Modernising the skin cancer MDT using novel technologies

Swansea University
Prifysgol Abertawe

Medical School
Ysgol Feddygaeth

Volume 1(2)

Mr Stephen Ali
BM MMedSc(Hons) PGCert(MedEd) PGCert(Mgmt) MAcadMEd
FHEA MRCS(Eng)

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy in Medical and Health Care Studies

Swansea University

2024

Summary

Aims

Multidisciplinary team meetings (MDTs) are an integral component of contemporary cancer care. However, variations in treatment uptake still exist and as such, MDTs have not been entirely successful in their aim of reducing variation in access to care. There are significant direct and indirect costs of MDT working and evidence demonstrates that some MDTs function more effectively than others. Regular meetings to discuss patients also present an opportunity cost to the National Health Service (NHS). The evidence base for cancer treatment is accumulating constantly, with the National Institute for Health and Care Excellence (NICE) regularly updating their advice. Innovative solutions to these problems would have health economic benefit, free up specialist time and improve the reproducibility of evidence-based decision making. It is the aim of this thesis to develop and validate methods of automating primary and secondary functions of the MDT using natural language processing (NLP) techniques.

Methods

Initial steps included conducting scoping surveys to understand the current state and challenges of conventional and remote skin cancer MDTs. A systematic review and meta-analysis of existing literature on ruled based NLP clinical decision support systems (CDSS) in cancer care were also undertaken to compare the effectiveness of these systems against human clinicians in medical decision-making. Further, development and validation of an NLP-based CDSS for

basal cell carcinoma (BCC) were carried out, focusing on providing treatment recommendations post-primary surgical interventions via a virtual MDT (vSMDT) platform. Additionally, an NLP-based information extraction system for BCC was created and validated to improve quality assurance and outcome benchmarking post-surgery.

Results

Survey responses uncovered prevailing practices and variances within specialist skin cancer multidisciplinary team meetings (SSMDTs) in the UK. It was discovered that only 26.0% of the SSMDTs were quorate by membership, with the major obstacle being the absence of clinical oncology presence. There was also a notable 69.0% achieving quoracy by meeting frequency, revealing considerable discrepancies and emphasising the significant need for standardisation and adherence to NICE quoracy standards. Further, the thesis provides insights into the effectiveness of MDT meetings, revealing a uniform belief in the importance of risk stratification and prioritisation of complex cases. There is a consensus on the need for protocolised treatment pathways and enhancements in meeting preparation and attendance. Evaluation of remote skin cancer MDTs in the post-COVID-19 era illustrated a comparative efficacy in communication and decision-making between virtual and in-person meetings, with a preference for a hybrid format for future interactions, emphasising the necessity for improvements in connectivity and integration. Innovations in automated information extraction in BCC histopathology were explored using an NLP system, which displayed high precision and recall, offering promising applications in improving the quality of cancer registry data. Moreover, the automation of a web-based model of care demonstrated high accuracy comparable with the literature, supporting the feasibility of a fully automated, virtual, web-based service model for hosting the skin MDT. Additionally, advancements like ChatGPT have

proven the ability to generate clinically accurate and human-like clinical letters, paving the way for a significant reduction in clinical and administrative workloads and a standardization in the dissemination of patient information.

Conclusion

The synthesis of the findings in this thesis underscores the transformative potential of advancements in computational methods like NLP in surgical practice. It illustrates the critical role of restructuring MDT discussions, emphasising the importance of tumour-specific guidance and protocolised treatment pathways. The insights derived from this research support the transformative potential of virtual MDTs and AI-driven solutions in refining service provision, clinical communications, and decision-making processes. It also highlights the continuous need for reassessing and revising clinical standards through innovative technologies to achieve optimal excision rates and enhance patient care. Overall, this thesis advocates for the accelerated integration of novel technologies like NLP, with a human in the loop to enable clinicians to give more time to care for patients in the digital era of healthcare.

Declaration and Statements

Declaration

This work has not been previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed 

Date 02/10/2023

Statement 1

This thesis is the results of my own work, except when otherwise stated.
Other sources are acknowledged by appropriate citation. A bibliography is appended.

Signed 

Date 02/10/2023

Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed 

Date 02/10/2023

Statement 3

The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed 

Date 02/10/2023

Table of Contents

Acknowledgments

I extend my sincere gratitude to Professor Iain Whitaker, Professor Hayley Hutchings, and Mr. Thomas Dobbs for their invaluable expertise, guidance, knowledge, and insightful ideas throughout my candidature. Their commitment to academic and scientific rigor has immensely contributed to my professional and personal development.

Special thanks to Dr. Owen Pickrell's team, in particular Mr. Huw Strafford, NLP Research Assistant at Health Data Research UK, Swansea University Medical School, for his comprehensive training in Natural Language Processing. His patience and expertise have been pivotal in my research.

I am grateful to Mr. Matt Jovic, Data Scientist and Research Assistant at the Reconstructive Surgery & Regenerative Medicine Research Centre, Swansea University Medical School, for his extensive training in SQL which was instrumental in managing and querying complex datasets effectively.

My appreciation extends to Dr. Namor Wyn Williams from the Department of Cellular Pathology, Morriston Hospital, Swansea, for facilitating access to the laboratory information management system, which was crucial for the retrieval and analysis of pathology data.

I want to express my thanks to Mr. Rhys Whelan, Library Services Manager at Swansea Bay University Health Board, for his invaluable advice on developing a robust literature search strategy, ensuring the comprehensive and relevant inclusion of existing works in my research.

I am also thankful to Mr. Octavian Parkes, Research Programme Manager at the Reconstructive Surgery & Regenerative Medicine Research Centre, for his critical review of the thesis, providing invaluable feedback and insights that significantly enhanced the quality of my work.

Finally, my deepest appreciation goes to my wife, Rosa, for her unwavering support, patience, and encouragement throughout the journey of my Ph.D. Her constant support and understanding have been my source of strength and inspiration.

Funding Acknowledgment

This work was supported by the Welsh Clinical Academic Training Fellowship, the Paton Masser grant from the British Association of Plastic, Reconstructive, and Aesthetic Surgeons, and the Topol fellowship. The Reconstructive Surgery & Regenerative Medicine Research Centre's pivotal work is funded by The Scar Free Foundation and Health and Care Research Wales. I am particularly thankful to The Scar Free Foundation, a unique medical research charity focused on scarring, for sharing the vision of achieving scar-free healing within a generation and for their substantial support in realising this vision through research.

List of Tables

List of Figures

List of Abbreviations

1,25(OH)2D3 - 1,25-Dihydroxyvitamin D3

25(OH)D3 - 25-Hydroxyvitamin D3

AI - Artificial Intelligence

AL - Active Learning

ANOVA - Analysis of Variance

AoMRC - Academy of Medical Royal Colleges

APACHE II - Acute Physiology and Chronic Health Evaluation II

API - Application Programming Interface

ASA - American Society of Anesthesiologists

ATP - Adenosine Triphosphate

BAD - British Association of Dermatologists

BAPRAS - British Association of Plastic, Reconstructive and Aesthetic Surgeons

BCC - Basal Cell Carcinoma

BRAF - B-Raf Proto-Oncogene, Serine/Threonine Kinase

CARET - Classification And REgression Training

CDKN2A - Cyclin-Dependent Kinase Inhibitor 2A

CDSS - Clinical Decision Support System

CI - Confidence Interval

CLND - Completion Lymphadenectomy

CNNs - Convolutional Neural Networks

CNS - Clinical Nurse Specialist

COP - Consultant Outcomes Publication

CPT - Current Procedural Terminology

CRUK - Cancer Research UK

CSV - Comma-Separated Values

DNA - Deoxyribonucleic Acid

DRG - Diagnosis-Related Group

EGFR - Epidermal Growth Factor Receptor

EHR - Electronic Health Record

FOI - Freedom of Information

GATE - General Architecture for Text Engineering

GDPR - General Data Protection Regulation

GLM - Generalised Linear Model

GP - General Practitioner

GPT - Generative Pre-trained Transformer

GUI - Graphical User Interface

HQIP - Healthcare Quality Improvement Partnership

HRA - Health Research Authority

HSD - Honestly Significant Difference

ICD - International Classification of Diseases

ICT - Information Communication Technology

JAPE - Java Annotation Patterns Engine

LLM - Large Language Model

LLMs - Large Language Models

LSMDT - Local Skin Cancer Multidisciplinary Teams

MAL-PDT - Methyl Aminolevulinate Photodynamic Therapy

MAPK - Mitogen-Activated Protein Kinase

MC1R - Melanocortin 1 Receptor

MDT - Multidisciplinary Team

MeSH - Medical Subject Headings

ML - Machine Learning

MM - Malignant Melanoma

MMS - Mohs Micrographic Surgery

MSCs - Mesenchymal Stromal Cells

MSS - Melanoma-Specific Survival

NCAPOP - National Clinical Audit and Patient Outcomes Programme

NCCN - National Comprehensive Cancer Network

NCPR - National Cancer Peer Review

NCRAS - National Cancer Registration and Analysis Service

NER - Named Entity Recognition

NHS - National Health Service

NICE - National Institute for Health and Care Excellence

NLP - Natural Language Processing

NMSC - Non-Melanoma Skin Cancer

NPV - Negative Predictive Value

NRAS - Neuroblastoma RAS Viral Oncogene Homolog

OMFS - Oral and Maxillofacial Surgery

OPCS-4 - Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures Version 4

OR - Odds Ratio

OXPHOS - Oxidative Phosphorylation

P-POSSUM - Portsmouth-Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity

PD-L1 - Programmed Death-Ligand 1

PLASTA - Plastic Surgery Trainees Association

PPV - Positive Predictive Value

PTH - Parathyroid Hormone

QST - Quality Surveillance Team

RCPath - Royal College of Pathologists

RDMS - Relational Database Management System

RNNs - Recurrent Neural Networks

ROS - Reactive Oxygen Species

SCC - Squamous Cell Carcinoma

SE - Standard Error

SIAG - Special Interest and Advisory Group

SIGN - Scottish Intercollegiate Guidelines Network

SLNB - Sentinel Lymph Node Biopsy

SNOMED RT - Systematised Nomenclature of Medicine Reference Terminology

SNOMED-CT - Systematised Nomenclature of Medicine Clinical Terms

SoC - Standard of Care

SSMDT - Skin Cancer Specialist Multidisciplinary Team

TLR-7 - Toll-Like Receptor 7

TNM8 - Tumour–Nodes–Metastasis 8th edition

TSG - Tumor Suppressor Gene

UICC8 - Union for International Cancer Control 8th edition

UKIACR - UK and Ireland Association of Cancer Registries

UMLS - Unified Medical Language System

USDHHS - United States Department of Health and Human Services

UV – Ultraviolet

VEGF - Vascular Endothelial Growth Factor

vSMDT - Virtual Specialist Multidisciplinary Team

WHO - World Health Organization

WLE - Wide Local Excision

Publications and Presentations

Peer reviewed publications

ALI SR, Dobbs TD, Whitaker IS. Proposal of a new model of national skin audit and data submission. **Clinical and Experimental Dermatology**. 2024 Jan 20:llae024.

PMID: 38245000

ALI SR, Dobbs TD, Jovic M, Strafford H, Lacey AS, Williams N, Pickrell WO, Hutchings HA, Whitaker IS. Revisiting basal cell carcinoma clinical margins: leveraging natural language processing and multivariate analysis with updated Royal College of Pathologists histological reporting standards. **Journal of Plastic, Reconstructive & Aesthetic Surgery**. 2024 Jan 1;88:443-51.

PMID: 38091687

ALI SR, Dobbs TD, Tarafdar A, Strafford H, Fonferko-Shadrach B, Lacey AS, Pickrell WO, Hutchings HA, Whitaker IS. Natural language processing to automate a web-based model of care and modernise skin cancer multidisciplinary team meetings? **British Journal of Surgery**. 2024 Jan;111(1):znad347

PMID: 38198154

ALI SR, Dobbs TD, Whitaker IS. The digital surgeon: unleashing the power of NLP in modern surgery. **Bulletin of the Royal College of Surgeons of England**. 2023 Nov;105(8):384-6.

ALI SR, Dobbs TD, Mohamedbhai H, Hutchings HA, Whitaker IS. Evaluating remote skin cancer multidisciplinary team meetings in the United Kingdom post-COVID-19. **Journal of Plastic, Reconstructive & Aesthetic Surgery**. 2023 Apr 19.

PMID: 37352621

ALI SR, Dobbs TD, Jovic M, Hutchings HA, Whitaker IS. Improving the effectiveness of multidisciplinary team meetings in skin cancer: Analysis of national Cancer Research UK survey responses**. Journal of Plastic, Reconstructive & Aesthetic Surgery**. 2023 Jul 1;82:141-51.

PMID: 37167715

ALI SR, Dobbs TD, Jovic M, Strafford H, Fonferko-Shadrach B, Lacey AS, Williams N, Pickrell WO, Hutchings HA, Whitaker IS. Validating a novel natural language processing pathway for automated quality assurance in surgical oncology: incomplete excision rates of 34 955 basal cell carcinomas. **British Journal of Surgery**. 2023 Mar 20:znad055.

PMID: 36935397

ALI SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. **The Lancet Digital Health**. 2023 Mar 7.

PMID: 36894409

ALI SR, Strafford H, Dobbs TD, Fonferko-Shadrach B, Lacey A, Pickrell OW, Whitaker ISW. Development and validation of an automated basal cell carcinoma histopathology information extraction system using natural language processing. **Frontiers in Surgery – Reconstructive & Plastic Surgery**. 2022. 1085.

PMID: 36439548

ALI SR, Dobbs TD, Hutchings HA, Whitaker IS. Composition, Quoracy and Cost of Specialist Skin Cancer Multidisciplinary Team Meetings in the United Kingdom. **Journal of Plastic, Reconstructive & Aesthetic Surgery**. 2021 Jun 6; S1748-6815(21)00271-0.

PMID:34187763

Presentations

Invited poster presentation at Plastic Surgery The Meeting 2023 Austin, TX, USA. ALI SR, Dobbs TD, Jovic M, Strafford H, Fonferko-Shadrach B, Lacey AS, Williams N, Pickrell WO, Hutchings HA, Whitaker IS. The use of natural language processing to enhance quality assurance in plastic surgery – the largest global study of basal cell carcinoma incomplete excision rates in 34,955 lesions.

Invited oral presentation at BAPRAS & NVPC Combing Meeting Amsterdam, Netherlands 2023. ALI SR. The use of big data, AI and natural language processing to enhance quality assurance in oncological surgery. Amsterdam, Netherlands

Oral presentation at BAPRAS & NVPC Combing Meeting 2023 Amsterdam, Netherlands. ALI SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters.

Oral presentation at 14th Meeting of the European Plastic Surgery Research Council 2023 Stockholm, Sweden. ALI SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters.

Oral presentation at BAPRAS x NCVP Free Paper Webinar 2023. ALI SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters.

Oral presentation at BAPRAS Winter Meeting Nottingham 2020. ALI SR, Dobbs T, Whitaker I. Composition, quoracy and cost of Specialist Skin Cancer Multidisciplinary Team Meetings in the United Kingdom.

Chapter One:

Introduction, Research Question Aims and Objectives

1.1 The embryology, anatomy, and physiology of human skin

The human skin, or "cutis" in Latin, is the largest organ in the body, comprising approximately 15-20.0% of an individual's total body weight (Proksch et al., 2008). It has been recognised as a vital structure for centuries, with ancient Greek and Roman physicians such as Galen and Celsus making observations on its protective and sensory functions. The term "integumentum" in Latin, which refers to a covering or an outer layer, also represents the concept of the skin as a protective barrier.

Throughout history, the importance of the skin as a multifunctional organ became more apparent, with increasing understanding of its roles in thermoregulation, immunological defence, and vitamin D synthesis. Notably, the Renaissance period saw significant advancements in the study of human anatomy, with scholars such as Leonardo da Vinci and Andreas Vesalius documenting the structure and function of the skin in great detail. As an essential structure, the skin provides protection from mechanical injury, pathogens, and environmental stressors, while also aiding in thermoregulation, sensation, immunological defence, and vitamin D homeostasis.

In the 19th and 20th centuries, advances in microscopy and histological techniques provided new insights into the skin's cellular and molecular organization. This led to a deeper understanding of the skin's embryonic development, its various layers (epidermis, dermis, and hypodermis), and the appendages it houses (e.g., hair follicles, sebaceous and sweat glands).

1.1.1 Embryology of human skin

The development of human skin begins during the early stages of embryogenesis. Skin development can be divided into three main stages: ectodermal differentiation, dermo-epidermal junction formation, and stratification/maturation.

1.1.2 Ectodermal differentiation

During the third week of embryogenesis, the outer layer of the embryo, known as the ectoderm, differentiates into the neural ectoderm and the surface ectoderm. The surface ectoderm will give rise to the epidermis, hair, nails, and various glands, while the neural ectoderm forms the central nervous system (Keith L. Moore et al., 2018).

1.1.3 Dermo-epidermal junction formation

The dermo-epidermal junction is formed during the fourth week of embryogenesis. The mesodermal layer, located between the ectoderm and endoderm, differentiates into the dermis, which is composed of connective tissue, blood vessels, and nerves. The basement membrane, a specialised extracellular matrix, is formed at this stage, serving as the interface between the developing epidermis and dermis (Keith L. Moore et al., 2018).

1.1.4 Stratification and maturation

During the fifth to seventh weeks of embryogenesis, the developing epidermis undergoes stratification and maturation, giving rise to its characteristic multi-layered structure. The epidermis differentiates into four primary layers: the stratum basale, stratum spinosum, stratum granulosum, and stratum corneum (Keith L. Moore et al., 2018).

1.1.5 Anatomy of human skin

The skin is divided into three primary layers: the epidermis, dermis, and hypodermis (subcutaneous tissue) (Proksch et al., 2008).

1.1.6 Epidermis

The epidermis is the outermost layer of skin, consisting of keratinocytes, melanocytes, Langerhans cells, and Merkel cells. The stratum basale is the deepest layer, containing proliferating keratinocytes and melanocytes. The stratum spinosum, stratum granulosum, and stratum corneum follow, with the latter forming a critical barrier against environmental stressors and pathogens (Holbrook & Odland, 1975; Venus et al., 2010).

1.1.7 Dermis

The dermis is a dense connective tissue layer, containing fibroblasts, mast cells, macrophages, blood vessels, and nerves. It is divided into the papillary dermis, which is closely associated with the epidermis, and the deeper reticular dermis. The dermis provides structural support and nutrition to the epidermis and houses various appendages, such as hair follicles, sweat glands, and sebaceous glands (Proksch et al., 2008; Venus et al., 2010).

1.1.8 Hypodermis

The hypodermis, also known as the subcutaneous tissue, is the innermost layer of skin, consisting primarily of adipose tissue. It provides insulation, energy storage, and cushioning, while also serving as a conduit for nerves and blood vessels (Proksch et al., 2008; Venus et al., 2010).

1.1.9 Skin appendages

In addition to the primary skin layers, specialised derivatives from the epidermis and dermis play crucial roles in various functions. Hair follicles produce hair shafts for protection, sensation, and thermoregulation. Sebaceous glands, associated with hair follicles, secrete sebum to lubricate hair and skin. Sweat glands are of two types: eccrine, which are distributed throughout the body, and apocrine, which are found mainly in the axillary and genital regions. Nails, made of keratinised epidermal cells, protect fingertips, aid in grasping objects, and serve as tools (Proksch et al., 2008; Venus et al., 2010).

1.1.10 Physiology of human skin

The skin is involved in numerous physiological processes, such as barrier function, thermoregulation, sensation, immune defence and vitamin D homeostasis.

1.1.11 Barrier function

The stratum corneum serves as a formidable barrier against environmental stressors, pathogens, and water loss. It is composed of flattened, dead keratinocytes called corneocytes, which are surrounded by a lipid matrix. This matrix restricts the permeability of the stratum corneum, ensuring that only small, lipophilic molecules can pass through the skin (Proksch et al., 2008; Venus et al., 2010).

1.1.12 Thermoregulation

The skin plays a crucial role in maintaining the body's temperature. Blood vessels in the dermis dilate or constrict to regulate heat loss, while sweat glands produce sweat to dissipate heat through evaporative cooling. Additionally, the hypodermis acts as an insulating layer, helping to retain heat in cold environments (Venus et al., 2010).

1.1.13 Sensation

The skin is densely innervated by sensory nerve endings, allowing it to detect and transmit information about touch, pressure, pain, and temperature. Merkel cells, located in the epidermis, contribute to the sensation of light touch. Free nerve endings, associated with nociception and thermal sensation, are present throughout the dermis and epidermis. Other specialised nerve endings, such as Meissner's corpuscles and Pacinian corpuscles, are responsible for detecting vibrations and deep pressure, respectively (Venus et al., 2010).

1.1.14 Immune defence

The skin is an essential component of the body's immune system, acting as a physical barrier and providing an immunological defence against invading pathogens. Langerhans cells, a type of dendritic cell found in the epidermis, are key players in the skin's immune response. They capture and process antigens, then migrate to nearby lymph nodes to activate T cells, initiating an adaptive immune response (Venus et al., 2010). Additionally, the skin's microbiome, composed of a diverse array of commensal bacteria, plays a vital role in maintaining skin health and preventing pathogenic colonization (Grice & Segre, 2011).

1.1.15 Vitamin D homeostasis

Vitamin D homeostasis is integral to skin physiology and overall health. As a unique vitamin, it can be synthesised in the skin upon exposure to ultraviolet B (UVB) radiation from sunlight, and it plays a vital role in various physiological processes, such as calcium absorption, bone health, and immune function.

The synthesis of vitamin D is initiated in the epidermis when 7-dehydrocholesterol, a cholesterol precursor, absorbs UVB radiation and undergoes photolysis to form pre-vitamin

D3. Pre-vitamin D3 isomerizes to cholecalciferol, or vitamin D3, which is then transported to the liver and hydroxylated to form 25-hydroxyvitamin D3 [25(OH)D3], also known as calcifediol. Subsequently, 25(OH)D3 is further hydroxylated in the kidneys, producing the biologically active hormone 1,25-dihydroxyvitamin D3 [1,25(OH)2D3], also known as calcitriol.

Vitamin D homeostasis is maintained by feedback mechanisms involving parathyroid hormone (PTH), calcium, and phosphate levels. Low serum calcium levels prompt the parathyroid glands to secrete PTH, stimulating the conversion of 25(OH)D3 to 1,25(OH)2D3 in the kidneys. The active 1,25(OH)2D3 promotes calcium absorption in the intestines and collaborates with PTH to regulate calcium and phosphate homeostasis in the body (Bikle, 2014).

In the context of skin physiology, vitamin D plays a crucial role. The skin serves as the primary site of vitamin D synthesis, with factors such as sun exposure, latitude, season, skin pigmentation, and sun protection influencing cutaneous production. Furthermore, the skin functions as a target tissue for the active form of vitamin D, 1,25(OH)2D3, which modulates epidermal cell proliferation, differentiation, and immune function. These regulatory actions contribute to the maintenance of the skin barrier function and overall skin health (Holick, 2004).

1.2 Skin cancer pathophysiology

1.2.1 Classification

There are two main categories of skin cancer: non-melanoma skin cancer (NMSC) and malignant melanoma (MM). NMSC arises from keratinocytes, which are the most common cells in the epidermis. It includes basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), which arise from the basal and squamous cells, respectively, in the epidermis. BCC is the most common type of skin cancer and arises from the basal cells in the stratum basale of the epidermis. SCC arises from the keratinocytes in the stratum spinosum of the epidermis. On the other hand, MM arises from melanocytes which are found in the basal layer of the epidermis. The most common type of MM arises in the skin, but it can also arise in other parts of the body, such as the eyes and mucous membranes (Netscher et al., 2011).

1.2.2 Pathophysiology

The pathophysiology of NMSC can be complex, as there are various factors that can contribute to its development. Prolonged exposure to ultraviolet (UV) radiation is a major cause, as it can lead to the accumulation of DNA damage in keratinocytes (Narayanan et al., 2010). This damage can cause mutations that activate oncogenes or inactivate tumour suppressor genes (TSGs), leading to uncontrolled cell growth and division. An oncogene is a gene that has the potential to cause cancer when it is mutated or overexpressed. Normally, oncogenes are involved in regulating cell growth, division, and differentiation. However, when they are activated or amplified, they can stimulate uncontrolled cell growth and division, leading to the development of a tumour. A TSG, on the other hand, is a gene that normally inhibits cell proliferation and promotes cell death to prevent the development of cancer. When a TSG is mutated or deleted, it can no longer perform its normal function, allowing cells to grow and

divide uncontrollably. TSGs can also be inactivated by epigenetic changes such as DNA methylation or histone modification (Hanahan & Weinberg, 2000). In BCC, the key driver of tumorigenesis is the activation of the Sonic hedgehog signalling pathway, which promotes the proliferation of basal cells (Epstein, 2008). In SCC, mutations in the p53 TSG are commonly observed, leading to the loss of cell cycle control and DNA repair capacity (Brash et al., 1991).

The pathophysiology of MM involves the transformation of melanocytes into malignant cells due to a combination of genetic and environmental factors. Mutations in genes such as B-Raf proto-oncogene, serine/threonine kinase (BRAF) and neuroblastoma RAS viral oncogene homolog (NRAS), which are part of the mitogen-activated protein kinase (MAPK) signalling pathway that regulates cell proliferation and survival, are commonly observed in MM (Akbani et al., 2015; Davies et al., 2002). In addition, prolonged exposure to UV radiation can induce DNA damage in melanocytes, leading to the formation of pyrimidine dimers that can cause mutations in genes such as p16 and p53, which are involved in cell cycle regulation and DNA repair (Hocker & Tsao, 2007).

1.2.3 Relevance to the hallmarks of cancer

The hallmark features of cancer described by Hanahan and Weinberg provide a comprehensive framework for understanding the pathophysiology of skin cancer (Hanahan & Weinberg, 2000). In the case of NMSC, mutations in oncogenes and TSGs that promote cell growth and division sustain proliferative signalling, while mutations in genes such as p53 that normally regulate cell cycle and DNA repair contribute to the evasion of growth suppressors and resistance to cell death. Additionally, the activation of signalling pathways such as the Sonic hedgehog pathway in BCC and the MAPK pathway in both SCC and MM can contribute to the enabling of replicative immortality (Brash et al., 1991; Davies et al., 2002; Epstein, 2008).

Angiogenesis, or the formation of new blood vessels to supply nutrients and oxygen to tumour cells, is also an important hallmark of NMSC (Hanahan & Weinberg, 2000). Tumours must induce angiogenesis to grow beyond a certain size, and the expression of angiogenic factors such as vascular endothelial growth factor (VEGF) has been observed in NMSC (Bowden et al., 2002). Invasion and metastasis are key features of advanced NMSC, with SCC being more aggressive in this regard than BCC. Studies have reported differences in the invasive and metastatic potential of these two types of skin cancer. BCC rarely metastasises, with an estimated rate of 0.0028% to 0.55% (Rubin et al., 2005). The median interval between the appearance of the tumour and metastasis is estimated to be nine years (Rubin et al., 2005). In contrast, SCC has a higher metastatic potential, with an overall risk of metastasis estimated at 4.0% (Brantsch et al., 2008).

In MM, sustained proliferative signalling is driven by mutations in genes such as BRAF and NRAS, which activate the MAPK pathway. Additionally, mutations in genes such as p16 and p53 contribute to the evasion of growth suppressors and resistance to cell death. The avoidance of immune destruction is also an important hallmark of MM, as melanoma cells can express proteins that inhibit the immune response, such as programmed death-ligand 1 (PD-L1) (Ribas & Wolchok, 2018). Replicative immortality is enabled by the upregulation of telomerase activity, which allows cells to continue dividing beyond their normal limits (Shay & Wright, 2019).

Deregulating cellular energetics is another emerging hallmark of cancer, as tumour cells often exhibit altered metabolism to support their high proliferation rates (Fouad & Aanei, 2017). This can include a shift towards glycolysis even in the presence of oxygen, known as the

Warburg effect, a phenomenon first described by Otto Heinrich Warburg (1883-1970), a German physiologist and Nobel laureate, in 1924 (Warburg, 1924). Warburg observed that cancer cells tend to consume higher levels of glucose and produce more lactate compared to normal cells, despite the availability of oxygen. This metabolic reprogramming, now referred to as the Warburg effect, allows cancer cells to support rapid growth and generate building blocks required for their high proliferation rates. Although initially it was believed that the Warburg effect in tumour cells was solely associated with the upregulation of glycolysis to produce ATP under normoxic conditions, scientific advancements in the last 15 years have revealed the importance of oxidative phosphorylation (OXPHOS) alongside glycolysis in malignant cells. Melanoma cells exhibit metabolic plasticity, dynamically switching between glycolysis and OXPHOS, which provides them with a survival advantage, enabling adaptation to harsh conditions and chemoresistance pathways. Moreover, the concurrent upregulation of both OXPHOS and glycolysis (metabolic symbiosis) has been proven crucial for melanoma progression. The tumour microenvironment (TME) plays a vital role in promoting melanoma progression, invasion, and metastasis. Mesenchymal stromal cells (MSCs) within the TME display a symbiotic relationship with melanoma, shielding tumour cells from apoptosis and granting chemoresistance (Kumar et al., 2021).

1.3 Epidemiology of skin cancer

1.3.1 Incidence

In the United Kingdom (UK), NMSC is the most common cancer whilst MM is the fifth most common cancer (Cancer Research UK, 2021-e). Since the early 1990s, the incidence of skin cancer has been steadily increasing. As reported by Cancer Research UK (CRUK), an average of 155,985 new NMSC cases were diagnosed per year during 2016-2018 (Cancer Research

UK, 2021-e). The age-standardised incidence rate for NMSC is substantially higher than that of MM at 252.4 per 100,000 people in the UK during 2016-2018 (Cancer Research UK, 2021-e). A recent systematic review highlights the incidence of BCC in the UK has been increasing annually by up to 4.0% (Ibrahim et al., 2023). In Wales, the rates have been rising even more rapidly, with annual increases of up to 6.6% and 1.6% for BCC and SCC, respectively (Ibrahim et al., 2023). Furthermore, an inverse relationship has been observed between the incidence of BCC/SCC and social deprivation. While the elderly population continues to be the most at risk, the 30-49 age group has demonstrated a significant growth rate of approximately 4.0% in recent years (Ibrahim et al., 2023). The Welsh data demonstrates one of the highest published incidences within the UK and Europe with European age-standardised incidence for BCC in 2018 224.6 per 100,000 person-years (Ibrahim et al., 2023).

MM ranks as the fifth most common cancer in the UK and the second most common type of skin cancer (Cancer Research UK, 2021-a). CRUK data reveals that there was an average of 16,744 new MM cases diagnosed per year during 2016-2018, accompanied by an age-standardised incidence rate of 26.8 per 100,000 people (Cancer Research UK, 2021-a). The incidence of MM in the UK has experienced a dramatic surge since the early 1990s, with rates more than doubling during this period. CRUK data reveals that in 1993-1995, the European age-standardised incidence for MM was 11.2 cases per 100,000 people (Cancer Research UK, 2021-a). By 2016-2018, this figure had risen to 26.8 cases per 100,000 people, highlighting the significant growth in MM incidence over the past few decades (Cancer Research UK, 2021-a).

The rising incidence of NMSC and MM can be attributed to several interconnected factors. Increased sun exposure due to changes in outdoor activities, clothing styles, and attitudes towards tanning and sunbathing has heightened the risks of developing skin cancer (Lomas,

Leonardi-Bee, et al., 2012). The growing popularity of tanning beds, which emit artificial UV radiation, has further contributed to the surge in skin cancer cases (Nikolaou & Stratigos, 2014). Additionally, as the population ages, the cumulative effects of sun exposure over their lifetimes place the elderly at a higher risk of developing NMSC and MM (Lomas, Leonardi-Bee, et al., 2012). Given these factors, the future demand on healthcare services is likely to be driven by further increases in skin cancer incidence, as the long-term effects of sun exposure from several generations of 'beach culture' become evident in the clinic.

1.3.2 Risk factors

Skin cancer remains a significant public health challenge in the UK, and identifying the risk factors associated with its development is crucial for implementing effective prevention, early detection, and treatment strategies.

The primary risk factor for skin cancer is exposure to UV radiation, which originates from both natural sources, such as sunlight, and artificial sources, such as tanning beds (Armstrong & Kricker, 2001). UV radiation triggers DNA damage in skin cells, leading to the formation of cyclobutane pyrimidine dimers and 6-4 photoproducts (Cadet et al., 2005). While some of these lesions can be repaired by cellular mechanisms like nucleotide excision repair, incomplete or inaccurate repair can result in mutations that contribute to skin cancer development (Marteijn et al., 2014).

Thomas Fitzpatrick (1919-2003), a Harvard Medical School dermatologist and researcher, developed the Fitzpatrick skin type classification system in 1975 to better understand individual variations in skin colour and their sensitivity to UV radiation (Fitzpatrick, 1975). Those with Fitzpatrick Skin Types I and II, characterised by fair skin, light hair, and light eyes,

have a reduced ability to produce protective melanin, which acts as a natural sunscreen by absorbing and scattering UV radiation. Consequently, they are more susceptible to UV-induced DNA damage and skin cancer.

UVA and UVB are two types of UV radiation that can damage the skin (de Gruijl, 2002). UVA radiation, with wavelengths between 315 and 400 nm, penetrates deeper into the skin, reaching the dermis and causing long-term damage (Narayanan et al., 2010). Although UVA radiation is less energetic than UVB, it is more abundant in sunlight and can still cause DNA damage (Narayanan et al., 2010). UVA radiation primarily causes indirect DNA damage through the generation of reactive oxygen species (ROS), such as singlet oxygen, superoxide anion, and hydroxyl radicals, which can lead to oxidative base modifications, DNA strand breaks, and mutations (Cadet et al., 2015).

UVB radiation, with wavelengths between 280 and 315 nm, primarily affects the epidermis and is responsible for sunburn and a significant contributor to skin cancer (Narayanan et al., 2010). UVB radiation induces direct damage to DNA by causing the formation of cyclobutane pyrimidine dimers and pyrimidine (6-4) pyrimidone photoproducts, which can lead to mutations if left unrepaired (Narayanan et al., 2010). Additionally, UVB can cause indirect damage through the generation of ROS that can oxidize DNA bases and proteins, potentially leading to DNA strand breaks and further mutations (Cadet et al., 2015).

A history of sunburn, particularly severe or frequent sunburns during childhood and adolescence, is indicative of acute UV damage to the skin, primarily from UVB radiation (Whiteman et al., 2001). This damage can accumulate over time, leading to the activation of

oncogenes or inactivation of TSGs, ultimately increasing the risk of skin cancer (Dennis et al., 2008).

While the risk of skin cancer increases with age due to the cumulative effects of sun exposure, it is important to note that skin cancer can also affect younger individuals, particularly those with other risk factors. A family history of skin cancer, particularly MM, can significantly increase an individual's risk, partly due to shared genetic factors, such as mutations in the cyclin-dependent kinase Inhibitor 2A (CDKN2A) and melanocortin 1 receptor (MC1R) genes, which influence melanocyte function and MM susceptibility (Fargnoli et al., 2010). The presence of numerous moles or atypical naevi is associated with an increased risk of skin cancer, particularly MM (Gandini et al., 2005). Dysplastic naevi often exhibit histological and molecular features that resemble MM, including increased cell proliferation, atypia, and alterations in key signalling pathways (Gandini et al., 2005). Genodermatoses, a group of inherited skin disorders, can predispose individuals to skin cancer (Itin & Fistarol, 2004). These genetic conditions often involve defects in DNA repair mechanisms or other cellular pathways, leading to an increased susceptibility to UV-induced DNA damage and skin cancer development (Itin & Fistarol, 2004). Examples of genodermatoses include xeroderma pigmentosum, which is characterised by defects in nucleotide excision repair (Cleaver, 1968), and Gorlin syndrome, caused by mutations in the PTCH1 gene (Gorlin, 2004). Chronic wounds, such as non-healing ulcers, can also heighten the risk of skin cancer, particularly SCC, due to the ongoing inflammation and tissue damage that promote cancerous changes in the affected skin cells. One example of this is Marjolin's ulcers, which are malignant transformations of long-standing non-healing ulcers or scars (Kowal-Vern & Criswell, 2005). Named after French surgeon and pathologist Jean-Nicolas Marjolin (1780-1850), who first described the condition in the 19th century, these ulcers are associated with an increased risk

of developing SCC. The chronic inflammatory environment in these wounds can lead to the production of pro-inflammatory cytokines, growth factors, and ROS, which can drive genomic instability, cell proliferation, and angiogenesis, ultimately contributing to skin cancer development (Schäfer & Werner, 2008).

1.3.3 Mortality

While MM is less common than BCC and SCC, it is responsible for the majority of skin cancer-related deaths in the UK. During 2017-2019, there were approximately 2,341 MM-related deaths per year, with an age-standardised mortality rate of 3.7 per 100,000 people (Cancer Research UK, 2021-b). The survival rate for MM has improved significantly over the years, as evidenced by CRUK data from England and Wales. Between 1971-1972, the 5-year net survival for adults was 52.1%, increasing to 90.4% by 2010-2011 (Cancer Research UK, 2021-c). This improvement is also reflected in the 5-year survival rates for different stages of the disease, according to Cancer Research UK data for adults diagnosed 2013-2017, followed up to 2018. In stage I MM, the cancer is confined to the epidermis or has just begun to invade the upper part of the dermis. The 5-year survival rate for this stage is now at 99.6%, highlighting the limited growth and lack of metastasis at this stage. For stage II MM, the cancer has grown deeper into the dermis but has not spread to the lymph nodes or distant organs. The 5-year survival rate for this stage is 80.4%. Although the prognosis remains relatively favourable, it is somewhat diminished compared to stage I MM. The 5-year survival rate for stage III MM is 70.6%. At this stage, the MM has metastasised to one or more regional lymph nodes, but not to distant organs. The prognosis is highly dependent on the number of lymph nodes involved and whether the cancer has spread to nearby tissues. Lastly, stage IV MM has the poorest prognosis, with a one-year survival rate of 56.4%. This advanced stage is characterised by the

spread of the MM to distant lymph nodes, organs, or other parts of the body, such as the lungs, liver, brain, or bones (Cancer Research UK, 2021-d).

1.4 Contemporary management of skin cancer

1.4.1 Diagnosis

Skin cancer is a ubiquitous and complex medical condition that demands multidisciplinary team (MDT) management. In the UK, a diverse group of medical professionals, including general practitioners, dermatologists, plastic surgeons, otolaryngologists, oral and maxillofacial surgeons (OMFS) and general surgeons, are involved in the management of skin cancer. The National Institute for Health and Care Excellence (NICE) guidelines serve as a framework for the diagnosis and referral of skin cancer patients. Suspected cases of MM and SCC should be referred urgently for cancer evaluation within two weeks, while BCC cases require routine referral for diagnosis and treatment (National Institute for Health and Care Excellence, 2015).

A comprehensive assessment of patients is essential to determine the risk factors for skin cancer development and inform treatment decisions. The assessment should begin with a thorough history, focusing on the patient's exposure to risk factors and factors that may influence treatment choice. The examination should concentrate on the presenting lesion, including features such as fixation to underlying structures, encroachment on cosmetically sensitive or functionally important areas (such as the nose or lip), available reconstructive options, and the assessment of regional lymph node basins. In addition, a complete skin examination should be performed to identify other lesions, and an abdominal examination for organomegaly is

recommended for potential MM patients (National Institute for Health and Care Excellence, 2015).

The clinical diagnosis of cutaneous MM with the naked eye has limited accuracy, with studies estimating its accuracy to be only around 60.0% (Kittler et al., 2002). However, the use of dermoscopy, a non-invasive imaging technique that allows for the microscopic examination of pigmented skin lesions, has emerged as a valuable tool for improving diagnostic accuracy. A meta-analysis of 27 studies found that the diagnostic accuracy for melanoma was significantly higher with dermoscopy than without it (Kittler et al., 2002). The use of dermoscopy increased the sensitivity of MM diagnosis by 49.0%, compared to naked eye examination alone (log odds ratio [OR] 4.0 [95% [confidence interval] CI 3.0 to 5.1] versus 2.7 [1.9 to 3.4]; p = 0.001). NICE guidelines recommend "assess all pigmented skin lesions that are either referred for assessment or identified during follow-up in secondary or tertiary care, using dermoscopy carried out by healthcare professionals trained in this technique" (National Institute for Health and Care Excellence, 2015). Proper training and expertise in the use of dermoscopy are crucial for accurate diagnosis and assessment of skin lesions. As such, healthcare professionals involved in the management of skin cancer should receive adequate training on the use of dermoscopy to maximise its diagnostic accuracy and utility.

Biopsy of the lesion and subsequent pathological analysis remains the gold standard for diagnosis. Biopsy can take the form of partial sampling, such as a punch biopsy, or complete removal, such as an excision biopsy.

1.4.2 Treatment

The treatment of skin cancer varies depending on the type of cancer and its location. Non-surgical options, such as topical therapies, can be used for low-risk superficial BCCs, Bowen's disease, and other in situ diseases.

Imiquimod, a synthetic imidazoquinoline, activates Toll-like receptor 7 (TLR-7) on dendritic cells, which in turn activates the immune system through the production of cytokines, including interferon-α and tumour necrosis factor-α. This leads to a cell-mediated immune response to the lesion and subsequent destruction of the tumour (Hemmi et al., 2002). Similarly, 5-fluorouracil is a pyrimidine analogue that inhibits thymidylate synthase, leading to decreased DNA synthesis and cell death (Goette, 1981). Topical treatments such as imiquimod and 5-fluorouracil are a convenient option for patients, as they avoid the pain and scarring associated with surgical management (Tanghetti & Werschler, 2007). However, these treatments are not without limitations. Almost all patients treated with these topical agents develop a transient localised inflammatory reaction, which can be uncomfortable and unsightly (Micali et al., 2010). Other non-surgical interventions include cryotherapy, radiotherapy, photodynamic therapy and laser (C. A. Morton et al., 2014).

Radiotherapy has been identified as an effective non-surgical treatment option for BCC, albeit with a higher likelihood of reduced cosmetic outcomes and increased recurrence compared to surgical excision. In contrast, nonsurgical treatments such as imiquimod and methyl aminolevulinate photodynamic therapy (MAL-PDT), although less effective than surgical treatments, still offer acceptable recurrence rates and should be considered when surgery is not feasible. While imiquimod may lead to more recurrences than surgical excision, it may result in improved observer-rated cosmetic outcomes. On the other hand, moderate-certainty

evidence suggests that imiquimod is likely to result in fewer recurrences than MAL-PDT, but there may be little difference between these treatments in terms of observer-rated cosmetic outcomes (Thomson et al., 2020). Very little evidence exists from randomised controlled trials comparing the efficacy of different interventions for primary cutaneous non-metastatic SCCs (Lansbury et al., 2010).

The surgical management of skin cancer is still the favoured option and is the only option of choice for the primary treatment of MM. This is in part due to their accessibility (you can see them), contiguity (you can cut around them) and their phenotypical margins being true margins (you can tell when you have cut it out). Surgical intervention can be divided into two main categories: destructive and non-destructive techniques. Destructive techniques refer to procedures where the cancerous tissue is destroyed through various methods, such as curettage and cautery, and non-destructive to procedures where the lesion is excised.

Nasr et al. found that recurrence rates for curative treatments, namely standard surgical excision, Mohs micrographic surgery (MMS) and radiotherapy, were very low (point estimates for recurrence rates range from 1.1% to 3.6% for <5 years; 2.4% to 12.4% for ≥5 years), while the rates for non-curative treatments, namely cryosurgery, curettage and cautery, PDT and topical therapy, were higher (15.3% to 16.5% for <5 years; 18.6% to 23.0% for >5 years) (Nasr et al., 2021).

Pooled estimates of recurrence of SCCs were lowest after cryotherapy (0.8% (95% CI 0.1% to 2.0%)) and curettage and electrodesiccation (1.7% (0.5% to 3.4%)), but most treated SCCs were small, low risk lesions. After MMS, the pooled estimate of local recurrence during variable follow-up periods was 3.0% (2.2% to 3.9%), which was non-significantly lower than

the pooled average local recurrence of 5.4% (2.5% to 9.1%) after standard surgical excision (12 studies), and 6.4% (3.0% to 11.0%) after radiotherapy. After an apparently successful initial response of SCCs to PDT, pooled average recurrence of 26.4% (12.3% to 43.7%) was significantly higher than other treatments (Lansbury et al., 2013).

The major disadvantage of destructive techniques is the lack of assessment of the margins around the lesion, which can lead to inadequate primary clearance and increased risk of recurrence if no further treatment is undertaken. Non-destructive or excision techniques, on the other hand, involve complete removal of the lesion along with a margin of healthy tissue to ensure complete clearance of the cancerous cells. This technique provides a more accurate assessment of the margins, which results in a lower risk of recurrence. Bread loafing is a technique used in the pathological analysis of skin cancer specimens that have undergone surgical excision with a pre-determined margin. In this technique, the specimen is sliced in a manner similar to slicing a loaf of bread, creating multiple thin slices. Each slice is then examined under a microscope to assess the margins of the specimen for the presence of cancerous cells (Abide et al., 1984). En face sectioning is another technique that can be used to assess the margins of a skin cancer lesion. This technique involves taking a vertical slice of the lesion, parallel to the skin's surface, and examining the entire cross-section under a microscope. It evaluates 100% of margins, whereas bread-loaf sectioning allows a sample analysis of only 1 to 2% (Gastman, 2016). En face sectioning allows for a comprehensive evaluation of the margins. However, en face sectioning is more time-consuming and requires more expertise than bread loafing, and it may not always be necessary for routine pathological analysis.

In order to achieve complete clearance and minimise the risk of recurrence, surgical excision with appropriate macroscopic margins is recommended for treating BCCs. The British Association of Dermatologists (BAD) guidelines recommend a 4mm peripheral clinical surgical margin for excising low-risk BCCs (Nasr et al., 2021). For primary BCCs with a high-risk factor a larger margin of at least 5mm is recommended to ensure complete removal of the cancerous cells. In addition to peripheral margins, adequate excision at the deep margin to a clear plane is also necessary for complete clearance, including a fat layer if present and other deeper structures if needed. The BAD similarly provide recommendations for the excision margins of SCC. Excision should be performed with a clinical peripheral surgical margin of ≥ 4 mm for a low-risk SCC, ≥ 6 mm for a high-risk SCC, and ≥ 10 mm for a very high-risk SCC. For mobile lesions, the deep margin should be within the next clear surgical plane, and on the scalp, the excision should include the galea (Keohane et al., 2021). In cases of deeply infiltrating or fixed lesions at any site, achieving an uninvolved deep histological margin may require inclusion of one or more of the following - fascia, muscle, bone or other underlying structure - which may be determined clinically or by imaging, or both. SCC excision should ensure at least a 1mm histological clearance at all margins by including sufficient peripheral and deep tissues (Keohane et al., 2021). Consideration should be given to excision of a further, orientated, deep-margin specimen where possible, if there is clinical concern at the time of resection that the resection is close or incomplete.

MM excision margins are based on the Breslow thickness of the lesion (the depth of invasion from the granular layer of epidermis to the deepest part of the melanoma). Data from eight randomised controlled trials have been used to inform the NICE recommendation for margin size (National Institute for Health and Care Excellence, 2022-b). There is, however, still uncertainty over the optimum excision margins, balancing reduction in local recurrence with

increased tissue preservation and reduced surgical morbidity. NICE currently recommend a clinical margin of at least 0.5cm for stage 0 MM. For stage I MM, a clinical margin of 1cm is recommended. For stage II MM, a clinical margin of 2cm is recommended but a 1cm margin may be used in cases where a 2cm margin would cause unacceptable disfigurement or morbidity (National Institute for Health and Care Excellence, 2015).

Frederick Mohs (1910-2002) was an American general surgeon who developed the technique of MMS. MMS has a role to play in the surgical management of NMSC, especially in cosmetically sensitive areas or high-risk lesions. MMS is a process of continual, sequential intra-operative margin assessment until all margins are deemed histologically clear (Mohs, 1978). It is time consuming and expensive, but due to high cure rates guidelines currently recommend MMS as a first-line treatment option for high-risk BCC in adults with poorly defined clinical margins or at high-risk anatomical sites. MMS should also be considered for primary BCC with at least one high-risk factor, recurrent BCC with at least one other high-risk factor, and advanced BCC (Nasr et al., 2021). For selected individuals with SCC, MMS may be considered after discussion with an MDT, particularly when tumour margins are difficult to delineate or tissue conservation is important for function (Keohane et al., 2021).

A number of adjuvant treatment options exist, primarily in the management of MM. In 1992, Morton and colleagues published a landmark paper describing the use of vital blue dye to identify the sentinel lymph node in patients with MM (D. L. Morton et al., 1992). Sentinel lymph node biopsy (SLNB) is a minimally invasive procedure used to determine whether cancer has spread beyond a primary tumour site to nearby lymph nodes. This technique was refined for use in breast cancer in 1993 with the introduction of lymphoscintigraphy, a procedure that uses radioactive tracers to locate the sentinel node (Krag et al., 1993). SLNB

serves as a critical diagnostic tool, furnishing crucial insights into the existence of pathological involvement in the regional lymph node basin. This information not only enables accurate prognostication, but also informs the development of targeted strategies for regional disease management. It assists in the identification of patients who may derive benefit from adjuvant pharmacological treatment. NICE issued updated guidelines in 2022 for the staging of MM using SLNB. NICE recommends against offering imaging or SLNB to people with stage IA MM. SLNB is recommended for people with MM with a Breslow thickness greater than 1.0 mm and with certain features such as ulceration, lymphovascular invasion or a mitotic index of 2 or more (National Institute for Health and Care Excellence, 2022-a). The use of SLNB must be weighed against the complications associated with it, such as lymphoedema and anaphylaxis to blue dyes (Perenyei et al., 2021). It is also important to make it clear to patients that there is no convincing evidence that it offers a survival benefit. This was investigated by the Multicentre Selective Lymphadenectomy Trial 1 (MLST-1) trial which evaluated the impact of SLNB on melanoma-specific survival (MSS) in patients with intermediate thickness (Breslow thickness 1.2-3.5mm) MM. In patients undergoing wide local excision (WLE) for MM and SLNB followed by immediate completion lymphadenectomy (CLND) there was no significant difference in 10-year MSS between the SLNB group (81.4%) and the observation group (78.3%) (HR for death = 0.84; 95% CI 0.64-1.09; p =0.18) (Faries et al., 2017). However, the subsequent MSLT-II trial did not find a MSS benefit for patients with sentinel-node metastases who underwent CLND compared to nodal observation, although CLND did reduce disease free survival. In light of the MLST-1 and MLST-2 trials, NICE advises against the routine use of CLND for patients with stage III MM and micrometastatic nodal disease identified through SLNB. This procedure should only be considered in cases where recurrent nodal disease management is difficult due to specific factors, such as head and neck MM, contraindicated stage III adjuvant therapies, or an inability to have regular follow-up. The

decision to proceed should be made collaboratively with the patient and the skin cancer specialist multidisciplinary team (SSMDT) (National Institute for Health and Care Excellence, 2022-b).

After primary treatment of high-risk BCC, standard surgical re-excision is recommended for adults with involved histological margins. For those with close histological margins, referral for MDT discussion of management options, including surgical re-excision, MMS, radiotherapy, or monitoring, is recommended. Routine follow-up is not recommended for adequately treated isolated BCC, but a postoperative review and yearly follow-up should be considered in certain cases (Nasr et al., 2021). In the case of SCC, the risk status of the tumour should be documented, and histology should be reviewed for those with involved or clear-but-close margins. Treatment options may include observation, further wide local excision, MMS, or adjuvant radiotherapy, depending on the risk factors for the patient and tumour. Immunocompromised individuals with SCC should receive active treatment and surveillance. Those with symptomatic perineural invasion and/or radiological evidence of perineural invasion should be discussed at a SSMDT, with the preferred option of aggressive surgical excision of the involved nerve followed by consideration of adjuvant radiotherapy, if technically possible (Keohane et al., 2021).

Systemic therapy for MM has historically been ineffective. Following a brief stint of interferon being the only available treatment option (Kirkwood et al., 1996), the emergence of newer, more effective therapies has shifted the treatment landscape for MM and has rendered interferon obsolete. Chemotherapy for MM primarily utilises dacarbazine, an alkylating agent. Although it lacks survival benefits and offers response rates of only 13-20.0% with a short-lived duration, it remains the main drug for treatment (Eggermont & Kirkwood, 2004). NICE

recommend that dacarbazine can be considered as a treatment option for people with untreated or previously treated stage IV or unresectable stage III MM for whom immunotherapies and targeted therapies are contraindicated, unsuitable, or unacceptable. It can also be used in the context of a clinical trial for people who have had previous treatment with dacarbazine (National Institute for Health and Care Excellence, 2015).

NICE currently recommend that dabrafenib can be used as a targeted therapy for untreated BRAF-mutant melanoma. Encorafenib plus binimetinib or dabrafenib plus trametinib can be offered if immunotherapies are contraindicated or there is not enough time for an immune response. If both are unsuitable, dabrafenib or vemurafenib can be offered, or chemotherapy or best supportive care can be considered. For untreated BRAF-wild type MM, chemotherapy or best supportive care can be considered. For previously treated MM, immunotherapies, targeted therapies, or chemotherapy depending on contraindications and suitability, or best supportive care should be considered (National Institute for Health and Care Excellence, 2015).

Recent advancements in immunotherapy with checkpoint inhibitors have significantly improved treatment options for MM patients (Reynolds & Fennelly, 2022). Currently, patients with resected stage III and IV melanoma are eligible for adjuvant treatment (National Institute for Health and Care Excellence, 2015). Programmed cell death protein 1 (PD-1) and its counterpart, programmed death ligand 1 (PD-L1), play vital roles within the immune checkpoint network. Positioned on the surface of T cells, PD-1 operates to suppress the immune response upon activation. PD-L1, its activating ligand, is frequently upregulated on malignant cell surfaces, facilitating immune evasion for these cells (Reynolds & Fennelly, 2022). The use of these against the PD-1 protein on T cells can overcome the mechanism of tumour

proliferation. Adjuvant pembrolizumab and nivolumab, both anti-PD1 agents, have also shown significant improvements in overall survival in resected stage III MM patients (Eggermont et al., 2018; Weber et al., 2017). These therapies are all approved by NICE as an option for the adjuvant treatment of completely resected stage III MM with lymph node involvement in adults (National Institute for Health and Care Excellence, 2022-c; National Institute of Health and Care Excellence, 2021). Stage IIIa represents the most delicate balance of risks and benefits for adjuvant therapy, necessitating thorough patient discussions. Patients need to be willing to undergo 1 year of drug treatment and accept a 1.0% risk of serious side effects including permanent endocrine dysfunction with nivolumab and pembrolizumab (Eggermont et al., 2018; Weber et al., 2017). Additionally, adjuvant ipilimumab has been shown to prolong survival in stage III MM patients. NICE recommends that nivolumab plus ipilimumab can be offered to patients with untreated stage IV or unresectable stage III MM (National Institute for Health and Care Excellence, 2016).

Signal transduction inhibitors like vemurafenib target the mutated BRAF gene present in about half of MMs (Chapman et al., 2011). In line with the NICE recommendations, BRAF analysis of MM tissue samples should be considered for people with stage IIA or IIB primary MM and carried out for those with stage IIC to IV primary MM (National Institute for Health and Care Excellence, 2015). This analysis helps identify the presence of the BRAF V600 mutation, allowing for the targeted use of signal transduction inhibitors like vemurafenib. Combination therapy with adjuvant dabrafenib (BRAF inhibitor) and trametinib (MEK inhibitor) has also been shown to improve overall survival for BRAF mutants (Long et al., 2017). NICE recommends this as an option for the adjuvant treatment of resected stage III BRAF V600 mutation-positive MM in adults (National Institute for Health and Care Excellence, 2018).

Chemoradiotherapy, which involves the use of chemotherapy and radiation therapy in combination, is another treatment option that can be effective for certain types of SCCs. For patients with metastatic SCC who are not candidates for immune checkpoint inhibitors, systemic chemotherapy or epidermal growth factor receptor (EGFR) inhibitors may be considered. However, it is important to note that EGFR inhibitors are not currently licensed for SCC in the UK. In palliative settings, electrochemotherapy may be considered for patients with locally advanced SCC when other local or systemic therapies are not appropriate. This approach can be especially useful for patients who are unable to tolerate more aggressive treatments (Keohane et al., 2021). In rare cases where BCCs have metastasised, vismodegib, a systemic hedgehog signalling pathway inhibitor, may be used. However, this treatment option is subject to availability and should only be offered after discussion at an MDT meeting (Nasr et al., 2021).

The reconstructive ladder is a hierarchical model that outlines different options for reconstructing defects or deformities in the body, ranging from simple techniques like primary closure or skin grafting to more complex procedures like tissue expansion and microvascular surgery (Janis et al., 2011). In the context of skin cancer excision, the reconstructive ladder provides a valuable guide for selecting the most appropriate technique for each patient's individual needs. While primary closure may be suitable for small defects, larger and more complex defects often require more advanced techniques such as skin grafting, local tissue rearrangement, regional flaps, or free tissue transfer. Each of these techniques has its advantages and disadvantages, and the decision about which technique to use should be based on factors such as the size and location of the defect, the patient's medical history, and their individual goals and expectations for the outcome of the procedure. In addition to the reconstructive ladder, other factors to consider in skin cancer excision reconstruction include

the type and stage of the cancer, the patient's skin type and colour, and their age and overall health.

1.5 History of cancer MDTs

Cancer care has undergone a paradigm shift over the past few decades, moving from a fragmented, discipline-specific approach to a more comprehensive and integrated model. This transition has been facilitated by the establishment of cancer MDTs and tumour boards. MDTs and tumour boards bring together healthcare professionals from different specialties, enabling coordinated decision-making and improving patient outcomes.

The UK has played a pivotal role in the establishment and development of cancer MDTs. In the early 1990s, the Calman-Hine report (1995), commissioned by the UK Department of Health, identified the need for integrated, multidisciplinary cancer care. The report highlighted the fragmented nature of cancer services and recommended the formation of cancer MDTs to enhance coordination, communication, and decision-making. As a result, the NHS Executive Group (1996) mandated that all new cases of cancer should be managed by an MDT. MDTs have since been a core component of patient management in the UK.

Following the Calman-Hine report, the UK government introduced the NHS Cancer Plan in 2000, which set out a strategy for improving cancer care through the development of MDTs (Department of Health, 2000). This plan required all cancer patients to have their care discussed and planned by an MDT. This ground-breaking initiative led to the widespread establishment of cancer MDTs across the UK, contributing to improved patient outcomes and standardising care.

Parallel to the development of MDTs in the UK, tumour boards emerged as an essential element of cancer care in the United States of America (USA). These boards served a similar function to MDTs, providing a forum for healthcare professionals from various disciplines to review and discuss individual cancer cases, ensuring a comprehensive and coordinated approach to patient care. The American College of Surgeons' Commission on Cancer now mandates that all accredited cancer programs hold regular multidisciplinary cancer conferences, where cases are prospectively reviewed and management decisions are collaboratively deliberated upon (American College of Surgeons, 2012).

Inspired by the UK's progress in MDTs and the USA's success with tumour boards, the concept of collaborative, MDT cancer care spread to other countries, resulting in the establishment of similar models of care.

1.5.1 Governance

In the UK, NICE provides guidance on the operation of MDTs and the conduct of MDT meetings. The NICE guidelines emphasise the importance of patient-centeredness, the need for clear communication, and the need for standardization of MDT processes and procedures.

The NICE improving outcomes guidance describes six levels of care for skin cancer which differ in their member composition and case-mix (National Institute for Health and Care Excellence, 2006). These have subsequently been incorporated into the Manual for Cancer Services: Skin Measures v 2.0 National Cancer Peer Review (NCPR) (National Cancer Peer Review-National Cancer Action Team, 2011). MDT structure and characteristics exist to standardise care. Level 1 care can be provided by any general practitioner (GP) in the community to manage benign lesions, actinic keratoses or SCC in situ. Level 2 care is provided

by a listed community skin cancer clinician associated with a named MDT to manage low risk BCC. Level 3 and 4 care is provided by local skin cancer multidisciplinary teams (LSMDTs) whilst SSMDTs provide level 5 care (Table 1) whilst level 6 care is provided by supra-network MDTs who manage cutaneous T-cell lymphoma for the whole of their host network and other named networks to offer total surface electron beam therapy and photopheresis. Recommendations for the case-mix of each skin MDT from the BAD National Reform of Cancer MDT Meetings report (British Association of Dermatologists, 2018) are shown in Table 1.

Table 1: Core membership and case-mix of each type of skin MDT.

Care level	Person or team	Core membership	Case mix/procedure
3	LSMDT, hospital staff core team member (may be core member of SSMDT acting as 'local' LSMDT). Without mandatory individual case review by MDT	• Two dermatologists • Histopathologist • Skin nurse specialist • MDT co-ordinator/secretary • An NHS-employed member of the core or extended team should be nominated as having specific responsibility for users' issues and information for patients and carers; • A member of the core team nominated as the person responsible for ensuring that recruitment into clinical trials and other well-designed studies is integrated into the function of the MDT	• Low risk BCC — incompletely or narrowly (<1mm) excised, perineural invasion • Low risk BCCs excised by non-accredited GPs in the community • High risk BCC — incompletely or narrowly (<1mm) excised, perineural invasion* • High risk BCCs excised by GPs in the community* • SCCs excised by GPs in the community
4	LSMDT, hospital staff core team member(s), with mandatory individual case review by LSMDT (may be the SSMDT and its core members acting as 'local' LSMDT)		• BCCs - recurrent after previous excision and BCCs persistent (i.e. having histologically positive resection margins) after excision • BCCs - which are sited such that excision poses a potential risk to important underlying structures, areas where difficult excision may lead to a poor cosmetic result and areas where primary closure may be difficult (lips, nose, nasofacial sulci, nasofacial folds, periorbital areas and ears). • SCC — incompletely or narrowly excised (<1mm), perineural or lymphovascular invasion, thickness 6mm or more, pT2 or above, poorly differentiated tumours, specific histological subtypes (clear cell, desmoplastic, verrucous, carcinosarcoma, adenosquamous)

Care level	Person or team	Core membership	Case mix/procedure
			• SCCs from special or high-risk sites (ear, lip, eyelid/canthus) • MM — new, single primary, adult, non-metastatic, not for approved trial entry, up to and including stage IIa • MM excised or biopsied in primary care • Radiotherapy if attendance by clinical oncologist at LSMDT • Lesion where diagnosis is uncertain but may be malignant • Incompatible clinical and histological findings
5	SSMDT hospital staff core team member(s) with mandatory individual case review by SSMDT. May have been previously reviewed by LSMDT or rapidly referred without prior review). For some cases — only one agreed SSMDT, if more than one in the network	• Two dermatologists • Two surgeons, at least one of whom should be a consultant surgeon trained in plastic and reconstructive surgery • Skin nurse specialist • Two histopathologists • Imaging specialist • Clinical oncologist • Medical oncologist • MDT co-ordinator/secretary • An NHS-employed member of the core or extended team should be nominated as having specific responsibility for users' issues and information for patients and carers • A member of the core team nominated as the person responsible for ensuring that recruitment into clinical trials and other well-	• Selected BCCs and SCCs needing plastic/reconstructive surgery by SSMDT core member (as per network clinical guidelines) • Radiotherapy (as per network clinical guidelines). If not discussed and treated by LSMDT clinical oncology core team member • Metastatic SCC on presentation or newly metastatic • MM — stage IIb or more, or <19 years or metastatic on presentation or newly metastatic or recurrent or for approved trial entry or positive excision margins • Patients for sentinel lymph node biopsy • Positive sentinel lymph node biopsies • Patients with positive lymph nodes following lymph node clearance

Care level	Person or team	Core membership	Case mix/procedure
		designed studies is integrated into the function of the MDT	• Any cases for adjuvant therapy (as per network clinical guidelines) • Histology opinion from SSMDT core pathology team member • Mohs surgery — designated SSMDT regionally • Skin cancer in immunocompromised patients including organ transplant recipients • Skin cancer in genetically predisposed patients including Gorlin's Syndrome including BCCs. • Tumours associated with burns, albinism, xeroderma, post-irradiation • Rare skin tumours — sebaceous carcinoma, malignant pilomatrixoma, neuroendocrine carcinoma — designated regional SSMDT • Cutaneous sarcoma superficial to the deep fascia

The quorum for frequency and attendance has been previously defined by cancer indicators derived from the Manual for Cancer Services. They define a quorum as an SSMDT 1) Occurring weekly and 2) Core membership attendance based on 1 x dermatologist, 1 x surgeon, 1 x clinical oncologist, 1 x medical oncologist, 1 x histopathologist, 1 x imaging specialist, 1 x skin nurse specialist and 1 x MDT coordinator.

1.5.2 Impact

MDT working has been shown to standardise practice with evidence-based decisions, reduce waiting times for treatment, improve the patient experience and confer economic benefits (Gabel et al., 1997; Mazzaferro & Majno, 2011; Murray et al., 2003; Stephens et al., 2006; Winters et al., 2021). There is also a body of evidence to support MDTs improving overall survival in patients with colorectal cancer, lung cancer, and gastrointestinal malignancies (Basta et al., 2017; Bilfinger et al., 2018; MacDermid et al., 2009). A meta-analysis of 11 studies with 30,814 patients found that MDTs improved overall survival in colorectal cancer patients (Peng et al., 2021). Another study found that lung cancer patients whose cases had been discussed at an MDT meeting had a mean survival duration of 280 days compared with 205 days for those whose cases were not discussed (Bilfinger et al., 2018). However, a meta-analysis of five studies on the impact of MDT on survival in lung cancer patients found that only two studies reported a modest 1-year survival increase in inoperable patients while the others did not disclose any advantage in terms of survival after the introduction of MDT meetings (Coory et al., 2008). Variations in treatment uptake also still exist and as such, MDTs have not been entirely successful in their aim of reducing variation in access to care (Munro, 2015). There are significant direct and indirect costs of MDT working and evidence demonstrates that some MDTs function more effectively than others (Fleissig et al., 2006; Taylor et al., 2012). Regular meetings to discuss patients also present an opportunity cost to

the NHS (Kane et al., 2007). No current evidence exists on the impact on skin cancer MDT on skin cancer surgery outcomes, which is reflected in the recent James Lind Alliance (JLA) Priority Setting Partnership in skin cancer (2022) highlighting this as an area for future research.

1.6 The digital surgeon: unleashing the power of NLP in modern surgery

The landscape of surgical practice is undergoing a dynamic shift, as the capabilities of natural language processing (NLP) intertwine with routinely collected electronic health record (EHR) data. Historically, there existed a notable disconnect: NLP experts in broader domains often lacked hands-on experience with EHR data, as well as the opportunity to collaborate with clinicians. Furthermore, accessing clinical datasets posed significant challenges. Today, the burgeoning realm of big data has unlocked doors for NLP, facilitating its interaction with EHRs. Rich data found locked away in ‘unstructured’ formats such as clinic letters, operation notes and histopathology reports had traditionally been relatively inaccessible for analysis, however with the advent of NLP this is no longer the case.

1.7 Unpacking the terminology

Big data is best described by the '5 V’s': the staggering *Volume* of data, its swift *Velocity* of generation and processing, the diverse *Variety* of data types, its consistent *Veracity* or trustworthiness, and the potential *Value* it brings to the healthcare sector (Gibson, Dobbs, et al., 2021a). NLP sits at the intersection of this data-rich environment and advanced computational capabilities, enhancing the accessibility and utility of diverse data types. Imagine a scenario where, during surgery, a clinician can vocally request a patient's comprehensive history and instantaneously receive a synthesised report from extensive EHRs. This is no longer blue sky thinking, but a reality being sculpted by progressive NLP research.

For instance, IBM Watson's utilisation in oncology, where it reviews and interprets patient data, medical literature, and clinical trial data to support treatment decisions, represents a real-world use case of NLP in surgery (Jie et al., 2021).

1.7.1 Natural language processing

NLP can be defined as a set of techniques used to convert written text into interpretable datasets (Harrison & Sidey-Gibbons, 2021). Further exploration into NLP reveals that research has utilised various methodologies (Mellia et al., 2021). There are those that centre around manually curated rule-based processing, whereas others leverage machine learning (ML) toolkits. Some studies, recognising the strengths of both paradigms, have opted for a hybrid approach, combining rule-based methods with other techniques like machine or deep learning. While the tools and techniques might differ, a unified theme emerges: the relentless pursuit of extracting meaningful and accurate insights from complex clinical data. To fully appreciate the breadth of NLP, it is useful to define key terminology commonly as outlined in Table 2 (Cunningham, 2002; Jurafsky, 2008).

Table 2: NLP terminology table.

Term	Description	Example
Annotation	The manual or automated process of labelling text with metadata, such as entities or POS tags.	Annotated histopathology reports where "BCC" is linked to "Basal Cell Carcinoma" and "perineural invasion" is labelled as a prognostic factor. See Appendix 3: Annotation guide.

Corpus	A large collection of text used for linguistic analysis and NLP tasks.	
Document	A single unit of text within a corpus, such as a histopathology report.	
Gazetteer	A predefined list of words or phrases used for entity recognition in text.	A gazetteer list of procedural terms like ["Mohs surgery", "wide local excision", "excision biopsy"]. See Appendix 3: Annotation guide pages 16-19.
Lexicon	A structured vocabulary or dictionary of words with meanings, usage, or metadata.	WHO Classification of Soft Tissue Tumours.
Ontology	A structured knowledge representation defining concepts and their relationships.	SNOMED CT, UMLS, ICD-10, DSM, MeSH.
Named entity recognition	An NLP task that identifies and classifies entities in text (e.g., names, locations, diseases).	In "The patient was diagnosed with malignant melanoma at Morriston Hospital," NER identifies: - Organization: Morriston Hospital - Condition: Malignant Melanoma.
Part of speech tagging	Assigning a grammatical category (e.g., noun, verb, adjective) to each word in a sentence.	For "The lesion was excised with clear margins," POS tagging labels "lesion (NOUN)", "excised (VERB)", "clear (ADJ)".

Phrase chunking	Grouping words into meaningful phrases based on syntax.	Training an NLP classifier to predict recurrence risk in melanoma patients based on histopathology reports.
Token	The smallest meaningful unit of text, typically a word, punctuation mark, or subword unit.	For "The patient was diagnosed with malignant melanoma," tokens would be: ["The", "patient", "was", "diagnosed", "with", "melanoma", "."]
Tokenisation	Splitting text into words or phrases.	'The patient was admitted.' → ['The', 'patient', 'was', 'admitted', '.']

1.7.2 Rule based systems

Before the meteoric rise of modern ML-driven NLP, rule-based systems were the vanguard of NLP. These systems function based on a meticulously crafted set of linguistic rules. Think of them as the guardians of linguistic grammar and structure, sifting through data using predefined syntactic and semantic rules. In surgery, rule-based NLP could standardise the extraction of specific data from EHRs or identify patterns based on explicit criteria. While perhaps less adaptive in the face of wildly varying data compared to their ML counterparts, their clarity, consistency, and transparency in processing language make them indispensable, especially when interpretability is paramount. Rule-based systems present a distinct advantage: they are easily tailored, straightforward to implement, and inherently interpretable. Such characteristics render them particularly apt for the biomedical domain.

1.7.3 Machine learning

Before exploring ML, it is essential to understand its two core concepts: supervised and unsupervised learning. Supervised learning is analogous to learning in a traditional classroom, where models, like students, learn from provided correct answers or labelled data, enabling them to make predictions on new, unseen data. Unsupervised learning, conversely, is more exploratory and akin to solving a puzzle without knowing the final picture, where the algorithm uncovers patterns and structures in the data without predefined answers or labels. ML, defined as a field within artificial intelligence that empowers computers to 'learn' from data using statistical techniques without explicit programming, can be likened to a diligent pupil. Imagine a scenario where a pupil, presented with a textbook of data, discerns patterns and memorises them. When faced with new challenges, the pupil draws upon this knowledge for answers. In the surgical realm, this 'diligent pupil' proves invaluable. ML-based NLP is being harnessed for predictive analytics such as identifying patients at risk of post-operative complications. Moreover, by scrutinising clinical notes, it provides surgeons with clinical decision support, aiding diagnosis and suggesting potential treatment pathways.

1.7.4 Deep learning

Delving deeper into ML, one encounters deep learning. This subfield focuses on algorithms inspired by the structure and function of our brain, specifically artificial neural networks. Picture a devoted scholar with a range of specialised notebooks, each representing a different layer of understanding. As these layers come together, they form an intricate neural network that mirrors the complexity of human cognition. In the medley of neural structures, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) hold a spotlight, each decoding a different facet of medical data: CNNs focusing on images, and RNNs unravelling the sequences and texts. Focusing on RNNs, they are critical for performing

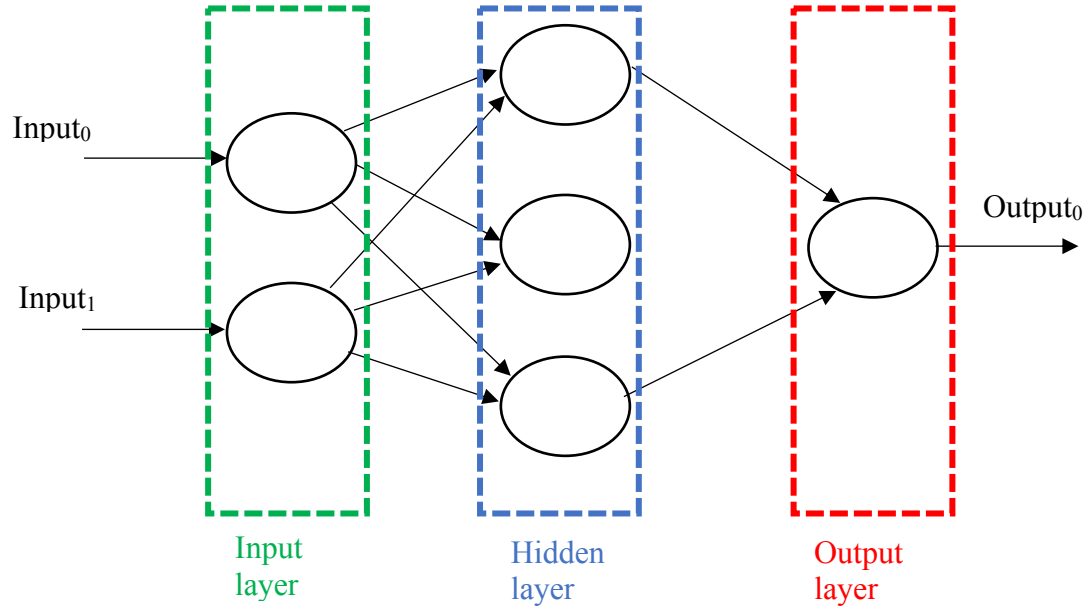
intricate analysis on sequential and text data and indispensable for extracting coherent patterns and meaningful insights from unstructured textual information. One proprietary example of leveraging deep learning in healthcare is Amazon Comprehend Medical (Amazon Web Services, 2024). Imagine, in an MDT meeting, a surgeon employs Amazon Comprehend Medical to rapidly synthesise and analyse a patient's EHR. This enables immediate access to detailed medical histories, including prior surgery and treatments, facilitating quicker, more informed, and precise decision-making. There are many deep learning techniques which are widely used for analysing patients' medical data. A few of them are discussed here in detail.

1.7.4.1 Neural networks

A neural network is a computational framework inspired by the structure and function of the human brain (Bishop, 1995a). It consists of multiple layers of artificial neurons, each performing mathematical operations to process and transform input data into meaningful predictions (Widrow and Hoff, 1960; Rosenblatt, 1962; Rumelhart *et al*., 1986). These layers play specific roles in extracting patterns and insights from data.

The input layer serves as the gateway for raw data. Each neuron in this layer represents a feature of the input data, ensuring relevant information is passed through the network. Figure 1 illustrates this architecture, where the input layer (green box) transmits information through weighted connections to hidden layers (blue box) and finally to the output layer (red box).

Figure 1: Structure of a neural network showing the input layer (green), hidden layers (blue), and output layer (red). Image adapted from DataCamp.

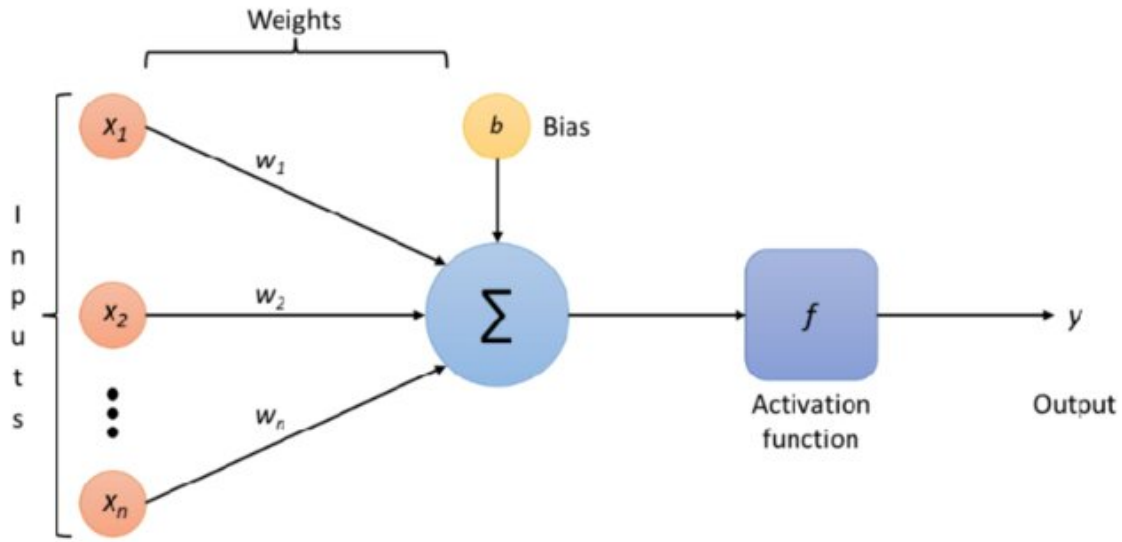


In neural network, learning occurs at hidden layers. These layers apply mathematical transformations using weights and biases, allowing the network to recognise patterns and relationships in the data. The deeper the network, the more complex features it can learn, which is why deep learning models with multiple hidden layers are referred to as deep neural networks (Witten et al., 2011).

Figure 2 provides a more detailed look at structure of a neuron within a neural network, illustrating weighted connections (w_1,w_2…w_n) and bias terms (b) that influence neuron activation. Each connection from one neuron to another has an associated weight, denoted as 'w'. Data flows through the network via feed-forward propagation, moving from the input layer to the output layer without looping back. Each neuron in a layer receives weighted inputs from the previous layer, computes a weighted sum, adds a bias, and applies an activation function to introduce non-linearity. This transformation ensures that complex patterns in the data can be captured. The output layer generates the final result, depending on the task at hand. For binary classification, the output layer consists of a single neuron that assigns a probability

score. In multiclass classification, the output layer contains multiple neurons corresponding to different categories.

Figure 2: Structure of a neuron in a neural network. Reproduced from Sánchez et al. 2022.



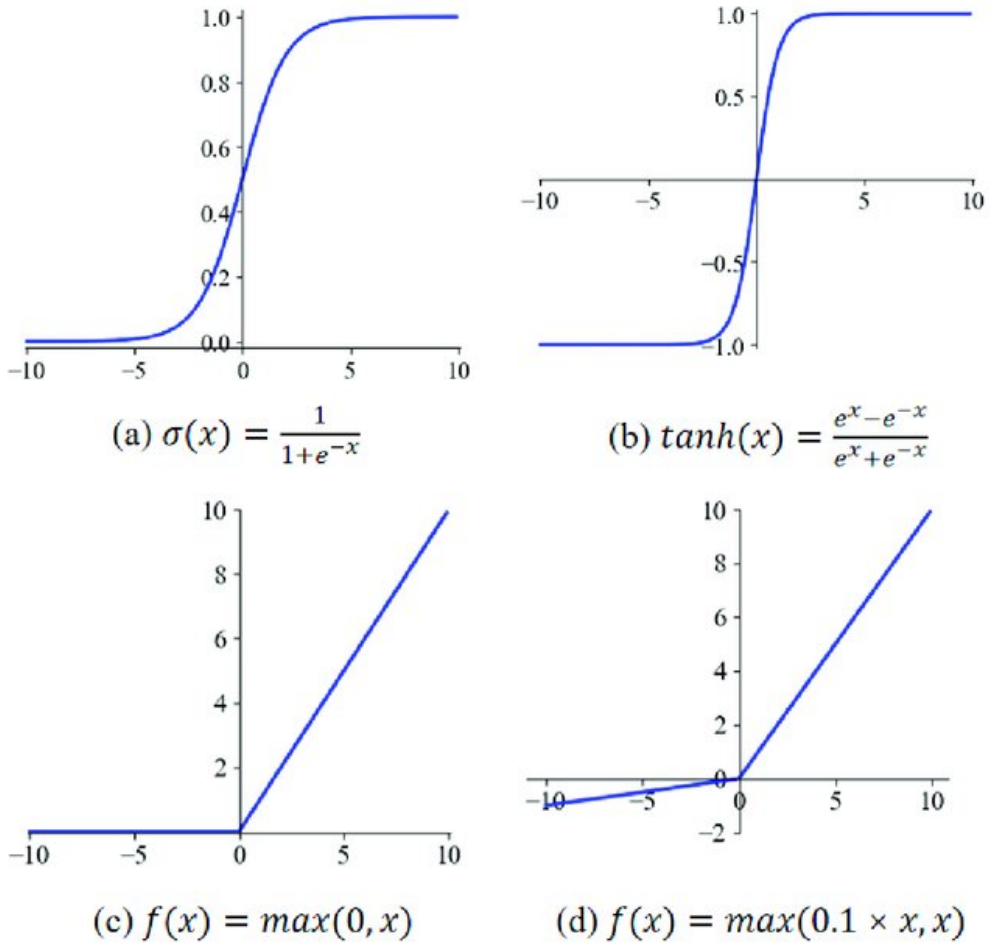
1.7.4.1 Activation functions

Activation functions are crucial for neural networks to make meaningful predictions (Bishop, 2006). Without them, the model would behave like a simple linear regression, unable to capture intricate relationships in the data. Different activation functions are used depending on the nature of the task.

The Rectified Linear Unit (ReLU) (Nair & Hinton, 2010) is widely used in deep neural networks due to its computational efficiency and ability to mitigate the 'vanishing gradient' problem (Bengio et al., 1994). The vanishing gradient problem occurs when gradients become so small during backpropagation that early layers in deep networks hardly update, stalling effective learning. ReLU outputs the input value if it is positive and zero otherwise, allowing neural networks to learn complex patterns efficiently. However, ReLU can sometimes suffer

from the ‘dying ReLU’ problem (He et al., 2015), where neurons become inactive if they always output zero due to predominantly negative inputs, halting their learning.

Figure 3: Activation function: (a) Sigmoid; (b) Tanh; (c) ReLU; (d) LeakyReLU. Reproduced from Yang et al. 2023.



The sigmoid function (Han & Morag, 1995) is commonly used in binary classification problems, mapping values to a range between 0 and 1. It helps interpret outputs as probabilities but can suffer from vanishing gradients when inputs are very large or very small. The softmax function (Bishop, 2006) is used in multiclass classification, converting raw scores into probability distributions across multiple categories. Softmax ensures that the sum of all outputs equals one, making it useful for categorical decision-making. The tanh (hyperbolic tangent)

function (Goodfellow et al., 2016) is similar in shape to the sigmoid function but varies between -1 and 1, allowing for stronger gradient signals. The leaky ReLU (Maas et al., 2013) is a modified version of ReLU that allows small negative values instead of zero, helping to mitigate the dying ReLU problem. All these activation functions are shown in Figure 3.

The selection of right activation function depends on the problem at hand, the architecture of the network, and empirical experimentation. ReLU is generally a good starting point due to its efficiency, alternatives like leaky ReLU, tanh, and softmax should be considered based on the specific needs of the task. By systematically testing different activation functions and analysing performance metrics, one can optimize neural network models for better accuracy and generalisation (Nair & Hinton, 2010 ; He et al., 2015). Changing the activation function in the hidden layer of a model influences its learning behaviour and classification performance. For instance, a sigmoid activation function in a binary classification model may result in misclassified data points due to a smooth decision boundary, whereas tanh can separate different categories more effectively (Goodfellow et al., 2016). ReLU provides sharper classification boundaries, while leaky ReLU prevents neurons from becoming completely inactive (Maas et al., 2013).

1.7.4.2 Learning in neural networks

Learning in neural networks involves tuning the weights or parameters to produce the desired output. One way to achieve this is through the gradient descent algorithm (Le Cun., 1998).

Figure 4: Gradient descent optimisation process illustrating how weight updates move towards the global cost minimum. Image courtesy of DataCamp.

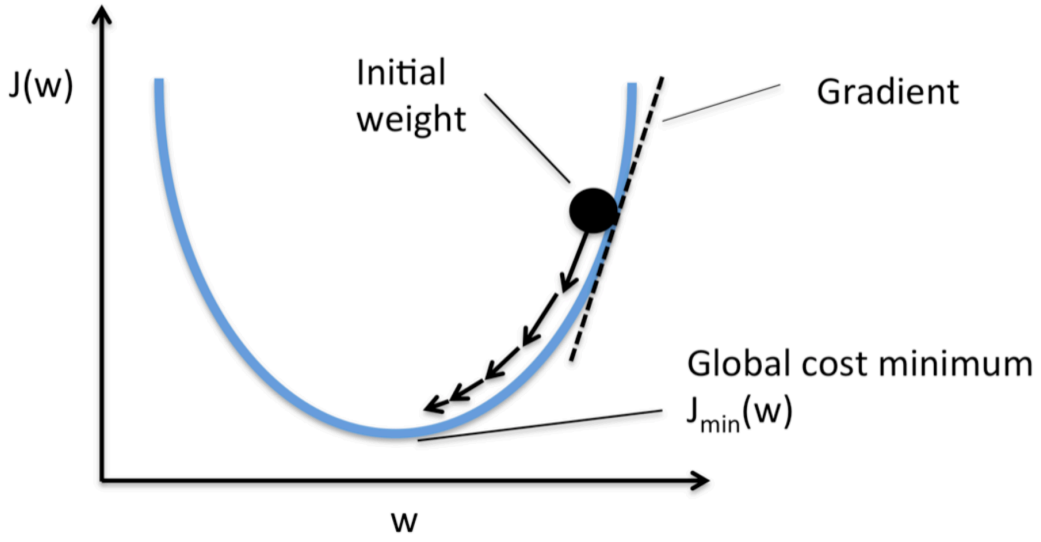


Figure 4 illustrates the gradient descent process, where the cost function J(w) is minimized by iteratively adjusting the weights. Gradient descent works by computing the derivative of the cost function with respect to each weight and updating the weights in the direction that reduces the error. This process continues until the global cost minimum $J_{min}(w)$ is reached, ensuring optimal network performance (Bengio et al., 1994). Moreover, another efficient way to compute the gradient of a neural network is called backpropagation of the gradient, that obtains a local minimizer easily (Goodfellow et al., 2016; Rumelhart et al., 1986). Backpropagation ensures weight updates are made in the correct direction, allowing the model to refine its predictions over time. It calculates how much each weight contributes to the overall error and adjusts them accordingly to reduce the cost function. This iterative process enables the network to gradually improve accuracy and refine predictions.

1.7.4.3 Word embeddings techniques and applications in neural networks

Word embedding or word vectors, are numerical representations of words that allow computers to perform complex tasks using text data. It is the process of mapping words into an n-dimensional vector space. These vectors are usually produced using deep learning models and

huge amounts of data (Li & Yang, 2017). Traditional methods, such as the ‘bag-of-words’ (Manning, Raghavan, & Schütze, 2008) approach, convert words into unique numbers but fail to capture contextual relationships. For example, "He is a doctor" versus "He is a physician. represented similarly because "doctor" and "physician" are nearly synonymous, conveying the same meaning despite being different words.

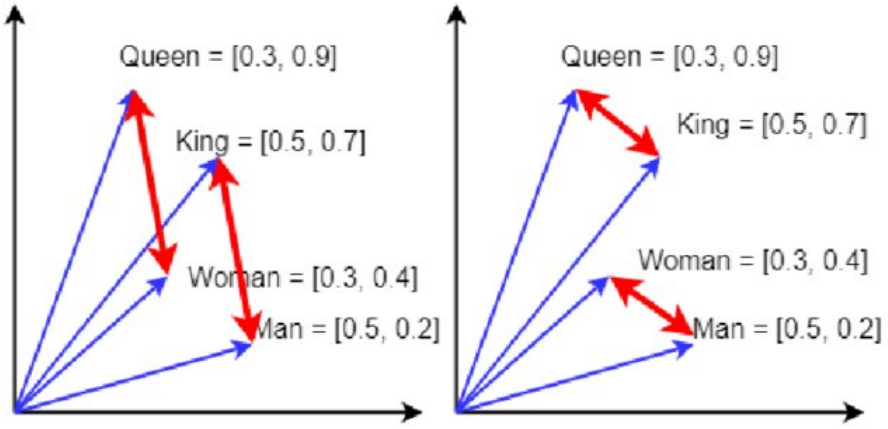
Recent methodologies such as word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017) generate word embeddings that encode semantic relationships. These approaches analyse word co-occurrences in large text corpora to understand similarities between words numerically. For instance, a 7-dimensional word vector can help differentiate between animals and objects based on their contextual usage. Word2vec, developed by Google in 2013, is one of the most widely used methods for generating word embeddings. It represents words in a dense vector space where similar words are positioned closer together. Figure 5 illustrates how words like "pork," "beef," and "chicken" cluster together, highlighting semantic similarity.

Figure 5: Large vector space. Image courtesy of DataCamp.

To understand the concept of words as numbers, we can consider how neural networks encode words as embeddings. Each word is assigned a vector of numbers, where the weights of the

network determine the placement of words in the vector space (Mikolov et al., 2013). For example, in word2vec, words frequently appearing together have similar weight distributions. Figure 6 demonstrates how weights influence word embeddings, showing that "king - man + woman = queen" due to the numerical relationships between words. This mathematical association suggests that king – man is approximately equivalent to queen – woman, as indicated by the red arrows. The diagram highlights four potential analogies, all following the same directional pattern. While this visualisation is idealised, the vectors do not need to be identical – only the closest match among all available word vectors. The alignment of the red arrows signifies a consistent relational meaning between the terms.

Figure 6: Example of neural word embeddings in 2D. Sutor P, Aloimonos Y, Fermuller C, Summers-Stay D. Image courtesy of Metaconcepts: isolating context in word embeddings. In 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) 2019 Mar 28 (pp. 544-549). IEEE.



11.7.5.2 Structural components of the transformer architecture

Transformers process input sequences and generate meaningful output, whether in language translation, text summarisation, or other generative tasks (Vaswani et al., 2017). It consists of two fundamental blocks: the encoder and the decoder.

1. *The encoder block*

The encoder processes the input sequence to capture contextual relationships between words. It consists of multiple identical layers, each using multi-head self-attention to understand token relationships and feed-forward layers to transform this knowledge into abstract representations (Lin, Wang, Liu, & Qiu, 2022). The encoder layers generate context-rich numerical representations of the input, making it easier for the model to interpret meaning. Instead of reading words sequentially, the encoder looks at all words simultaneously, allowing it to learn complex dependencies efficiently. Encoder-only transformers simplify the architecture to focus solely on understanding and input data representation. These models are often used in tasks like text classification.

2. *The decoder block*

The decoder takes the encoded input representation and generates an output. It uses self-attention to focus on the words it has generated so far while simultaneously incorporating information from the encoder through encoder-decoder attention (Vaswani et al., 2017). This allows the model to generate coherent and contextually relevant text.

Decoder-only Transformers specialise in tasks like text generation, where each token is predicted based only on the preceding tokens. They differ from full encoder-decoder models by incorporating masked multi-head self-attention, which ensures that each token only attends to previously generated tokens, preventing the model from "cheating" by looking ahead.

3. *Structural components of the transformer architecture*

Beyond the core encoder and decoder blocks, transformers architecture is built on several critical components that drive their performance:

a) Embedding layers and positional encoding

The Transformer architecture begins by embedding input tokens as vectors, which are numerical representations of words or sub-words. Each token has a unique ID within the model's vocabulary, which is mapped to an embedding vector. This vector's length, or dimensionality, determines how much information it can store about the token.

Since Transformers do not process words in order like RNNs, they need a way to retain word positioning. Positional encoding provides this information by adding numerical values to word embeddings that indicate their place in the sequence. This ensures the model understands word order while maintaining the benefits of parallel computation. Positional embeddings are generated using sin and cosine functions, which assign unique patterns to each position. These mathematical functions allow positional encodings to generalise across sequences of varying lengths (Li, Miao, Ma, Shuang, & Huang, 2023).

b) Self-attention mechanism and multi-head Attention

The self-attention mechanism allows transformers to determine how important each word is relative to every other word in a sentence (Vaswani et al., 2017). Rather than processing words one at a time, the model considers all words simultaneously, assigning attention scores to highlight important relationships. This mechanism is essential for capturing long-range dependencies and improving contextual understanding.

Multi-head attention extends this by enabling the model to focus on different linguistic features—such as grammar, meaning, and syntax—in parallel. This results in richer token representations and improves overall model accuracy.

c) Feed-forward networks and output layers

Once self-attention assigns importance to different words, the model applies feed-forward networks to refine the information. These networks independently transform each token's embedding, helping the model generate more complex representations that improve its understanding of language. Feed-forward layers are applied to each token position-wise, ensuring that each word's representation is adjusted independently based on its surrounding context (Devlin et al., 2018).

11.7.5.3 The transformer encoder-decoder structure and cross-attention

A full encoder-decoder transformer is used for sequence-to-sequence tasks like language translation. The encoder processes the input, and its outputs are passed into the decoder, where cross-attention mechanisms enable the decoder to focus on relevant parts of the encoded input while generating an output sequence (Lin, Wang, Liu, & Qiu, 2022).

Cross-attention is crucial for linking the encoder and decoder components, ensuring that information from the input is properly integrated into the output generation process (Gheini et al., 2021).

Comparison of neural networks and transformers

To better understand how transformers differ from traditional neural networks, the following table highlights key differences:

Table 3: Comparison of RNNs and transformers

Aspect	**Recurrent Neural Networks**	**Transformers**
Processing Style	Sequential; each step depends on the previous one.	Parallel; processes entire sequences simultaneously.
Handling Long-Range Dependencies	Struggles with long-range dependencies due to information loss over time.	Uses self-attention to capture dependencies across long distances.
Training Efficiency	Slow training due to sequential nature; difficult to parallelise.	Highly parallelisable, efficient training.
Positional Information	Maintained through recurrence and hidden states.	Uses positional encoding to retain word order information.

1.7.4 Clinical decision support systems

Drug discovery is a systematic, iterative process that begins by identifying and validating molecular targets implicated in disease pathology (Drews, 2000). Next, lead compounds are commonly discovered -often via high-throughput screening or computational approaches−to isolate molecules demonstrating specific biological effects (Trzeciak, 2001). These leads undergo stringent *in vitro* and *in vivo* preclinical testing before advancing into clinical trials (phases I–III). After a new therapy is authorized for market, phase IV surveillance continues in real-world settings to monitor both safety signals and long-term effectiveness. In a parallel manner, clinical decision support systems (CDSS) evolve along a similar trajectory (Sim et al., 2001; Sutton et al., 2020). Initially introduced as simple rule-based aids, modern CDSS now integrate both published research and local practice-based data (Demner-Fushman et al.,

2009; Eguia et al., 2024). The process begins with data acquisition and preprocessing−often relying on NLP to extract clinically relevant information from EHRs. Once validated on retrospective datasets (akin to preclinical testing), CDSS undergo real-world evaluations, paralleling phases I–III in drug development, to confirm accuracy and clinical utility (Demner-Fushman et al., 2009; Eguia et al., 2024). Ultimately, they enter a refinement stage analogous to phase IV, where new evidence or user feedback prompts updates to system logic. Extensive studies show that well-designed CDSS can improve clinical performance and patient outcomes (Hunt et al., 1998; Kawamoto et al., 2005). In particular, Kawamoto et al. (2005) highlight key design elements−such as presenting actionable recommendations automatically within the clinician's workflow and delivering them in real time at the point of care−that consistently boost effectiveness. Building on these concepts, Bates et al. (2003) propose "ten commandments" for effective clinical decision support, emphasising the importance of system speed, unobtrusive yet relevant prompts, and minimal extra data entry for clinicians. When these elements converge, CDSS can raise guideline adherence and reduce prescribing errors (Kawamoto et al., 2005; Bates et al., 2003). Nonetheless, practical hurdles remain, including "alert fatigue," gaps in interoperability, and the effort of continuously updating system guidelines (Sutton et al., 2020; Eguia et al., 2024).

1.8 Research question

This research seeks to address the central question: "Can NLP be employed to analyse EHR documents effectively and automate the primary and secondary functions of the skin cancer MDT, performing at a level that is comparable to human clinicians?"

1.9 Aim

The aim of this research is to investigate the application of NLP techniques in analysing extensive healthcare datasets. The insights derived from this exploration will inform the development of a CDSS platform for a vSMDT.

1.10 Objectives

To accomplish the stated aim, the following specific objectives were established:

- Conduct scoping surveys to delineate the current state, challenges, and potential areas for improvement in skin MDTs and virtual MDT meetings.
- Conduct a systematic review and meta-analysis of existing literature focusing on NLP CDSSs in cancer care, benchmarking them against the performance of human clinicians.
- Develop and validate an NLP-based CDSS for BCC capable of providing treatment recommendations after primary surgical treatment hosted on a vSMDT platform.
- Develop and validate an NLP-based information extraction system for BCC capable of improving quality assurance for benchmarking outcomes after primary surgical treatment.

Chapter Two:

Composition, Quoracy and Cost of Specialist Skin Cancer Multidisciplinary Team Meetings in the United Kingdom

2.1 Introduction

Since the Calman-Hine report in 1995 highlighted inadequacies in cancer care, and the National Health Service (NHS) Executive Group mandated that all cases of cancer should be managed by an MDT; they have been a core component of patient management in the United Kingdom (UK) (Calman & Hine, 1995; Department of Health, 2000; NHS Executive Group, 1996). MDT working has been shown to reduce variations in care, by standardising practice with evidence-based decisions, reduce waiting times for treatment, and improve the patient experience (Gabel et al., 1997; Mazzaferro & Majno, 2011; Murray et al., 2003; Stephens et al., 2006). Despite these advantages, the current MDT process has been shown to be time consuming, expensive, and inefficient (Chinai et al., 2013; De Ieso et al., 2013). In England, over 55,000 cancer MDT meetings take place each year (Munro, 2015). The contribution of NHS consultants' time to these meetings has been estimated at 1.2 million hours or 550 full-time equivalents, with a total cost of £154.3 million (National Cancer Peer Review Programme, 2011) . There is insufficient contemporary evidence to determine whether MDT working is cost-effective within the context of secondary care due to the small number of studies and their high risk of bias, in addition to poor reporting on the definition of MDT operationalisation and transparent costs (administering, preparing and attending) (Ke et al., 2013). A common example of MDT inefficiency is multiple case re-discussions, with over 10.0% of new patients requiring more than one discussion (National Cancer Action Team, 2011). Reasons for re-discussion include lack of contemporaneous clinical or specialist advice, imaging or investigations at the primary meeting, referral from one site-specific MDT to another or abnormal test results during the course of their management or regular follow-up. Composition and quoracy are key features of cancer MDT operationalisation that are yet to be adequately addressed by the literature (Birchall et al., 2004; Houssami & Sainsbury, 2006; Junor et al., 1994; Kelly et al., 2013; Licitra et al., 2016).

Skin cancer is the most common cancer in the UK, comprising at least 25.0% of all new cancer diagnoses (Cancer Research UK, 2024-d, 2024-c). There are over 16,000 new melanoma and 147,000 new non-melanoma cases diagnosed every year, with over 200,000 annual excisions performed at significant cost to the NHS (Cancer Research UK, 2024-c, 2024-d; Kerr, 2019). Incidence rates for melanoma skin cancer are projected to rise by 7.0% in the UK between 2014 and 2035, resulting in 32 cases per 100,000 people by 2035 (Cancer Research UK, 2024-c). The large caseloads of Local Skin Cancer Multidisciplinary Teams (LSMDTs) and Specialist Skin Cancer Multidisciplinary Teams (SSMDTs) contribute enormously to the national MDT workload. Despite this, there is a paucity of literature within the field of skin oncology examining costs, quality, and efficiency of skin cancer MDT work. If the NHS is to cope with the increased number of patients and associated cost implications, then it is important that these data are available to help the NHS consider how services can be optimised and re-configured to meet demand. Central to understanding these operational aspects of MDT working include ensuring that the different specialties represented are in attendance and able to participate in discussion. The primary aim of this study was to assess national SSMDT composition and attendance against a quorum with the secondary aim to analyse the total costs of administering, preparing and attending a SSMDT.

2.2 Patients and methods

Freedom of information (FOI) requests were sent to all National Health Service trusts across the United Kingdom to identify those with a SSMDT in July 2019. Any site holding a SSMDT was then sent a more detailed electronic FOI request asking for details regarding frequency, attendance, preparation time, running time, dissemination time, number of new cases, number of re-discussions and whether videoconferencing was used between sites (Table 4). The

quorum for frequency and attendance has been previously defined by cancer indicators derived from the Manual for Cancer Services: Skin Measures v 2.0 National Cancer Peer Review (NCPR), National Cancer Action Team (2011) which is based on the National Institute for Health and Clinical Excellence (NICE) standard Improving outcomes for people with skin tumours including melanoma and Cancer Services Guidelines (2006). In the context of this study, quorum was therefore defined as a SSMDT 1) Occurring weekly and 2) Core membership attendance based on ≥1 x dermatologist, ≥1 x surgeon, ≥1 x clinical oncologist, ≥1 x medical oncologist, ≥1 x histopathologist, ≥1 x imaging specialist, 1 x skin nurse specialist and 1 x MDT coordinator.

Table 4: FOI request sent to all SSMDTs.

Venue
All staff present based on attendance records
Time spent by each team member in preparing for the SSMDT meeting
Time spent by team coordinator preparing for and disseminating SSMDT outcome
Running time in minutes of the SSMDT
Overheads as a percentage of total operating costs from the most recent financial year
Number of new cases discussed
Number of re-discussions discussed (re-discussion defined as any patient discussed at the same point in their pathway but following an additional test or any patient brought back to MDT for re-discussion of the same test results)
Were any videoconferencing facilities used across sites?

Our costing methodology was based on core membership attendance records. Hourly rates were calculated from the Standard NHS Agenda for Change salary points, with preparation, running, and dissemination time factored in (NHS Employer Services, 2018). Overheads including heating, lighting and information technology support were expressed as a percentage of total operating costs. Total SSMDT cost and cost per patient was calculated based on the calculation of overheads + (mean hourly salary cost core SSMDT members x total preparation, running and dissemination time SSMDT core members).

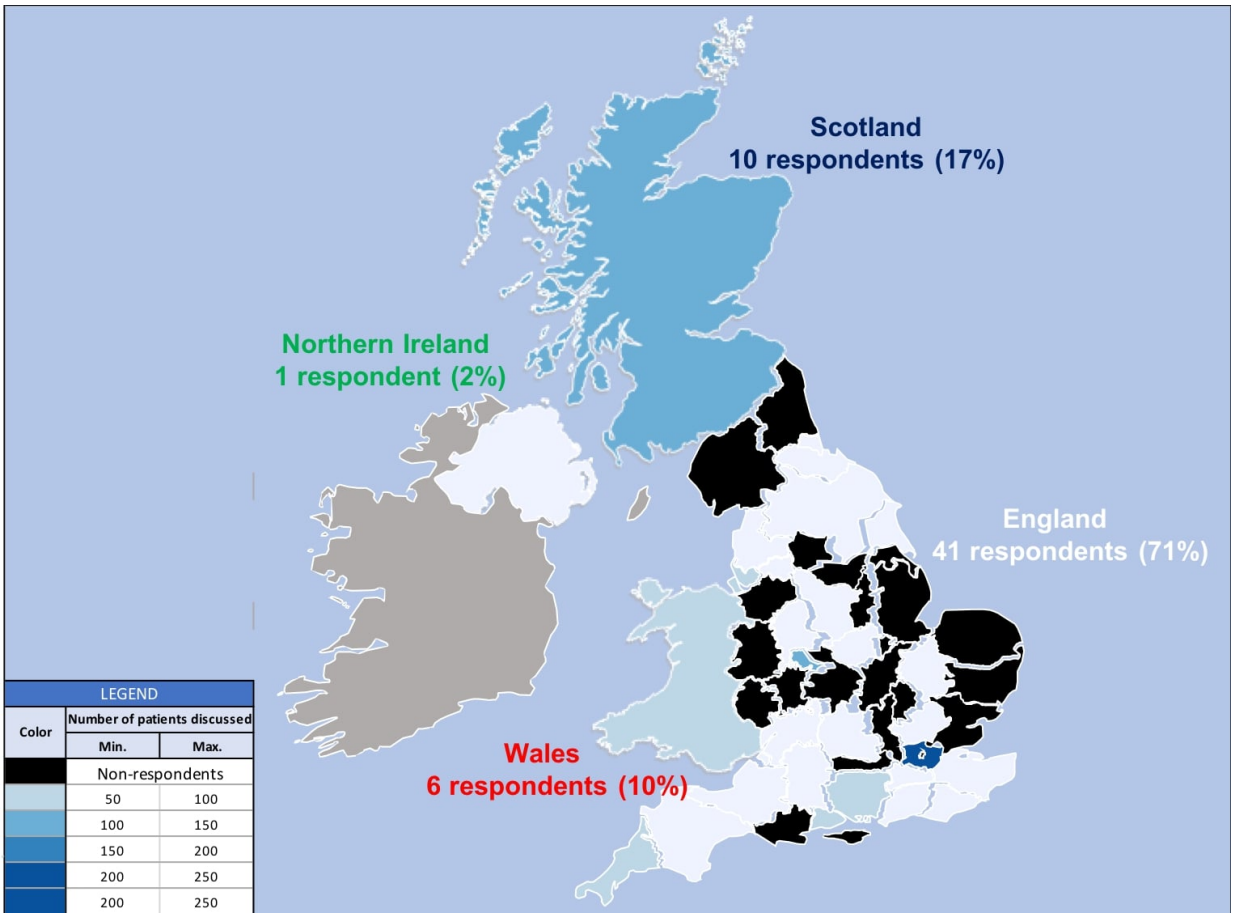
Descriptive and statistical data analysis were performed using Microsoft Excel 2010® (Microsoft Cooperation, Redmond, Washington USA) and IBM SPSS Statistics for Windows (version 24, IBM Corp, Armonk, NY) respectively. The chi-square (X^2) test was used to test the hypothesis of no difference in frequencies among England, Wales, Scotland and Northern Ireland. Furthermore, the F-statistic was used for testing the hypothesis of no differences among the means of total cost and cost per patient. $p < 0.05$ was considered statistically significant.

The study was approved by the clinical audit department of Morriston Hospital, Swansea (ID number: MH2).

2.3 Results

There were 58 respondents (89.0.0% response rate) to the FOI request. A heat map demonstrating national distribution of SSMDT respondents with regional variation in total patients discussed at SSMDTs is shown in Figure 7. The majority (69.0%) of SSMDTs were held in venues holding teaching hospital status.

Figure 7: Heat map that demonstrates distribution of SSMDT respondents and total number of patients discussed at SSMDTs regionally.



2.3.1 Composition and quoracy

Of the respondents 18 (31.0%) reported on parent speciality of the SSMDT chair: 1 oral and maxillofacial surgery (OMFS) (6.0%), 2 clinical oncology (11.0%), 3 plastic surgery (17.0%) and 12 dermatology (67.0%). With respect to surgical presence, the frequency of ≥1 consultant in attendance at the SSMDT from plastic surgery, OMFS, ear, nose and throat and general surgery was 43 (74.0%), 12 (21.0%) and 4 (7.0%) and 2 (3.0%) respectively. Only 15 SSMDTs (26.0%) were quorate by membership. Forty SSMDTs (69.0%) were quorate by meeting frequency. Ascending frequency of core membership role non-attendance is shown in Table 5. The most common reason for a lack of membership quoracy was a lack of clinical oncology presence. There was no geographical variation in frequency or attendance quoracy, and this

was not statistically significant (X^2 = 6.71, p= 0.08). Fifteen (26.0%) SSMDTs hosted videoconferencing facilities at the time of the data request. Post-hoc analysis showed weak positive correlation with attendance quoracy and SSMDTs who had videoconferencing facilities, although this was not statistically significant (Phi coefficient = +0.13, p = 0.07). There was a statistically significant weak positive correlation with the specific attendance quoracy for clinical oncology and teaching hospital status (Phi coefficient = +0.28807, p = 0.02825).

Table 5: Frequency of core membership role non-compliance contributing to non-quorate SSMDT's.

Core membership role	**Absolute non-compliance frequency count**	**Percentage expressed as proportion of all SSMDTs**
Consultant Clinical Oncologist (<1)	34	59.0%
Consultant Medical Oncologist (<1)	13	22.0%
Consultant Radiologist (<1)	11	19.0%
MDT Co-ordinator (<1)	6	10.0%
Consultant Dermatologist (<1)	3	5.0%
Total Consultant Surgeon (<1)	3	5.0%
Skin cancer CNS (<1)	2	3.0%
Consultant Histopathologist (<1)	1	2.0%

CNS; clinical nurse specialist.

2.3.2 Cost analysis

Twenty-four SSMDTs were able to provide data towards a costing analysis. The mean total cost for an SSMDT was £3,963.68 (£946.12-£9,353.94). The mean cost per patient was £132.68 (£31.67-£313.10). There was a weak positive correlation with cost per patient and total number of patients discussed (Pearson correlation coefficient -0.2767, p = 0.0766), suggesting that those SSMDTs who discuss more people at each meeting bring the unit cost per patient down, although this was not statistically significant. There was a large range in overheads, mean 12.7% (8.9%-17.6%). The geographical variations in costing and total SSMDT running times are shown in Table 6 and 7. The mean number of total cases discussed was 30 (12-50), with 10 re-discussed cases (32.7%) (0-41 [0-88.0%]). There was no geographical variation in either total or per patient cost (F= 1.82, p = 0.169 and F= 0.40, p = 0.757 respectively). The baseline cost of a quorate SSMDT was calculated at £2,380.41 based on these mean values. Assuming a 52 week per year service, this mean value was used to extrapolate the total annual cost of all quorate SSMDTs in the UK as £8,045,802.70.

Table 6: Breakdown of costing analysis by geography.

Costing Measure	Overall	England	Wales	Scotland	Northern Ireland
Mean overhead % (£)	12.65 (444.02)	12.64 (431.31)	12.75 (317.10)	12.08 (122.39)	12.75 (736.34)
Mean total SSMDT cost (£)	3,963.68	3852.90	2,804.19	1,135.98	6,511.58
Mean per patient cost (£)	132.68	145.84	155.79	53.14	166.96

Table 7: Total SSMDT time across the UK.

	Preparation time (hours)	Running time (hours)	Dissemination time (hours)
Mean	10.2	1.45	5
Minimum	0	0.5	0
Maximum	38.5	3	37.5

2.4 Discussion

MDT working is considered the gold standard for cancer patient management, with the aim of promoting best practice and reducing variation in access to treatment (Independent Cancer Taskforce, 2015). Data validation, audit and education are other important secondary functions. The healthcare landscape has drastically changed since their inception with increasing demands placed upon MDTs with complex treatment pathways, and an increasing range of imaging modalities and pharmacotherapeutic options. This has resulted in MDTs becoming resource intensive and costly to the NHS. This is reflected in the total cost of MDTs rising rapidly, from £88 million in 2011/12 to £159 million in 2014/15, driven by a rapid rise in activity (Cancer Research UK, 2024-b). This has not been matched by a growth of overall staff numbers attending MDTs (Workforce Team NHS England, 2024). Costs in this study were related to the number of people discussed at each meeting. The mean unit cost per discussion at the SSMDT in this study was £132.68 which is comparable to previous work reporting cost per discussion varying between £91.84 for breast cancer and £132.95 for colorectal cancer (Cancer Research UK, 2024-b). There was a relatively high rate of case re-discussion in this study which highlights some of the inefficiencies of cancer MDT working. Previous work has shown that over 10.0% of new patients need more than one discussion due to insufficient clinical information at the initial meeting and complex presentations needing multiple site-specific MDT input and to discuss abnormal investigations (Munro, 2015).

Previous work has demonstrated that skin cancer MDTs are costly, especially considering they last for an average of two hours, with opportunity costs ranging from £129,134 (± 25.0% range, £96,851 to £161,418) to £258,268 (± 25.0% range £193,701 to £322,835) per network per annum (National Institute for Health and Care Excellence, 2006). This cost is due to the

interdisciplinary nature and the co-location of senior specialists in several spheres, which also impacts on several consultant job plans.

In addition to cost, other shortcomings of current MDT practice have been highlighted in the Cancer Research UK (CRUK) commission investigating the effectiveness of MDT meetings in cancer services which was aimed at improving the effectiveness of MDT meetings in cancer services (Cancer Research UK, 2024-b). The crucial findings from this report included the lack of sufficient time to discuss complex patients, suboptimal meeting attendance, not utilising the right information to inform discussions, and MDTs not being able to fulfil their secondary roles. One of the most valuable features of MDT meetings is the diversity in terms of specialities represented, contributing to a meaningful discussion and a range of opinion and interpretation. With an estimated 50,000 shortfall in NHS clinical staff in England reported as of 2014, recruitment for specific clinical positions becomes harder (Cancer Research UK, 2024-b). Specifically, this was expanded on by CRUK in relation to imaging, endoscopy, and pathology capacity (H. Brown et al., 2015; Cake et al., 2016; Janssen et al., 2018). These challenges increase pressure on MDT meetings drastically, specifically among non-surgical specialities, including oncologists, pathologists and radiologists who are often core members of multiple MDTs (Stephens et al., 2006). This could explain the pattern of core membership quoracy differences observed in this study.

The recent COVID-19 pandemic has resulted in the need to take a fresh view of current practice, and due to the economic restrictions and reduction of footfall in clinical areas, it is timely to review efficiency and seek out novel avenues to deliver services. Institutions are now routinely performing video conferencing and using modern technologies to support MDT functioning. It is intuitive that information communication technology (ICT) should positively

impact on cancer MDT meetings, however there is little evidence to support this conventional wisdom.

Qualitative work has reported that improving real time data collection and feedback has the potential to improve quality, care coordination, and patient-centered models of care (Janssen et al., 2018). The Multidisciplinary meeting Assistant and Treatment sElector (MATE) is an example of a clinical decision support (CDS) system with potential to be more widely used (Patkar et al., 2012). CDS systems are defined as systems that are designed to directly aid clinical decision-making, with the characteristics of an individual patient on an electronic health record matched to a computerised clinical knowledge base, and patient-specific assessments or recommendations then presented to the clinician(s) and/or the patient for a decision (Department of Health, 2000).

2.4.1 Limitations

The Manual for Cancer Services: Skin Measures v 2.0 National Cancer Peer Review (NCPR), National Cancer Action Team 2011 was used, which is based on the NICE standard 'Improving outcomes for people with skin tumours including melanoma and Cancer Services Guidelines 2006'. It is acknowledged that the national requirements for MDTs vary across the four UK nations, as do the processes for their assessment, albeit with some elements of similarity. According to the NCPR, now referred to as the Quality Surveillance Team (QST), the attendance at each individual MDT should constitute a quorum, for 95.0% or more, of the meetings (National Institute for Health and Care Excellence, 2006). Additionally, there are limitations inherent to the study design. This study is a cross-sectional study and only calculated quoracy at one given time point. An accurate measure of quoracy reflects calculation

across an entire calendar year prior to the internal annual quality assurance assessment (National Institute for Health and Care Excellence, 2006).

2.5 Conclusion

This novel study, using an FOI study design to report observational data of national SSMDT working, is the first study to objectively measure the composition, quoracy and cost of Specialist Skin Cancer Multidisciplinary Team (SSMDT) meetings in the United Kingdom. The majority of SSMDTs in the UK are meeting at the required frequency, however they are not meeting the standards for attendance. The most common reason is the lack of clinical oncology presence. SSMDTs are costly to the NHS, and strategies need to be developed to mitigate this or revolutionise efficiency. This should be taken into consideration by the NHS England's QST, the Scottish Intercollegiate Guidance Network, and the National Cancer Standards for Wales and Northern Ireland Cancer Registry for future policy changes. Future research will build on these findings by using a comparative analysis with different models of MDT working to investigative cost-effectiveness and improve efficiency.

Chapter Three:

Improving the effectiveness of multidisciplinary team meetings in skin cancer – analysis of national Cancer Research UK survey responses

3.1 Introduction

MDTs are an integral component of contemporary cancer care. Since their introduction to the UK in the 2000's by the NHS Cancer Plan, evidence has shown that MDTs improve outcomes for patients with cancer (Back et al., 2007; Bydder et al., 2009; Department of Health, 2000; Forrest et al., 2005; Kesson et al., 2012; MacDermid et al., 2009; Saini et al., 2012; Stephens et al., 2006). However, variations in treatment still exist and as such, MDTs have not been entirely successful in their aim of reducing variation in access to the best, evidenced-based care (Munro, 2015). As the previous chapter highlighted, there are significant direct and indirect costs of MDT working and evidence demonstrates that some MDTs function more effectively than others (Fleissig et al., 2006; Taylor et al., 2012). Regular meetings to discuss patients also present an opportunity cost to MDT members who have other critical roles in the NHS, negatively impacting on workflows in their respective departments e.g. pathology and radiology (Kane et al., 2007). With an ageing population, longer life expectancy, higher patient expectations and the long-term impact of a cultural shift to increased sun exposure now being realised, it is unlikely that the health service will be able to cope unless care services are designed and planned more effectively. Innovative solutions to new ways of MDT working are therefore warranted.

Despite MDTs being central to the management of patients with skin cancer, previous work investigating the functionality and financial impact of SSMDT meetings in the UK has deemed that they are costly and do not currently meet NICE quoracy standards (Ali, Dobbs, et al., 2021). The mean unit cost per discussion at an SSMDT has been quoted as £132.68 vs £91.84 for breast cancer, the most common cancer in the UK (Ali, Dobbs, et al., 2021; Cancer Research UK, 2024-a). Whilst only 26.0% of SSMDTs are quorate by membership with a lack of clinical oncology presence as the most common reason for failure (Ali, Dobbs, et al., 2021). It is evident

that there is a need to standardise operationalisation to reduce variations in cost. Identifying alternative methods to enhance the effectiveness of MDT meetings in skin cancer care services puts us in a position to recommend redesigns for a service that is safer, more effective, more convenient, and more cost-efficient for patients and the NHS

The 2017 Cancer Research UK (CRUK) commission investigated the effectiveness of MDT meetings in cancer services with the aim of improving MDT effectiveness (Cancer Research UK, 2024-b). The crucial findings from this report included the lack of sufficient time to discuss complex patients, suboptimal meeting attendance, not utilising the right information to inform discussions, and MDTs not being able to fulfil their secondary roles, such as in audit and education. A number of approaches to streamlining MDT working have been identified, including the development of tumour-specific guidance (Cancer Research UK, 2024-b, 2024-a). However, no attempts have been made by the skin cancer community to undertake an evidence-based approach to identifying areas for future development.

The aim of this study was to identify skin cancer specific variation in views of current MDT practices and suggestions for refocusing MDT meetings. The culmination of this will provide a better understanding of existing skin cancer care in the UK, which will serve as a basis for developing new and more effective ways of SSMDT working.

3.2 Methods

CRUK distributed two surveys to all MDT members in the UK in 2017 (Cancer Research UK, 2024-b). The first of these surveys asked respondents to provide their opinion on the importance of 13 different areas to MDT working (Table 8) and current compliance of their MDT with each of these on a 6-point Likert scale (1 = not very important or never done to 6 =

extremely important or always done). Anonymised survey data files data were provided by the CRUK policy research manager upon request.

Table 8: Suggested key areas for cancer MDT working.

Areas
Stratify patients based on risk
Prioritise more complex cases
Incorporate discussion on patient preferences
Have results present for patient discussions
Audit decisions made by team
Discuss patients on 14-day pathway if investigations do not show cancer
Discuss patients at all stages in pathway
Enter patient details into database in real time
Ensure all required members are present
Ensure sufficient time to discuss patients
Circulate agenda in advance of meetings
Meeting owner takes charge of discussions
Time allocated for preparation in job plans

The second survey asked respondents to rank on a 5-point Likert scale the degree to which they agreed or disagreed with several recommendations for changes to MDT working (1 = strongly disagree to 5 = strongly agree). Responses to both surveys from all skin MDT members (LSMDT) and (SSMDT) were for analysis. Results were reported with median values and range. Responses from members belonging to other MDTs across the UK (brain, haematological, gynaecology, upper gastrointestinal, head and neck, colorectal, lung, urology and breast) were included for comparison. Statistical data analyses were performed using RStudio (R Core Team, R Foundation for Statistical Computing, Vienna, Austria) and Mathematica Version 12.3 (Wolfram Research, Inc., Champaign, Illinois, USA).

3.3 Results

3.3.1 Characteristics of respondents

The first CRUK survey had 2,294 responses, of which 181 were from skin MDT members from a range of professional groups (Table 7). The second CRUK survey had 1,269 responses, of which 101 were from skin MDT members (Table 9). Respondents covered all areas in the UK. Overall there was a response rate of 50.0% to both surveys from skin MDT sites.

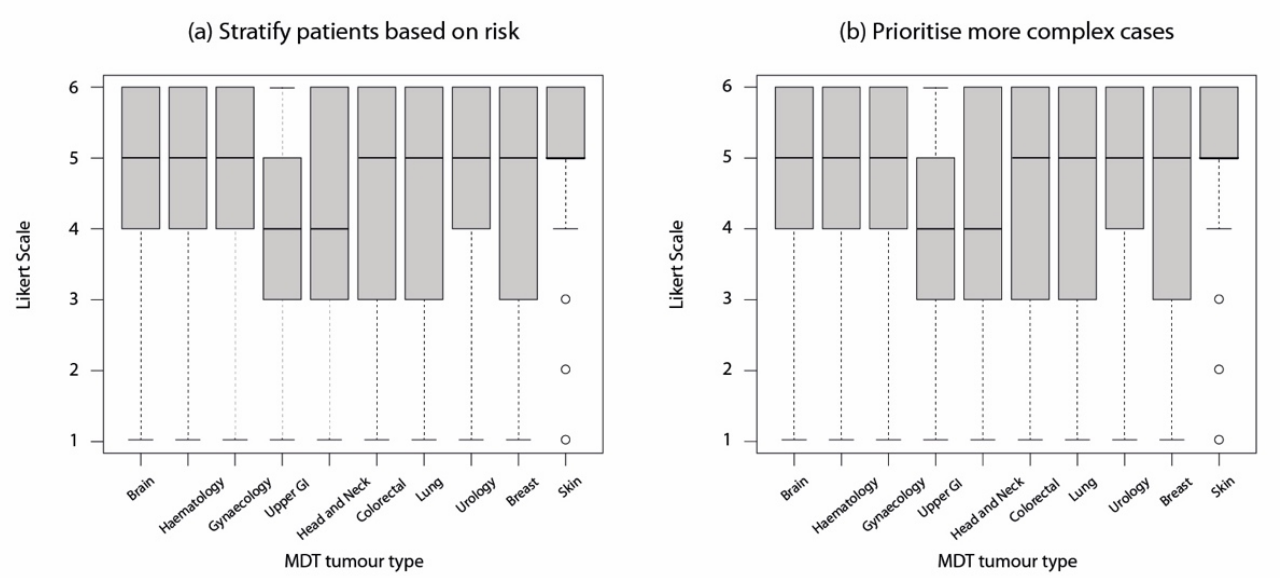
Table 9: Number of survey respondents for major cancer types.

	Number of respondents (% of total)	
Tumour type	**Survey 1**	**Survey 2**
Brain	81 (3.5%)	43 (3.4%)
Haematology	161 (7.0%)	77 (6.1%)
Gynaecology	160 (7.0%)	89 (7.0%)
Upper GI	185 (8.1%)	90 (7.1%)
Head and Neck	178 (7.8%)	124 (9.8%)
Colorectal	293 (12.8%)	132 (10.4%)
Lung	260 (11.3%)	141 (11.1%)
Urology	263 (11.5%)	160 (12.6%)
Breast	322 (14.0%)	177 (13.9%)
Skin	**181 (7.9%)**	**101 (8.0%)**

3.3.2 Current practise

Of the 13 questions in the first survey, there was more uniformity (Likert score 5 [range 4-6]) amongst skin respondents in the belief that stratification on risk and prioritising more complex cases were the most important factors compared to other MDTs (Figure 8).

Figure 8: Extent of importance of key areas of MDTs. MDT members ranked on a Likert scale their current level of implementation of the following factors (1 = low to 6 = high. Bold bar represents median; boxes represent interquartile range; whiskers represent overall range. Outliers are represented by dots.



For factors considered potential targets for future improvement, there was considerable variation in the extent that they are implemented (Figure 9). There was a mismatch between the majority of importance and implementation Likert scores (Figure 10).

Figure 9: Extent of implementation of key areas of multidisciplinary teams. MDT members ranked on a Likert scale their current level of implementation of the following factors (1 = low to 6 = high. Bold bar represents median; boxes represent interquartile range; whiskers represent overall range. Outliers are represented by dots.

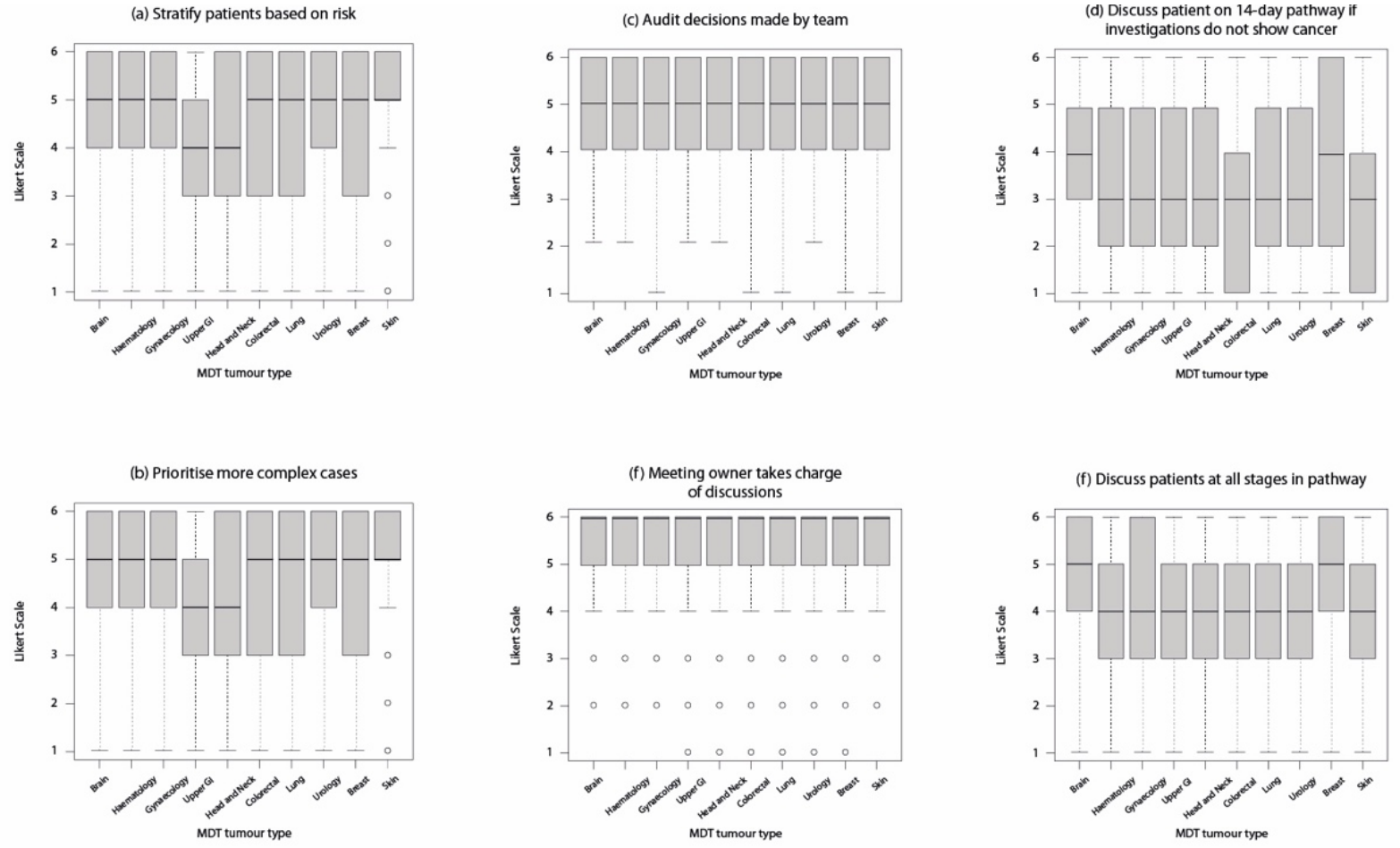
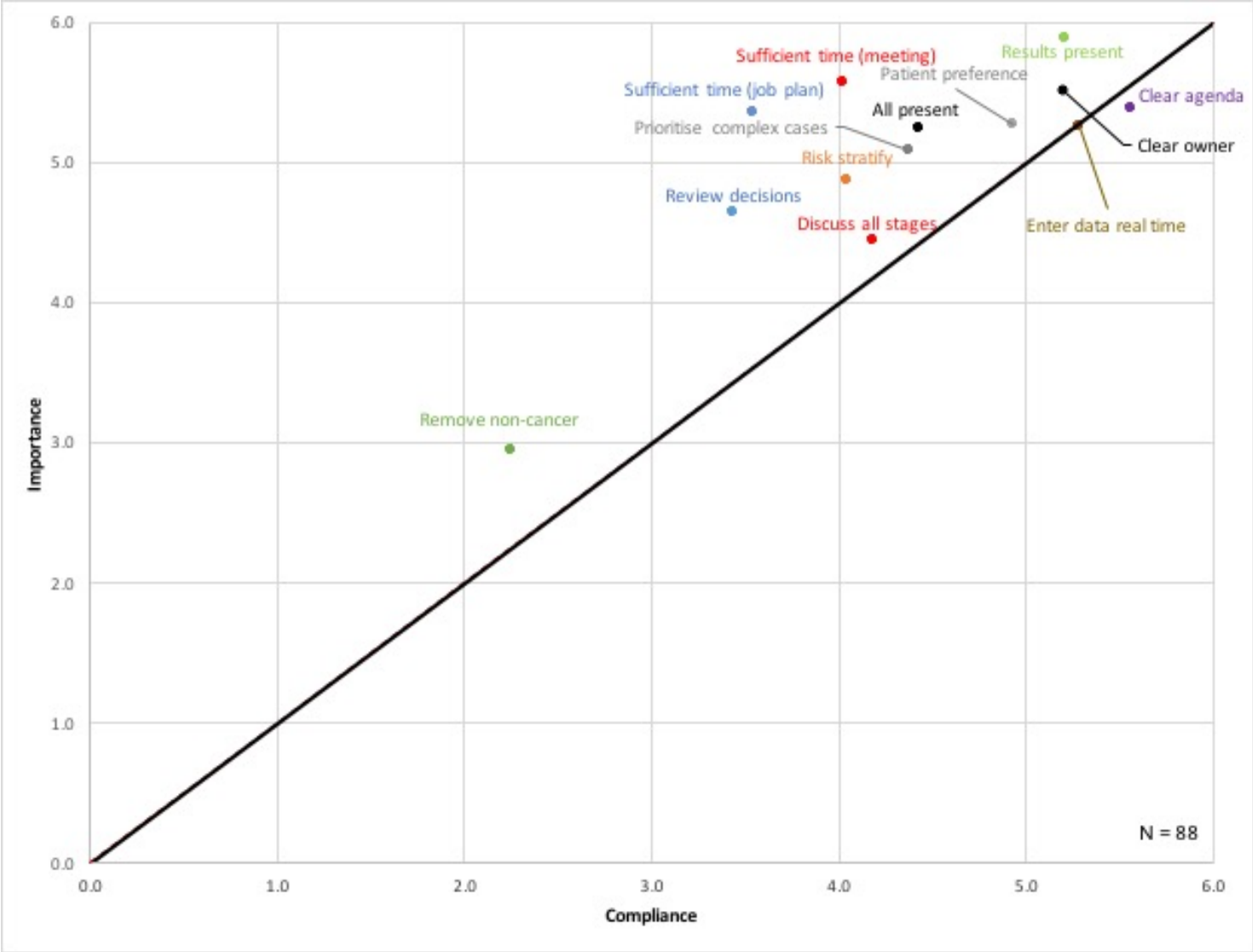


Figure 10: Plot of compliance vs importance Likert scores from skin respondents.



3.3.3 Recommendation for changes to skin MDT working

The most important priorities in the changes necessary to MDT working deemed by the skin MDT were 1) imaging and pathology results ready for the meeting 2) time to discuss patients in detail 3) clear meeting owner in charge and 4) clear agenda, in advance of the meeting. These results closely mirror those of other MDT respondents except that other MDTs ranked 'clear meeting owner in charge' higher (Table 10).

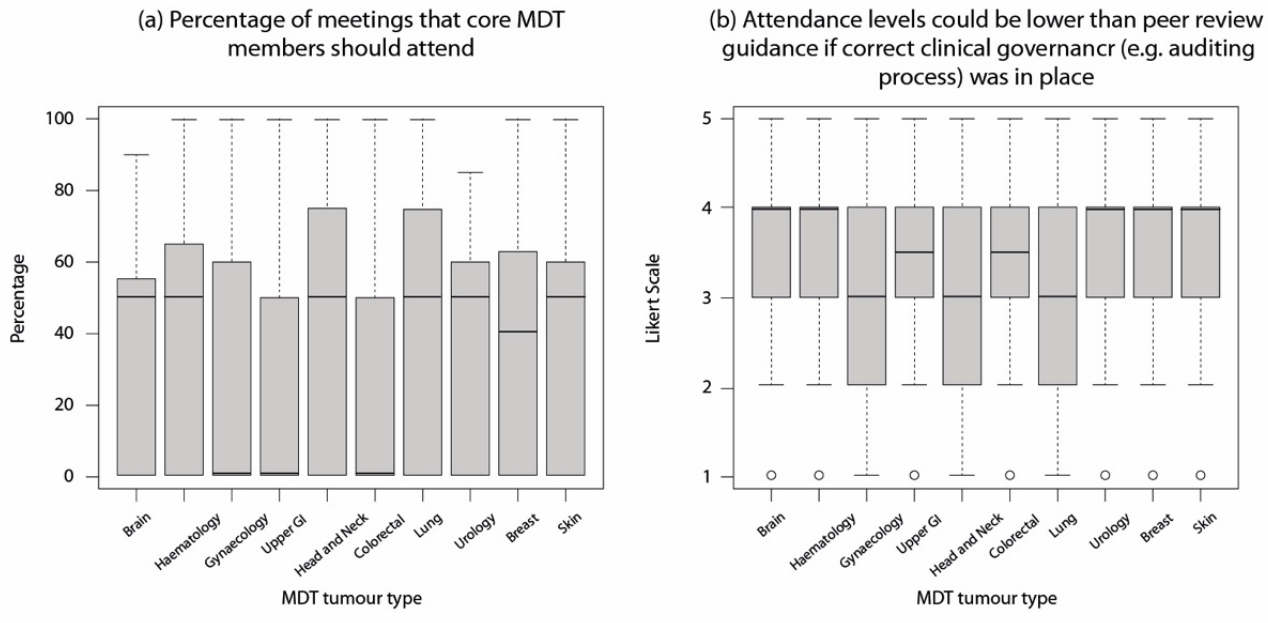
Table 10: Summary of priorities of importance for across different MDTs.

Tumour group	Priority 1	Priority 2	Priority 3	Priority 4
Brain	Imaging, pathology results ready	Clear meeting owner in charge	All required members present	Time to discuss patients in detail
Breast	Imaging, pathology results ready	Time to discuss patients in detail	All required members present	Clear meeting owner in charge
Children and Young People	Time to discuss patients in detail	Clear meeting owner in charge	Clear agenda, in advance	All required members present
Colorectal	Imaging, pathology results ready	Time to discuss patients in detail	All required members present	Clear agenda, in advance
Gynaecology	Imaging, pathology results ready	Time to discuss patients in detail	Prep time in job plan	Clear meeting owner in charge
Haematology	Imaging, pathology results ready	Time to discuss patients in detail	Clear agenda, in advance	Clear meeting owner in charge
Head and Neck	Imaging, pathology results ready	Time to discuss patients in detail	Clear agenda, in advance	Clear meeting owner in charge
Lung	Imaging, pathology results ready	Time to discuss patients in detail	Clear meeting owner in charge	All required members present
Sarcoma	Imaging, pathology results ready	All required members present	Time to discuss patients in detail	Clear meeting owner in charge
Skin	Imaging, pathology results ready	Time to discuss patients in detail	Clear meeting owner in charge	Clear agenda, in advance
Upper GI	Imaging, pathology results ready	Time to discuss patients in detail	Clear agenda, in advance	All required members present
Urology	Imaging, pathology results ready	Time to discuss patients in detail	Clear meeting owner in charge	All required members present
Summary of most important priorities	**Imaging, pathology results ready**	**Time to discuss patients**	**Clear agenda, in advance**	**Clear meeting owner in charge**

3.3.4 Clinician attendance levels

Skin MDT respondents agreed that core-MDT members should attend more than 50% of the total meetings annually (Figure 11a) as defined by cancer indicators derived from the Manual for Cancer Services: Skin Measures v 2.0 National Cancer Peer Review (NCPR), National Cancer Action Team 2011 (National Cancer Peer Review-National Cancer Action Team, 2011) which is based on the NICE standard 'Improving outcomes for people with skin tumours including melanoma and Cancer Services Guidelines 2006' (National Institute for Health and Care Excellence, 2006). However, skin MDT respondents agreed (median Likert score 4) that attendance levels could be lower than peer review guidance if correct clinical governance (e.g. auditing process) was in place (Figure 11b).

Figure 11: Differences in view of skin MDT respondents on clinician attendance levels.



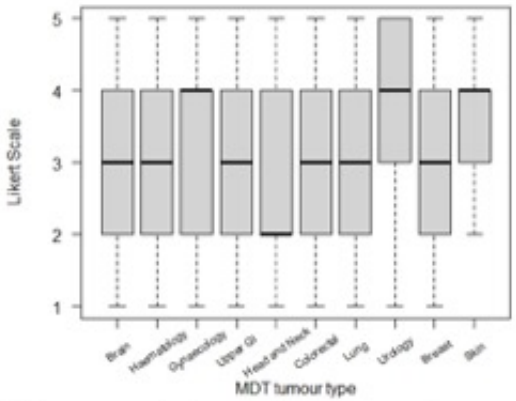
3.3.5 Protocolised streamlining

There was agreement (median Likert score 4) amongst skin MDT respondents that some patients should be placed on protocolised treatment pathways and do not need to be discussed at the meeting at all, whilst the other main specialties were uncertain (median Likert score 3) about this (Figure 6). Most specialities including skin agreed (median Likert score 4) that if protocolised streamlining were to take place that it should take place in advance of the main MDT meeting in order to decide which patients should be discussed (Figure 6). Skin respondents felt that this could allow more straightforward cases to be progressed more quickly, rather than waiting for a weekly meeting (Figure 6) and deemed that 30.0% of cases could potentially be resolved outside of the meeting (Figure 6) which was higher compared to other specialities. Skin respondents agreed with the majority of other specialties (median Likert score 4) that if patients followed treatment protocols or had recommendations made by a

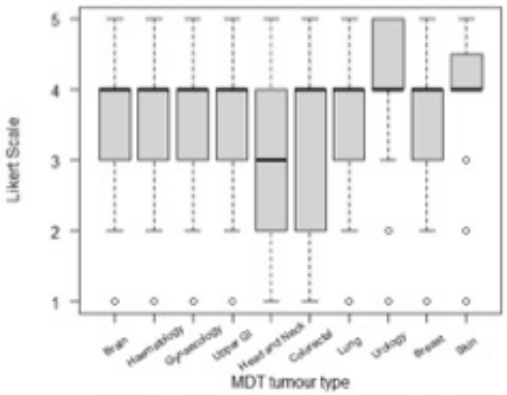
smaller team, the full MDT reviewing a selection of these patients would provide sufficient governance of this process (Figure 12).

Figure 12: Differences in view of skin MDT respondents on protocolised streaming.

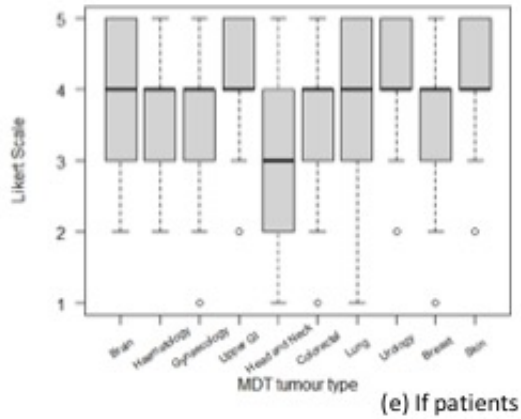
(a) Patients should be placed on protocolised treatment pathways and are not needed to be discussed at the meeting at all.



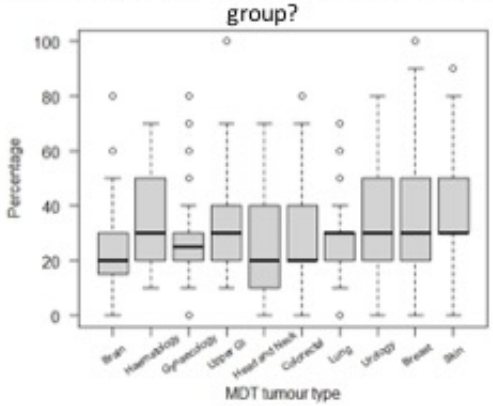
(b) The streamlining of patient discussions should be performed in advance of the main MDT meeting in order to decide which patients should be discussed at the meeting, and which should receive a protocolised treatment plan.



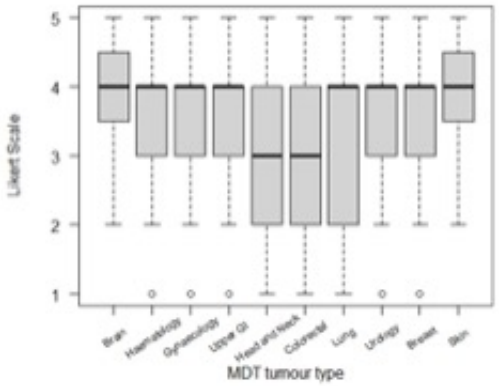
(c) This approach of streamlining patient discussions could allow more straightforward cases to be progressed more quickly, rather than waiting for the weekly meeting.



(d) What percentage of patients do you feel could be resolved outside of the meeting, for example, through clearly defined treatment protocols and review by a smaller group?



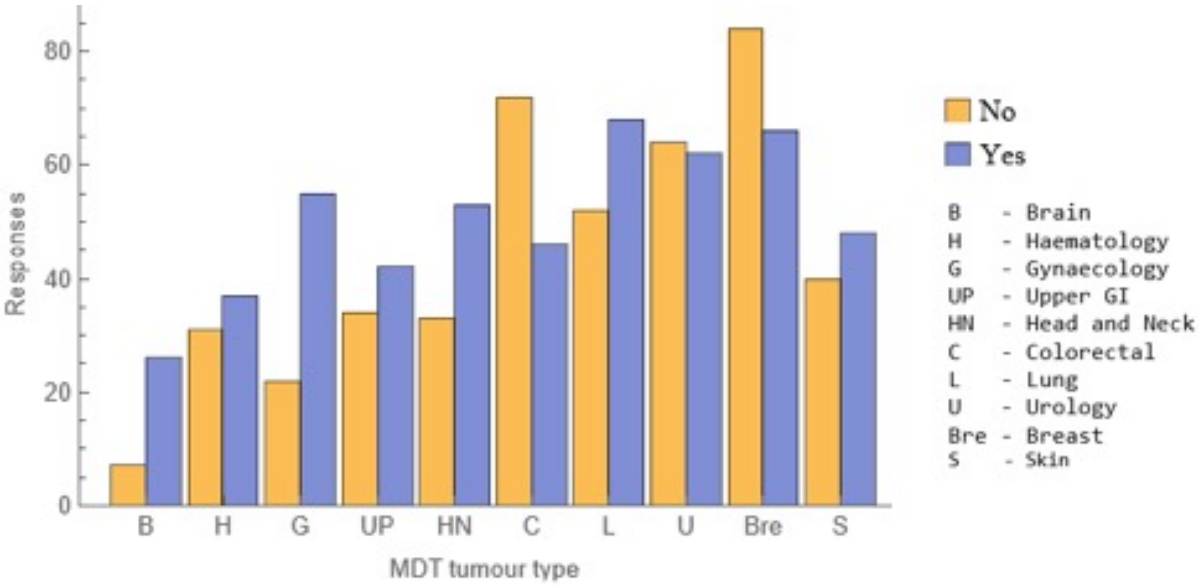
(e) If patients followed treatment protocols or had recommendations made by a smaller team, the full MDT reviewing a selection of these patients would provide sufficient governance of this process



3.3.6 Meeting preparation

Most MDTs use a form of checklist or proforma to inform referrals to the MDT (Figure 7). All respondents felt that use of a checklist would help improve MDT meetings (median Likert score 4), however skin respondents had the widest range in this response (1-5).

Figure 13: Responses to the question **"**Do you currently use any form of checklist or proforma to inform referrals to your MDT?"



3.3.7 Identifying case complexity

Participants were asked to identify the importance of a number of potential issues that could increase case complexity and require discussion in a full MDT. There were a range of medical, surgical, psychological and treatment factors considered important (Likert score ≥4) to skin MDT members (Table 11) which closely mirrored that of other specialities.

Table 11: Factors suggested by skin respondents which would increase case complexity and require escalation to full MDT discussion.

Category	Factor
Medical	Patient discussed in the meeting has unusual or rare tumour type
	Patient has a poor performance status (i.e., they are frail and/or need assistance with care/mobility)
	Patient has significant physical co-morbidity (e.g. diabetes, congestive heart failure, kidney or vascular disease, immunocompromised or suppressed).
Surgical	Patient has a significant past surgical history (e.g. relevant previous surgeries that may affect surgical options)
Psychological	Patient has a significant mental health or cognitive co-morbidity (e.g. they are sanctioned under the Mental Health Act, have schizophrenia, dementia from stroke or Alzheimer's disease)
Treatment	Patient has treatment failure (i.e., there is cancer progression despite current treatment)
	Patient experienced treatment toxicity and/or contraindications to standard treatment
	There is a conflict of opinion regarding the best treatment option for a patient
	Guidelines/pathway do not account for patients' specific situation, (i.e. exceptional case)

3.4 Discussion

Fieldwork and observations that underpinned the CRUK report into 'Improving the Effectiveness of Multidisciplinary Team Meetings in Cancer Services' demonstrate that there is not enough time to discuss complex patients, attendance is not optimal, the right information is often not used to inform discussions and that MDTs are unable to fulfil their secondary roles in data validation, audit and education (Cancer Research UK, 2024-b). The evidence base for cancer treatment is accumulating constantly, with NICE regularly updating their advice. Innovative solutions to these problems identified by CRUK would have health economic benefit, free up specialist time and improve the reproducibility of evidence-based decision making.

Whilst it is incumbent to refresh the format of MDT meetings to reflect the changing nature of cancer care and increased demand for services, solutions will not be the same for every MDT or every specialty. Previous national data investigating composition, quoracy and cost of SSMDT meetings has shown that SSMDTs in the UK are not currently meeting NICE quoracy (Ali, Dobbs, et al., 2021). In addition to this there is large variation in mean cost per patient, time (preparation, running and dissemination), total cases discussed and case re-discussion across SSMDTs (Ali, Dobbs, et al., 2021). Whilst it is clear SSMDTs are costly and there is a need to improve MDT meeting efficiency without losing the considerable benefits associated with regular meetings, there are no previous studies that provide an evidence-based approach to identifying areas for future development, specific to the skin MDT. It is important to understand whether tumour site affects the areas of MDT practice. If MDTs are relatively homogeneous in terms of practices and priorities then recommendations can address the issues and priorities of a broad range of MDTs, and therefore be broadly applicable.

In the present study it was identified that members of skin LDMDTs and SSMDTs considered the ability to stratify patients based on risk and prioritise more complex cases to be more important than other MDT members. This likely represents either the fact that NMSC is the most common group of cancers in the UK and therefore results in a high volume of patients processed by skin MDTs, or that there is a feeling that many of these NMSCs are less serious and therefore those that are more so should be prioritised. Clearly a one-size fits all approach for all tumour types is not appropriate for protocolised streaming. Previous studies support the development of tumour-specific guidance for streamlining MDT discussions considering a range of approaches (Cancer Research UK, 2024-b; Hoinville et al., 2019).

Historically, UK guidelines have recommended that all cases of high-risk SCC and MM are discussed at the skin MDT and omit any recommendations on referral for BCC (Keohane et al., 2021; National Institute for Health and Care Excellence, 2015; Telfer et al., 2008). However, the 2021 BAD UK BCC guidelines now highlight the pivotal role of the MDT in the management of high-risk BCC (Nasr et al., 2021). Given these new recommendations, the caseloads for LSMDTs and SSMDTs are only expected to increase, making timely solutions to these factors critical for the health service's ability to manage the growing demand for peer review.

3.4.1 Meeting attendance

MDT respondents felt that there should be no change to clinician attendance levels. They did however agree that that attendance levels could be lower than peer review guidance if correct clinical governance was in place. This would potentially have the ability to make better use of some of the specialties' time commitments e.g. radiologists, oncologists and histopathologists who may only need to attend part of the meeting to discuss specific cases. Previous work has demonstrated reduced core membership quoracy at the SSMDT in these specialities who have critical roles in the NHS for imaging, endoscopy and pathology capacity. MDTs should bring together staff with the necessary knowledge, skills and experience to ensure high quality diagnosis, treatment and care. Each patient will require a different set of core team members to discuss their case. Treatment options may be lost if some specialties are not in attendance at the MDT (Cancer Research UK, 2024-b). However, pre-MDT structured listing of patients and grouping could allow specific MDT members to stay only for the required period as recognised by the BAD multi-stakeholder workshop that discussed and proposed recommendations for changes to the structure and function of the skin MDT (British Association of Dermatologists,

2018). National guidance on quoracy standards therefore needs to be reviewed in respect to this.

3.4.2 Meeting preparation and protocolised streaming

In the current study skin respondents considered 'imaging and pathology results ready for the meeting' as the top priority for changes necessary to MDT working. Fieldwork and observation of 624 MDT discussions by CRUK has revealed that 7.0% are deferred (Cancer Research UK, 2024-b). A main contributor to this is missing diagnostic information. Delays may occur at different stages of the cancer diagnostic journey and secondary care delay (delay in being first seen in secondary care to diagnosis) due to missing diagnostic information is both distressing for the patient but also wastes valuable MDT discussion time. Fifty four percent of MDTs currently use a proforma however usage is inconsistent and there is no national guidance on proforma use (Cancer Research UK, 2024-b). All respondents in the study felt that use of a checklist would help improve MDT meetings and CRUK in fact now recommend that MDTs should mandate a completed proforma for incoming MDT referrals prior to discussion (Cancer Research UK, 2024-b). This would go some way to mitigate deferral and secondary care delay due to incomplete or missing diagnostic information.

As evidenced here there is a general acceptance by skin respondents of the benefits of protocolised treatment. This would reduce some of the work associated with the more straightforward cases and ensure that quoracy could be met and allow MDTs to undertake important secondary roles in data validation, audit and education. CRUK recommend that the UK's health services should work with NICE and Scottish Intercollegiate Guidelines Network (SIGN) to identify where a protocolised treatment pathway could be applied and develop a set of treatment recommendations for each of these to be implemented across the four nations

(Cancer Research UK, 2024-b). The response by the BAD to the suggestions of protocolised treatment is that individual MDTs should consider formalised management protocols for routine cases that can be managed on a treatment pathway without the need for formal discussion by the full MDT (British Association of Dermatologists, 2018).

The Association of Breast Surgery, British Society of Breast Radiology, Association of Breast Pathology and the UK Breast Cancer Group have jointly produced the Breast MDT meeting Toolkit (Association of Breast Surgery, 2024). This toolkit is a comprehensive resource and includes guidance on ways to conduct a pre-MDT triage meeting as one component of the toolkit. The pre-MDT triage meeting can be an effective way of reducing the number of cases requiring formal discussion. A defined smaller group of MDT members meet with the MDT coordinator in advance of the meeting to determine cases that should 1) be listed for formal discussion, 2) managed without formal discussion and 3) be suitable for management by protocolisation to a standard of care (SoC). The SoC can be defined as a point in the pathway of patient management where there is a recognised international, national, regional or local guideline on the intervention(s) which should be made available to a patient (Association of Breast Surgery, 2024).

With the guidance on streamlining according to clinical complexity and guidelines having been published by NHS England and NHS Improvement in 2020 – discussing all cancer cases is? now unnecessary (NHS England and NHS Improvement, 2020). There is an urgent need for evidence-based approaches that can be used by the skin MDT to streamline services, while maintaining the safety and quality of patient care. Existing validated tools can be applied at different points along the MDT pathway to build a protocolised streaming programme to ensure the delivery of excellent cancer care whilst safety is maintained. These tools can facilitate pre-

meeting case selection, intra-multidisciplinary team meeting streamlining and team reflection, assessment and team building (Soukup et al., 2020).

Another novel solution to support protocolised streaming includes a machine-learning approach to clinical decision support. Andrew et al developed a supervised machine-learning algorithm to predict MDT decisions for Mohs micrographic surgery (MMS) vs conventional surgery or radiotherapy (Andrew et al., 2022). By using their model, 37.5% of patients were able to be triaged to MMS and reduce the overall MDT workload by 45.1%. Whilst the authors determined that this approach would provide more time for MDT members to consider more complex patients the predictive accuracy precludes their use as a fully autonomous system and a 'human in the loop' would still be required to review treatment decisions and account for the shortcomings in system performance.

3.4.3 Complex patients

In surgical oncology, the traditional approach has been to use tumour factors to predict patient outcomes and determine treatment pathways. Patient factors have been considered secondarily, mainly to decide on the suitability of a treatment pathway, especially in cases where the treatment carries a risk of toxicity or morbidity. There are many non-invasive measures of patient status such as the American Society of Anesthesiologists (ASA) physical status classification grade, Rockwood clinical frailty scale, World Health Organization (WHO) performance and Karnofsky performance status which predict clinical frailty, and Charlson comorbidity index, Portsmouth-Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (P-POSSUM), the Acute Physiology and Chronic Health Evaluation (APACHE II) which predict 30-day perioperative mortality (Moonesinghe et al., 2013). Routine use of these patient-based risk stratification tools at the outset of an MDT could allow patient screening and prioritisation of complex cases and to facilitate pre-meeting

case selection. There are authors who have commented on how the adoption of the Rockwood clinical frailty scale can potentially improve the quality and focus of MDT discussions for complex patients at their skin MDT (Moncrieff & Nobes, 2021).

3.4.4 Limitations

The primary data did not record whether skin respondents belonged to either LSMDT or the SSMDT. LSMDT and SSMDT have similar but different roles. The NHS Cancer Plan recommends that 'the care of all patients with cancer should be formally reviewed by a specialist team'. The BAD recognises that LSMDTs currently keep clinicians 'honest' and act as a 'safety net' allowing for increased scrutiny and for personalised care of individuals with increased local knowledge (British Association of Dermatologists, 2018). Accordingly, it would be valuable to identify variations between LSMDTs and SSMDT practice and if there are any specific suggestions that could be tailored to these members and ergo patients which were not able to investigated in the current study.

While internet based surveys offer significant advantages, including broad accessibility, cost-effectiveness, and the ability to collect large datasets efficiently, certain methodological limitations should be considered when interpreting the findings. Internet access itself is unlikely to be a barrier for NHS clinicians, however engagement with online surveys may be influenced by competing clinical responsibilities, survey fatigue, and self-selection bias. Respondents who choose to participate may be those with more flexible schedules or stronger opinions about MDT functionality, potentially skewing the representativeness of the dataset. The lack of direct researcher participant interaction also limits opportunities for clarification, increasing the risk of varied interpretations of survey items. Additionally, self-reported data may be influenced by recall bias, with respondents potentially overestimating adherence to best

practices or underreporting inefficiencies. While the insights gained from this survey provide a valuable overview of MDT member perspectives, they should be interpreted alongside complementary data sources, such as direct MDT observations, meeting audits, and qualitative interviews, to ensure a comprehensive and balanced understanding of MDT functionality and areas for improvement.

3.5 Conclusion

The views of skin MDT members in the current study support changes to meeting attendance, preparation and protocolised streaming. There is now a mandate for protocolised streaming at a national level. Clearly a one-size fits all approach for all tumour types is not appropriate for protocolised treatment and tumour-specific guidance for streamlining the skin MDT is needed. Encouraging those seeking to implement change in skin MDT practise to consider the views held by respondents identified in this study and make use of the wide range of evidence-based tools available to test, develop and re-assess changes in MDT practice is needed.

Chapter Four:

Evaluating remote skin cancer multidisciplinary team meetings in the United Kingdom post-COVID-19

4.1 Introduction

Previous work has established that SSMDTs are costly when compared to other MDTs yet often do not meet NICE standards for composition and quoracy (Ali, Dobbs, et al., 2021). Innovative solutions as to how the skin cancer community in the UK can improve the effectiveness of SSMDTs is therefore a priority for patients and clinicians alike. Analysis of national CRUK survey responses from skin MDT members highlights that there is support for change to the way meetings are attended, prepared for and patient management to be protocolised (Ali, Dobbs, Jovic, Hutchings, et al., 2023). With the incidence of skin cancer rising faster than that of any other malignancy and the indications for MDT referrals continuing to expand, it is imperative that a time critical evidence-based assessment of the fundamentals of the skin MDT is made (Cancer Research UK, 2024-d, 2024-c). Prior to the COVID-19 pandemic some authors commented on how the virtual or remote MDT was a niche concept and unlikely to replace the more traditional face-to-face MDT. However, the sudden shift to virtual meetings at the beginning of the COVID-19 pandemic has been one of the most dramatic changes since the inception of the MDT concept and one that has persisted even with the relaxation of social distancing rules.

Despite this persistent change no studies have evaluated the impact this has had on the skin MDT. In this study a virtual or remote MDT was defined as being any MDT meeting which had been conducted over a digital/virtual conferencing medium, and where the majority of participants connected remotely. The aim was to understand how effectively skin MDTs have been conducted since the move to virtual meetings (focusing on data security, confidentiality, decision making, efficiency, and organisation), to understand how members perceived their experiences of the virtual meetings (with a focus on teamwork, training, and engagement), and

also to determine underlying issues with virtual MDT meetings and elicit possible areas for improvement.

4.2 Methods

In a previous study prior to the pandemic, FOI requests were sent to all NHS trusts across the UK to identify those with a SSMDT (Ali et al., 2021). Any sites holding a SSMDT were then sent an electronic survey designed using Google Forms (Google LLC, Menlo Park, California, USA) during April 2022. The survey was also distributed to all members of the British Association of Plastic, Reconstructive and Aesthetic Surgeons (BAPRAS) Skin Oncology Special Interest and Advisory Group (SIAG).

The survey design was informed by a previous study in head and neck surgery, with questions pertaining to communication, data protection, decision making, training, engagement, technology, and organisation/coordination (Appendix 1). This was adapted for the skin MDT based on findings from previous work which highlight ways of improving the effectiveness specific to the skin MDT (Ali et al., 2023).. Standardisation was achieved through clear instructions to ensure the questionnaire was administered in a consistent and uniform manner to all participants. For all aspects of MDT functionality, questions were asked that required participants to record ordinal e.g. better, about the same, worse or nominal e.g. don't know, responses respectively. To specifically record satisfaction in the area of technology and security a Likert scale was included ranging from 0 to 5 (very unsatisfied to very satisfied respectively). Descriptive and statistical data analysis were performed using Microsoft Excel 2010® (Microsoft Cooperation, Redmond, Washington, USA).

4.3 Results

4.3.1 Characteristics of respondents

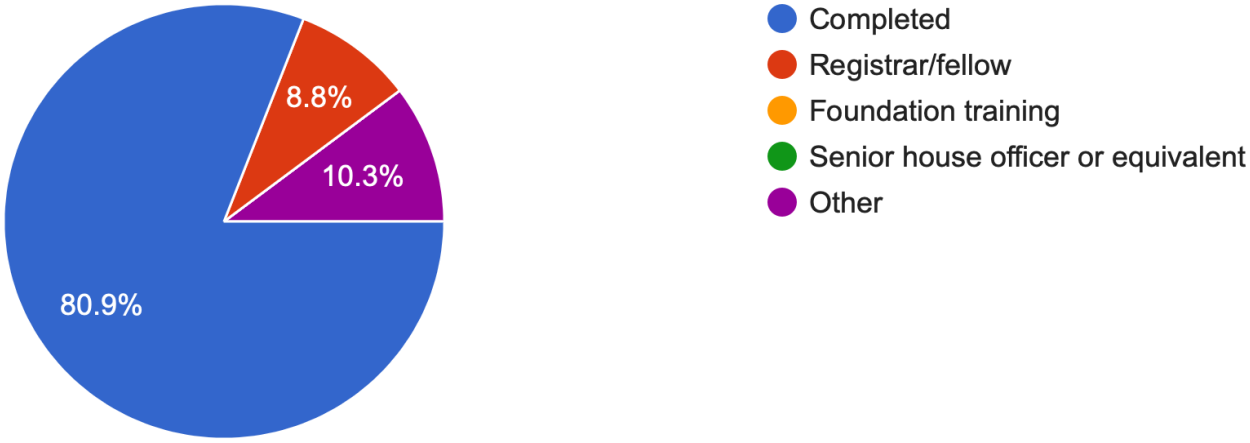
There were 68 responses from members of 36 different SSMDTs. This represented a response rate of 55.3% of all SSMDTs. These included teams from all four countries of the UK. Responses included all core members of the SSMDT (Table 12).

Table 12: Roles of survey respondents.

Role in MDT	Number of respondents (%)
Surgeon	21 (30.9)
Dermatologist	16 (23.5)
Nurse	15 (22.1)
Oncologist	5 (7.4)
Radiologist	4 (5.9)
Co-ordinator/administrative	3 (4.4)
Pathologist	3 (4.4)
GP with an Extended Role in Dermatology	1 (1.5)
Total	**68**

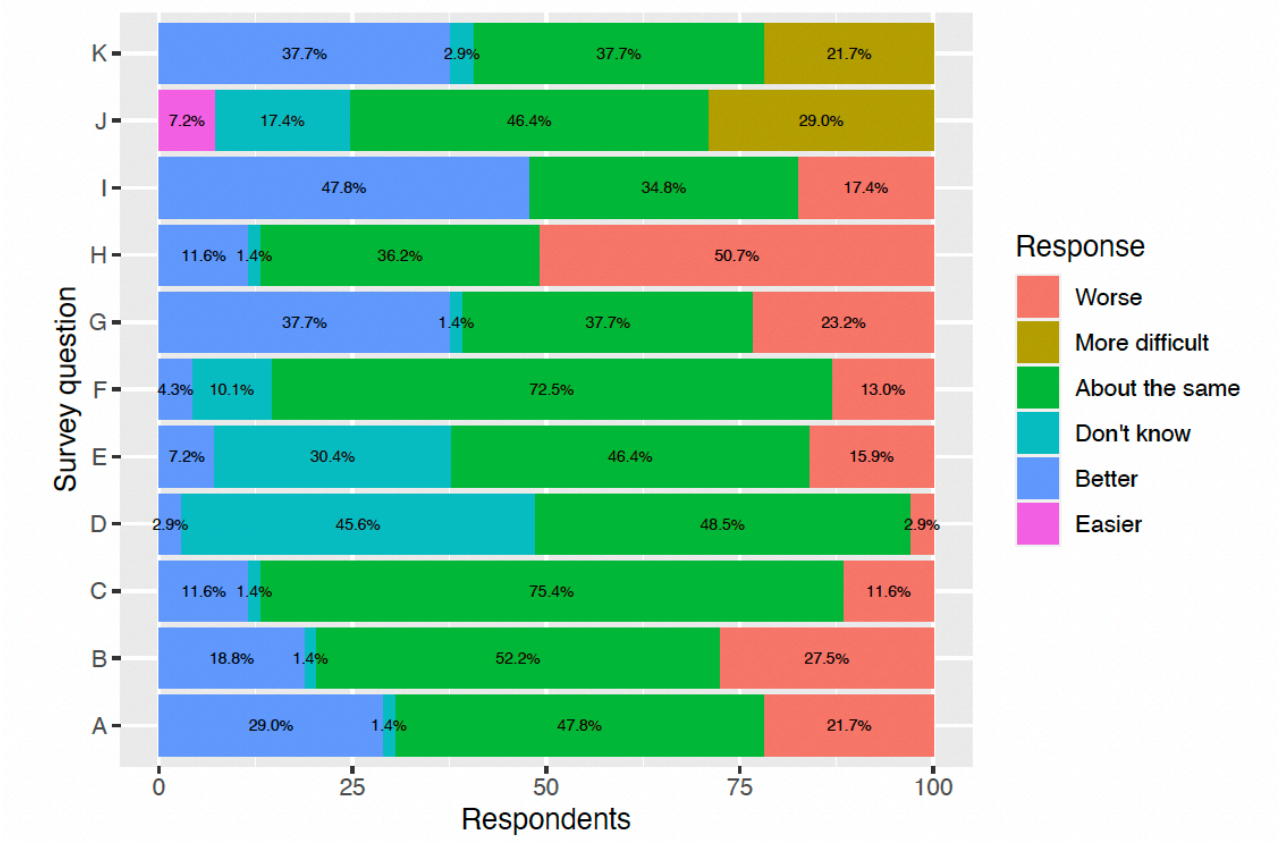
There was a range of training levels from respondents in clinical roles (Figure 14) with the majority having completed training.

Figure 14: Training levels of survey respondents.



Results of all aspects of MDT functionality are summarised in Figure 15.

Figure 15: Stacked bar chart summarising responses from questions surrounding all aspects of MDT functionality.



4.3.3 Communication, chairing and decision making

46.4% (the largest group) felt that chairing the meeting and recording outcomes in a virtual MDT was about the same as in-person. However, a significant proportion of respondents, 29.0%, felt it was more difficult. Similarly, 47.8% of respondents felt that communication overall was about the same. However, among the remaining respondents who did not choose "about the same", opinions were mixed - 29.0% felt it was better, 21.7% felt it was worse, and 1.4% were uncertain. Communication was comparable to decision-making with the majority of respondents (75.4%) feeling that this too remained about the same. Since moving to remote

working most respondents felt that time efficiency and organisation was better (47.8%) or about the same (34.8%).

4.3.4 Team working and engagement

50.7% believed that interpersonal relationships and team working were worse since moving to virtual meetings. Despite this, 37.7% felt that engagement of all specialties with the MDT process was better. This is in comparison to only 37.7% and 21.7% who felt that engagement by all specialties was about the same or was worse, respectively.

4.3.5 Training, clinical trials recruitment and audit

The largest group (52.2%) reported that training was about the same in virtual MDTs, compared to in-person meetings. However, a significant proportion (27.5%) reported a decline in training quality, with only a minority (18.8%) expressing an improvement. Responses to the impact of virtual conferencing software on viewing clinical images, dermoscopic images, pathology slides and radiological imaging demonstrate that MDTs felt this was about the same (37.7%) or was better (37.7%), with a smaller percentage (23.2%) feeling it was worse. Most respondents felt that recruitment to clinical trials and data validation and audit was about the same (largest responses groups 48.5% and 46.4% respectively) since the move to virtual MDTs.

4.3.6 Technology and security

Prior to COVID-19 only 66.7% of MDTs utilised virtual conferencing software, with this increasing to 100.0% in the post COVID-19 era. In 68.1% of respondents, there was agreement (median Likert score 4) that adequate technology skills and resources were in place to meet virtual MDT requirements despite 18.8% having never used virtual conferencing software prior

to the pandemic. 72.5% of respondents felt that since using virtual conferencing software data security and patient confidentiality during a virtual MDT is about the same when compared to conventional face to face MDTs.

4.3.7 The future

The majority of respondents' preference for the format of the MDT in the post COVID-19 era was face to face with the option to attend virtually (87.0%), whilst only 8.7% wished to return to a fully virtual MDT and 4.3% wished to return to a fully face to face MDT. There was no statistically significant association between geographical location (England, Wales, Scotland or Northern Ireland) and preference for MDT format (X^2=2.6833, df=6, p=0.8474). There were 17 responses for the free text responses available for recommendations on how virtual MDTs could be improved, representing a 25.0% response rate to this qualitative exploration. Descriptive analysis of these responses revealed four distinct themes (Table 13) summarised as audio-visual, engagement, secondary MDT functions and general MDT.

Table 13: Major themes and sub-themes following descriptive analysis of free text comments.

Theme	Extract(s)
Audio visual	
Better internet connectivity	"Better bandwidth would help with better sound and visual resolutions"
Provision of real time IT support	"Someone present who knows how to use the technology well and can troubleshoot should there be issues" "We have weekly issues with the software our hospital uses, sometimes meaning as MDT members are not present then phone calls have to be made instead between consultants, greatly reducing communication between the team"
Engagement	
Provide virtual MDT training	"Find the speed that some presenters speak is too fast" "The simplest improvement would be for members to remember to mute their microphones when not contributing"
Enhance non-verbal cues during virtual MDT	"The IT platform used must allow for 'hand up' alerts to the chair for people to speak, currently our system […] does not and people can talk over each other unknowingly, others' voices drown out and it can be hard to hear all the points being made. Hands up alerts (like Microsoft Teams or Zoom) permit the chair to invite everyone to speak in turn"
Enhance integration of new staff	"MDT worked well together when switched to completely virtual because we all had well-established relationships. With new staff joining whom I've not met in person, I notice that it's a little more challenging due to not knowing each other" "Ultimately it helps to have real life eye contact with your team but it works well enough with some definite efficiencies if systems are working well, terrible when not. Teams should also aim to meet physically perhaps quarterly to ensure good rapport and morale"
Secondary MDT functions	
Training	"Allocated education time at the end"
General MDT	
Preparation	"Clinicians [should] prepare for the meeting and know the patient"
	"More time dedicated towards discussion of cancer patients in job plans for MDT core members and as chair of MDT"

Despite the quantitative responses highlighting how communication overall was about the same, the qualitative responses under the audio-visual theme signify a general need for better internet connectivity and provision of real time IT support. Analysis of the engagement theme revealed how there was a need to provide virtual MDT training, enhance non-verbal cues during virtual MDT and enhance integration of new staff. All of which could drive

improvement in interpersonal relationships and team working which the quantitative responses highlighted a need for. Despite the majority of respondents feeling that training was about the same in virtual MDTs the qualitative responses in the secondary MDT function theme suggest that allocated education time at the end of virtual MDTs could improve training for staff. Finally, respondents felt that MDT preparation should be a key area for future improvement common to both virtual and face to face meetings.

4.4 Discussion

The dramatic and rapid shift to remote working since March 2020 has seen, not just in healthcare, but the wider world, a move to virtual meetings. During the height of the pandemic this had obvious benefits, however since the relaxation of social distancing this new way of working has persisted, and is likely to remain in some form long-term. This has been accepted by the NHS tacitly with their recent priorities and operational planning guidance for 2022/23 (NHS England, 2024-a). Within it, NHS England state that we should consistently adopt new models of care that exploit the full potential of digital technologies. This study aimed to determine the experiences of skin MDTs in the UK with this new remote working and meeting. It aimed to reflect on the benefits, limitations and ultimately inform on how best to develop and improve this novel way of meeting. The first key finding of this study is that despite relaxation of social distancing rules, the vast majority of skin MDTs have chosen to continue to meet remotely. While most results showed no difference between virtual and in-person MDT meetings, it is important to acknowledge that there were some areas where virtual meetings were found to be inferior. The most notable result was a deterioration in interpersonal relationships and team working, which was reported by 50.7% of respondents. This is an important finding that warrants further consideration. The lack of face-to-face interaction and non-verbal cues during virtual meetings may make it more difficult for team members to build

rapport and develop a sense of trust and camaraderie. This is further demonstrated in these free-text responses:

"MDT worked well together when switched to completely virtual because we all had well-established relationships. With new staff joining whom I've not met in person, I notice that it's a little more challenging due to not knowing each other" and "Ultimately it helps to have real life eye contact with your team but it works well enough with some definite efficiencies if systems are working well, terrible when not."

Responses suggest that these issues have already been flagged and addressed by some MDTs. For example, some respondents suggested "Teams should also aim to meet physically perhaps quarterly to ensure good rapport and morale". In addition, obvious rules such as "the simplest improvement would be for members to remember to mute their microphones when not contributing".

Another area where virtual meetings may present challenges is effective chairing. 29.0% of respondents reported that decision making, chairing and communication was more difficult during virtual meetings, which suggests that virtual meetings may not be performing as well as in-person meetings in this area. Although 46.0% reported that chairing was the same in both virtual and in-person meetings, it is important to recognise that the number of respondents who found it more difficult is not insignificant. It is acknowledged that virtual meetings may present certain challenges when it comes to effective chairing, such as technical difficulties and lack of non-verbal cues.

In addition to the areas of team work and chairing, training is another area where virtual meetings were found to be inferior by a significant proportion of respondents. While virtual training offers flexibility and convenience, it may not be as effective as face-to-face training in some areas. In this study, 27.5% of respondents reported that virtual MDT training was worse, which suggests that trainees may not be getting the same level of learning experience as they would in a face-to-face setting. However, 18.8% of respondents stated that virtual training was improved, which may reflect learning style preferences. It is also important to note that the quality of virtual training can vary depending on factors such as the technology used and the quality of the internet connection. Responses to the questions surrounding training as a secondary function of the MDT could, however, be confounded by the majority of respondents being from those who have completed clinical training. Future work involving trainee associations such as the Plastic Surgery Trainees Association (PLASTA) and trainee members from the BAD would therefore be useful for a deeper look into these perspectives

By recognising the potential downsides of virtual MDT meetings, steps can be taken to work towards addressing them and optimising the overall benefits. After all, the best improvements are made by identifying areas that are working poorly and taking action to make them better. However, while it is important to acknowledge the potential downsides of virtual MDT meetings, the overall benefits should also be taken into consideration.

It is clear that there are significant benefits such as the ability to more easily share clinical images and radiology, with 37.0% of respondents believing this was better using a digital platform. This does suggest improved communication by harnessing this technology. It perhaps also may aid in decision making by allowing all the MDT to easily and quickly understand a lesion and its extent. Despite team working and interpersonal relationships among respondents

were found to be worse, engagement was actually better. Furthermore, virtual meetings may offer benefits such as increased accessibility and convenience for team members who may be located in different geographic locations or who have other constraints that make in-person meetings difficult.

The study findings are similar to other published reports, in particular in head and neck (Mohamedbhai et al., 2021). The difference is that this study was conducted later, 2 years since the beginning of widespread remote working, therefore it may be that teams and individuals have developed familiarity with remote meeting, and may be more adept. Skin MDTs are often less centralised than head and neck cancer, and thus there may be differences in the actualised benefits in terms of efficiency e.g. having to meet in another unit.

4.4.1 Strengths and limitations

When the response rate of the current study is compared to a systematic review of response rates in patient and health care professional surveys in surgery it can be noted that the response rate for this study was satisfactory for an online web-based survey (Meyer et al., 2022). There was also wide coverage across the UK and from all members of the skin MDT. This study also allows for fair evaluation of the virtual MDT without the additional challenges and biases that came during the various COVID lockdowns.

The low inter-rater reliability of the questionnaire should be noted. This suggests that the questionnaire may need to be revised or clarified to improve agreement between raters. To address this issue, it may be necessary to provide detailed instructions and guidelines for raters, as well as ensure that all raters have similar levels of training and experience. Additionally, it may be helpful to conduct a factor analysis to ensure that the questionnaire items are measuring

what they are intended to measure. This can help to determine whether the questionnaire items are measuring a single construct or multiple constructs. These limitations should be taken into consideration when interpreting the results of the study. Another constraint of this study is the lower response rate from the radiology and pathology specialties, which are particularly relevant to the challenges of image sharing in MDTs. While the data from this population were not available in the current survey, it is acknowledged that there is potential value of gathering data from these specialties in future research to better understand their specific needs and challenges.

Cancer MDTs often hold post-meeting MDT clinics, where core members and allied health professionals are able to review patients together. Data on the post-meeting MDT clinic were not collected in this study. It is important to note that the move to virtual MDT meetings may have had an impact on the post-meeting MDT clinic. The virtual setting may have disrupted the usual attendance of the core members and allied health professionals, resulting in a less comprehensive and less effective clinic. Further research is needed to explore the impact of virtual MDT meetings on the post-meeting MDT clinic and to ensure that this vital aspect of cancer care is maintained in the virtual setting.

While previous research has demonstrated the high costs and low compliance with NICE quoracy standards of SSMDT meetings in the UK, the present study did not specifically examine cost or quoracy (Ali, Dobbs, et al., 2021). However, the virtual nature of these meetings may offer potential improvements in these areas. By eliminating the need for travel and conference room bookings, virtual SSMDT meetings may reduce overhead costs and increase engagement, particularly among senior specialists who may be located at multiple

hospital sites. Further research is needed to formally assess the cost-effectiveness and quoracy of virtual SSMDT meetings compared to traditional in-person meetings.

4.5 Conclusion

Running an efficient MDT can be challenging due to variations in treatment, significant costs, and the opportunity cost for MDT members who have other important roles in the NHS. These challenges can potentially be addressed through the use of virtual MDT meetings, which allow for the participation of team members remotely and can potentially reduce costs and improve efficiency. However, virtual MDT meetings may also have their own challenges, such as the need for reliable technology and the potential for reduced interpersonal interaction. In the Socratic dialogue 'Republic', Plato famously wrote "our need will be the real creator" which was moulded over time into the English proverb "necessity is the mother of invention". The restrictions COVID-19 placed on society has resulted in the adaptation of services throughout healthcare including decision making in cancer care. Whilst social distancing feels like a distant memory to most, it seems that the virtual MDT is one product of the pandemic that is here to stay. Understanding the strengths and weaknesses of virtual MDTs and identifying ways to enhance them is crucial for future practices, especially given the increasing caseloads of skin MDTs and the negative consequences that follow. The results from this study provide a unique chance to reassess the service before COVID-19 and offer insights into opportunities for refining and optimising virtual skin MDTs in the UK after the COVID-19 era.

Chapter Five:

From text to treatment: a systematic review and meta-analysis of rule-based natural language processing algorithms in oncology clinical decision support

5.1 Introduction

Big data research, encompassing the interdisciplinary analysis of high volume, complex and diverse clinical and lifestyle information, stands at the forefront of modern healthcare transformation (Gibson, Dobbs, et al., 2021a). Identified as one of the UK Government's eight Great Technologies, and with the Medical Research Council having invested more than £100 million since 2012 in developing the Health Data Research UK infrastructure, big data has been recognised as a pivotal technological advancement (Department for Business, 2013). AI, a cornerstone of big data, equips physicians with advanced tools to enhance patient care.

In the specialised context of cancer care, rule-based NLP algorithms operate with notable precision, displaying an advanced understanding and control of the specialist language and terminologies therein. Governed by pre-set linguistic rules and domain-specific vocabularies, these rule-based systems can accurately interpret complex medical terms and contexts, so are ideally suited to navigate the complexities inherent in oncological terminology. Tailored application of rule-based algorithms can uncover intricate relationships and deliver medical insights that might otherwise go unnoticed. This confers significant advantage in the multifactorial and interconnected landscape of oncological care and research, enabling the discovery of the nuanced insights that are crucial for exact and personalised patient care.

Unlike ML models, rule-based approaches do not require substantial, annotated training datasets, so lack the logistical and ethical challenges associated with obtaining this data. Moreover, the superior transparency offered by rule-based approaches harmonises with clinical reasoning processes, giving healthcare practitioners the ability to decode, validate, and comprehend the foundational logic behind each piece of extracted information or recommendation. This transparency builds trust and facilitates the integration of these systems

into clinical workflows, positioning them as comprehensible and reliable assistive technologies.

While both rule-based and ML approaches have exhibited notable capabilities in NLP applications within the medical domain, their comparative efficacy can hinge on the specificity and requirements of the task at hand. Rule-based approaches, which often excel in name entity recognition (NER) tasks where the language structure and terminology are precise and consistent, might offer advantages in scenarios like extracting standard medical terms from clinical texts. For instance, Savova et al. showcased that their rule-based system achieved an F1-score (harmonic mean of precision and recall) of 0.715 for exact NER matches and system-level evaluation (Savova et al., 2010). NLP performance metrics are discussed further in Chapter Six. Conversely, ML-based models, particularly deep learning variants, tend to shine in text classification assignments owing to their adeptness in managing high-dimensional data and deciphering complex patterns within text. An exemplary case is a study by Jagannatha and Yu, 2016, where a recurrent neural network model was employed for automatic extraction of medical information from clinical narratives, achieving a best F1-score of 0.82 with a strict evaluation (Jagannatha & Yu, 2016). The choice between rule-based and ML approaches should be dictated by task specifics, availability of labelled data, and the need for model interpretability, underscoring the importance of a task-oriented strategy when deploying NLP solutions in medical contexts.

Despite encouraging advances, the application of NLP in oncology has remained confined to tasks such as case identification, staging, and outcome determination (Yim et al., 2016). Previous systematic reviews of NLP algorithms in this area have often been limited to particular tasks or fields unrelated to medical or surgical oncology, accompanied by variable

risk of bias and inconsistent adherence to reporting protocols (Dreisbach et al., 2019; Koleck et al., 2019; Kreimeyer et al., 2017; Pons et al., 2016; Sheikhalishahi et al., 2019; Young et al., 2019). As the AI landscape continues to unfold, a cohesive roadmap towards practical implementation becomes essential, encompassing robust development, validation studies, and randomised clinical trials that conform to evidentiary benchmarks. This systematic review and meta-analysis endeavour to fill existing voids by furnishing a present-day synopsis of rule-based NLP methodologies in oncology. It includes an in-depth examination of methods and reporting quality, provides a transparent comparison of algorithm performance against human clinicians, and identifies areas of improvement in future research into the development of novel oncological applications of NLP.

5.2 Methods

5.2.1 Search strategy

The identification of studies involved a comprehensive search across various databases and other platforms. Electronic searches were conducted in EMBASE, MEDLINE, CINAHL, Cochrane Library, Web of Science, and Collection of Computer Science Bibliographies. These searches extended from the date of inception to the 13th of April 2020. Details of the search strategies utilised are provided in Table 14.

Table 14: Search strategies.

	Source Criteria	Results
0.	**Medline**	
1.	exp Natural Language Processing/	4156
2.	"natural language processing".ti,ab.	2726
3.	"text mining".ti,ab.	2226
4.	"information extraction".ti,ab.	1124
5.	"unstructured clinical information".ti,ab.	3
6.	"data mining".ti,ab.	8404
7.	1 or 2 or 3 or 4 or 5 or 6	16158
8.	exp Neoplasms/	3305635
9.	cancer.ti,ab.	1655103
10.	oncology.ti,ab.	90326
11.	8 or 9 or 10	3767239
12.	7 and 11	1967
13.	limit 12 to humans	1419
0.	**Embase**	
1.	exp Natural Language Processing/	5088
2.	"natural language processing".ti,ab.	3480
3.	"text mining".ti,ab.	2383
4.	"information extraction".ti,ab.	1128
5.	"unstructured clinical information".ti,ab.	3
6.	"data mining".ti,ab.	10975
7.	1 or 2 or 3 or 4 or 5 or 6	19355
8.	exp Neoplasms/	4381195
9.	cancer.ti,ab.	2333611
10.	oncology.ti,ab.	178305
11.	8 or 9 or 10	4845543
12.	7 and 11	3209
13.	limit 12 to humans	2717
14.	limit 13 to (embase or medline)	1841
0.	**Cochrane Library**	
1.	MeSH descriptor: [Natural Language Processing] explode all trees	8
2.	"natural language processing"	104
3.	"text mining"	18
4.	"information extraction"	13
5.	"unstructured clinical information"	0
6.	"data mining"	108
7.	MeSH descriptor: [Neoplasms] explode all trees	76583

	Source	Criteria	Results
8.		Cancer	172234
9.		Oncology	67054
10.		{OR #1-#6}	224
11.		{OR #7-#9}	200054
12.		#10 AND #11	37
0.	**Web of Science**		
1.		(TS="natural language processing")	6895
2.		(TS="text mining") AND DOCUMENT TYPES: (Article)	5279
3.		(TS="information extraction") AND DOCUMENT TYPES: (Article)	3676
4.		(TS="unstructured clinical information") AND DOCUMENT TYPES: (Article)	2
5.		(TS="data mining") AND DOCUMENT TYPES: (Article)	29733
6.		#5 OR #4 OR #3 OR #2 OR #1	43406
7.		(TS=cancer) AND DOCUMENT TYPES: (Article)	1623335
8.		(TS=oncology) AND DOCUMENT TYPES: (Article)	95028
9.		#8 OR #7	1647811
10.		#9 AND #6	2219
0	**CINAHL**		
1		(MH "Natural Language Processing")	1783
2		TI "natural language processing" OR AB "natural language processing"	1084
3		TI "text mining" OR AB "text mining"	486
4		TI "information extraction" OR AB "information extraction"	284
5		TI "unstructured clinical information" OR AB "unstructured clinical information"	2
6		TI "data mining" OR AB "data mining"	1625
7		S1 OR S2 OR S3 OR S4 OR S5 OR S6	4,314
8		(MH "Neoplasms+")	(565,604)
9		TI cancer OR AB cancer	396,193
10		TI oncology OR AB oncology	42,789
9		TI cancer OR AB cancer	396,193
10		TI oncology OR AB oncology	42,789
11		S8 OR S9 OR S10	693,374
12		S7 AND S11	441
0	**The Collection of Computer Science Bibliographies**		
1		+"natural language processing" +cancer	149

5.2.2 Protocol and registration

The systematic review protocol was formulated in accordance with the Preferred Reporting for Items for Systematic Reviews and Meta-Analyses-Protocols (PRISMA-P) and was registered with PROSPERO (CRD42020180676).

5.2.3 Review focus and study selection criteria

The focus of the review was on NLP algorithms applied to specific clinical problems in cancer, defined as situations where a clinician's intervention is essential for improving or managing patient health e.g. clinical decision support during a cancer multidisciplinary team meeting. The context of the review emphasised the potential of NLP to revolutionise service delivery within healthcare, particularly in the National Health Service, by providing robust clinical decision support tools. The included population comprised human patients, and the input data consisted of real human information, either retained during routine clinical care for research purposes or specially generated for the study, reflecting real-world applications in expert clinical practice. Studies were included if they utilised an NLP algorithm to address a clinical problem in cancer, with comparators including human clinical healthcare professionals. The types of studies considered included peer-reviewed full-text journal articles, while excluding reviews, conference proceedings, books, and theses.

5.2.4 Screening, data extraction and quality assessment

Screening and data extraction were performed using the Covidence screening and data extraction tool (Veritas Health Innovation, Melbourne, Australia), guided by the Cochrane handbook for systematic reviews of interventions. The extraction process involved two authors working independently, with disagreements resolved through consultation with a third author.

Reporting quality assessment was gauged against the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.

5.2.5 Statistical analysis

Extraction and back-transformation of raw data from each paper into binary outcome parameters i.e. true positive (TP), false positive (FP), false negative (FN), true negative (TN) was undertaken in order to allow meta-analysis of diagnostic test accuracy. Meta-analysis was performed in Stata 17 (Stata Corp LLC, Texas, USA) using these variables, with exclusion criteria detailed for studies with insufficient information, using a random-effects model. For the studies that were eligible for meta-analysis, sensitivity and specificity were assessed through forest plots. A Hierarchical Summary Receiver Operating Characteristic (HSROC) curve was generated accounting for study sample size. Deek's effective sample size weighted regression test was applied to assess publication bias. Heterogeneity statistics were evaluated using the I^2 statistic. In order to synthesise and evaluate the diagnostic accuracy derived from the various studies, a unified hierarchical model was developed and applied for the meta-analysis of diagnostic accuracy studies. The hierarchical structure of the model was built upon two distinct statistical distribution levels. On the initial, lower level, the model contemplated the cell counts that configure the 2x2 contingency tables - namely true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) - utilising binomial distributions to proficiently account for the within-study variability. Moving to the upper level, the model encapsulated the between-study variability, aggregating disparities across different studies. Further, summary receiver operating characteristic (ROC) curves were rendered to effectively illustrate the NLP performance through visual means. Univariate meta-regressions were performed to explore temporal trends in specificity, sensitivity, accuracy, and F1 score. The year of publication served as the independent variable. To ensure robust analysis, the Freeman-

Tukey double arcsine transformation was employed on these proportion-based measures, enhancing variance stability. (Cancer Research UK, 2024-d)

5.3 Results

5.3.1 Study selection

Following the de-duplication process, a total of 3,223 titles and abstracts were screened. Subsequently, 271 articles were subjected to a full-text review, with 89 studies meeting the criteria for inclusion in the systematic review (Figure 15). Publication years ranged from 1993 to 2020. A variety of cancer types were studied, with the most common being breast, colorectal, and lung (Figure 16).

Figure 15: PRISMA flowchart.

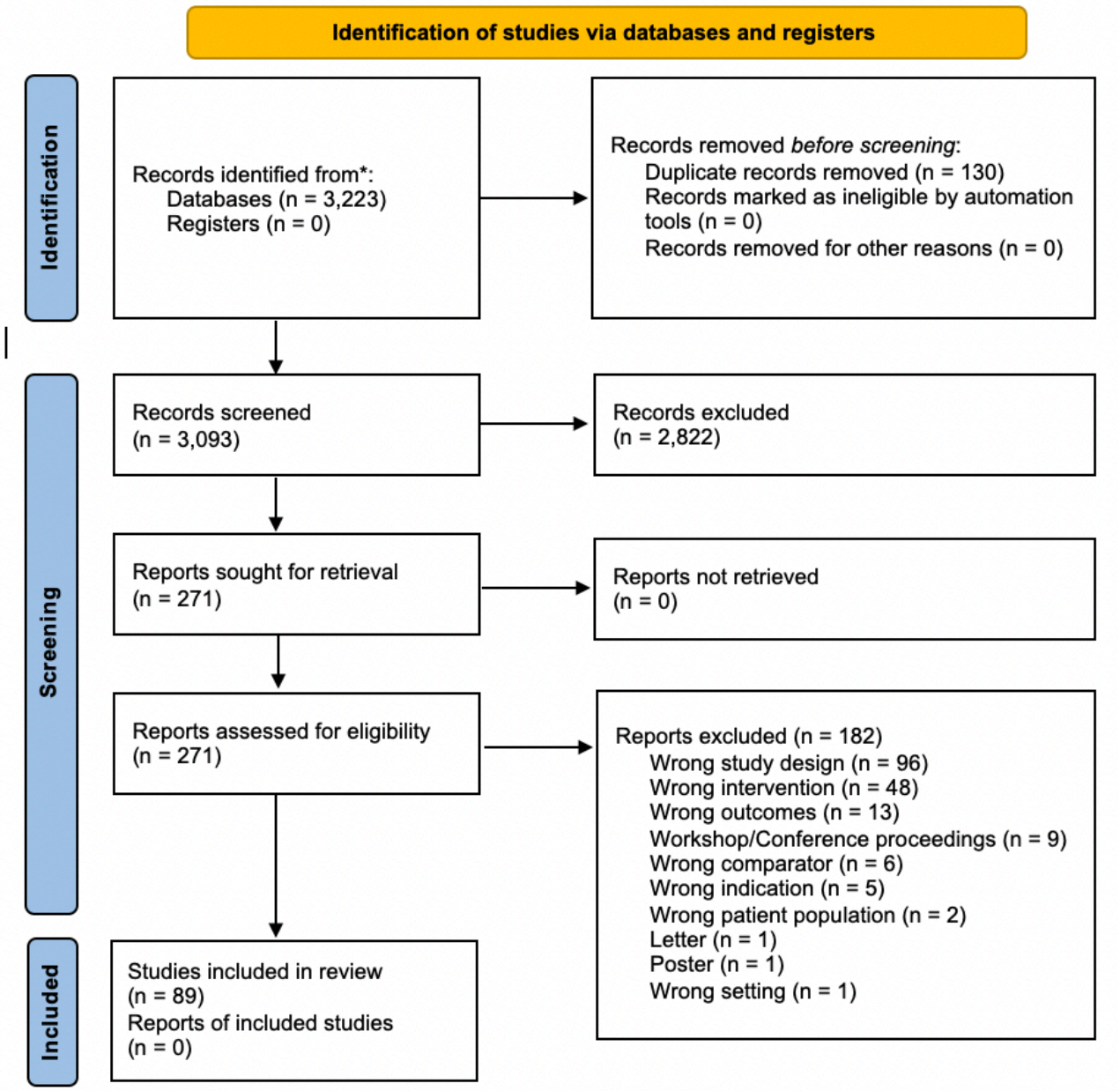
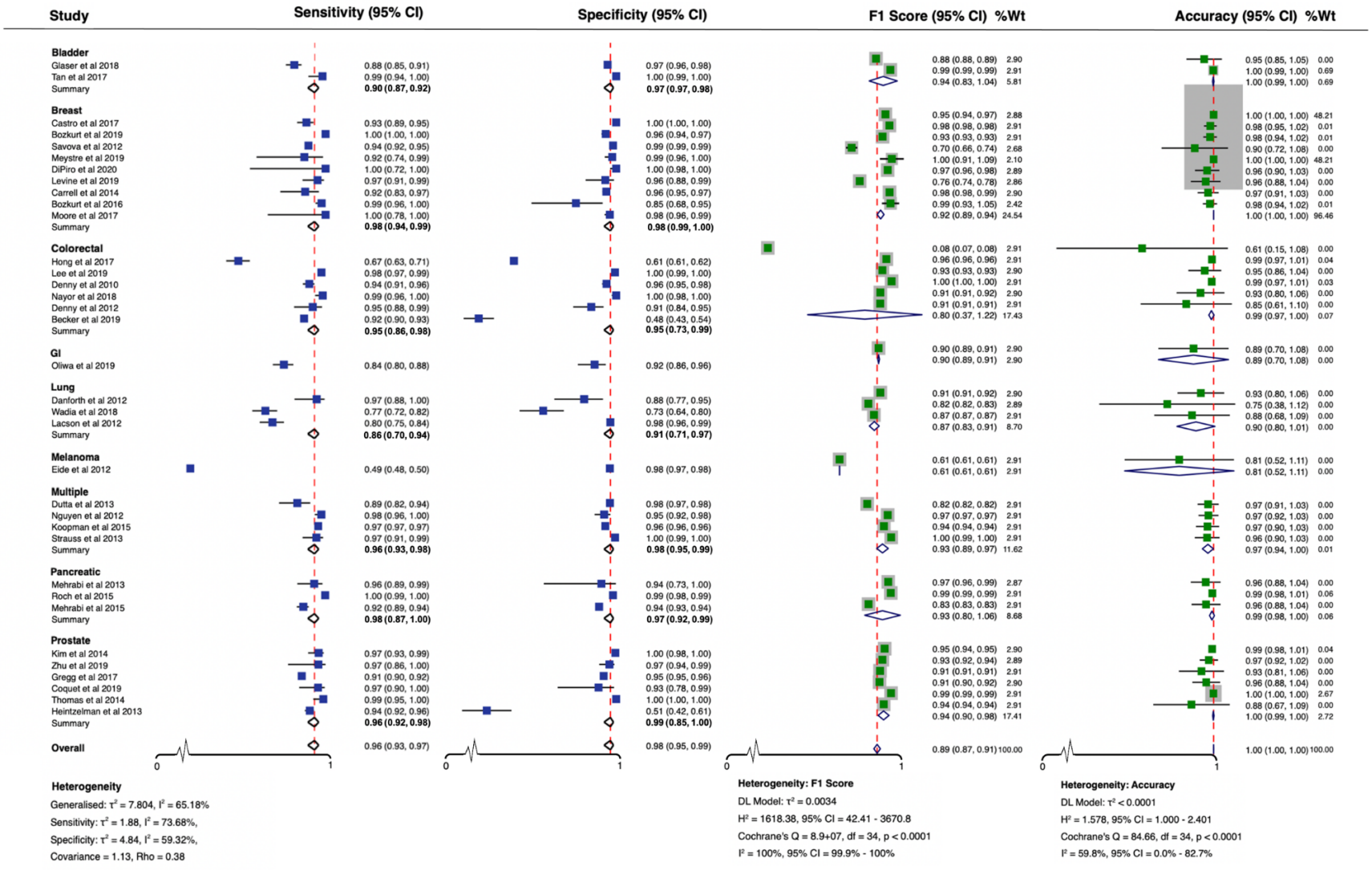


Figure 16: Bivariate random effects forest plots of pooled sensitivity, specificity, F1 score, and accuracy stratified by pathology.



5.3.2 Study characteristics

In terms of text-based classification, the studies predominantly focused on information extraction, text classification, or a combination of both. Task-based classifications were diverse, covering tasks such as case identification, clinical decision support, and quality assurance. A comprehensive summary of text- and task-based classifications can be found in Table 13.

Table 15: Summary of text and task-based classifications.

Study	Text based classification	Task based classification
Kim et al 2014	Information Extraction	Clinical Decision Support
Hong et al 2017	Information Extraction	Clinical Decision Support, Risk Assessment
Lee et al 2019	Information Extraction	Clinical Decision Support
Mehrabi et al 2013	Information Extraction	Clinical Decision Support, Disease Surveillance
Gawron et al 2014	Information Extraction	Quality Assurance
Castro et al 2017	Information Extraction	Quality Assurance
Bozkurt et al 2019	Information Extraction	Clinical Decision Support
Dutta et al 2013	Information Extraction	Clinical Decision Support, Radiology Report Analysis
Savova et al 2012	Information Extraction	Drug Discovery
Sippo et al 2013	Information Extraction	Clinical Decision Support
Lee et al 2018	Information Extraction	Clinical Decision Support
Glaser et al 2018	Information Extraction	Clinical Decision Support
Danforth et al 2012	Information Extraction	Clinical Decision Support
Roch et al 2015	Information Extraction	Clinical Decision Support
Percha et al 2012	Text Classification	Clinical Decision Support
Liu et al 2019	Information Extraction, Text Classification	Clinical Decision Support
Koopman et al 2018	Text Classification	Clinical Decision Support
Beyer et al 2017	Information Extraction	Clinical Decision Support
Meystre et al 2019	Text Classification	Clinical Trial Matching
Coden et al 2009	Information Extraction	Clinical Decision Support
Zhu et al 2019	Information Extraction	Clinical Decision Support
Odisho et al 2019	Information Extraction	Clinical Decision Support
Gregg et al 2017	Information Extraction	Clinical Decision Support, Risk Assessment
Nguyen et al 2012	Text Classification	Cancer Registry Notification
Wagholikar et al 2012	Information Extraction	Clinical Decision Support
Imler et al 2014	Information Extraction	Clinical Decision Support
DiPiro et al 2020	Text Classification	Case Identification

Study	Text based classification	Task based classification
Giess et al 2017	Text Classification	Case Identification
Wadia et al 2018	Text Classification	Clinical Decision Support, Risk Assessment
Coquet et al 2019	Information Extraction	Quality Assurance
Patel et al 2017	Information Extraction	Clinical Decision Support
Harkema et al 2011	Information Extraction	Quality Assurance
Soysal et al 2019	Information Extraction	Case Identification
Tan et al 2017	Information Extraction	Case Identification
Schroeck et al 2017	Information Extraction	Case Identification
Walker et al 2019	Information Extraction	Clinical Decision Support
Karwa et al 2019	Information Extraction	Clinical Decision Support, Quality Assurance
Poort et al 2020	Information Extraction	Clinical Decision Support
Breitenstein et al 2018	Information Extraction	Clinical Decision Support, Risk Assessment
Malke et al 2019	Information Extraction	Case Identification
Baldwin K. B. 2008	Information Extraction	Clinical Decision Support
Loda et al 2019	Information Extraction	Real-World Evidence Extraction
Thomas et al 2014	Information Extraction	Case Identification
Denny et al 2010	Text Classification	Clinical Decision Support
Khor et al 2019	Information Extraction	Case Identification
Miao et al 2018	Information Extraction	Case Identification
Lacson et al 2012	Text Classification	Clinical Decision Support
Warner et al 2016	Text Classification	Case Identification
Piotrkowicz et al 2019	Text Classification	Clinical Decision Support
Wagholikar et al 2013	Text Classification	Clinical Decision Support
Mehrabi et al 2015	Text Classification	Clinical Decision Support
Strauss et al 2013	Text Classification	Case Identification
Ni et al 2015	Text Classification	Clinical Decision Support
Ping et al 2013	Information Extraction, Text Classification	Clinical Decision Support
Bozkurt et al 2019	Text Classification	Quality Assurance
Aramaki et al 2019	Information Extraction	Clinical Decision Support
Levine et al 2019	Information Extraction	Clinical Decision Support

Study	Text based classification	Task based classification
Heintzelman et al 2013	Information Extraction	Clinical Decision Support
Zingmond et al 1993	Text Classification	Clinical Decision Support, Quality Assurance
Imler et al 2015	Text Classification	Clinical Decision Support, Quality Assurance
Nayor et al 2018	Text Classification	Clinical Decision Support, Quality Assurance
Imler et al 2013	Text Classification	Clinical Decision Support, Quality Assurance
Banerjee et al 2019	Information Extraction	Clinical Decision Support
Raju et al 2015	Text Classification	Clinical Decision Support, Quality Assurance
Al-Haddad et al 2010	Information Extraction	Case Identification
Denny et al 2012	Information Extraction	Case Identification
Nobel et al 2020	Text Classification	Case Identification
Matthias et al 2019	Information Extraction	Clinical Decision Support, Quality Assurance
Lindvall et al 2019	Text Classification	Clinical Decision Support, Quality Assurance
Brizzi et al 2020	Text Classification	Clinical Decision Support, Quality Assurance
Oliwa et al 2019	Text Classification, Named Entity Recognition (NER), Relation-extraction heuristics	Case Identification
Han et al 2019	Information Extraction	Case Identification, Clinical Decision Support, Quality Assurance
Giess et al 2019	Information Extraction	Case Identification
Mehrotra et al 2018	Text Classification	Clinical Decision Support, Quality Assurance
Lott et al 2018	Information Extraction	Case Identification, Clinical Decision Support, Quality Assurance
Carrell et al 2015	Text Classification	Case Identification
Chen et al 2020	Text Classification	Clinical Decision Support
Gould et al 2015	Text Classification	Case Identification
Maehara et al 2014	Text Classification	Clinical Decision Support, Quality Assurance
Patel et al 2018	Text Classification	Clinical Decision Support, Quality Assurance
Liu et al 2019	Information Extraction	Case Identification, Clinical Decision Support
Coopey et al 2012	Information Extraction	Case Identification
Goldbraich et al 2015	Text Classification	Clinical Decision Support, Quality Assurance
Ashish et al 2014	Information Extraction	Case Identification

Study	Text based classification	Task based classification
Hripcsak et al 2002	Information Extraction	Case Identification
Bozkurt et al 2016	Information Extraction	Clinical Decision Support
Moore et al 2017	Information Extraction	Case Identification
Eide et al 2012	Information Extraction	Case Identification
Udelsman et al 2020	Information Extraction	Clinical Decision Support, Quality Assurance

The studies employed various designs, with 83 being retrospective cohort studies, 3 being prospective cohort studies, and 3 being case-control studies (Appendix 2). The language or corpora used in the overwhelming majority of studies was English, with two studies in German and a single study in each of Italian, Dutch and Chinese. Details about language usage in each study are outlined in Appendix 2. Regarding the NLP systems, the studies featured a variety of methods including commercially available systems, internally developed rule-based methods, and hybrid methods including ML techniques. Detailed information about the NLP systems used, including their training and validation sizes, can be found in Appendix 2. The total dataset size varied significantly between the studies, from a minimum of 19 reports to a maximum of 6,100,000 reports. Often, dataset sizes were reported in terms of 'number of patients', where one patient might have multiple reports. Various types of input data were used by the NLP systems across the studies, predominantly pathology reports, free-text clinical notes, and EHRs. The most used sources of input were pathology reports, followed closely by EHRs (Appendix 2).

5.3.3 Quality assessment

Quality assessment of studies using the TRIPOD criteria revealed a median score of 19 (range 0-26), suggesting that there is some disparity in reporting quality between different studies (Table 16). The percentage adherence to the TRIPOD criteria was calculated for each study to aid interpretation. This percentage represents the proportion of items in the TRIPOD criteria that were adequately addressed by the study. There was variability in adherence among studies, with some achieving high adherence percentages (e.g., 92.0%) and others scoring lower (e.g., 48.0%). Notably, the introduction and discussion sections of the TRIPOD criteria were often well addressed, whereas the methods and results sections posed challenges for many studies.

Table 16: Individual TRIPOD scores and breakdown for each study.

Study	TRIPOD																									TRIPOD score	Adherence
	Introduction				Methods											Results						Discussion						
	1	2	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8	10a	10b	13a	13b	14a	15a	15b	16	18	19b	20	21	22		
Kim et al 2014	0	1	1	1	1	1	0	0	1	1	1	2	1	1	1	1	0	2	0	1	1	1	1	1	1	0	18	75.0%
Hong et al 2017	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	0	1	1	1	1	0	2	22	92.0%
Lee et al 2019	0	1	1	1	1	1	1	0	2	1	1	2	1	2	2	1	1	1	2	2	1	1	1	1	1	1	18	90.0%
Mehrabi et al 2013	0	1	1	1	1	0	1	0	1	0	1	2	1	1	1	1	1	1	1	1	1	0	0	1	1	1	19	76.0%
Gawron et al 2014	1	1	1	1	1	1	1	1	1	0	0	2	1	1	0	1	1	1	0	0	0	1	1	1	1	1	19	76.0%
Castro et al 2017	0	1	1	1	1	1	1	1	2	1	1	2	1	1	1	1	0	0	0	1	1	1	1	1	1	1	20	83.0%
Bozkurt et al 2019	0	1	1	1	1	0	0	0	1	1	1	2	1	1	1	1	0	1	1	1	1	1	1	1	0	0	18	72.0%
Dutta et al 2013	0	1	1	1	1	0	1	1	2	1	1	2	1	0	1	0	0	1	0	0	1	1	1	1	0	1	16	67.0%
Savova et al 2012	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	1	1	1	0	1	1	1	1	1	1	1	22	88.0%
Sippo et al 2013	0	1	1	0	1	1	1	1	1	0	1	2	1	1	1	1	0	1	0	0	1	1	1	1	0	0	17	68.0%
Lee et al 2018	0	1	1	0	1	1	1	0	1	1	1	2	1	1	1	1	0	1	1	0	1	1	1	1	1	1	20	80.0%
Glaser et al 2018	0	1	1	1	1	1	1	0	1	1	1	2	1	1	1	1	1	1	0	0	1	1	1	1	0	1	20	80.0%
Danforth et al 2012	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	1	1	0	0	1	1	1	1	0	0	18	72.0%

Study	TRIPOD																										TRIPOD score	Adherence
	Introduction				Methods											Results						Discussion						
	1	2	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8	10a	10b	13a	13b	14a	15a	15b	16	18	19b	20	21	22		
Roch et al 2015	0	1	1	1	1	1	1	0	1	1	1	2	1	1	1	0	0	1	0	1	1	1	0	1	0	1	18	72.0%
Percha et al 2012	0	0	1	1	1	0	0	0	1	1	1	2	1	1	1	0	0	1	0	0	0	1	1	1	0	1	14	56.0%
Liu et al 2019	0	1	1	1	1	1	0	1	1	1	1	2	1	1	1	0	0	1	0	0	1	1	1	1	0	0	17	68.0%
Koopman et al 2018	0	1	1	1	1	1	1	2	1	1	1	2	0	1	1	2	2	1	0	1	1	0	1	1	0	0	18	82.0%
Beyer et al 2017	0	1	1	1	1	1	1	1	1	0	1	2	1	0	0	0	1	0	0	0	1	1	1	1	0	0	15	60.0%
Meystre et al 2019	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	0	0	1	0	1	1	1	1	1	0	1	20	80.0%
Coden et al 2009	0	1	1	1	1	0	0	0	1	1	1	2	1	1	1	0	0	0	0	1	1	0	1	1	0	0	14	56.0%
Zhu et al 2019	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	0	1	0	1	1	1	1	1	1	1	1	22	88.0%
Odisho et al 2019	0	1	1	1	1	1	1	0	1	0	1	2	1	1	1	0	0	1	0	1	1	1	1	1	0	1	18	72.0%
Gregg et al 2017	0	1	1	1	1	1	1	0	1	0	1	2	1	1	1	0	0	1	1	1	1	1	1	1	0	1	19	76.0%
Nguyen et al 2012	0	1	1	1	1	0	0	0	1	0	0	2	0	1	1	0	0	0	1	1	1	0	1	1	0	0	12	48.0%
Wagholikar et al 2012	0	0	1	1	0	0	0	0	1	1	1	2	1	1	1	0	0	1	0	1	0	1	1	1	0	1	14	56.0%
Imler et al 2014	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	0	1	1	1	1	1	1	1	0	1	22	88.0%
DiPiro et al 2020	0	1	0	1	1	1	1	1	1	1	0	2	1	1	1	1	0	1	0	0	1	1	1	1	0	0	17	68.0%

Study	TRIPOD																										TRIPOD score	Adherence
	Introduction				Methods											Results						Discussion						
	1	2	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8	10a	10b	13a	13b	14a	15a	15b	16	18	19b	20	21	22		
Giess et al 2017	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	0	1	1	1	1	1	0	0	21	84.0%
Wadia et al 2018	0	1	1	1	1	1	0	0	1	1	1	2	0	1	1	1	1	1	1	1	1	1	1	1	0	1	20	80.0%
Coquet et al 2019	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	23	92.0%
Patel et al 2017	0	1	1	1	1	1	1	0	1	0	1	2	0	1	1	0	1	1	0	0	1	1	1	1	0	1	17	68.0%
Harkema et al 2011	1	1	1	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	24	100.0%
Soysal et al 2019	1	1	1	1	1	1	1	0	1	1	1	2	0	1	0	1	0	0	1	1	1	0	1	1	2	1	20	83.0%
Tan et al 2017	1	1	1	1	1	1	1	0	1	0	1	2	1	1	1	0	0	0	1	0	1	1	1	1	0	1	18	72.0%
Schroeck et al 2017	1	1	0	1	1	1	1	1	1	0	0	2	1	1	1	0	0	1	2	2	1	1	1	1	1	1	18	78.0%
Walker et al 2019	1	1	1	1	1	1	1	2	1	1	1	2	1	1	1	0	0	0	0	0	1	1	0	0	0	1	16	67.0%
Karwa et al 2019	1	1	1	1	1	0	1	1	1	1	1	2	1	1	1	1	1	1	1	1	0	0	0	1	1	0	20	80.0%
Poort et al 2020	0	1	1	0	1	1	1	1	1	0	1	2	1	1	1	0	1	1	0	1	0	1	1	1	0	1	18	72.0%
Breitenstein et al 2018	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	0	1	0	1	0	0	1	1	1	0	1	17	68.0%
Malke et al 2019	0	1	1	1	1	1	1	2	1	1	1	2	1	1	1	0	0	1	0	1	1	1	1	1	0	1	21	88.0%
Baldwin K. B. 2008	0	0	1	1	1	1	1	1	1	0	1	2	1	1	1	1	1	1	0	0	1	0	1	1	0	0	17	68.0%

Study	TRIPOD																										TRIPOD score	Adherence
	Introduction				Methods											Results						Discussion						
	1	2	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8	10a	10b	13a	13b	14a	15a	15b	16	18	19b	20	21	22		
Loda et al 2019	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	24	96.0%
Thomas et al 2014	1	1	1	1	1	1	1	0	1	1	1	2	1	1	1	0	0	1	0	1	1	1	1	1	0	1	20	80.0%
Denny et al 2010	0	1	1	1	1	1	0	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	0	1	22	88.0%
Khor et al 2019	0	1	1	1	1	0	1	0	1	0	1	2	1	1	1	0	0	1	1	1	1	1	1	1	0	1	18	72.0%
Miao et al 2018	0	1	1	1	1	0	1	0	1	0	1	2	1	1	1	0	0	1	1	1	1	0	1	1	0	1	17	68.0%
Lacson et al 2012	0	0	1	0	1	1	1	1	1	0	1	2	1	1	1	1	1	1	1	1	1	1	1	1	0	1	20	80.0%
Warner et al 2016	0	1	1	1	1	1	1	0	1	0	1	2	1	1	1	0	0	1	1	1	1	1	1	1	1	1	20	80.0%
Piotrkowicz et al 2019	0	1	1	1	1	1	1	0	1	0	1	2	0	1	1	0	0	0	1	1	1	1	1	1	1	1	18	72.0%
Wagholikar et al 2013	1	1	1	1	1	1	1	2	1	1	1	2	1	1	1	1	1	1	1	1	0	1	1	1	0	0	21	88.0%
Mehrabi et al 2015	0	1	1	1	1	1	1	0	1	1	0	2	0	0	1	0	0	0	1	1	1	0	1	1	0	0	14	56.0%
Strauss et al 2013	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	0	0	1	1	1	1	1	1	1	1	20	80.0%
Ni et al 2015	0	1	1	1	1	1	1	1	1	0	0	2	1	0	1	0	1	1	0	1	1	1	1	1	1	1	19	76.0%
Ping et al 2013	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	1	0	0	1	1	0	1	1	0	0	17	68.0%
Bozkurt et al 2019	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	1	1	1	0	1	1	1	1	1	1	1	22	88.0%

Study	TRIPOD																										TRIPOD score	Adherence
	Introduction				Methods											Results						Discussion						
	1	2	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8	10a	10b	13a	13b	14a	15a	15b	16	18	19b	20	21	22		
Aramaki et al 2019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0%
Levine et al 2019	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	1	1	1	0	1	1	1	1	1	1	21	84.0%
Heintzelman et al 2013	0	1	1	1	1	1	1	0	1	1	1	2	1	1	1	0	1	1	0	1	1	1	1	1	1	1	21	84.0%
Zingmond et al 1993	0	1	1	1	1	1	1	0	1	0	1	2	1	1	1	1	0	1	0	1	0	0	1	1	0	0	18	72.0%
Imler et al 2015	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	0	1	1	1	1	1	1	1	0	0	21	84.0%
Nayor et al 2018	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	0	1	1	1	1	1	1	1	0	1	22	88.0%
Imler et al 2013	0	0	1	1	1	1	1	1	1	0	1	2	1	1	1	1	0	1	1	1	1	1	1	1	0	1	20	80.0%
Banerjee et al 2019	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	1	0	1	1	1	1	1	1	1	1	1	22	88.0%
Raju et al 2015	0	0	1	1	1	1	1	1	1	0	1	2	1	1	1	0	0	1	0	1	1	1	1	1	0	1	18	72.0%
Al-Haddad et al 2010	1	1	1	1	1	1	1	0	1	0	1	2	1	1	1	1	0	1	1	1	1	1	1	1	0	1	21	84.0%
Denny et al 2012	0	1	1	1	1	0	1	1	1	0	1	2	1	1	1	1	1	1	1	1	1	1	1	1	0	0	20	80.0%
Nobel et al 2020	0	1	1	1	1	0	0	1	1	0	1	2	1	0	1	0	0	0	0	1	1	1	1	1	1	0	15	60.0%
Matthias et al 2019	0	1	1	1	1	0	1	1	1	0	1	2	1	1	1	1	1	1	0	1	1	0	1	1	1	1	20	80.0%
Lindvall et al 2019	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	1	0	0	1	1	1	1	1	0	1	19	76.0%

Study	TRIPOD																										TRIPOD score	Adherence
	Introduction				Methods											Results						Discussion						
	1	2	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8	10a	10b	13a	13b	14a	15a	15b	16	18	19b	20	21	22		
Brizzi et al 2020	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	1	1	1	1	1	1	1	1	0	1	21	84.0%
Oliwa et al 2019	0	1	1	1	1	1	0	0	1	0	1	2	1	1	1	1	0	1	0	0	1	1	1	1	1	1	18	72.0%
Han et al 2019	0	1	1	0	1	1	1	1	1	0	1	2	1	0	0	1	1	1	0	0	1	1	1	1	0	1	17	68.0%
Giess et al 2019	0	1	1	1	1	1	1	1	1	0	0	2	1	0	0	0	1	1	0	0	1	1	1	1	0	1	16	64.0%
Mehrotra et al 2018	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	1	1	0	1	1	1	1	1	1	1	21	84.0%
Lott et al 2018	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	0	1	1	1	1	1	1	1	1	1	21	84.0%
Carrell et al 2015	0	0	1	1	1	1	1	1	1	0	1	2	1	1	1	0	1	1	1	1	1	1	1	1	0	1	20	80.0%
Chen et al 2020	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	1	1	1	1	1	1	1	1	1	1	22	88.0%
Gould et al 2015	0	1	1	1	1	1	1	1	1	0	1	2	1	1	0	0	1	1	0	1	1	1	1	1	0	0	18	72.0%
Maehara et al 2014	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	1	0	1	1	1	1	1	1	1	0	1	21	84.0%
Patel et al 2018	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	0	1	1	1	0	1	0	1	1	0	0	19	76.0%
Liu et al 2019	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	0	0	1	0	1	1	1	1	1	1	1	21	84.0%
Coopey et al 2012	0	1	1	1	1	1	1	1	1	0	1	2	1	1	0	0	0	1	0	1	1	1	1	1	0	0	17	68.0%
Goldbraich et al 2015	0	1	1	0	1	1	1	1	1	0	0	2	1	0	1	1	1	1	0	0	1	1	1	1	0	0	16	64.0%

Study	TRIPOD																										TRIPOD score	Adherence
	Introduction				Methods											Results						Discussion						
	1	2	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8	10a	10b	13a	13b	14a	15a	15b	16	18	19b	20	21	22		
Ashish et al 2014	0	0	1	0	1	0	0	0	1	0	0	2	1	1	1	0	0	1	1	1	1	0	1	1	0	1	13	52.0%
Hripcsak et al 2002	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	0	1	0	1	1	1	1	1	0	0	18	72.0%
Bozkurt et al 2016	0	1	1	1	1	0	0	0	1	0	0	2	1	1	1	1	0	1	0	0	1	1	1	1	0	0	14	56.0%
Moore et al 2017	0	1	1	1	1	1	1	1	1	0	1	2	1	1	1	0	0	1	1	1	1	1	1	1	0	0	19	76.0%
Eide et al 2012	1	1	1	0	1	1	1	1	0	2	1	0	1	1	0	1	0	1	2	2	1	1	1	1	0	1	17	74.0%
Udelsman et al 2020	0	1	1	1	1	1	1	1	1	2	1	2	1	2	1	0	0	1	2	2	1	1	0	1	1	1	17	81%

5.3.4 Meta-analysis of NLP algorithm performance

The analysis of heterogeneity statistics for the forest plot revealed various insights into the differences across the included studies. The overall heterogeneity was captured by a covariance of 1.134 and a correlation coefficient of 0.376. In the generalised assessment, tau-squared (τ^2) was 7.804, and the I^2 statistic was 65.2%, indicating a substantial level of inconsistency across studies. Sensitivity analysis produced a τ^2 of 1.878 and an I^2 of 73.7%, indicating a more significant heterogeneity compared to the generalised analysis. Specificity analysis revealed a τ^2 of 4.84 and an I^2 of 59.3%, showing a moderate inconsistency between studies. F1 Score analysis returned a τ^2 of 0.0034 and an I^2 of 100.0%, pointing to maximum heterogeneity. Accuracy analysis provided a τ^2 of < 0.0001 and an I^2 of 59.8%, signifying moderate inconsistency between studies. These observed heterogeneity statistics underline the diversity among the studies in different aspects, including generalised, sensitivity, and specificity analyses. The significant differences in tau-squared and I^2 statistics among the different analyses emphasise the need for cautious interpretation of the findings.

In an exhaustive evaluation of NLP algorithms for diagnostic accuracy in oncology, the data revealed a hierarchy of performance across tumour types (Table 17). Algorithms designed for breast tumours emerged as the most reliable, with an unparalleled sensitivity of 0.98 (95% CI: 0.94–0.99) and specificity of 0.98 (95% CI: 0.99–1.00). Pancreatic tumours followed closely with a sensitivity of 0.98 (95% CI: 0.87–1.00) and specificity of 0.97 (95% CI: 0.92–0.99). Notably, algorithms for colorectal tumours exhibited a high sensitivity of 0.95 (95% CI: 0.86–0.98) but a more variable specificity, ranging from 0.73 to 0.99. For lung tumours, the profile presented sensitivities and specificities ranging from 0.77 to 0.97 and 0.71 to 0.97, respectively. Moreover, malignant melanoma stood as an outlier, with a markedly low sensitivity of 0.49

(95% CI: 0.48–0.50), yet maintaining a high specificity of 0.98 (95% CI: 0.97–0.98). Multiple and prostate tumours also demonstrated high accuracies, albeit marginally lower than those for breast and pancreatic pathology, with respective sensitivities and specificities of 0.96 (95% CI: 0.93–0.98) and 0.98 (95% CI: 0.95–0.99) for multiple tumours, and 0.96 (95% CI: 0.92–0.98) and 0.99 (95% CI: 0.85–1.00) for prostate tumours. The data for bladder and gastrointestinal tumours were also notable, with bladder tumours showing a sensitivity of 0.90 (95% CI: 0.87–0.92) and specificity of 0.97 (95% CI: 0.97–0.98), and gastrointestinal tumours presenting a sensitivity of 0.84 (95% CI: 0.80–0.88) and specificity of 0.92 (95% CI: 0.86–0.96.

Table 17: Summary data for forest plot stratified by pathology.

Tumour type/study	Sensitivity			Specificity		
	Estimate	(95% CI)		Estimate	(95% CI)	
Bladder						
Glaser et al 2018	0.88	0.85	0.91	0.97	0.96	0.98
Tan et al 2017	0.99	0.94	1.00	1.00	0.99	1.00
Summary	0.90	0.87	0.92	0.97	0.97	0.98
Breast						
Castro et al 2017	0.93	0.89	0.95	1.00	1.00	1.00
Bozkurt et al 2019	1.00	1.00	1.00	0.96	0.95	0.97
Savova et al 2012	0.94	0.92	0.95	0.99	0.99	0.99
Meystre et al 2019	0.92	0.74	0.99	0.99	0.96	1.00
DiPiro et al 2020	1.00	0.72	1.00	1.00	0.98	1.00
Levine et al 2019	0.97	0.91	0.99	0.96	0.89	0.99
Carrell et al 2014	0.92	0.83	0.98	0.96	0.95	0.97
Bozkurt et al 2016	0.99	0.96	1.00	0.85	0.68	0.95
Moore et al 2017	1.00	0.78	1.00	0.98	0.96	0.99
Summary	0.98	0.94	0.99	0.98	0.99	1.00
Colorectal						
Hong et al 2017	0.67	0.63	0.72	0.62	0.61	0.62
Lee et al 2019	0.98	0.97	0.99	1.00	0.99	1.00
Denny et al 2010	0.94	0.91	0.96	0.96	0.95	0.98
Nayor et al 2018	0.99	0.96	1.00	1.00	0.98	1.00
Denny et al 2012	0.95	0.88	0.99	0.91	0.84	0.95
Becker et al 2019	0.92	0.91	0.93	0.48	0.43	0.54
Summary	0.95	0.86	0.98	0.95	0.73	0.99
Gastrointestinal						
Oliwa et al 2019	0.84	0.80	0.88	0.92	0.86	0.96
Lung						
Danforth et al 2012	0.97	0.88	1.00	0.88	0.77	0.95
Wadia et al 2018	0.77	0.72	0.82	0.73	0.64	0.80
Lacson et al 2012	0.80	0.75	0.84	0.98	0.96	0.99
Summary	0.86	0.70	0.94	0.91	0.71	0.97
Malignant melanoma						
Eide et al 2012	0.49	0.48	0.50	0.98	0.97	0.98
Multiple						
Dutta et al 2013	0.89	0.83	0.94	0.98	0.97	0.98
Nguyen et al 2012	0.98	0.96	1.00	0.96	0.92	0.98
Koopman et al 2015	0.97	0.97	0.97	0.96	0.96	0.96
Strauss et al 2013	0.97	0.91	0.99	1.00	0.99	1.00
Summary	0.96	0.93	0.98	0.98	0.95	0.99

Tumour type/study	Sensitivity			Specificity		
	Estimate	(95% CI)		Estimate	(95% CI)	
Pancreatic						
Mehrabi et al 2013	0.96	0.89	0.99	0.94	0.73	1.00
Roch et al 2015	1.00	0.99	1.00	0.99	0.98	0.99
Mehrabi et al 2015	0.92	0.89	0.94	0.94	0.93	0.94
Summary	0.98	0.87	1.00	0.97	0.92	0.99
Prostate						
Kim et al 2014	0.97	0.93	0.99	1.00	0.98	1.00
Zhu et al 2019	0.97	0.86	1.00	0.98	0.94	0.99
Gregg et al 2017	0.91	0.90	0.92	0.95	0.95	0.96
Coquet et al 2019	0.97	0.90	1.00	0.93	0.78	0.99
Thomas et al 2014	0.99	0.95	1.00	1.00	1.00	1.00
Heintzelman et al 2013	0.94	0.92	0.96	0.51	0.42	0.61
Summary	0.96	0.92	0.98	0.99	0.85	1.00
Overall	0.96	0.93	0.97	0.98	0.95	0.99

A unified hierarchical model was developed for meta-analysis of diagnostic accuracy of studies and used to plot summary ROC curves for NLP performance (Figure 17). The hierarchical model involves statistical distributions at two different levels. At the lower level, it models the cell counts that form the 2x2 contingency tables (TP, TN, FP, and FN) by using binomial distributions. This accounts for the within-study variability. At the higher level, it models the between-study variability across studies. The hierarchical summary ROC (HSROC) figure provides estimates of average sensitivity and specificity across included studies with a 95% confidence region of the summary operating point and the 95% prediction region, which represents the confidence region for forecasts of sensitivity and specificity in a future study (Table 18).

Figure 17: HSROC curve with prediction region and accounting for study sample size.

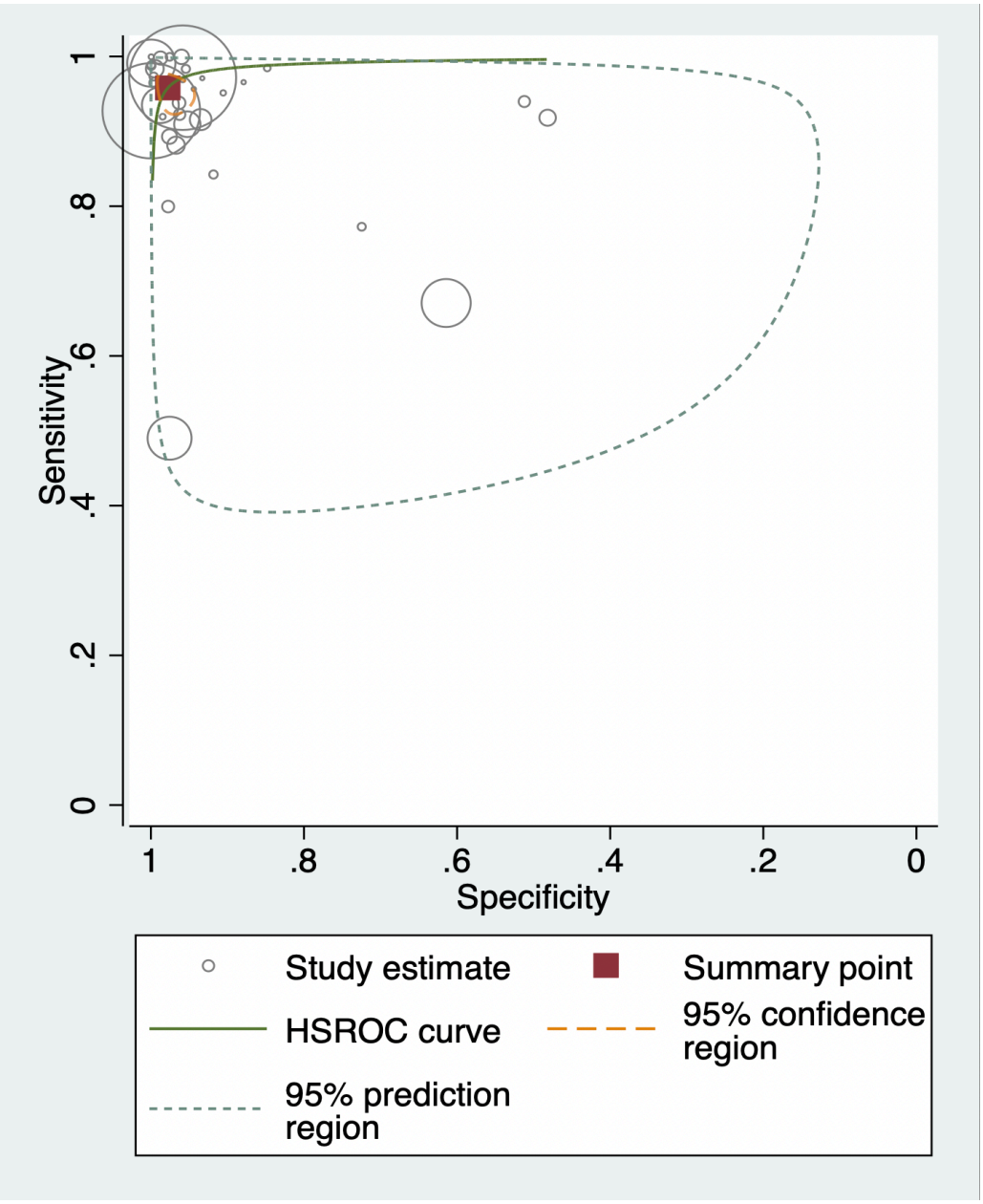


Table 18: Summary data for meta-analysis of diagnostic accuracy using HSROC curves.

Metric	Coefficient	SE	Z	P > Z	95% CI	
Bivariate						
E (logit Sensitivity)	3.13	0.25			2.63	3.6
E (logit Specificity)	3.79	0.39			3.03	4.5
Var (logit Sensitivity)	1.87	0.54			1.07	3.3
Var (logit Specificity)	4.82	1.31			2.83	8.2
Correlation (logit Sensitivity and logit Specificity)	0.38	0.16			0.03	0.6
HSROC						
Lambda	6.95	0.51			5.94	8.0
Theta	0.48	0.34			-0.19	1.2
Beta	0.47	0.19	2.54	0.01	0.11	0.8
s2alpha	8.28	2.24			4.87	14.1
s2theta	0.94	0.26			0.55	1.6
Summary point						
Sensitivity	0.96	0.01			9328285.00	1.0
Specificity	0.98	0.01			0.95	1.0
Diagnostic odds ratio	1007.65	531.10			358.65	2831.1
Positive likelihood ratio	43.33	16.49			20.55	91.4
Negative likelihood ratio	0.04	0.01			0.03	0.1
Inverse negative likelihood ratio	23.26	5.70			14.39	37.6

The analysis encompasses 35 studies and demonstrates a log-likelihood of 309.40. A covariance of 0.03 is observed between the estimates of Expected (logit Sensitivity) and Expected (logit Specificity).

Post-hoc model diagnostics were performed to evaluate possible model misspecification, goodness of fit, and to identify outlying and possibly influential data points. Goodness-of-fit (A) and Bivariate normality (B) showed that random effects bivariate model was suitable. Four studies (Castro et al 2017, Bozkurt et al 2019, Eide et al 2012, Thomas et al 2014) were identified as outliers and were found to exert significant influence on the pooled estimates (i.e.,

higher % weight) (Figure 18). Sensitivity analysis excluding these four studies generated a pooled sensitivity of 0.95, specificity of 0.97, and AUC of 0.99 (0.98 - 0.99) (Figure 19)

Figure 18: Model diagnostics to aid sensitivity analysis.

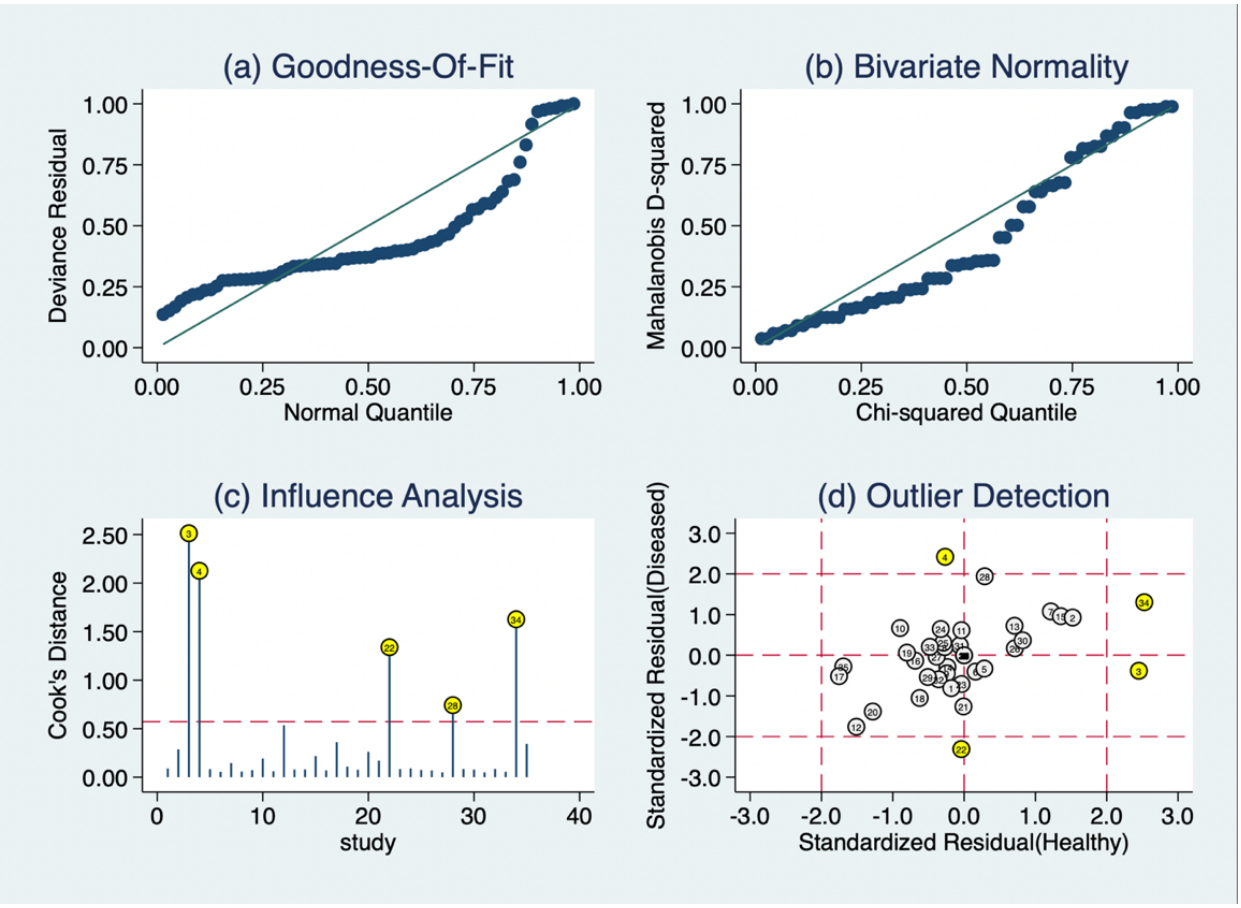
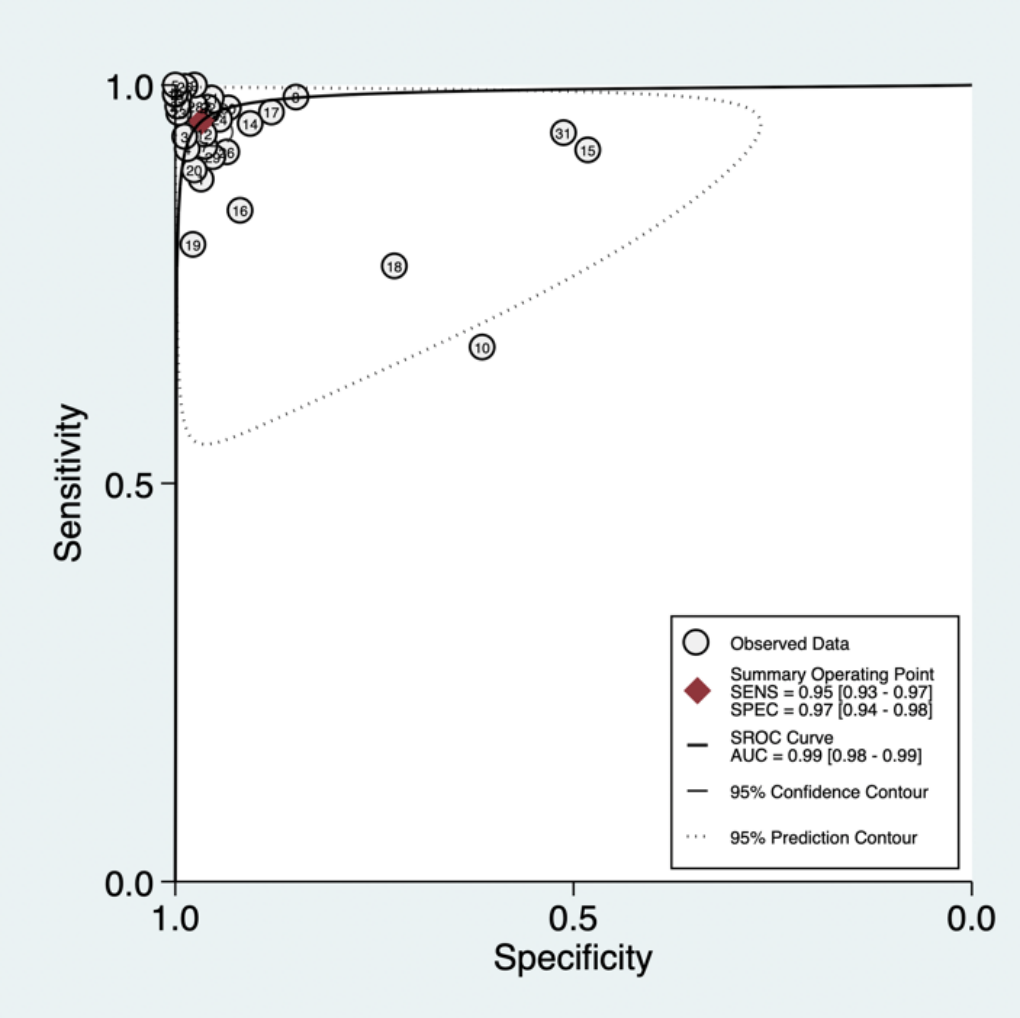


Figure 19: Sensitivity analysis excluding outliers.

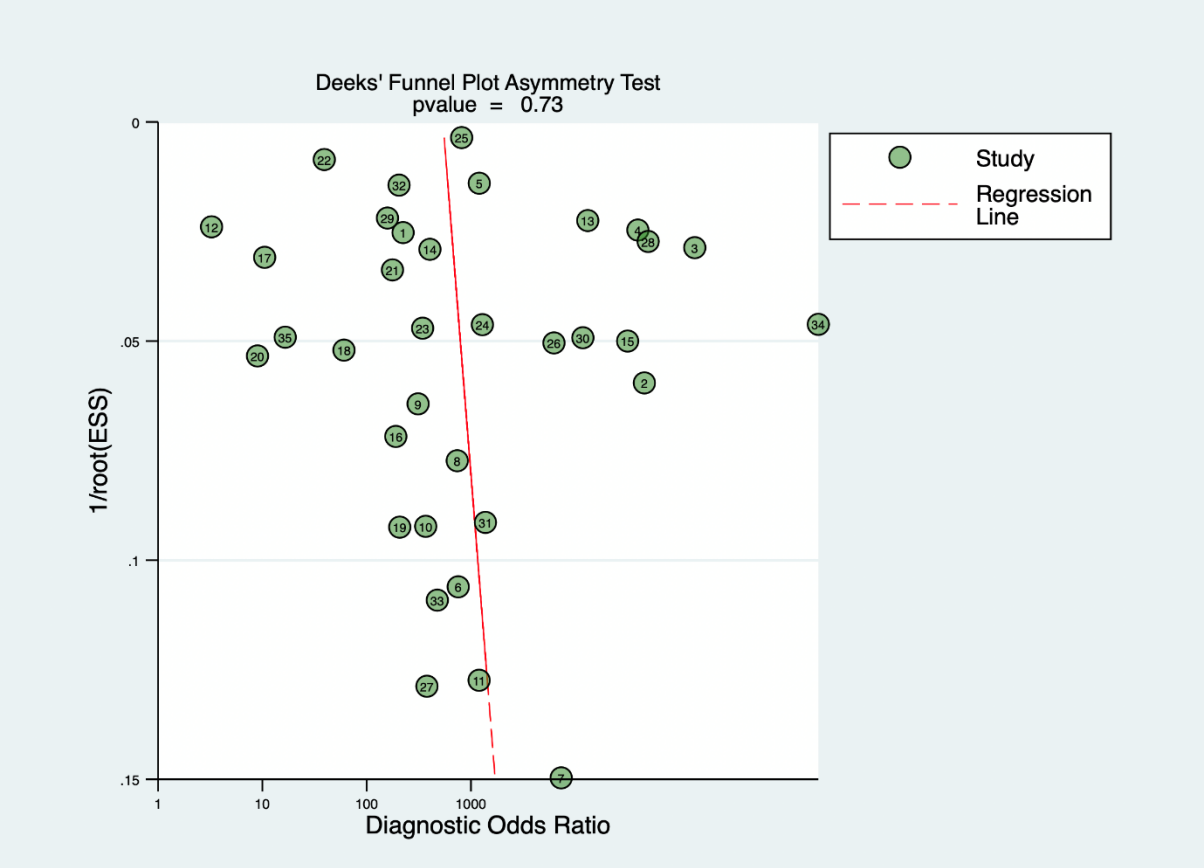


(Beadle et al., 1982; Cancer Research UK, 2024-d; Ali et al., 2022)

5.3.6 Publication bias

Using the Deek's effective sample size weighted regression test, there was no evidence of publication bias (t=0.35, p= 0.73) with a symmetrical funnel plot on visual inspection (Figure 20).

Figure 20: Deek's funnel-plot asymmetry test.



5.4 Discussion

This systematic review and meta-analysis navigate through the labyrinthine landscape of studies, exploring the application of rule-based NLP algorithms to aid clinical decision support across diverse scenarios in oncology over a considerable 27-year timespan. Given the pronounced variations in aspects like dataset size and input data types across the included studies, it is evident that rule-based NLP systems have undergone a dynamic evolution over this period. Variations in study design and the utilisation of both commercially available and internally developed NLP systems have created a rich and complex tapestry of methodological approaches. Notwithstanding this methodological variation, the overarching emphasis of the included studies substantiates NLP's efficacy in generating pivotal insights from colossal, text-

rich medical data, particularly on tasks relating to information extraction and text classification. This facilitates a cascade of possibilities: from fortifying case identification to crafting informed clinical decisions.

The observed diversity in the reporting quality of studies, as highlighted by the TRIPOD scores, emphasises the pressing need for a robust and standardised reporting framework. A unified and rigorously adhered-to reporting standard will be vital in delivering reliable, transparent, and replicable outcomes in future research endeavours exploring the application of NLP in oncology. The importance of clear reporting of research studies has been noted across the literature, with many reporting standards having been developed for other study or research types. Notable examples include the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement and the TRIPOD Statement, although both have limitations for use as reporting guidelines for NLP studies. Alsawas et al. attempted to address these shortcomings through the development of a set of critical appraisal criteria derived from common risk of bias domains and from the NLP literature (Alsawas et al., 2016). These criteria focus on the methods of sample selection, coding, the gold standard, algorithm training, algorithm testing and measures of accuracy. The scope of this tool is limited to outlining the broad domains for evidenced based-medicine critical appraisal and does not aim to serve as a reporting guideline for studies utilising NLP. There is therefore a gap in the literature, with the reporting of NLP studies currently not standardised and falling between reporting guidelines. The utilisation of a Delphi study to formulate reporting standards for NLP research could offer a tailored and consensus-driven solution. By pooling insights from NLP researchers, clinical experts, and standards specialists, the diverse clinical NLP research community could bridge the identified gaps between existing guidelines. The iterative nature of the Delphi method would ensure that all potential nuances specific to NLP studies are covered, and contentious

areas are refined until a broad agreement is reached. Furthermore, this process can instil greater confidence in the clinical NLP research community, knowing that the standards have been formulated with collective expertise and thorough deliberation. This would promote consistency, comparability, and transparency in NLP studies, ensuring that results are both robust and replicable, enhancing the overall quality and trustworthiness of clinical NLP research.

The complexity and richness of meta-analysis results from this study provide a multifaceted view of the performance of various NLP models in the context of algorithms to address a clinical problem in cancer. A key aspect of the analysis was the evaluation of the degree of heterogeneity across the included studies. As indicated by a covariance of 1.134 and a correlation coefficient of 0.376, there exists a definitive variability. This heterogeneity is further reflected in the I^2 statistic of 65.2% during generalised assessment, surging to 73.7% during sensitivity analysis. In this context, the utility of a unified hierarchical model cannot be overstated. By offering a two-tiered structure, this captures both within-study variability through binomial distributions of contingency tables, as well as between-study variability. The HSROC curve, accordingly, paints a comprehensive picture of the average sensitivity and specificity across studies. A rigorous diagnostic evaluation of the model further strengthened the robustness of the study findings. The goodness-of-fit and bivariate normality assessments endorsed the suitability of the selected model. Interestingly, the evaluation unveiled four studies that emerged as outliers. Their removal not only illuminated their profound influence on pooled estimates but also resulted in a heightened pooled sensitivity, specificity, and an AUC nearing perfection. The study findings emphasise the importance of ensuring a balanced algorithmic output from complex input data sources like histopathology reports, where both positive and negative identifications have profound implications in real-world clinical

scenarios and where misidentifications could have significant repercussions. The implication is clear: there is a compelling need for further refinement and recalibration of the algorithm, ensuring it aligns more closely with the demands and rigours of real-world applications. This naturally segues into a pertinent discussion on the potential integration of human-in-the-loop approaches in enhancing and safeguarding algorithmic outputs in a clinical context where the stakes of misclassification or erroneous prediction can be high. A judicious integration of human-in-the-loop embeds human expertise into the algorithmic workflow, positioning clinicians not merely as end-users but as active participants who supervise, validate, and refine algorithmic predictions. In this light, NLP algorithms become supplementary tools that sift through voluminous data, extracting and synthesising clinically relevant insights that are scrutinised and contextualised by human experts to ensure accuracy, relevance, and ethical adherence within real-world clinical applications.

A key challenge in this meta-analysis was the significant clinical outcome heterogeneity across included studies, which introduces complexities in synthesising findings and drawing definitive conclusions about the performance of text and task-based classifications outlined in Table 13. The variation in reported performance metrics, such as sensitivity, specificity, accuracy, and F1-score, reflects differences in study design, NLP methodology, dataset characteristics, and gold standard comparators. Some studies utilised structured pathology reports as input data, while others analysed free-text clinical notes, introducing variability in data quality and content. Furthermore, differences in outcome definitions such as whether studies measured algorithm performance against human annotations, registry data, or clinical decision support recommendations further contribute to inconsistencies in reported effectiveness. These variations impact the generalisability of pooled estimates, as performance metrics may not be directly comparable across studies with differing

methodological frameworks. The use of a hierarchical model and HSROC curves helps mitigate some of these challenges by accounting for within- and between-study variability.

5.4.1 Strengths and limitations

NLP's historical evolution is showcased and its multifaceted applicability in cancer care. However, this is somewhat obfuscated by considerable variance in methodology and reporting standards amongst the examined studies. Although confronted with hurdles such as disparate performance metrics and occasionally inaccessible raw data, the robust meta-analysis undertaken yields a nuanced appraisal of algorithmic efficacies in this context. Thus, the study findings spotlight a critical, impending necessity for the instatement of standardised reporting and methodological frameworks in future NLP research endeavours. The spectre of publication bias, a persistent concern in meta-analyses, was carefully addressed. The Deek's effective sample size weighted regression test, coupled with the symmetrical appearance of the funnel plot, dispelled any underlying apprehensions about potential publication bias. This affirmation of objectivity and fairness lends an added layer of credibility to the findings.

5.4.2 Future directions

In navigating forward in the arena of automated oncological clinical decision support, there is a compelling need to meticulously orchestrate head-to-head studies that compare the respective and relative efficacies of rule-based NLP systems with alternative technologies such as machine learning, deep learning, transfer learning, and large language models.

5.5 Conclusion

This study illuminates both the potent capabilities and challenges encountered across a diverse spectrum of studies over 27 years, confirming the critical role of rule-based NLP algorithms in

extracting essential insights from voluminous medical free-text data and enhancing oncological clinical decision-making. Amidst the promising horizons of various NLP algorithms in fortifying clinical decision support, a conspicuous heterogeneity in methodology and reporting has surfaced, sounding an urgent call for the adoption of a standardised, universally adhered-to reporting framework. The meta-analysis, while exposing noteworthy disparities across studies, concurrently offers data that may serve as a benchmark for evaluating and refining NLP algorithms in this area. The trajectory ahead invites a synthesis of technological adeptness and standardised reporting, facilitating the judicious evolution of NLP in oncological applications.

Chapter Six:

Development and validation of an automated basal cell carcinoma histopathology information extraction system using natural language processing

6.1 Introduction

Routinely collected healthcare data has the potential to significantly impact skin oncology research, identifying new disease associations, treatment modalities, outcomes and healthcare delivery planning (Gibson, Cordaro, et al., 2021; Gibson et al., 2020). Current coding methods and therefore data capture for NMSC grossly underestimate the true burden of disease, by up to 50.0.0% for BCC and 30% for SCC data (Ibrahim et al., 2021). As per the UK and Ireland association of cancer registries (UKIACR), most cancer registries currently only record a single BCC or SCC per patient lifetime (Ibrahim et al., 2021). Efforts are underway to enhance data capture methodologies, with a transition towards a first per patient per annum approach. This evolution is evident in initiatives such as the transformation of Welsh Cancer Intelligence and Surveillance Unit (WCISU) data to first per patient per annum (FPPPA) and a similar trend observed in the National Cancer Registration and Analysis Service (NCRAS). The adoption of this methodology facilitates the classification of ‘non-genital BCC’ and ‘non-genital cutaneous SCC’ into distinct categories: 'first tumour' and 'subsequent tumour' (NHS England, 2024-b).. While only first tumours are fully registered, subsequent tumours are often imputed from pathology reports, resulting in poorer data quality. Consequently, these subsequent tumours are generally excluded from international incidence rates.

The classification of tumours into 'first' and 'subsequent' enables meaningful comparisons with international statistics and facilitates the identification of cohorts with robust data quality for further analysis. Patients are identified based on unique patient IDs, with the first tumour defined as the one registered with the earliest diagnosis date, starting from 1995 onwards. All tumours registered or imputed with later diagnosis dates are categorised as 'subsequent tumours'. Notably, the methodology restricts the inclusion of only one tumour per individual in each year, ensuring the integrity and accuracy of the data analysis.

The limitations of current data collection practices extend beyond the underreporting of metachronous, synchronous, and recurrent lesions (Ibrahim et al., 2021). Not only is a significant volume of data therefore not collected, but the depth and quality of data are also lacking. Clinical coding using International Classification of Diseases Version 10 (ICD-10), Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures Version 4 (OPCS-4) and other coding systems can mis-represent the true disease and treatment burden (Beadle et al., 1982; de Vries et al., 2012; Venables et al., 2019). Furthermore, error can be introduced by those tasked with coding disease or treatment episodes (Daniels et al., 2021). This practice of underestimating the true incidence of NMSC is not isolated to the UK alone (Lomas, Leonardi-Bee, et al., 2012). In order to gain a better understanding of the disease and the burden it places on both patients and the NHS, this needs to be addressed.

Rich data can be found locked away in 'unstructured' formats such as clinic letters, hand written clinical notes and histopathology reports which form part of EHRs (Gibson, Dobbs, et al., 2021b). Manual review of the free text in EHRs has been the mainstay of data capture in this setting. This is a process which is labour intensive, costly and open to error and bias. Traditionally, EHR data have been inaccessible in 'unstructured' free text format, however with the advent of NLP this is no longer the case. Information extraction using NLP describes a set of techniques used to convert passages of written text into interpretable datasets through either rule-based or machine learning models (Harrison & Sidey-Gibbons, 2021). It can be used in healthcare for named entity recognition (NER) or feature extraction using unstructured EHR data. The stages can conceptually be broken down and summarised into five steps – text extraction, text processing, system task, performance evaluation and implementation

(Nadkarni et al., 2011). First an unstructured free text report is converted into a series of features such as part of speech tags, tokens and phrase chunks which are then algorithmically processed. The NLP algorithm can then be set to task. After the system is validated it can finally be applied to unstructured text to extract data in the setting of its intended purpose. The nascent field of NLP has the potential to enrich routinely collected healthcare data with detailed disease-specific information and harness the power of big data to ensure that these data are accurate, comprehensive and easily accessible.

The importance of big data and modern analytical techniques in medical research have been identified by the UK Government in their Eight Great Technologies drive, by the Medical Research Council in 2016/17 who plan to invest £37.5 million in health informatics over the next five years and by The Royal College of Surgeons of England in their 'Future of Surgery' commission (Department for Business, 2013; Kerr, 2019; Medical Research Council, 2024). Additionally, the King's Fund report into dermatology services in the UK commissioned by the British Association of Dermatologists recognises the need for improved dermatology service data collection and accessible real-time information (Edwards & Imison, 2024). The use of NLP in clinical outcomes research is accelerating. In a recent systematic review and meta-analysis of NLP-based data capture versus conventional administrative methods of data capture (current procedural terminology [CPT] codes, international classification of diseases [ICD] codes, patient safety indicators based on discharge coding and diagnosis-related group [DRG] database code) postoperative complications were identified with higher sensitivity whilst specificity was comparable (Mellia et al., 2021). NLP models may be reliably used for both confirming and ruling out documentation of outcomes/diagnoses whilst conventional methods of data capture demonstrate clinical utility for confirming documentation of outcomes/diagnoses alone (Mellia et al., 2021). Applications of NLP to the EHR continue to

expand and include novel phenotype discovery, clinical trial screening, pharmacogenomics, drug-drug interaction and adverse drug event detection and genome-wide and phenome-wide association studies (Zeng et al., 2019).

This wave of enthusiasm can be built upon to improve data collection and therefore research in skin oncology. In this study a rule-based NLP pipeline was developed and internally validated to extract BCC primary histopathological data, with the ultimate aim to improve cancer registry data, support service planning and enhance the quality of research using routinely collected data.

6.2 Material and methods

6.2.1 Study population

Manually de-identified and pseudonymised BCC histopathology reports from Swansea Bay University Health Board were used for the study. Forty-one histopathology reports from 2015 were used to develop, train and test rule sets. This training corpus of 41 reports contained 62 individual BCCs. Training histopathology reports were written by ten consultant histopathologists. The gold standard validation corpus consisted of 200 histopathology reports from 2016 to 2018 and contained 299 individual BCCs. The validation histopathology reports were written by 20 consultant histopathologists.

6.2.2 Annotation

Each free text histopathology report was annotated using the open-source, web-based annotation tool Markup (https://www.getmarkup.com/) (Dobbie et al., 2021). Markup incorporates NLP and Active Learning (AL) technologies to enable rapid and accurate

annotation using custom user configurations, predictive annotation suggestions, and automated mapping suggestions to both domain-specific ontologies, such as the Unified Medical Language System (UMLS), and custom, user-defined ontologies.

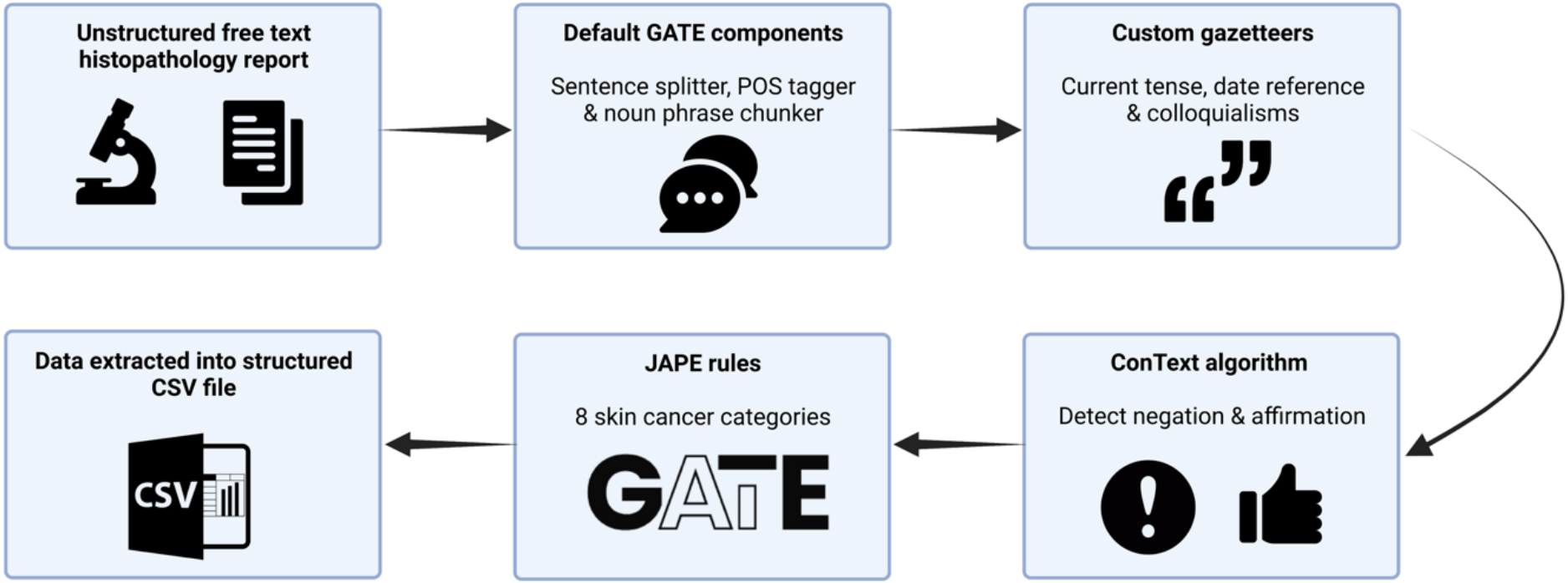
6.2.3 Annotation guidelines

An annotation guideline (Appendix 3) and data definition dictionary (Appendix 3) were developed as an aide for clinicians when annotating histopathology reports as the ground truth. An iterative approach to guideline development was taken, with a first draft containing general guidelines updated and re-tested following its implementation by two clinicians on an initial 20 histopathology reports and compared to a gold standard defined by consensus agreement between two expert skin cancer clinicians.

6.2.4 Algorithm construction

General Architecture for Text Engineering (GATE) Developer 9.0 (University of Sheffield, UK) was used for the study, an established open source toolkit for NLP, to build an information extraction system using rule-based techniques from histopathology reports (Cunningham, 2002) (Figure 21).

Figure 21: Schematic representation of the rule-based NER NLP system for BCC histopathology reports.



GATE can be defined as a Java based infrastructure for developing and deploying software components that process human language (Cunningham, 2002). GATE as an architecture can be broken down into various types of component, known as resources (Cunningham, 2002). These resources include language resources (lexicons, corpora or ontologies), processing resources (parsers, generators or ngram modellers) and visual resources (visualisation and editing components involved with graphical user interfaces) (Cunningham, 2002). The 139 entities the study set out to extract are summarised in Appendix 3. Custom gazetteers (native dictionaries used within GATE) were developed being informed by the World Health Organisation classification of tumours of soft tissue and bone tumours, to map clinical terms to UMLS concepts (Fletcher et al., 2013). The ConText algorithm was deployed to detect negation of extracted terms e.g. 'there was *no* residual disease' and to detect affirmation of normal prognostic factors, such as 'tumour *confined* to the dermis'. Finally, Java Annotation Patterns Engine (JAPE) scripting language was used to define rules based on varying combinations of UMLS and custom lookups to extract eight broad information categories. In total, 80 separate gazetteers and 445 JAPE rule files were created in order to annotate the variables of interest, establish context and to remove certain annotations from the output. Data were outputted into a comma-separated values (CSV) file by using the Groovy scripting language.

6.2.5 Determining the number of documents needed for a gold standard validation corpus

There is no agreed standard for determining the size of a validation set in NLP (Alsawas et al., 2016; Johnson et al., 2018; Juckett, 2012). String matching and a modified version of the method outlined by Juckett et al were therefore used to determine the number of documents required for the validation set to ensure all 139 concepts were represented well enough to validate them (Juckett, 2012). A further 1000 manually de-identified and pseudonymised BCC

histopathology reports generated from Swansea Bay University Health Board in 2015 were used for this. The capture probability of a token (which is any sequence of alphanumeric characters, beginning with a letter and occurring between spaces, slashes, brackets, braces, parentheses, quotation marks or punctuation marks that is found in the study's 139 concepts) occurring in the validation set from this further 1000 reports was calculated. This is explained in detail here.

Assuming a corpus of 1000 documents and a word appears in 3 documents the probability that at least one document has the word when randomly selecting 100 documents is shown by Equation 1.

Equation 1: Probability that at least one document has the word when randomly selecting 100 documents.

$$P(x > 0) = 1 - \frac{997}{1000} \times \frac{996}{999} \times \ldots \times \frac{888}{901}$$

Or more generally, assuming a corpus of N documents and a word appears in n_{doc}documents the probability that at least one document has the word when randomly selecting n documents is shown by Equation 2.

Equation 2: The probability that at least one document has the word when randomly selecting n documents.

$$P(x > 0) = 1 - \frac{N - n_{doc}}{N} \times \frac{N - n_{doc} - 1}{N - 1} \times \ldots \times \frac{N - n_{doc} - n + 1}{N - n + 1}$$

By setting N = 1000 and calculating the probability for various n and n_{doc} a plot of capture probability vs document frequency can be generated (Figure 22). This plot shows the capture probability of a word with a given document frequency in a 1000 document corpus. It is not surprising that the capture probability of a word with a smaller document frequency is lower. A valid token for this process was defined as any sequence of alphanumeric characters, beginning with a letter and occurring between spaces, slashes, brackets, braces, parentheses, quotation marks or punctuation marks. To capture 95% of tokens of document frequency = 1, 950 documents would be needed. Less than 500 documents would be needed if the document frequency = 5 and less than 300 if the document frequency = 10. Note that a document frequency = 10 means the word only appears in 1.0% of the documents in the corpus. The capture probability of all valid and relevant tokens manually picked by expert skin cancer clinicians from the working corpus was then calculated. Assuming n_{token} unique tokens and knowing the capture probability of each token p_i the percentage of tokens being captured can be shown by Equation 3.

Equation 3: Percentage probability of token being captured.

$$\sum_{i=1}^{n_{token}} p_i / n_{token}$$

The document frequency of valid and relevant tokens in the corpus was then calculated. By only keeping words with a minimum length, small token sizes with a low document frequency can be eliminated (Figure 23). Only tokens with nine characters were used as a cut off for this task (Figure 24). A minimum document frequency of 5 or above was also set as it was deemed that rare tokens would not be used in the employed JAPE rules. In order to achieve a capture probability of 90.0% a validation set of 200 documents was required.

Figure 22: Plot of capture probability versus document frequency.

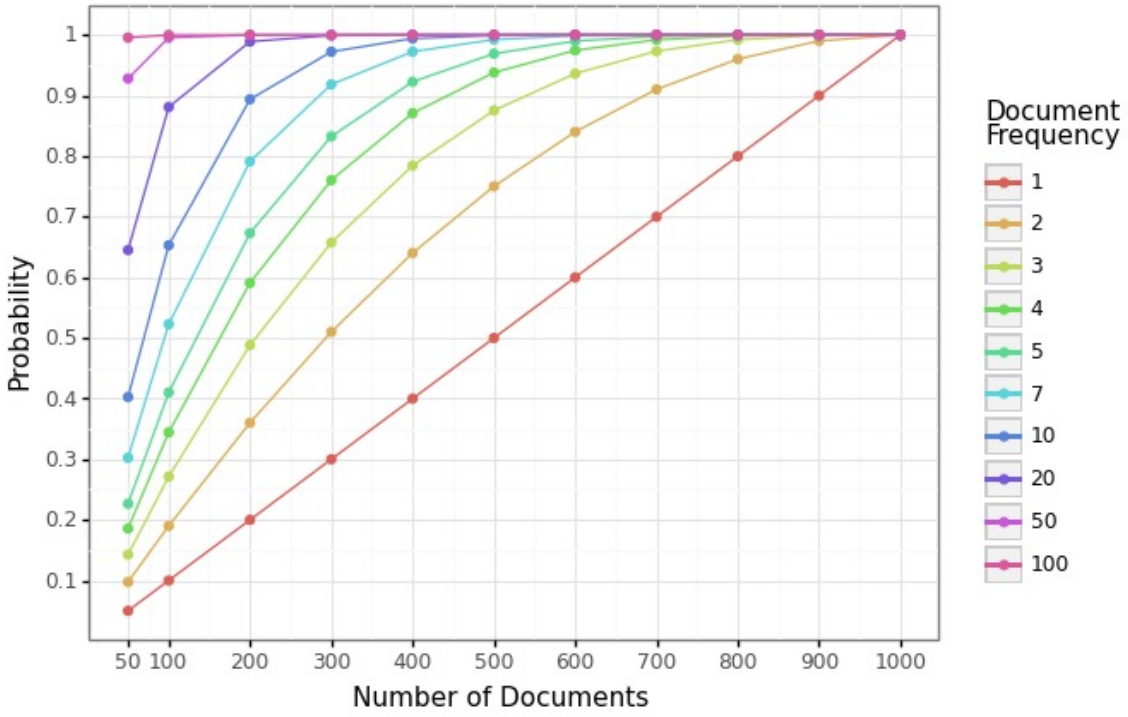


Figure 23: Distribution of document frequency in the working corpus.

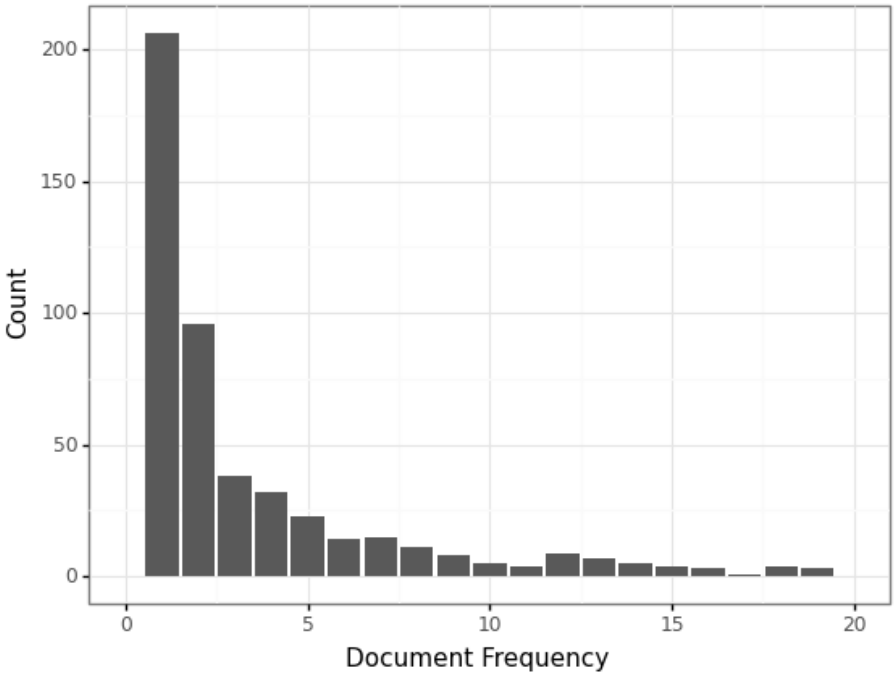
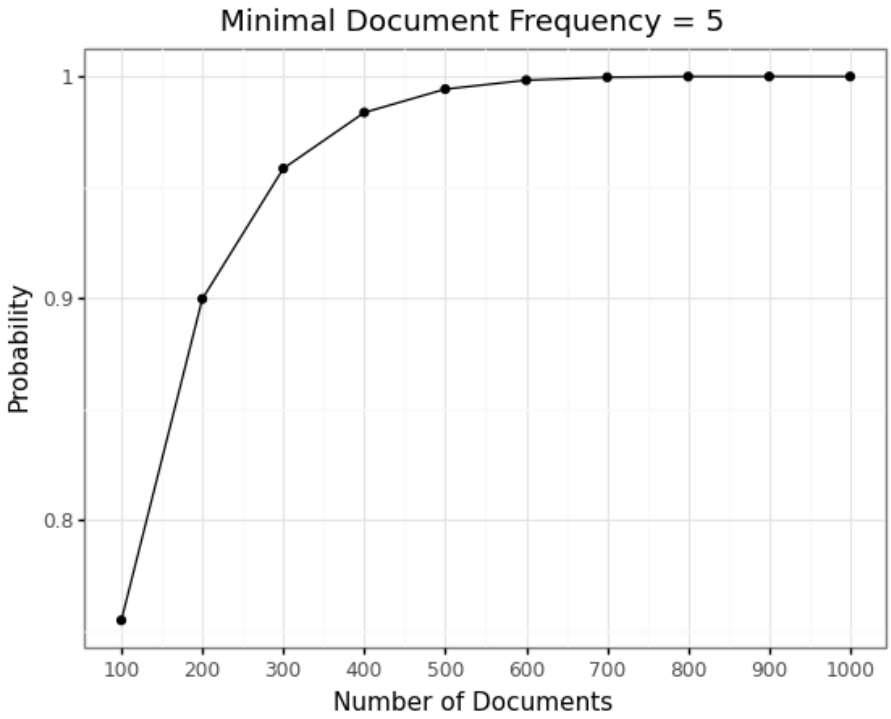


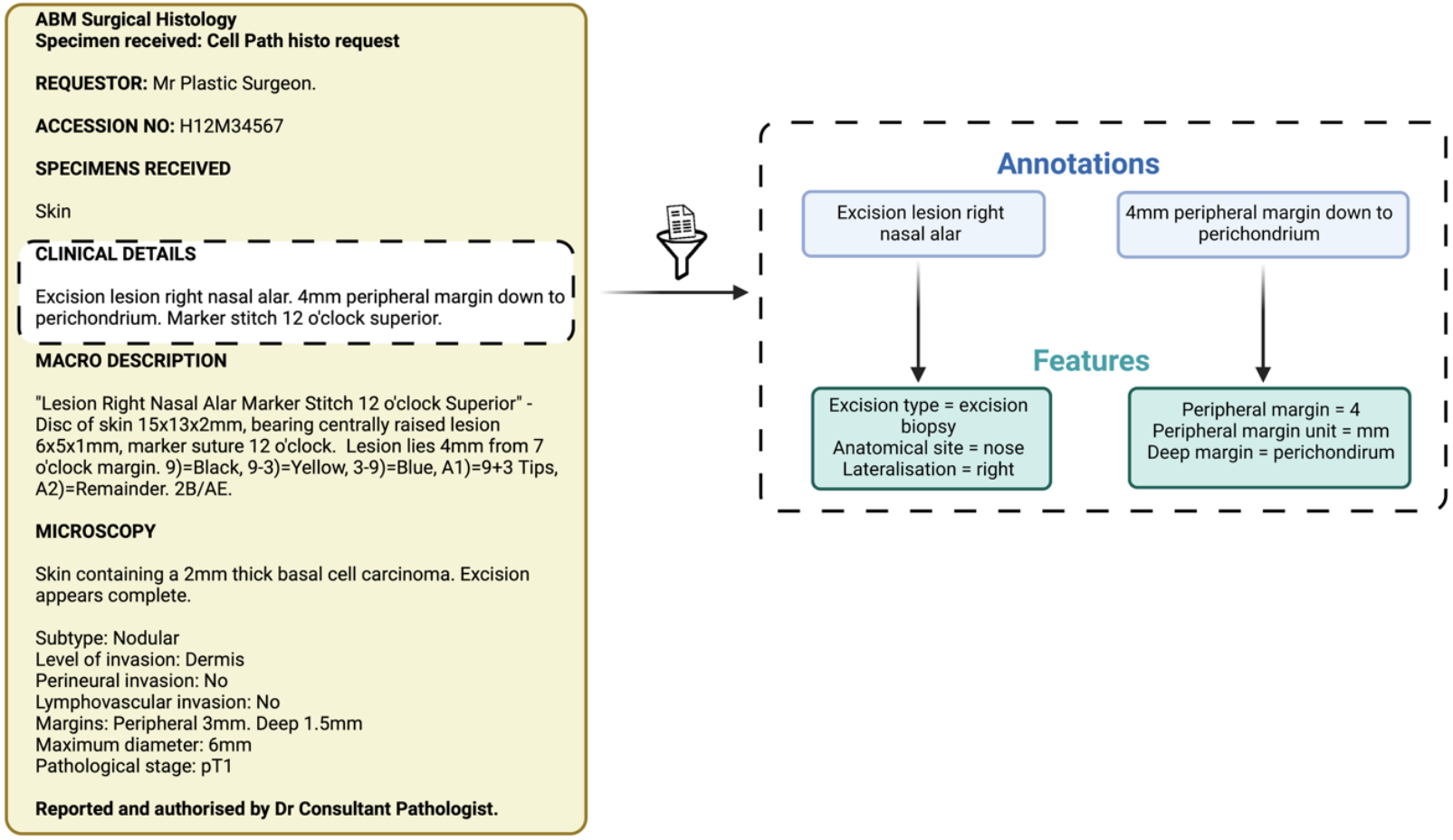
Figure 24: Plot demonstrating the aggregate capture probability when keeping tokens appearing in ≥ 5 histopathology reports.



6.2.6 Analysis and statistical tests

Items of information extracted by the NLP pipeline were compared with those extracted by manual review performed by two independent expert skin cancer clinicians who had access to the annotation guidelines only (Figure 25).

Figure 25: Schematic representation of annotations and features. These items of information were extracted by both the NLP pipeline and expert manual review.



The most widely adopted measures in the literature were used to evaluate an NLP pipeline; precision, recall and F1 score to calculate the accuracy of the NLP pipeline when compared to clinician assessment (Table 19) (Cunningham, 2002).

Table 19: Formulae used to calculate precision, recall and F1 score.

Measure	Formula
Precision	$= \frac{Correct + \frac{1}{2}\, partial}{Correct + spurious + partial}$
Recall	$= \frac{Correct + \frac{1}{2}\, partial}{Correct + missing + partial}$
F1 score	$= \frac{2 \times precision \times recall}{precision + recall}$

Partial; two annotations are partially compatible if they overlap and if the features of one (usually the ones from the key) are included in the features of the other (response), Spurious; a response annotation is spurious if either it is not coextensive or overlapping, or if one or more features from the key are not included in the response annotation, Missing; a key annotation is missing if either it is not coextensive or overlapping, or if one or more features are not included in the response annotation.

GATE's definition of precision was used as the number of correctly identified items expressed as a percentage of the number of items identified, recall as the number of correctly identified items expressed as a percentage of the total number of correct items and the F1 score as the harmonic mean of precision and recall (aiming to achieve a balance between precision and recall) (Table 17) (Cunningham, 2002). Precision is analogous to positive predictive value (PPV) and aims to measure how many of the items identified by the application are actually correct, irrespective of whether it also failed to retrieve correct items (Table 17) (Cunningham, 2002). Recall is analogous to sensitivity or the true positive rate and aims to measure how many

of the items that should have been identified actually were identified, regardless of how many spurious or false positive identifications were made (Table 17) (Cunningham, 2002).

The assessment of partially correct annotations can differ in GATE depending on the intended task of the pipeline (Table 17) (Cunningham, 2002). 'Strict', 'average' and 'lenient' are graded approaches to deriving an F1 score depending on how well the clinician annotation matches or spans that of the data extracted by the algorithm. A 'strict' F1 score considers only perfectly matching annotations to be correct, a 'lenient' F1 score considers partially matching annotations as correct whilst an 'average' F1 score derives a value from the average of the strict and lenient F1 scores. The span of the annotation was not considered as an important goal in the generation of the study's pipeline and a lenient scoring approach was used to calculating the F-measure.

Each clinician's set of annotations was compared to one another using the Python programming language to calculate 1) difference in F1 score and 2) inter-annotator agreement (IAA). The single gold standard validation corpus was initially annotated by two independent and blinded expert clinicians involved in skin cancer care. Following this a single gold standard validation corpus was then produced following a meeting between the clinician annotators who discussed and resolved disagreements in their annotations to achieve consensus through strict adherence to the annotation guideline and data dictionary. Statistical data analyses were performed using RStudio (R Core Team, R Foundation for Statistical Computing, Vienna, Austria).

6.3 Results

6.3.1 Training

2591 items of information were identified in 41 histopathology reports with overall precision, recall and F1 score of 94.9% (95% CI 88.9-100.0), 95.1% (95% CI 89.9-100) and 94.8% (95% CI 89.1-100.0), respectively, when assessed against a single clinician. Table 20a summarises the performance of the NLP pipeline in identifying these items of information.

Table 20: Information extracted from (a) training corpus of 41 BCC histopathology reports and (b) validation corpus of 200 BCC histopathology reports.

Entity	Number of annotations created per report by clinician		Number of features in clinician annotation		Mean number of features per clinician annotation		Number of annotations created per report by algorithm		Number of features extracted by algorithm		Mean number of features per algorithm extraction		Match		Missing		Spurious		Partial	
	Training (a)	Validation (b)	Training (a)	Validation (b)	Training (a)	Validation (b)	Training (a)	Validation (b)	Training (a)	Validation (b)	Training (a)	Validation (b)	Training (a)	Validation (b)	Training (a)	Validation (b)	Training (a)	Validation (b)	Training (a)	Validation (b)
Accession number	38	199	38	199	1.0	1.0	38	199	38	199	1.0	1.0	0	0	0	0	0	0	38	199
Excision date	19	5	57	15	3.0	3.0	19	5	57	15	3.0	3.0	0	0	0	0	0	0	19	5
Clinical details	89	345	234	1054	2.6	3.1	86	330	258	915	3.0	2.8	49	37	7	117	4	101	33	191
Macroscopic details	134	660	503	2811	3.8	4.3	136	752	546	2787	4.0	3.7	51	12	24	227	30	309	59	431
Microscopic details	125	682	272	1530	2.2	2.2	125	624	289	1322	2.3	2.1	45	11	18	289	24	214	62	399
Microscopic measurements	170	757	538	2433	3.2	3.2	170	648	626	3032	3.7	4.7	85	6	10	213	12	94	75	548
Report details	40	157	103	172	2.6	1.1	40	152	103	167	2.6	1.1	27	0	0	43	0	33	13	119
Requestor	20	5	40	10	2.0	2.0	20	5	40	10	2.0	2.0	20	0	0	0	0	0	0	5
Supplementary report	0	10	0	38	0	3.8	0	18	0	44	0	2.4	0	0	0	4	0	12	0	6

Match; fully matching clinician annotation and algorithm extraction.

There was variance in performance amongst the canonical structure of the histopathology report signifying areas where the NLP algorithm was able to undertake NER with less error and reflecting feature complexity. Precision and recall for the categories were: accession number (100.0%, 100.0%), excision date (100.0%, 100.0%), clinical details (96.5%, 93.1%), macroscopic details (81.1%, 83.4%), microscopic details (87.0%, 89.1%), microscopic measurements (94.8%, 95.2%), report details (100.0%, 100.0%), and requestor (100.0%, 100.0%) (Table 21a).

Table 21: Performance of the NLP pipeline on (a) training and (b) corpus compared to clinician assessment. Values calculated per document across annotation types and averaged across the corpus, displayed with 95% CIs.

Entity	Precision %		Recall %		F1 score %	
	Training (a)	Validation (b)	Training (a)	Validation (b)	Training (a)	Validation (b)
Accession number	100.0	100.0	100.0	100.0	100.0	100.0
Excision date	100.0	100.0	100.0	100.0	100.0	100.0
Clinical details	96.5 (90.3-100)	75.4 (70.2-80.7)	93.1 (86.8-99.4)	73.3 (68.0-78.5)	94.1 (88.2-100)	73.6 (68.4-78.84)
Macroscopic details	81.1 (72.3-89.9)	60.8 (56.3-65.3)	83.4 (75.1-92.3)	66.4 (62.3-70.4)	82.1 (73.4-90.8)	63.0 (58.7-67.3)
Microscopic details	87.0 (79.8-94.3)	74.2 (69.8-78.6)	89.1 (82.7-95.5)	66.9 (62.5-71.3)	87.5 (80.7-94.3)	69.1 (64.8-73.5)
Microscopic measurements	94.8 (91.2-98.5)	83.0 (79.1-86.8)	95.2 (91.2-99.3)	72.8 (68.5-77.1)	94.8 (91.0-98.5)	75.9 (71.8-79.9)
Report details	100.0	83.5 (78.3-88.7)	100.0	81.4 (76.1-86.7)	100.0	81.7 (76.4-87.0)
Requestor	100.0	100.0	100.0	100.0	100.0	100.0

6.3.2 Validation

11,224 items of information were identified across 200 histopathology reports in the validation set. Table 18b summarises the performance of the NLP pipeline in identifying these items of information, with Table 19b showing the performance relative to the gold standard. The mean precision, recall and F1 scores were 86.0% (95% CI 75.1-96.9), 84.2% (95% CI 72.8-96.1) and 84.5% (95% CI 73.0-95.1) respectively. The overall difference between mean clinician annotator F1 scores was 7.9% (Table 22) in comparison to 15.5% between the NLP pipeline and the gold standard corpus.

Table 22: Differences in (a) F1 score between annotators on the validation corpus and (b) inter-annotator agreement on the validation corpus.

Entity	(a) Difference in F1 score (%)	(b) Specific agreement
Accession number	0.0	1.00
Excision date	0.0	1.00
Clinical details	1.7	0.95
Macroscopic details	3.9	0.90
Microscopic details	9.5	0.89
Microscopic measurements	2.6	0.97
Report details	0.9	0.99
Requestor	0.0	1.00
Supplementary report	52.4	0.44

A confusion matrix was used to identify tokens identified by one annotator but not by the other (Table 23).

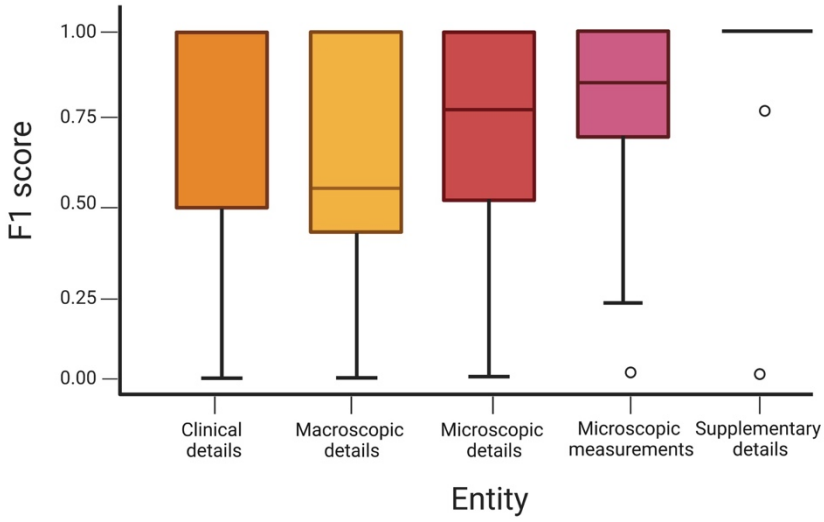
Table 23: Confusion matrix from validation corpus data with columns representing annotations by the first clinician and the rows representing annotations by the second clinician**.** Not identified label is used to label a token identified by one annotator but not by the other.

Entity	Accession number	Excision date	Clinical details	Macroscopic details	Microscopic details	Microscopic measurements	Report details	Requestor	Supplementary report	Not identified
Accession number	199	0	0	0	0	0	0	0	0	0
Excision date	0	4	0	0	0	0	0	0	0	0
Clinical details	0	0	302	0	0	0	0	0	0	23
Macroscopic details	0	0	0	602	0	0	0	0	0	80
Microscopic details	0	0	0	0	589	0	0	0	0	58
Microscopic measurements	0	0	0	0	0	734	0	0	0	25
Report details	0	0	0	0	0	0	156	0	0	2
Requestor	0	0	0	0	0	0	0	5	0	0
Supplementary report	0	0	0	0	0	0	0	0	2	2
Not identified	0	0	10	61	88	18	2	0	3	0

IAA between each clinician and the gold standard corpus F1 scores was calculated using specific agreement (Table 20b). A token was characterised by line range, entity and attributes and assigned to the component of the histology report it belonged to e.g. Clinical Details. The overall Cohen's Kappa score on annotated tokens and F1 score, calculated using pairwise comparisons, were 0.85 and 0.88 respectively.

Sub-analysis of the features extracted within the different report entities shows a variance in performance by which these data are extracted (Figure 26).

Figure 26: Boxplot of F1 score across entities demonstrating variance with 95% CIs.



For example, if the pipeline is tasked to look at the ‘tag’ feature within the microscopic details entity to calculate BCC incidence and surgical volume, an F1 score of 86.3% (95% CI 83.7-88.9) is achieved. However, if looking at values for ‘tumour thickness’, ‘tumour diameter’, ‘peripheral clearance’ and ‘deep clearance’ within the entity of microscopic measurements, a lower F1 score of 78.5% (95% CI 74.6-82.3) is achieved.

Post-hoc analysis revealed 14 reports (7.0%) used a form of structured template in the validation corpus whilst none were used in the training corpus. There were 3 forms of template used but no report utilised the 'cutaneous basal cell carcinoma removed with therapeutic intent' proforma produced by The Royal College of Pathologists (RCPath) in their minimum dataset reporting guideline (Slater & Barrett, 2019-a). The difference in performance of the study's pipeline with and without a template is shown in (Table 24). A marginal overall increase in precision (2.5%), recall (1.8%) and F1 score (1.1%) with proforma use is demonstrated.

Table 24: Performance (a) with and (b) without the use of a report template during validation.

Entity	Precision %		Recall %		F1 score (%)	
	With template (a)	Without template (b)	With template (a)	Without template (b)	With template (a)	Without template (b)
Accession number	100.0	100.0	100.0	100.0	100.0	100.0
Excision date	-	99.7 (99.2-100.0)	-	100.0	-	99.8 (99.5-100.0)
Clinical details	98.2 (94.4-100.0)	81.8 (77.2-86.4)	98.2 (94.4-100.0)	83.6 (79.2-88.1)	98.2 (94.4 - 100.0)	81.6 (77.1-86.1)
Macroscopic details	91.2 (82.0-100.0)	85.8 (82.2-89.4)	77.9 (66.7-89.1)	76.4 (72.9-79.9)	83.4 (73.8-93.0)	79.8 (76.5-83.2)
Microscopic details	64.7 (49.7-79.7)	74.8 (70.8.8-78.8)	72.0 (58.2-85.9)	82.1 (78.3-85.8)	66.6 (51.7-81.5)	76.7 (73.0-80.5)
Microscopic measurements	86.9 (77.6-96.2)	71.5 (66.7-76.2)	91.1 (83.3-98.9)	82.4 (78.1-86.7)	87.7 (80.6-94.7)	74.9 (70.4-79.4)
Report details	85.8 (64.7-100.0)	95.3 (92.5-98.1)	100.0	96.8 (94.2-99.3)	85.8 (64.7-100.0)	95.7 (93.0-98.4)
Requestor	-	99.7 (99.2-100.0)	-	100.0	-	99.8 (99.5-100.0)
Supplementary report	100.0	100.0	92.9 (77.4-100)	98.2 (96.4-100)	92.9 (77.4-100)	98.2 (96.4-100)

6.4 Discussion

A novel NLP pipeline was developed and validated for BCC histopathology reports. This is the first reported use of NLP for NMSC NER which differs from other more primitive NLP models used for text classification and case identification.

The overall performance of this pipeline was good and most importantly it compared well with clinician review (difference between mean F1 score: 15.5% vs 7.9%). As to be expected the pipeline performed better at certain tasks compared to others. For example, when extracting data on accession number, excision date, report details and requestor an F1 score of 100% was achieved. Given that these fields consist of fixed format dates which are easier to extract this is to be expected. More complex entities performed less well, although still with performance close to that expected of an experienced clinician. In terms of disease-specific information, the pipeline performed best in identifying microscopic measurements with an F1 score of 75.9%. These items are frequently mentioned and presented in a relatively standard format e.g. peripheral clearance 1mm at 12 o'clock. The entities with the highest drop off in F1 score between validation and training were clinical details, macroscopic details and microscopic details. In training, F1 scores were 94.1%, 82.1% and 87.5% respectively, however during validation they fell to 73.6%, 63.0% and 69.1% respectively.

There are a number of explanations for this. A model that performs poorly on internal validation can be described as under-fit or having high bias, commonly caused by insufficient data, or an overly simplistic model. It is generally accepted that more data produces better accuracy and higher quality data (closer domain, less noise) (Beleites et al., 2013; Cho et al., 2015). Exactly how much data or what quality of data is required to achieve a given performance goal however, is unclear in the context of NLP as an engineering discipline.

Specifically, there are no universally agreed criteria for the required number of documents needed for a gold standard validation corpus (Alsawas et al., 2016; Johnson et al., 2018; Juckett, 2012). The training set was generated in 2015 whilst the validation set represented data from 2016 to 2018. Additionally, histopathology reports were written by a group of ten individuals in the training set and 20 in the validation set. This may account for linguistic differences in sentence structure and performance of the JAPE rules designed in this study. The population characteristics of the test set i.e. the reporting style of the pathology reports should mirror the target population for the algorithm (Kuo et al., 2021). While population characteristics were generally similar, subtle differences are inevitable and it is difficult to account for this phenomenon whilst reducing selection bias. Despite no proforma being used in the training corpus, post hoc analysis of the validation data suggests that a structured reporting template such as that developed by the RCPath could improve the performance of the study's pipeline. It is intuitive that designing JAPE rules on a larger training set which includes variants of such a proforma should yield higher performance.

Studies assessing error rates in manual clinical data entry demonstrate that rates vary between 0.2% and 26.9% depending on the complexity of interpretation and abstraction of individual data elements (Arts et al., 2002; Goldberg et al., 2008). Current NMSC-based NLP pipelines report validation on smaller subsets and are centred around text classification and case identification rather than true NER. Lott et al used NLP on 80,368 histopathology reports to investigate the frequency and percentage of NMSC vs melanocytic histologic diagnoses and frequency and percentage of melanocytic proliferations classified according to the Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis (MPATH-Dx) reporting schema (Lott et al., 2018). 289 original histopathology reports were independently reviewed and classified into the MPATH-Dx system by two dermatologists and any cases with disagreements were

reviewed in conjunction to reach consensus. This NLP system yielded a PPV of 82.4%, sensitivity of 81.7% and F1 score of 82.0%. Eide et al used NLP to validate NMSC claims-data cases at a large health care system provider and its affiliated health maintenance organization (Eide et al., 2012). They set out to define NMSC case volume only and did not report any feature extraction from their NLP pipeline. A comparison of 909 electronic pathology reports to the NLP pipeline showed a sensitivity of 98.3%, specificity of 99.6%, negative predictive value (NPV) of 99.6% and PPV of 98.2% for this task. The performance of both these different pipelines reflect how performance can vary depending on the complexity of the intended system task – calculating incidence is a much simpler task compared to pathological diagnosis. The pipeline was designed with the intended purpose of increasing the accuracy and to enrich data contained in cancer registries. The remit of the JAPE rules created was therefore quite broad which inherently means a narrower focus of system task on the same validation set would perform better.

One disadvantage with rule-based NLP is the significant time investment in developing and validating such a pipeline. Building rules, testing, refining and retesting is a significant undertaking, with the potential for a large amount of further work if crucial aspects are either not included, or poorly thought out from the beginning. Therefore, there has been growing interest in the use of ML, instead of rule-based NLP, for information extraction. ML can be broadly categorised into supervised, unsupervised, reinforcement learning and deep learning. Supervised ML encompasses a model based on a dataset with labelled examples that can be used to solve a classification problem, unsupervised ML is based on a dataset with no labelled examples and reinforcement learning is a branch of artificial intelligence (AI) concerned with the generation of models that aim to maximize the receipt of rewards from an environment by learning to perform specific actions (Kuo et al., 2021). ML can be used in NLP for the purposes

of classification (group instances into predefined categories), clustering (group instances into undefined categories) and regression (predict numeric variables) (Harrison & Sidey-Gibbons, 2021). However, other authors have commented how ML algorithms have not been able to demonstrate superior performance in comparison to the rule-based techniques as described in this study, are poorly reported and raise concerns about interpretability and external generalisability (Harrison & Sidey-Gibbons, 2021).

6.4.1 Strengths and limitations

A significant strength of this study is that a comprehensive NLP algorithm has been developed for NER on 134 features from 139 possible entities in BCC histopathology reports. The study additionally went beyond basic NER since it also captured entity attributes – enriching quality of the data collected. Therefore, a vast amount of data can be extracted from a single histopathology report. The diagnoses within these reports are linked to Unified Medical Language System terminologies mapped to Concept Unique Identifier codes. This platform can be mapped to external coding vocabularies such as the International Classification of Diseases (ICD), Medical Subject Headings (MeSH) and Systematised Nomenclature of Medicine Clinical Terms (SNOMED-CT). This approach, which uses more than 600 dictionaries available within UMLS, provides broader term coverage and flexible interoperability compared to relying solely on a single vocabulary such as SNOMED-CT or OMOP. Nonetheless, in view of the NHS-wide move toward SNOMED-CT adoption, our pipeline retains the capacity for direct SNOMED-CT mapping and remains compatible with other widely used standards such as the OMOP Common Data Model. An additional design consideration pertains to multiple CUIs that can occasionally be associated with a single term; to address this, our system applies a practical method that either prioritizes SNOMED-CT–

inherited codes or selects the first alphanumeric CUI. This ensures maximal term capture without sacrificing semantic precision.

As with any early stage innovation there are some limitations. Data for training and validation were only obtained from one health board in a single country and therefore the external generalisability of the current pipeline to other hospitals in the UK or differing health services around the world may be limited. The pipeline needs to be iteratively re-tested and re-developed on larger internal and external datasets to enable future iterations to exhibit improvements in accuracy and allow validation across differing healthcare settings.

6.5 Conclusion

This project is novel within the field of skin oncology in the UK. The information extracted with this system such as tumour subtype, prognostic factors and margin status is often missing from routinely collected data. The algorithm has the promise to bridge this data gap enabling further skin cancer research opportunities and in clinical practice to record patient information in a structured manner. Future work will require large scale external validation of this system with blinded clinician assessment on a gold standard corpus with high inter-annotator agreement.

Chapter Seven:

Can natural language processing be used to automate a web-based model of care and modernise skin cancer multidisciplinary team meetings?

7.1 Introduction

MDTs have been shown to be an integral component in the management of skin cancer. However, there is mounting evidence to support reform in how they operate. The skin MDT has been shown to be costly when compared to other specialities as well as poorly attended, with only 26.0% quorate by membership and 69.0% quorate by meeting frequency (Ali, Dobbs, et al., 2021). With an estimated 50,000 shortfall in NHS clinical staff in England reported as of 2021, this is unlikely to improve (British Medical Association, 2024). The skin MDT is also unique in that the incidence of skin cancer is rising faster than that of any other malignancy (Cancer Research UK, 2024-c, 2024-d). Furthermore, the remit of the skin MDT is expanding. Historically, UK guidelines have recommended that all cases of high-risk SCC and MM were discussed at the SSMDT but omitted any recommendations on referral for BCC (Keohane et al., 2021; Nasr et al., 2021; National Institute for Health and Care Excellence, 2015). This has now changed, with the most up to date BAD UK BCC guidelines highlighting the pivotal role of the MDT in the management of high-risk BCC. Given these new recommendations, caseloads of LSMDTs and SSMDTs are only set to rise. Urgent solutions are therefore required to address the workload of the skin MDT. In addition to these challenges the skin cancer community has identified a number of potential areas for MDT improvement (Ali, Dobbs, Jovic, Hutchings, et al., 2023). One specific area was the need for protocolised treatment pathways, reflecting the mandate for protocolised streaming at a national level, with guidance on streamlining according to clinical complexity issued by NHS England and NHS Improvement in 2020 (NHS England and NHS Improvement, 2020).

Innovative solutions to these problems are needed. Since the COVID-19 pandemic the concept of the virtual MDT has expanded, with many MDT meetings moving to an online teleconferencing format. In a recent study this move was shown to maintain or improve

standards in the domains of 1) communication, chairing and decision making, 2) training, clinical trials recruitment and audit, 3) data security and patient confidentiality when compared to a face to face MDT (Ali, Dobbs, Mohamedbhai, et al., 2023). It would also be expected to improve attendance of all MDT members, allowing them to be involved without having to travel between sites for face-to-face meetings.

The effectiveness of the virtual MDT could be further enhanced by the inclusion of novel technologies. This could reduce the burden on skin cancer MDTs by facilitating the protocolisation of treatment pathways and supporting management decisions for 'simple' cases. The nascent fields of data science, AI and NLP all represent huge opportunities. NLP can be defined as a set of techniques used to convert written text into interpretable datasets through either rule-based or machine learning models (Harrison & Sidey-Gibbons, 2021). Using NLP to extract surgical outcomes from EHR is accelerating across disciplines and clinical outcomes research, aiming to improve outcomes, increase safety and aid service planning (Mellia et al., 2021). In the previous chapter an automated clinical text extraction system was developed and validated that can accurately extract pathological data from BCC histopathology reports (Ali et al., 2022). Use of this platform has demonstrated the feasibility of generalising a validated NLP pipeline to new data within a web application framework (Ali, Dobbs, Jovic, Strafford, et al., 2023). The objective in the current study was to use NLP techniques to harness the power of big data and build a novel web-based platform capable of transforming MDT working in the UK. The primary aim of this study was to validate an NLP based platform to automate evidence-based decisions in a CDSS capable of multiclass classification aligned with national guidelines for skin cancer care, with BCC as a use case – the vSMDT. We envisage the vSMDT being deployed before the main MDT meeting, overseen by the MDT chair and coordinator to apply protocolised streaming of cases. By filtering out “simple” or routine BCC cases in

advance, the main MDT would then have the capacity to dedicate more focused time on the "complex" cases requiring deeper multidisciplinary discussion. Through this mechanism of automated decision-making and protocolised streaming, the vSMDT aims to reduce the mounting workload on skin cancer MDTs while maintaining or even enhancing the quality of patient care.

7.2 Methods

7.2.1 Study design

A multi-centre (Morriston Hospital, Singleton Hospital and Neath Port Talbot Hospital, Wales, UK), pan-speciality, consecutive retrospective analysis of patients with a diagnosis of BCC over a 6-month period from 01/03/2021 to 20/09/2021 was undertaken. Lesions were examined by a consultant histopathologist using the bread loafing cross-section technique (Abide et al., 1984).

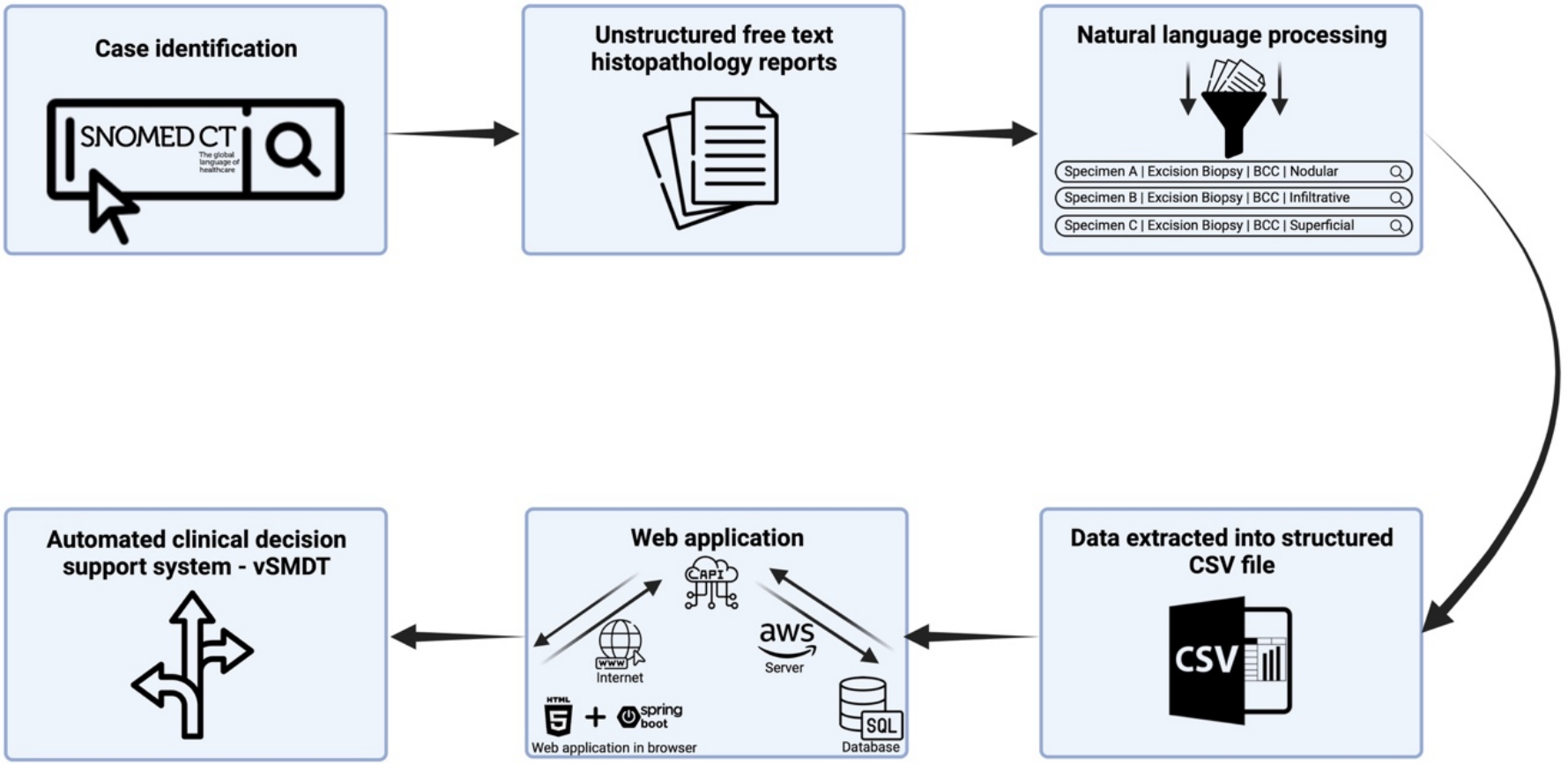
7.2.2 Case identification, inclusion and exclusion criteria

Cases were retrospectively identified from InterSystems TrakCare Lab Laboratory Information Management System (InterSystems TrakCare Lab, Cambridge, Massachusetts, USA), using SNOMED reference term (RT) codes for BCC. SNOMED is a logic-based health care terminology used in an EHR. It is a consistent vocabulary for recording patient clinical information (NHS Digital, 2024). All those patients with a SNOMED RT code for BCC during the study period who were managed by surgery (either non-definitive diagnostic sampling biopsy (punch biopsy, incision biopsy, shave biopsy or curettage) or surgical excision using a pre-determined margin) were included. CSV text files were generated from the respective canonical subheadings of the histopathology report.

7.2.3 General framework

The vSMDT consists of two main parts: the extraction of the right text from the pathology report using the NLP algorithm and then producing the right recommendation within a web application given the data extracted from the histopathology report. This process is schematically represented in Figure 27.

Figure 27: Schematic representation of the vSMDT.



7.2.4 NLP algorithm

A detailed description and validation of the NLP algorithm was described in the previous chapter. To recapitulate, the GATE framework was used to build an NLP information extraction system using rule-based techniques. This was validated on previously unseen, de-identified and pseudonymised BCC histopathological reports at the same institution as the current study. The mean precision, recall and F1 score were 86.0% (95% CI, 75.1-96.9), 84.2% (95% CI 72.8-96.1) and 84.5% (95% CI 73.0-95.1) respectively. Notably, the pipeline's performance closely approaches that of the gold standard clinician review, with a mean F1 score difference of 7.9%, compared to 15.5%.

7.2.5 Web application

Java™ Spring Boot (VMware Incorporated, Palo Alto, California, USA) was used to develop a web application hosted on Amazon Web Services (Amazon.com Incorporated, Seattle, Washington, USA) in EC2. Respective CSV files generated from the NLP pipeline were imported into a relational database management system (RDMS). MySQL Workbench and MySQL Community Server (Oracle Corporation, Austin, Texas, USA) was used as the platform for the RDMS. Within the application framework, an application programming interfaces (APIs) was developed to automate a CDSS that was mapped to and adapted from the recommendations for management following primary treatment in the 2021 BAD guidelines for the management of adults with BCC (Table 25).

Table 25: vSMDT recommendations for histopathology report outcome following primary surgical treatment.

Histopathology report outcome	Recommendation
Completely excised single BCC	No follow-up
Incompletely excised BCC lesion (peripheral margin, deep margin or both)	Offer re-excision. If declined follow-up 6 monthly for 2 years
Multiple BCCs	Follow-up 6 monthly for 5 years
Recurrent BCC	Follow-up 6 monthly for 5 years
Supplemental BCC peripheral margin = positive	Offer re-excision. If declined follow-up 6 monthly for 2 years
Supplemental BCC peripheral margin = negative	No-follow-up
Supplemental BCC deep margin = positive	Offer re-excision. If declined follow-up 6 monthly for 2 years
Supplemental BCC deep margin = negative	No follow-up
Punch biopsy, incision biopsy, shave biopsy or curettage AND cancer type = BCC	Further excisional surgery, destructive surgical or non-surgical technique recommended to obtain oncological clearance
Any margin outcome (complete/incomplete) for a benign lesion or non-specific	No follow-up
If cancer type = any other cancer, other in situ or other intermediate	Other cancerous, in situ or intermediate lesion. Review of histopathology free text required to guide management

A letter template is then generated ready to be sent to the patient to communicate the diagnosis(es) and provide recommendations for the next course of action. The letter included a BAD patient information sheet (PIS) that explained the diagnosis, as well as a Melanoma UK PIS on self-examination. The outcome from the CDSS model was binary (i.e. 1/0) rather than using probabilities of the predicted class. A training set of 100 histopathology reports was used to develop the API with 100.0% accuracy to ensure that the correct predictions were made given the data inputted prior to validation.

7.2.6 Variables

Peri-operative tumour factors (primary versus recurrent) and surgical factors (excision type, diagnosis and margin status) were recorded.

7.2.7 Outcomes

The primary endpoints were histological margin status. Margin status was defined as either clear (>0mm) or involved (0mm) in line with current RCPath histopathological reporting standards for primary BCC (Slater & Barrett, 2019-b).

7.2.8 Statistical analysis

Validation was undertaken retrospectively by two independent and blinded expert clinicians acting as 'the MDT'. These two clinicians decided on the management outcome after surgery for each patient based on guidelines from the BAD and used by the hospital MDT (Table 23).

These were taken as the gold standard to which the performance of the vSMDT decision tool was compared. Disagreements in management outcomes were resolved by case discussion until a consensus was reached. Cohen's kappa was used to assess inter-observer agreement and ensure consistency of decision making between cases. Statistical analysis was undertaken in R version 4.1.1 (R Core Team, R Foundation for Statistical Computing, Vienna, Austria). Sensitivity, specificity, PPV and negative predictive value (NPV) was calculated to evaluate the performance of the multiclass CDSS. The overall system performance was summarised using the Yardstick package in R. Micro-averaging and macro-averaging are two ways to combine the results of multiple binary classification models into a single multiclass classification model (Vaughan, 2024). In micro averaging, the metric is calculated for each individual class and then averaged across all classes. In macro-averaging, the metric is

calculated for each individual class and then the unweighted average is taken across all classes. This can be problematic in imbalanced datasets, because it can result in the model being overly influenced by the majority class. Given the likelihood for class distribution imbalance a micro-averaging approach was adopted to ensure that each class was given equal consideration in the final model. A one-versus-rest method was used to evaluate each individual model within the multi-classifier by comparing each class against all the others at the same time using the Classification And REgression Training (CARET) package in R. An *a priori* sample size calculation was undertaken. 40 cases and 760 controls were required on the assumption of a sensitivity of 0.95, significance level of 0.95, desired precision (delta) of 0.10 and prevalence of 0.05. Controls could be the other patients not meeting the singular reference value being tested at that time. The controls can then rotate with case patients when the other reference value is being tested. A disease prevalence of 0.05 was assumed for the reference values, based on previous work that established the proportion of incomplete excision in the same study population as 5.5% (Ali et al., 2023). It was deemed that there would be a similar prevalence amongst the reference values. A complete-case analysis approach was used to handle any missing data. $p < 0.05$ was deemed statistically significant.

7.3 Results

There were 893 patients (1045 lesions) who had a SNOMED RT diagnosis of BCC within their histopathology report. The mean patient age was 74 years (26-102). From the 1045 observations, there were only 10 disagreements, yielding an agreement rate of 99.2% between the two independent clinicians. The Cohen's Kappa coefficient, computed assuming an equal probability of random agreement for the two outcomes, was found to be 0.99, indicating a very high level of agreement that significantly surpasses chance. These results highlight the quality and reliability of the reference standard.

There were 745 instances where the reference group and prediction matched when the no follow-up criteria were met (Table 26).

Table 26: Baseline confusion matrix of reference (clinician) and prediction recommendations (vSMDT).

	Reference				
Prediction	Follow-up 6 monthly for 5 years	Further excisional surgery, destructive surgical or non-surgical technique recommended to obtain oncological clearance	No follow-up	Offer re-excision. If declined follow-up 6 monthly for 2 years	Other cancerous, in situ or intermediate lesion. Review of histopathology free text required to guide management
Follow-up 6 monthly for 5 years	138	7	12	11	3
Further excisional surgery, destructive surgical or non-surgical technique recommended to obtain oncological clearance	0	17	1	1	0
No follow-up	8	7	745	20	4
Offer re-excision. If declined follow-up 6 monthly for 2 years	0	0	5	12	0
Other cancerous, in situ or intermediate lesion. Review of histopathology free text required to guide management	0	0	7	1	46

This demonstrated a strong concordance between the vSMDT and human decision-making processes in certain scenarios. When considering the entire dataset, the vSMDT showed promising performance. The overall performance of producing the same recommendation as a clinician given the histopathology report was good with an accuracy of 0.92. These results imply that the vSMDT was highly capable of correctly interpreting and making decisions based on the histopathology reports across the vast majority of cases. Overall summary statistics are shown in Table 27a and provide an in-depth overview of these performance metrics.

Table 27: Performance of (a) overall and (b-f) individual recommendations by the vSMDT.

	Prediction					
Statistics	(a) Overall [a]	(b) Follow-up 6 monthly for 5 years	(c) Further excisional surgery, destructive surgical or non-surgical technique recommended to obtain oncological clearance	(d) No follow-up	(e) Offer re-excision. If declined follow-up 6 monthly for 2 years	(f) Other cancerous, in situ or intermediate lesion. Review of histopathology free text required to guide management
Accuracy	0.92	0.99	0.99	0.98	0.97	0.99
Sensitivity	0.92	0.95	0.55	0.97	0.27	0.87
Specificity	1.00	1.00	1.00	1.00	1.00	1.00
PPV	1.00	1.00	1.00	1.00	1.00	1.00
NPV	0.98	0.99	0.99	0.92	0.97	0.99

NPV; negative predictive value, PPV; positive predictive value. [a] Overall performance calculated by micro-averaging pairwise comparisons.

Upon further examination, it was observed that the performance of the vSMDT varied depending on the specific type of recommendation being issued. Table 27b-f presents a detailed breakdown of the vSMDT's performance per recommendation type.

The vSMDT demonstrated high specificity across all recommendation categories, implying that the issued recommendations were typically correct. Specificity values stood at 1.00 across all categories (Table 27b-f), revealing the system's strong performance in avoiding false positives. On the other hand, the system's sensitivity, reflecting its ability to correctly identify cases necessitating a specific recommendation, was lower. Notably, two recommendations stood out with markedly lower sensitivity scores. The recommendation for "Further excisional surgery, destructive surgical or non-surgical technique recommended to obtain oncological clearance" exhibited a sensitivity of 0.55 (Table 27c). Similarly, the recommendation "Offer re-excision. If declined follow-up 6 monthly for 2 years" had a sensitivity score of 0.27 (Table 27e).

Interestingly, of the 22 patients who met the 2021 BAD guideline recommendation of *"following discussion at an MDT, offer further standard surgical re-excision to adults with excised high-risk BCC with involved histological margin unless there is a contraindication"* only 22.7% were actually discussed at a Local Skin Cancer MDT or Specialist Skin Cancer MDT. This finding exposes potential gaps in the application of guideline recommendations in the real-world clinical context, underscoring the potential utility of the vSMDT in supporting clinical decision-making.

7.4 Discussion

This study validates a fully automated, virtual, web-based service model to host the skin MDT. High specificity, as demonstrated here, would make the system suitable for use as a 'diagnostic test' in the context of a vSMDT. Highly specific tests are used for ruling in a disease, as it rarely misclassifies those without a disease, which is desirable from a CDSS. However, given

the scope for potential downstream clinical error with even the highest specificity the vSMDT lends itself to being an adjunct to MDT decision making and facilitating protocolised streaming with a 'human in the loop' as opposed to a fully autonomous system. This aligns with the Topol Review 'Preparing the healthcare workforce to deliver the digital future' which anticipates that genomics, digital medicine, AI and robotics will not replace healthcare professionals, but will enhance them, giving them more time to care for patients (Topol, 2019).

Arguably one of the most high-profile use cases of AI as a method of automation in healthcare is the international validation of an AI system for breast cancer screening (McKinney et al., 2020). McKinney et al found that they were able to develop a system capable of surpassing human experts in breast cancer prediction. In their simulation, the AI system participated in the double-reading process that is used in the UK and found that the AI system maintained non-inferior performance and reduced the workload of the second reader by 88.0%. The NICE Medtech innovation briefing on 'Artificial intelligence in mammography' recognises the value of how AI technologies may improve performance and save time in interpreting mammograms (National Institute for Health and Care Excellence, 2021). NHS trusts are beginning to adopt this technology with the AI algorithm Transpara (ScreenPoint Medical, Nijmegen, Netherlands), which uses deep learning convolutional neural networks, feature classifiers and image analysis algorithms, in use in 4 NHS trusts as at January 2021 (National Institute for Health and Care Excellence, 2021). At present the NHS Breast Screening Programme (BSP) uses a system of 2 readers, and arbitration to interpret mammograms (Public Health England, 2015). However, it is currently facing a shortage of qualified people, especially radiologists. AI technologies in this setting could reduce workloads by replacing 1 of the 2 readers, or by performing triage according to the likelihood of an image being malignant. It could also be used to automatically classify images showing a low likelihood of malignancy as normal, and

remove these from the images to be reviewed. Similar to the use of AI mammography in breast screening, the vSMDT could deliver comparable benefits and efficiency savings for skin cancer services. In particular vSMDT mediated protocolised streaming of 'low risk' cases could take place in advance of the main face-to-face skin MDT. This would go some way to addressing the issues identified by the CRUK report into 'Improving the Effectiveness of Multidisciplinary Team Meetings in Cancer Services' which demonstrates that there is not enough time to discuss complex patients, attendance is not optimal, the right information is often not used to inform discussions and that MDTs are unable to fulfil their secondary roles in data validation, audit and education.

Current 'black-box' AI models may offer superior performance but model outcomes are not easily explained, causing some to question their suitability for use in high-stakes scenarios such as medicine (Ghassemi et al., 2021). The more transparent, 'glass-box' nature of explainable AI helps clinicians and patients understand and trust the behaviour of the model and more easily allows for debugging and improvements to model performance. In this study a rule-based method was used to develop the NLP algorithm. This has the benefit of the rules contained in the model being precise and easy to customise, often simple to implement whilst being interpretable and explainable. These factors make rule-based methods eminently suitable to the biomedical domain. However, rules can become incredibly complex, order matters and maintenance is complicated. Rules are unsuitable for some error types e.g. semantic errors and require language specific knowledge. Despite these limitations, the performance of the rule-based NLP pipeline was maintained when an API was designed to use these data to predict treatment pathways.

With a highly specific model like the vSMDT, one would usually expect a trade-off with a lower sensitivity. However, two rules in the model displayed markedly lower sensitivity compared to the other three. In the NLP model that underpins the data extraction of the vSMDT there were 80 separate gazetteers and 445 JAPE rule files in total. Isolating the specific data extraction error here that gets incorrectly transduced into the web application and subsequently contributes to poor sensitivity is likely to be the best strategy to increase performance in this area. The initial NLP pipeline was designed with the aim of improving the quality of routinely collected data for research as well as supporting the vSMDT. A scaled back NLP pipeline with a smaller number of rules designed to extract the minimum amount of data needed by the vSMDT to make a clinical recommendation would simplify 'explainability' and improve the transparency of the decision-making model. General approaches to increasing NLP pipeline performance by increasing the volume and quality of training data is also valid here. The API produced a binary outcome in the clinical recommendation outputted. Instead, using probabilities of the predicted class on a scale e.g. 0-100 may be a more nuanced approach and would allow clinicians to define an optimal cut-point value using ROC curve analysis.

An alternative strategy to extracting data, converting this into a structured database and creating a series of rules that use this data to predict MDT recommendations (binary or scale) could use ML to directly predict clinical decision pathway(s) from free text histopathology reports e.g. text classification. This approach could also easily be combined with historical data to improve predictions. A rapidly evolving area of NLP is the evolution of deep learning models which can be used for this purpose. The two particular types of structure of neural networks used in deep learning are CNNs and RNNs with the latter being used for analysis of sequential data such as text. Medical research output in the field of deep learning has predominantly focussed on RNNs in imaging rather than CNNs and text analysis, where the images themselves are

more easily de-identified and able to be shared as a public dataset for use amongst medical researchers around the world. This is at present probably the most significant barrier to creating high quality, large volume training datasets that are necessary for creating and deploying deep learning models in healthcare. However, a rapidly evolving area of research that has the potential to change this is the transformer. A transformer represents a new type of AI language model that does not use the traditional methods of recurrent or convolutional neural networks (Vaswani et al., 2017). Instead, it uses attention mechanisms, which are a different way of processing information. This makes the transformer simpler and more efficient than other models, and it can be trained faster with fewer resources. Other ML algorithms used in skin cancer MDTs have not been able to demonstrate superior performance in comparison to rule-based techniques. This is illustrated by Andrew et al who developed a supervised machine-learning algorithm utilising a decision-tree model trained on a routinely collected SSMDT dataset from a single institution to predict MDT decisions for Mohs micrographic surgery (MMS) vs conventional surgery or radiotherapy (Andrew et al., 2022). Their model was only able to triage 45.1% of patients to a treatment plan.

It is acknowledged that the external generalisability of the CDSS to other tumour types beyond BCC remains a limitation of the current study, especially when considering more complex surgical tumour types such as pancreatic cancer, oesophageal cancer, sarcoma or even metastatic disease. However, even within these more complex tumour types, there exists a subset of cases that are simpler and could be effectively managed with a protocolised approach. The CDSS could play a crucial role in standardising treatment recommendations for these cases, allowing MDTs to focus more of their time and resources on the truly complex cases that require more nuanced discussion and decision-making. The quality of histopathological reporting can also significantly impact the external generalisability of the model. As

highlighted by Barrett & Barrett, compliance with minimum dataset reporting in NMSC tends to be lower compared to melanoma, indicating potential variances in reporting structure (Barrett & Barrett, 2015). Paradoxically, there could be better performance when the tool is applied to other tumour types with perceived higher morbidity and mortality, as more complex cases often warrant more thorough histopathological reporting, providing more consistent and comprehensive data for the model.

While the vSMDT demonstrates significant potential for improving efficiency in diagnosing and treating BCC, it has certain limitations. A key restriction is its current dependence on the information contained in histopathology reports or EHRs. This means the system's efficacy is linked to the comprehensive nature and detail within these reports. In its current iteration, the system may not fully consider the wider clinical context of a patient, particularly previous diagnoses like melanoma. The decision-making algorithm, while robust for BCC, does not yet comprehensively integrate patient's prior medical history to influence its recommendations. For instance, even in cases of low-risk lesions, patients with a history of melanoma may require continued surveillance and should not be prematurely discharged. Moving forward, the aim is to enhance the sophistication and utility of the system by incorporating algorithms that can parse through and learn from a broader range of data sources. This would include information on a patient’s prior diagnoses and other relevant clinical details. By doing so, the system will offer more individualised and contextually appropriate recommendations that consider each patient's unique health status and history. Such an enhancement, could elevate the decision support system's role in delivering personalised, efficient, and safe patient care.

A CDSS such as the vSMDT that employs a human-in-the-loop approach can serve as a powerful adjunct to decision-making processes and protocolised streaming in MDTs.

Nevertheless, such systems, especially if classified as medical devices under specific criteria, must adhere to stringent regulations and standards. According to the UK Medical Device Regulations 2002 (UK MDR 2002), a medical device is defined as "any instrument, apparatus, appliance, software, material or other article, whether used alone or in combination, together with any accessories, including the software intended by its manufacturer to be used specifically for diagnosis or therapeutic purposes or both and necessary for its proper application…" (The National Archives, 2024). If a CDSS aligns with these criteria, it must comply with the regulations outlined by the Medicines and Healthcare products Regulatory Agency (MHRA) in the UK. The path to compliance starts at the conceptualisation stage, involving the identification of the appropriate legislation applying to the device. This process demands careful consideration and thorough documentation, encompassing details such as product specifications, evidence of safety and effectiveness, results of clinical trials, and information about the manufacturing process. Subsequently, the medical device requires an assessment, the nature of which hinges on the associated risk level. Low-risk devices (class 1) can opt for self-certification, while higher-risk devices (class 2a, 2b, and 3) necessitate an evaluation by an MHRA-approved body (Digital Regulations Innovation, 2024). The vSMDT with human involvement would likely be classified as a class I device. Class I devices typically includes non-invasive devices like a CDSS, which support, rather than dictate, clinical decisions and as such the vSMDT would fall into this category. This assessment guarantees that the device's benefits outweigh the minimised risks, with a successful evaluation culminating in the assignment of a UKCA mark, denoting compliance with UK MDR 2002. Integral to this process is the implementation of a quality management system (QMS) (Digital Regulations Innovation, 2024). Legally required for medical device development, a QMS delineates processes minimising the production, deployment, and surveillance risks of medical devices. A robust QMS presents structure for essential company processes revolving around

device safety and efficacy. The QMS should be certified to a recognised standard like ISO 13485, which ensures comprehensive document management, risk assessment, sign-off procedures, and decision records.

As outlined in chapter 1, implementation of the vSMDT mirrors the iterative processes observed in drug discovery, starting with retrospective proof-of-concept work (akin to preclinical testing) and progressing through prospective, real-world trials to confirm safety, efficacy, and scalability. The CDSS work in this chapter is based on retrospective validation and analogous to phases I–III in drug development. We envisage the vSMDT being deployed before the main MDT meeting, overseen by the MDT chair and coordinator, to automatically filter out "simple" or routine BCC cases. This protocolised streaming frees clinicians to focus on complex cases requiring deeper multidisciplinary input, thus optimising resources while maintaining or enhancing care quality. In the future the vSMDT should be designed to integrate with existing EHR systems, thereby minimising additional IT burdens and preserving a human-in-the-loop approach that keeps final decisions under expert clinical oversight. In terms of user experience (UX), MDT members across dermatology, surgery, pathology, and oncology should be engaged in iterative design and early usability testing, ensuring that the vSMDT's interface is clear, that clinical tasks map logically, and that the system seamlessly merges with existing data flows.

Whilst the vSMDT has shown promise in this initial retrospective assessment it is important to note that these findings should be considered preliminary and hypothesis-generating. The next essential step towards establishing the validity of the tool would be to apply it prospectively in an independent cohort, allowing for real-time evaluation of its performance and reliability. This

will provide a robust confirmation of the initial findings, and, if externally validated, could have significant implications for practice.

Patient and public involvement is now an established part of medical research study design. Whilst the focus this study was on validation, a future aspiration of the vSMDT is to include patient choice into the recommendations. Co-referencing of clinic letters in the NLP algorithm will endeavour to factor in this critical facet to aid decision making when planning after primary treatment.

7.5 Conclusion

A fully automated, virtual, web-based service model was validated to host the skin MDT with good system performance. This project is novel within the field of skin oncology and aligns with current NHS initiatives to digitally transform services. It is feasible that the vSMDT platform could be used to support clinical decision making during the skin MDT as a 'human in the loop' approach to aid protocolised streaming in the context of the skin MDT, similar to other high-profile use cases of AI in the NHS. Scott's parabola describes the phenomenon for technologic adoption in medicine, outlining the progression from a promising idea to ubiquity and later to obsolescence. Instead of this ill-fated rise and fall it is hoped that the vSMDTs real-world resource and efficiency savings will underpin the business case for its widespread adoption across the NHS and its integration into day-to-day clinical practice. Future work to externally validate this model in SCC and MM, where clinical guidelines are more complex will be key to realising its full potential.

Chapter Eight:

Using artificial intelligence to write patient clinic letters: a study using ChatGPT

8.1 Introduction

In the Chapter Seven, the discussion centred on the implementation of a CDSS for the vSMDT. This system efficiently delivers treatment recommendations post-primary surgery based on histopathology reports, generating personalised letters to communicate diagnoses and recommend actions. This chapter explores the potential integration of generative pre-trained transformer (GPT) based algorithms, such as ChatGPT, to directly generate patient letters based on CDSS outcomes. This shift from template-based letters to dynamically generated, patient-specific correspondence represents a potential significant advancement in medical communication. By combining the recommendation outcomes of the CDSS with NLP capabilities, the goal is to streamline patient communication and enhance the patient experience throughout the cancer diagnosis and treatment pathway.

The appropriate recording and communication of clinical information between clinicians and with their patients is of paramount importance. Over recent years there has been a much-needed drive to improve the information that is shared with patients, giving them "the information they want or need to know in a way they can understand" (General Medical Council, 2024). In the UK guidelines exist to outline the appropriate use of clinical letters in healthcare, covering issues such as the provision of clear and accurate information to patients, the standards for the management of health information and the regulation of personal data processing. In 2018 the Academy of Medical Royal Colleges (AoMRC) launched the 'Please, Write to Me' campaign (*Please, Write to Me. Writing Outpatient Clinic Letters to Patients Guidance.*, 2018), where they endorsed the practice of writing all clinic letters directly to the patient and copying them to other healthcare professionals involved in their care (Academy of Medical Royal Colleges, 2018).

The preparation of clinical letters can be time-consuming for clinicians and secretarial staff, especially at a time when health services are so stretched. They also risk including errors or omissions and due to their non-standardised nature can be difficult for patients to read and understand. As part of the drive to write to patients it is also important that they are able to understand the information included in clinical letters. However, a significant portion of adults in the UK, as estimated by the National Literacy Trust, possess literacy skills below those expected of a student in Year 6 (age 11) (National Literacy Trust, 2024). Similarly, data from the US suggests that 52.0% of the population has literacy abilities at or below the level of a US 4th or 5th grader (age 10-11) (Wylie Communications, 2024). In the US, it is recommended that patient-facing health literature be written at or below a 6th grade level (age 11-12) (US Department of Health and Human Services Office of Disease Prevention and Health Promotion, 2010), although there are no specific guidelines on this in the UK. Despite, these recommendations, previous published work has demonstrated that letters to patients are often written with poor readability and adherence to plain English writing standards (Drury et al., 2021).

Traditionally letters are dictated by the clinician involved in the patient's care and then transcribed. While there has been an increase in the use of letter templates and voice recognition systems, with the aim of improving efficiency, novel technologies such as NLP and AI have the power to revolutionise this area of practice. NLP algorithms are designed to recognise and understand the structure and meaning of human language, including words, phrases, and sentences, and can be trained on large datasets of text to learn how to identify and extract important information, classify texts according to their content or purpose and generate responses that are appropriate and coherent (Jurafsky & Martin, 2008). In medicine, NLP is often used for tasks such as information extraction from electronic health records, text

summarization of clinical reports, and language translation (Harrison & Sidey-Gibbons, 2021). AI generated clinical letters that use NLP have the potential to improve efficiency, consistency, accuracy, patient satisfaction, and cost savings in a healthcare system.

OpenAI's ChatGPT chatbot was launched in November 2022 and utilises NLP technology to generate human-like text. The generative pre-trained transformer (GPT) language model is based on a transformer architecture, which allows it to process large amounts of text data and generate coherent text outputs by learning the relationships between input and output sequences (Vaswani et al., 2017). The GPT language model has been trained on large datasets of human language and is able to generate human-like text that is difficult to distinguish from text written by a human. The performance of the GPT language model has been evaluated in a number of studies and has generally been found to be very good at generating high-quality and coherent text outputs (Brown et al., 2020; Raffel et al., 2019; Vaswani et al., 2017). Despite no published performance data to date, it is suggested that ChatGPT may outperform previously developed NLP algorithms in the generation of medical clinic letters due to its state-of-the-art transformer-based technology. These attributes may enable ChatGPT to create letters of high quality that are easily comprehended by patients and accurately communicate complex medical concepts. In this study the early adoption and evaluation of ChatGPT was described to generate clinical letters to patients with limited clinical input. The aim of this study was to evaluate the readability, factual correctness and humanness of ChatGPT generated clinical letters to patients using the example of skin cancer as the most common human cancer.

8.2 Methods

A series of different clinical scenarios were created for patients who had either a BCC, SCC or MM awaiting imaging or histopathology results to be communicated to them following an

MDT outcome. To simulate how clinicians might use the ChatGPT (December 15 Version, OpenAI, San Francisco, California, USA) graphical user interrace (GUI) in the clinical environment shorthand instructions were created to input in the chatbot which were defined as ‘limited clinical input‘ considering the relative amount of natural free text one would otherwise be required to write or dictate to generate a clinical letter (Appendix 4). All letters were instructed to be written at a reading age of 11-12 years (US 6th grade) (US Department of Health and Human Services Office of Disease Prevention and Health Promotion, 2010). The capabilities of ChatGPT were evaluated by presenting it with a series of instructions of increasing complexity (Figures 28 to 32).

Figure 28: Example of a clinical letter informing a patient of a completely excised BCC.

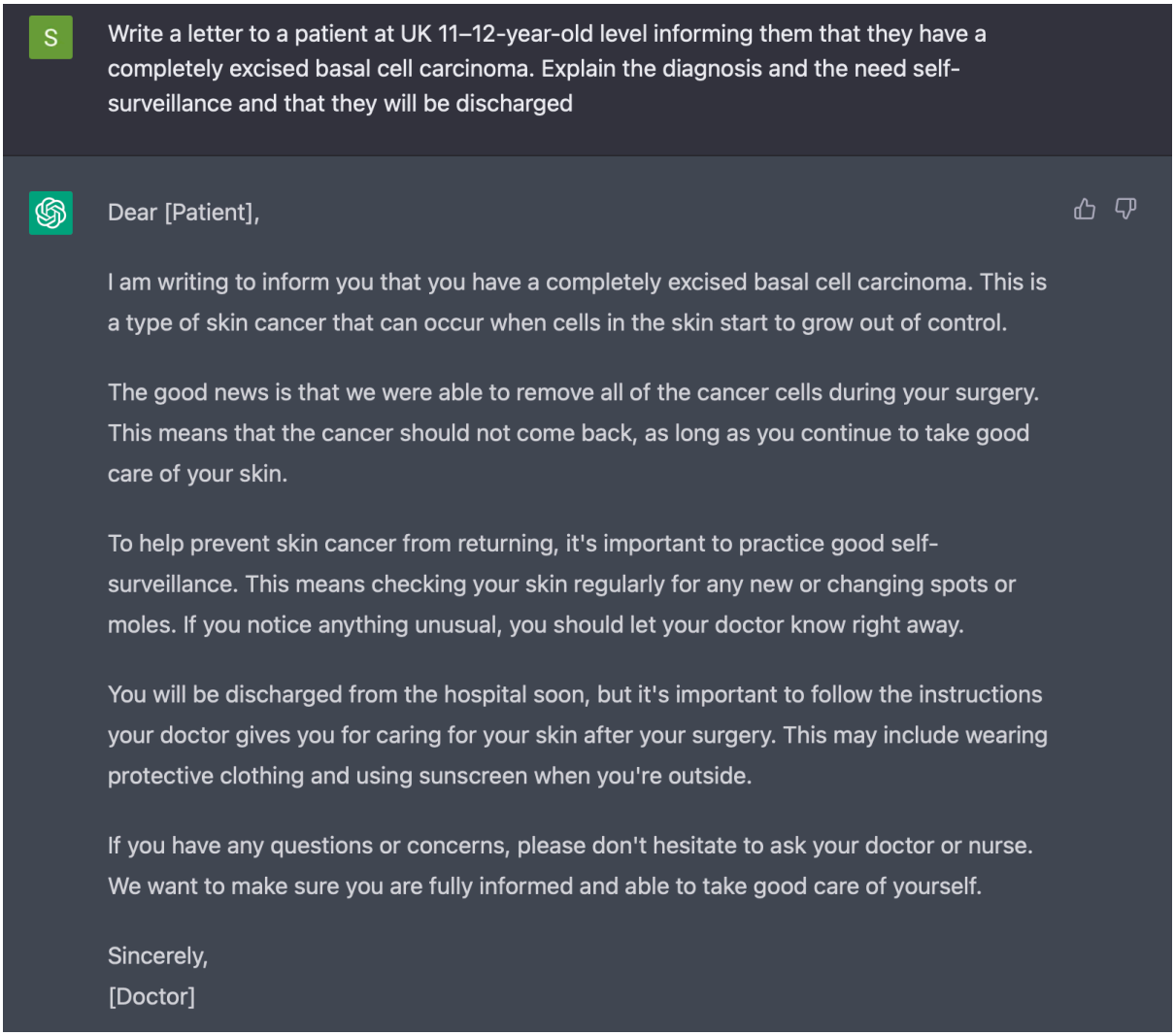


Figure 29: Example of a clinical letter informing a patient of an incompletely excised BCC.

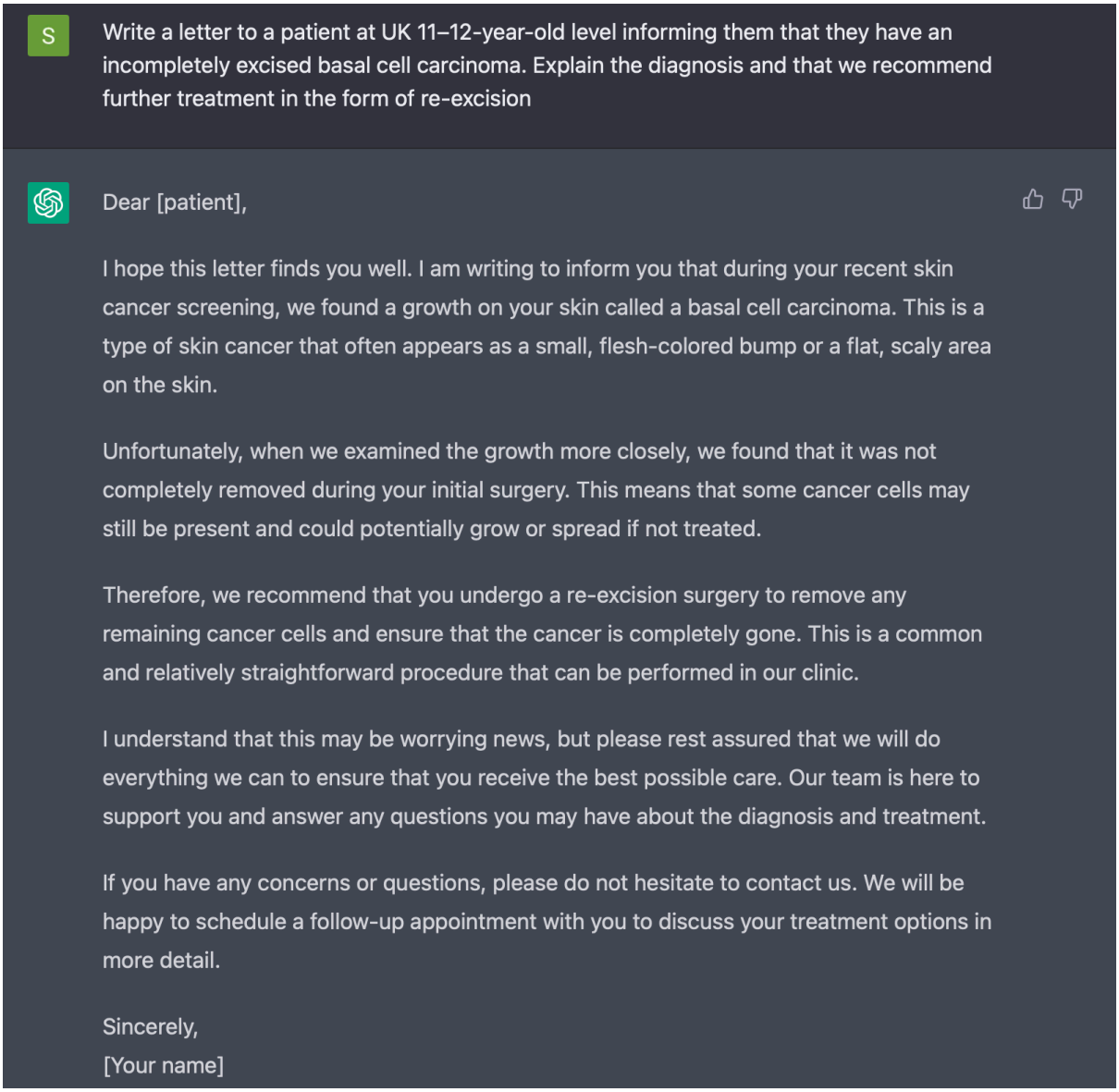


Figure 30: Example of a clinical letter informing a patient of an incompletely excised BCC and management of warfarin peri-operatively referencing British Society of Haematology guidelines.

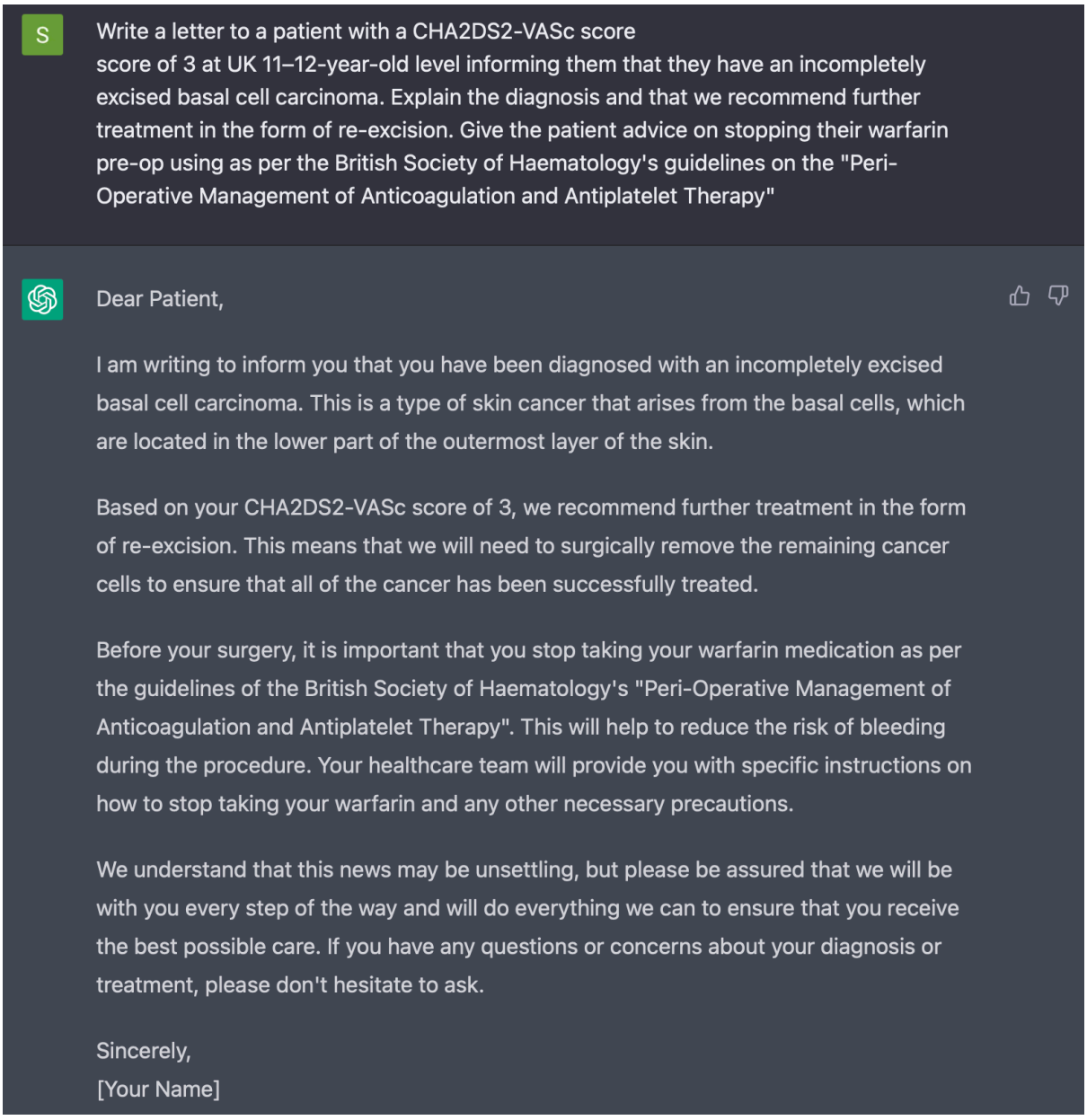


Figure 31: Example of a clinical letter informing a patient of an incompletely excised BCC and management of warfarin peri-operatively prompting ChatGPT to consult the literature and find the latest guidance.

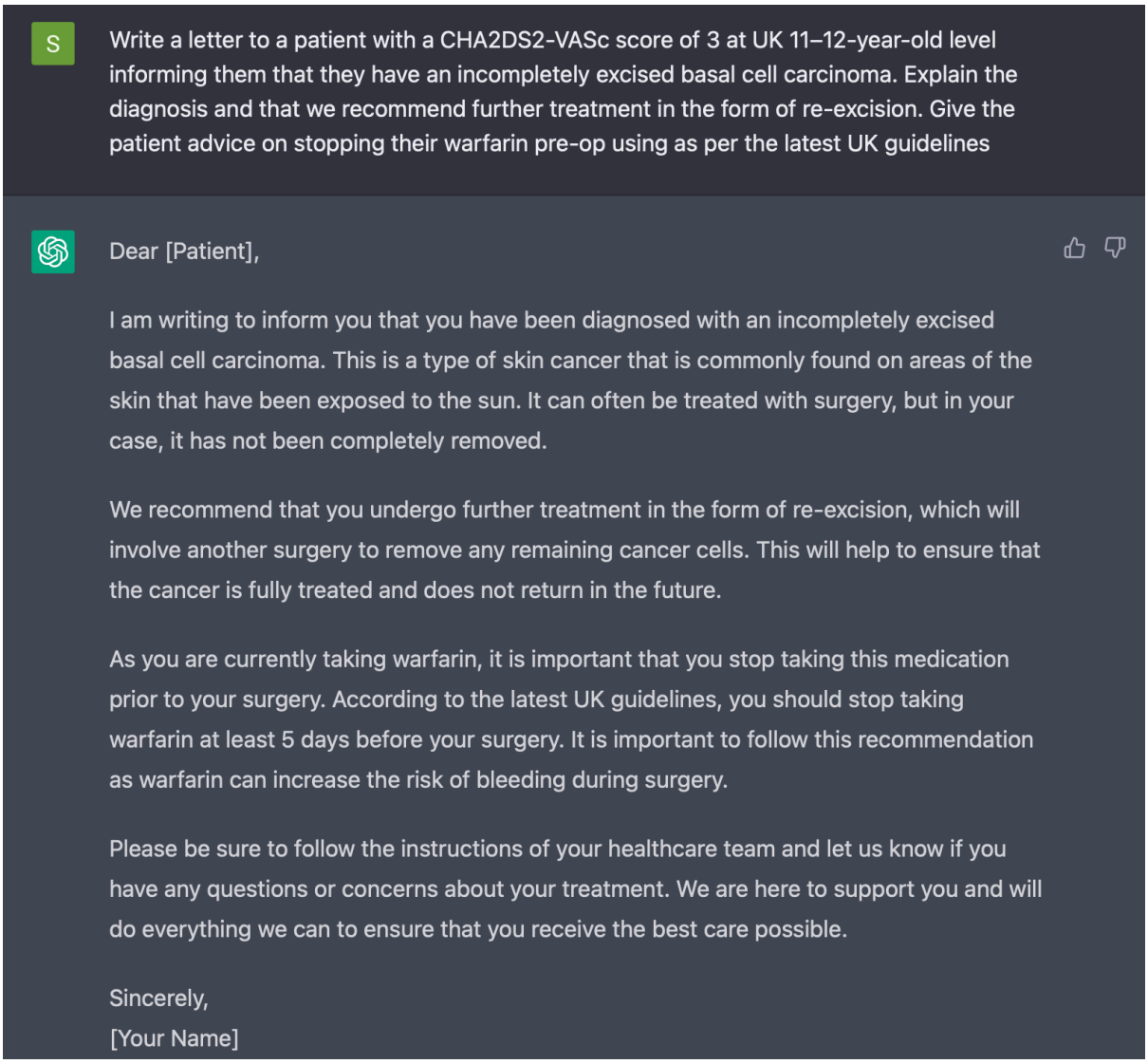
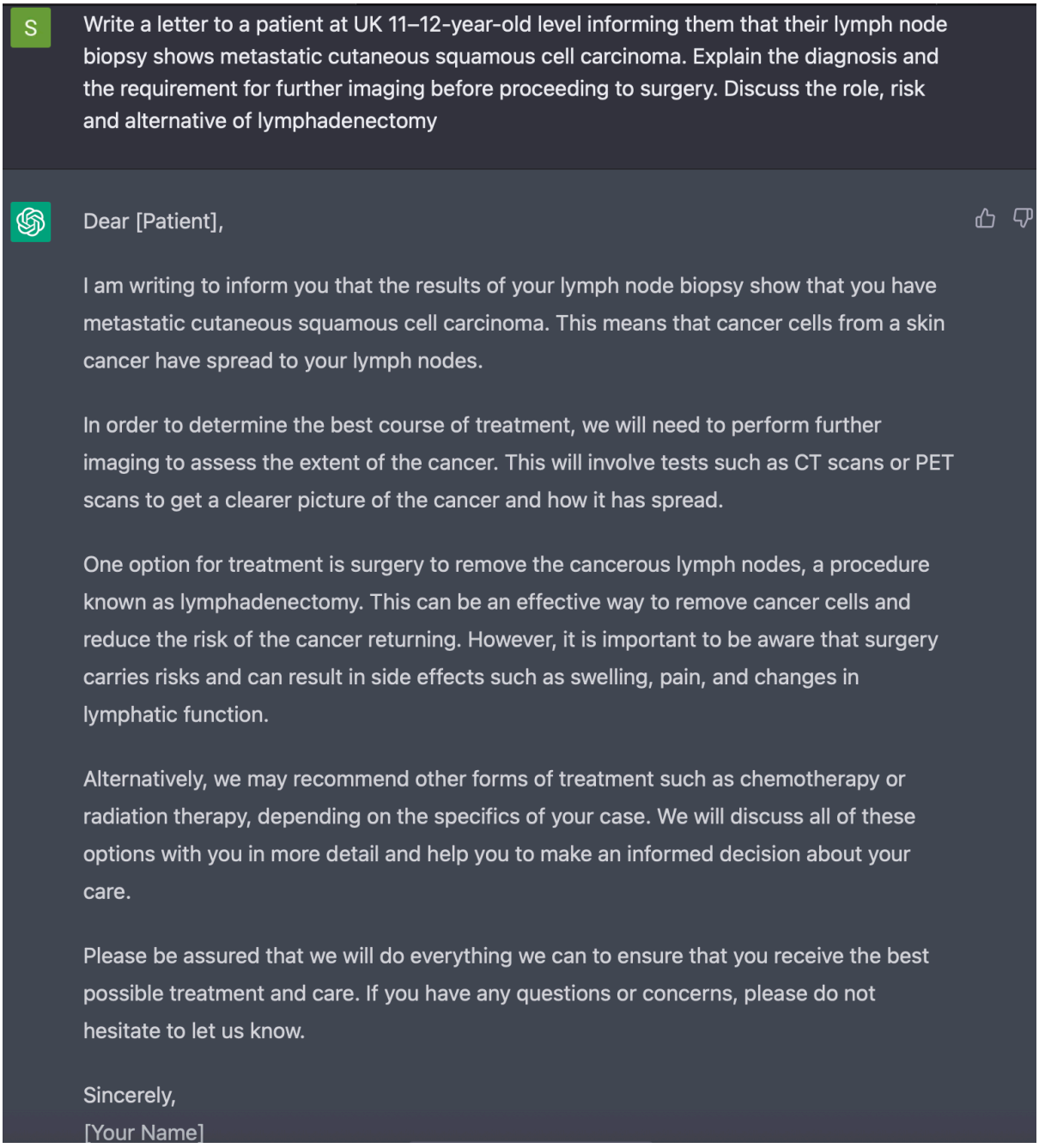


Figure 32: Example of a clinical letter informing a patient of a regionally metastatic SCC and explaining lympadenectomy.



These instructions ranged from simply following specific directions to using national guidelines and data from these guidelines to provide clinical advice in a clinical letter e.g. management of anticoagulation peri-operatively. Completely and incompletely excised BCC/SCC were modelled as well as patients with all stages of MM including those with positive sentinel node biopsy and distant metastatic disease. This allowed testing of the limits of ChatGPT's ability to process and respond to a variety of different types of input. After submitting these instructions, ChatGPT then generated a response in the form of a clinical letter to be issued to the patient. To evaluate the readability of letters generated by ChatGPT, these were analysed for reading difficulty using a variety of techniques, as described in Table 28.

Table 28: Techniques used to analyse the reading difficulty of ChatGPT-generated letters.

Test	Factors assessed
Automated Readability Index	Letters per word
Coleman-Liao Index	Letters per word
Flesch Reading Ease	Syllables per word
Flesch-Kincaid Grade Level	Syllables per word
Fry Grade Score	Sentence length
Gunning Fog Grade Score	Sentence length
Linsear Write Formula	Word difficulty
SMOG Grade Score	Sentence length

(Dale & Chall, 1948)

Some of these techniques involved the use of whole text to form an estimate of readability, while others involved sampling, where multiple random sections of text, usually 100 words in length, were analysed using a set of readability scores calculated with standardised formulas. These formulae were used with online calculators at https://readable.com to calculate the readability of the text. Readable™ is a website that enables users to input a piece of text and apply various readability formulae to assess its comprehensibility. This tool can be useful for

ensuring that written materials, such as patient-facing health literature, are easily understood by the intended audience. It has been found that in order to ensure complete understanding of health-related literature, it is advisable to utilise a combination of two or more readability formulas, including the SMOG formula, for accurate readability assessment (T. B. Brown et al., 2020; Raffel et al., 2019). The majority of these formulae provide a predicted "grade score" of readability, indicating the level of education, as determined by the year group system employed in USA public schools, that is expected for the text to be comprehended. To evaluate both factual correctness and humanness of letters generated by ChatGPT, each letter was analysed by two independent clinicians who regularly manage skin cancer as part of their practice. The analysis of factual correctness and humanness was conducted on a Likert scale ranging from 0 to 10, with 0 representing completely incorrect or inhuman and 10 representing completely correct and human, respectively. The clinician assessed each letter individually and rated it on the Likert scale based on their expert judgment of the accuracy and human-like qualities of the text. Statistical analysis was undertaken in R version 4.1.1 (R Core Team, R Foundation for Statistical Computing, Vienna, Austria). $P < 0.001$ was deemed statistically significant.

8.3 Results

In total 38 hypothetical clinical scenarios were created. Seven of these pertained to BCC, 11 to SCC and 20 to MM. The characteristics of these are shown in Table 29.

Table 29: Characteristics of included clinical scenarios.

Skin cancer	Instructions (a)		Specific (b)		General (c)	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**
BCC	2	5	6	1	6	1
SCC	8	3	6	5	7	4
MM	20	0	10	10	10	10

There was increasing complexity using (a) general commands issued to ChatGPT3 to write a letter (b), referencing specific national guidelines and asking ChatGPT to follow these or (c) prompting ChatGPT3 to consult the literature to find the latest available guidelines and follow these but without specifically mentioning the title of the guideline(s) in question. Overall, the readability scores suggest that the text may be suitable for a varying reading ability (Table 30), but with the average readability for the text generated by ChatGPT at USA 9th grade (UK 14-15 years old) and considered by the USDHHS as 'average difficulty'.

Table 30: Summary of overall readability scores for clinical letters generated by ChatGPT.

Readability metric	Flesch Reading Ease	Flesch Kincaid Grade Level	Gunning Fog Score	Coleman Liau Index	SMOG Index	Automated Readability Index	Fry Grade Level	Lensear Write
Mean score	55.63	9.81	12.58	10.51	12.50	9.70	8.79	74.68

Overall median correctness of the clinical information contained in the letter was 7 (range 1-9). Overall median humanness of the writing style was 7 (range 5-9). The weighted kappa was for correctness and humanness was 0.8 ($p < 0.001$) and 0.77 ($p < 0.001$) respectively.

One-way ANOVA was used to compare the mean values of the response variables, median correctness and median humanness, among the cancer types of BCC, SCC and MM. For

median correctness, the overall results of the ANOVA showed a statistically significant difference among the groups ($F(2,35) = 10.1$, $p < 0.001$). The Tukey honestly significant difference (HSD) test showed that the mean difference between the MM and BCC groups was statistically significant (-2.71, 95% CI -4.32 to -1.11, $p < 0.001$), with the mean of the MM group being lower than the mean of the BCC group. However, the mean differences between the SCC and BCC group (1.08 95% CI -2.84 to 0.69) and SCC and MM group (1.64 95% CI 0.27 to 3.02) were not statistically significant ($p = 0.31$ and $p = 0.02$, respectively).

For median humanness, the overall results of the ANOVA showed a statistically significant difference among the groups ($F(2,35) = 27.76$, $p < 0.001$). The Tukey HSD test showed that there was a statistically significant difference between the MM and BCC groups (-1.63 95% CI -2.17 to -1.09, $p < 0.001$), with the mean of the MM group being lower than the mean of the BCC group. There was also a statistically significant difference between the SCC and BCC groups (-1.43 95% CI -2.03 to -0.83, $p < 0.001$), with the mean of the SCC group being lower than the mean of the BCC group. However, there was not a statistically significant difference between the SCC and MM groups ($p = 0.55$).

Two separate generalised linear models (GLMs) were used to investigate the effect of the predictor variables ‘cancer type’ (BCC as reference category), ‘general commands’, ‘specific guidelines’, and ‘general guidelines’ on the outcome variables, ‘median humanness’ and ’median correctness’ in the first and second GLM respectively. The results of the GLM for median humanness showed that cancer type was a significant predictor, with the MM and SCC coefficients being -1.45 ± 0.30 standard error (SE) and -1.32 ± 0.28 SE, respectively (both $p < 0.001$). The variables ‘general commands’, ‘specific guidelines’, and ‘general guidelines’ were not significant predictors of median humanness. The multiple R-squared value of 0.622

indicated that the model explained 62.2% of the variance in median humanness, and the F-statistic of 10.55 and corresponding $p < 0.001$ indicated that the model was significant overall.

In the GLM for median correctness, the MM predictor was found to be significantly associated with median correctness with a coefficient of -2.64 ± 0.90 SE ($p < 0.001$). The other predictors were not significantly associated with median correctness. The multiple R-squared value of 0.3729 suggested that the model explained approximately 37.3% of the variance in median correctness. The F-statistic and corresponding p-value were used to test the overall significance of the model, and $p < 0.001$ indicated that the model is significantly different from a model with no predictors.

8.4 Discussion

This pilot assessment of ChatGPT3 aimed to assess the feasibility of using an LLM to generate clinic letters that were readable, factually correct and with a human-like quality, when presented with only limited human commands. A total of 38 clinical scenarios were created, covering some of the most common scenarios a clinician might face during the workflow of communicating histopathology results to patients for BCC, SCC and MM as part of a skin cancer practice. The AI generated letters were assessed for readability and quality by two independent reviewers using several readability formulae and a Likert scoring system respectively.

The main findings from this study show that AI generated letters had a high overall correctness and humanness score. Nevertheless, there was variance in these scores amongst the cancer type the clinical letter was issued for, with a lower correctness score for MM when compared with BCC and a lower humanness score for both SCC and MM when compared with BCC. For

example, one problem encountered with MM letters was the generality of the information provided from national guidelines, which while not factually incorrect, could be misleading and create disproportionate concern for someone with a thin, low stage MM. GLMs showed that cancer type was a significant predictor for both humanness and correctness, with MM and SCC having lower scores compared to BCC. 'General commands', 'specific guidelines', and 'general guidelines' were not significant predictors of median humanness. However, MM was a significant predictor of negative median correctness. The comparatively lower requirements for histopathological reporting and the limited range of treatment options may contribute to the higher level of humanness and correctness observed for BCC and may allow for more accurate and concise communication in clinical letters compared to SCC and MM (Slater & Barrett, 2019-b, 2019-c; Slater & Cook, 2019). Further research is needed to evaluate the impact of these factors on the quality of clinical communication and to identify strategies for improving the humanness and correctness of clinical letters for all types of cancer independent of reporting structure and treatment options. OpenAI has stated that the ChatGPT model has limited knowledge of events and the world after 2021. This limitation should be considered when evaluating the generalisability of this method and when clinical guidelines are updated in the future. Without information about the training data used to develop ChatGPT, the external validity and performance on novel data remains uncertain until further studies are conducted.

Despite instructing the chatbot to write at a UK reading age of 11-12 years the average reading age for the text generated by ChatGPT was at USA 9th grade (UK reading age of 14-15 years old) which is somewhat above the USDHHS recommendation of at or below the 6th-grade level (UK reading age of 11-12 years). Previous work investigating the readability of plastic surgery showed clinic letters sent to patients had a median USA grade score of 9.4 (Drury et al., 2021). As such, the method employed in this study appears to generate results that are

broadly comparable to current real-world practice in the host institutional setting. However, further testing with instructions to ChatGPT to further lower the reading age is needed to facilitate understanding of clinic letters relaying results to patients amongst the general population.

The potential for AI to generate clinical letters as an alternative to those written by clinicians raises important considerations for the quality and effectiveness of healthcare communication. While AI has the potential to improve the accuracy and efficiency of clinical letters, there are also potential risks associated with its use in healthcare. One such risk is the possibility of errors or omissions in the information provided by the AI system, which can have serious downstream consequences for patient care. For example, incorrect or incomplete information about a patient's cancer histopathology results or treatment recommendations could lead to incorrect diagnoses and inappropriate treatment decisions, potentially impacting patient survival and quality of life. Additionally, there is a risk of downstream clinical errors when healthcare providers rely on the information provided by the AI system without adequately verifying its accuracy or completeness. In order to mitigate these risks, it is important for the use of AI in healthcare, including the automated generation of clinical letters, to be carefully regulated and monitored. This may involve the development of guidelines and standards for the use of AI in healthcare, as well as ongoing training and education for healthcare providers to ensure they are aware of the potential risks and limitations of AI systems. It may also be desirable to adopt a human-in-the-loop approach, whereby AI systems are used to assist rather than replace healthcare providers, and the accuracy and completeness of the information provided by the AI system is carefully verified by healthcare providers. To responsibly incorporate ChatGPT into a clinical workflow, one approach could be to use voice to text recognition software with limited human input, followed by rapid editing by authors. This

could be a feasible starting point for exploring the potential applications of this technology while also addressing any potential risks. While this study provides insight into the potential of AI generated clinical letters, the small sample size and use of a single AI system for letter generation should be considered as limitations. Further studies are needed to assess the effectiveness of AI generated clinical letters in real-world clinical settings and to compare the performance of various AI systems against one another, as it is likely that one developed with a greater medical focus will yield improved results. In addition, the quality and writing style of input provided to any chatbot should be assessed in a range of settings, including different languages, resource levels and cultural contexts.

8.5 Conclusion

The use of AI-generated clinical letters has the potential to revolutionise healthcare communication, improving efficiency, consistency, accuracy, patient satisfaction, and deliver cost savings. However, it is crucial to exercise caution as there are potential risks to patient care that must be carefully considered. In advance of widespread adoption of AI technologies like ChatGPT in clinical medicine, it will be vital to carefully regulate and monitor their use to ensure the safety and quality of patient care. It is also essential that clinicians stay up to date with the evolution of AI technology and participate in the development and implementation of guidelines and standards for its use in healthcare, as well as ongoing training and education to stay aware of the potential risks and limitations of AI systems.

Chapter Nine:

Validating a novel automated population-based approach to quality assurance in surgical oncology with basal cell carcinoma incomplete excision rates using natural language processing

9.1 Introduction

Measuring clinician and patient-reported outcomes alongside treatment morbidity, recurrence and survivorship are key metrics to consider in the quality assurance of any cancer service. This evidence-based approach can assist patients and clinicians to critically appraise the most suitable treatment, validate prognostication systems used in surgical and medical oncology, and can be used to compare and assess operative or oncological protocols. National cancer databases represent powerful tools that help us to improve our understanding and management of cancer via transparent reporting of these outcomes. Within the UK, the National Cancer Registration and Analysis Service (NCRAS), Welsh Cancer Intelligence and Surveillance Unit, Scottish Cancer Registry and Northern Ireland Cancer Registry quality assures and analyses data on all people diagnosed with cancer. Similarly, the United States of America has invested in the development of the National Cancer Data Base (The American College of Surgeons, 2024).

However, the 2017 CRUK commission into the effectiveness of MDT meetings in cancer services found that MDTs are not currently able to fulfil their important secondary roles of data validation and audit (Cancer Research UK, 2024-b). These are pivotal for auditing cancer services and facilitating information flows to national cancer registries. CRUK recommend that MDTs should use a database or proforma to enable documentation of recommendations in real time. Auditing of outcomes including incomplete excision rates in surgical oncology is generally undertaken manually and retrospectively – both at surgeon, departmental and national levels. This is an inefficient use of skilled clinician time, especially when large volumes of data need extracting. It is crucial that there is sufficient time to discuss complex patients and thus a novel automated method of data capture that frees up specialist time during

the MDT is desirable so that MDTs can fulfil their primary function of patient centred, evidence-based decision making (Ali et al., 2021).

Routinely collected EHR data is a powerful research resource but often lacks detailed disease-specific information collected in clinical free text e.g. clinic letters, operative notes or histopathology reports. Historically, data from these records have been relatively inaccessible when compared to data stored in a structured database. However, NLP represents a technique which means this is no longer the case. NLP techniques enable automated extraction of detailed clinical information from unstructured free text EHR data. NLP itself can be defined as a set of techniques used to convert written text into interpretable datasets through either rule-based or machine learning models (Harrison & Sidey-Gibbons, 2021).

The use of NLP in surgical outcomes research is accelerating. In a recent systematic review and meta-analysis of NLP-based versus conventional administrative methods of data capture (CPT codes, ICD codes, patient safety indicators based on discharge coding and DRG database code) postoperative complications were identified with higher sensitivity and comparable specificity in the NLP group (Mellia et al., 2021). While the application of NLP to EHR data continues to expand, no studies to date describe its use for determining incomplete excision rates in surgical oncology. A population level data analysis tool that can interrogate routinely collected free text histopathology reports from the EHR could be used to benchmark standards and allow confidential feedback to service providers and users.

Standard surgical excision is an effective treatment for the majority of primary BCCs, with reported 5-year recurrence rates of 3-8.0%. However, the excision margin, or the amount of normal tissue removed around the BCC, plays a crucial role in achieving a balance between

cure and minimising morbidity (Mosterd et al., 2008; Smeets et al., 2004; Thissen et al., 1999). In order to optimise treatment outcomes, the BAD guidelines currently recommend that low-risk BCCs should be excised using a 4mm peripheral clinical margin, while primary BCCs with high-risk factors should be excised using at least a 5mm peripheral clinical margin (Nasr et al., 2021). Additionally, they highlight the importance of adequate excision at the deep margin, with recommendations to excise to a clear plane, including a fat layer where present, and other deeper structures if needed.

Since the publication of previously reported data on the likelihood of achieving clear peripheral and deep margins with standard surgical excision, the RCPath have updated their guidance on the reporting of BCC histological subtype (Ali, Abdulla, et al., 2021; Slater & Barrett, 2019-b; Warren et al., 2022). They now advise that there is no clinical value in distinguishing between the infiltrative, sclerosing, morphoeic and micronodular subtypes and that these should all be regarded as histological features indicating a high-risk lesion. This allows the majority of BCCs to be categorised as either low-risk or high-risk.

Chapter six described the validation of an automated histopathology information extraction system using rule-based NLP techniques capable of named entity recognition and feature extraction (Ali et al., 2022). It is hypothesised that a web application that imports data from this NLP algorithm and automates downstream data analysis would provide an accessible real-time quality assurance data tool that increases the speed and reduces the cost of the audit process. The primary aim of this study was to use this platform to bridge the data gap between EHR and cancer registry data and to provide a novel approach to assessing incomplete excision rates in surgical oncology at scale with the most common human cancer (BCC) as a use case.

Previous research is expanded upon by conducting an analysis using data that is categorised according to the updated RCPath standards. These standards were used to develop the current BAD guidelines and provide recommendations for treatment according to low-risk and high-risk BCC criteria. However, it is worth noting that the systematic review conducted by the BAD to inform their guideline development included studies that were conducted before the implementation of the RCPath standards. As a result, there are no studies to date that investigate the impact of the updated RCPath criteria on treatment efficacy for low-risk and high-risk BCC. It is clear that excision margins still remain an important and often contentious issue, with the recent James Lind Alliance (JLA) Priority Setting Partnership in skin cancer surgery identifying excision margins in the surgical management of BCC as one of their research priorities (James Lind Alliance, 2022). The secondary aim was to clarify the BCC types that should be treated with wider or deeper clinical margins and inform further guideline updates.

9.2 Methods

9.2.1 Study design

A multi-centre (Morriston Hospital, Singleton Hospital and Neath Port Talbot Hospital), pan-speciality, retrospective analysis was undertaken of consecutive patients with a BCC over a 17-year period from 2004 to 2021, managed with surgical excision using a pre-determined margin at Swansea Bay University Health Board, Swansea, United Kingdom. All lesions were examined by a consultant histopathologist using the bread loafing cross-section technique (Abide et al., 1984). Primary, recurrent and previously excised lesions were grouped together for analysis.

9.2.2 Inclusion criteria

Patients with BCC that were managed by surgical excision, using a pre-determined margin, within the study period.

9.2.3 Exclusion criteria

Diagnostic biopsies including punch biopsy, incision biopsy, shave biopsy and curettage and patients managed using Mohs micrographic surgery were excluded from the study.

9.2.4 NLP algorithm

A detailed description and validation of the NLP algorithm has previously been described in chapter six. To reiterate, the GATE framework was used to build an NLP information extraction system using rule-based techniques. This was validated on previously unseen, de-identified and pseudonymised BCC histopathological reports at the same institution as the current study. The mean precision, recall and F1 score of 86.0% (95% CI, 75.1-96.9), 84.2% (95% CI 72.8-96.1) and 84.5% (95% CI 73.0-95.1) respectively. The difference between clinician annotator F1 scores was 7.9% in comparison to 15.5% between the NLP pipeline and the gold standard corpus.

9.2.5 Case identification, data extraction and processing

Cases were retrospectively identified from InterSystems TrakCare Laboratory Information Management System (InterSystems TrakCare Lab, Cambridge, Massachusetts, USA), using SNOMED RT codes for BCC (M-80983, M-80903, M-80943, M-80933, M-80923, M-80943, M-80973, M-80913). Once cases were identified from this prospectively maintained database, free text pathological reports were retrieved and saved in text file format. A rule-based NLP pipeline was then run on this corpus. CSV text files were generated from the respective canonical subheadings of the pathology report. Complete case analysis was used as an approach

to the treatment of missing data. Given the structured nature of the data output by the study's NLP algorithm no data cleaning was required per se. However, when there was more than one output for the same piece of free text the first annotation was used. If a numeric value was extracted the worst prognostic value was selected e.g. in the following statement 1mm would be selected as the peripheral margin value: 'peripheral margin 1mm at 9 o'clock, 3mm 12 o'clock, 5mm 3 o'clock and 3mm 6 o'clock'. Custom Python scripts were used for both these tasks.

9.2.6 Web application

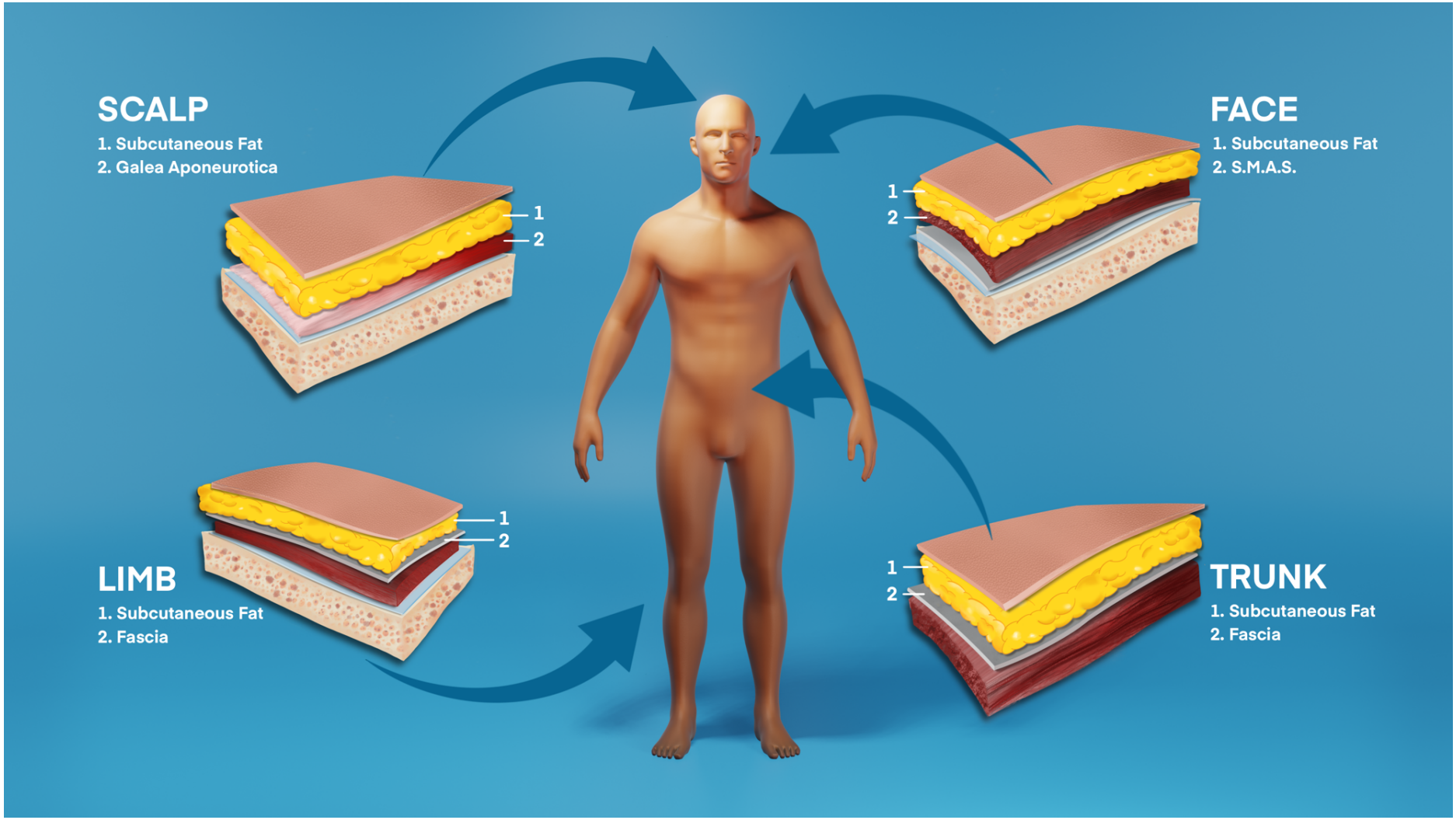
A Java™ Spring Boot (VMware Incorporated, Palo Alto, California, USA) was developed using a web application hosted on Amazon Web Services (Amazon.com Incorporated, Seattle, Washington, USA) in EC2. Respective CSV files were imported into a RDMS. MySQL Workbench and MySQL Community Server (Oracle Corporation, Austin, Texas, USA) were used as the platform for the study's RDMS. Within the application framework separate APIs were developed to automate: 1) descriptive analysis for generating incomplete excision rates (on a surgeon and speciality basis) and 2) data visualisation in the form of a histogram to identify outlying surgeons and specialities (Figure 35).

9.2.7 Variables

Peri-operative tumour factors (primary versus recurrent anatomical site, borders, perineural invasion, lymphovascular invasion, tumour thickness, diameter, level of invasion, subtype, differentiation, stage), patient factors (immunosuppression and previous radiotherapy) and surgical factors (pre-operative peripheral and deep margin and speciality of clinician undertaking surgery) were recorded. The documented deep clinical margin was recorded based

on the anatomical plane the lesion was excised to, with subcutaneous fat recorded as plane 1 (Figure 33). This was defined as depth level.

Figure 33: Anatomical planes of deep margin excision demonstrated in different sites across the body.



SMAS; superficial musculoaponeurotic system.

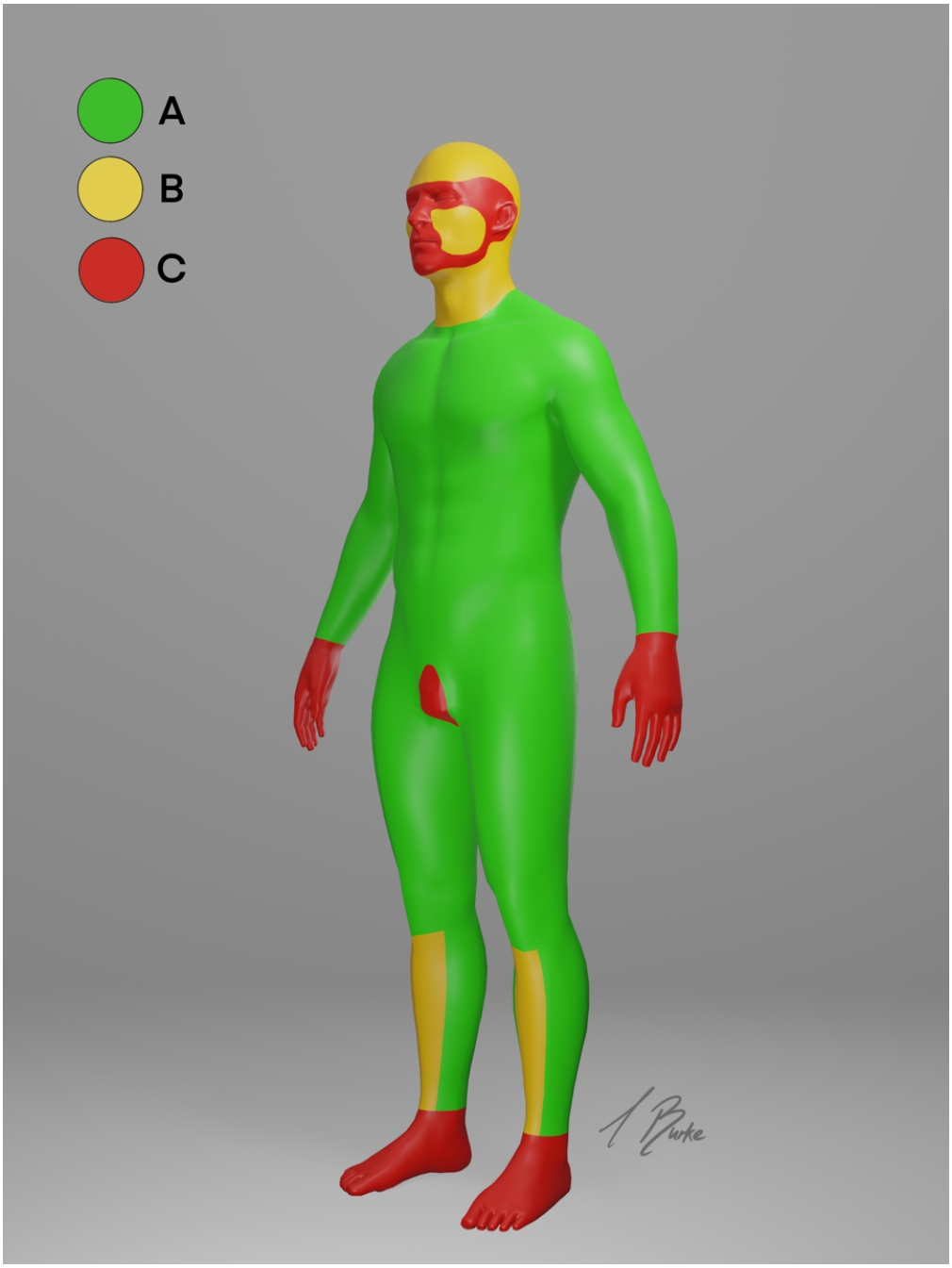
The BAD use an adapted National Comprehensive Cancer Network (NCCN) table in their guidelines on the treatment of BCC to define of criteria for low-risk and high-risk BCC (Nasr et al., 2021). This was modified this for use in the current study to categorise BCCs clinicopathologically into low-risk and high-risk, equating to the Union for International Cancer Control 8th edition (UICC8) version of TNM8 (tumour–nodes–metastasis) and RCPath dataset (Table 31 and Figure 34) (Amin et al., 2017; Gospodarowicz et al., 2016). The overall clinical risk status of a mixed subtype BCC was judged from the highest risk subtype(s) present, irrespective of percentage or location, in line with current RCPath reporting standards (Slater & Barrett, 2019-b).

Table 31: Criteria for low-risk and high-risk BCC.

Criteria	**Low risk**	**High risk**
Location and size	Area A ≤ 20 mm (maximum clinical diameter)	Area A > 20 mm (maximum clinical diameter)
	Area B ≤ 10 mm (maximum clinical diameter)	Area B > 10 mm (maximum clinical diameter)
		Area C
Borders	Well defined	Poorly defined
Primary vs. recurrent	Primary	Recurrent
Immunosuppression	No	Yes
Site of prior radiotherapy	No	Yes
Growth pattern	Nodular, cystic, superficial, fibroepithelial	Infiltrative (infiltrating, morphoeic, micronodular and multinodular)
Differentiation: basosquamous	Absent	Present (with or without lymphovascular invasion)
Level of invasion	Dermis, subcutaneous fat	Beyond subcutaneous fat
Depth (thickness)	≤ 6 mm	> 6 mm
Perineural invasion	Absent	Present
Pathological TNM stage	pT1	pT2

TNM, Tumour–Nodes–Metastasis. One or more criteria satisfies the criteria for high risk.

Figure 34: Topographical areas used for classification of low-risk and high-risk BCC that corresponds with Table 1.



Area A, trunk and extremities but excluding hands, nail units, genitalia, pretibia, ankles and feet; Area B, cheeks, forehead, scalp, neck and pretibial; Area C, 'mask areas' of face [central face, eyebrows, periorbital, eyelids, nose, lips (cutaneous and vermilion), chin, mandible, preauricular, postauricular, temple, ears]; genital areas; hands, nail units, ankles and feet.

9.2.8 Outcomes

The primary endpoints were histological margin status, risk status and speciality of the operating surgeon. The margin status was defined as either clear (≥1mm) or involved (0mm).

9.2.9 Validation

An *a priori* sample size calculation was undertaken based on pilot data to calculate the size of the internal validation cohort using the method described by Cantor et al (Cantor, 1996). The probability of 0.05 for both clinician and NLP algorithm derived incomplete excision rate was used since the BAD states that a ≥95% complete excision rate is acceptable and a normal distribution between the two ratings was assumed. Testing the null hypothesis: Cohen's kappa = 0.7 versus the alternative hypothesis: Cohen's kappa > 0.7 given a kappa of 0.85 derived from the study's pilot data with a desired type I error of 0.05 and power of 0.9, a sample size of 2010 was required to reject the null hypothesis. Retrospective analysis of the baseline variables and outcomes was undertaken retrospectively by two independent and blinded expert clinicians. This clinical pathway served as the reference standard for the study. There were no differences from the development data compared to the validation in terms of setting, eligibility criteria, outcome or predictors. Disagreements were resolved by case discussion until a consensus was reached. The single consensus clinician-derived outputs were then compared against NLP-derived outputs for analysis.

9.2.10 Statistical analyses

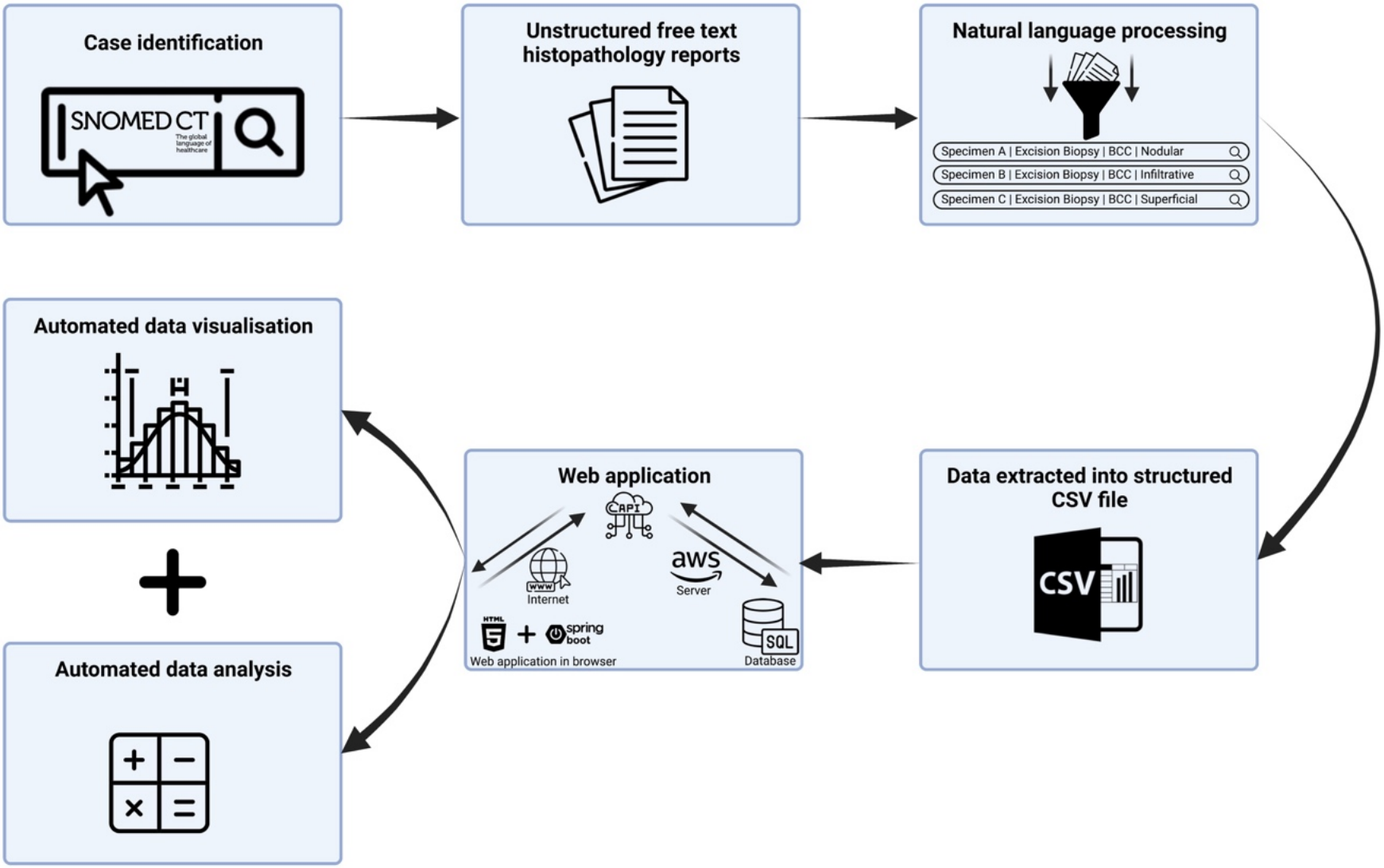
Percentage agreement and Cohen's kappa were used as measures of agreement between NLP-derived and clinician-derived completeness of excision, risk status and speciality of the operating surgeon. Percentage agreement computes simple agreement among raters whilst Cohen's kappa provides a coefficient to interpret the level of agreement: poor if $k < 0.20$, fair

if $0.21 \leq k \leq 0.40$, moderate if $0.41 \leq k \leq 0.60$, substantial if $0.61 \leq k \leq 0.80$ and good if $k > 0.80$ (Landis & Koch, 1977).

A logistic regression model was used to examine the relationship between the pre-determined clinical peripheral margin value (mm) and complete histological peripheral margin clearance. The model included an interaction term between peripheral margin value (mm) and risk, which was included as a covariate. Predictions of complete histological peripheral margin clearance were made for different levels of clinical peripheral margin value, stratified by risk. Finally, a plot of the probability of complete histological peripheral margin clearance as a function of clinical peripheral margin value was created for high-risk, low-risk and all BCCs separately. Similarly, a second logistic regression model was fitted to investigate the relationship between complete histological deep margin clearance and the predictor variables surgical depth level, risk, and tumour thickness.

Statistical analysis was undertaken in R version 4.1.1 (R Core Team, R Foundation for Statistical Computing, Vienna, Austria). This process is schematically represented in Figure 29. *Post-hoc* analysis was performed to investigate model the relationship between uninvolved margin and speciality. $p < 0.001$ was deemed statistically significant.

Figure 35: Schematic representation of the automated population-based quality assurance model.



9.3 Results

9.3.1 Quality assurance of incomplete excision rates at scale

There were a total of 34,955 lesions in 15,657 patients that met the inclusion criteria for the primary aim of the study. Each patient received a separate histopathology report describing the lesions excised. Baseline characteristics are shown in Table 32. The overall incomplete excision rate in this cohort was 5.5%.

Table 32: Baseline characteristic of included patients.

Specialty	**High risk (n = 8855)**	**Low risk (n = 6802)**	**Total patients**	***p value***
Dermatology	3581	4546	8127	< 0.001
Plastic surgery	2648	951	3599	< 0.001
Oral and maxillofacial surgery	1251	368	1619	< 0.001
Other	589	308	897	< 0.001
Ear, nose and throat	402	82	484	< 0.001
General practice	174	292	466	< 0.001
Ophthalmology	190	237	427	< 0.001
General surgery	20	18	38	0.7456

p-values are for chi-square tests for differences in the proportion of high/low risk patients between speciality.

The overall incomplete excision rate was calculated and then stratified by risk (high or low) and margin (peripheral or deep) (Table 33).

Table 33: Incomplete excision rates for all specialities stratified by risk and margin.

Risk status	Peripheral margin incomplete rate (%)	Deep margin incomplete rate (%)	Overall incomplete rate (%)
Low risk	0.9	1.2	2.1
High risk	1.7	1.7	3.4
Total	2.6	3.0	5.5

The descriptive incomplete excision rate stratified by risk and margin by the web application is shown in Table 3. There were 6,152 histopathology reports in the validation cohort (representing 17.6% of the data) that compared NLP-derived and clinician-derived completeness of excision, risk status and speciality of the operating surgeon. There was 0.99 agreement (95% CI, 0.98 to 0.99) (Cohen's kappa = 0.74 (95% CI, 0.68 to 0.80), $p < 0.001$) for completeness of excision. There was 0.99 agreement (95% CI, 0.99 to 0.99) (Cohen's kappa = 0.73 (95% CI, 0.69 to 0.77), $p < 0.001$) for speciality of the operating surgeon.

Using a MacBook Apple M1 Pro with 16 GB RAM operating macOS Monterey it took the NLP-pipeline 22.7 minutes to extract 2,184,309 items of information and transform this into a structured format ready for analysis, representing a data extraction rate of 689.7 cases/minute. The data extraction rate for a single clinician was 0.25 cases/minute on the validation cohort. By extrapolating this rate from the validation cohort over 15,657 histopathology reports using an 8-hour working day, Monday to Friday, with a 30-minute rest break every 4 hours it would have taken a single clinician 29.8 weeks to extract the same amount of data. Using this latter figure, it is estimated that there was a time saving of 208 days with this novel approach.

Post-hoc binary logistic regression (risk status as covariate) demonstrated that plastic surgeons were more likely to get a clear margin when compared to all other specialities (Table 34).

Table 34: *Post-hoc* binary logistic regression to model the relationship between uninvolved margin and speciality.

Specialty	**Odds ratio (95% CI)**	**Incomplete excision rate (%) (95% CI)**	***p value***
Oral and maxillofacial surgery	0.62 (0.52-0.75)	7.78 (6.96-8.60)	< 0.001
Other	0.57 (0.46-0.70)	8.28 (7.30-9.27)	< 0.001
Dermatology	0.53 (0.47-0.61)	8.26 (7.16-9.37)	< 0.001
Ear, nose and throat surgery	0.43 (0.33-0.56)	11.09 (10.08-12.11)	< 0.001
General practice	0.22 (0.17-0.29)	17.47 (15.41-19.53)	< 0.001
Ophthalmology	0.21 (0.17-0.27)	18.29 (16.12-20.46)	< 0.001
General surgery	0.00	0.00	0.934

Plastic surgery as reference specialty. Risk as covariate.

Incomplete excision rates were calculated from the population data by computing the log odds corresponding to several discrete values in the model. Conversion of each log odds (probability = exp[log odds]/1+exp[log odds]) allowed the derivation of the probability of complete excision for each category of specialty and risk on a 0 to 1 scale (Hrabač & Trkulja, 2019). This was undertaken on significant values only. These complete excision probabilities were then converted to incomplete excision rates found in Table 35. This approach was adopted, instead of using "raw" incomplete excision rates, to factor risk when comparing between specialities.

Table 35: Incomplete excision rate across specialities stratified by risk status in relation to plastic surgery.

Specialty	High risk incomplete excision rate (%)	Low risk incomplete excision rate (%)
Oral and maxillofacial surgery	8.23	6.28
Other	8.96	6.86
Dermatology	9.50	7.28
Ear, nose and throat	11.56	8.90
General practice	20.22	15.93
Ophthalmology	20.79	16.40

Plastic surgery as reference specialty. Risk as covariate.

9.3.2 Multivariate analysis with updated RCPath standards using NLP

1,447 lesions had complete data and met the inclusion criteria for the secondary aim of the study. Each patient received a separate histopathology report describing the lesions excised. Baseline characteristics are shown in Table 36.

Table 36: Baseline characteristic of included lesions.

Margin	Risk	Margin (mm)	Number of lesions
Peripheral	Low	= 0	28
Peripheral	Low	> 0	319
Peripheral	High	= 0	55
Peripheral	High	> 0	745
Total peripheral	-	-	1147
Deep	Low	= 0	4
Deep	Low	> 0	48
Deep	High	= 0	19
Deep	High	> 0	229
Total deep	-	-	300
Total peripheral and deep	-	-	1447

9.3.3 Peripheral clearance

Binary logistic regression demonstrated that whether a lesion was high or low-risk was not influential in determining if the peripheral margin was histologically clear (Table 37, Figure 36 and 37).

Table 37: Odds ratios for obtaining histological peripheral clearance based on low-risk versus high risk and peripheral clinical margin.

Coefficients	OR	95% CI	p-value
Peripheral Margin	1.270	(1.066 - 1.551)	0.0128
Low-risk	0.901	(0.563 - 1.471)	0.670

OR; = odds ratio.

Figure 36: Probability of achieving a complete peripheral margin for *high-risk* BCC.

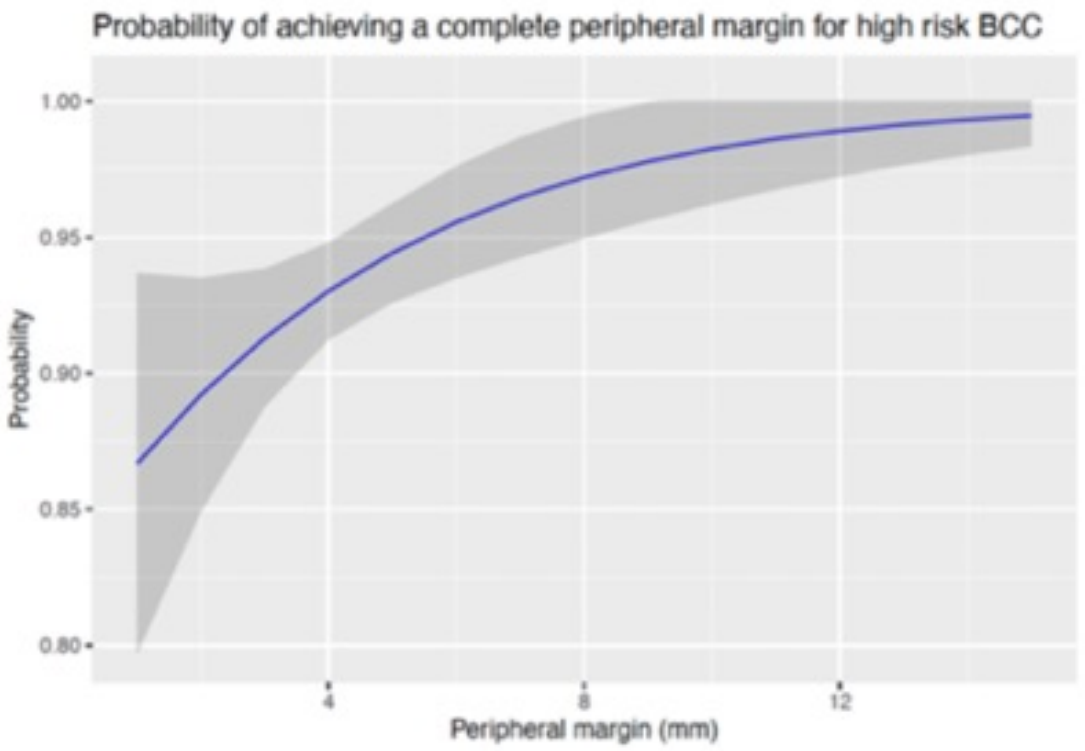
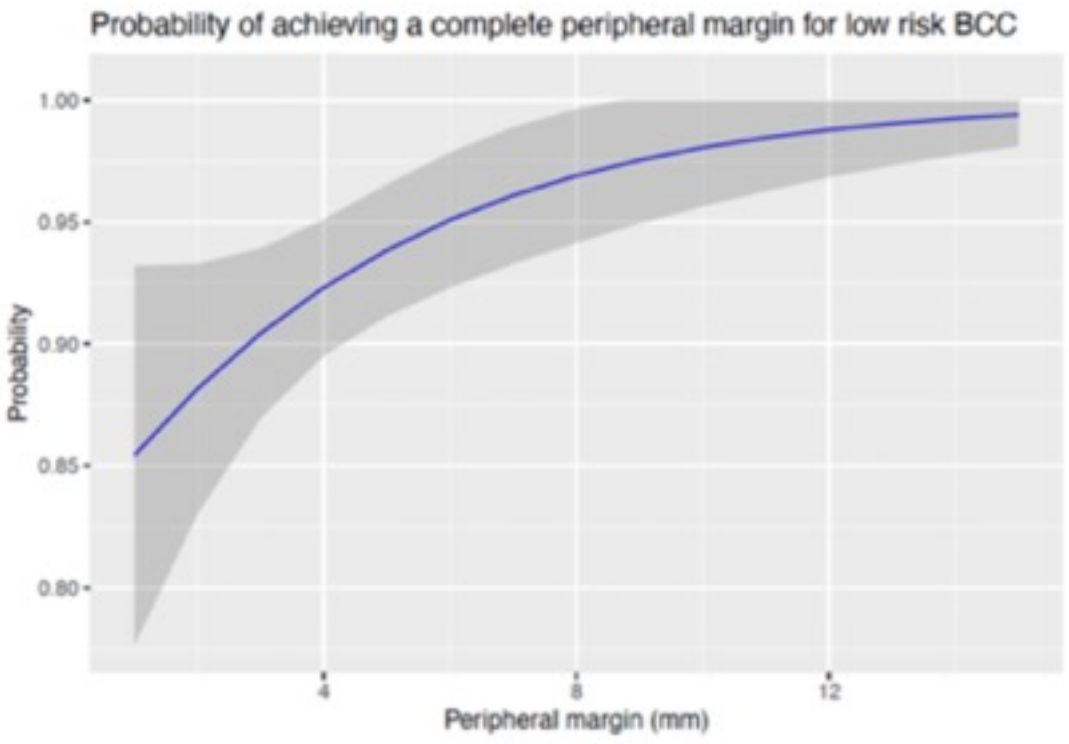


Figure 37: Probability of achieving a complete histological peripheral margin for *low-risk* BCC.



However, the clinical peripheral margin used was found to be statistically significant in determining if a lesion's histological peripheral margin was clear. As expected, increasing the clinical peripheral margin increased the chance of obtaining histological peripheral margin clearance. To assess the degree to which increasing the clinical peripheral margin had on the chance of complete histological clearance, conversion of log odds to probabilities for varying clinical peripheral margins was undertaken stratified by risk. At a clinical peripheral margin of 6mm a 95.0% histological clearance rate was achieved, this being the same for both high and low-risk BCCs (Figure 38). At a clinical peripheral margin of 11mm there was a plateauing of results, where there after the probability of obtaining clearance stayed at 99.0%. Data across all clinical margins and risk status is summarised in Table 38.

Figure 38: Probability of achieving a complete histological peripheral margin for all BCC.

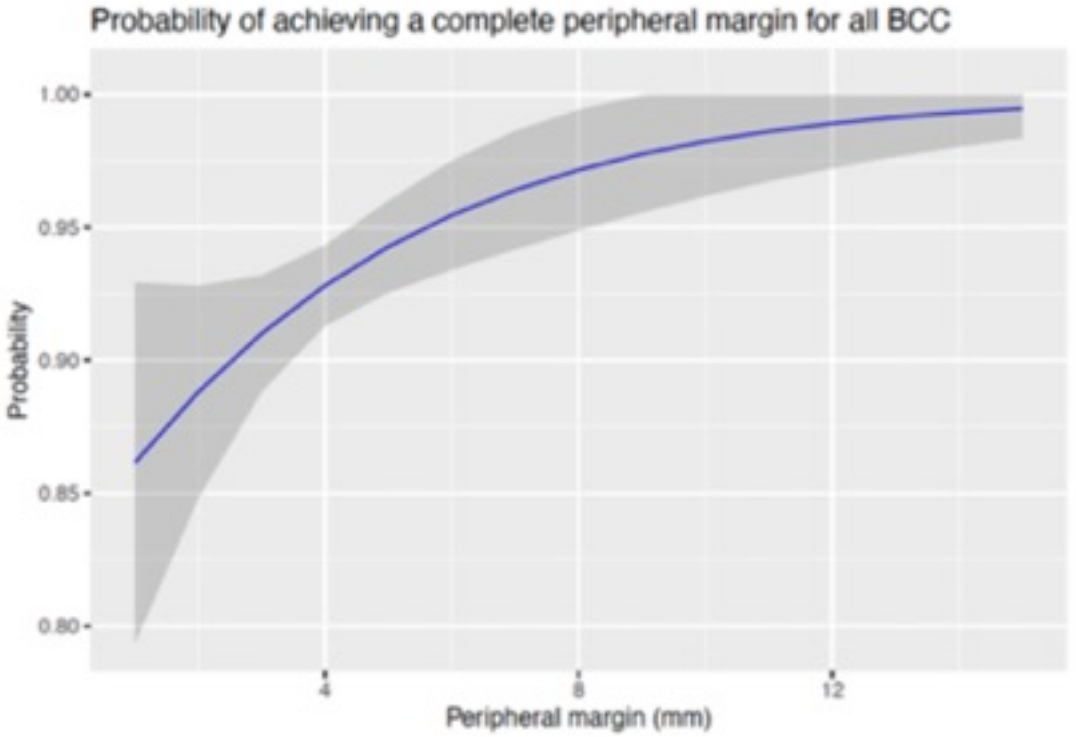


Table 38: Probability 95% CI of achieving a complete peripheral histological margin.

Peripheral margin (mm)	Risk	Probability	Low limit	High limit
1	High	0.87	0.80	0.94
2	High	0.89	0.85	0.94
3	High	0.91	0.89	0.94
4	High	0.93	0.91	0.95
5	High	0.94	0.93	0.96
6	High	0.96	0.93	0.98
7	High	0.96	0.94	0.99
8	High	0.97	0.95	0.99
9	High	0.98	0.96	1.00
10	High	0.98	0.96	1.00
11	High	0.99	0.97	1.00
12	High	0.99	0.97	1.00
13	High	0.99	0.98	1.00
14	High	0.99	0.98	1.00
15	High	0.99	0.98	1.00
1	Low	0.85	0.78	0.93
2	Low	0.88	0.83	0.93
3	Low	0.90	0.87	0.94
4	Low	0.92	0.90	0.95
5	Low	0.94	0.91	0.97
6	Low	0.95	0.92	0.98
7	Low	0.96	0.93	0.99
8	Low	0.97	0.94	1.00
9	Low	0.98	0.95	1.00
10	Low	0.98	0.96	1.00
11	Low	0.98	0.96	1.00
12	Low	0.99	0.97	1.00
13	Low	0.99	0.97	1.00
14	Low	0.99	0.98	1.00
15	Low	0.99	0.98	1.00
1	Overall	0.86	0.79	0.93
2	Overall	0.89	0.85	0.93
3	Overall	0.91	0.89	0.93
4	Overall	0.93	0.91	0.94
5	Overall	0.94	0.93	0.96
6	Overall	0.95	0.93	0.98
7	Overall	0.96	0.94	0.99
8	Overall	0.97	0.95	0.99
9	Overall	0.98	0.96	1.00

Peripheral margin (mm)	Risk	Probability	Low limit	High limit
10	Overall	0.98	0.96	1.00
11	Overall	0.99	0.97	1.00
12	Overall	0.99	0.97	1.00
13	Overall	0.99	0.98	1.00
14	Overall	0.99	0.98	1.00
15	Overall	0.99	0.98	1.00

Whether the addition of ulceration to upgrade low risk lesions to high risk would impact on the likelihood of histological clearance on *post-hoc* analysis was investigated. The odds of achieving clearance were found to decrease by a factor of 0.553 when ulceration was present ($p = 0.0257$). This reduced the probability of achieving complete histological peripheral margin clearance for all clinical margins (Table 39). Whether low risk BCCs and differing anatomical sites and diameters could be managed with smaller clinical margins on *post-hoc* analysis was also investigated. Univariate analysis of low risk BCCs (satisfying low risk criteria *except* high risk anatomical site) at area C showed that the clinical margin was a significant predictor of clearance (estimate = 0.8033, standard error = 0.3876, z-value = 2.072, $p = 0.0382$) but not in other anatomical areas with smaller margins (<10mm or 10-20mm). In this group the probability of achieving a complete peripheral margin was higher for the majority of clinical margins (Table 40) with a 4mm margin giving a 95.0% clearance rate.

Table 39: Probability 95% CI of achieving a complete peripheral margin using ulcerated to upgrade low risk lesions to high risk lesions.

Peripheral margin (mm)	Risk	Probability	Low limit	High limit
1	High	0.82	0.71	0.93
2	High	0.85	0.77	0.93
3	High	0.87	0.82	0.93
4	High	0.89	0.85	0.93
5	High	0.91	0.87	0.95
6	High	0.92	0.89	0.96
7	High	0.94	0.90	0.98
8	High	0.95	0.91	0.99
9	High	0.96	0.91	1.00
10	High	0.96	0.92	1.00
11	High	0.97	0.93	1.00
12	High	0.98	0.94	1.00
13	High	0.98	0.94	1.00
14	High	0.98	0.95	1.00
15	High	0.99	0.96	1.00
1	Low	0.89	0.83	0.96
2	Low	0.91	0.87	0.95
3	Low	0.93	0.90	0.95
4	Low	0.94	0.92	0.96
5	Low	0.95	0.93	0.97
6	Low	0.96	0.93	0.98
7	Low	0.96	0.94	0.99
8	Low	0.97	0.94	1.00
9	Low	0.98	0.95	1.00
10	Low	0.98	0.95	1.00
11	Low	0.98	0.96	1.00
12	Low	0.99	0.96	1.00
13	Low	0.99	0.97	1.00
14	Low	0.99	0.97	1.00
15	Low	0.99	0.98	1.00
1	Overall	0.78	0.64	0.92
2	Overall	0.81	0.71	0.92
3	Overall	0.84	0.76	0.92
4	Overall	0.87	0.80	0.93
5	Overall	0.89	0.82	0.95

Peripheral margin (mm)	Risk	Probability	Low limit	High limit
6	Overall	0.90	0.84	0.97
7	Overall	0.92	0.86	0.98
8	Overall	0.93	0.87	0.99
9	Overall	0.94	0.88	1.00
10	Overall	0.95	0.90	1.00
11	Overall	0.96	0.91	1.00
12	Overall	0.97	0.92	1.00
13	Overall	0.97	0.93	1.00
14	Overall	0.98	0.93	1.00
15	Overall	0.98	0.94	1.00

Table 40: Probability of achieving a complete peripheral margin for low risk BCCs <10mm at area C.

Peripheral margin (mm)	Probability	Low limit	High limit
1	0.65	0.25	1.00
2	0.81	0.64	0.98
3	0.90	0.85	0.96
4	0.95	0.92	0.99
5	0.98	0.95	1.00
6	0.99	0.97	1.00
7	1.00	0.98	1.00
8	1.00	0.99	1.00
9	1.00	1.00	1.00
10	1.00	1.00	1.00
11	1.00	1.00	1.00
12	1.00	1.00	1.00
13	1.00	1.00	1.00
14	1.00	1.00	1.00
15	1.00	1.00	1.00

9.3.4 Deep clearance

Assessment of tumour thickness with depth level stratified by risk was not possible due to the low frequency of lesions which were low-risk and had an incomplete deep margin. To avoid bias risk was not added as a variable in the model. After removing outliers, tumour thickness up to a maximum of 6mm was analysed.

Regression analysis indicated that increasing tumour thickness decreased the chance of obtaining deep histological clearance (OR 0.720, 95% CI, 0.525 – 0.991, $p < 0.05$ [Table 41]). This is reflected in the results arising on the analysis of the depth levels (Table 42). The ORs of the depth levels fluctuated (with, in most cases, large CIs). As in the peripheral margin example, probabilities of deep clearance were computed for varying tumour thickness stratified by depth level. The results can be seen in Figure 39. All depths levels showed a decreasing chance of deep clearance with increasing tumour thickness. Depth level 2 had the greatest probability of achieving deep clearance (97.0%) at all tumour thickness, followed by depth level 1 (92.0%).

Figure 39: Probability of achieving a complete deep margin when considering tumour thickness for *all* BCCs stratified by depth level.

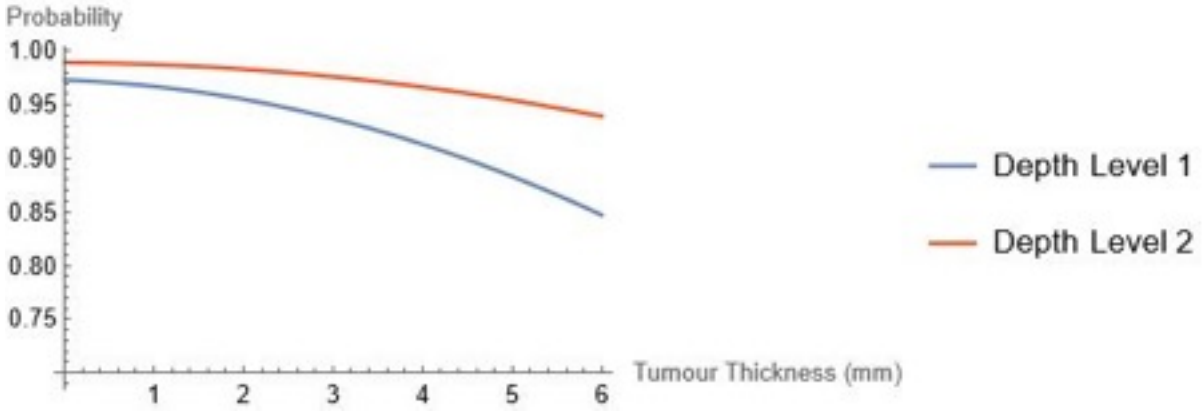


Table 41: Summary of logistic regression results assessing deep clearance subject to depth level and tumour thickness.

Coefficients	OR	95% CI	p-value
Depth level 2	2.726	(0.659 – 18.428)	0.213
Tumour thickness	0.720	(0.525 – 0.991)	0.041

OR; = odds ratio.

Table 42: Probability 95.0% CI of achieving a complete deep margin using a logistic regression model of depth level and tumour thickness for high-risk and low-risk lesions combined.

Depth level	Tumour thickness (mm)	Probability	Low limit	High limit
1	1	0.97	0.93	1.00
1	2	0.95	0.92	0.99
1	3	0.94	0.90	0.98
1	4	0.91	0.86	0.98
1	5	0.89	0.79	0.98
1	6	0.85	0.70	1.00
2	1	0.99	0.97	1.00
2	2	0.98	0.96	1.00
2	3	0.98	0.94	1.00
2	4	0.97	0.92	1.00
2	5	0.95	0.89	1.00
2	6	0.94	0.84	1.00

The ratio of depth level to tumour thickness was used as a variable in the model with risk, probabilities were computed based on the relationship of these variables to demonstrate the distribution of this ratio relative to the likelihood of achieving a complete of deep margin. While both coefficients indicated no statistical significance (Table 43) the probability plot showed an increasing trend of ratio and probability in an almost linear trend (Figure 35). It was clear that no real change was evident between the high and low-risk probabilities (Table 44).

For lower values of ratio (i.e. low values of depth level, and high values of tumour thicknesses) there was a lower probability of achieving complete deep clearance. For higher values of the ratio (i.e. high values of depth level and low values of tumour thickness) there was an increase of likelihood of a deep clearance. Therefore, the results suggest that increasing depth level relative to decreasing tumour thickness increases the probability of complete deep clearance, however this result is not statistically significant.

Figure 40: Probability of achieving a complete deep margin for *all* BCCs when considering the ratio of depth level: tumour thickness (mm).

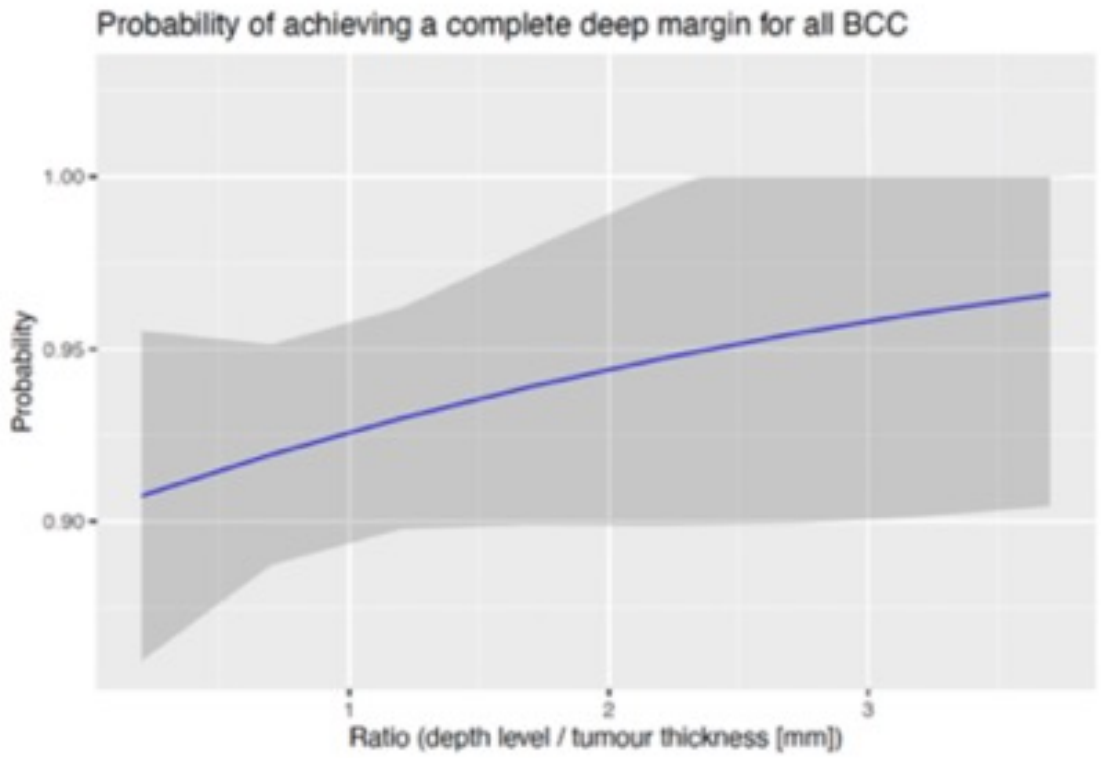


Table 43: Summary of logistic regression results assessing deep clearance subject to ratio and risk status.

Coefficients	OR	95% CI	p-value
Ratio	1.366	(0.778 – 2.850)	0.355
Low-risk	0.901	(0.294 – 3.101)	0.847

Table 44: Probability 95% CI of achieving a complete deep margin using a logistic regression model of ratio = depth level/tumour thickness for high-risk and low-risk lesions combined.

Ratio	Prediction	Low limit	High limit
0	0.90	0.84	0.96
0.5	0.91	0.88	0.95
1	0.93	0.90	0.96
1.5	0.94	0.90	0.97
2	0.94	0.90	0.99
2.5	0.95	0.90	1.00
3	0.96	0.90	1.00
3.5	0.96	0.90	1.00
4	0.97	0.91	1.00

9.3.5 Sensibility analysis based on variability in F1 score

Given the pivotal role of the F1 score in appraising the equilibrium between precision and recall, a sensibility analysis was undertaken to envisage the potential implications of its variability on study outcomes. For this analysis, two primary modelling assumptions were adopted. Firstly, that the influence of the F1 score on clearance is linear. Secondly, that fluctuations in the F1 score would yield proportional modifications in the reported outcomes. This assumption can be mathematically represented as:

Best-case scenario: $Adjustment\ factor = \frac{Upper\ CI\ of\ F1}{Reported\ F1}$.

Worse-case scenario: $Adjustment\ factor = \frac{Lower\ CI\ of\ F1}{Reported\ F1}$.

The reported F1 score in validation was 84.5%, with a 95% CI spanning from 73.0% to 95.1%. Drawing upon this data, two hypothetical scenarios were devised. In the best-case scenario, with the F1 score at its upper bound of 95.1%, the adjustment factor was determined to be 1.125. Under this assumption, the peripheral clearance for a 6mm clinical margin, previously cited at a 95.0% clearance rate, would essentially top out at 100.0% after adjustment. Similarly, the 11mm clinical margin, initially reported at 99.0%, would also peak at 100.0% following the adjustment. Pertaining to deep clearance, depth level 2's original clearance of 97.0% would be adjusted to 100.0%, and depth level 1, which had an earlier clearance rate of 92.0%, would also escalate to 100.0% post-adjustment. Conversely, the worst-case scenario emanates from the lower threshold of the F1 score's CI, 73.0%. The corresponding adjustment factor for this scenario is 0.8637. For peripheral clearance at a 6mm clinical margin, the original 95.0% clearance rate would be adjusted downwards to 82.1%. The 11mm clinical margin would see

its clearance rate diminish from the original 99.0% to 85.5%. In terms of deep clearance, depth level 2's clearance would recede from 97.0% to 83.9%, whilst depth level 1's clearance would drop from the initial 92.0% to 79.5%. These observations highlight the susceptibility of the results of this study to alterations in the F1 score.

9.4 Discussion

In this study it is shown that a novel and valid automated population-based approach to quality assurance in surgical oncology using state of the art NLP techniques to extract margin status from histopathology reports at scale with the most common human cancer as a use case. A web application was used to automate the analysis of incomplete excision rates, stratifying margin and risk before undertaking *post-hoc* analysis to investigate the relationship between uninvolved margin and speciality. Manual review of the free text in EHRs has been the mainstay of data capture and analysis in this setting. This is a process which is labour intensive, costly and open to error and bias. Chapter six demonstrates that comparing the NLP-based method with blinded expert clinicians can achieve high levels of percentage agreement (>90%). (Ali et al., 2022). This approach permits automated, rapid and large-scale health data interrogation associated with a significant time saving. Given the potential for downstream clinical impact of errors that may arise from independent use of the study model it is advocated that this model could be used as an adjunct with a 'human in the loop' embedded at the MDT rather than a fully autonomous system. This aligns with the Topol Review 'Preparing the healthcare workforce to deliver the digital future' which anticipates that genomics, digital medicine, AI and robotics will not replace healthcare professionals, but will enhance them, giving them more time to care for patient (Topol, 2019).

This is also the first study to date to use NLP for large volume assessment of the completeness of excision margins for BCC. Complete tumour excision is the cornerstone of the surgical management of cancer and despite the availability of destructive and topical treatments, surgical excision with a pre-determined clinical margin has been the mainstay of management of BCCs for decades. As microscopic tumour extension is not identifiable at the time of surgical excision the aim of a pre-determined clinical excision margin is to gain complete histological clearance in as high a number of cases as possible, whilst balanced against the functional and aesthetic considerations of larger excisions. These clinical margins serve only as a guide however, aimed at achieving a 1mm histological margin for oncological clearance. This means that while they provide a helpful benchmark, they do not guarantee complete excision in all cases.

9.4.1 Quality assurance of incomplete excision rates at scale

There is joint guidance from NICE and the BAD on the treatment of BCC which includes the planned surgical margins (Nasr et al., 2021; National Institute for Health and Care Excellence, 2006). The BAD states that individual operators and units should regularly audit their outcomes with a target of ≥95% complete excision rate being defined as acceptable. Accordingly, when considering the 5.5% rate of incomplete excision in this study the approach is externally valid. However, recent systematic review evidence suggests that the proportion of incomplete excisions is higher than previously reported (11.0% [95% CI 9.7–12.4]), with significant variation amongst specialists (Nolan et al., 2021). *Post-hoc* analysis revealed that plastic surgery had the lowest risk of an involved margin of any speciality which again mirrors previous work (Nolan et al., 2021).

There are multiple previous reports which describe text classification and information extraction for cancer using NLP. However, these predominantly focus on extracting common cancer characteristics (grade, stage and laterality etc) (Santos et al., 2022). In a recent systematic review of automatic classification of cancer pathology reports only one model was capable of extracting tumour margin status. Soysal et al described a hybrid approach combining rule-based NLP and machine learning-based NLP to extract tumour involvement in excisional margins with precision, recall and F1 score of 0.80, 0.79 and 0.80 respectively on a test set of 93 histopathology reports (Soysal et al., 2019). However, there are currently no reports of leveraging this approach on a population-based scale for ensuring quality assurance in surgical oncology (Mellia et al., 2021).

The principle of audit participation is to provide a means of measuring performance against agreed standards to support best practice. The Royal College of Surgeons of England state that surgeons should play an active role in quality assuring the data submitted to and reported by an audit, to ensure that the results accurately reflect their practice (Royal College of Surgeons England, 2024). Additionally, audit is an integral part of revalidation in the UK with the analysis of outcomes provides one piece of the supporting information required for revalidation. The Healthcare Quality Improvement Partnership (HQIP) is responsible for several national healthcare quality improvement programmes in the UK, including the National Clinical Audit and Patient Outcomes Programme (NCAPOP) on behalf of NHS England, the Welsh government and in some cases other devolved authorities (Healthcare Quality Improvement Partnership, 2024-b). Within the UK there is a growing trend for transparency of outcomes from surgery. Consultant Outcomes Publication (COP) is an NHS England initiative launched with ten audits in 2013 and managed by HQIP (Healthcare Quality Improvement Partnership, 2024-a). The number of audits publishing quality measures at the level of

individual consultant doctor expanded to twelve in 2014 and the data were also published on MyNHS website. Most quality assurance programs in the USA use easily obtainable claims or billing data. However, claims data are limited, inconsistent and subject to interpretation when used to measure quality (Steinberg et al., 2008). The American College of Surgeons National Surgical Quality Improvement Program® has been in operation since late 2004 and evaluates surgical quality and safety by feeding back valid, timely, risk-adjusted outcomes, which providers use to improve care (The American College of Surgeons, 2024).

Risk adjustment and case-mix adjustment are important factors to consider when reporting surgical outcome data (Cohen et al., 2013). Consider the following scenario, comparing two consultant (attending) surgeons. Consultant A has a low incomplete excision rate (2.0%) while Consultant B has a higher incomplete excision rate (10.0%). These unadjusted raw rates do not consider the patient, tumour or surgical factors (risk), or underlying comorbidities (case-mix) of the patients treated by each consultant. A well thought out risk adjustment approach therefore has to consider the underlying patient characteristics to adjust important outcome measures (Steinberg et al., 2008). In the present study risk was included as a covariate in the regression model when comparing incomplete excision rates across specialities and generating ORs.

The development and validation of rule-based NLP models require significant time investment. Alternative methods of data extraction are therefore being explored, such as ML, instead of rule-based NLP. However, other authors have commented how ML algorithms have not been able to demonstrate superior performance in comparison to rule-based techniques as described in this study (Harrison & Sidey-Gibbons, 2021). A nascent area of NLP research is the evolution of deep learning models. Only 3.0% of models in a recent systematic review and

meta-analysis of NLP in surgical outcomes research utilised deep learning, yet this is an area of AI research that is currently generating most interest (Mellia et al., 2021). The medical community has been relatively slow to adopt deep learning technology which is largely due to a short supply of anonymised, accessible and good quality healthcare specific data to train and test these models on.

Pretrained language models that use transformer-based technologies are one solution for zero-shot learning on a wide variety of NLP tasks without the need for annotated data. A transformer is a model that heavily relies on attention to boost the speed with which these models can be trained (Vaswani et al., 2017). There are several state of the art large open-source transformer language models available e.g. Clinical BERT and Med7 that are poised to increase the reach of deep learning in the medical community (Alsentzer et al., 2019; Kormilitzin et al., 2021).

9.4.2 Multivariate analysis with updated RCPath standards using NLP

In their systematic review, the BAD aimed to evaluate appropriate clinical margins for standard surgical excisions of BCC (Nasr et al., 2021). They compared non-standard clinical margins (<4mm, >5mm) to specified clinical margins (4-5mm) and assessed incomplete excision rates. Quazi et al. included 15 case-series encompassing 3,843 BCC lesions. Complete excision rates were 94.7% (5mm), 92.2% (4mm), 90.3% (3mm), and 88.2% (2mm). The review also included 29 non-comparative studies on high-risk BCC (head and neck) and whole-body BCC, with varying lesion numbers. Head and neck studies showed incomplete excision rates ranging from 3.5% to 20.7% for <4mm margins. The reference group (4-5mm) consisted of four small studies, while the >5mm group had two studies. Schell et al. found that a 4.75mm margin for low-risk and 8mm margin for high-risk tumours achieved at least a 95.0% complete excision rate. Meta-analyses showed that 4-5mm margins had lower incomplete excision rates (3.7% in

high-risk BCC and 3.7% in whole-body BCC) compared to smaller margins. Overall, the findings of these studies suggest that standard excision margins of 4-5mm are appropriate for surgical excisions of BCC. However, the data using the RCPath reporting schema suggests that a 4mm and 5mm peripheral margin would give an incomplete excision rate of 8.0% and 6.0% respectively. Given that BAD guidance provides a target rate of ≥95.0% for complete excision this study suggests that a peripheral margin of 6mm for both high and low-risk BCCs, providing a clearance rate of 95.0%, would instead be more appropriate which is at odds with current BAD guidelines.

Prior to the publication of the 2019 RCPath dataset for histopathological reporting of primary cutaneous BCC, the World Health Organization (WHO) classification of skin tumours, which was published in 2018, grouped the infiltrative, micronodular, and sclerosing/morphoeic subtypes of BCC together under the category of "high-risk BCC" (Elder et al., 2018). This was based on evidence that these subtypes have a higher potential for local recurrence, invasion, and metastasis compared to the more common nodular and superficial subtypes of BCC. The RCPath initially published National Minimum Datasets for the histopathological reporting of skin cancer in February 2002 (Slater & McKee, 2002). Early editions described the need to report histological subtype, the degree of differentiation, the presence or absence of mitotic activity and necrosis, and the involvement of margins. However, the guidelines did not specifically address the reporting of the infiltrative, micronodular, or sclerosing/morphoeic subtypes of BCC or provide guidance on grouping these subtypes as high-risk. Each dataset has been reviewed and issued as a standard with some subsequent adjustments to allow better conformity to the National Cancer Intelligence Network dataset. In 2014, the RCPath published an updated guideline on the reporting of skin tumours, which included more detailed recommendations for the reporting of BCC. The updated guideline recommended that the

histological subtype of BCC should be reported, and that additional features such as the depth of invasion, the presence of perineural invasion, and the degree of differentiation should also be included in the report. However, the guideline did not group any specific subtypes of BCC as high-risk. The 2019 Dataset for histopathological reporting of primary cutaneous BCC, which was developed by an international working group, provided more specific and comprehensive guidance for the reporting of BCC, including the reporting of the infiltrative, micronodular, and sclerosing/morphoeic subtypes, as well as recommendations for grouping certain subtypes as high-risk (Slater & Barrett, 2019-b). This dataset builds upon the previous guidelines from the WHO and the RCPath, and represents an important step forward in improving the accuracy and consistency of histopathological reporting of primary BCC. However, despite these calls to reform BCC histological reporting the results of the binary logistic regression analysis indicated that the risk status of the lesion (high or low) did not have a significant influence on whether the peripheral margin was clear. Interestingly these guidelines do not use ulceration as a negative prognostic factor. However, in this study it is observed that presence of ulceration has a negative impact on the likelihood of clearance. In addition, low risk BCCs <10mm in size at body area C could be managed with smaller margins to achieve the same clearance rates. This may explain why the observations did not show any difference in the clearance rates between low and high-risk lesions. These findings have important implications for clinical management and underscore the need for further refinement of risk stratification in the management of BCCs.

While clinical peripheral margin size has been relatively well studied, the depth to which a BCC should be excised to gain complete histological clearance is poorly evidenced. BAD guidelines recommend excision to a clear plane at the deep margin, including any deeper structures if necessary. However, this statement is not supported by evidence in the guidelines

(Nasr et al., 2021). Some studies suggest intraoperative assessment of deep margin status, although this is likely to be inaccurate when using clinical assessment alone, or expensive and time consuming if using intra-operative frozen section histology or MMS. It is therefore important to gain good quality evidence as to the appropriate clinical depth of excision to ensure adequate clearance in the majority of cases. Kiely et al. showed that excising to the first underlying anatomical plane resulted in uninvolved margins in 95.0.% of infiltrative or mixed infiltrative BCC, while subcutaneous fat was adequate for clearance in 95.0% of nodular, superficial, and mixed non-infiltrative BCC (Kiely & Patel, 2019). The data would support this. However, risk was not added in the regression model due to the low frequency of lesions that were both low-risk and had an incomplete deep margin. It is therefore difficult to comment on the impact of the RCPath standards on the recommended excision plane differentially for high-risk and low-risk lesions. However, it is shown that depth level 2 has the greatest probability of achieving deep clearance at all tumour thickness included in the analysis. In the investigation of the relationship between the probability of complete deep margin clearance of a skin cancer and the anatomical plane it is excised at, it was found that tumour depth potentially confounds this relationship.

In the pursuit of increased oncological safety with lower incomplete excision rates, the BAD recommend that individual operators and units should regularly audit their outcomes, with a target of ≥ 95.0%. This is rooted in empirical findings from national audits on NMSC excisions (Keith et al., 2017, 2020). Yet, while these recommendations provide clarity and a unified direction, they also bring forth questions about the broader implications of achieving such high rates. Would pushing for a higher complete excision rate necessitate wider and deeper clinical margins? And if so, at what cost? Every excision inherently brings forth risks. Therefore, decisions about excision margins must be made with a holistic perspective. The onus falls on

both the plastic surgery and dermatology communities to question whether the reduction of an incomplete excision rate by a few percentages justifies a potentially increased risk of complications or a compromise in aesthetic/functional outcomes. Secondary procedures, even if they afford a smaller initial clinical margin, bring with them added emotional stress, potential financial implications, and risks associated with surgery. On the contrary, for some patients, particularly those concerned about scarring or functional impairment, a secondary excision or even radiotherapy might be a more palatable trade-off than a larger primary excision or reconstruction. The debate on margins for BCCs should be seen not just as a matter of millimetres, but as a multifaceted issue interlacing oncological safety, aesthetic/functional considerations, patient preferences, and quality of life. The challenge lies in crafting an approach that seamlessly integrates all these facets, epitomising the essence of holistic patient care.

9.4.3 Strengths and limitations

Strengths of the current study include the very high volume of lesions included in the study over a 17-year period representing the largest global study of incomplete BCC excision rates. The largest previous single study to date was by Dhepnorrarat et al who reported 21,677 excised BCCs. Data were collected prospectively over a period of 6 years by a defined group of 25 plastic surgeons in Western Australia (Dhepnorrarat et al., 2009). The size of the datasets improves statistical power and reduces the risk of type II error. From an administrative perspective, the big data approach adopted in this study enabled analysis to be performed at a much faster rate, that is more cost efficient. Using NLP, the study provides novel insights into the importance of assessing peripheral margins in the context of the RCPath 2019 standards for the completeness of BCC excisions which no other study has performed to date. By utilising new statistical modelling techniques, added to the literature on the optimal deep margin for

complete excision of BCCs. Even with considerable missing data, the research includes a complete dataset for 1,447 lesions. This sizable dataset compares favourably with many studies cited in the BAD systematic review, which spanned 29 studies with an average lesion count of 744 (Nasr et al., 2021).

Analyses were automated for the purposes of descriptive statistical analysis. Binary logistic regression was performed on a manual *post-hoc* basis to investigate the relationship between uninvolved margin and speciality. By undertaking more complex back-end programming a fully automated analysis using the study's web-based application would be possible. Furthermore, other factors were not considered which may be associated with the risk of incomplete excision other than speciality e.g. grade of surgeon, use of operating loupes, the margin of normal tissue excised with the lesion and the *a priori* plan for reconstruction. These factors were not available for extraction from histopathology reports but could potentially negatively impact on incomplete excision rates. There is no specific pathway for allocating specialities for BCC surgery in Swansea Bay University Health Board. The risk and case-mix of the patients treated by each speciality is likely to be different but was accounted for in the regression model. This difference in case mix is seen nationally in many publications (Fleischer et al., 2001; Keith et al., 2020; Salmon et al., 2010; Twist, 2009).

The tumour status following treatment is described by the residual tumour (R) classification: R0, no residual tumour; R1, microscopic residual tumour; R2, macroscopic residual tumour (Hermanek & Wittekind, 1994). Oncological clearance or an R0 resection is key to preventing local recurrence in many tumour types. However, there are other tumour factors that may in addition to completeness of excision be relevant for prognostication and determining further management. As such these other factors may be subject to quality assurance scrutiny and may

need to be considered if the approach should be transferred and deployed across other tumour types for benchmarking outcomes. For example, the prognosis of potentially curable completely resected gastric cancer is primarily determined by pathologic T and N staging criteria. According to the 8th edition of the American Joint Committee on Cancer (AJCC) staging manual, the retrieval of at least 16 lymph nodes is the minimal requirement for lymph node dissection, and retrieval of 30 lymph nodes is more desirable (Amin et al., 2017).

Another factor to consider is the variance in the reporting structure and how this can potentially impact the success of any rule-based approach to NLP. There is evidence to that there are lower levels of compliance with minimum dataset reporting in NMSC when compared to melanoma (Barrett & Barrett, 2015). Performance of this approach could therefore differ when applied to other tumour types treated in surgical oncology. The base NLP model may need updating with fine-tuning on local data before this technology is scaled across other hospitals with possible different histopathological reporting structures.

Despite utilising an NLP algorithm on over 34,000 BCCs over a 17-year period, only 5.0% had complete data. The credibility of the results hinges on consistent and unbiased data documentation. There exists potential variability in how surgeons measure and define clinical margins and depth level, introducing inconsistencies in the data. To address this limitation, there should be consideration of the use of the UK National Histopathology Request Form for skin biopsies, developed by the Public Health England (PHE) Skin Site-Specific Reference Group which has been approved by the BAD (Slater & Barrett, 2019-b). Although the measured clinical peripheral margin is a non-core clinical item in this form, incorporating this information would significantly improve the completeness of data for future studies using NLP on histopathology reports for quality assurance in skin oncology. Despite the sensibility

analysis, it is paramount to emphasise that the study's foundation rests heavily on the accuracy of the NLP algorithm that fuels the multivariate analysis. As such, any inaccuracies or biases within the algorithm can cascade into the multivariate model, potentially impacting the validity of the conclusions. Thus, the findings should be interpreted with an added degree of caution.

9.5 Conclusion

NLP opens up a whole new realm of possibilities within healthcare. The development of tools to analyse text to the current standard has only happened relatively recently. Through clinical work at the frontline of cancer management, surgeons are in a unique position to make key hypothesis generating observations, and to subsequently address and drive the research agenda. Despite NLP being increasingly being used for a wide range of clinical applications, there is currently a considerable gap between NLP as an engineering discipline and translation into surgery. By collaborating with data scientists this gulf can be narrowed though increased access to data where NLP and clinical expertise can be shared between disciplines. The present study was able to generalise a validated rule-based NLP pipeline to new data within a Java™ Spring Boot web application framework. This has been incorporated into a clinical workflow of the largest global study of incomplete BCC excision rates as a use case that other surgical specialities can adopt. This study has shed light on the impact of updated RCPath histological reporting standards on clearance rates in the surgical management of BCC. The findings suggest that peripheral margin clearance is not influenced by RCPath risk criteria, and that larger peripheral margins may be necessary to achieve complete excision rates of ≥95.0%. The findings suggest that risk stratification should include ulceration in the future and future guidelines updates should consider the requirement for smaller clinical margins depending on anatomical sites and tumour diameter. This novel, big data approach to automated, rapid and large-scale health data interrogation has a large role to play in improving patient outcomes and

maintaining quality assurance in surgical oncology. This is a potentially scalable and transferrable method that can be deployed across a range of tumour types for benchmarking outcomes in surgical practice. Future models may wish to leverage pre-trained, zero-shot or fine-tuned unsupervised models to mitigate development cost and time required for development of rule-based NLP or supervised ML methods. To fully leverage the potential of NLP, future studies should consider using multiple algorithms on different data sources to feed logistic regression models. Overall, the study highlights the importance of continually reassessing and updating clinical standards to improve patient care.

Chapter Ten:

Final Discussion, Future Work and Conclusion

10.1 Final discussion

The amalgamation of findings across the studies contained within this thesis highlights critical challenges, opportunities and potential advancements in the arena of skin cancer MDT management. This research journey was underpinned by the pivotal question - "Can NLP be employed to analyse EHR documents effectively and automate the primary and secondary functions of the MDT, performing at a level that is comparable to human clinicians?" The aim was to investigate the utility and efficacy of NLP in sifting through healthcare datasets and subsequently inform the creation of a CDSS platform for a vSMDT through systematic reviews, development, and validation processes. Herein, I delve deeper into each aspect to comprehensively understand and contextualise the impact, implications and prospective directions.

The current state of SSMDTs in the UK reveals a pronounced incongruity with the quoracy standards established by NICE. A predominant number of SSMDTs demonstrate non-compliance, with a mere 26.0% aligning with quorate by membership. This inconsistency primarily stems from the notable absence of clinical oncology presence, elucidating a critical gap in membership composition. Moreover, substantial variations in operational costs are indicative of a lack of uniformity in service provision, impacting the overall effectiveness and efficiency of skin cancer management within the NHS. The geographical variations observed in cost and quoracy between England, Wales, Scotland, and Northern Ireland further accentuate the need for a standardised operational framework, aimed at aligning service provisions and management practices. To drive skin cancer MDT management within the NHS, strategic interventions focusing on stringent adherence to operational protocols are imperative, ensuring that patient care is not only uniform and cost-efficient but also of the highest quality.

The ongoing transformation in the format of MDT meetings in the wake of the COVID-19 pandemic has underscored the viability of a hybrid format, integrating both virtual and in-person interactions. The virtual format, although fraught with challenges in team working and the collegiate nature of MDT working, has showcased commendable resilience in maintaining crucial aspects such as recruitment, data security, and patient confidentiality. The collective preference emerging for a hybrid format delineates the perceived value in retaining the flexibility and accessibility offered by virtual interactions. To fully harness the advantages of this transformation in MDT delivery, efforts must be channelled toward refining connectivity, IT support, training, and staff integration. Enhancements in these domains are paramount to fostering a harmonious blend of virtual and in-person interactions, ensuring that the integrative approach is conducive to optimal team working and decision-making processes.

The integration of novel technologies like NLP represent an innovation solution to the issue highlighted by the scoping reviews. NLP, with its rule-based approach to NER in BCC has manifested potential in refining quality assurance, re-appraising guidelines and re-designing services. The conception of a fully automated, virtual, web-based service model is a testament to the transformative power of AI in augmenting clinical decision-making during MDTs. This development heralds a 'human in the loop' approach, conducive to facilitating protocolised streaming of ‘simple’ cases and mitigating human error. The inherent adaptability and scalability of such technological advancements make them prime candidates for widespread implementation across various tumour types and settings in the NHS.

The advent of GPT-based systems, such as ChatGPT, in clinical communication is a groundbreaking innovation, exemplifying the capability to generate clinical letters with high readability, correctness, and human-like qualities. The ability of AI to accurately convey

complex medical information in an accessible manner is crucial in bridging the communication gap between clinicians and patients. This innovation not only promises a substantial reduction in the workload on clinicians and secretaries but also augurs well for the standardisation and accessibility of information disseminated to patients.

The utilisation of automated systems and big data in health data interrogation is a paradigm shift in maintaining quality assurance in surgical oncology. The approach's potential scalability and transferability across various tumour types signify its potential to revolutionise outcomes benchmarking in surgical practice, ultimately improving patient outcomes. It is this integration of automation and large-scale data analysis that holds the promise of a new era in healthcare.

The interrogation of clinical margins in BCC through NLP and multivariate analysis has unearthed pivotal insights into the potential need for modifications in clinical peripheral margins. The continuous reassessment and recalibration of clinical standards are quintessential to adapting to the evolving landscapes of reporting schema and new treatments in skin cancer.

Reflecting on my use of regression analysis in this research, I can see several areas where a more structured and considered approach would have strengthened my conclusions. In both Chapter Eight, where I examined AI-generated clinic letters, and Chapter Nine, where I analysed incomplete excision rates in basal cell carcinoma using NLP, I had clear expectations about how different factors would relate to my outcomes. However, these expectations were often implicit rather than formally stated in advance. A more rigorous approach would have involved clearly setting out my hypotheses beforehand or *a priori*. A common problem in statistical analysis is data dredging – running multiple tests on the data without a clear hypothesis and then selectively reporting only significant findings. I now recognise that my

approach in some areas risked falling into this trap. For example, in Chapter Nine, when exploring predictors of incomplete excision, I included multiple variables in my regression model. This increases the likelihood of identifying false associations simply by chance. The use of directed acyclic graphs (DAGs), which help to visualise the relationships between variables and ensure that the right factors are adjusted for in a regression model, would have improved my approach. Had I constructed DAGs before running my analyses, I would have been better able to distinguish between confounders, mediators, and colliders, ensuring that my models adjusted for the correct factors without introducing bias. For example, in Chapter Eight, I examined correctness and humanness scores separately, but these two outcomes are likely related. Adjusting for correctness in the humanness model may have distorted the relationships between other variables. Similarly, in Chapter Nine, while I adjusted for risk status when comparing incomplete excision rates across specialties, I needed to be careful about collider bias, which can arise when adjusting for a variable that is itself influenced by multiple other factors. Another key issue in my approach was multicollinearity, which occurs when two or more predictor variables in a regression model are highly correlated. This can make it difficult to determine the individual effect of each variable and can lead to inflated standard errors, reducing the reliability of the estimated coefficients. In my analysis, this may have been a concern when adjusting for multiple clinical and surgical factors that could be interdependent. For instance, in Chapter Nine, the relationship between tumour characteristics, surgeon specialty, and margin status may have been influenced by overlapping variables such as anatomical site and tumour size. If multicollinearity was present, the regression model may have struggled to separate the independent effects of these factors. One way to detect multicollinearity would have been to calculate the variance inflation factor (VIF) for each predictor variable. If a VIF value was above a certain threshold (commonly 5 or 10), I would have needed to reconsider whether both variables should remain in the model. Potential

solutions include removing one of the correlated variables, combining them into a single composite variable, or using principal component analysis (PCA) to reduce dimensionality. A better approach would have been to first conduct univariable analysis, identifying which variables were strongly associated with the outcome, and then selecting the most relevant ones for inclusion in the multivariable model. This would have helped to avoid overfitting, where the model picks up random noise in the data rather than meaningful relationships, and would have ensured that my results were more robust and generalisable. A related issue is the so-called "Table 2 fallacy," where adjusted regression results are presented without enough context. Simply reporting which variables remain significant in a multivariable model does not necessarily provide insight into cause-and-effect relationships. In Chapter Nine, for instance, I adjusted for risk status when comparing incomplete excision rates across specialties, but this did not fully account for differences in surgical techniques, patient selection, or other unmeasured factors that may have influenced the results. Overcoming the Table 2 fallacy requires a shift in focus from simply reporting significant associations to interpreting results within a broader clinical and methodological framework. One way to do this is to avoid placing undue emphasis on p-values and instead present effect sizes and confidence intervals, which provide a clearer indication of the strength and precision of associations. Another strategy is to conduct sensitivity analyses to assess whether results remain consistent under different model specifications. This could include running models with and without certain variables, using alternative definitions for key predictors, or applying different statistical techniques such as propensity score matching or inverse probability weighting to account for confounding.

10.2 Future directions

The continued evolution of AI models will necessitate rigorous development and fine-tuning, aligning them with the continual advancements in medical knowledge and clinical practices.

Comparative evaluations involving various architectures, including transformers and LLMs, will be instrumental in fine-tuning and prompt engineering, optimising their applicability in diverse clinical contexts. This will entail a meticulous examination and integration of models to ensure they remain state-of-the-art, assimilating evolving medical knowledge and adjusting to emerging clinical practices, hence maintaining congruence with the intricate demands and pace of clinical practice.

Regulation and standards will indeed play a pivotal role in this journey, enforcing meticulous adherence to certifications and compliance with data protection norms such as General Data Protection Regulation (GDPR), thus underlining the paramount importance of patient privacy and data protection. The application of NHS Digital's Clinical Risk Management standards and Health Research Authority (HRA) definitions will be imperative, contributing to the lawful and ethical deployment of AI software and securing the domains against potential negligence and liabilities in clinical decision-making processes.

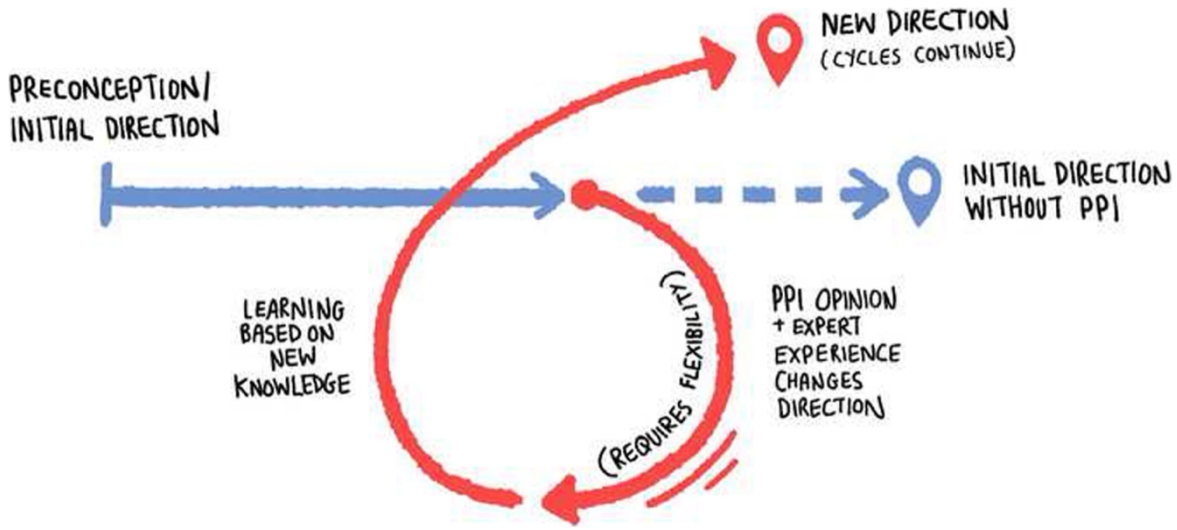
The validation and evaluation process of AI applications will undergo refinement, emphasising the critical appraisal of AI algorithms and the implementation of robust post-deployment monitoring, aimed at the continual refinement of systems. Prospective multi-centre clinical studies and stringent local model validations will become the bedrock, guaranteeing the safety, efficacy, and reliability of AI in real-world scenarios, and ensuring the systematic detection, management, and reporting of any adverse effects or discrepancies related to AI. Elevating the capabilities of NLP in clinical settings, particularly concerning training, development, and validation of models, calls for a cohesive international effort, standardised data labelling practices, and rich, accessible datasets. Having a substantial, well-labelled dataset is crucial, as it allows models to understand and generate medical text effectively. This warrants an

international collaboration to pool together anonymised and diverse clinical text data, which would enable models to be trained on varied and extensive information, thereby enhancing their predictive and analytical accuracy across different languages and medical specialties. Concurrently, adhering to common ontologies, or a universal system for data classification and labeling, ensures that the gathered data are coherent, interpretable, and useful for researchers and models across the globe. Together, these focal points will ensure the development of NLP models that are not only technically proficient but externally generalisable. There is also an absence of a standardised reporting approach in clinical NLP studies in cancer which compromises the consistency, comparability, and replicability of research outcomes. The fragmentation in NLP research reporting underscores the imperative for a Delphi study, which through a structured, iterative process, could harmonise expert insights from NLP researchers, clinical experts, and standards specialists to formulate widely applicable and consensus-driven reporting standards.

Engaging patients and the public in the development of the vSMDT is essential for ensuring that automated recommendations align with real-world needs (NIHR, n.d.; UK Research and Innovation, n.d.) (Figure 41). The UK Standards for Public Involvement outline key principles—such as flexibility, sharing, learning, and mutual respect—to elevate both quality and consistency in involvement practices (UK Standards for Public Involvement, n.d.). These standards emphasize accessible opportunities, valuing contributions, support and learning, clear communication, governance, and measuring impact, ensuring that involvement is meaningful and productive. Resources from INVOLVE highlight the importance of co-creation with patient advisory panels from the earliest stages, ensuring practical input on interfaces, workflows, and decision thresholds (INVOLVE, 2012). Evidence from digital health projects

also shows that sustained PPIE fosters stronger acceptability, trust, and relevance (Staniszewska et al., 2012; Ocloo & Matthews, 2016).

Figure 41: More than a method: trusting relationships, productive tensions, and two-way learning as mechanisms of authentic co-production. Courtesy of Knowles et al 2021.



In the area of integration and systems impact, the meticulous design and re-design of clinical workflows will be needed to facilitate the amalgamation of AI software with existing healthcare IT systems. Comprehensive evaluations will be needed to ascertain AI's impact on various domains such as service efficiency, health economic measures, and patient outcomes, with a clear understanding of the interplay between human cognitive biases and AI interactions. This focus aims to derive insights on how biases such as automation bias and alert fatigue might impinge upon clinical reasoning and decision-making processes.

A parallel focus on ethical and social implications is of paramount importance, propelling the narrative towards establishing ethical norms and frameworks that are ingrained with principles of patient autonomy, justice, beneficence, and non-maleficence. Continuous scrutiny and

ethical debates are necessitated to address and resolve concerns related to equity, accessibility, and inclusivity, reinforcing ethical norms and societal values in AI-enhanced healthcare. Beyond establishing ethical norms and frameworks, deeper engagement with key ethical dimensions such as prospective and retrospective responsibility, foreseeability, and the distribution of moral accountability is essential (Smith, Birchley, & Ives, 2024). Prospective responsibility entails anticipating and mitigating potential harms before deployment, ensuring that AI-driven recommendations uphold patient safety and clinical integrity. Conversely, retrospective responsibility addresses accountability when AI-driven decisions result in adverse outcomes, raising questions about liability and oversight, particularly when human oversight is limited. The risks associated with AI-driven decision-making have been particularly highlighted in healthcare insurance, where LLMs have been increasingly deployed to automate claims processing and risk assessment. Reports have linked AI-driven denial of care with widening inequities, exacerbating patient distrust, and even contributing to high-profile controversies (Ali & Dobbs, 2025). LLMs, while capable of improving efficiency, remain vulnerable to bias, hallucinations, and ethical blind spots, particularly when financial considerations take precedence over patient well-being. These concerns extend to CDSS in clinical settings, where the risk of algorithmic opacity, automated decision-making errors, and biases in training data could inadvertently reinforce healthcare disparities (Mittelstadt et al., 2016). Ensuring fairness requires careful dataset curation, ongoing algorithmic auditing, and ethical oversight to prevent biased outputs that disproportionately disadvantage certain patient populations (Morley et al., 2020). Additionally, AI models designed with built-in safeguards, such as uncertainty quantification and explainable AI (XAI) methods, could help ensure that CDSS recommendations remain transparent, clinically sound, and ethically defensible. As LLMs continue to evolve, their integration into clinical workflows should be guided by robust

ethical frameworks, continuous interdisciplinary oversight, and a commitment to equitable, patient-centered care.

The exploration into the strategic and cultural integration of AI in healthcare will be significant in future developments. The formation of multidisciplinary collaborations is emphasised, marking the confluence of academia and industry in bringing innovations to clinical frontiers. This synergetic convergence is pivotal in shaping the structural framework for AI software in healthcare, ensuring an aligned and holistic integration that adheres to the evolving needs and advancements in medical sciences.

10.3 Conclusion

The exploration conducted within this thesis has highlighted the transformative potential and challenges of employing NLP within skin cancer MDT management, particularly in enhancing and automating processes in EHR analysis and decision-making. Through systematic review, development, and validation processes aimed at optimising a CDSS for a vSMDT, significant insights into the current disparities and opportunities within SSMDTs in the UK, have been unveiled. The hybridisation of MDT meetings, leveraging both virtual and in-person formats, alongside the integration of innovative technological solutions like NLP and GPT-based systems, demonstrates a pivotal shift towards augmenting clinical communication and decision-making, whilst maintaining crucial elements of team functionality and patient confidentiality. The strategic and ethical incorporation of these automated systems necessitates continuous evolution, ensuring regulatory compliance, ethical integrity, and alignment with ongoing advancements in medical practices and knowledge. Future endeavours should focus on refining AI models and systems, enhancing international collaborations for data pooling,

establishing universal data classification systems, and perpetuating the enhancement and ethical oversight of AI applications within the clinical landscape.

Bibliography

Abide, J. M., Nahai, F., & Bennett, R. G. (1984). The meaning of surgical margins. *Plastic and Reconstructive Surgery*, *73*(3), 492–497. https://doi.org/10.1097/00006534-198403000-00030

Academy of Medical Royal Colleges. (2018). *Please, write to me. Writing outpatient clinical letters to patients guidance*. https://www.aomrc.org.uk/wp-content/uploads/2018/09/Please_write_to_me_Guidance_010918.pdf

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., Arachchi, H., Arora, A., Auman, J. T., Ayala, B., Baboud, J., Balasundaram, M., Balu, S., Barnabas, N., Bartlett, J., Bartlett, P., Bastian, B. C., Baylin, S. B., Behera, M., … Zou, L. (2015). Genomic Classification of Cutaneous Melanoma. *Cell*, *161*(7), 1681–1696. https://doi.org/10.1016/j.cell.2015.05.044

Ali, S. R., Abdulla, M., Ibrahim, N., Dobbs, T. D., Haj-Basheer, M., & Whitaker, I. S. (2021). The incidence and risk of involved margins in surgically resected basal cell carcinoma - A multi-centre consecutive case series. *Journal of Plastic, Reconstructive & Aesthetic Surgery : JPRAS*, *74*(11), 3196–3211. https://doi.org/10.1016/j.bjps.2021.08.012

Ali, S. R., Dobbs, T. D., Hutchings, H. A., & Whitaker, I. S. (2021). Composition, Quoracy and Cost of Specialist Skin Cancer Multidisciplinary Team Meetings in the United Kingdom. *Journal of Plastic, Reconstructive & Aesthetic Surgery : JPRAS*, *74*(12), 3335–3340. https://doi.org/10.1016/j.bjps.2021.05.019

Ali, S. R., Dobbs, T. D., Jovic, M., Hutchings, H. A., & Whitaker, I. S. (2023). Improving the effectiveness of multidisciplinary team meetings on skin cancer: Analysis of the National Cancer Research UK survey responses. *Journal of Plastic, Reconstructive & Aesthetic Surgery : JPRAS*, *82*, 141–151. https://doi.org/10.1016/j.bjps.2023.01.002

Ali, S. R., Dobbs, T. D., Jovic, M., Strafford, H., Fonferko-Shadrach, B., Lacey, A. S., Williams, N., Pickrell, W. O., Hutchings, H. A., & Whitaker, I. S. (2023). Validating a novel natural language processing pathway for automated quality assurance in surgical oncology: incomplete excision rates of 34 955 basal cell carcinomas. *The British Journal of Surgery*, *110*(9), 1072–1075. https://doi.org/10.1093/bjs/znad055

Ali, S. R., Dobbs, T. D., Mohamedbhai, H., Whitaker, S., Hutchings, H. A., & Whitaker, I. S. (2023). Evaluating remote skin cancer multidisciplinary team meetings in the United Kingdom post-COVID-19. *Journal of Plastic, Reconstructive & Aesthetic Surgery : JPRAS*, *84*, 250–257. https://doi.org/10.1016/j.bjps.2023.04.052

Ali, S. R., Strafford, H., Dobbs, T. D., Fonferko-Shadrach, B., Lacey, A. S., Pickrell, W. O., Hutchings, H. A., & Whitaker, I. S. (2022). Development and validation of an automated basal cell carcinoma histopathology information extraction system using natural language processing. *Frontiers in Surgery*, *9*, 870494. https://doi.org/10.3389/fsurg.2022.870494

Ali, S., & Dobbs, T. (2025). The promise and peril of AI in healthcare: A cautionary tale of UnitedHealthcare. Bulletin of the Royal College of Surgeons, 44. https://doi.org/10.1308/rcsbull.2025.44

Alsawas, M., Alahdab, F., Asi, N., Li, D. C., Wang, Z., & Murad, M. H. (2016). Natural language processing: use in EBM and a guide for appraisal. *Evidence-Based Medicine*, *21*(4), 136–138. https://doi.org/10.1136/ebmed-2016-110437

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). *Publicly Available Clinical BERT Embeddings*.

Amazon Web Services. (2024). *Amazon Comprehend Medical*. Retrieved March 6, 2024, from https://aws.amazon.com/comprehend/medical/

American College of Surgeons. (2012). *Cancer Program Standards 2012: Ensuring Patient-Centered Care*. www.facs.org

Amin, M. B., Greene, F. L., Edge, S. B., Compton, C. C., Gershenwald, J. E., Brookland, R. K., Meyer, L., Gress, D. M., Byrd, D. R., & Winchester, D. P. (2017). The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalised " approach to cancer staging. *CA: A Cancer Journal for Clinicians*, *67*(2), 93–99. https://doi.org/10.3322/caac.21388

Andrew, T. W., Hamnett, N., Roy, I., Garioch, J., Nobes, J., & Moncrieff, M. D. (2022). Machine-learning algorithm to predict multidisciplinary team treatment recommendations in the management of basal cell carcinoma. *British Journal of Cancer*, *126*(4), 562–568. https://doi.org/10.1038/s41416-021-01506-7

Armstrong, B. K., & Kricker, A. (2001). The epidemiology of UV induced skin cancer. *Journal of Photochemistry and Photobiology B: Biology*, *63*(1–3), 8–18. https://doi.org/10.1016/S1011-1344(01)00198-1

Arts, D. G. T., De Keizer, N. F., & Scheffer, G.-J. (2002). Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association : JAMIA*, *9*(6), 600–611. https://doi.org/10.1197/jamia.m1087

Association of Breast Surgery. (2024). *Breast MDTM Toolkit*. Retrieved March 6, 2024, from https://associationofbreastsurgery.org.uk/professionals/clinical/breast-mdtm-toolkit/

Back, M. F., Ang, E. L. L., Ng, W.-H., See, S.-J., Lim, C. C. T., Tay, L.-L., & Yeo, T.-T. (2007). Improvements in quality of care resulting from a formal multidisciplinary tumour clinic in the management of high-grade glioma. *Annals of the Academy of Medicine, Singapore*, *36*(5), 347–351.

Barrett, P. D., & Barrett, H. E. (2015). An audit into use of minimum dataset reporting of skin cancers in the North of England Cancer Network. *Pathogenesis*, *2*(1–2), 5–8. https://doi.org/10.1016/j.pathog.2015.05.003

Basta, Y. L., Bolle, S., Fockens, P., & Tytgat, K. M. A. J. (2017). The Value of Multidisciplinary Team Meetings for Patients with Gastrointestinal Malignancies: A Systematic Review. *Annals of Surgical Oncology*, *24*(9), 2669–2678. https://doi.org/10.1245/s10434-017-5833-3

Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, Spurr C, Khorasani R, Tanasijevic M, Middleton B. (2003). Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. Journal of the American Medical Informatics Association, 10(6), 523–530.

Beadle, P. C., Bullock, D., Bedford, G., Leach, J. F., Webb, R. A., Dent, N. A., & Burton, J. L. (1982). Accuracy of skin cancer incidence data in the United Kingdom. *Clinical and Experimental Dermatology*, *7*(3), 255–260. https://doi.org/10.1111/j.1365-2230.1982.tb02424.x

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2), 157-166.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta*, *760*, 25–33. https://doi.org/10.1016/j.aca.2012.11.007

Bikle, D. D. (2014). Vitamin D metabolism, mechanism of action, and clinical applications. *Chemistry & Biology*, *21*(3), 319–329. https://doi.org/10.1016/j.chembiol.2013.12.016

Bilfinger, T. V, Albano, D., Perwaiz, M., Keresztes, R., & Nemesure, B. (2018). Survival Outcomes Among Lung Cancer Patients Treated Using a Multidisciplinary Team Approach. *Clinical Lung Cancer*, *19*(4), 346–351. https://doi.org/10.1016/j.cllc.2018.01.006

Bishop, C. M. (1995a). Neural networks for pattern recognition. Oxford University Press.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Birchall, M., Bailey, D., King, P., & South West Cancer Intelligence Service Head and Neck Tumour Panel. (2004). Effect of process standards on survival of patients with head and neck cancer in the south and west of England. *British Journal of Cancer*, *91*(8), 1477–1481. https://doi.org/10.1038/sj.bjc.6602118

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Bowden, J., Brennan, P. A., Umar, T., & Cronin, A. (2002). Expression of vascular endothelial growth factor in basal cell carcinoma and cutaneous squamous cell carcinoma of the head and neck. *Journal of Cutaneous Pathology*, *29*(10), 585–589. https://doi.org/10.1034/j.1600-0560.2002.291003.x

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage Publications.

Brantsch, K. D., Meisner, C., Schönfisch, B., Trilling, B., Wehner-Caroli, J., Röcken, M., & Breuninger, H. (2008). Analysis of risk factors determining prognosis of cutaneous squamous-cell carcinoma: a prospective study. *The Lancet Oncology*, *9*(8), 713–720. https://doi.org/10.1016/S1470-2045(08)70178-5

Brash, D. E., Rudolph, J. A., Simon, J. A., Lin, A., McKenna, G. J., Baden, H. P., Halperin, A. J., & Pontén, J. (1991). A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proceedings of the National Academy of Sciences*, *88*(22), 10124–10128. https://doi.org/10.1073/pnas.88.22.10124

British Association of Dermatologists. (2018). *NHS England Reform of Cancer Multidisciplinary Team (MDT) Meetings: The Skin MDT Response Report* . https://cdn.bad.org.uk/uploads/2022/02/29200024/National-Reform-of-Cancer-MDTMs-Skin-MDT-Response.pdf

British Medical Association. (2024). *NHS Medical Staffing Data Analysis*. Retrieved March 6, 2024, from https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/workforce/nhs-medical-staffing-data-analysis

Brown, H., Wyatt, S., Croft, S., Gale, N., Turner, N., & Mulla, A. (2015). *Scoping the future: an evaluation of endoscopy capacity across the NHS in England*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901).

Bydder, S., Nowak, A., Marion, K., Phillips, M., & Atun, R. (2009). The impact of case discussion at a multidisciplinary team meeting on the treatment and survival of patients with inoperable non-small cell lung cancer. *Internal Medicine Journal*, *39*(12), 838–841. https://doi.org/10.1111/j.1445-5994.2009.02019.x

Cadet, J., Douki, T., & Ravanat, J.-L. (2015). Oxidatively generated damage to cellular DNA by UVB and UVA radiation. *Photochemistry and Photobiology*, *91*(1), 140–155. https://doi.org/10.1111/php.12368

Cadet, J., Sage, E., & Douki, T. (2005). Ultraviolet radiation-mediated damage to cellular DNA. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *571*(1–2), 3–17. https://doi.org/10.1016/j.mrfmmm.2004.09.012

Cake, R., Cavanagh, P., & Gordon, B. (2016). *Horizon scanning: an evaluation of imaging capacity across the NHS in England*.

Calman, K., & Hine, D. (1995). *A policy framework for commissioning cancer services. A report by the expert advisory group on cancer to the chief medical officers of England and Wales: guidance for purchasers and providers of cancer services*.

Cancer Research UK. (2024-a). *Cancer incidence for common cancers*. Retrieved March 6, 2024, from https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared#heading-Zero

Cancer Research UK. (2024-b). *Improving the effectiveness of multidisciplinary team meetings in cancer services*.

Cancer Research UK. (2024-c). *Melanoma skin cancer statistics*. Retrieved March 6, 2024, from https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer#heading-Zero

Cancer Research UK. (2024-d). *Non-melanoma skin cancer statistics*. Retrieved March 6, 2024, from https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-melanoma-skin-cancer

Cancer Research UK. (2021a). *Melanoma skin cancer incidence statistics*. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer/incidence

Cancer Research UK. (2021b). *Melanoma skin cancer mortality statistics*. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer/mortality

Cancer Research UK. (2021c). *Melanoma skin cancer survival statistics*. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer/survival#heading-Two

Cancer Research UK. (2021d). *Melanoma skin cancer survival statistics*. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer/survival#heading-Three

Cancer Research UK. (2021e). *Non-melanoma skin cancer incidence statistics*. Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-melanoma-skin-cancer/incidence

Cantor, A. B. (1996). Sample-size calculations for Cohen's kappa. *Psychological Methods*, *1*(2), 150–153. https://doi.org/10.1037/1082-989X.1.2.150

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., Hogg, D., Lorigan, P., Lebbe, C., Jouary, T., Schadendorf, D., Ribas, A., O'Day, S. J., Sosman, J. A., Kirkwood, J. M., … BRIM-3 Study Group. (2011). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England Journal of Medicine*, *364*(26), 2507–2516. https://doi.org/10.1056/NEJMoa1103782

Chinai, N., Bintcliffe, F., Armstrong, E. M., Teape, J., Jones, B. M., & Hosie, K. B. (2013). Does every patient need to be discussed at a multidisciplinary team meeting? *Clinical Radiology*, *68*(8), 780–784. https://doi.org/10.1016/j.crad.2013.02.011

Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2015). *How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?*

Cleaver, J. E. (1968). Defective repair replication of DNA in xeroderma pigmentosum. *Nature*, *218*(5142), 652–656. https://doi.org/10.1038/218652a0

Cohen, M. E., Ko, C. Y., Bilimoria, K. Y., Zhou, L., Huffman, K., Wang, X., Liu, Y., Kraemer, K., Meng, X., Merkow, R., Chow, W., Matel, B., Richards, K., Hart, A. J., Dimick, J. B., & Hall, B. L. (2013). Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. *Journal of the American College of Surgeons*, *217*(2), 336-46.e1. https://doi.org/10.1016/j.jamcollsurg.2013.02.027

Coory, M., Gkolia, P., Yang, I. A., Bowman, R. V, & Fong, K. M. (2008). Systematic review of multidisciplinary teams in the management of lung cancer. *Lung Cancer (Amsterdam, Netherlands)*, *60*(1), 14–21. https://doi.org/10.1016/j.lungcan.2008.01.008

Cunningham, H. (2002). GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, *36*(2), 223–254. http://www.jstor.org/stable/30204529

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, *27*(1), 11–28. http://www.jstor.org/stable/1473169

Daniels, A., Muloiwa, R., Myer, L., & Buys, H. (2021). Examining the reliability of ICD-10 discharge coding in Red Cross War Memorial Children's Hospital administrative database. *South African Medical Journal = Suid-Afrikaanse Tydskrif Vir Geneeskunde*, *111*(2), 137–142. https://doi.org/10.7196/SAMJ.2021.v111i2.14686

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W., Davis, N., Dicks, E., Ewing, R., Floyd, Y., Gray, K., Hall, S., Hawes, R., Hughes, J., Kosmidou, V., … Futreal, P. A. (2002). Mutations of the BRAF gene in human cancer. *Nature*, *417*(6892), 949–954. https://doi.org/10.1038/nature00766

de Gruijl, F. R. (2002). Photocarcinogenesis: UVA vs. UVB Radiation. *Skin Pharmacology and Physiology*, *15*(5), 316–320. https://doi.org/10.1159/000064535

De Ieso, P. B., Coward, J. I., Letsa, I., Schick, U., Nandhabalan, M., Frentzas, S., & Gore, M. E. (2013). A study of the decision outcomes and financial costs of multidisciplinary team meetings (MDMs) in oncology. *British Journal of Cancer*, *109*(9), 2295–2300. https://doi.org/10.1038/bjc.2013.586

de Vries, E., Micallef, R., Brewster, D. H., Gibbs, J. H., Flohil, S. C., Saksela, O., Sankila, R., Forrest, A. D., Trakatelli, M., Coebergh, J. W. W., Proby, C. M., & EPIDERM Group. (2012). Population-based estimates of the occurrence of multiple vs first primary basal cell carcinomas in 4 European regions. *Archives of Dermatology*, *148*(3), 347–354. https://doi.org/10.1001/archdermatol.2011.2244

Demner-Fushman D, Chapman WW, McDonald CJ. (2009). What can natural language processing do for clinical decision support? Journal of Biomedical Informatics, 42(5), 760–772.

Dennis, L. K., Vanbeek, M. J., Beane Freeman, L. E., Smith, B. J., Dawson, D. V, & Coughlin, J. A. (2008). Sunburns and risk of cutaneous melanoma: does age matter? A comprehensive meta-analysis. *Annals of Epidemiology*, *18*(8), 614–627. https://doi.org/10.1016/j.annepidem.2008.04.006

Department for Business, I. and S. (2013). *Eight Great Technologies*.

Department of Health. (2000). *The NHS Cancer Plan: A Plan for Investment, A Plan for Reform*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Dhepnorrarat, R. C., Lee, M. A., & Mountain, J. A. (2009). Incompletely excised skin cancer rates: a prospective study of 31,731 skin cancer excisions by the Western Australian Society of Plastic Surgeons. *Journal of Plastic, Reconstructive & Aesthetic Surgery : JPRAS*, *62*(10), 1281–1285. https://doi.org/10.1016/j.bjps.2008.04.028

Digital Regulations Innovation. (2024). *Understanding UK MDR 2002 regulations for medical devices*. Retrieved March 6, 2024, from https://www.digitalregulations.innovation.nhs.uk/developers-guidance/all-developers-guidance/uk-mdr-2002-understanding-regulations-medical-devices/

Dobbie, S., Strafford, H., Pickrell, W. O., Fonferko-Shadrach, B., Jones, C., Akbari, A., Thompson, S., & Lacey, A. (2021). Markup: A Web-Based Annotation Tool Powered by Active Learning. *Frontiers in Digital Health*, *3*, 598916. https://doi.org/10.3389/fdgth.2021.598916

Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, *125*, 37–46. https://doi.org/10.1016/j.ijmedinf.2019.02.008

Drews J. (2000). Drug discovery: a historical perspective. Science, 287(5460), 1960–1964.

Drury, D. J., Kaur, A., Dobbs, T., & Whitaker, I. S. (2021). The Readability of Outpatient Plastic Surgery Clinic Letters: Are We Adhering to Plain English Writing Standards? *Plastic Surgical Nursing : Official Journal of the American Society of Plastic and Reconstructive Surgical Nurses*, *41*(1), 27–33. https://doi.org/10.1097/PSN.0000000000000356

Edwards, N., & Imison, C. (2024). *How can dermatology services meet current and future patient needs, while ensuring quality of care is not compromised and access is equitable across the UK?* Retrieved March 6, 2024, from https://silo.tips/download/nigel-edwards-and-candace-imison

Eggermont, A. M. M., Blank, C. U., Mandala, M., Long, G. V., Atkinson, V., Dalle, S., Haydon, A., Lichinitser, M., Khattak, A., Carlino, M. S., Sandhu, S., Larkin, J., Puig, S., Ascierto, P. A., Rutkowski, P., Schadendorf, D., Koornstra, R., Hernandez-Aya, L., Maio, M., … Robert, C. (2018). Adjuvant Pembrolizumab versus Placebo in Resected Stage III Melanoma. *New England Journal of Medicine*, *378*(19), 1789–1801. https://doi.org/10.1056/NEJMoa1802357

Eggermont, A. M. M., & Kirkwood, J. M. (2004). Re-evaluating the role of dacarbazine in metastatic melanoma: what have we learned in 30 years? *European Journal of Cancer (Oxford, England : 1990)*, *40*(12), 1825–1836. https://doi.org/10.1016/j.ejca.2004.04.030

Eguia H, Sánchez-Bocanegra CL, Vinciarelli F, Alvarez-Lopez F, Saigí-Rubió F. (2024). Clinical decision support and natural language processing in medicine: systematic literature review. Journal of Medical Internet Research, 26, e55315.

Eide, M. J., Tuthill, J. M., Krajenta, R. J., Jacobsen, G. R., Levine, M., & Johnson, C. C. (2012). Validation of claims data algorithms to identify nonmelanoma skin cancer. *The Journal of Investigative Dermatology*, *132*(8), 2005–2009. https://doi.org/10.1038/jid.2012.98

Elder, D., Massi, D., Scoyler, R., & Willemze, R. (2018). *WHO Classification of Skin Tumours* (4th ed., Vol. 11).

Epstein, E. H. (2008). Basal cell carcinomas: attack of the hedgehog. *Nature Reviews Cancer*, *8*(10), 743–754. https://doi.org/10.1038/nrc2503

Fargnoli, M. C., Gandini, S., Peris, K., Maisonneuve, P., & Raimondi, S. (2010). MC1R variants increase melanoma risk in families with CDKN2A mutations: a meta-analysis.

European Journal of Cancer (Oxford, England : 1990), *46*(8), 1413–1420. https://doi.org/10.1016/j.ejca.2010.01.027

Faries, M. B., Thompson, J. F., Cochran, A. J., Andtbacka, R. H., Mozzillo, N., Zager, J. S., Jahkola, T., Bowles, T. L., Testori, A., Beitsch, P. D., Hoekstra, H. J., Moncrieff, M., Ingvar, C., Wouters, M. W. J. M., Sabel, M. S., Levine, E. A., Agnese, D., Henderson, M., Dummer, R., … Elashoff, R. M. (2017). Completion Dissection or Observation for Sentinel-Node Metastasis in Melanoma. *The New England Journal of Medicine*, *376*(23), 2211–2222. https://doi.org/10.1056/NEJMoa1613210

Fitzpatrick, T. (1975). Soleil et peau [Sun and skin]. *Journal de Medecine Esthetique*, *2*, 33–34.

Fleischer, A. B., Feldman, S. R., Barlow, J. O., Zheng, B., Hahn, H. B., Chuang, T. Y., Draft, K. S., Golitz, L. E., Wu, E., Katz, A. S., Maize, J. C., Knapp, T., & Leshin, B. (2001). The specialty of the treating physician affects the likelihood of tumour-free resection margins for basal cell carcinoma: results from a multi-institutional retrospective study. *Journal of the American Academy of Dermatology*, *44*(2), 224–230. https://doi.org/10.1067/mjd.2001.110396

Fleissig, A., Jenkins, V., Catt, S., & Fallowfield, L. (2006). Multidisciplinary teams in cancer care: are they effective in the UK? *The Lancet. Oncology*, *7*(11), 935–943. https://doi.org/10.1016/S1470-2045(06)70940-8

Fletcher, C., Bridge, J., Hogendoorn, P., & Mertens, F. (2013). *WHO Classification of Tumours of Soft Tissue and Bone* (5th ed., Vol. 4). IARC Press.

Forrest, L. M., McMillan, D. C., McArdle, C. S., & Dunlop, D. J. (2005). An evaluation of the impact of a multidisciplinary team, in a single centre, on treatment and survival in patients with inoperable non-small-cell lung cancer. *British Journal of Cancer*, *93*(9), 977–978. https://doi.org/10.1038/sj.bjc.6602825

Fouad, Y. A., & Aanei, C. (2017). Revisiting the hallmarks of cancer. *American Journal of Cancer Research*, *7*(5), 1016–1036.

Gabel, M., Hilton, N. E., & Nathanson, S. D. (1997). Multidisciplinary breast cancer clinics. Do they work? *Cancer*, *79*(12), 2380–2384.

Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Abeni, D., Boyle, P., & Melchi, C. F. (2005). Meta-analysis of risk factors for cutaneous melanoma: I. Common and atypical naevi. *European Journal of Cancer (Oxford, England : 1990)*, *41*(1), 28–44. https://doi.org/10.1016/j.ejca.2004.10.015

Gastman, B. R. (2016). *Cutaneous Malignancies: A Surgical Perspective*.

General Medical Council. (2024). *Domain 3: Communication partnership and teamwork*. Retrieved March 6, 2024, from https://www.gmc-uk.org/professional-standards/professional-standards-for-doctors/good-medical-practice/domain-3---communication-partnership-and-teamwork#paragraph-31

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet. Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gheini, M., Ren, X., & May, J. (2021). Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation. arXiv preprint arXiv:2104.08771. Retrieved from https://arxiv.org/abs/2104.08771

Gibson, J. A. G., Cordaro, A., Dobbs, T. D., Griffiths, R., Akbari, A., Whitaker, S., Hutchings, H. A., Lyons, R. A., & Whitaker, I. S. (2021). The association between immunosuppression and skin cancer in solid organ transplant recipients: a control-matched cohort study of 2,852 patients. *European Journal of Dermatology : EJD*, *31*(6), 712–721. https://doi.org/10.1684/ejd.2021.4108

Gibson, J. A. G., Dobbs, T. D., Griffiths, R., Song, J., Akbari, A., Whitaker, S., Watkins, A., Langan, S. M., Hutchings, H. A., Lyons, R. A., & Whitaker, I. S. (2020). The association of smoking and socioeconomic status on cutaneous melanoma: a population-based, data-linkage, case-control study. *The British Journal of Dermatology*, *182*(5), 1136–1147. https://doi.org/10.1111/bjd.18526

Gibson, J. A. G., Dobbs, T. D., Kouzaris, L., Lacey, A., Thompson, S., Akbari, A., Hutchings, H. A., Lineaweaver, W. C., Lyons, R. A., & Whitaker, I. S. (2021a). Making the Most of Big Data in Plastic Surgery: Improving Outcomes, Protecting Patients, Informing Service Providers. *Annals of Plastic Surgery*, *86*(3), 351–358. https://doi.org/10.1097/SAP.0000000000002434

Gibson, J. A. G., Dobbs, T. D., Kouzaris, L., Lacey, A., Thompson, S., Akbari, A., Hutchings, H. A., Lineaweaver, W. C., Lyons, R. A., & Whitaker, I. S. (2021b). Making the Most of Big Data in Plastic Surgery: Improving Outcomes, Protecting Patients, Informing Service Providers. *Annals of Plastic Surgery*, *86*(3), 351–358. https://doi.org/10.1097/SAP.0000000000002434

Goette, D. K. (1981). Topical chemotherapy with 5-fluorouracil. *Journal of the American Academy of Dermatology*, *4*(6), 633–649. https://doi.org/10.1016/S0190-9622(81)80196-X

Goldberg, S. I., Niemierko, A., & Turchin, A. (2008). Analysis of data errors in clinical research databases. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2008*, 242–246.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Gorlin, R. J. (2004). Nevoid basal cell carcinoma (Gorlin) syndrome. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *6*(6), 530–539. https://doi.org/10.1097/01.gim.0000144188.15902.c4

Gospodarowicz, M. K., Brierley, J. D., & Wittekind, C. (2016). *TNM Classification of Malignant Tumours* (8th ed.). John Wiley & Sons.

Grice, E. A., & Segre, J. A. (2011). The skin microbiome. *Nature Reviews. Microbiology*, *9*(4), 244–253. https://doi.org/10.1038/nrmicro2537

Han, J., & Morag, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. In J. Mira & F. Sandoval (Eds.), From Natural to Artificial Neural Computation (Lecture Notes in Computer Science, Vol. 930, pp. 195–201). Springer. https://doi.org/10.1007/3-540-59497-3_175

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, *100*(1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

Harrison, C. J., & Sidey-Gibbons, C. J. (2021). Machine learning in medicine: a practical introduction to natural language processing. *BMC Medical Research Methodology*, *21*(1), 158. https://doi.org/10.1186/s12874-021-01347-1

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Healthcare Quality Improvement Partnership. (2024-a). *Consultant Outcomes Publication*. Retrieved March 6, 2024, from https://www.hqip.org.uk/national-programmes/clinical-outcomes-publication/

Healthcare Quality Improvement Partnership. (2024-b). *Healthcare Quality Improvement Partnership*. Retrieved March 6, 2024, from https://www.hqip.org.uk/

Hemmi, H., Kaisho, T., Takeuchi, O., Sato, S., Sanjo, H., Hoshino, K., Horiuchi, T., Tomizawa, H., Takeda, K., & Akira, S. (2002). Small anti-viral compounds activate immune cells via the TLR7 MyD88–dependent signaling pathway. *Nature Immunology*, *3*(2), 196–200. https://doi.org/10.1038/ni758

Hermanek, P., & Wittekind, C. (1994). Residual tumour (R) classification and prognosis. *Seminars in Surgical Oncology*, *10*(1), 12–20. https://doi.org/10.1002/ssu.2980100105
Hocker, T., & Tsao, H. (2007). Ultraviolet radiation and melanoma: a systematic review and analysis of reported sequence variants. *Human Mutation*, *28*(6), 578–588. https://doi.org/10.1002/humu.20481
Hoinville, L., Taylor, C., Zasada, M., Warner, R., Pottle, E., & Green, J. (2019). Improving the effectiveness of cancer multidisciplinary team meetings: analysis of a national survey of MDT members' opinions about streamlining patient discussions. *BMJ Open Quality*, *8*(2), e000631. https://doi.org/10.1136/bmjoq-2019-000631
Holbrook, K. A., & Odland, G. F. (1975). The fine structure of developing human epidermis: light, scanning, and transmission electron microscopy of the periderm. *The Journal of Investigative Dermatology*, *65*(1), 16–38. https://doi.org/10.1111/1523-1747.ep12598029
Holick, M. F. (2004). Sunlight and vitamin D for bone health and prevention of autoimmune diseases, cancers, and cardiovascular disease. *The American Journal of Clinical Nutrition*, *80*(6), 1678S-1688S. https://doi.org/10.1093/ajcn/80.6.1678S
Houssami, N., & Sainsbury, R. (2006). Breast cancer: multidisciplinary care and clinical outcomes. *European Journal of Cancer (Oxford, England : 1990)*, *42*(15), 2480–2491. https://doi.org/10.1016/j.ejca.2006.05.023
Hrabač, P., & Trkulja, V. (2019). What are the odds you will read this article? *Croatian Medical Journal*, *60*(1), 53–54. https://doi.org/10.3325/cmj.2019.60.53
Hunt DL, Haynes RB, Hanna SE, Smith K. (1998). Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA, 280(15), 1339–1346.
Ibrahim, N., Ali, S. R., Dobbs, T. D., Gibson, A. G. J., Hutchings, H. A., & Whitaker, S. I. (2023). The incidence of non-melanoma skin cancer in the UK and the Republic of Ireland: a systematic review. *European Journal of Dermatology : EJD*, *33*(3), 218–229. https://doi.org/10.1684/ejd.2023.4496
Ibrahim, N., Gibson, J., Ali, S., Dobbs, T., & Whitaker, I. S. (2021). Is poor quality non-melanoma skin cancer data affecting high quality research and patient care? *Journal of Plastic, Reconstructive & Aesthetic Surgery : JPRAS*, *74*(6), 1355–1401. https://doi.org/10.1016/j.bjps.2020.12.036
Ibrahim, N., Jovic, M., Ali, S., Williams, N., Gibson, J. A. G., Griffiths, R., Dobbs, T. D., Akbari, A., Lyons, R. A., Hutchings, H. A., & Whitaker, I. S. (2023). The epidemiology, healthcare and societal burden of basal cell carcinoma in Wales 2000-2018: a retrospective nationwide analysis. *The British Journal of Dermatology*, *188*(3), 380–389. https://doi.org/10.1093/bjd/ljac090
Independent Cancer Taskforce. (2015). *Achieving world-class cancer outcomes: a strategy for England 2015 – 2020.*
INVOLVE (2012). Briefing notes for researchers: Public involvement in NHS, public health and social care research. INVOLVE. Retrieved March 2, 2025, from https://ncto.ie/wp-content/uploads/2022/10/1.-Briefing-notes-for-researchers-involving-the-public-in-NHS-public-health-and-social-care-research-by-INVOLVE-2012-1.pdf
Itin, P. H., & Fistarol, S. K. (2004). Ectodermal dysplasias. *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics*, *131C*(1), 45–51. https://doi.org/10.1002/ajmg.c.30033
Jagannatha, A., & yu, hong. (2016). Structured prediction models for RNN based sequence labeling in clinical text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 856–865. https://doi.org/10.18653/v1/D16-1082

James Lind Alliance (JLA). (2022). *Top 10 research priorities for skin cancer surgery*. https://www.jla.nihr.ac.uk/priority-setting-partnerships/skin-cancer-surgery/top-10-priorities.htm

Janis, J. E., Kwon, R. K., & Attinger, C. E. (2011). The new reconstructive ladder: modifications to the traditional model. *Plastic and Reconstructive Surgery*, *127 Suppl 1*, 205S-212S. https://doi.org/10.1097/PRS.0b013e318201271c

Janssen, A., Robinson, T., Brunner, M., Harnett, P., Museth, K. E., & Shaw, T. (2018). Multidisciplinary teams and ICT: a qualitative study exploring the use of technology and its impact on multidisciplinary team meetings. *BMC Health Services Research*, *18*(1), 444. https://doi.org/10.1186/s12913-018-3242-3

Jie, Z., Zhiying, Z., & Li, L. (2021). A meta-analysis of Watson for Oncology in clinical application. *Scientific Reports*, *11*(1), 5792. https://doi.org/10.1038/s41598-021-84973-5

Johnson, M., Anderson, P., Dras, M., & Steedman, M. (2018). Predicting accuracy on large datasets from smaller pilot data. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 450–455. https://doi.org/10.18653/v1/P18-2072

Juckett, D. (2012). A method for determining the number of documents needed for a gold standard corpus. *Journal of Biomedical Informatics*, *45*(3), 460–470. https://doi.org/10.1016/j.jbi.2011.12.010

Junor, E. J., Hole, D. J., & Gillis, C. R. (1994). Management of ovarian cancer: referral to a multidisciplinary team matters. *British Journal of Cancer*, *70*(2), 363–370. https://doi.org/10.1038/bjc.1994.307

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Vol. 2).

Kane, B., Luz, S., O'Briain, D. S., & McDermott, R. (2007). Multidisciplinary team meetings and their impact on workflow in radiology and pathology departments. *BMC Medicine*, *5*, 15. https://doi.org/10.1186/1741-7015-5-15

Kawamoto K, Houlihan CA, Balas EA, Lobach DF. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ, 330(7494), 765.

Ke, K. M., Blazeby, J. M., Strong, S., Carroll, F. E., Ness, A. R., & Hollingworth, W. (2013). Are multidisciplinary teams in secondary care cost-effective? A systematic review of the literature. *Cost Effectiveness and Resource Allocation : C/E*, *11*(1), 7. https://doi.org/10.1186/1478-7547-11-7

Keith, D. J., Bray, A. P., Brain, A., Mohd Mustapa, M. F., Barrett, H. E., Lane, S., Emmerich, M., Jakes, A., Barrett, P. D., & de Berker, D. A. R. (2020). British Association of Dermatologists (BAD) National Audit on Non-Melanoma Skin Cancer Excision 2016 in collaboration with the Royal College of Pathologists. *Clinical and Experimental Dermatology*, *45*(1), 48–55. https://doi.org/10.1111/ced.14034

Keith, D. J., de Berker, D. A. R., Bray, A. P., Cheung, S. T., Brain, A., & Mohd Mustapa, M. F. (2017). British Association of Dermatologists' national audit on nonmelanoma skin cancer excision, 2014. *Clinical and Experimental Dermatology*, *42*(1), 46–53. https://doi.org/10.1111/ced.12990

Keith L. Moore, T. V. N. Persaud, & Mark G. Torchia. (2018). *The Developing Human Clinically Oriented Embryology* (11th ed.). Elsevier.

Kelly, S. L., Jackson, J. E., Hickey, B. E., Szallasi, F. G., & Bond, C. A. (2013). Multidisciplinary clinic care improves adherence to best practice in head and neck cancer. *American Journal of Otolaryngology*, *34*(1), 57–60. https://doi.org/10.1016/j.amjoto.2012.08.010

Keohane, S. G., Botting, J., Budny, P. G., Dolan, O. M., Fife, K., Harwood, C. A., Mallipeddi, R., Marsden, J. R., Motley, R. J., Newlands, C., Proby, C., Rembielak, A., Slater, D. N., Smithson, J. A., Buckley, P., Fairbrother, P., Hashme, M., Mohd Mustapa, M. F., & Exton, L. S. (2021). British Association of Dermatologists guidelines for the management of people with cutaneous squamous cell carcinoma 2020*. *British Journal of Dermatology*, *184*(3), 401–414. https://doi.org/10.1111/bjd.19621

Kerr, R. (2019). The future of surgery. *The Bulletin of the Royal College of Surgeons of England*, *101*(7), 264–267. https://doi.org/10.1308/rcsbull.2019.264

Kesson, E. M., Allardice, G. M., George, W. D., Burns, H. J. G., & Morrison, D. S. (2012). Effects of multidisciplinary team working on breast cancer survival: retrospective, comparative, interventional cohort study of 13 722 women. *BMJ (Clinical Research Ed.)*, *344*, e2718. https://doi.org/10.1136/bmj.e2718

Kiely, J. R., & Patel, A. J. K. (2019). A retrospective study of 694 Basal Cell Carcinoma excisions to quantify deep margin documentation and clearance compared to histological type and surgical margin. *Journal of Plastic, Reconstructive & Aesthetic Surgery : JPRAS*, *72*(11), 1805–1812. https://doi.org/10.1016/j.bjps.2019.06.002

Kirkwood, J. M., Strawderman, M. H., Ernstoff, M. S., Smith, T. J., Borden, E. C., & Blum, R. H. (1996). Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group Trial EST 1684. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *14*(1), 7–17. https://doi.org/10.1200/JCO.1996.14.1.7

Kittler, H., Pehamberger, H., Wolff, K., & Binder, M. (2002). Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, *3*(3), 159–165. https://doi.org/10.1016/S1470-2045(02)00679-4

Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*, *26*(4), 364–379. https://doi.org/10.1093/jamia/ocy173

Kormilitzin, A., Vaci, N., Liu, Q., & Nevado-Holgado, A. (2021). Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, *118*, 102086. https://doi.org/10.1016/j.artmed.2021.102086

Kowal-Vern, A., & Criswell, B. K. (2005). Burn scar neoplasms: a literature review and statistical analysis. *Burns : Journal of the International Society for Burn Injuries*, *31*(4), 403–413. https://doi.org/10.1016/j.burns.2005.02.015

Krag, D. N., Weaver, D. L., Alex, J. C., & Fairbank, J. T. (1993). Surgical resection and radiolocalization of the sentinel lymph node in breast cancer using a gamma probe. *Surgical Oncology*, *2*(6), 335–339; discussion 340. https://doi.org/10.1016/0960-7404(93)90064-6

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, *73*, 14–29. https://doi.org/10.1016/j.jbi.2017.07.012

Kumar, P. R., Moore, J. A., Bowles, K. M., Rushworth, S. A., & Moncrieff, M. D. (2021). Mitochondrial oxidative phosphorylation in cutaneous melanoma. *British Journal of Cancer*, *124*(1), 115–123. https://doi.org/10.1038/s41416-020-01159-y

Kuo, R. Y. L., Harrison, C. J., Jones, B. E., Geoghegan, L., & Furniss, D. (2021). Perspectives: A surgeon's guide to machine learning. *International Journal of Surgery (London, England)*, *94*, 106133. https://doi.org/10.1016/j.ijsu.2021.106133

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lansbury, L., Bath-Hextall, F., Perkins, W., Stanton, W., & Leonardi-Bee, J. (2013). Interventions for non-metastatic squamous cell carcinoma of the skin: systematic review and pooled analysis of observational studies. *BMJ (Clinical Research Ed.)*, *347*, f6153. https://doi.org/10.1136/bmj.f6153

Lansbury, L., Leonardi-Bee, J., Perkins, W., Goodacre, T., Tweed, J. A., & Bath-Hextall, F. J. (2010). Interventions for non-metastatic squamous cell carcinoma of the skin. *The Cochrane Database of Systematic Reviews*, *4*, CD007869. https://doi.org/10.1002/14651858.CD007869.pub2

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. IEEE Communications Magazine, 27(11), 41–46.

Li, Yang & Yang, Tao. (2017). Word Embedding for Understanding Natural Language: A Survey. 10.1007/978-3-319-53817-4.

Li, Y., Miao, N., Ma, L., Shuang, F., & Huang, X. (2023). Transformer for object detection: Review and benchmark. Engineering Applications of Artificial Intelligence, 126(Part C), 107021. https://doi.org/10.1016/j.engappai.2023.107021

Licitra, L., Keilholz, U., Tahara, M., Lin, J.-C., Chomette, P., Ceruse, P., Harrington, K., & Mesia, R. (2016). Evaluation of the benefit and use of multidisciplinary teams in the treatment of head and neck cancer. *Oral Oncology*, *59*, 73–79. https://doi.org/10.1016/j.oraloncology.2016.06.002

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open, 3*, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

Lomas, A., Leonardi-Bee, J., & Bath-Hextall, F. (2012). A systematic review of worldwide incidence of nonmelanoma skin cancer. *British Journal of Dermatology*, *166*(5), 1069–1080. https://doi.org/10.1111/j.1365-2133.2012.10830.x

Lomas, A., Leonardi-Bee, J., & Bath-Hextall, F. (2012). A systematic review of worldwide incidence of nonmelanoma skin cancer. *The British Journal of Dermatology*, *166*(5), 1069–1080. https://doi.org/10.1111/j.1365-2133.2012.10830.x

Long, G. V, Hauschild, A., Santinami, M., Atkinson, V., Mandalà, M., Chiarion-Sileni, V., Larkin, J., Nyakas, M., Dutriaux, C., Haydon, A., Robert, C., Mortier, L., Schachter, J., Schadendorf, D., Lesimple, T., Plummer, R., Ji, R., Zhang, P., Mookerjee, B., … Kirkwood, J. M. (2017). Adjuvant Dabrafenib plus Trametinib in Stage III BRAF-Mutated Melanoma. *The New England Journal of Medicine*, *377*(19), 1813–1823. https://doi.org/10.1056/NEJMoa1708539

Lott, J. P., Boudreau, D. M., Barnhill, R. L., Weinstock, M. A., Knopp, E., Piepkorn, M. W., Elder, D. E., Knezevich, S. R., Baer, A., Tosteson, A. N. A., & Elmore, J. G. (2018). Population-Based Analysis of Histologically Confirmed Melanocytic Proliferations Using Natural Language Processing. *JAMA Dermatology*, *154*(1), 24–29. https://doi.org/10.1001/jamadermatol.2017.4060

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*.

MacDermid, E., Hooton, G., MacDonald, M., McKay, G., Grose, D., Mohammed, N., & Porteous, C. (2009). Improving patient survival with the colorectal cancer multi-disciplinary team. *Colorectal Disease : The Official Journal of the Association of Coloproctology of Great Britain and Ireland*, *11*(3), 291–295. https://doi.org/10.1111/j.1463-1318.2008.01580.x

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Marteijn, J. A., Lans, H., Vermeulen, W., & Hoeijmakers, J. H. J. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Reviews Molecular Cell Biology*, *15*(7), 465–481. https://doi.org/10.1038/nrm3822

Mazzaferro, V., & Majno, P. (2011). Principles for the best multidisciplinary meetings. *The Lancet. Oncology*, *12*(4), 323–325. https://doi.org/10.1016/S1470-2045(11)70010-9

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). Addendum: International evaluation of an AI system for breast cancer screening. *Nature*, *586*(7829), E19. https://doi.org/10.1038/s41586-020-2679-9

Medical Research Council. (2024). *Medical Research Council Annual Report and Accounts 2016/2017*. Retrieved March 6, 2024, from https://assets.publishing.service.gov.uk/media/5a820203e5274a2e87dc0a8b/mrc-annual-report-accounts-2016-17-web.pdf

Mellia, J. A., Basta, M. N., Toyoda, Y., Othman, S., Elfanagely, O., Morris, M. P., Torre-Healy, L., Ungar, L. H., & Fischer, J. P. (2021). Natural Language Processing in Surgery: A Systematic Review and Meta-analysis. *Annals of Surgery*, *273*(5), 900–908. https://doi.org/10.1097/SLA.0000000000004419

Meyer, V. M., Benjamens, S., Moumni, M. El, Lange, J. F. M., & Pol, R. A. (2022). Global Overview of Response Rates in Patient and Health Care Professional Surveys in Surgery: A Systematic Review. *Annals of Surgery*, *275*(1), e75–e81. https://doi.org/10.1097/SLA.0000000000004078

Micali, G., Lacarrubba, F., Dinotta, F., Massimino, D., & Nasca, M. R. (2010). Treating skin cancer with topical cream. *Expert Opinion on Pharmacotherapy*, *11*(9), 1515–1527. https://doi.org/10.1517/14656566.2010.481284

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (Vol. 26, pp. 3111–3119).

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (Third). SAGE Publications.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1-21.

Mohamedbhai, H., Fernando, S., Ubhi, H., Chana, S., & Visavadia, B. (2021). Advent of the virtual multidisciplinary team meeting: do remote meetings work? *The British Journal of Oral & Maxillofacial Surgery*, *59*(10), 1248–1252. https://doi.org/10.1016/j.bjoms.2021.05.015

Mohs, F. E. (1978). Chemosurgery for the microscopically controlled excision of cutaneous cancer. *Head & Neck Surgery*, *1*(2), 150–166. https://doi.org/10.1002/hed.2890010209

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods, and research to translate principles into practices. Science and Engineering Ethics, 26(4), 2141-2168.

Moncrieff, M. D., & Nobes, J. P. (2021). Commentary on the British Association of Dermatologists UK basal cell carcinoma guidelines 2021: all together now …. *The British Journal of Dermatology*, *185*(5), 877. https://doi.org/10.1111/bjd.20694

Moonesinghe, S. R., Mythen, M. G., Das, P., Rowan, K. M., & Grocott, M. P. W. (2013). Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. *Anesthesiology*, *119*(4), 959–981. https://doi.org/10.1097/ALN.0b013e3182a4e94d

Morton, C. A., Birnie, A. J., & Eedy, D. J. (2014). British Association of Dermatologists' guidelines for the management of squamous cell carcinoma in situ (Bowen's disease)

2014. *The British Journal of Dermatology*, *170*(2), 245–260. https://doi.org/10.1111/bjd.12766

Morton, D. L., Wen, D. R., Wong, J. H., Economou, J. S., Cagle, L. A., Storm, F. K., Foshag, L. J., & Cochran, A. J. (1992). Technical details of intraoperative lymphatic mapping for early stage melanoma. *Archives of Surgery (Chicago, Ill. : 1960)*, *127*(4), 392–399. https://doi.org/10.1001/archsurg.1992.01420040034005

Mosterd, K., Krekels, G. A. M., Nieman, F. H., Ostertag, J. U., Essers, B. A. B., Dirksen, C. D., Steijlen, P. M., Vermeulen, A., Neumann, H., & Kelleners-Smeets, N. W. J. (2008). Surgical excision versus Mohs' micrographic surgery for primary and recurrent basal-cell carcinoma of the face: a prospective randomised controlled trial with 5-years' follow-up. *The Lancet. Oncology*, *9*(12), 1149–1156. https://doi.org/10.1016/S1470-2045(08)70260-2

Movano. (2025). AI-driven accuracy in medical chatbots: Lessons from EvieAI. The Verge. https://www.theverge.com/2025/1/7/24337403/movano-evie-ring-ai-chatbot-ces-health-tech-wearables

Munro, A. J. (2015). Multidisciplinary Team Meetings in Cancer Care: An Idea Whose Time has Gone? *Clinical Oncology (Royal College of Radiologists (Great Britain))*, *27*(12), 728–731. https://doi.org/10.1016/j.clon.2015.08.008

Murray, P. V, O'Brien, M. E. R., Sayer, R., Cooke, N., Knowles, G., Miller, A. C., Varney, V., Rowell, N. P., Padhani, A. R., MacVicar, D., Norton, A., Ashley, S., & Smith, I. E. (2003). The pathway study: results of a pilot feasibility study in patients suspected of having lung carcinoma investigated in a conventional chest clinic setting compared to a centralised two-stop pathway. *Lung Cancer (Amsterdam, Netherlands)*, *42*(3), 283–290. https://doi.org/10.1016/s0169-5002(03)00358-1

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. Proceedings of the 27th International Conference on Machine Learning (ICML'10), 807–814. Omnipress, Madison, WI, USA.

Narayanan, D. L., Saladi, R. N., & Fox, J. L. (2010). Review: Ultraviolet radiation and skin cancer. *International Journal of Dermatology*, *49*(9), 978–986. https://doi.org/10.1111/j.1365-4632.2010.04474.x

Nasr, I., McGrath, E. J., Harwood, C. A., Botting, J., Buckley, P., Budny, P. G., Fairbrother, P., Fife, K., Gupta, G., Hashme, M., Hoey, S., Lear, J. T., Mallipeddi, R., Mallon, E., Motley, R. J., Newlands, C., Newman, J., Pynn, E. V., Shroff, N., … Ezejimofor, M. C. (2021). British Association of Dermatologists guidelines for the management of adults with basal cell carcinoma 2021*. *British Journal of Dermatology*, *185*(5), 899–920. https://doi.org/10.1111/bjd.20524

National Cancer Action Team. (2011). *National Cancer Peer Review Programme Report 2010/2011*.

National Cancer Peer Review Programme. (2011). *National Cancer Peer Review Programme Report 2010/2011*.

National Cancer Peer Review-National Cancer Action Team. (2011). *Manual for Cancer Services: Service User Partnership Group Measures*.

National Institute for Health and Care Excellence. (2006). *Improving outcomes for people with skin tumours including melanoma Cancer service guideline [CSG8]*. https://www.nice.org.uk/guidance/csg8

National Institute for Health and Care Excellence. (2015). *Melanoma: assessment and management NICE guideline [NG14]*. www.nice.org.uk/guidance/ng14

National Institute for Health and Care Excellence. (2016). *Nivolumab in combination with ipilimumab for treating advanced melanoma. Technology appraisal guidance [TA400]*. https://www.nice.org.uk/guidance/ta400
National Institute for Health and Care Excellence. (2018). *Dabrafenib with trametinib for adjuvant treatment of resected BRAF V600 mutation-positive melanoma Technology appraisal guidance [TA544]*. https://www.nice.org.uk/guidance/ta544/chapter/1-Recommendations
National Institute for Health and Care Excellence. (2021). *Artificial intelligence in mammography*. https://www.nice.org.uk/advice/mib242/resources/artificial-intelligence-in-mammography-pdf-2285965629587653
National Institute for Health and Care Excellence. (2022-a). *Melanoma: assessment and management [B] Evidence review for the use of sentinel lymph node biopsy in people with melanoma NICE guideline NG14*.
National Institute for Health and Care Excellence. (2022-b). *Melanoma: assessment and management [C] Evidence review for surgical and histological excision margins for people with stage 0 to II melanoma NICE guideline NG14*.
National Institute for Health and Care Excellence. (2022-c). *Pembrolizumab for adjuvant treatment of completely resected stage 3 melanoma. Technology appraisal guidance [TA766]*. https://www.nice.org.uk/guidance/ta766/chapter/1-Recommendations
National Institute for Health and Care Excellence (NICE). (2015, June). *Suspected cancer: recognition and referral [NG12]*. https://www.nice.org.uk/guidance/ng12/chapter/Recommendations-organised-by-site-of-cancer#skin-cancers
National Institute of Health and Care Excellence. (2021). *National Institute for Health and Care Excellence. Nivolumab for treating squamous non-small-cell lung cancer after chemotherapy. Technology appraisal guidance [TA684]* . https://www.nice.org.uk/guidance/ta684/chapter/1-Recommendations
National Institute for Health and Care Research (n.d.). Working with People and Communities. National Institute for Health and Care Research. Retrieved March 2, 2025, from https://www.nihr.ac.uk/research-funding/application-support/working-with-people-and-communities
National Literacy Trust. (2024). *What is literacy?* Retrieved March 6, 2024, from https://literacytrust.org.uk/information/what-is-literacy/
Netscher, D. T., Leong, M., Orengo, I., Yang, D., Berg, C., & Krishnan, B. (2011). Cutaneous Malignancies: Melanoma and Nonmelanoma Types. *Plastic and Reconstructive Surgery*, *127*(3), 37e–56e. https://doi.org/10.1097/PRS.0b013e318206352b
NHS Digital. (2024). *SNOMED CT*. Retrieved March 6, 2024, from https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct
NHS Employer Services. (2018). *Pay and Conditions Circular (M&D) 3/2018*.
NHS England. (2024-a). *2022/23 priorities and operational planning guidance*. Retrieved March 6, 2024, from https://www.england.nhs.uk/wp-content/uploads/2022/02/20211223-B1160-2022-23-priorities-and-operational-planning-guidance-v3.2.pdf
NHS England. (2024-b). Get Data Out: Skin Grouping. Retrieved March 6, 2024, from https://www.cancerdata.nhs.uk/getdataout/Skin_Grouping#:~:text=All%20tumours%20registered%20or%20imputed,have%20two%20%27First%20tumours%27.
NHS England and NHS Improvement. (2020). *Streamlining Multi-Disciplinary Team Meetings: Guidance for Cancer Alliances*.

NHS Executive Group. (1996). *Guidance on commissioning cancer services: improving outcomes in breast cancer - the manual*.
Nikolaou, V., & Stratigos, A. J. (2014). Emerging trends in the epidemiology of melanoma. *British Journal of Dermatology*, *170*(1), 11–19. https://doi.org/10.1111/bjd.12492
Nolan, G. S., Kiely, A. L., Totty, J. P., Wormald, J. C. R., Wade, R. G., Arbyn, M., & Jain, A. (2021). Incomplete surgical excision of keratinocyte skin cancers: a systematic review and meta-analysis. *The British Journal of Dermatology*, *184*(6), 1033–1044. https://doi.org/10.1111/bjd.19660
Ocloo J, Matthews R. (2016). From tokenism to empowerment: progressing patient and public involvement in healthcare improvement. BMJ Quality & Safety, 25(8), 626–632.
Patkar, V., Acosta, D., Davidson, T., Jones, A., Fox, J., & Keshtgar, M. (2012). Using computerised decision support to improve compliance of cancer multidisciplinary meetings with evidence-based guidance. *BMJ Open*, *2*(3). https://doi.org/10.1136/bmjopen-2011-000439
Peng, D., Cheng, Y.-X., & Cheng, Y. (2021). Improved Overall Survival of Colorectal Cancer under Multidisciplinary Team: A Meta-Analysis. *BioMed Research International*, *2021*, 5541613. https://doi.org/10.1155/2021/5541613
Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543).
Perenyei, M., Barber, Z. E., Gibson, J., Hemington-Gorse, S., & Dobbs, T. D. (2021). Anaphylactic Reaction Rates to Blue Dyes Used for Sentinel Lymph Node Mapping: Systematic Review and Meta-analysis. *Annals of Surgery*, *273*(6), 1087–1093. https://doi.org/10.1097/SLA.0000000000004061
Please, write to me. Writing outpatient clinic letters to patients guidance. (2018, September). Academy of Medical Royal Colleges. https://www.aomrc.org.uk/wp-content/uploads/2018/09/Please_write_to_me_Guidance_010918.pdf
Pons, E., Braun, L. M. M., Hunink, M. G. M., & Kors, J. A. (2016). Natural Language Processing in Radiology: A Systematic Review. *Radiology*, *279*(2), 329–343. https://doi.org/10.1148/radiol.16142770
Proksch, E., Brandner, J. M., & Jensen, J.-M. (2008). The skin: an indispensable barrier. *Experimental Dermatology*, *17*(12), 1063–1072. https://doi.org/10.1111/j.1600-0625.2008.00786.x
Public Health England. (2015). *Breast screening: programme overview*. https://www.gov.uk/guidance/breast-screening-programme-overview
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*.
Reynolds, I. S., & Fennelly, D. (2022). Immunotherapy in surgical oncology. *The British Journal of Surgery*, *110*(1), 3–5. https://doi.org/10.1093/bjs/znac385
Ribas, A., & Wolchok, J. D. (2018). Cancer immunotherapy using checkpoint blockade. *Science*, *359*(6382), 1350–1355. https://doi.org/10.1126/science.aar4060
Rosenblatt, F. (1962). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.Spartan.
Royal College of Surgeons England. (2024). *National Audit*. Retrieved March 6, 2024, from https://www.rcseng.ac.uk/standards-and-research/support-for-surgeons-and-services/audit/national-audit/
Rubin, A. I., Chen, E. H., & Ratner, D. (2005). Basal-Cell Carcinoma. *New England Journal of Medicine*, *353*(21), 2262–2269. https://doi.org/10.1056/NEJMra044151

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. Mc- Clelland, and the PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations, pp. 318–362. MIT Press. Reprinted in Anderson and Rosenfeld (1988).

Saini, K. S., Taylor, C., Ramirez, A.-J., Palmieri, C., Gunnarsson, U., Schmoll, H. J., Dolci, S. M., Ghenne, C., Metzger-Filho, O., Skrzypski, M., Paesmans, M., Ameye, L., Piccart-Gebhart, M. J., & de Azambuja, E. (2012). Role of the multidisciplinary team in breast cancer management: results from a large international survey involving 39 countries. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, *23*(4), 853–859. https://doi.org/10.1093/annonc/mdr352

Salmon, P., Mortimer, N., Rademaker, M., Adams, L., Stanway, A., & Hill, S. (2010). Surgical excision of skin cancer: the importance of training. *The British Journal of Dermatology*, *162*(1), 117–122. https://doi.org/10.1111/j.1365-2133.2009.09548.x

Sánchez-Gutiérrez, M. E., & González-Pérez, P. P. (2022). Multi-Class Classification of Medical Data Based on Neural Network Pruning and Information-Entropy Measures. *Entropy*, *24*(2), 196. https://doi.org/10.3390/e24020196

Santos, T., Tariq, A., Gichoya, J. W., Trivedi, H., & Banerjee, I. (2022). Automatic Classification of Cancer Pathology Reports: A Systematic Review. *Journal of Pathology Informatics*, *13*, 100003. https://doi.org/10.1016/j.jpi.2022.100003

Savova, G. K., Masanz, J. J., Ogren, P. V, Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, *17*(5), 507–513. https://doi.org/10.1136/jamia.2009.001560

Schäfer, M., & Werner, S. (2008). Cancer as an overhealing wound: an old hypothesis revisited. *Nature Reviews Molecular Cell Biology*, *9*(8), 628–638. https://doi.org/10.1038/nrm2455

Shay, J. W., & Wright, W. E. (2019). Telomeres and telomerase: three decades of progress. *Nature Reviews Genetics*, *20*(5), 299–309. https://doi.org/10.1038/s41576-019-0099-1

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics*, *7*(2), e12239. https://doi.org/10.2196/12239

Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, Tang PC. (2001). Clinical decision support systems for the practice of evidence-based medicine. Journal of the American Medical Informatics Association, 8(6), 527–534.

Slater, D., & Barrett, P. (2019-a). *Dataset for histopathological reporting of primary cutaneous basal cell carcinoma*.

Slater, D., & Barrett, P. (2019-b). *Dataset for histopathological reporting of primary cutaneous basal cell carcinoma*. https://www.rcpath.org/static/53688094-791e-4aaa-82cec42c3cb65e35/Dataset-for-histopathological-reporting-of-primary-cutaneous-basal-cell-carcinoma.pdf

Slater, D., & Barrett, P. (2019-c). *Dataset for histopathological reporting of primary invasive cutaneous squamous cell carcinoma and regional lymph nodes*.

Slater, D., & Cook, M. (2019). *Dataset for the histological reporting of primary cutaneous malignant melanoma and regional lymph nodes.*

Slater, D., & McKee, P. (2002). *Minimum dataset for the histopathological reporting of common skin cancers*.

Smeets, N. W. J., Krekels, G. A. M., Ostertag, J. U., Essers, B. A. B., Dirksen, C. D., Nieman, F. H. M., & Neumann, H. A. M. (2004). Surgical excision vs Mohs'

micrographic surgery for basal-cell carcinoma of the face: randomised controlled trial. *Lancet (London, England)*, *364*(9447), 1766–1772. https://doi.org/10.1016/S0140-6736(04)17399-6

Smith, H., Birchley, G., & Ives, J. (2024). Artificial intelligence in clinical decision-making: Rethinking personal moral responsibility. Bioethics, 38(1), 78-86.

Soukup, T., Lamb, B. W., Sevdalis, N., & Green, J. S. (2020). Streamlining cancer multidisciplinary team meetings: challenges and solutions. *British Journal of Hospital Medicine (London, England : 2005)*, *81*(3), 1–6. https://doi.org/10.12968/hmed.2020.0024

Soysal, E., Warner, J. L., Wang, J., Jiang, M., Harvey, K., Jain, S. K., Dong, X., Song, H.-Y., Siddhanamatha, H., Wang, L., Dai, Q., Chen, Q., Du, X., Tao, C., Yang, P., Denny, J. C., Liu, H., & Xu, H. (2019). Developing Customizable Cancer Information Extraction Modules for Pathology Reports Using CLAMP. *Studies in Health Technology and Informatics*, *264*, 1041–1045. https://doi.org/10.3233/SHTI190383

Staniszewska S, Brett J, Mockford C, Barber R. (2012). The GRIPP checklist: strengthening the quality of patient and public involvement reporting in research. International Journal of Technology Assessment in Health Care, 28(3), 235–240.

Steinberg, S. M., Popa, M. R., Michalek, J. A., Bethel, M. J., & Ellison, E. C. (2008). Comparison of risk adjustment methodologies in surgical quality improvement. *Surgery*, *144*(4), 662–667; discussion 662-7. https://doi.org/10.1016/j.surg.2008.06.010

Stephens, M. R., Lewis, W. G., Brewster, A. E., Lord, I., Blackshaw, G. R. J. C., Hodzovic, I., Thomas, G. V, Roberts, S. A., Crosby, T. D. L., Gent, C., Allison, M. C., & Shute, K. (2006). Multidisciplinary team management is associated with improved outcomes after surgery for esophageal cancer. *Diseases of the Esophagus : Official Journal of the International Society for Diseases of the Esophagus*, *19*(3), 164–171. https://doi.org/10.1111/j.1442-2050.2006.00559.x

Sutor, P., Aloimonos, Y., Fermuller, C., & Summers-Stay, D. (2019). Metaconcepts: Isolating context in word embeddings. In 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 544–549). IEEE. https://doi.org/10.1109/MIPR.2019.00110

Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digital Medicine, 3, 17.

Tanghetti, E., & Werschler, P. (2007). Comparison of 5% 5-fluorouracil cream and 5% imiquimod cream in the management of actinic keratoses on the face and scalp. *Journal of Drugs in Dermatology : JDD*, *6*(2), 144–147.

Taylor, C., Atkins, L., Richardson, A., Tarrant, R., & Ramirez, A.-J. (2012). Measuring the quality of MDT working: an observational approach. *BMC Cancer*, *12*, 202. https://doi.org/10.1186/1471-2407-12-202

Telfer, N. R., Colver, G. B., Morton, C. A., & British Association of Dermatologists. (2008). Guidelines for the management of basal cell carcinoma. *The British Journal of Dermatology*, *159*(1), 35–48. https://doi.org/10.1111/j.1365-2133.2008.08666.x

The American College of Surgeons. (2024). *National Cancer Database*. Retrieved March 6, 2024, from https://www.facs.org/quality-programs/cancer-programs/national-cancer-database/

The National Archives. (2024). *The Medical Devices Regulations 2002*. Retrieved March 6, 2024, from https://www.legislation.gov.uk/uksi/2002/618/contents

Thissen, M. R., Neumann, M. H., & Schouten, L. J. (1999). A systematic review of treatment modalities for primary basal cell carcinomas. *Archives of Dermatology*, *135*(10), 1177–1183. https://doi.org/10.1001/archderm.135.10.1177

Thomson, J., Hogan, S., Leonardi-Bee, J., Williams, H. C., & Bath-Hextall, F. J. (2020). Interventions for basal cell carcinoma of the skin. *The Cochrane Database of Systematic Reviews*, *11*(11), CD003412. https://doi.org/10.1002/14651858.CD003412.pub3

Topol, E. (2019). *The Topol Review: Preparing the healthcare workforce to deliver the digital future*. https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf

Trzeciak JR. (2001). High-throughput screening strategies in drug discovery. Drug Discovery Today, 6(2), 69–78.

Twist, M. (2009). Rate of incomplete excision of basal cell carcinomas by General Practitioners with Special Interest. *The British Journal of Dermatology*, *161*(1), 187. https://doi.org/10.1111/j.1365-2133.2009.09129.x

UK Research and Innovation (n.d.). Patients and the Public. Retrieved March 2, 2025, from https://www.ukri.org/councils/epsrc/guidance-for-applicants/what-to-include-in-your-proposal/health-technologies-impact-and-translation-toolkit/stakeholder-engagement/patients-and-the-public/

UK Standards for Public Involvement (n.d.). Better public involvement for better health and social care research. Developed by the UK Public Involvement Standards Development Partnership. https://sites.google.com/nihr.ac.uk/pistandards/home

US Department of Health and Human Services Office of Disease Prevention and Health Promotion. (2010). *National Action Plan to Improve Health Literacy*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems*.

Vaughan, D. (2024). *Multiclass Averaging*. Retrieved March 6, 2024, from https://yardstick.tidymodels.org/articles/multiclass.html?q=micro%20a

Venables, Z. C., Nijsten, T., Wong, K. F., Autier, P., Broggio, J., Deas, A., Harwood, C. A., Hollestein, L. M., Langan, S. M., Morgan, E., Proby, C. M., Rashbass, J., & Leigh, I. M. (2019). Epidemiology of basal and cutaneous squamous cell carcinoma in the U.K. 2013-15: a cohort study. *The British Journal of Dermatology*, *181*(3), 474–482. https://doi.org/10.1111/bjd.17873

Venus, M., Waterman, J., & McNab, I. (2010). Basic physiology of the skin. *Surgery (Oxford)*, *28*(10), 469–472. https://doi.org/10.1016/j.mpsur.2010.07.011

Warburg, O. (1924). Über den Stoffwechsel der Carcinomzelle. *Die Naturwissenschaften*, *12*(50), 1131–1137. https://doi.org/10.1007/BF01504608

Warren, H., Ali, S. R., Abdulla, M., Ibrahim, N., Haj-Basheer, M., Dobbs, T. D., & Whitaker, I. S. (2022). What depth of surgical excision results in adequate histological deep margin clearance in basal cell carcinoma? A retrospective cohort study of 1126 basal cell carcinomas. *Journal of Plastic, Reconstructive & Aesthetic Surgery : JPRAS*, *75*(7), 2387–2440. https://doi.org/10.1016/j.bjps.2022.02.066

Weber, J., Mandala, M., Del Vecchio, M., Gogas, H. J., Arance, A. M., Cowey, C. L., Dalle, S., Schenker, M., Chiarion-Sileni, V., Marquez-Rodas, I., Grob, J.-J., Butler, M. O., Middleton, M. R., Maio, M., Atkinson, V., Queirolo, P., Gonzalez, R., Kudchadkar, R. R., Smylie, M., … Ascierto, P. A. (2017). Adjuvant Nivolumab versus Ipilimumab in Resected Stage III or IV Melanoma. *New England Journal of Medicine*, *377*(19), 1824–1835. https://doi.org/10.1056/NEJMoa1709030

Whiteman, D. C., Whiteman, C. A., & Green, A. C. (2001). Childhood sun exposure as a risk factor for melanoma: a systematic review of epidemiologic studies. *Cancer Causes & Control : CCC*, *12*(1), 69–82. https://doi.org/10.1023/a:1008980919928

Widrow, B. and M. E. Hoff (1960). Adaptive switching circuits. In IRE WESCON Convention Record, Volume 4, pp. 96–104. Reprinted in Anderson and Rosenfeld (1988).

Winters, D. A., Soukup, T., Sevdalis, N., Green, J. S. A., & Lamb, B. W. (2021). The cancer multidisciplinary team meeting: in need of change? History, challenges and future perspectives. *BJU International*, *128*(3), 271–279. https://doi.org/10.1111/bju.15495

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: Practical machine learning tools and techniques (3rd ed.). Morgan Kaufmann.

Workforce Team. (2024). *NHS Workforce Statistics*. Retrieved March 6, 2024, from https://digital.nhs.uk/data-and-information/publications/statistical/nhs-workforce-statistics/september-2023

Wylie Communications. (2024). *What's the latest U.S. literacy rate?* Retrieved March 6, 2024, from https://www.wyliecomm.com/2021/08/whats-the-latest-u-s-literacy-rate/

Yang, Xiaomei & Hu, Miao & Chen, Xin & Teng, Shuai & Chen, Gongfa & Bassir, David. (2023). Predicting Models for Local Sedimentary Basin Effect Using a Convolutional Neural Network. Applied Sciences. 13. 9128. 10.3390/app13169128.

Yim, W.-W., Yetisgen, M., Harris, W. P., & Kwan, S. W. (2016). Natural Language Processing in Oncology: A Review. *JAMA Oncology*, *2*(6), 797–804. https://doi.org/10.1001/jamaoncol.2016.0213

Young, I. J. B., Luz, S., & Lone, N. (2019). A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *International Journal of Medical Informatics*, *132*, 103971. https://doi.org/10.1016/j.ijmedinf.2019.103971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008).

Zeng, Z., Deng, Y., Li, X., Naumann, T., & Luo, Y. (2019). Natural Language Processing for EHR-Based Computational Phenotyping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(1), 139–153. https://doi.org/10.1109/TCBB.2018.2849968

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., & Amodei, D. (2019). Fine-Tuning Language Models from Human Preferences. arXiv preprint arXiv:1909.08593.