EGFR mutation status? *Medicine* **98,** e13963 (Jan. 2019).

138. Jia, T. Y. *et al.* Identifying EGFR mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling. *European Radiology* **29,** 4742–4750 (Sept. 2019).

139. Li, S., Ding, C., Zhang, H., Song, J. & Wu, L. Radiomics for the prediction of EGFR mutation subtypes in non-small cell lung cancer. *Medical physics* **46,** 4545–4552 (Oct. 2019).

140. Tu, W. *et al.* Radiomics signature: A potential and incremental predictor for EGFR mutation status in NSCLC patients, comparison with CT morphology. *Lung cancer (Amsterdam, Netherlands)* **132,** 28–35 (June 2019).

141. Yang, X. *et al.* Computed Tomography-Based Radiomics Signature: A Potential Indicator of Epidermal Growth Factor Receptor Mutation in Pulmonary Adenocarcinoma Appearing as a Subsolid Nodule. *The oncologist* **24,** e1156–e1164 (Nov. 2019).

142. Wang, X. *et al.* Decoding tumor mutation burden and driver mutations in early stage lung adenocarcinoma using CT-based radiomics signature. *Thoracic cancer* **10,** 1904–1912 (Oct. 2019).

143. Bak, S. H. *et al.* Imaging genotyping of functional signaling pathways in lung squamous cell carcinoma using a radiomics approach. *Scientific Reports* **8,** 1–9 (Feb. 2018).

144. Gu, Q. *et al.* Machine learning-based radiomics strategy for prediction of cell proliferation in non-small cell lung cancer. *European journal of radiology* **118,** 32–37 (Sept. 2019).

145. Song, S. H. *et al.* Imaging Phenotyping Using Radiomics to Predict Micropapillary Pattern within Lung Adenocarcinoma. *Journal of Thoracic Oncology* **12,** 624–632 (Apr. 2017).

146. Chen, X. *et al.* A Radiomics Signature in Preoperative Predicting Degree of Tumor Differentiation in Patients with Non–small Cell Lung Cancer. *Academic Radiology* **25,** 1548–1555 (Dec. 2018).

147. She, Y. *et al.* The predictive value of CT-based radiomics in differentiating indolent from invasive lung adenocarcinoma in patients with pulmonary nodules. *European radiology* **28,** 5121–5128 (Dec. 2018).

148. Yang, B. *et al.* Radiomic signature: a non-invasive biomarker for discriminating invasive and non-invasive cases of lung adenocarcinoma. *Cancer management and research* **11,** 7825–7834 (2019).

149. Patil, R., Mahadevaiah, G. & Dekker, A. An Approach Toward Automatic Classification of Tumor Histopathology of Non–Small Cell Lung Cancer Based on Radiomic Features. *Tomography* **2,** 374–377 (Dec. 2016).

150. Wu, W. *et al.* Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. *Frontiers in Oncology* **6,** 1–11 (2016).

151. Zhu, X. *et al.* Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. *European radiology* **28,** 2772–2778 (July 2018).

152. E, L. *et al.* Radiomics for Classifying Histological Subtypes of Lung Cancer Based on Multiphasic Contrast-Enhanced Computed Tomography. *Journal of Computer Assisted Tomography* **43,** 300–306 (2019).

153. Liu, J. *et al.* Multi-subtype classification model for non-small cell lung cancer based on radiomics: SLS model. *Medical Physics* **46,** 3091–3100 (July 2019).

154. Zhou, B., Xu, J., Tian, Y., Yuan, S. & Li, X. Correlation between radiomic features based on contrast-enhanced computed tomography images and Ki-67 proliferation index in lung cancer: A preliminary study. *Thoracic Cancer* **9,** 1235–1240 (Oct. 2018).

155. FDA Device Classification. *QuantX* https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?ID=DEN170022. *Date accessed: 19/02/2020.*

156. Feedback Medical. *TexRAD* https://fbkmed.com/texrad-landing-2/. *Date accessed: 22/02/2020.*

157. Kim, Y. J., Lee, H. J., Kim, K. G. & Lee, S. H. The Effect of CT Scan Parameters on the Measurement of CT Radiomic Features: A Lung Nodule Phantom Study. *Computational and Mathematical Methods in Medicine* **2019,** 8790694 (2019).

158. Zhovannik, I. *et al.* Learning from scanners: Bias reduction and feature correction in radiomics. *Clinical and Translational Radiation Oncology* **19,** 33–38 (Nov. 2019).

159. Tunali, I. *et al.* Stability and reproducibility of computed tomography radiomic features extracted from peritumoral regions of lung cancer lesions. *Medical Physics* **46,** 5075–5085 (Sept. 2019).

160. Mahon, R. N., Ghita, M., Hugo, G. D. & Weiss, E. ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. *Physics in Medicine and Biology* **65,** 15010 (Jan. 2020).

161. Fave, X. *et al.* Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? *Med. Phys* **42,** 6784–6797 (Dec. 2015).

162. Hepp, T. *et al.* Effects of simulated dose variation on contrast-enhanced CT-based radiomic analysis for Non-Small Cell Lung Cancer. *European journal of radiology* **124,** 108804 (Jan. 2020).

163. Kakino, R. *et al.* Comparison of radiomic features in diagnostic CT images with and without contrast enhancement in the delayed phase for NSCLC patients. *Physica Medica* **69,** 176–182 (Jan. 2020).

164. Mackin, D. *et al.* Effect of tube current on computed tomography radiomic features. *Scientific Reports* **8,** 2354 (Feb. 2018).

165. Mahmood, U., Apte, A. P., Deasy, J. O., Schmidtlein, C. R. & Shukla-Dave, A. Investigating the robustness neighborhood gray tone difference matrix and gray level co-occurrence matrix radiomic features on clinical computed tomography systems using anthropomorphic phantoms: Evidence from a multivendor study. *Journal of Computer Assisted Tomography* **41,** 995–1001 (2017).

166. Midya, A., Chakraborty, J., Gönen, M., Do, R. K. G. & Simpson, A. L. Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. *Journal of Medical Imaging* **5,** 1 (2018).

167. Yasaka, K. *et al.* Precision of quantitative computed tomography texture analysis using image filtering. *Medicine (United States)* **96,** e6993 (2017).

168. Lafata, K. *et al.* Spatialoral variability of radiomic features and its effect on the classification of lung cancer histology. *Physics in Medicine and Biology* **63,** 225003 (Nov. 2018).

169. Du, Q. *et al.* Radiomic feature stability across 4D respiratory phases and its impact on lung tumor prognosis prediction. *PloS one* **14,** e0216480 (2019).

170. Larue, R. T. *et al.* Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncologica* **56,** 1544–1553 (Nov. 2017).

171. Mackin, D. *et al.* Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS ONE* **12,** e0178524 (Sept. 2017).

172. Shafiq-Ul-Hassan, M. *et al.* Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Medical physics* **44,** 1050–1062 (Mar. 2017).

173. Lu, L., Ehmke, R. C., Schwartz, L. H. & Zhao, B. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PloS one* **11,** e0166550 (2016).

174. Shafiq-Ul-Hassan, M. *et al.* Voxel size and gray level normalization of CT radiomic features in lung cancer. *Scientific Reports* **8,** 10545 (July 2018).

175. Li, Y. *et al.* CT Slice Thickness and Convolution Kernel Affect Performance of a Radiomic Model for Predicting EGFR Status in Non-Small Cell Lung Cancer: A Preliminary Study. *Scientific Reports* **8,** 17913 (Dec. 2018).

176. Park, S. *et al.* Deep learning algorithm for reducing ct slice thickness: Effect on reproducibility of radiomic features in lung cancer. *Korean Journal of Radiology* **20,** 1431–1440 (Oct. 2019).

177. Shafiq-ul-Hassan, M. *et al.* Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra. *Journal of Medical Imaging* **5,** 1 (Jan. 2017).

178. Zhao, B. *et al.* Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific Reports* **6,** 23428 (Mar. 2016).

179. Haga, A. *et al.* Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis. *Radiological Physics and Technology* **11,** 27–35 (Mar. 2018).

180. Huang, Q. *et al.* Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status. *Journal of Medical Imaging* **5,** 1 (Jan. 2017).

181. Kalpathy-Cramer, J. *et al.* Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. *Tomography* **2,** 430–437 (Dec. 2016).

182. Owens, C. A. *et al.* Lung tumor segmentation methods: Impact on the uncertainty of radiomics features for non-small cell lung cancer. *PLoS ONE* **13,** 1–22 (Oct. 2018).

183. Parmar, C. *et al.* Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS ONE* **9,** e102107 (July 2014).

184. Pavic, M. *et al.* Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncologica* **57,** 1070–1074 (Aug. 2018).

185. Wang, H. Y. *et al.* The stability of imaging biomarkers in radiomics: A framework for evaluation. *Physics in Medicine and Biology* **64,** 165012 (Aug. 2019).

186. Fave, X. *et al.* Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. *Translational Cancer Research* **5,** 349–363 (2016).

187. Welch, M. L. *et al.* Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology* **130,** 2–9 (Jan. 2019).

188. Choi, W. *et al.* Technical Note: Identification of CT Texture Features Robust to Tumor Size Variations for Normal Lung Texture Analysis. *International Journal of Medical Physics, Clinical Engineering and Radiation Oncology* **07,** 330–338 (Aug. 2018).

189. Larue, R. T. *et al.* 4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers. *Radiotherapy and Oncology* **125,** 147–153 (Oct. 2017).

190. Zwanenburg, A. *et al.* Assessing robustness of radiomic features by image perturbation. *Scientific Reports* **9,** 614 (Jan. 2019).

191. Tanaka, S. *et al.* Investigation of thoracic four-dimensional CT-based dimension reduction technique for extracting the robust radiomic features. *Physica Medica* **58,** 141–148 (Feb. 2019).

192. Van Timmeren, J. E. *et al.* Test–Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography* **2,** 361–365 (Dec. 2016).

193. Haga, A. *et al.* Standardization of imaging features for radiomics analysis. *Journal of Medical Investigation* **66,** 35–37 (Feb. 2019).

194. Zhang, Y., Oikonomou, A., Wong, A., Haider, M. A. & Khalvati, F. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. *Scientific Reports* **7,** 46349 (Apr. 2017).

195. Sun, W., Jiang, M., Dang, J., Chang, P. & Yin, F. F. Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. *Radiation Oncology* **13,** 197 (Oct. 2018).

196. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific Reports* **5,** 13087 (Oct. 2015).

197. Bogowicz, M. *et al.* Post-radiochemotherapy PET radiomics in head and neck cancer – The influence of radiomics implementation on the reproducibility of local control tumor models. *Radiotherapy and Oncology* **125,** 385–391 (2017).

198. Foy, J. J., Armato, S. G. & Al-Hallaq, H. A. Effects of variability in radiomics software packages on classifying patients with radiation pneumonitis. *Journal of Medical Imaging* **7,** 1 (Jan. 2020).

199. Deist, T. M. *et al.* Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers. *Medical Physics* **45,** 3449–3459 (July 2018).

200. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *European Urology* **67,** 1142–1151 (2015).

201. Lambin, P. *et al.* Radiomics: The bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **14,** 749–762 (2017).

202. Vallières, M. *et al.* Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Scientific Reports* **7,** 10117 (2017).

203. Ger, R. B. *et al.* Effects of alterations in positron emission tomography imaging parameters on radiomics features. *Plos One* **14,** e0221877 (2019).

204. He, L. *et al.* Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Scientific Reports* **6,** 1–10 (2016).

205. Apte, A. P. *et al.* Technical Note: Extension of CERR for computational radiomics: A comprehensive MATLAB platform for reproducible radiomics research. *Medical Physics* **45,** 3713–3720 (Aug. 2018).

206. Zhang, L. *et al.* IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics. *Medical Physics* **42,** 1341–53 (2015).

207. Van Griethuysen, J. J. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Research* **77,** e104–e107 (Nov. 2017).

208. Nioche, C. *et al.* Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Research* **78,** 4786–4789 (Aug. 2018).

209. Court, L. E. *et al.* Computational resources for radiomics. *Translational Cancer Research* **5,** 340–348 (Aug. 2016).

210. Wilkinson, M. D. *et al.* Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018 (2016).

211. Foy, J. J. *et al.* Variation in algorithm implementation across radiomics software. *Journal of Medical Imaging* **5,** 044505 (2018).

212. Liang, Z. G. *et al.* Comparison of radiomics tools for image analyses and clinical prediction in nasopharyngeal carcinoma. *British Journal of Radiology* **92,** 20190271 (Oct. 2019).

213. Hatt, M., Vallieres, M., Visvikis, D. & Zwanenburg, A. IBSI: an international community radiomics standardization initiative. *Journal of Nuclear Medicine* **59,** 287 LP –287 (2018).

214. Faivre-Finn, C. *et al.* Concurrent once-daily versus twice-daily chemoradiotherapy in patients with limited-stage small-cell lung cancer (CONVERT): an open-label, phase 3, randomised, superiority trial. *The Lancet Oncology* **18,** 1116–1125 (2017).

215. Pfaehler, E., Zwanenburg, A., de Jong, J. R. & Boellaard, R. RACAT: An open source and easy to use radiomics calculator tool. *PLoS ONE* **14,** 1–26 (Feb. 2019).

216. Szczypiński, P. M., Strzelecki, M., Materka, A. & Klepaczko, A. MaZda-A software package for image texture analysis. *Computer Methods and Programs in Biomedicine* **94,** 66–76 (2009).

217. Fang, Y. H. D. *et al.* Development and evaluation of an open-source software package "cGITA" for quantifying tumor heterogeneity with molecular images. *BioMed Research International* **2014,** 248505 (2014).

218. Dinapoli, N. *et al.* Moddicom: A complete and easily accessible library for prognostic evaluations relying on image features. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* **2015,** 771–774 (2015).

219. Echegaray, S., Bakr, S., Rubin, D. L. & Napel, S. Quantitative Image Feature Engine (QIFE): an Open-Source, Modular Engine for 3D Quantitative Feature

Extraction from Volumetric Medical Images. *Journal of Digital Imaging* **31,** 403–414 (2018).

220. Götz, M., Nolden, M. & Maier-Hein, K. MITK Phenotyping: An open-source toolchain for image-based personalized medicine with radiomics. *Radiotherapy and Oncology* **131,** 108–111 (2019).

221. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* **15,** 155–163 (2016).

222. R Core Team. *R: A Language and Environment for Statistical Computing* https://www.r-project.org/. *Date accessed: 26/05/2023.*

223. Gamer, M., Lemon, J., Fellows, I. & Singh, P. *irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.* https://cran.r-project.org/package=irr. *Date accessed: 09/09/2019.*

224. Therneau, T. M. *A Package for Survival Analysis in R. R package version 3.1-12* https://cran.r-project.org/package=survival. *Date accessed: 26/05/2023.*

225. Yip, S. S. & Aerts, H. J. Applications and limitations of radiomics. *Physics in Medicine and Biology* **61,** R150–R166 (2016).

226. Amadasun, M. & King, R. Texural Features Corresponding to Texural Properties. *IEEE Transactions on Systems, Man and Cybernetics* **19,** 1264–1274 (1989).

227. Sun, C. & Wee, W. G. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing* **23,** 341–352 (Sept. 1983).

228. MacHtay, M. *et al.* Defining local-regional control and its importance in locally advanced non-small cell lung carcinoma. *Journal of Thoracic Oncology* **7,** 716–722 (2012).

229. Diwanji, T. P. *et al.* Advances in radiotherapy techniques and delivery for non-small cell lung cancer: Benefits of intensity-modulated radiation therapy, proton therapy, and stereotactic body radiation therapy. *Translational Lung Cancer Research* **6,** 131–147 (Apr. 2017).

230. Brown, S., Banfill, K., Aznar, M. C., Whitehurst, P. & Finn, C. F. The evolving role of radiotherapy in non-small cell lung cancer. *British Journal of Radiology* **92** (2019).

231. Adizie, J. B. *et al.* Stage III Non-small Cell Lung Cancer Management in England. *Clinical Oncology* **31,** 688–696 (Oct. 2019).

232. Chan, C., Lang, S., Rowbottom, C., Guckenberger, M. & Faivre-Finn, C. Intensity-modulated radiotherapy for lung cancer: Current status and future developments. *Journal of Thoracic Oncology* **9,** 1598–1608 (Nov. 2014).

233. Chun, S. G. *et al.* Impact of intensity-modulated radiation therapy technique for locally advanced non-small-cell lung cancer: A secondary analysis of the NRG oncology RTOG 0617 randomized clinical trial. *Journal of Clinical Oncology* **35,** 56–62 (Jan. 2017).

234. Johnson, C. *et al.* A method to combine target volume data from 3D and 4D planned thoracic radiotherapy patient cohorts for machine learning applications. *Radiotherapy and Oncology* **126,** 355–361 (Feb. 2018).

235. Ball, D. L. *et al.* The complex relationship between lung tumor volume and survival in patients with non-small cell lung cancer treated by definitive radiotherapy: A prospective, observational prognostic factor study of the Trans-Tasman Radiation Oncology Group (TROG 99.05). *Radiotherapy and Oncology* **106,** 305–311 (Mar. 2013).

236. Zhang, J. *et al.* Relationship between tumor size and survival in Non-Small-Cell Lung Cancer (NSCLC): An analysis of the Surveillance, Epidemiology, and End Results (SEER) registry. *Journal of Thoracic Oncology* **10,** 682–690 (Apr. 2015).

237. Mcdonald, F., Senan, S., Faivre-Finn, C. & Hatton, M. Curative Radiotherapy for Non-small Cell Lung Cancer: Practice Changing and Changing Practice? *Clinical Oncology* **31,** 678–680 (Oct. 2019).

238. The Royal College of Radiologists. *Radiotherapy for lung cancer RCR consensus statements* https://www.rcr.ac.uk/publication/radiotherapy-lung-cancer-rcr-consensus-statements. *Date accessed: 28/07/2021.*

239. Yom, S. S. *et al.* Initial Evaluation of Treatment-Related Pneumonitis in Advanced-Stage Non-Small-Cell Lung Cancer Patients Treated With Concurrent Chemotherapy and Intensity-Modulated Radiotherapy. *International Journal of Radiation Oncology Biology Physics* **68,** 94–102 (May 2007).

240. Ball, D. *et al.* Routine use of intensity-modulated radiotherapy for locally advanced non-small-cell lung cancer is neither choosing wisely nor personalized medicine. *Journal of Clinical Oncology* **35,** 1492 (May 2017).

241. Hu, X. *et al.* Is IMRT superior or inferior to 3DCRT in radiotherapy for NSCLC? A meta-analysis. *PLoS ONE* **11** (Apr. 2016).

242. Banfill, K. *et al.* Cardiac Toxicity of Thoracic Radiotherapy: Existing Evidence and Future Directions. *Journal of Thoracic Oncology* **16,** 216–227 (Feb. 2021).

243. Christodoulou, M., Bayman, N., McCloskey, P., Rowbottom, C. & Faivre-Finn, C. New radiotherapy approaches in locally advanced non-small cell lung cancer. *European Journal of Cancer* **50,** 525–534 (Feb. 2014).

244. Richards, M., Anderson, M., Carter, P., Ebert, B. L. & Mossialos, E. The impact of the COVID-19 pandemic on cancer care. *Nature Cancer* **1,** 565–567 (2020).

245. Spencer, K. *et al.* The impact of the COVID-19 pandemic on radiotherapy services in England, UK: a population-based study. *The Lancet Oncology* **22,** 309–320 (Mar. 2021).

246. Venkatesulu, B. P. *et al.* A Systematic Review and Meta-Analysis of Cancer Patients Affected by a Novel Coronavirus. *JNCI Cancer Spectrum* **5,** 1–11 (2021).

247. Liu, W. *et al.* What is the optimal radiotherapy utilization rate for lung cancer?—a systematic review. *Translational Lung Cancer Research* **8,** S163–S169 (Sept. 2019).

248. Delaney, G. P. & Barton, M. B. Evidence-based Estimates of the Demand for Radiotherapy. *Clinical Oncology* **27,** 70–76 (Feb. 2015).

249. Passaro, A. *et al.* Severity of COVID-19 in patients with lung cancer: evidence and challenges. *Journal for Immunotherapy of Cancer* **9,** 2266 (Mar. 2021).

250. Faivre-Finn, C. *et al.* Reduced Fractionation in Lung Cancer Patients Treated with Curative-intent Radiotherapy during the COVID-19 Pandemic. *Clinical Oncology* **32,** 481–489 (Aug. 2020).

251. Iyengar, P. *et al.* Accelerated Hypofractionated Image-Guided vs Conventional Radiotherapy for Patients with Stage II/III Non-Small Cell Lung Cancer and Poor Performance Status: A Randomized Clinical Trial. *JAMA Oncology* **7,** 1497 (2021).

252. Zeng, K. L. *et al.* Accelerated Hypofractionated Radiotherapy for Centrally Located Lung Tumours Not Suitable for Stereotactic Body Radiotherapy or Chemoradiotherapy. *Clinical Oncology* **35,** e173–e181 (Feb. 2023).

253. Brada, M., Forbes, H., Ashley, S. & Fenwick, J. Improving Outcomes in NSCLC: Optimum Dose Fractionation in Radical Radiotherapy Matters. *Journal of Thoracic Oncology* **17,** 532–543 (2022).

254. Maguire, J. *et al.* SOCCAR: A randomised phase II trial comparing sequential versus concurrent chemotherapy and radical hypofractionated radiotherapy in patients with inoperable stage III Non-Small Cell Lung Cancer and good performance status. *European Journal of Cancer* **50,** 2939–2949 (Nov. 2014).

255. O'Rourke, N., Roqué i Figuls, M., Farré Bernadó, N. & Macbeth, F. Concurrent chemoradiotherapy in non-small cell lung cancer. *Cochrane Database of Systematic Reviews* (June 2010).

256. NICE. *COVID-19 rapid guideline: Managing COVID-19* https://www.nice.org.uk/guidance/ng191/resources/covid19-rapid-guideline-managing-covid19-pdf-51035553326. *Date accessed: 26/05/2023.*

257. Pignon, J. P., Stewart, L. A. & Marino, P. Randomized trials of radiotherapy alone versus combined chemotherapy and radiotherapy in Stages IIIa and IIIb nonsmall cell lung cancer: A meta- analysis. *Cancer* **77,** 2413–2414 (1996).

258. Conibear, J. *et al.* The National Lung Cancer Audit: The Impact of COVID-19. *Clinical Oncology* **34,** 701–707 (2022).

259. Willink, R. & White, R. *Disentangling Classical and Bayesian Approaches to Uncertainty Analysis* tech. rep. (2011), 1–19.

260. Colquhoun, D. The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science* **4,** 171085 (Dec. 2017).

261. Goodman, S. A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology* **45,** 135–140 (2008).

262. Altman, D. G. & Bland, J. M. Absence of evidence is not evidence of absence. *BMJ* **311,** 485 (Aug. 1995).

263. O'Hagan, T. Dicing with the unknown. *Significance* **1,** 132–133 (Sept. 2004).

264. Ranstam, J. Why the P-value culture is bad and confidence intervals a better alternative. *Osteoarthritis and Cartilage* **20,** 805–808 (Aug. 2012).

265. Hoekstra, R., Morey, R. D., Rouder, J. N. & Wagenmakers, E. J. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review* **21,** 1157–1164 (2014).

266. Sim, J. & Reid, N. Statistical inference by confidence intervals: Issues of interpretation and utilization. *Physical Therapy* **79,** 186–195 (Feb. 1999).

267. Dinnes, J. *et al.* Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. *Cochrane Database of Systematic Reviews* **2021** (Mar. 2021).

268. Johnson-Hart, C. N., Price, G. J., Faivre-Finn, C., Aznar, M. C. & van Herk, M. Residual Setup Errors Towards the Heart After Image Guidance Linked With Poorer Survival in Lung Cancer Patients: Do We Need Stricter IGRT Protocols? *International Journal of Radiation Oncology Biology Physics* **102,** 434–442 (Oct. 2018).

269. Bürkner, P. C. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80,** 1–28 (2017).

270. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2016,** e55 (Apr. 2016).

271. Goodman, S. N. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine* **130,** 1005–1013 (June 1999).

272. Schonbrodt, F. D. & Wagenmakers, E.-j. Bayes factor design analysis : Planning for compelling evidence. *Psychonomic Bulletin and Review* **25,** 128–142 (2018).

273. Tyldesley, S. *et al.* Estimating the need for radiotherapy for lung cancer: An evidence-based, epidemiologic approach. *International Journal of Radiation Oncology Biology Physics* **49,** 973–985 (Mar. 2001).

274. Bradley, J. D. *et al.* Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial p. *The Lancet Oncology* **16,** 187–199 (Feb. 2015).

275. Antonia, S. J. *et al.* Durvalumab after Chemoradiotherapy in Stage III Non–Small-Cell Lung Cancer. *New England Journal of Medicine* **377,** 1919–1929 (Nov. 2017).

276. Burnet, N. G. *et al.* Methodological Considerations in the Evaluation of Radiotherapy Technologies. *Clinical Oncology* **24,** 707–709 (Dec. 2012).

277. Cook, J. A., Ramsaya, C. R. & Fayers, P. Statistical evaluation of learning curve effects in surgical trials. *Clinical Trials* **1,** 421–427 (Oct. 2004).

278. Bak, K., Dobrow, M. J., Hodgson, D. & Whitton, A. Factors affecting the implementation of complex and evolving technologies: Multiple case study of intensity-modulated radiation therapy (IMRT) in Ontario, Canada. *BMC Health Services Research* **11,** 178 (Dec. 2011).

279. *Characteristics of the Weibull Distribution* https://www.weibull.com/hotwire/issue14/relbasics14.htm. *Date accessed: 19/01/2019.*

280. Teshnizi, S. H. & Ayatollahi, S. M. T. Comparison of cox regression and parametric models: Application for assessment of survival of pediatric cases of acute leukemia in southern Iran. *Asian Pacific Journal of Cancer Prevention* **18,** 981–985 (Apr. 2017).

281. Igl, W. *Calculation of hazard ratios of parametric survival models in R - A tutorial* https://www.gnu.org/licenses/gpl-3.0.en.html. *Date accessed: 15/01/2021.*

282. Bürkner, P. C. Advanced Bayesian multilevel modeling with the R package brms. *R Journal* **10,** 395–411 (2018).

283. Johnson-Hart, C. *et al.* Impact of small residual setup errors after image guidance on heart dose and survival in non-small cell lung cancer treated with curative-intent radiotherapy. *Radiotherapy and Oncology* **152,** 177–182 (2020).

284. Grimshaw, J., Campbell, M., Eccles, M. & Steen, N. Experimental and quasi-experimental designs for evaluating guideline implementation strategies. *Family Practice* **17,** S11–S16 (Feb. 2000).

285. Brink, C. *et al.* Causal relation between heart irradiation and survival of lung cancer patients after radiotherapy. *Radiotherapy and Oncology* **172,** 126–133 (July 2022).

286. Zampieri, F. G., Casey, J. D., Shankar-Hari, M., Harrell, F. E. & Harhay, M. O. Using bayesian methods to augment the interpretation of critical care trials. an overview of theory and example reanalysis of the alveolar recruitment for acute respiratory distress syndrome trial. *American Journal of Respiratory and Critical Care Medicine* **203,** 543–552 (Mar. 2021).

287. Depaoli, S., Winter, S. D. & Visser, M. The Importance of Prior Sensitivity Analysis in Bayesian Statistics: Demonstrations Using an Interactive Shiny App. *Frontiers in Psychology* **11,** 3271 (Nov. 2020).

288. Li, B., Sun, Z., He, Q., Zhu, Y. & Qin, Z. S. Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes. *Bioinformatics* **32,** 682–689 (Mar. 2016).

289. Van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B. & Lesaffre, E. Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research* **27,** 3167–3182 (Oct. 2018).

290. Ryan, E. G. *et al.* Using Bayesian adaptive designs to improve phase III trials: A respiratory care example. *BMC Medical Research Methodology* **19,** 99 (May 2019).

291. Greenland, S. *et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* **31,** 337–350 (Apr. 2016).

292. Huang, E. P. *et al.* Criteria for the translation of radiomics into clinically useful tests. *Nature Reviews Clinical Oncology* **20,** 69–82 (Feb. 2023).

293. Park, J. E. *et al.* Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. *European Radiology* **30,** 523–536 (Jan. 2020).

294. Paquier, Z. *et al.* Radiomics software comparison using digital phantom and patient data: IBSI-compliance does not guarantee concordance of feature values. *Biomedical Physics and Engineering Express* **8** (Nov. 2022).

295. Van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. & Baessler, B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging* **11,** 1–16 (Dec. 2020).

296. McWilliam, A. *et al.* Radiation dose to heart base linked with poorer survival in lung cancer patients. *European Journal of Cancer* **85,** 106–113 (Nov. 2017).

297. McWilliam, A. *et al.* Novel Methodology to Investigate the Effect of Radiation Dose to Heart Substructures on Overall Survival. *International Journal of Radiation Oncology Biology Physics* **108,** 1073–1081 (Nov. 2020).

298. Putora, P. M. & De Ruysscher, D. K. Is Hypofractionation a Good Idea in Radiotherapy for Locally Advanced NSCLC? *Journal of Thoracic Oncology* **17,** 487–488 (Apr. 2022).

299. McGinnis, J. M., Fineberg, H. V. & Dzau, V. J. Advancing the Learning Health System. *New England Journal of Medicine* **385,** 1–5 (2021).

300. Dash, D. *et al.* Building a Learning Health System: Creating an Analytical Workflow for Evidence Generation to Inform Institutional Clinical Care Guidelines. *Applied Clinical Informatics* **13,** 315–321 (Jan. 2022).

301. He, L. *et al.* Bayesian regression analyses of radiation modality effects on pericardial and pleural effusion and survival in esophageal cancer. *Radiotherapy and Oncology* **121,** 70–74 (Oct. 2016).

302. *Radiomics Quality Score - RQS 2.0* https://www.radiomics.world/rqs. *Date accessed: 21/06/2023.*

303. Collins, G. S. *et al.* Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11,** e048008 (July 2021).

304. Ni, K., Chu, H., Zeng, L., Li, N. & Zhao, Y. Barriers and facilitators to data quality of electronic health records used for clinical research in China: A qualitative study. *BMJ Open* **9,** e029314 (July 2019).

305. Thuraisingam, S. *et al.* Assessing the suitability of general practice electronic health records for clinical prediction model development: a data quality assessment. *BMC Medical Informatics and Decision Making* **21,** 1–11 (Dec. 2021).

306. Gianicolo, E. A., Eichler, M., Muensterer, O., Strauch, K. & Blettner, M. Methods for evaluating causality in observational studies: Part 27 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International* **117,** 101–107 (Feb. 2020).

307. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* **183,** 758–764 (Apr. 2016).

308. Price, G. *et al.* Can Real-world Data and Rapid Learning Drive Improvements in Lung Cancer Survival? The RAPID-RT Study. *Clinical Oncology* **34,** 407–410 (June 2022).

309. Fornacon-Wood, I., O'Connor, J., Faivre-Finn, C. & Price, G. PH-0652: Standardization influences repeatability and prognostic value of radiomic features. *Radiotherapy and Oncology* **152,** S362–S363 (Nov. 2020).

310. Fournier, L. *et al.* Incorporating radiomics into clinical trials: expert consensus endorsed by the European Society of Radiology on considerations for data-driven compared to biologically driven quantitative biomarkers. *European Radiology* **31,** 6001–6012 (Aug. 2021).

Appendix A

Changes in the Management of Patients having Radical Radiotherapy for Lung Cancer during the First Wave of the COVID-19 Pandemic in the UK

This chapter has been published in Clinical Oncology 2022 Volume 34 p19-27. The section numbering, references and formatting have been modified to be consistent with the rest of the thesis.

Authors

Kathryn Banfill[1,2], William Croxford[2], Isabella Fornacon-Wood[1], Kate Wicks[1], Shahreen Ahmad[3], Anna Britten[4], Carrie Carson[5], Nicole Dorey[6], Matthew Hatton[7], Crispin Hiley[8], Kamalram Thippu Jayaprakash[9], Apurna Jegannathen[10], Pek Koh[11], Niki Panakis[12], Clive Peedell[13], Anthony Pope[14], Ceri Powell[15], Claire Stilwell[16], Betsan Thomas[17], Elizabeth Toy[18], Victoria Wood[19], Sundus Yahya[20], Zhou Y Suyun[21], Gareth Price[1] and Corinne Faivre-Finn[1,2]

Affiliations

[1] The University of Manchester, Manchester, UK.

[2] The Christie NHS Foundation Trust, Manchester, UK.

[3] Guy's and St Thomas' NHS Foundation Trust, London, UK.

[4] Brighton and Sussex University Hospitals NHS Trust, Brighton, UK.

[5] The Northern Ireland Cancer Centre, Belfast, UK.

[6] Torbay and South Devon NHS Foundation Trust, Torquay, UK.

[7] Weston Park Hospital, Sheffield, UK.

[8] University College London Hospitals, London, UK.

[9] Oncology Centre, Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK.

[10] University Hospitals North Midlands, Stoke-on-Trent, UK.

[11] Royal Wolverhampton NHS Trust, Wolverhampton, UK.

[12] Oxford Universities NHS Trust, Oxford, UK.

[13] The James Cook University Hospital, Middlesbrough, UK.

[14] Clatterbridge Cancer Centre, Bebington, UK.

[15] Velindre Cancer Centre, Cardiff, UK.

[16] Aberdeen Royal Infirmary, Aberdeen, UK.

[17] Swansea Bay University Hospital, Swansea, UK.

[18] Royal Devon and Exeter NHS Foundation Trust, Exeter, UK.

[19] University Hospitals Southampton NHS Foundation Trust, Southampton, UK.

[20] University Hospitals Birmingham, Birmingham, UK.

[21] Beatson West of Scotland Cancer Centre, Glasgow, UK.

Author contributions

I wrote the R script to perform the statistical analysis, and created Figures 1 and 2. I reviewed the manuscript written by K.B..

Abstract

Aims

In response to the COVID-19 pandemic, guidelines on reduced fractionation for patients treated with curative-intent radiotherapy were published, aimed at reducing the number of hospital attendances and potential exposure of vulnerable patients to minimise the risk of COVID-19 infection. We describe the changes that took place in the management of patients with stage I–III lung cancer from April to October 2020.

Materials and Methods

Lung Radiotherapy during the COVID-19 Pandemic (COVID-RT Lung) is a prospective multicentre UK cohort study. The inclusion criteria were: patients with stage I–III lung cancer referred for and/or treated with radical radiotherapy between 2nd April and 2nd October 2020. Patients who had had a change in their management and those who continued with standard management were included. Data on demographics, COVID-19 diagnosis, diagnostic work-up, radiotherapy and systemic treatment were collected and reported as counts and percentages. Patient characteristics associated with a change in treatment were analysed using multivariable binary logistic regression.

Results

In total, 1553 patients were included (median age 72 years, 49% female); 93 (12%) had a change to their diagnostic investigation and 528 (34%) had a change to their treatment from their centre's standard of care as a result of the COVID-19 pandemic. Age $\geq$ 70 years, male gender and stage III disease were associated with a change in treatment on multivariable analysis. Patients who had their treatment changed had a median of 15 fractions of radiotherapy compared with a median of 20 fractions in those who did not have their treatment changed. Low rates of COVID-19 infection were seen during or after radiotherapy, with only 21 patients (1.4%) developing the disease.

Conclusion

The COVID-19 pandemic resulted in changes to patient treatment in line with national recommendations. The main change was an increase in hypofractionation. Further

work is ongoing to analyse the impact of these changes on patient outcomes.

A.1 Introduction

At the outset of the coronavirus pandemic in early 2020, there was paucity of data about the risk of COVID-19 in patients with lung cancer. Initial reports suggested that patients with cancer had both a higher risk of COVID-19 and an increased incidence of intensive care unit admissions and death [1]. There were concerns that immunosuppression due to chemotherapy and thoracic radiotherapy [2] would increase the likelihood of severe COVID-19 infection in patients with lung cancer. Moreover, cigarette smoking and comorbidities, such as hypertension and chronic obstructive pulmonary disease, that are common in patients with lung cancer, increased the risk of hospital admission or death from COVID-19 [3]. Consequently, in the UK, patients undergoing radical radiotherapy for lung cancer were identified as clinically extremely vulnerable and advised to shield [4].

Radiotherapy is used in the primary treatment of 20–60% of patients with lung cancer [5]. It was therefore crucial to find ways to balance continued access of cancer patients to radiotherapy while minimising the risk of COVID-19 infection. To this end, at the start of the first wave of the COVID-19 pandemic in the UK, a group of clinical oncologists published guidelines based on current literature and practice, on how to safely reduce the number of fractions (and therefore hospital visits) when delivering curative-intent radiotherapy in patients with lung cancer [6].

Following the publication of the UK recommendations [6], it is now important to assess their impact on practice and on the clinical outcome of lung cancer patients considered for radiotherapy in the UK during the COVID-19 pandemic. Lung Radiotherapy during the COVID-19 Pandemic (COVID-RT Lung) is a UK-wide cohort study established to record and analyse changes in lung cancer management and outcomes, with a specific focus on radical radiotherapy treatment. The study aims to assess the effect of the COVID-19 pandemic on the diagnostic and treatment pathways for patients with lung cancer. Here we outline the initial results from COVID-RT Lung.

A.2 Methods and Materials

COVID-RT Lung is a national, multicentre prospective registry cohort study. Radiotherapy centres in the UK were sent a letter of invitation to participate in COVID-RT Lung. Interested centres then registered the project locally as a clinical audit. Local approval to enter anonymous patient data was obtained by each participating centre from their Caldicott Guardian.

A.2.1 Patient Cohort

Patients were eligible for inclusion in COVID-RT Lung if they had stage I–III lung cancer and were referred for curative-intent radiotherapy (biologically equivalent dose $>$ 50 Gy) at a participating centre between 2nd April and 2nd October 2020. Participating radiotherapy centres prospectively collected data from the patient records on all patients referred for and treated with radical radiotherapy during the inclusion period, whether or not their treatment was changed due to the COVID-19 pandemic. Radiotherapy centres were divided by UK region.

A patient was defined as having a change to their diagnostic investigations if standard investigations that would usually be carried out prior to radiotherapy at their treating centre were not undertaken as a result of the COVID-19 pandemic. A patient was defined as having a change to treatment if they had a different treatment to their centre's standard of care treatment, taking into account individual patient characteristics, such as performance status, and tumour characteristics, such as stage.

The following information was collected for each patient: age at the time of treatment; gender; histology; stage; baseline Eastern Cooperative Oncology Group Performance Status (PS) and comorbidities; radiotherapy dose and fractionation; dates of radiotherapy; chemotherapy or immunotherapy delivery. Stereotactic ablative body radiotherapy (SABR) was defined as radiotherapy delivered in $\leq$ 8 fractions of more than 6 Gy per fraction. Specific details on the chemotherapy drugs used were not collected. A systemic therapy dose reduction was defined as a reduction in the number of planned cycles of therapy and/or a reduction in the dose for any single cycle.

The presence of the following comorbidities was recorded: ischaemic heart disease; congestive heart failure; cardiac arrhythmia, hypertension; chronic obstructive pulmonary disease; chronic kidney disease; diabetes; stroke; dementia and any previous malignancy prior to the current lung cancer diagnosis. Other comorbidities were recorded as free text.

Data were collected on COVID-19 diagnosis. Patients were classified as having COVID-19 if they had a positive reverse transcriptase polymerase chain reaction (RT-PCR) nasopharyngeal swab or if they had a clinical diagnosis of COVID-19 in the absence of an RT-PCR swab.

In addition, the following data were collected if available: Rockwood clinical frailty scale; smoking history; administration of granulocyte colony stimulating factor (G-CSF) during treatment; neutrophil and lymphocyte count in the final week of radiotherapy.

Study data were collected and managed using the Research Electronic Data Capture (REDCap) cloud platform (nPhase Inc, CA, USA) [7] administered by the University of Manchester Clinical Trials Unit.

A.2.2 Statistical Analysis

Key baseline characteristics were summarised as categorical variables and reported as counts and percentages. Adjusted odds ratios and 95% confidence intervals for a change to diagnostic investigations and treatment from the centre's standard of care were estimated by multivariable binary logistic regression. Age was dichotomised at 70 years in line with the UK Government's shielding advice. The Rockwood frailty score was excluded from multivariable analysis as more than 50% of the data were missing. North West England was chosen as the base factor in regional analysis as it had the largest number of patients. The median number of radiotherapy fractions were compared using the Wilcoxon signed rank test. The statistical analysis was carried out in R studio version 3.6.3.

A.3 Results

Data on 1553 patients were available for analysis on 17 March 2021. The median age was 72 years (37–93 years), 762 (49%) were female. There were 906 patients (58.3%) with non-small cell lung cancer (NSCLC) and 482 (31%) had a radiological diagnosis of cancer. Only 167 patients (10.8%) had no comorbidities prior to their diagnosis of lung cancer and 624 patients (40.2%) had three or more comorbidities. The most common comorbidity was chronic obstructive pulmonary disease, recorded in 667 patients (42.9%). A list of participating radiotherapy centres and their region is given in Supplementary Table A.5. Baseline characteristics are summarised in Table A.1 .

Table A.1: Baseline characteristics stratified by change to treatment [n (%)]

	No change	**Changed**	**Total**
Total n (%)	1025 (66.0)	528 (34.0)	1553
Age (years)			
< 70	405 (39.5)	203 (38.4)	608 (39.2)
≥ 70	610 (59.5)	325 (61.6)	935 (60.2)
Missing	10 (1.0)	0 (0.0)	10 (0.6)
Gender			
Female	524 (51.1)	238 (45.1)	762 (49.1)
Male	495 (48.3)	290 (54.9)	785 (50.5)
Missing	6 (0.6)	0 (0.0)	6 (0.4)
Performance status			
0	118 (11.5)	96 (18.2)	214 (13.8)
1	509 (49.7)	309 (58.5)	818 (52.7)
2–3	390 (38.0)	123 (23.3)	513 (33.0)
Missing	8 (0.8)	0 (0.0)	8 (0.5)
Clinical frailty scale			
1	18 (1.8)	14 (2.7)	32 (2.1)
2	72 (7.0)	59 (11.2)	131 (8.4)
3	143 (14.0)	115 (21.8)	258 (16.6)
4	101 (9.9)	62 (11.7)	163 (10.5)
5	56 (5.5)	22 (4.2)	78 (5.0)
6	27 (2.6)	10 (1.9)	37 (2.4)
7	7 (0.7)	1 (0.2)	8 (0.5)
Missing	601 (58.6)	245 (46.4)	846 (54.5)
Smoking status			
Current smoker	298 (29.1)	148 (28.0)	446 (28.7)
Ex-smoker	594 (58.0)	317 (60.0)	911 (58.7)

Table A.1 continued from previous page

	No change	Changed	Total
Never smoker	29 (2.8)	22 (4.2)	51 (3.3)
Missing	104 (10.1)	41 (7.8)	145 (9.3)
Histology			
NSCLC	576 (56.2)	330 (62.5)	906 (58.3)
SCLC	87 (8.5)	70 (13.3)	157 (10.1)
Radiological diagnosis	354 (34.5)	128 (24.2)	482 (31.0)
Missing	8 (0.8)	0 (0.0)	8 (0.5)
Stage			
I	473 (46.1)	189 (35.8)	662 (42.6)
II	164 (16.0)	71 (13.4)	235 (15.1)
III	380 (37.1)	265 (50.2)	645 (41.5)
Missing	8 (0.8)	3 (0.6)	11 (0.7)
Region			
North West England	166 (16.2)	159 (30.1)	325 (20.9)
Yorkshire & North East England	161 (15.7)	106 (20.1)	267 (17.2)
South East England	151 (14.7)	77 (14.6)	228 (14.7)
London	46 (4.5)	17 (3.2)	63 (4.1)
South West England	50 (4.9)	28 (5.3)	78 (5.0)
Midlands	123 (12.0)	49 (9.3)	172 (11.1)
Northern Ireland	88 (8.6)	30 (5.7)	118 (7.6)
Wales	63 (6.1)	42 (8.0)	105 (6.8)
Scotland	177 (17.3)	20 (3.8)	197 (12.7)
IHD			
No IHD	819 (79.9)	439 (83.1)	1258 (81.0)
IHD	206 (20.1)	89 (16.9)	295 (19.0)
CHF			
No CHF	964 (94.0)	506 (95.8)	1470 (94.7)
CHF	61 (6.0)	22 (4.2)	83 (5.3)
Cardiac arrhythmia			
No arrhythmia	912 (89.0)	471 (89.2)	1383 (89.1)
Arrhythmia	113 (11.0)	57 (10.8)	170 (10.9)
Hypertension			
No hypertension	660 (64.4)	354 (67.0)	1014 (65.3)
Hypertension	365 (35.6)	174 (33.0)	539 (34.7)
COPD			
No COPD	574 (56.0)	312 (59.1)	886 (57.1)
COPD	451 (44.0)	216 (40.9)	667 (42.9)
CKD			
No CKD	961 (93.8)	507 (96.0)	1468 (94.5)
CKD	64 (6.2)	21 (4.0)	85 (5.5)

Table A.1 continued from previous page

	No change	Changed	Total
Diabetes			
No diabetes	859 (83.8)	449 (85.0)	1308 (84.2)
Diabetes	166 (16.2)	79 (15.0)	245 (15.8)
Stroke/TIA			
No stroke	930 (90.7)	493 (93.4)	1423 (91.6)
Stroke	95 (9.3)	35 (6.6)	130 (8.4)
Dementia			
No dementia	1011 (98.6)	522 (98.9)	1533 (98.7)
Dementia	14 (1.4)	6 (1.1)	20 (1.3)
Previous malignancy			
No previous malignancy	772 (75.3)	423 (80.1)	1195 (76.9)
Previous malignancy	253 (24.7)	105 (19.9)	358 (23.1)

Abbreviations: *CHF*, congestive heart failure; *COPD*, chronic obstructive pulmonary disease; *CKD*, chronic kidney disease; *IHD*, ischaemic heart disease; *NSCLC*, non-small cell lung cancer; *SCLC*, small cell lung cancer; *TIA*, transient ischaemic attack.

One hundred and ninety-three patients (12%) had their diagnostic investigations affected by the pandemic (Table A.2). The characteristics of patients who had their diagnostic investigations changed are listed in Supplementary Table A.6. The most common change was not obtaining histology prior to treatment in 66 patients.

Table A.2: Changes to diagnostic investigations.

Change to diagnostic investigations	Patients n=1553
Histology not obtained	66 (4.3%)
No nodal sampling	38 (2.5%)
No pulmonary function tests	29 (1.9%)
No brain imaging	32 (2.1%)
No PET-CT or PET-CT out of date*	50 (3.2%)
Delays in diagnosis	11 (0.7%)

*Defined by the local clinical teams. Abbreviations: *PET-CT*, positron emission tomography-computed tomography.

Radiotherapy details were not recorded for 11 patients. In 33 patients (2.1%), a watch and wait approach was adopted, 26 of whom went on to have radiotherapy at a later date. Three patients (0.2%) had best supportive care instead of radical treatment and 26 patients referred for radical treatment had a palliative radiotherapy schedule. In

patients with stage I–II lung cancer, 579 (64.5%) had SABR, 296 (33%) had fractionated curative-intent radiotherapy. Eight patients had single fraction SABR with 34 Gy. In patients with stage III lung cancer, 356 patients (55%) had sequential or concurrent chemoradiotherapy and 264 patients (40.9%) had curative-intent radiotherapy alone.

Changes to treatment due to the pandemic are shown in Table 3. The most common change was to the centre's standard radiotherapy dose or fractionation. The median number of fractions of radiotherapy received by patients who had their treatment changed was significantly lower than those without a change to their standard radiotherapy (15 fractions versus 20 fractions, P<0.001).

Table A.3: Changes made to patients' treatment according to lung cancer stage (information on stage was missing for four patients).

Change to treatment	**Stage I–II (n=897)**	**Stage III (n=645)**
Any change	260 (29%)	265 (41.1%)
Change to radiotherapy dose/fractionation	144 (16.1%)	126 (19.5%)
Radiotherapy given instead of surgery	85 (9.5%)	18 (2.8%)
Chemotherapy dose reduced	12 (1.3%)	59 (6.8%)
Chemotherapy omitted	9 (1%)	69 (10.7%)
Immunotherapy dose reduced or omitted	0	8 (1.2%)
Watch and wait	31 (3.5%)	2 (0.3%)
Best supportive care	1 (0.1%)	2 (0.3%)
Other	2 (0.2%)	4 (0.6%)

A higher proportion of patients with small cell lung cancer (SCLC) had their treatment changed compared with patients with NSCLC (44.6% versus 36.4%). The median radiotherapy dose per fraction for patients with SCLC was 2.67 Gy to a total of 40 Gy, delivered once daily. This schedule was used with concurrent chemotherapy in 18 patients, with sequential chemotherapy in 65 patients and without chemotherapy in 10 patients. The schedule of 45 Gy in 30 fractions bi-daily with concurrent chemotherapy was delivered in 25 patients with SCLC.

The median radiotherapy dose per fraction for patients with NSCLC was 2.75 Gy to a total of 55 Gy, delivered with sequential chemotherapy in 142 patients, concurrent

chemotherapy in 146 patients and without chemotherapy in 616 patients. In patients with a radiological diagnosis of lung cancer (assumed to be NSCLC by the multidisciplinary team), the median radiotherapy dose per fraction was 11 Gy to a total of 55 Gy; five patients with a radiological diagnosis had chemotherapy.

Five hundred and twenty-eight patients (34%) had their treatment changed from their centre's standard of care treatment due to the COVID-19 pandemic. The North West and Yorkshire/North East of England had the highest proportion of patients who had their treatment changed from their centre's standard of care. Multivariable analysis revealed that male gender, age $\geq$ 70 years and stage III lung cancer were associated with a change in treatment (Table A.4). Patients of performance status 2–3 were less likely to have their treatment changed. A change in diagnostic investigations was associated with a radiological diagnosis of lung cancer, chronic kidney disease and treatment in Northern Ireland.

Table A.4: Adjusted odds ratio (aOR) of baseline factors with change to treatment and change to diagnostic investigations.

	Change to treatment aOR (95% CI)	**Change to investigations aOR (95% CI)**
Age (years) <70 versus $\geq$ 70	1.34 (1.03–1.74)	0.95 (0.64–1.42)
Gender, female versus male	1.36 (1.07–1.73)	0.90 (0.63–1.29)
Performance status, versus 0		
1	0.78 (0.55–1.12)	1.00 (0.57–1.82)
2–3	0.37 (0.25–0.56)	0.72 (0.38–1.39)
Smoking status, versus current smoker		
Ex-smoker	0.96 (0.73–1.25)	0.74 (0.50–1.10)
Never smoker	1.51 (0.78–2.86)	0.65 (0.20–1.77)
Histology, versus NSCLC		
SCLC	1.27 (0.86–1.89)	0.97 (0.46–1.90)
Radiological diagnosis	0.88 (0.64–1.20)	3.83 (2.44–6.09)
Stage, versus stage I		
II	0.95 (0.65–1.38)	1.25 (0.73–2.10)
III	1.56 (1.15–2.13)	1.42 (0.88–2.30)
Region, versus North West England		
Yorkshire & North East England	0.62 (0.43–0.90)	0.81 (0.43–1.46)
South East England	0.41 (0.27–0.61)	0.34 (0.13–0.75)
London	0.34 (0.18–0.63)	0.29 (0.05–1.02)
South West England	0.41 (0.23–0.71)	1.62 (0.75–3.36)
Midlands	0.41 (0.26–0.65)	0.64 (0.29–1.31)
Northern Ireland	0.33 (0.20–0.54)	13.87 (8.05–24.47)

Table A.4 continued from previous page

	Change to treatment aOR (95% CI)	Change to investigations aOR (95% CI)
Wales	0.62 (0.39–1.01)	2.36 (1.22–4.48)
Scotland	0.09 (0.05–0.14)	0.44 (0.21–0.88)
Ischaemic heart disease, no versus yes	0.94 (0.68–1.28)	1.22 (0.78–1.89)
Congestive heart failure, no versus yes	0.98 (0.54–1.74)	0.54 (0.20–1.27)
Arrhythmia, no versus yes	1.07 (0.71–1.60)	1.04 (0.56–1.86)
Hypertension, no versus yes	0.80 (0.62–1.03)	1.15 (0.79–1.68)
COPD, no versus yes	1.24 (0.96–1.60)	1.29 (0.88–1.88)
Chronic kidney disease, no versus yes	0.78 (0.43–1.36)	2.13 (1.03–4.19)
Diabetes, no versus yes	0.99 (0.71–1.39)	1.08 (0.65–1.75)
Stroke, no versus yes	0.83 (0.52–1.30)	1.03 (0.54–1.87)
Dementia, no versus yes	1.19 (0.39–3.27)	0.91 (0.18–3.36)
Previous malignancy, no versus yes	0.86 (0.64–1.15)	1.12 (0.72–1.70)

Abbreviations: *CI*, confidence interval; *COPD*, chronic obstructive pulmonary disease; *NSCLC*, non-small cell lung cancer; *SCLC*, small cell lung cancer.

Figure A.1 shows the changes in radiotherapy dose per fraction for patients who did and did not have their treatment changed. A higher proportion of the change to treatment group were treated with 3–5.9 Gy/fraction across all stages compared with the no change group (27.2% versus 5.1%). No patient in the change to treatment group received a radiotherapy schedule with <2 Gy/fraction.

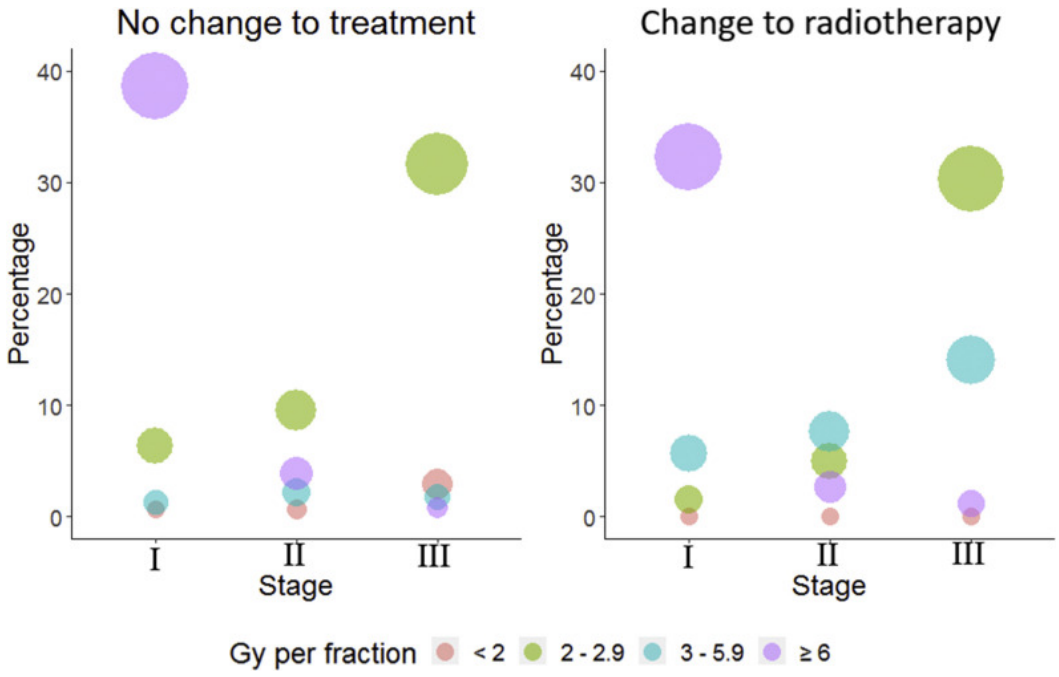

Figure A.1: Bubble plot of radiotherapy dose per fraction by stage for patients who had standard of care treatment and those who had their treatment changed.

Figure A.2 shows how the changes to treatment varied between April and October 2020. In April 2020, 105 of 180 (37%) patients had their treatment changed; in May 2020, 154 of 345 (45%) had their treatment changed. The total number of patients treated and the proportion of patients who had a change to their treatment decreased from June to August 2020.

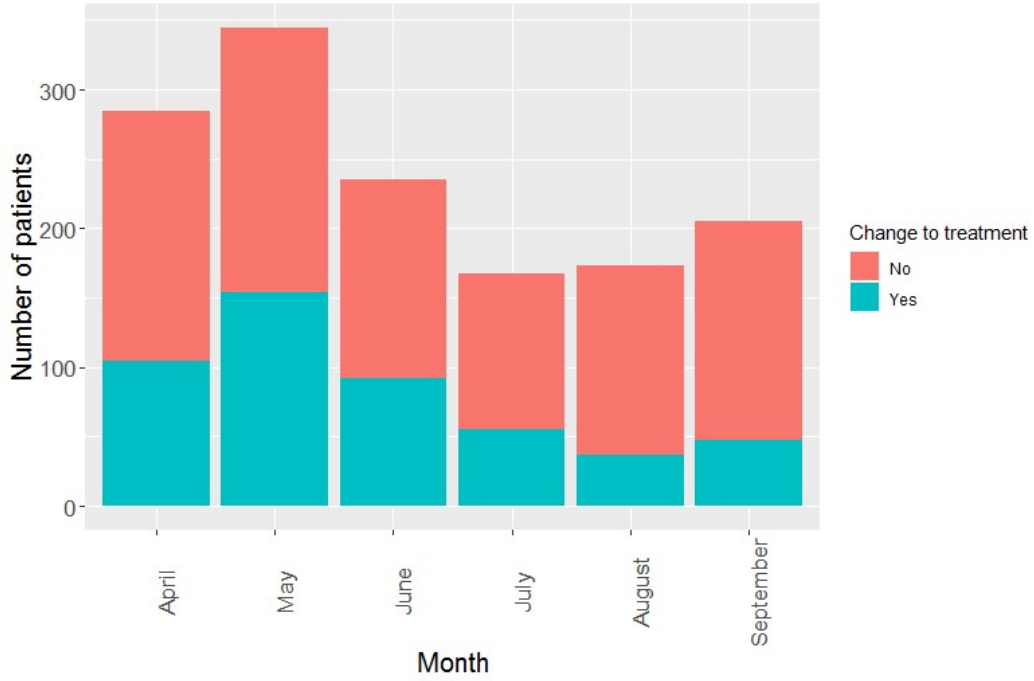

Figure A.2: Monthly number of patients referred for radical radiotherapy for lung cancer and the number who had a change to their treatment from April to September 2020.

A.3.1 Lymphocyte Count at End of Radiotherapy

Lymphocyte count in the final week of radiotherapy was available for 90 patients with stage I–II lung cancer (10%) and 210 patients with stage III lung cancer (32.6%). Sixty patients for whom counts were available had sequential chemoradiotherapy and 114 had concurrent chemoradiotherapy. The median lymphocyte count in the last week of radiotherapy for patients who had any chemotherapy was 0.6 $\times$ 109/l (0.2–2.4 $\times$ 109/l) and in patients who did not have chemotherapy the median lymphocyte count was 0.7 $\times$ 109/l (0.2–2.3 $\times$ 109/l). Fourteen patients out of 33 who had a diagnosis of COVID-19 had a lymphocyte count from the last week of radiotherapy available for analysis and nine of these patients had a lymphocyte count $\leq$ 0.5 $\times$ 109/l.

Seventy-eight patients in COVID-RT Lung received G-CSF during treatment. Sixty-one patients who received G-CSF underwent concurrent chemoradiotherapy; 43 for NSCLC and 18 for SCLC. Three patients who received G-CSF had a diagnosis of COVID-19 during their radiotherapy.

A.3.2 COVID-19 Diagnosis and Treatment Delays

Thirty-three (2.1%) patients had a diagnosis of COVID-19, 26 of whom had RT-PCR swab confirmation. Twelve patients were diagnosed with COVID-19 prior to starting treatment for lung cancer, six were diagnosed during radiotherapy and 15 were diagnosed after the end of radiotherapy. Of the 21 patients who had COVID-19 during or after radiotherapy, seven patients had stage I–II lung cancer (0.8% of all early stage patients) and 14 had stage III disease (2.2% of all stage III patients). Six patients died with COVID-19 at a median of 175 days (21–279 days) after the start of radiotherapy.

The median duration of treatment interruption for patients with confirmed or suspected COVID-19 was 4 days (1–16 days). In total, 83 patients (5.3%) had an interruption to radiotherapy for any reason, 18 of whom had their treatment stopped early. No patient with a diagnosis of COVID-19 had their treatment stopped early. The most common reasons for treatment interruptions, apart from COVID-19, were other infections (10 patients) or treatment toxicity (10 patients). One patient (who had RT-PCR-confirmed COVID-19) had a radiotherapy delay compensation with the addition of two extra fractions at the end of radiotherapy, making a total dose of 60.5 Gy in 22 fractions.

A.4 Discussion

This first analysis of the UK COVID-RT Lung cohort study has shown that the COVID-19 pandemic led to a subsequent change to treatment in a third of lung cancer patients referred for radical radiotherapy between April and October 2020. In addition, the diagnostic pathway was altered in 12% of patients of the same cohort.

The most common change to treatment was the use of a different radiotherapy dose/fractionation from the centre's usual standard of care in 17.5% of patients, resulting in a higher proportion of patients with lung cancer treated with hypofractionated radiotherapy. These changes are in line with a national UK guidance document of hypofractionated radiotherapy [6]. Greater use of hypofractionation in all cancers during the COVID-19 pandemic was reported in a population-based study analysing data from the UK National Radiotherapy Dataset [8]. The national dataset showed an increase in moderately hypofractionated radiotherapy for patients with lung cancer (2.5–4.9 Gy/fraction) but little change in ultra-hypofractionated radiotherapy ($\geq$ 5 Gy per fraction). In COVID-RT Lung, there was an increase in the use of 3–5.9 Gy/fraction regimens in patients who had their treatment changed for all stages of lung cancer. This change reflects the increased use of 15-fraction schedules during the pandemic. The evidence for the use of 60 Gy in 15 fractions comes from a phase II trial in patients with T1-3 N0 M0 NSCLC, which reported an overall survival of 68.7% at 2 years with a low rate of grade 3+ toxicity [9].

Radiotherapy was suggested as an alternative to surgery at the start of the COVID-19 pandemic if there was pressure on operating and anaesthetic resources [8, 10]. We found that radiotherapy was used as an alternative treatment to surgery in 9.5% of stage I–II and 2.8% of stage III operable patients between April and October 2020. The number of patients who had changes to treatment changed over time with, the largest number of treatment changes in April and May 2020. Our results show a fall in the number of referrals for radical radiotherapy in June, July and August 2020 which could be explained by the fall in suspected lung cancer referrals [11].

Multivariable analysis of baseline factors found that male gender, age $\geq$ 70 years and stage III lung cancer were associated with patients having a change to their treatment from their centre's standard of care. Spencer et al. [8] also found that there was more of a decrease in the number of radiotherapy treatment courses for all cancer patients in patients aged $\geq$ 70 compared with patients $<$ 70 years. Older age and male gender have consistently been associated with a higher morbidity and mortality with COVID-19 [3], leading to adjustment in treatments to mitigate the risk.

When interpreting the results of COVID-RT Lung it should be noted that standard of care treatment for stage I–III lung cancer varies considerably between UK centres. For this reason, the central question of the analysis was whether the patient's treatment had been changed from their centre's standard of care.

Results from the National Lung Cancer Audit 2016 reported that 65% of patients who received radical radiotherapy for stage III lung cancer had sequential or concurrent chemotherapy [12]. Only 55% of patients with stage III disease in COVID-RT Lung had chemotherapy in addition to radical radiotherapy and 10.7% had their chemotherapy omitted because of the pandemic. The lower rates of chemotherapy in our study may be due to the perceived risk of COVID-19 in patients who are immunosuppressed [13]. The combination of lower rates of chemotherapy and more hypofractionated radiotherapy resulted in patients with stage III disease having more changes to treatment than those with stage I disease. Furthermore, we found that patients with PS 2–3 were less likely to have their treatment changed. Patients with PS 2–3 often receive curative-intent radiotherapy alone rather than surgery or chemoradiotherapy [12, 14], leaving less scope for their treatment to be changed as a result of the pandemic. No specific comorbidity was associated with a change in treatment, although patients with chronic kidney disease were more likely to have a change to their diagnostic investigations.

COVID-RT Lung demonstrates geographical variation in treatment changes, with the North of England having the greatest proportion of patients with changes to their treatment. These regions had some of the highest rates of COVD-19 infection in the UK [15]. Pre-pandemic regional treatment variations may also explain why some areas of the country recorded a lower proportion of patients with a change to treatment.

The incidence of COVID-19 infections in the COVID-RT Lung cohort was low (2.1% of patients in total and only 1.4% were infected during or after radiotherapy). Variations in COVID-19 testing policies between centres and over time, especially the slow roll out of COVID-19 testing nationally during the initial months of the pandemic, will influence the reported incidence rate in this study. Nevertheless, the low COVID-19 rate is reassuring given the high rates of lymphopenia reported during the last

week of radiotherapy in a subgroup of patients. The low rates of COVID-19 infection and death may reflect the UK Government's shielding advice [4] for patients having thoracic radiotherapy. In addition the use of hypofractionated radiotherapy, as per UK and international guidance [6, 16], may have reduced the patients' exposure to COVID-19.

Only six patients in our study died with COVID-19, in contrast to the high rate of death from COVID-19 reported in the TERAVOLT [17] registry. TERAVOLT was a small cohort of 200 patients, most with stage IV disease, and included a skewed population of symptomatic patients on active treatment who presented to oncological services. The UK Coronavirus Cancer Monitoring project is a larger UK-based registry of patients with COVID-19 and cancer, which did not find an increased case-fatality rate due to COVID-19 in patients with lung cancer [18]. There were also concerns in April 2020 that patients having radiotherapy for lung cancer would have treatment delays as a result of COVID-19 and therefore the Royal College of Radiologists produced guidance on compensating for treatment gaps [19]. We found that 5.3% of patients in the COVID-RT Lung cohort had a treatment gap, but this was more often due to treatment toxicity or an infection other than COVID-19. The median treatment gap for patients with suspected COVID-19 during treatment was 4 days (range 1–16 days), which implies that most patients continued their treatment during the self-isolation period, in order to maximise the potential for cure [20]. Nevertheless, it is surprising that only one patient in this cohort had treatment compensation for a gap in radiotherapy.

Our analysis has limitations as it only includes data from 30 UK radiotherapy centres across the whole of the UK and participating centres had not completed data collection on all treated patients at the time of this initial analysis. Consequently, the denominator of patients with lung cancer receiving radiotherapy during this period of the pandemic is not known. The analysis of lymphopenia following radiotherapy is limited by the small proportion of patients in COVID-RT Lung for whom lymphocyte count was available. COVID-RT Lung demonstrates the same pattern of radiotherapy hypofractionation as reported in national datasets during the pandemic [8], which indicates that it is probably representative of changes across the country. Our study

provides more granular detail on the changes to diagnostic pathways, radiotherapy and systemic therapy in patients with lung cancer and specifically asks if the patient had a change to treatment compared with the local standard of care rather than inferring this from changes in national datasets over time.

We have shown that the risk of developing COVID-19 in lung cancer patients receiving radical radiotherapy was low during the first wave of the pandemic, showing that the measures put in place by radiotherapy departments to protect patients [16] were adequate. We have described the characteristics of patients who had changes to their centre's standard of care management and the regional differences in the management of patients with lung cancer. An important next step is to report the outcomes of patients treated during the pandemic in order to assess the effect of radiotherapy and chemotherapy adaptations on survival and toxicity. Outcome data are being collected as data matures. Given the current concerns regarding the cancer backlog and National Health Service pressures as a consequence of the pandemic, our study will provide valuable information to the oncology community to help guide optimal treatment for lung cancer patients going forward.

A.5 Conflicts of interest

C. Peedell reports a relationship with Elekta that includes: speaking and lecture fees; a relationship with AstraZeneca Pharmaceuticals LP that includes: speaking and lecture fees; a relationship with Boston Scientific Corp that includes: consulting or advisory and speaking and lecture fees. K. Banfill reports a relationship with AstraZeneca that includes: speaking and lecture fees. Kamalram Thippu Jayaprakash reports a relationship with AstraZeneca that includes: travel reimbursement. K.Thippu Jayaprakash acknowledges the following funding support outside the submitted work: research grant from the UK National Institute of Health Research and educational grants from Bayer UK, Janssen Oncology, Pfizer, Roche, and Takeda.Elizabeth Toy is a member of Lung Cancer Expert Reference Group and Clinical Lead GIRFT Lung Cancer workstream. Crispin Hiley acknowledges the following funding support outside the submitted work: research grant and speaking fees from AstraZeneca. Crispin Hiley reports a relationship with Roche that includes: speaking and lecture fees.

A.6 Acknowledgements

This work was supported by Cancer Research UK RadNet Manchester (grant number C1994/A28701) and NIHR Manchester Biomedical Research Centre (grant number BRC-1215-20007). The authors would like to acknowledge the following people for their assistance with this project: Jacqui Parker, Lee Whiteside, Lucy Davies, Josephine Sanders, Louise McHugh, Philip Teles Amaro, Amy Irwin, Yash Choudhary, Victoria Harrop, Rebekah Shingler, Emma Wingate, Liliam Ross, Lynn Bell, Jasima Latif, Chloe Wilkinson, Stephen Harrow, Adam Peters, Paula Robson, Keith Harland, Asia Sarwar, Jolyne O'Hare, Jonathan McAleese, Ruth Eakin, Linda Young, Nicola Hill, Charis Thompson, C.L. Lee, Hannah Bainbridge, Mike Bayne, Eleanor Weir, Sam Guglani, Hannah Lord, Dila Mokhtar, Lynne White, Sarah Treece, Jennifer Poole.

A.7 Supplementary materials

Table A.5: Number of patients in COVID-RT Lung from each participating centre.

Participating centre	**N**
Scotland	
Beatson West of Scotland Cancer Centre	161
Aberdeen Royal Infirmary	24
Ninewells Hospital	12
Wales	
Swansea Bay University Hospital Trust	39
Velindre Cancer Centre	66
Northern Ireland	
Northern Ireland Cancer Centre	118
North West England	
The Christie NHS Foundation Trust	257
Clatterbridge Cancer Centre	68
North East England	
Leeds Cancer Centre	14
Weston Park Cancer Centre	120
James Cook University Hospital	133
Midlands	
University Hospital Birmingham	115
Wolverhampton	32
University Hospital North Midlands	25

Table A.5 continued from previous page

Participating centre	**N**
South West England	
Cheltenham Hospital	6
Plymouth Oncology Centre	12
Poole Hospital	14
Royal Devon and Exeter Hospital	15
Taunton and Somerset Hospital	11
Torbay and South Devon Hospital	15
University Hospitals Bristol and Weston	5
South East England	
Brighton and Sussex University Hospital	20
Cambridge University Hospital	44
University Hosptial Southampton	82
Oxford Universities NHS Foundation Trust	66
Peterborough City Hospital	11
Portsmouth Hospital	5
London	
Guy's and St Thomas' NHS Foundation Trust	20
Royal Free London NHS Foundation Trust	16
University College London Hospitals	27

Table A.6: Baseline characteristics stratified by change to diagnostic investigations [n (%)]

	No change	**Changed**	**Total**
Total n (%)	1361 (87.6)	192 (12.4)	1553
Age (years)			
< 70	531 (39.0)	77 (40.1)	608 (39.2)
≥ 70	821 (60.3)	114 (59.4)	935 (60.2)
Missing	9 (0.7)	1 (0.5)	10 (0.6)
Gender			
Female	663 (48.7)	99 (51.6)	762 (49.1)
Male	692 (50.8)	93 (48.4)	785 (50.5)
Missing	6 (0.4)	0 (0.0)	6 (0.4)
Performance status			
0	194 (14.3)	20 (10.4)	214 (13.8)
1	713 (52.4)	105 (54.7)	818 (52.7)
2–3	446 (32.8)	67 (34.9)	513 (33.0)
Missing	8 (0.6)	0 (0.0)	8 (0.5)
Clinical frailty scale			
1	29 (2.1)	3 (1.6)	32 (2.1)
2	119 (8.7)	12 (6.2)	131 (8.4)
3	236 (17.3)	22 (11.5)	258 (16.6)

Table A.6 continued from previous page

	No change	Changed	Total
4	146 (10.7)	17 (8.9)	163 (10.5)
5	74 (5.4)	4 (2.1)	78 (5.0)
6	35 (2.6)	2 (1.0)	37 (2.4)
7	8 (0.6)	0 (0.0)	8 (0.5)
Missing	714 (52.5)	132 (68.8)	846 (54.5)
Smoking status			
Current smoker	380 (27.9)	66 (34.4)	446 (28.7)
Ex-smoker	798 (58.6)	113 (58.9)	911 (58.7)
Never smoker	46 (3.4)	5 (2.6)	51 (3.3)
Missing	137 (10.1)	8 (4.2)	145 (9.3)
Histology			
NSCLC	820 (60.2)	86 (44.8)	906 (58.3)
SCLC	145 (10.7)	12 (6.2)	157 (10.1)
Radiological diagnosis	388 (28.5)	94 (49.0)	482 (31.0)
Missing	8 (0.6)	0 (0.0)	8 (0.5)
Stage			
I	571 (42.0)	91 (47.4)	662 (42.6)
II	203 (14.9)	32 (16.7)	235 (15.1)
III	576 (42.3)	69 (35.9)	645 (41.5)
Missing	11 (0.8)	0 (0.0)	11 (0.7)
Region			
North West England	290 (21.3)	35 (18.2)	325 (20.9)
North East England	244 (17.9)	23 (12.0)	267 (17.2)
South East England	219 (16.1)	9 (4.7)	228 (14.7)
London	61 (4.5)	2 (1.0)	63 (4.1)
South West England	65 (4.8)	13 (6.8)	78 (5.0)
Midlands	161 (11.8)	11 (5.7)	172 (11.1)
Northern Ireland	52 (3.8)	66 (34.4)	118 (7.6)
Wales	85 (6.2)	20 (10.4)	105 (6.8)
Scotland	184 (13.5)	13 (6.8)	197 (12.7)
IHD			
No IHD	1113 (81.8)	145 (75.5)	1258 (81.0)
IHD	248 (18.2)	47 (24.5)	295 (19.0)
CHF			
No CHF	1285 (94.4)	185 (96.4)	1470 (94.7)
CHF	76 (5.6)	7 (3.6)	83 (5.3)
Cardiac arrhythmia			
No arrhythmia	1210 (88.9)	173 (90.1)	1383 (89.1)
Arrhythmia	151 (11.1)	19 (9.9)	170 (10.9)
Hypertension			
No hypertension	892 (65.5)	122 (63.5)	1014 (65.3)

Table A.6 continued from previous page

	No change	**Changed**	**Total**
Hypertension	469 (34.5)	70 (36.5)	539 (34.7)
COPD			
No COPD	789 (58.0)	97 (50.5)	886 (57.1)
COPD	572 (42.0)	95 (49.5)	667 (42.9)
CKD			
No CKD	1290 (94.8)	178 (92.7)	1468 (94.5)
CKD	71 (5.2)	14 (7.3)	85 (5.5)
Diabetes			
No diabetes	1148 (84.3)	160 (83.3)	1308 (84.2)
Diabetes	213 (15.7)	32 (16.7)	245 (15.8)
Stroke/TIA			
No stroke	1249 (91.8)	174 (90.6)	1423 (91.6)
Stroke	112 (8.2)	18 (9.4)	130 (8.4)
Dementia			
No dementia	1344 (98.8)	189 (98.4)	1533 (98.7)
Dementia	17 (1.2)	3 (1.6)	20 (1.3)
Previous malignancy			
No previous malignancy	1053 (77.4)	142 (74.0)	1195 (76.9)
Previous malignancy	308 (22.6)	50 (26.0)	358 (23.1)

Abbreviations: *CHF*, congestive heart failure; *COPD*, chronic obstructive pulmonary disease; *CKD*, chronic kidney disease; *IHD*, ischaemic heart disease; *NSCLC*, non-small cell lung cancer; *SCLC*, small cell lung cancer; *TIA*, transient ischaemic attack.

References

1. Liang, W. *et al.* Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. *The Lancet Oncology* **21,** 335–337 (Mar. 2020).

2. Abravan, A., Faivre-Finn, C., Kennedy, J., McWilliam, A. & van Herk, M. Radiotherapy-Related Lymphopenia Affects Overall Survival in Patients With Lung Cancer. *Journal of Thoracic Oncology* **15,** 1624–1635 (Oct. 2020).

3. Clift, A. K. *et al.* Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *The BMJ* **371** (Oct. 2020).

4. *NHS Digital Clinical inclusion criteria* https://digital.nhs.uk/coronavirus/shielded-patient-list/methodology/rule-logic. *Date accessed: 28/06/2021.*

5. Cancer Research UK. *Lung cancer treatment statistics* https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/diagnosis-and-treatment#heading-Two. *Date accessed: 28/06/2021.*

6. Faivre-Finn, C. *et al.* Reduced Fractionation in Lung Cancer Patients Treated with Curative-intent Radiotherapy during the COVID-19 Pandemic. *Clinical Oncology* **32,** 481–489 (Aug. 2020).

7. Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics* **95** (July 2019).

8. Spencer, K. *et al.* The impact of the COVID-19 pandemic on radiotherapy services in England, UK: a population-based study. *The Lancet Oncology* **22,** 309–320 (Mar. 2021).

9. Cheung, P. *et al.* Phase II study of accelerated hypofractionated three-dimensional conformal radiotherapy for stage T1-3 N0 M0 non-small cell lung cancer: NCIC CTG BR.25. *Journal of the National Cancer Institute* **106** (Aug. 2014).

10. Antonoff, M. *et al.* COVID-19 Guidance for Triage of Operations for Thoracic Malignancies: A Consensus Statement From Thoracic Surgery Outcomes Research Network. *Annals of Thoracic Surgery* **110,** 692–696 (Aug. 2020).

11. United Kingdom Lung Cancer Coalition. *A review of the impact of COVID-19 on the lung cancer pathway and opportunities for innovation emerging from the health system response to the pandemic.* https://www.uklcc.org.uk/our-reports/october-2020/covid-19-matters. *Date accessed: 28/06/2021.*

12. Adizie, J. B. *et al.* Stage III Non-small Cell Lung Cancer Management in England. *Clinical Oncology* **31,** 688–696 (Oct. 2019).

13. Derosa, L. *et al.* The immuno-oncological challenge of COVID-19. *Nature Cancer* **1,** 946–964 (Oct. 2020).

14. Phillips, I., Sandhu, S., Lüchtenborg, M. & Harden, S. Stereotactic Ablative Body Radiotherapy Versus Radical Radiotherapy: Comparing Real-World Outcomes in Stage I Lung Cancer. *Clinical Oncology* **31,** 681–687 (Oct. 2019).

15. UK Government. *Coronavirus (COVID-19) in the UK* https://coronavirus.data.gov.uk/details/cases. *Date accessed: 28/06/2021.*

16. Guckenberger, M. *et al.* Practice recommendations for lung cancer radiotherapy during the COVID-19 pandemic: An ESTRO-ASTRO consensus statement. *Radiotherapy and Oncology* **146,** 223–229 (May 2020).

17. Garassino, M. C. *et al.* COVID-19 in patients with thoracic malignancies (TERAVOLT): first results of an international, registry-based, cohort study. *The Lancet Oncology* **21,** 914–922 (July 2020).

18. Lee, L. Y. *et al.* COVID-19 prevalence and mortality in patients with cancer and the effect of primary tumour subtype and patient demographics: a prospective cohort study. *The Lancet Oncology* **21,** 1309–1316 (Oct. 2020).

19. Fenwick, J. D., Faivre-finn, C., Franks, K. N., Hatton, M. Q. F. & Hospital, W. P. *Managing treatment gaps in radiotherapy of lung cancer during the COVID-19 pandemic* tech. rep. May (2020), 1–14.

20. Cox, J. D. *et al.* Interruptions of high-dose radiation therapy decrease longterm survival of favorable patients with unresectable nonsmall cell carcinoma of the lung: analysis of 1244 cases from 3 radiation therapy oncology group (RTOG) trials. *International Journal of Radiation Oncology, Biology, Physics* **27,** 493–498 (Oct. 1993).

Appendix B

In Regard to Fornacon-Wood et al.

This is a Letter to the Editor received in response to Chapter 6, 'Understanding the Differences Between Bayesian and Frequentist Statistics', published in the International Journal of Radiation Oncology-Biology-Physics 2022 Volume 115 Issue 1 p249-250.

Authors

Amit K. Chowdhry[1], Deborah Mayo[2], Stephanie L. Pugh[3], John Park[4], Clifton David Fuller[5] and John Kang[6]

Affiliations

[1] Department of Radiation Oncology, University of Rochester Medical Center, Rochester, New York.

[2] Department of Philosophy, Virginia Tech, Blacksburg, Virginia.

[3] NRG Oncology Statistical and Data Management Center, American College of Radiology, Philadelphia, Pennsylvania.

[4] Department of Radiation Oncology, Kansas City VA Medical Center, Kansas City, Missouri.

[5] Department of Radiation Oncology, MD Anderson Cancer Center, Houston, Texas.

[6] Department of Radiation Oncology, University of Washington, Seattle, Washington.

To the Editor:

We appreciate the authors bringing attention to controversies surrounding the use of Bayesian and frequentist statistics [1]. There are many benefits to frequentist statistics and disadvantages of Bayesian statistics which were not discussed in the referenced article. We write this accompanying letter to aim for a more balanced presentation of Bayesian and frequentist statistics.

With frequentist statistical significance tests, we can learn whether the data indicate there is a genuine effect or difference in a statistical analysis, as they have the ability to control type I and type II error probabilities [2]. Posteriors and Bayes factors do not ensure that the method rarely reports one treatment is better or worse than the other erroneously. A well-known threat to reliable results stems from the ease of using high powered methods to data-dredge and try to hunt for impressive-looking results that fail to replicate with new data. However, the Bayesian assessment is not altered by things like stopping rules—at least not without violating inference by Bayes theorem [3]. The frequentist account [4], by contrast, is required to take account of such selection effects in reporting error probabilities. Another caution for those unfamiliar with practical Bayesian research is that estimation of a prior distribution is nontrivial. The priors they discuss are subjective degrees of belief, but there is considerable disagreement about which beliefs are warranted, even among experts. Furthermore, should conclusions differ if the prior is chosen by a radiation oncologist or a surgeon? [5] These considerations are some of the reasons why most phase 3 studies in oncology rely on frequentist designs.

The article equates frequentist methods with simple null hypothesis testing without alternatives, thereby overlooking hypothesis testing methods that control both type I and II errors. The frequentist takes account of type II errors and the corresponding notion of power. If a test has high power to detect a meaningful effect size, then failing to detect a statistically significant difference is evidence against a meaningful effect. Therefore, a P value that is not small is informative.

The authors write that frequentist methods do not use background information, but this is to ignore the field of experimental design and all of the work that goes into specifying the test (eg, sample size, statistical power) and critically evaluating the connection between statistical and substantive results. An effect that corresponds to a clinically meaningful effect, or effect sizes well warranted from previous studies, would clearly influence the design.

Although their article engenders important discussion, these differences between frequentist and Bayesian methods may help readers understand why so many researchers around the world still prefer the frequentist approach.

References

1. Fornacon-Wood, I. *et al.* Understanding the Differences Between Bayesian and Frequentist Statistics. *International Journal of Radiation Oncology Biology Physics* **112,** 1076–1082 (Apr. 2022).

2. Mayo, D. G. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (Cambridge University Press, 2018).

3. Ryan, E. G., Brock, K., Gates, S. & Slade, D. Do we need to adjust for interim analyses in a Bayesian adaptive trial design? *BMC Medical Research Methodology* **20,** 1–9 (June 2020).

4. Jennison, C. & Turnbull, B. W. *Group Sequential Methods with Applications to Clinical Trials* (CRC Press, 1999).

5. Stark, P. B. Comment on "The statistics wars and intellectual conflicts of interest" by D. Mayo. *Conservation Biology* **36,** e13861 (2022).

Appendix C

In Reply to Chowdhry et al.

This is a reply to the Letter to the Editor (Appendix B) written in response to Chapter 6, 'Understanding the Differences Between Bayesian and Frequentist Statistics', published in the International Journal of Radiation Oncology-Biology-Physics 2022 Volume 115 Issue 1 p250-251.

Authors

Isabella Fornacon-Wood[1], Hitesh Mistry[1], Corinne Johnson-Hart[2], Corinne Faivre-Finn[1,3], James P B O'Connor[1,4] and Gareth J Price[1]

Affiliations

[1] Division of Cancer Sciences, University of Manchester, Manchester, UK.

[2] Department of Medical Physics, The Christie NHS Foundation Trust, Manchester, UK.

[3] Department of Radiation Oncology, The Christie NHS Foundation Trust, Manchester, UK.

[4] Department of Radiology, The Christie NHS Foundation Trust, Manchester, UK.

To the Editor:

We thank the authors for their response [1] to our "statistics for the people" article [2] that aimed to introduce perhaps unfamiliar readers to Bayesian statistics and some potential advantages of their use. We agree that frequentist statistics are a useful and widespread statistical analytical approach, and we are not aiming to revisit the frequentist versus Bayesian arguments that have been well articulated in the literature [3–5]. However, there are a couple of points we would like to make.

First, we acknowledge that the majority of phase 3 studies use frequentist designs, and this has the advantage of facilitating meta-analyses using established techniques. However, we would argue that the reason such frequentist designs are so prevalent is likely to have as much to do with convention (from funders/regulators as well as from researchers themselves), the relative exposure of the 2 approaches in educational materials, and the historic difficulties in calculating Bayesian posteriors as it does with the arguments the authors make [6, 7].

Second, although we agree with Chowdhry et al that there are many challenges associated with the estimation of prior probability distributions, we note that similar arguments apply to effect size estimation, which they cite as a strength of the Neyman-Pearson/null hypothesis significance testing approach (ie, the use of power calculations to limit the risk of type II errors) [8, 9]. We would also re-enforce the point we make in the article about the importance of testing the influence of the prior (represented as the divergent beliefs of the hypothetical radiation oncologist and surgeon in the communication by Chowdhry et al) in the analysis results. If the data are strong enough, the posterior distributions will be in close enough agreement to convince both parties. As we noted, it is also possible to undertake Bayesian analyses without prior information, using an uninformative prior, in which case the analysis is driven directly by the data, as for a frequentist calculation. As an aside, there is continued debate about the relative merits and deficiencies of the different frequentist approaches to significance testing, particularly around the widespread use of the hybrid Neyman-Pearson/null hypothesis significance testing approach [10].

There have, undoubtedly, been important practice changing studies delivered using

frequentist approaches, but equally there is often erroneous interpretation of frequentist results [11, 12]. Indeed, the American Statistical Society had to issue guidance on the misuse of frequentist significance testing [13]. We would argue that the often-counterintuitive nature of null hypothesis significance testing likely makes such interpretation errors inevitable. One of the principal strengths of the Bayesian approach we discuss in the article is that the researcher can directly ask the question they are interested in, that is, what is the probable effect size and uncertainty of an intervention compared with an alternative.

Finally, as we note in our original concluding paragraph, "both frequentist and Bayesian approaches are useful for data analysis as long as they are interpreted correctly" and that "however data are analyzed, it is of utmost importance to be transparent and to correctly interpret the results in a manner consistent with... limitations in how data were collected." That is, that the quality of the whole study design and its execution, including its assumptions and data collection approaches, is likely more important to the inferences one can make than the analytical approach itself.

References

1. Chowdhry, A. K. *et al.* In Regard to Fornacon-Wood et al. [Letter to the Editor]. *International Journal of Radiation Oncology*Biology*Physics* **115,** 249–250 (Jan. 2023).

2. Fornacon-Wood, I. *et al.* Understanding the Differences Between Bayesian and Frequentist Statistics. *International Journal of Radiation Oncology Biology Physics* **112,** 1076–1082 (Apr. 2022).

3. Dienes, Z. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science* **6,** 274–290 (May 2011).

4. Harrell, F. E. *My Journey From Frequentist to Bayesian Statistics* https://www.fharrell.com/post/journey/. *Date accessed: 25/07/2022.*

5. Zampieri, F. G., Casey, J. D., Shankar-Hari, M., Harrell, F. E. & Harhay, M. O. Using bayesian methods to augment the interpretation of critical care trials. an overview of theory and example reanalysis of the alveolar recruitment for

acute respiratory distress syndrome trial. *American Journal of Respiratory and Critical Care Medicine* **203,** 543–552 (Mar. 2021).

6. Natanegara, F. *et al.* The current state of Bayesian methods in medical product development: survey results and recommendations from the DIA Bayesian Scientific Working Group. *Pharmaceutical statistics* **13,** 3–12 (Jan. 2014).

7. Clark, J. *et al.* Why are not There More Bayesian Clinical Trials? Perceived Barriers and Educational Preferences Among Medical Researchers Involved in Drug Development. *Therapeutic Innovation and Regulatory Science* **1,** 1–9 (Jan. 2022).

8. Fern, E. F. & Monroe, K. B. Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research* **23,** 89–105 (1996).

9. Marino, M. J. The use and misuse of statistical methodologies in pharmacology research. *Biochemical Pharmacology* **87,** 78–92 (Jan. 2014).

10. Perezgonzalez, J. D. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology* **6** (2015).

11. Hoekstra, R., Morey, R. D., Rouder, J. N. & Wagenmakers, E. J. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review* **21,** 1157–1164 (2014).

12. Greenland, S. *et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* **31,** 337–350 (Apr. 2016).

13. Wasserstein, R. L. & Lazar, N. A. The ASA's Statement on p-Values: Context, Process, and Purpose. *American Statistician* **70,** 129–133 (Apr. 2016).