UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Longitudinal Health Transformer for Cancer Pathways Modelling

by Leah Gerrard

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of A/Prof. Guodong Long, Dr. Xueping Peng, and Dist. Prof. Chengqi Zhang

University of Technology Sydney
Faculty of Engineering and Information Technology

August 2025

Certificate of Original Authorship

I, Leah Gerrard, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by an Australian Government Research Training Program (RTP) Scholarship https://doi.org/10.82133/C42F-K220

Signature: Production Note:
Signature removed prior to publication.

Date: 11/02/2026

Abstract

The success of artificial intelligence in modelling complex data, from natural language to visual inputs, has inspired its application to healthcare. In cancer care, the cancer pathway consists of a sequence of longitudinal health data that combines development of the disease with interactions between patients and various healthcare providers. Modelling both the disease progression and human interactions within cancer pathways remains a significant challenge. As a result, there is a pressing need for approaches that can better capture and understand complex cancer data and pathways for unified solutions. This includes considering the unique characteristics of cancer pathways being multi-outcome, captured within various data sources, and often suffering from limited data and labels. To address these issues, this research focuses on developing deep learning-based methods for cancer pathways modelling, utilising transformer-based models with longitudinal health data. A series of models are proposed that leverage strategies including multi-task learning; longitudinal patient modelling; and transfer learning; that are tailored to the cancer context. Experimental testing of these models demonstrates their ability to improve predictions for cancer patients and provide more effective, flexible, and data-efficient approaches. This work illustrates the value of the Transformer in capturing intricate relationships between cancer patients and the healthcare system, offering a promising foundation for advancing the modelling of cancer pathways and improving the care and outcomes for cancer patients.

Keywords: Cancer, Cancer Pathways Modelling, Transformer, Deep Learning, Healthcare, Longitudinal Health Data.

Dissertation directed by A/Prof. Guodong Long (Principal Supervisor), Dr. Xueping Peng, and Dist. Prof. Chengqi Zhang. The Australian Artificial Intelligence Institute, Faculty of Engineering and IT, University of Technology Sydney.

Acknowledgements

First, I would like to express my sincere gratitude to my research supervisors. To my principal supervisor A/Prof. Guodong Long, thank you for your expert guidance and invaluable feedback, which was essential in shaping and improving this research. Your encouragement to pursue my own research interests has also been greatly appreciated along this research journey. To my co-supervisor Dr. Xueping Peng, thank you for your advice and for sharing your technical and methodological knowledge. You helped me acquire competence in the development of deep learning models. I also thank you for presenting our research at the ADMA conference. To my co-supervisor Distinguished Prof. Chengqi Zhang, thank you for your support and encouragement for cross-discipline partnerships that helped make this research a success. To my industry supervisor, Dr. Allison Clarke, thank you for your continued support in this research, for sharing your healthcare expertise and insights, and for facilitating the collaborations in this research. This helped to align the work with real-world challenges and enhance its industry relevance.

A special thanks goes to my co-authors and colleagues – particularly Dr. Alvin Wang. Thank you for sharing your knowledge and perspectives, and for making the day-to-day a memorable experience. I also thank all the project stakeholders involved in this research who have shared their expertise to enhance this work. I express my appreciation to the Australian Artificial Intelligence Institute and the Department of Health, Disability and Ageing, for supporting this research. I particularly want to thank the Health Economics and Research Division for enabling access to real-world health and cancer data to enrich the significance of this research. I also acknowledge the Victorian Department of Health, the Australian Institute of Health and Welfare, and the Australian Bureau of Statistics, for their respective data, ethics, and infrastructure contributions.

Last, but certainly not least, I express my heartfelt thanks to my family and friends. To my soon-to-be husband, I am deeply grateful for your encouragement, understanding, and

kindness. You have been a constant support and comfort not only during this research, but in every aspect of life during this time. I could not have done this without you. To my parents, thank you for your endless love and reassurance. Without you, I never would have pursued doctoral research, and your advice and emotional support helped me persevere through the challenging moments. To my sister, thank you for always making me laugh and for motivating me when I needed it most. This process was far more enjoyable because of you, and I thank you for always being around as a sounding board and a reviewer. And to my brother, thank you for always cheering me on and providing much needed moral support. Finally, thanks to all my wonderful friends for your encouragement and for bringing joy and balance to this research.

Leah Gerrard

Canberra, Australia, 2025

List of Publications

Conference Papers

1. **Gerrard, L.**, Peng, X., Clarke, A., Schlegel, C., Jiang, J.: Predicting Outcomes for Cancer Patients with Transformer-Based Multi-task Learning. In: Australasian Joint Conference on Artificial Intelligence. pp. 381-392. Springer (2022)

2. **Gerrard, L**., Peng, X., Clarke, A., Long, G.: Multi-level Transformer for Cancer Outcome Prediction in Large-Scale Claims Data. In: International Conference on Advanced Data Mining and Applications. pp. 63-78. Springer (2023)

3. **Gerrard, L**., Peng, X., Clarke, A., Long, G.: Claimsformer: Pretrained Transformer for Administrative Claims Data to Predict Chronic Conditions. In: Australasian Joint Conference on Artificial Intelligence. pp. 348-362. Springer (2024)

Other Paper Contributions

1. Wang, Y., Long, G., Peng, X., Clarke, A., Stevenson, R., **Gerrard, L**.: Interactive Deep Metric Learning for Healthcare Cohort Discovery. In: Australasian Conference on Data Mining. pp. 208-221. Springer (2019)

2. Long, G., Shen, T., Tan, Y., **Gerrard, L.,** Clarke, A., and Jiang, J.: Federated Learning for Privacy-Preserving Open Innovation Future on Digital Health, In: Humanity Driven-AI: productivity, well-being, sustainability and partnership. pp. 113-133. Springer (2021)

3. Xie, M., Jiang, J., Shen, T., Wang, Y., **Gerrard, L.,** and Clarke, A.: A Green Pipeline for Out-of-Domain Public Sentiment Analysis. In: International Conference on Advanced Data Mining and Applications. pp. 190-202. Springer(2022)

Contents

Certificate i

Abstract ii

Acknowledgments iii

List of Publications v

Acronyms xii

List of Figures xv

List of Tables xvi

1 Introduction 1

1.1 Background and Motivation 1

1.2 Research Problems 3

1.3 Research Aim and Objectives 6

1.4 Research Significance 7

1.5 Research Contributions 9

1.6 Thesis Outline and Structure 11

2 Literature Review 14

2.1 Context and Overview 14

2.2 Transformers for Longitudinal Health Data 16

2.2.1 Overview of Transformers and Longitudinal Health Data 16

2.2.2 Benefits of Transformers for Longitudinal Health Data 21

2.2.3 Early Transformer Models for Longitudinal Health Data 23

2.2.4 Summary 25
2.3 Multi-task Learning with Longitudinal Health Data 26
2.3.1 Brief Description of Multi-task Learning 26
2.3.2 Multi-Task Learning with Longitudinal Health Data 27
2.3.3 Application to Cancer Prediction Problems 29
2.3.4 Summary 30
2.4 Transformer-based Patient Longitudinal Modelling 30
2.4.1 Integrating Additional Patient Features 30
2.4.2 Temporal and Hierarchical Modelling in EHR 32
2.4.3 Temporal and Hierarchical Modelling in Claims Data 35
2.4.4 Applications to Cancer Predictions 37
2.4.5 Summary 39
2.5 Transfer Learning with Longitudinal Health Data 40
2.5.1 Transfer Learning and Pretrained Models 40
2.5.2 Pretrained Models for EHR 41
2.5.3 Pretrained Models for Claims Data 44
2.5.4 Pretrained Models with Cancer-related Fine-tuning 46
2.5.5 Summary 48
2.6 Conclusion 48

3 Multi-Outcome Prediction for Cancer Patients with Transformer-Based Multi-Task Learning 50

3.1 Introduction 51
3.2 Related Work 53
3.3 Methodology 54

3.3.1 Notations . . . 54
3.3.2 Model Overview . . . 54
3.3.3 Common Representations . . . 55
3.3.4 Transformer Model . . . 57
3.3.5 Prediction via Multi-task Learning . . . 58
3.4 Experiments . . . 59
3.4.1 Datasets . . . 59
3.4.2 Data Preprocessing . . . 60
3.4.3 Experimental Setup . . . 60
3.4.4 Results . . . 62
3.5 Conclusion . . . 63

4 Modelling Longitudinal Cancer Pathways in Claims Data with Hierarchical Transformer 65

4.1 Introduction . . . 66
4.2 Related Work . . . 68
4.3 Methodology . . . 70
4.3.1 Model Overview . . . 70
4.3.2 Claims Item Embedding . . . 70
4.3.3 Feature Encoder Block . . . 72
4.3.4 Patient Representation . . . 73
4.3.5 Cancer Outcome Prediction . . . 74
4.4 Experiments . . . 75
4.4.1 Dataset . . . 75
4.4.2 Data Preparation . . . 76

4.4.3 Cohort Definitions 77
4.4.4 Experimental Setup 78
4.4.5 Results 80
4.5 Conclusion 83

5 Pretrained Transformer for Claims Data to Predict Chronic Conditions for Cancer Patients with Limited Data 84

5.1 Introduction 85
5.2 Related Work 87
5.3 Methodology 87
5.3.1 Model Overview 87
5.3.2 Input Embedding 88
5.3.3 Transformer Block 90
5.3.4 Model Pretraining 90
5.3.5 Fine-tuning on Downstream Condition Tasks 92
5.4 Experiments 94
5.4.1 Dataset 94
5.4.2 Data Preprocessing 94
5.4.3 Experimental Setup 97
5.4.4 Results 98
5.5 Conclusion 102

6 Discussion and Conclusion 103

6.1 Summary of Contributions 103
6.2 Research Implications 104
6.3 Limitations 114

6.4 Future Directions 116
6.5 Further Challenges and Opportunities 123
6.6 Concluding Remarks 132

A Supplementary Material **133**

Bibliography **137**

Acronyms

ACP	Australian Cancer Plan
AI	Artificial Intelligence
ATC	Anatomical Therapeutic Chemical
AUPRC	Area Under the Precision-Recall Curve
AUC	Area Under the Receiver Operating Characteristic Curve
BP	back pain
BEHRT	BERT for EHR
BERT	Bidirectional Encoder Representations from Transformers
BRLTM	Bidirectional Representation Learning model with a Transformer on Multimodal data
BYOL	Bootstrap Your Own Latent
CPUs	Central Processing Units
CLS	classification
Claim-PT	Claim Pre-Training
CF	Claimsformer
CCS	Clinical Classifications Software
CFIR	Consolidated Framework for Implementation Research
COPD	Chronic Obstructive Pulmonary Disease
CORE-BEHRT	Carefully Optimised and Rigourously Evaluated BEHRT
DT	Decision Tree
EHR	Electronic Health Record
ExMed-BERT	Extended Med-BERT
FEMRs	Foundation models for Electronic Medical Records

GRU	Gated Recurrent Unit
GANs	Generative Adversarial Networks
GenAI	Generative AI
GPT	Generative Pretrained Transformer
GB	Gradient Boosting Classifier
GPU	Graphics Processing Unit
HSVD	heart, stroke and vascular disease
HC	high cholesterol
HT	hypertension
ICU	Intensive Care Unit
ICD	International Classification of Diseases
IRSD	Index of Relative Socioeconomic Disadvantage
LLMs	Large Language Models
LMP	Last Month Prediction
LR	Logistic Regression
LSTM	Long Short-Term Memory
MedHMP	Hierarchical Multimodal Pretraining framework for the Medical domain
MCM	Masked Claims Modelling
MLM	Masked Language Modelling
MBS	Medicare Benefits Schedule
MIMIC	Medical Information Mart for Intensive Care
MH	mental health
MOTOR	Many Outcome Time Oriented Representations
MM	multimorbidity
MTL	Multi-Task Learning
NHS	National Health Survey
NLP	Natural Language Processing

NMP	Next Month Prediction
OCPs	Optimal Care Pathways
PBS	Pharmaceutical Benefits Scheme
PLIDA	Person Level Integrated Data Asset
PLOS	Prolonged Length of Stay
RF	Random Forest
ROC	Receiver Operating Characteristic
RNN	Recurrent Neural Network
RPs	Research Problems
SARD	Self-Attention with Reverse Distillation
SEP	separation
SEIFA	Socio-Economic Indexes for Areas
SHAP	SHapely Additive exPlanations
STL	Single-Task Learning
SURCs	Symptom and Urgent Review Clinics
TMAE	Transformer-based Multimodal Autoencoder
XAI	eXplainable AI

List of Figures

1.1 Relationship between research aim, objectives and contributions. 11

1.2 Thesis structure and chapter relations. 13

2.1 The Transformer architecture, copied from Vaswani et al. [232]. 17

3.1 The proposed TransMT model. 56

3.2 Area Under the Receiver Operating Characteristic Curve (AUC) of hospital readmission on two datasets. 63

4.1 The proposed Claims-MLT model. 71

4.2 Impact of static patient features on model performance 82

5.1 The proposed Claimsformer model. 88

5.2 Input from a hypothetical person showing how the Claimsformer model sees the claims data. 89

5.3 Convergence analysis comparing performance of Claimsformer and naive Transformer. 101

A.1 Performance during training for the MCM Pretraining Objective. 134

List of Tables

3.1 Notations for TransMT. 55

3.2 Statistics of the MIMIC datasets for the two cancer cohorts. 60

3.3 Performance comparison of prediction tasks. 62

3.4 Relationship between future diagnosis and readmission. 63

4.1 Patient static features. 77

4.2 Statistics of the cancer data 78

4.3 Model performance for survival prediction. 81

4.4 Model performance for heart disease prediction. 81

5.1 Summary of the pretraining strategies implemented for the Claimsformer model. 92

5.2 Statistics of the fine-tuning cohorts for multimorbidity prediction. 96

5.3 Statistics of the fine-tuning cohorts for single condition prediction. 96

5.4 Hyperparameters for pretraining and fine-tuning. 98

5.5 Comparison of pretraining strategies on multimorbidity prediction. 99

5.6 Average model performance metrics on single condition prediction for the cancer cohort. 100

5.7 Average model performance metrics on single condition prediction for the mental health cohort. 101

5.8 Model performance metrics for diabetes prediction on various fine-tuning data sizes. 102

A.1 Results of the objectives for each pretraining strategy, presented as average precision or accuracy @ top k. 133

A.2 Precision, recall and accuracy metrics for the multimorbidity task. 134

A.3 Standard deviations of F1 and AUC metrics on single condition predictions for the cancer cohort. 135

A.4 Standard deviations of F1 and AUC metrics on single condition predictions for the mental health cohort. 135

A.5 Precision, recall and accuracy metrics on single condition predictions for the cancer cohort. 136

A.6 Precision, recall and accuracy metrics on single condition predictions for the mental health cohort. 136

Chapter 1

Introduction

This chapter provides the background and motivation for this research by describing the current challenges and problems of cancer pathways, and the modelling of longitudinal health data with transformers. It also presents the research aim and objectives, significance, contributions, and provides an outline of the thesis structure.

1.1 Background and Motivation

Cancer is a cellular disease characterised by abnormal cells that grow within the body uncontrollably [205, 231, 171]. Because of this uncontrollable growth, cancer can spread from its original site to other parts of the body, in a process known as 'metastasis' [181, 182]. The continuous and evolving nature of cancer, with its ability to evade the immune response and resist therapies, makes it challenging to develop broadly applicable treatment approaches [182, 271]. This results in numerous treatment options and ongoing development of new therapies [64, 61], driving complex and heterogenous cancer pathways.

Cancer pathways consist of the events related to the progression, treatment and monitoring of cancer, as well as co-morbid (co-occurring) conditions, resulting in intricate interactions between patients and healthcare providers. In addition to the unique characteristics of individual cancer tumours, there is wide diversity in cancer stages, subtypes, and patient populations [264, 152, 3]. Modelling both the disease and human interactions within a single system remains a significant challenge for oncology (the study and treatment of cancer), as the patient 'state' is continually evolving and there are various possible pathways and outcomes that can occur under differing conditions. Despite the presence of existing risk prediction models, clinical guidelines, and Optimal Care Pathways (OCPs), it remains difficult to predict outcomes for cancer patients. Innovations are needed that

can better understand and model cancer pathways to enable more effective and universal approaches for oncology.

The digitisation of healthcare information has provided great opportunity for developing improved risk prediction models. The growth of longitudinal health data, which typically can be found in Electronic Health Record (EHR) and administrative claims (claims) data, has enabled the capture of long-term and repeated patient activity over time [214]. Longitudinal health data is useful in modelling cancer pathways and making cancer-related predictions as it reveals information about cancer treatment (e.g., surgery, chemotherapy regimen), as well as patient outcomes (e.g., hospitalisation, death). However, despite longitudinal health data showing benefits for predicting cancer outcomes [192], the temporal patterns within this data are under-exploited in oncology [147]. This is partly because it is challenging to analyse longitudinal health data due to its high-dimensional, heterogeneous, temporal, and irregular data characteristics [269, 154, 76, 210, 10, 147].

Artificial Intelligence (AI) and deep learning methods have shown great promise in analysing complex data, such as image and natural language. AI, which can be defined as the development of computer systems that simulate human intelligence, has been increasingly applied in healthcare [17, 161]. AI consists of several sub-fields, including machine learning, which refers to the study of algorithms that automatically learn through experience, and deep learning, which is classified by the use of large, layered neural networks [17]. The main advantage of deep learning compared to traditional machine learning is the development of meaningful data representations that can capture complex, non-linear relationships in data [22]. This is advantageous for longitudinal health data, with early deep learning models based on the Recurrent Neural Network (RNN) outperforming traditional machine learning approaches [128, 16, 251].

More recently, the Transformer model has evolved to be the next-generation neural architecture for deep learning, establishing its effectiveness in text and computer vision applications [52, 101]. Transformers have also been introduced into the healthcare and medical domain, demonstrating potential to successfully capture complex relations within large-scale health data via pretrained models BEHRT [119] and Med-BERT [188], and

improve patient-related prediction tasks [118, 142, 66, 262]. This emphasises the suitability of transformers to extract patterns from longitudinal health data and the need to adopt these models in the healthcare sector, particularly with the continual growth of health data, computational resources, and advanced technologies such as Generative AI (GenAI).

However, the application of transformers to cancer in longitudinal health data is relatively unexplored [243]. Moreover, the complexity of cancer pathways pose additional challenges to transformers, including being multi-outcome, captured within various data sources (e.g., EHR and claims data), and often suffering from data and label insufficiencies. Developing effective, flexible, and data-efficient approaches to overcome these shortcomings is crucial for improving the modelling of cancer pathways and progressing towards generative capability. Recent exploration of transformers with Multi-Task Learning (MTL), longitudinal patient modelling, and transfer learning, have provided opportunities to better model patient care and predict health outcomes. However, there is little investigation of these approaches with transformers in longitudinal health data for cancer, and so transformers need to be adapted to the specific cancer data and context for tailored solutions.

This research is therefore motivated by the opportunity to enhance cancer pathways modelling in longitudinal health data with transformers, and demonstrate the adaptability and effectiveness of these models for cancer. Developing more effective transformer models will not only provide opportunity to enhance cancer care and outcomes by supporting clinical and healthcare policy decision making, but support the development and transition towards more comprehensive and generative approaches that are pivotal in realising the benefits of AI in oncology.

1.2 Research Problems

This research focuses on the challenges of modelling complex cancer pathways in longitudinal health data with transformers. Specifically, it is driven by three primary Research Problems (RPs) as described below.

RP1: How to extend transformers to model multiple related cancer outcomes in longitudinal health data?

Cancer is a multi-outcome or multi-objective disease [6] due to multiple care phases (as defined in the OCPs[1]) and numerous cancer treatment options. While cancer prognosis and survival are key outcomes, cancer patients experience many healthcare events and outcomes along the cancer pathway that are of clinical interest and important for patient care [62, 92, 69, 75, 83, 172, 238, 84]. Some of these include risk of hospital readmission; future disease; complications, adverse events or side effects; and cancer recurrence (when cancer returns after treatment). Research has shown that there can be associations or correlations between these outcomes (e.g., common diseases and hospitalisation, cancer stage and recurrence) [27, 159, 53, 46, 93, 230]. However, many existing cancer-related prediction models are built for single predictions (e.g., survival *or* readmission) [3, 58, 69, 83, 89, 220, 77], and so do not consider the potential relationships between outcomes. This contrasts with real-world cancer care where clinicians must consider the risk of many outcomes for a patient simultaneously.

Existing evidence in deep learning has demonstrated that MTL, which involves the joint training of multiple tasks, can provide better predictions than single task approaches and help address data heterogeneity [79, 121, 198]. However, implementation of MTL with transformers is limited with longitudinal health data in the cancer context, and studies on EHR data have shown inconsistent results [139, 172]. This is despite wide use of MTL for cancer studies that involve image data [141, 155, 32]. Further work is needed with longitudinal health data to determine the potential benefit of MTL with transformers and for cancer prediction problems. Models that can simultaneously predict outcomes would also more closely reflect real cancer care. This research explores the simultaneous prediction of future disease and hospital readmission for cancer patients using transformers with EHR data.

[1] https://www.cancervic.org.au/get-support/for-health-professionals/optimal-care-pathways

RP2: How to effectively model longitudinal relations in claims data with transformers for predicting cancer outcomes?

Cancer care occurs across diverse settings and is captured within various healthcare data [44]. For example, in Australia, healthcare services and pharmaceutical use is captured within claims data, while cancer diagnosis and clinical features are contained within cancer registries. Approaches for modelling cancer pathways therefore need to be flexible and able to leverage these differing datasets. In addition, cancer care is inherently long-term, temporal, and irregular. This is due to multiple phases of care and treatment pathways, as well as variability in timing of healthcare interactions between patients [271, 44, 147]. Models thus also need to be able to capture these long-range temporal dependencies within cancer pathways in claims data.

Existing application of transformers to longitudinal health data has demonstrated promise in developing effective longitudinal patient representations via inclusion of additional patient features, and use of temporal and hierarchical modelling [66, 118, 83, 142, 211]. However, there are a number of current limitations, including lack of consensus on how best to encode and represent longitudinal health data for patient predictions; a focus primarily on EHR data, leaving claims data understudied; and limited cancer applications. In particular, there are no studies that have looked at Australian claims data, which contains extensive information related to cancer care but has a differing temporal structure to EHR. Further work is needed for modelling cancer pathways in claims data with transformers. This research studies the problem of modelling longitudinal claims data for cancer patients initiating chemotherapy treatment, using a hierarchical transformer to model both dynamic and static (demographic, cancer clinical features) patient information.

RP3: How to improve model performance in cancer prediction settings with limited data or labels?

Cancer studies often suffer from limited data or labels. It is costly and laborious to collect medical data [257]. It is also typically challenging to access healthcare datasets due to privacy and access concerns [212], leading to difficulties in obtaining large and diverse data for cancer prediction problems [181, 182, 269, 271, 202]. Moreover, many

cancer studies are derived from sub-populations, due to a focus on specific cancer types, or use of single-institution data [23], also resulting in small data size. In addition to limited data, labels can be limited as they often require specialised clinical (and oncological) expertise [245, 44, 271]. Furthermore, some elements that are related to, or impact cancer care are not comprehensive or routinely captured in longitudinal health data unless linked with other data sources (e.g., chronic conditions, cancer stage, cancer recurrence) [99]. Models for cancer must therefore be data-efficient and generalisable to operate effectively in limited data or label scenarios.

Transfer learning has been identified as an effective technique to improve predictions where datasets are of limited size or labels [52]. By developing pretrained transformer models, generalised relationships in large unlabelled healthcare data can be leveraged to help fine-tune predictions in scenarios where data is small but labels are available. Pretraining has demonstrated improvements in the prediction of patient outcomes [119, 118]. However, there is wide variability in pretraining approaches and evaluations, and limited exploration of claims data [142, 188]. There is opportunity to improve predictions for cancer patients and provide more generalisable models through further exploration of pretraining strategies and transfer learning. This research explores the challenge of predicting chronic conditions for cancer patients in low resource scenarios by developing pretrained transformers for claims data.

1.3 Research Aim and Objectives

Based on the aforementioned research problems, the aim of this work is to effectively model complex cancer pathways in longitudinal health data with transformers. This relies on developing more effective, flexible, and data-efficient transformer-based approaches, leveraging various data and machine learning training techniques. There are three main objectives, with each targeted at addressing a different research problem:

- **Objective 1** (to RP1): To integrate MTL with transformers and evaluate this methodology on EHR data for multi-outcome cancer predictions. The Transformer will leverage temporal relationships within cancer pathways and advance existing single-task approaches by enabling joint modelling of multiple outcomes for cancer pa-

tients.

- **Objective 2** (to RP2): To extend transformers to claims data, develop a model tailored to the characteristics of the data, and evaluate the impact of additional patient demographics and cancer features on predictions. This will provide a customised approach for modelling longitudinal claims data to address its differing temporal structure (to EHR). Extending transformers to claims data will enhance their flexibility for cancer pathways modelling by expanding their use to additional data sources, as well as through use of both dynamic and static patient information.

- **Objective 3** (to RP3): To develop a generalised pretrained transformer model that leverages relations between claims items in large-scale patient histories, and evaluate the model on various pretraining strategies, downstream tasks, fine-tuning data sizes, and model convergence. This will assist in providing more data-efficient and generalised models and help address the data and label insufficiency problem when predicting chronic conditions for cancer patients.

1.4 Research Significance

The development of more effective and comprehensive approaches for cancer pathways modelling and risk prediction is essential to face modern day cancer, data, and technological challenges. Cancer continues to be a global health priority with wide impacts on individuals and healthcare systems. Despite increased survival and availability of contemporary cancer treatments in Australia, cancer still accounts for an estimated one third of deaths (in 2024), and between 2000 and 2024, cancer cases in the country have increased about 93% [15]. While this is generally attributed to the growing and ageing population, this cannot fully explain the observed increase, and more recent evidence has suggested cancer incidence rates are on the rise for younger populations [15]. Overall, these trends are resulting in more individuals being diagnosed with, treated for, and dying of cancer, or living with cancer for longer periods of time. This represents a crucial need for more innovative solutions for cancer. This research explores the opportunities for such strategies by adapting the Transformer to model cancer pathways in longitudinal health data.

In addition to the increasing cancer burden, the size and complex nature of cancer data presents issues for traditional modelling approaches [154, 268]. These models often fail to consider temporal nuances in data and reliance on manual feature engineering can limit scalability, robustness, and performance [47, 268]. In order to enhance cancer pathways modelling at scale, it is necessary to adopt methods that can handle and learn from this complex information. The Transformer is well-placed for this task, however, being originally developed for natural language, it needs to be adapted to the cancer data and context. This research addresses this need by developing transformers that are tailored to the specific characteristics of longitudinal health data, overcoming limitations of traditional statistical and machine learning methods, and existing transformers.

As the volume and complexity of healthcare data has grown, this has also driven demand for more sophisticated methods that can effectively analyse this information to provide actionable insights [241]. The patterns within complex cancer pathways are not easily interpretable by clinicians to support care decisions, nor by government to inform the evidence-base [7]. Methods and tools are required that can extract meaningful trends from this data and provide it to decision makers for consideration. This is particularly important for the Australian Government, who holds large amounts of Australian claims data and have invested into linked data assets and enduring linkages. In this research, methods to support effective analysis of large longitudinal health data are investigated and the benefit of these approaches in predicting patient outcomes are explored. By utilising real-world longitudinal health data and focusing on key cancer outcomes related to survival, hospitalisation, and disease risk, this facilitates the use of these models in practice to support evidence-based, data-driven decision making, enhancing opportunity to benefit real-life cancer care and outcomes.

Finally, for cancer and the broader healthcare system, there is ongoing pressure to increase health services and improve care within constrained budgets [7]. AI has potential to help optimise the healthcare system and deliver more efficient healthcare [134, 193]. Research indicates that foundation and generative models can extract generalised patterns, correlations, and anomalies from extensive amounts of healthcare information [193, 178, 133]. These models can leverage this knowledge via efficient re-use of existing data to

benefit patient predictions, that can in turn help to optimise patient treatment, care, and outcomes [193, 243, 153]. This is important as oncology is increasingly moving towards value-based care [8, 144]. This research supports these challenges by developing generalisable and data-efficient transformers that can leverage relations in large-scale claims data via pretrained models. This is significant to informing the development of foundation models and GenAI for the Australian healthcare system, where these solutions can support the progression from discriminative models towards generative systems to better understand and interrogate cancer pathways in longitudinal health data to improve cancer care.

Ultimately, the core significance of this research lies in its interdisciplinary nature, with the unique combination of transformers, longitudinal health data (particularly claims data), and cancer. This research therefore bridges the gap between advanced deep learning architectures and real-world cancer data and prediction problems, with potential to influence AI-based cancer models, clinician and government practice, and public health initiatives, to ultimately improve cancer care. As cancer continues to have a significant individual and health system burden, efficient and scalable solutions that can move beyond pure predictability to generalised data understanding can offer opportunity to unlock the full potential of AI for oncology.

1.5 Research Contributions

This work proposes several transformer-based approaches for enhancing cancer pathways modelling in longitudinal health data. The contributions are summarised below and their connection to the primary purpose and objectives of this research are illustrated in Fig 1.1.

- **Contribution 1:** We initially explore MTL with transformers and propose the TransMT model to jointly predict two related outcomes for hospitalised patients with cancer: future diagnosis and 30-day hospital readmission. We leverage the transformer encoder to capture the inherent diagnosis and sequential visit dependencies, and apply MTL to simultaneously learn common low-level representations and task specific knowledge. We implement experiments on publicly available EHR

data, and compare performance of TransMT against two common RNN baselines and a single-task transformer. Results show the superiority of the TransMT model, demonstrating for the first time the potential of a multi-task transformer to enhance prediction of related cancer outcomes in EHR data.

- **Contribution 2:** We then adapt the Transformer to claims data, and propose the Claims-MLT model. The approach uses a hierarchical (multi-level) transformer encoder to learn effective longitudinal patient representations by considering the low-level claims item relationships (within a month) and sequential patterns in patient claims histories (between months). We evaluate the approach on a real-world cancer dataset from Australia containing individuals diagnosed with breast and/or colorectal cancer, and make the following predictions for patients initiating chemotherapy treatment: one-year survival and heart disease diagnosis. We compare the approach to a single-level transformer, as well as shallow machine learning methods, and show Claims-MLT outperforms baselines for both predictive tasks and cancer types. This is the first approach to combine transformer-based dynamic hierarchical claims modelling with static patient demographic and clinical features to predict cancer outcomes.

- **Contribution 3:** We further leverage claims item relations in the Claimsformer, a pretrained transformer model tailored to claims data. This two-stage transfer learning approach firstly captures generalised patterns from Australian medical services and prescriptions, and then applies this knowledge to predict chronic conditions for cancer and mental health cohorts. We provide a comprehensive investigation into pretraining on claims data by exploring various pretraining strategies to identify the optimal approach for the Claimsformer. We then compare the Claimsformer to a naive transformer (no pretraining) and shallow machine learning methods, and show it can offer improved chronic condition predictions, earlier model convergence, and effectiveness in low-resource settings. This is the first exploration of pretrained transformers on Australian claims data, validating the feasibility and usefulness of pretraining with this data source.

Together, this research establishes the foundation for a longitudinal health transformer

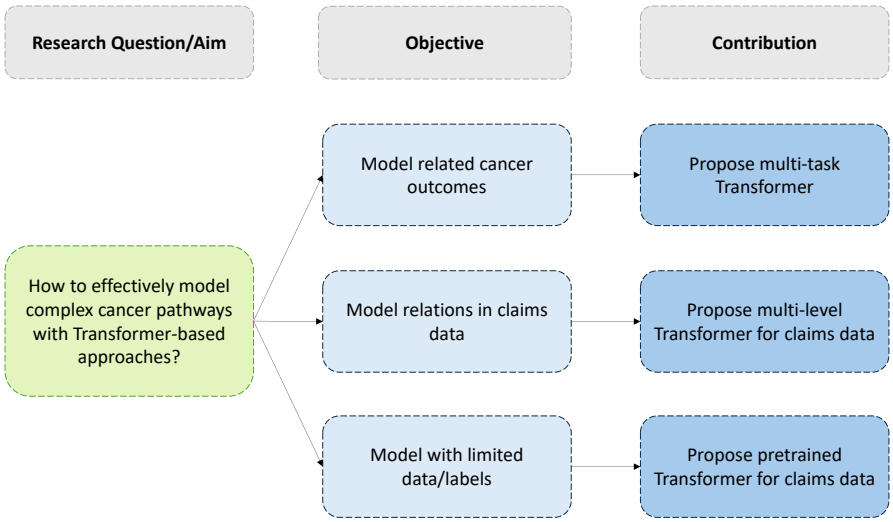

Figure 1.1 : Relationship between research aim, objectives and contributions.

for cancer pathways modelling. Novel transformer-based approaches are developed that are specifically designed for longitudinal health data in the cancer context that are more effective, flexible, and data-efficient. While each of the contributions relates to a distinct cancer objective, they collectively address the core challenge of this research to effectively model complex cancer pathways with transformers (see Fig 1.1). This work informs a comprehensive cancer pathways modelling framework, contributing to the development of more scalable and unified approaches, and edging us closer to more advanced AI tools and generative technologies for better cancer care and outcomes.

1.6 Thesis Outline and Structure

The remaining chapters of this thesis are summarised as follows:

- **Chapter 2:** presents a review of the literature related to this research. It focuses on studies related to MTL, longitudinal patient modelling, and transfer learning with transformers and longitudinal health data, with inclusion of available cancer studies.
- **Chapter 3:** presents the proposed multi-task transformer to enhance multi-outcome prediction for cancer patients with EHR data.

- **Chapter 4:** details the implementation of the transformer to claims data, to develop more effective and tailored longitudinal patient representations for modelling cancer pathways and predicting cancer-related outcomes.

- **Chapter 5:** considers the data and label insufficiency problem and presents a pretrained model for claims data that leverages generalised claims relations to improve chronic condition predictions for those with cancer.

- **Chapter 6:** summarises the contributions of this research, discusses the implications, outlines the research limitations, and presents avenues for future studies. It also provides a brief overview of challenges and opportunities when considering the real-world implementation of deep learning models in healthcare, before providing the concluding remarks.

A visual representation of the thesis structure, including relationships between chapters and longitudinal health data used, is depicted in Fig 1.2. As shown in the figure, Chapter 3 focuses on application to EHR data, with Chapters 4 and 5 utilising claims data. While Chapters 3 - 5 each focus on distinct cancer pathways challenges, and make differing cancer predictions, these components are linked and progressively build upon one another. Implementation of the transformer on EHR data in Chapter 3 is leveraged in Chapter 4, where the model is adapted to the specific characteristics of claims data. The modelling of claims item relations in Chapter 4 is then further developed in Chapter 5, where the transformer is extended to enable pretraining to provide a generalised model. Collectively, this provides integrated strategies to enhance existing cancer pathways modelling with transformers in longitudinal health data.

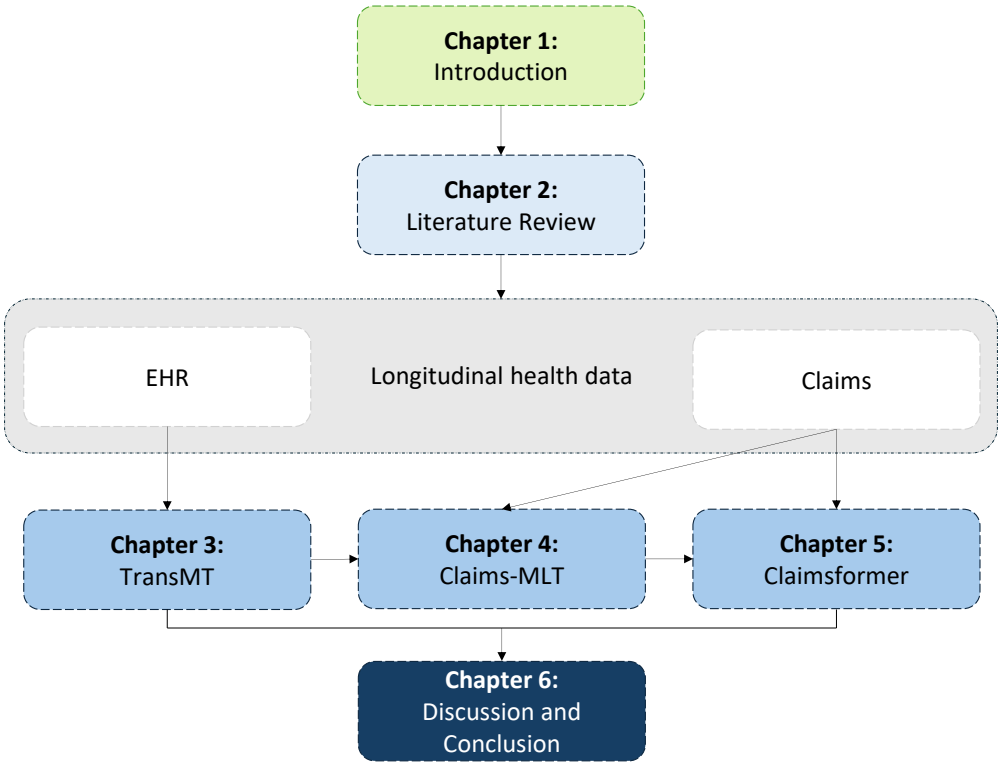

Figure 1.2 : Thesis structure and chapter relations.

Chapter 2

Literature Review

This chapter provides a review of the existing literature, focusing on the core methodological themes of MTL, longitudinal patient modelling, and transfer learning. The use of these methods with deep learning (primarily the Transformer) is explored on longitudinal health data, including applications to cancer. The chapter discusses developments in the field, patterns that have emerged, and existing gaps or inconsistencies that have informed the approaches used in this research.

2.1 Context and Overview

The last few years have seen much growth in the application of AI to the cancer field. A simple Google Scholar search of 'AI for cancer' reveals over a million results since 2020[1], highlighting the focus on applying AI to oncology in recent years. While there are numerous publications dedicated to the discussion of AI for cancer, much research to date has focused on cancer diagnosis and prognosis through use of image and omics data [207, 171, 229, 67, 122, 231, 257, 269]. As a consequence, this has left the analysis of longitudinal health data for cancer less mature, with a recent review suggesting that further research is needed on AI approaches that fully exploit the information in longitudinal health data for cancer [147].

Much of the research investigating transformers on longitudinal health data has explored health conditions other than cancer (e.g., heart disease and kidney disease are commonly used) or used more generic healthcare tasks (e.g., future disease, mortality, readmission). Additionally, as earlier noted, much of the existing literature has focused on EHR data, with claims data receiving less attention. A review exploring Foundation

[1] As at 18 August 2025.

models for Electronic Medical Records (FEMRs)[2] found that of the 34 FEMR models included in the review, only four used claims data, and only two had cancer-related predictions [243]. While this paper does not capture all the models or studies related to this research, it demonstrates a general trend towards use of EHR data, and less exploration of claims data and cancer prediction problems. Therefore, much of the literature related to this research is the use of transformers with EHR data.

The challenging characteristics of longitudinal health data [31] has inspired numerous deep learning studies to try and better model this healthcare information for predicting patient outcomes. Three areas which have demonstrated benefits to health-related predictions include: 1) the use of MTL to simultaneously predict multiple outcomes; 2) the integration of additional patient features and use of time-aware or hierarchical approaches to better model longitudinal patient pathways; and 3) the use of transfer learning and pretrained models to share generalised knowledge to enhance downstream prediction tasks. This review focuses on these methodological approaches as opportunities to improve cancer pathways modelling and the prediction of cancer-related outcomes. While many studies related to applications of transformers on EHR data, applications to claims data and cancer are included where available.

In Section 2.2, transformers and longitudinal health data are introduced. The benefits of transformers for this data, as well as early transformer models, are described. In Section 2.3, the literature related to MTL for longitudinal health data is summarised. This focuses on use of the transformer but includes mention of other deep learning methods where appropriate. In Section 2.4, the use of transformers for longitudinal patient modelling is discussed. Common trends are summarised including enhancing transformers using additional patient features (such as demographics), and temporal and hierarchical modelling in EHR and claims data, and for cancer. In Section 2.5, an overview of the related work on transfer learning and pretrained models using transformers is provided. Existing models for EHR data, claims data, and for cancer-related fine-tuning are described. Finally, the chapter concludes with a review of the key research gaps and an indication of how this research addresses the noted limitations.

[2] EMRs are synonymous to EHRs.

2.2 Transformers for Longitudinal Health Data

This section aims to provide an overview of the Transformer and longitudinal health data, and discuss the benefits of the transformer for this data. It also provides a summary of the early transformer models for EHR data, as these served as the inspiration for much of the subsequent research.

2.2.1 Overview of Transformers and Longitudinal Health Data

Transformers

The Transformer, developed by Vaswani and colleagues [232], is based on the neural network and has become the state-of-the-art model for deep learning. As shown in Fig 2.1, the Transformer is made up of two components, the encoder and the decoder, which use self-attention and fully connected layers. The purpose of the encoder is to extract features from sequence inputs and produce encoded representations, which are then leveraged by the decoder [214, 148, 156]. Transformers are created through stacking encoders and/or decoder blocks, and hence can be used in various ways, including encoder-only, decoder-only, and encoder-decoder models [214].

The layers within the encoder and decoder of the Transformer enables multiple layers of abstraction which helps improve the data representation [148]. However, the key to the Transformer is the attention mechanism. Self-attention is used to assign a weight to each element in a sequence, allowing the Transformer to consider the relevance of each input at the same time [148]. This enables the Transformer to attend to the most important information in the sequence, capturing complex and long-term relationships by processing entire sequences in parallel [210, 214, 132]. A more detailed description of the Transformer components is provided in Chapter 3. Readers are also referred to the original paper on the Transformer [232] for more detail on its architecture.

Two of the most successful transformers include the Bidirectional Encoder Representations from Transformers (BERT) and the Generative Pretrained Transformer (GPT). BERT is a well-known transformer adaptation that has achieved state-of-the-art performance on numerous natural language tasks [210]. It uses the encoder-only component to generate contextual representations of inputs in a sequence. Broadly speaking, encoder-

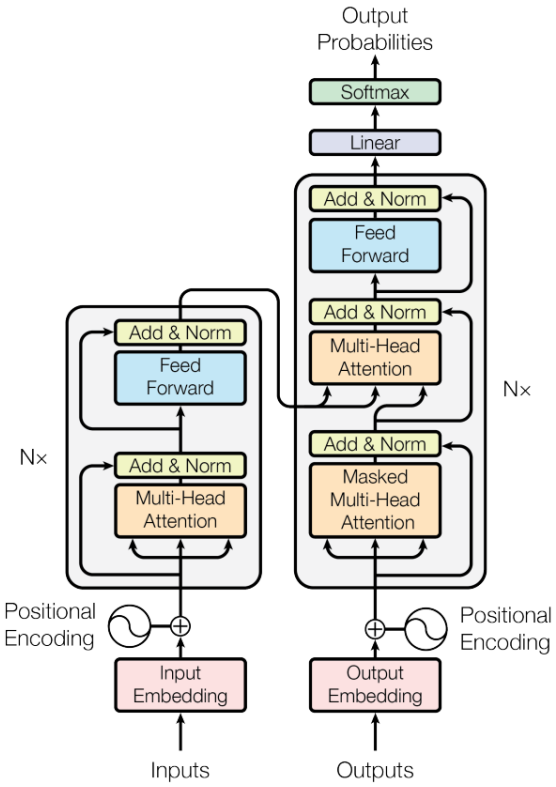

Figure 2.1 : The Transformer architecture, copied from Vaswani et al. [232].

only models like BERT are primarily used for language understanding including classification and question-answering tasks, and are thus termed discriminative models. In contrast to BERT, GPT is a decoder-only transformer that is used to generate the next word in the sequence and has been widely popularised via OpenAI's ChatGPT[3] tool [210]. GPT is primarily used for generative tasks, like the generation of text, and thus is referred to as a generative model [132]. More details on these models can be found in the original BERT [52] and GPT [179] papers.

Both BERT and GPT are considered to be foundation models [24]. Foundation models can be defined as large-scale models trained on vast amounts of information that can then be used for numerous tasks [156]. The success of foundation models is primarily due to the benefit of the two-step training paradigm, where the model (generally the Transformer) is firstly pretrained in a self-supervised fashion to develop a general-purpose model [156, 132]. In the second step, the model is fine-tuned using alternative data sources for numerous downstream tasks. Models that use the BERT architecture are

[3] https://chatgpt.com

typically pretrained via an approach called Masked Language Modelling (MLM), which refers to masking elements of a sequence and pretraining the model to predict the masked items. As indicated in the name, BERT operates in a bidirectional fashion during pretraining [156]. GPT-based models are pretrained using next item prediction, which refers to prediction of the next element in the sequence [156]. Here the model only has access to the previous elements in the sequence (hence pretraining is unidirectional) and predictions are made one element at a time in an auto-regressive fashion. Compared to traditional machine learning models, foundation models (such as BERT and GPT) typically offer better predictive performance, require less labelled data (if using pretraining/fine-tuning paradigm), and can handle multi-type inputs [243].

Large Language Models (LLMs) are a subset of foundation models trained largely on natural language to solve related tasks, and typically use the transformer architecture as the backbone [156, 132]. While BERT and GPT can both be classified as LLMs, LLMs are typically referred to as being a type of GenAI [37], which would fit the profile of GPT but not BERT (which is a discriminative model). GenAI is a type of machine learning that generates new data from the original data. It can synthesise large amounts of data to learn underlying data distributions and model the joint probability of many potential options or outcomes [193]. This enables generative models to create synthetic data, simulate scenarios, or test hypotheses under different conditions, in addition to performing traditional predictive tasks [193]. Literature suggests GenAI is revolutionising healthcare [153, 133], indicating the benefits of the Transformer architecture and large-scale pretrained models for learning relationships in healthcare data.

Longitudinal Health Data

Longitudinal health data can be defined as repeated health data collected over a patient's pathway or history [271]. Longitudinal health data is frequently captured in two key health data sources: EHR data, sometimes also called electronic medical records, and claims data. EHRs are typically found in hospitals or primary care centres and are a digital record for patient data generated during routine healthcare [226, 10]. The information within EHRs can be divided into structured (diagnoses, medications, procedures, demographics), and unstructured (clinical notes) data [226, 10]. This research is focused on

structured longitudinal health data, and hence any reference to longitudinal health data from this point refers to the structured aspect of the data unless otherwise indicated.

Compared to EHR which is often limited to a specific site or number of sites, claims data typically captures activity across various health infrastructures, and thus offers potential for larger and more diverse populations [243, 36]. This is because they are usually owned by governments or insurance companies, leading to different types of claims data, including administrative claims and insurance claims, respectively. This research utilises administrative claims data, specifically Australian Government administrative claims data, collected when persons access subsidised services in the Australian healthcare system [25]. The two main health-related claims data include the Medicare Benefits Schedule (MBS)[4] and the Pharmaceutical Benefits Scheme (PBS)[5]. The MBS captures use of healthcare services such as doctor and specialist appointments, pathology tests, and imaging, while the PBS captures use of subsidised medicines. Hospital data[6], which contains diagnosis codes and procedures, as well as cancer registry data[7], which records cancer-specific information like diagnosis and clinical features, are also considered longitudinal health information in this research as they capture patient information over time.

There are many advantages of longitudinal health data for pathways analysis and predictive modelling. A key benefit is the ability to capture and analyse temporal relationships, as this is not possible in cross-sectional analysis (collected at a single point in time) [271]. This feature is critical in healthcare, due to many conditions like cancer being long-term or chronic. Because of the historical information included, longitudinal health data are also useful in providing initial disease symptoms or early indications [226]. The longitudinal nature also allows for forecasting disease risks, capturing treatment, and monitoring patient pathways over time [154, 226]. In addition, longitudinal health data can offer opportunity to access large and diverse real-world populations across different parts of healthcare [260, 106]. It also provides a cheaper and more practical way to analyse patient pathways that reflect actual clinical care [31, 154, 215, 147].

---

4 https://www.mbsonline.gov.au/

5 https://www.pbs.gov.au/pbs/home

6 https://www.health.vic.gov.au/data-reporting/victorian-admitted-episodes-dataset

7 https://www.cancervic.org.au/research/vcr

For cancer, longitudinal health data can provide a more comprehensive capture of cancer treatment and symptoms [19]. This can include complications or adverse events, cancer treatment, cancer spread or metastasis, and cancer recurrence [271]. This extends into the understanding of cancer pathways and outcomes, effectiveness of cancer treatments, and how policy changes impact oncology outcomes and care [215]. Therefore, prediction models for cancer often will leverage longitudinal health data, due to the detailed information contained within these data sources [147, 97]. There are several examples of the Australian MBS and PBS data being used to explore risk factors for cancer, as well as variability in medicines and costs of cancer care [196, 115, 110, 252], indicating its suitability for capturing elements within cancer pathways.

While longitudinal health data does provide information on patient pathways, the characteristics of the data pose many challenges for predictive models. These challenges are also nearly always exacerbated within oncology. Longitudinal health data is high-dimensional [269]. There are thousands of medical codes reflecting diseases, medications and healthcare services, leading to many dimensions. This is apparent in oncology where growing data, treatment options, and advances in computation infrastructure have resulted in high-dimensional patient data [99]. In addition, as a patient progresses through the cancer care pathway, their data grows as new information is generated [99]. Moreover, longitudinal health data is highly heterogeneous, containing many types of information and diverse patient pathways and outcomes [147]. For cancer, outcomes include the state of the patient, determining whether a tumour is cancerous or not, cancer stage, treatment-related outcomes, and survival [6]. Different cancer tumour streams can be treated in different ways, each with distinct and complex care pathways, and there can be high variability between populations [257].

Furthermore, longitudinal health data is temporal and irregular. Healthcare information is delivered over time and there may be correlations between consecutive or repeated inputs [147]. Cancer care is inherently long-term and includes diagnostic, treatment, and follow-up or end of life care phases [44]. This contributes to it being a longitudinal disease than can span for years. Across patient pathways, diagnoses and treatments can occur irregularly, contributing to differences in timing of healthcare interations and wide varia-

tion between patients [76]. For cancer, individuals may receive differing cancer screening or follow up monitoring due to different risks for cancer or recurrence, contributing to the irregularity issue [271]. In order to address the above aspects of modelling cancer pathways in longitudinal health data, advanced deep learning methods such as transformers are needed [76, 154].

2.2.2 Benefits of Transformers for Longitudinal Health Data

Although originally developed for Natural Language Processing (NLP), the Transformer can be adapted to model longitudinal health data by considering sequences of healthcare information as a language problem [236]. For example, a patient pathway can be thought of as a sequence of healthcare events, where diagnosis codes can be treated as words, short-term relations such as a visit can be treated as a sentence, and a patient's history treated as a document [214]. Therefore, instead of transformers being used to learn relationships between words in a sentence within a document, they can be used to learn relationships between medical codes (such as diagnoses) in a visit within a patient's history [76]. Modelling the data in this way offers great opportunity to learn intricate relations in patient pathways.

The literature identifies many advantages in using deep learning methods including transformers with longitudinal health data. First is the ability to model the intricate, non-linear healthcare relationships. Statistical methods such as logistic regression have largely been the preferred approach for modelling longitudinal health data [271]. However, regression models lack consideration for the rich and complex relations in longitudinal health data [236]. In addition to this is the capture of temporal and sequential relationships within longitudinal health data with transformers [226], which is often not possible using traditional statistical or machine learning approaches [47]. Furthermore, research has indicated that for health data, repeated measures for an individual are often interrelated, and not considering these relationships could result in bias [31]. This further contributes to the advantages of transformers compared to statistical or traditional machine learning models [76].

Another benefit of transformers is the ability to process and analyse large data volumes, allowing models to learn trends and patterns that may not be apparent to clinicians,

enabling enhanced predictive capability [171, 268, 193]. While traditional statistical and machine learning methods may perform well when investigating associations between a small number of features, it is often beneficial to consider more complex relationships through deep learning when the feature space is large [31]. Therefore, being able to process large amounts of data means that transformers can be trained on the entire length of a patient's pathway to produce comprehensive but low-dimensional patient representations to improve predictions [243]. Additionally, deep learning methods require minimal feature engineering and can learn directly from the data to operate as end-to-end-systems [154, 76]. They also have the benefit of being able to accept and consider multimodal data in predictions, for example, through combination of medical history data with imaging and/or omics data [268, 269].

The Transformer also has advantages over recurrent networks, which are typically biased towards the most recent inputs and not as computationally efficient as transformers in handling sequences [236]. RNNs struggle to retain context when sequences become long because all information in the sequence must be processed before generating results, and hence suffer from the 'vanishing gradient' problem, where inputs are essentially 'forgotten' [214]. The Transformer avoids the vanishing gradient problem by using the attention mechanism to focus on the most important information in the sequence [214]. The architecture of the Transformer, containing multi-layer multi-head self-attention mechanism, allows parallel processing, enabling the learning of complex, longer-range dependencies [236]. Its structure is also more flexible in learning relations between elements of a sequence through use of positional encoding [236]. This makes the Transformer highly successful in efficiently modelling large, complex, longitudinal health data.

Finally, it has been suggested that the greatest benefit of transformers is the potential to continually improve over time [125]. This is supported by the earlier mentioned model pretraining, which can enhance model flexibility and generalisation [171]. This learning paradigm equips pretrained models (e.g., foundation models and GenAI systems) with a knowledge-rich starting point, enabling more accurate predictions across populations and time [74]. These models can also be repeatedly updated and fine-tuned to reflect new data and knowledge [258], enhancing opportunity for more real-time and rapid responses. This

is important for population health surveillance and point-in-time care decision making [227]. Moreover, pretrained models can provide multi-purpose use, reducing the need for extensive labelled data or training an independent model from scratch for every type of task (e.g., prediction, clustering), also enabling more rapid insights [227]. These benefits contribute to the Transformer being an ideal model to provide scalable solutions with longitudinal health data in oncology and healthcare.

2.2.3 Early Transformer Models for Longitudinal Health Data

The success of transformers in NLP naturally led to interest in application of these models on EHR data. Given these similarities between text and EHR, early approaches opted to leverage the BERT [52] training strategy. The work of Li et al. [119] and Rasmy et al. [188] pioneered the application of transformers to EHR, with the BERT for EHR (BEHRT) and Med-BERT models becoming seminal works in the field. Both these studies attempted to develop general-purpose EHR representations that could improve the prediction of diseases by leveraging a pretraining and fine-tuning framework. As earlier mentioned, this framework operates by the two-step training method, whereby pretraining occurs via self-supervised learning to create generalised embeddings, and these are then utilised in fine-tuning for specific tasks. Model pretraining and transfer learning is further discussed in Section 2.5.

The first of these works, BEHRT, was published in 2020. Using a population of nearly 1.6 million individuals from the United Kingdom, BEHRT was developed as a large-scale transformer to improve patient prediction tasks. To better suit EHR data, the original BERT [52] architecture was modified to include patient age information and introduce a segment embedding. The segment embedding was designed to distinguish between patient hospital visits, as discrete medical codes typically do not have individual positions within a visit (unlike words in a sentence). The study used the age information with diagnosis codes to develop a general-purpose representation using the Transformer, leveraging the original MLM task from BERT. The model was then evaluated on disease prediction tasks in the next visit, next six months, and next twelve months. For each of these prediction tasks, BEHRT obtained better AUC and average precision score metrics than two other popular non-transformer deep learning models at the time (Deepr [157] and RE-

TAIN [41]). This work provided initial evidence on the benefit of large-scale transformers for EHR and patient predictions, and argued that BEHRT provided more efficient training compared to a recurrent network.

Within a year of BEHRT, Rasmy et al. [188] published their transformer model for EHR called Med-BERT. This model was similar to BEHRT in that it aimed to produce generalised embeddings for patient predictions. However, it leveraged a population of 28.5 million individuals for pretraining, giving a much larger cohort and vocabulary. It also differed from BEHRT in that the model utilised a secondary domain-specific task for pretraining (besides MLM, which is referred to as the prediction of Prolonged Length of Stay (PLOS) (greater than 7 days) in a patient's history. Med-BERT was evaluated by predicting two diseases, one of which is a cancer task: prediction of heart failure among diabetes patients and onset of pancreatic cancer. The cancer prediction task was evaluated on both an internal and external dataset (of which the external dataset was claims data), indicating how Med-BERT could facilitate predictions in other data and in the cancer context. Experiments demonstrated that Med-BERT did provide improvements to prediction accuracy in terms of AUC scores for the two prediction tasks, validating the benefit of the modelling approach.

Albeit early models, both BEHRT and Med-BERT had a number of limitations. First, Med-BERT was trained on two disease classifications, using both International Classification of Diseases (ICD), Ninth Revision (ICD-9) and Tenth Revision (ICD-10) diagnosis codes, rather than mapping all codes to a single classification. This may have increased the vocabulary and learning complexity unnecessarily. In addition, it may have also affected the capture of historical diseases and patterns, thus impacting the learning of longitudinal pathways. BEHRT, on the other hand, does map all codes to a single classification. However, the authors map to Caliber[8] codes, which results in only 301 diagnosis codes for modelling [108, 50]. While mapping to Caliber may be appropriate for small datasets, it is likely suboptimal for training large transformer models due to loss of information. Both Med-BERT and BEHRT also considered limited features for training (primarily diagnoses), negating other aspects of the EHR.

[8] https://www.hdruk.ac.uk/case-studies/caliber/

Second, Med-BERT defines an ordering or position of diagnosis codes within a visit, which is not used for BEHRT. The ordering of diagnosis codes is non-standard in EHR data, as this information is usually not available. The authors also suggest in the paper that the code ordering may not be reliable [188]. While the Med-BERT study does not explore the impact of diagnosis code ordering, later work examining this suggests ordering does not impact model performance [199]. Third, the fine-tuning prediction tasks used for BEHRT are essentially equivalent (all predicting diagnoses) but for different timeframes and have been criticised for being too simplistic [188]. Med-BERT is designed to predict conditions, rather than diagnosis codes, and this is suggested to be superior to BEHRT as condition prediction is a more complex task, requiring understanding of disease trends or patterns, supporting the generalisability of the model [188].

Finally, the approach for the fine-tuning of Med-BERT and associated results make it difficult to draw concrete conclusions on the overall benefit of the Transformer architecture. For example, the authors implement Med-BERT on top of existing RNN models, finding that Bidirectional Gated Recurrent Unit (GRU) + Med-BERT outperforms Med-BERT alone. In addition, Med-BERT only appears to add value when the number of fine-tuning samples are 1000 or more, with logistic regression being the superior model with smaller sample sizes. This raises questions as to the optimal transformer-based architecture and data sizes for patient modelling.

2.2.4 Summary

In summary, the Transformer is well-suited to model complex cancer pathways in longitudinal health data and develop predictive models. The benefits of the Transformer have been demonstrated in early models such as BEHRT and Med-BERT, which provide initial evidence on the opportunity to model large-scale longitudinal health information for improved health-related prediction tasks. However, these early studies also highlight that adaption of BERT-based models to EHR is non-trivial, prompting further research to improve these models for longitudinal health data.

2.3 Multi-task Learning with Longitudinal Health Data

One approach that has been explored to enhance patient representations and predictions in longitudinal health data is use of MTL. The purpose of this section is to provide an overview of MTL and summarise its use with longitudinal health data and for cancer-related predictions.

2.3.1 Brief Description of Multi-task Learning

MTL is a machine learning technique that aims to improve predictions through the joint training of multiple related prediction tasks. The main avenue by which MTL offers improvements to predictions is through shared data representations, where information from one task can be shared to a second task to facilitate predictions [198]. MTL is a special type of multi-output learning, which focuses on predicting multiple outputs given an input [250]. While multi-output learning often uses the same training data and features for predictions, MTL can be trained using either the same, or different, data and features [250].

In comparison to Single-Task Learning (STL), MTL has several advantages for improving prediction models. It helps reduce model overfitting, leading to more generalisable models, and increases the size of the training data, which helps in instances where datasets are small [198]. Moreover, MTL can provide more relevant features in highly heterogeneous data by identifying useful features across tasks, helping to address data sparsity [198]. All these features can provide more robust models and lead to improvements in predictive tasks [267]. Further detail on MTL can be found in a number of summary papers [198, 267, 266, 48, 259].

In addition to the data representation and modelling benefits, MTL better resembles the real-world decision-making process for patients in healthcare systems [79]. Clinicians use a range of information to help inform care decisions, and often need to consider a patient's risk of several outcomes, such as survival, side-effects, quality of life, and treatment outcomes [150, 73]. This is of particular relevance to cancer patients; due to the long-term and complex treatment patterns they can experience. Finally, there is also evidence to suggest that there are relationships between patient outcomes, such as hospital

readmission and diagnosis [27], and hence modelling these tasks together with MTL may provide more accurate predictions.

2.3.2 Multi-Task Learning with Longitudinal Health Data

The use of MTL with deep learning algorithms has been growing for EHR data, with early approaches focused on RNN models. For example, in 2019, Harutyunyan et al. [79] conducted a detailed study on MTL with Long Short-Term Memory (LSTM) networks to simultaneously predict four hospitalisation outcomes in the Medical Information Mart for Intensive Care (MIMIC)-III data: in-hospital mortality, patient deterioration, length of stay and phenotype classification. The study found that MTL improved prediction performance of all four tasks compared to single task predictions. MTL has also been successfully used for improving mortality [213, 121, 224, 158] and diagnosis prediction [191, 40, 239]. While these studies indicate the potential of MTL with EHR data, they were all based on RNN architectures, not transformers.

McDermott et al. [139] did provide a comparison on the performance of RNNs with transformers in a MTL framework. This paper reported that the RNN based on GRUs provided the best model, outperforming the Transformer. It also found that multi-task prediction of health outcomes decreased model performance, questioning the benefit of MTL compared with STL. The authors, however, made a number of methodology decisions that make it difficult to interpret the study findings. These include providing a reduced number of samples to the Transformer, not using positional encoding, and the use of many tasks for MTL, without confirming if there were relationships between them. These aspects may have implications on the Transformer and multi-task prediction performance.

Other studies have also explored MTL with transformer models for predicting patient outcomes. Shickel et al. [211] used used data from the Intensive Care Unit (ICU) in the University of Florida Health to develop a flexible transformer model to predict seven clinical outcomes: two relating to hospital readmission, and five relating to mortality at different intervals. They expand BERT's classification (CLS) token to enable the multi-task prediction of patient outcomes and compare the method to GRU models. The results demonstrated the Transformer model was superior to the GRU models based on AUC

scores. In a similar study, Ma et al. [129] used a multi-task fine-tuning strategy with a transformer to predict nine post-operative complications from EHR data. For both of these studies, however, MTL was used as the sole prediction approach, with neither study evaluating the difference between multi-task and single task learning.

Hur et al. [88] proposed GenHPR, which leveraged a multi-task multi-source framework for predicting outcomes in MIMIC EHR data. The authors found their framework outperformed other comparative methods for both single-source and multi-source data, with multi-source generally offering performance benefits compared to single-source. While average and individual AUC metrics are provided for the twelve prediction tasks, this study, similar to those above, did not evaluate the impact or benefit of MTL.

In contrast, Chan and colleagues [33] evaluated their proposed model, MulT-EHR, against single-task baselines. MulT-EHR leveraged a heterogeneous graph transformer and a multi-task module to share representations among tasks in EHR data. The model had the best performance (in terms of AUC) on mortality, readmission, and length of stay prediction, as well as drug recommendation on the MIMIC-III and IV datasets. It was also consistently superior to single-task learning approaches including a vanilla transformer, verifying the benefit of multi-task learning.

Other approaches have leveraged MTL in the pretraining of transformer models. For instance, Med-BERT's [188] two-task pretraining strategy can be considered multi-task learning, as well as other models that have followed this methodology [116]. Similarly, in a claims data example, Zeng et al. [262] used a joint training approach consisting of medical event reconstruction and expenditure prediction to train a patient representation, and model relationships among paediatric claims data. While the use of multiple tasks is generally considered beneficial for pretrained representations, some works have indicated that this can impede pretraining tasks [199]. Although these studies above generally support the benefit of MTL with transformers for EHR, further work is needed due to the noted discrepancies in findings and frequent lack of MTL evaluation.

2.3.3 Application to Cancer Prediction Problems

When considering application of MTL to cancer-related prediction, there are several use cases demonstrating benefit in this area. Nebbia et al. [155] explored multiple tasks to enhance breast cancer diagnosis prediction. This study found that the addition of clinical knowledge through simultaneous prediction of tasks improved the performance of cancer diagnosis. Also focused on breast cancer, Sainz de Cea et al. [32] found that a multi-task approach facilitated the prediction of abnormal findings and need for additional cancer screening. Both these studies used mammogram data as model input. Focusing on a different cancer type, Meng et al. [141] proposed a MTL model to predict survival for patients with advanced nasopharyngeal carcinoma using image data. They found that use of tumour segmentation as a secondary task was able to enhance the prediction of survival for patients. While these studies demonstrate potential for MTL in cancer-related studies, they focused on image data as opposed to longitudinal health data.

One study that did use EHR data in a multi-task cancer prediction problem is that by Pham et al. [172]. Here, clinical features were used to simultaneously predict the onset of several cardiac complications for breast cancer survivors. The authors hypothesised that a MTL framework would provide the ability to better capture interactions among cardiac complications, leading to more precise predictions for cancer survivors. Performance comparisons for single and multi-task approaches found that MTL generally offered better predictions of complications. While the authors did not demonstrate this using the Transformer (they instead use a combination of different deep learning architectures), they do include a transformer as a baseline model. However, comparing the Transformer to the proposed approach is non-trivial, as the proposed model uses a multi-view encoding framework that is not used with the Transformer.

Adding to this, the Med-BERT [188] study also leveraged a cancer prediction problem as part of model evaluation. However, unlike the pretraining of the model which was done as multi-task problem, the prediction of pancreatic cancer is a single task prediction. This, and the aforementioned works, suggest that MTL can be beneficial for cancer pathways modelling and predictions. However, the limited MTL applications for cancer with longitudinal health data compared to other data types (such as image) emphasises

that further work is needed with transformers and this data in the cancer context.

2.3.4 Summary

Overall, the use of MTL with Transformers has been growing on longitudinal health data and it has demonstrated potential to improve prediction of patient outcomes. However, the inconsistencies in findings between studies, and the observed lack of evaluation of MTL in some scenarios, indicate further research is needed. Additionally, the limited MTL cancer applications in longitudinal health data compared to other types of data (e.g., image) warrants additional exploration in this data source. Application of MTL with transformers to longitudinal health data has opportunity to address the multi-outcome nature of cancer pathways and improve prediction of outcomes for cancer patients.

2.4 Transformer-based Patient Longitudinal Modelling

Following BEHRT [119] and Med-BERT [188], transformers emerged to improve on these methods and develop better patient representations for modelling longitudinal health data. Some of these approaches focused on the integration of more patient information, to either provide a more comprehensive patient representation, or to model both dynamic and static (e.g., demographic) patient features in a single representation. Other methods combined additional patient features with approaches to better model the longitudinal nature of the data, such as via temporal or hierarchical modelling, to produce more context-aware patient representations. The purpose of this section is to summarise these areas of the literature, separating the longitudinal modelling of EHR and claims data, and including studies on applications to cancer pathways.

2.4.1 Integrating Additional Patient Features

Many approaches have explored the inclusion of additional patient features or information to improve transformer-based models for longitudinal health data. This has often been referred to in the literature as multimodal learning, which involves developing models that can represent and join information from multiple modalities or data types [223]. While deep learning applications often consider multimodal learning to occur when combining different types of data (e.g., structured, image, text data), many healthcare studies

have described multimodal learning as the integration of different information from the EHR such as diagnoses, medications, procedures, and demographics. Compared to learning with a single modality, it is generally considered that multimodal learning can improve model prediction performance. This is because the integration of multiple modalities can provide additional knowledge and richer data representations for predictions [20].

Not long after the publication of BEHRT and Med-BERT, studies began integrating other modalities from the EHR in an effort to improve patient representations and predictions. While BEHRT and Med-BERT focused primarily on relationships between diagnosis (and age for BEHRT), subsequent models built upon this by adding in other medical code data and patient information. The authors of BEHRT extended their model from two modalities (diagnoses and age) to four modalities, which included expanding medical codes to includes medications and adding calendar year [186]. Ablation analyses were undertaken to determine the importance of different modalities, with the study indicating that all four modalities provided the best performance for heart failure prediction. Also related to the work of [119], Li et al. [118] further developed a transformer approach to incorporate information on diagnosis, medications, procedures, tests, blood pressure, drinking and smoking status, and body mass index. The authors report that the addition of more modalities generally leads to improvements in performance, and that the contribution of features appears to be related to their frequency in the data. However, many of the features used, such as blood pressure, drinking and smoking status, and body mass index, are not available in typical EHR or claims data.

Meng et al. [142] further expanded medical codes to include procedures in addition to diagnoses and medications, and included gender as part of patient demographic information. This Bidirectional Representation Learning model with a Transformer on Multimodal data (BRLTM), used a single vector to represent all the medical concepts (diagnoses, medications and procedures). Similar to BEHRT, the authors also included a segment embedding to indicate where codes occurred within the same or a different hospital visit. Wang et al. [242] also used discrete medical codes, together with demographics, patient observations (laboratory tests, vitals) and admission date, to develop a transformer-based model to predict chronic kidney disease. However, neither of these

studies evaluated the impact of patient sociodemographic features on predictions.

Extended Med-BERT (ExMed-BERT) sought to incorporate even more information from the EHR. This model, proposed by Lentzen et al. [116] added medications, gender, state of residency, and age. A different embedding was used for each of the five modalities, which was then concatenated with quantitative clinical measures (including weight, body mass index, body surface area, height, temperature, blood pressure and heart rate) to predict severe COVID-19 disease. ExMed-BERT was compared to shallow machine learning and deep learning approaches, and experimentation was undertaken with different classification prediction heads. The authors first considered model performance without inclusion of the quantitative clinical measures, finding all ExMed-BERT variants outperformed baselines in terms of AUC and Area Under the Precision-Recall Curve (AUPRC) metrics. When quantitative clinical measures were included, performance of all models improved, with ExMed-BERT still providing the best prediction.

Finally, in a claims data application, Fouladvand et al. [66] proposed a Multi-stream Transformer for Predicting Opioid Use Disorder using large-scale administrative claim records. The approach included diagnosis, medication, and demographic modalities and found the model provided better F1 and AUC metrics than the original transformer encoder block. Unlike BEHRT, where age information was added for each visit (or time step), the demographic features (age and gender) were added only to the last layer of the model (similar to the clinical features of ExMed-BERT). While this would have not provided relational learning between diagnoses and age (as in BEHRT), it was suitable for the purposes of the model which was to make predictions at the patient level. It is also one of the only approaches on longitudinal health data that grouped the data into monthly intervals for modelling, providing a suitable option for claims data, as it does not have strict visit level information like the EHR. Despite this, the work does not include ablation studies with differing modalities, and hence the impact of additional modalities is unclear.

2.4.2 Temporal and Hierarchical Modelling in EHR

In addition to leveraging more patient information from longitudinal health data, other research sought to better capture the temporal or hierarchical relations in the EHR to

improve patient representations. Key model developments that explore these approaches are discussed below.

Pang et al. [163] proposed CEHR-BERT, which incorporated artificial time tokens to better capture the temporality of EHR data. The tokens were defined based on the length of time between neighbouring visits and depended on whether the visits were less than 28 days apart, 29 to 365 days apart, or greater than 365 days. Tokens were also used to define the start and end of a visit. When compared to BEHRT and Med-BERT, CEHR-BERT provided the best performance (in terms of AUC and AUPRC metrics) in the four prediction tasks, which related to hospitalisation, death, new heart failure diagnosis, and readmission for heart failure patients. Through ablation studies, the authors confirmed the usefulness of embedding time tokens, as well as inclusion of the start and end visit tokens. CEHR-BERT did include different modalities (condition, drug, procedure, age) and similar to BRLTM, aggregated all medical codes into a single sequence. While the authors did evaluate differing embedding structures (without temporal concept embeddings, or time-age embeddings), and found these were useful to performance, they did not explore the impact of individual modalities.

Whilst including more medical codes (i.e., diagnosis, medication, procedure) can provide additional information to benefit patient predictions, it also increases the length of the EHR sequences and transformer model complexity. To address this limitation of CEHR-BERT and BRLTM, Rupp et al. [199] proposed ExBEHRT, to add more EHR modalities (including demographics, clinical characteristics vital signs, smoking status, diagnoses, procedures, medications, laboratory tests) without growing the length of the patient sequence. Similar to ExMed-BERT, ExBEHRT separately embedded each of the EHR concepts. This enabled the modalities to be stacked vertically in the input, preventing the length from exploding horizontally. The authors compared ExBEHRT to BEHRT and Med-BERT and found it outperformed all baseline methods in mortality and readmission prediction. However, the authors did not compare ExBEHRT to an equivalent but horizontally stacked model, making it unclear from a performance perspective on the effect of vertically stacked features. The authors did, however, conduct feature ablations to look at the impact of adding procedures, labs, or observations to BEHRT and found that the

inclusion of all features (i.e., in ExBEHRT) generally provides the best predictions, with procedures giving the most substantial performance benefit. This supports the findings of other studies on the benefits of additional modalities.

Other methods used hierarchical structures to capture temporal dependencies or model long sequences. Luo et al. [126] proposed HiTANet, a Hierarchical Time-Aware Attention Network for risk prediction on EHR. The approach used a hierarchical structure to model time information at the local and global level. At the local level, a time-aware transformer incorporated time information into the attention weights for each visit. At the global level, a time-aware attention mechanism was used to identify key timestamps in a patient's EHR. The attention at these two levels was then fused to obtain the final patient representation. Compared to baselines which included shallow machine learning, recurrent and attention-based models, HiTANet demonstrated superiority in terms of F1 scores for three disease onset predictions: Chronic Obstructive Pulmonary Disease (COPD), heart failure, and kidney disease. This demonstrated the benefit of hierarchically modelling temporal information, however only included diagnosis codes in visit inputs.

Wang et al. [240] proposed an approach for hierarchically multimodal EHR data. It included as input five modalities – including patient demographics, temporal clinical features, diagnosis codes, drug codes, and clinical notes. The model was trained to learn relations at different hierarchical levels of the EHR, including at the patient, admission and stay level, and was evaluated on health prediction tasks. The authors found that inclusion of their model framework with the HiTANet model boosted its performance in amnesia prediction (based on the AUPRC metric), supporting benefits of considering the EHR hierarchical structure. This approach did not evaluate the impact of different modalities; however, ablation studies did confirm the value of hierarchical learning.

Finally is the work of Li et al. [118], who extended BEHRT to create the hierarchical model Hi-BEHRT. Similarly to ExBEHRT, the approach aimed to address the issue of long sequences which increase the complexity of transformers. Based on the idea that medical events have stronger semantic relation when close in time, the authors proposed a sliding window over the EHR, splitting it into smaller segments. A transformer is applied to capture local features within each segment, and then all segment representations are

provided to a second transformer to aggregate the features prior to pooling and prediction. The model also received multimodal input including diagnoses, medications, procedures, tests, blood pressure measurement, drinking status, smoking status and body mass index. Hi-BEHRT was evaluated by predicting 5-year risk of four chronic diseases: heart failure, diabetes, chronic kidney disease, and stroke. Compared to baselines BEHRT, Med-BERT, and CEHR-BERT, Hi-BEHRT provided the best performance metrics on the four disease prediction tasks. Given the Hi-BEHRT architecture also allowed the model to accept larger sequence lengths (max = 1220), Hi-BEHRT's performance was also evaluated on longer sequences, where it remained the best performing approach.

2.4.3 Temporal and Hierarchical Modelling in Claims Data

Although modelling on EHR data has dominated the literature, there have been some application to claims data. For example, Kodialam et al. [105] proposed the Self-Attention with Reverse Distillation (SARD) model and demonstrated the potential for multimodal transformers to improve predictions on insurance claims data. The architecture, inspired by BEHRT, aimed to deal with data sparsity and irregularity by combining contextual and temporal embeddings with self-attention. Unlike BEHRT, which operates on code embeddings, SARD is applied to visit embeddings, and included additional modalities to BEHRT, including medications, procedures, and physician speciality. When compared to the recurrent model RETAIN [41] and BEHRT, SARD provided better predictions for risk of patient death, surgery, and hospitalisation.

Lahlou et al. [109] used a time-aware attention mechanism to learn temporal relationships in Medicare claims data. This study leveraged diagnostic, procedural, and demographic information, and demonstrated potential to improve the prediction of unplanned hospital readmission. The work also investigated smaller transformer architectures and found a BERT model with only two hidden and attention layers, and a hidden dimension of 512 was optimal according to the best-fit learning curve. This suggested that smaller-scale BERT models may be appropriate for claims data. However, the study did not evaluate the effectiveness of different data modalities on prediction.

Zeng et al. [262] proposed a Transformer-based Multimodal Autoencoder (TMAE) for patient representation learning from paediatric claims data. The approach used mul-

timodal input – considering information on medical codes, utilisation (code category or type – i.e., inpatient, outpatient or pharmacy), expenditure, and demographics (age and gender). Like SARD, the model aimed to tackle irregular time intervals and data sparsity, augmenting diagnosis codes with Clinical Classifications Software (CCS) groupings to alleviate the sparsity of rare medical codes. Embeddings were produced for each of the inputs, concatenated, and then a max-pooling layer was used to capture the relations within each visit. Demographic information was then added, and the initial element of the encoder served as the patient embedding (equivalent to the [CLS] token in BEHRT). TMAE was investigated via clustering tasks (not prediction tasks), and like Lahlou et al. [109], there is no ablation study on the contribution of different types of claims data or modalities. A later study [263] by the same authors and using the same data did look at the impact of different modalities and found that medical codes alone provided good performance, with small gains from addition of demographics and utilisation information. However, for both this model and TMAE, there is a lack of detail on how a visit was defined for the claims data, and whether these included claims items that occurred on the same date or grouping by an alternative method.

Another study that looked at rare conditions is RareBERT [176], an extension of Med-BERT to detect rare diseases from administrative claims data. It built on the Med-BERT architecture by addition of a 'type' embedding – representing whether a token relates to a diagnosis, procedure or treatment (i.e. claims type), and a 'temporal reference' embedding – incorporating event position relative to an index date in the patient pathway. RareBERT was evaluated on two rare diseases and outperformed Med-BERT (based on AUPRC metrics). The authors attributed the robustness of RareBERT to the use of the special [CLS] token to represent the patient, compared to the feed-forward layer used in Med-BERT. This was supported by studies comparing RareBERT with an equivalent model but with a feed-forward layer (i.e., no [CLS] token), which showed a 14% difference in performance. The authors also evaluated the impact of different inputs on RareBERT, finding that each embedding (token, event type, visit number, and temporal reference) provided useful information for predictions. However, RareBERT did not include any demographic information, and used the same mapping of diagnosis codes to CCS category as BEHRT, which may be too aggregated to learn low-level information.

Additionally, similar to TMAE, it is unclear how a visit is defined for claims data, indicating further detail is needed for modelling this type of longitudinal health data.

2.4.4 Applications to Cancer Predictions

While much research has focused on chronic conditions or general health predictions, there is some application of transformers to cancer-related predictions using longitudinal health data. For example, the earlier discussed ExBEHRT [199] model is evaluated on six- and twelve-month mortality predictions for cancer patients from time of diagnosis. The study reported that the ExBEHRT model, which includes additional modalities from the EHR, can outperform BEHRT. The authors also investigated predictability of ExBEHRT on the top ten most common cancers, as there can be variability in predictions by cancer type. The AUC metrics ranged from 66.5% to 72.6% for the six-month mortality task, and from 68.4% to 78.5% for the twelve-month mortality prediction. Although there was variability in the class labels for different cancer types, the results were close to the overall mortality results (AUC 6-month 71.5% and 12-month 74.3%) and hence the authors conclude no bias towards certain cancers.

Another previously mentioned model, BRTLM [142], focused on depression prediction for those with breast cancer. This model leveraged five modalities from the EHR and made predictions at multiple different time windows. Comparison of the approach to baselines (which included BEHRT) suggested additional data modalities offered improved predictions for depression (at short (2-week) and long-term (1-year)) for breast cancer. However only the clinical notes were evaluated in terms of impact on performance. In addition, only six months of patient history was used for each patient to avoid bias toward longer histories, and this is likely insufficient for cancer-based analysis as care and treatment tend to be lengthy. Furthermore, position embeddings are assigned to each code in the entire patient sequence, with no explicit position information provided at the visit level.

Another model that provided predictions for breast cancer patients is Multimodal BEHRT (M-BEHRT), proposed by Mbaye et al. [138] and based on the BEHRT architecture. This model focused on the prediction of disease-free survival for breast cancer patients treated with adjuvant chemotherapy, which is a key outcome of interest, and

is important in informing treatment plans. The approach included both structured and unstructured health data. The structured data component, termed Tabular BEHRT, included clinical features (age, cancer tumour details), biological measurements, therapies (surgery, chemotherapy etc.), and nature of visit information. The unstructured data component, termed Text BEHRT, included free-text information from medical reports. The M-BEHRT study found that the combined model which used both Tabular BEHRT and Text BEHRT components provided the best performance in terms of AUC, outperforming the commonly used prognostic tool for breast cancer, the Nottingham Prognostic Index by 10%. Ablation studies revealed that clinical features were most critical to performance, supporting the inclusion of this information in cancer prediction models. Ablation analysis also found that addition of Text BEHRT to Tabular BEHRT only provided a small increase in predictability (AUC 0.75% vs 0.77% with addition of text BEHRT), suggesting that Tabular BEHRT can capture majority of the features important for prognosis. While this study does utilise unique cancer-specific features and includes some procedure and treatment information, the data is dissimilar from other EHR models that leverage medical concepts (i.e., diagnoses, medications, procedures).

Ho et al. [83] developed a model to predict colorectal cancer recurrence with multimodal healthcare data. The authors explored multiple approaches to extract and integrate patient features from multiple modalities. They included both tabular data (demographic information, tumour characteristics, laboratory test results and treatment parameters), and time series data (using measured levels of carcinoembryonic antigen). The study found that using multiple modalities was better than using a single modality, and that the transformer model with multiple modalities outperformed other deep learning approaches. This study, like M-BEHRT, did include cancer-specific tumour characteristics, which provide detailed cancer-related information that are typically used in practice to help determine cancer prognosis. However, unlike M-BEHRT, the study did not quantify the contributions of individual modalities.

Iyer and colleagues [91] developed an ensemble transformer-based model for predicting Barret's eosophagus and esophagael adenocarcinoma from EHR data. The model used several patient static features including age, gender, race/ethnicity, family history

and smoking status. A number of temporal features related to medications, comorbidities, body mass index, lab tests and patient notes were also included. By modelling these inputs with the transformer, the study found this architecture provided better performance for the two conditions compared to risk factor-based scores, further supporting the benefits of transformers for cancer pathways modelling. Unlike Ho et al. [83] and M-BEHRT [138], this study did explore the impact of the input features to compare against established risk factors. However, the authors do not compare the model with any other machine learning or deep learning methods, nor include cancer-specific characteristics.

Finally, Chen et al. [36] developed a transformer-based model for predicting lung cancer in claims data. They used diagnosis and medication information, together with age and gender to build their model, which was based on the Vision Transformer. The model outperformed traditional machine learning approaches and recurrent methods in terms of AUC scores for lung cancer prediction at all stage and early stages. Compared to the earlier mentioned TMAE [262] and RareBERT [176] models where there is lack of clarity on how a visit is defined in the claims data, this work groups inputs together in 30-day windows. The rationale provided is that appointments tend to be scheduled monthly and so this timeframe is appropriate for modelling progression. While this study provides a proposed approach for modelling claims data for cancer predictions, there is potential for more information to be included in the model, such as additional patient demographics and cancer characteristics.

2.4.5 Summary

Collectively, existing research with transformers highlights trends in including more patient information from longitudinal health data and exploration of strategies to better capture the temporal and hierarchical data relations. While it is evident that integration of additional patient features generally offers better predictions, not all approaches evaluate the impact of including these features or modalities, and there are considerable differences in the types of patient features included and how they are integrated into model architectures. Similarly, while temporal and hierarchical modelling can better capture the relations within longitudinal health data, there is variability in the methods and approaches used, resulting in lack of clarity on how best to encode and represent longitudinal health

data. Furthermore, majority of existing work (including for cancer pathways) has been developed for EHR data, leaving claims data understudied in the cancer context. More research is needed to model cancer pathways in claims data that captures longitudinal and static patient level detail.

2.5 Transfer Learning with Longitudinal Health Data

Owing to the success of transformers with natural language, another key direction following BEHRT [119] and Med-BERT [188] has been the development of general-purpose transformers. These have been popularised as they can be used for a range of prediction tasks with minimal additional effort by way of transfer learning. The purpose of this section is to provide an overview of transfer learning, and to review the development of pretrained transformers for EHR, claims data, and for cancer prediction problems. Many of the models discussed appeared in earlier sections of the review but are also included here as they utilised a pretraining component.

2.5.1 Transfer Learning and Pretrained Models

Transfer learning aims to improve predictions by sequentially transferring knowledge from a related domain to a target domain [270]. In this way, it is possible to utilise large unlabelled data sources to train a model to learn generalised knowledge, which can later be leveraged in the target domain to improve predictions where labelled data is present. Readers are referred to a number of review papers for a more detailed overview of transfer learning [270, 85, 56].

One of the main benefits of transfer learning is its ability to improve model performance in instances where datasets are small, or where there is limited labelled data available [52]. This scenario is particularly relevant to healthcare data, including cancer-related data, as it can be difficult to obtain large datasets (due to privacy requirements) or to label data without input from clinical expertise [212, 245]. In addition, many cancer-related prediction studies utilise sub-cohorts from larger populations. For example, prediction models may utilise data from only those with cancer (thereby excluding the population without cancer), or limit data to a specific cancer type, stage, or treatment regimen. These types of cohort studies reduce the data samples for modelling, and as a result, can be

challenging for deep learning techniques.

The use of transfer learning and pretrained models in NLP gained much attention following the development of the BERT model [52]. As earlier described, BERT uses MLM, together with next sentence prediction to learn contextualised relationships in language from large amounts of text data. Once trained, BERT can be fine-tuned using only one additional modelling layer for downstream tasks, such as for question and answering, and language inference [52]. Since its introduction, BERT has been widely successful for language-based tasks, including those related to the healthcare domain [18, 170]. Its use is also growing for longitudinal health data, where there are now several pretrained transformer models based on BERT. However, due to differences between longitudinal health data and natural language, BERT-based models have explored modifications to better tailor them to the characteristics of EHR and claims data.

2.5.2 Pretrained Models for EHR

Several of the earlier mentioned works included a pretraining fine-tuning framework as part of their proposed transformer model. For example, CEHR-BERT [163] used pretraining for modelling relations in the EHR from 2.4 million patients. The authors introduced a second pretraining objective called Visit Type Prediction to predict whether a medical code was an inpatient or outpatient activity. This objective was trained with MLM, so the model included masking of medical codes and masking of visit types. CEHR-BERT was compared to an equivalent model that did not utilise pretrained weights, which indicated that pretraining did provide benefits in downstream tasks. Experiments were also conducted on differing amounts of fine-tuning data, showing CEHR-BERT could provide good performance with just 5% of training data, with other models needing up to 80% for comparable metrics.

While CEHR-BERT did outperform BEHRT [119] and Med-BERT [188] in downstream tasks, its multitude of differences from these approaches (pretraining strategy, number of modalities, model design) make it challenging to determine the impact of the specific pretraining approach. Indeed, the authors suggest that further investigation is required on the Visit Type Prediction pretraining task to confirm the benefit was not due to random variation. In addition, the model includes patients with only one visit. Simi-

lar models have traditionally adopted a threshold of five or more visits (e.g., ExBEHRT [199]) in order to provide sufficient temporal information to the model and enhance the ability to learn longitudinal relations.

Hi-BEHRT [118], which relates to the hierarchical BEHRT model, also used model pretraining. Due to the hierarchical nature of this model using a sliding window and multi-level data representations, it was not straightforward to apply MLM. To address this, the authors used a self-supervised framework based on contrastive learning called Bootstrap Your Own Latent (BYOL) to create a modified version of MLM. Pretraining involved two strategies: EHR and segment augmentation. EHR augmentation focused on masked tokens, while segment augmentation used BYOL to compare sequences with true and modified (masked) segments and minimise the loss. In addition to the model providing better predictions for heart failure, it also provided benefits when only a very small amount of fine-tuning data was available (1% when using pretraining compared to 5% needed without pretraining). This benefit was no longer evident at 50% of the training dataset, however this may be explained by the relatively large fine-tuning training data of nearly two million samples.

Another previously described hierarchical model, the Hierarchical Multimodal Pretraining framework for the Medical domain (MedHMP) [240], proposed multi-level pretraining. This framework was designed to consider the hierarchy of EHR data at the stay level (time-series clinical variables and monitoring), admission level (series of medical codes and clinical notes), and patient level (multiple visits). At the stay level, self-supervised learning was used to reconstruct the clinical features and predict acute respiratory failure, shock, and mortality prediction. At the admission level, self-supervised learning was used to model both intra-modality and inter-modality relations, via masked code prediction (like MLM) and contrastive learning (masking EHR concepts), respectively. The model is first pretrained at the stay level, and this is used to initialise the admission level pretraining, evaluated by 30-day readmission prediction. There are eight downstream tasks used for evaluation, as well as comparison with eight baseline models.

Experiments with MedHMP demonstrated the following: 1) MedHMP was superior to baselines at the stay and admission level tasks, and use of the pretrained model generally

improved the prediction of patient health conditions (heart failure, COPD, amnesia). 2) The pretrained model always provided better performance even when varying the fine-tuning data (1, 10, 50, 100%). 3) Pretraining enabled the downstream model to converge faster. 4) Inclusion of both stay and admission level pretraining objectives was beneficial, as well as the use of two tasks at the admission level. While this study was one of the more comprehensive in terms of exploring pretraining in EHR, the authors only include ICD-9 codes. Rather than mapping ICD-10 codes, they exclude these codes, resulting in information loss. In addition, the methodology is not directly implementable on claims data, as claims data does not contain the stay level characteristics noted for EHR.

TransformEHR [255] took a different approach for pretraining. The authors claimed that existing methods (Med-BERT [188], BEHRT [119] and BRLTM [142]) only predicted a small amount of diagnosis codes within each visit, while their approach TransformEHR aimed to predict all diseases and outcomes in a visit. The model was more architecturally similar to the GPT model and used an encoder-decoder framework with past visit information to predict future codes, rather than the bidirectionality used in BERT-based models. TransformEHR was pretrained to learn the true EHR sequences from distorted sequences, where visit masking was used instead of code masking, and diagnosis predictions were made in an autoregressive fashion. Compared to BERT [52], TransformEHR was superior in the pretraining evaluation tasks of predicting the top 10 most common and uncommon diseases. It also outperformed BERT in the two disease prediction tasks for pancreatic cancer and intentional self-harm among post-traumatic stress disorder patients. While TransformEHR outperformed BERT, it is unknown whether it would outperform other BERT-based Transformer models that have been optimised for EHR (such as BEHRT or Med-BERT). It was also built using Veteran's EHR data and so the relations learned may not be applicable to all cohorts or populations. Furthermore, only diagnosis codes were used as input for TransformEHR, neglecting other modalities.

A more recent model HERBERT (HEalthRecordBERT), proposed by Moore et al. [149], attempted to implement a second pretraining task similar to the next sentence prediction task of the original BERT [52] model. For this task, which the authors term Next Visit Prediction, the model is provided two EHR sequences, where 50% of the time it

is the next visit in the patient's EHR and 50% of the time it is from a different random patient. The model then has to predict which is more likely. The authors suggest that this task is more appropriate for capturing long-term dependencies than the PLOS task used in Med-BERT [188], which is more reflective of predicting patient health status than likely future activity [149]. To aid the Next Visit Prediction task, a separation (SEP) token is used only at the end of a series of EHR visits to differentiate historical from future visits, as opposed to BEHRT [119] which uses a [SEP] token at the end of each visit. HERBERT was evaluated on a prediction task for transition from chronic kidney disease to end stage renal disease at 3 different time windows – less than 1 year, 1-2 years and 2-5 years. While prediction performance between HERBERT and Med-BERT was similar at the 1-year prediction (HERBERT AUC 0.91 vs Med-BERT AUC 0.894), HERBERT substantially outperformed Med-BERT at the 2-year and 5-year predictions. However, HERBERT only evaluated the pretrained model on one downstream prediction task, and so whether it provides benefits for other tasks is unclear. Based on ablation studies, the authors also suggest that the impact of alternative secondary pretraining tasks is less apparent, as only a slight performance uplift is evident when compared to pretraining with only MLM. Both these issues suggest further work is needed in evaluation of pretraining and fine-tuning tasks.

2.5.3 Pretrained Models for Claims Data

While not as extensive as on EHR data, there has been some use of transfer learning and pretraining with claims data. Closely related to this research, is the work of Zeng et al. [263], who proposed the Claim Pre-Training (Claim-PT) framework. This study leveraged a million patient records from paediatric claims data to develop a general-purpose pretrained model. The study built off the previously discussed TMAE model, where the authors used the same data and claims inputs – medical codes, claim type information, service date, and expenditure. The model was pretrained using two objectives: Next Visit Prediction, a modified MLM task to learn relations between claims items, and Categorical Prediction, to predict the type of claims input and designed to input medical knowledge.

Claim-PT was evaluated on two patient-level fine-tuning tasks – suicide risk prediction and asthma exacerbation prediction. Experiments showed the Claim-PT model was

superior to baselines (based on AUC metric) on the two downstream tasks. While a transformer encoder is included as a baseline, the study does not evaluate the transformer used in the Claim-PT framework without the pretraining component. This would be useful given there are noted differences compared to the standard transformer encoder (e.g., the use of pooling in Claim-PT), and thus the individual contribution of pretraining on predictions is unclear. The authors do, however, explore the impact of single pretraining tasks and fine-tuning approaches for Claim-PT. They found that while use of both pretraining objectives was optimal for predicting suicide and asthma prediction, Next Visit Prediction alone provided good performance, with Categorical Prediction providing only marginal improvement. The downstream predictions were also influenced by the fine-tuning approach. Use of fine-tuning (i.e., few-shot) compared to no fine-tuning (zero-shot) provided better suicide and asthma predictions, indicating the benefit of fine-tuning downstream tasks.

Finally, the authors also investigated for ClaimPT how pretraining could benefit cross-domain tasks. They found the pretrained embeddings were more useful when applied to a hold-out subset of the original data compared to an external dataset (demonstrated through use of MIMIC-III data). This can likely be explained by the differing populations of these two datasets (paediatric vs. intensive care), however there is little detail provided on what the cross-domain prediction task is other than indicating it is an auto-diagnosis task.

The aforementioned RareBERT [176] also leveraged pretraining with their model for rare diseases. Similarly to Claim-PT, the model was pretrained via masked event prediction (like MLM) and type embedding (whether code is a diagnosis, procedure or treatment). The approach additionally introduced adaptive loss to address imbalanced classes, which is common for rare conditions. However, like Claim-PT, the authors only explore two downstream tasks (two rare diseases) and find differing results on whether adaptive loss provides the best performance. The study also aggregates medical codes by mapping to CCS category (Like BEHRT [119]), which potentially results in information loss. Furthermore, while the study does compare RareBERT to Med-BERT [188], it does not examine the performance of the RareBERT model without pretraining, despite having some architectural differences from Med-BERT (e.g., number of layers, embedding size,

[CLS] token).

ExMed-BERT [116], which extended Med-BERT by adding more modalities, is also a pretrained transformer. Unlike Med-BERT, ExMed-BERT was trained on EHR and insurance claims data, and used a smaller population for pretraining (28.5 million in Med-BERT vs 3.5 million in ExMed-BERT). ExMed-BEHRT leveraged the same hyperparameters and pretraining tasks as Med-BERT – i.e., MLM and PLOS. However, in contrast to Med-BERT, which used both [ICD-9 and ICD-10 diagnosis codes, ExMed-BERT mapped all ICD codes to Phecodes, reducing the feature space and possible coding biases. While the results suggest that ExMed-BERT offers performance improvements compared to shallow machine learning and recurrent baselines, ExMed-BERT was only evaluated on one downstream task of acute respiratory manifestations using medical history of COVID-19 patients. There was also no inclusion of deep learning models without pretraining for evaluation. The authors also looked at the transferability of ExMed-BERT to external hospital data, finding no difference in performance between the pretrained model and a random forest trained from scratch. This further identifies the challenges of transfer learning from pretrained models.

2.5.4 Pretrained Models with Cancer-related Fine-tuning

The use of transfer learning and pretrained models is a growing area for cancer. However, there is limited research with respect to longitudinal health data. A recent review exploring the use of transfer learning applied to non-image clinical data reported that the majority of cancer-related applications on tabular or time series data had utilised genetic data [56]. There are only a handful of pretrained models on longitudinal health data that have made predictions related to cancer. Many of these have been mentioned in earlier sections, such as Med-BERT [188] and TransformEHR [255], which both include pancreatic cancer prediction as a fine-tuning task.

In addition, the BRLTM model by Meng et al. [142], applied MLM pretraining with their transformer model to predict depression for breast cancer patients. While the model did outperform baselines, authors did not validate the benefit of transfer learning by performing experiments without pretraining. Furthermore, the earlier noted restriction of patient data to the most recent six months may be problematic for cancer patients. While

this approach may reduce bias towards patients with longer EHR histories (as the authors claim), it does not consider the longer-term perspective of the patient, and cancer treatment and monitoring often extends beyond a period of six months [90]. The use of additional data is therefore desirable for model pretraining and prediction for cancer outcomes.

ExBEHRT [199] also used model pretraining and evaluated it on a cancer-related prediction task. Like most other approaches, they pretrained using MLM, but appear to only apply this for diagnosis codes in the study and not for other medical codes. This is likely due to the architecture of the model vertically stacking modalities rather than as one sequence, however this approach for pretraining may result in some relational loss. One strength of the approach is that rather than using only ICD-9, or multiple diagnostic classifications, they map all ICD-9 codes to ICD-10. This means they only use one coding structure, allowing longitudinal relational understanding using the most current classification, while appropriately reducing the vocabulary and dimensionality. ExBEHRT is also implemented with and without the PLOS objective used in Med-BERT and found that this task did slightly harm the MLM performance. Nevertheless, they use both models in downstream predictions and find that while the PLOS task was not beneficial for cancer prediction, it was useful for predicting readmission for heart failure patients. This suggests that secondary pretraining objectives can potentially provide useful context for patient predictions even if they reduce MLM performance. Finally, ExBEHRT also investigated the ordering of medical codes within a visit given some approaches in the literature (e.g., Med-BERT) use code ordering while others (e.g., BEHRT) do not. The authors report similar performance is observed when varying the order of diagnosis, procedures and laboratory codes within a visit, suggesting low-level code order is not necessarily important for learning medical relations. This is important for claims and much EHR data which does not have information on code ordering.

Wang et al. [236] developed a pretrained model that was evaluated specifically for improving the diagnosis of suspected lung cancer. The authors used the Lite BERT model ALBERT [112] to pretrain their model using MLM on longitudinal patient histories called MedAlbert. A logistic regression classifier head is added on top of MedAlbert to make

predictions for lung cancer diagnosis, which outpeformed a logistic regression classifier without use of the pretrained model. This demonstrated that considering the temporal relations within cancer pathways for diagnosis provides higher performing models. While this supports the use of transformers for cancer pathways modelling, the authors only compare MedAlbert to a logistic regression model, and focus solely on EHR data. Similarity, He and colleagues [81] explored downstream cancer predictions from pretrained models. However, this study leveraged the existing pretrained Med-BERT [188] model and focused on ways to design the fine-tuning task to improve the prediction of pancreatic cancer. This work, therefore, didn't explore opportunities for performing the model pretraining, only fine-tuning. Furthermore, it also focused on EHR data.

2.5.5 Summary

Overall, existing research has shown that using transfer learning and pretrained models offers benefits for predicting health-related outcomes and model convergence when data is limited. However, there have been numerous approaches implemented for pretraining and fine-tuning. Many studies use differing pretraining strategies (e.g., what BERT special tokens or diagnosis code classifications they use), that typically leverage limited pretraining tasks, leading to debate in how best to pretrain and evaluate generalised models. In addition, some studies do not evaluate the pretrained model alongside an equivalent model trained from scratch, or only evaluate pretrained models on a small number of downstream tasks. This makes it challenging to validate the benefit of pretraining or the broader generalisability of models. Finally, there remains less of a focus on claims data compared to EHR data, particularly for cancer, highlighting a gap in pretrained models for this context. Further research is needed to determine if pretraining can be used successfully with claims data to improve predictions for cancer patients in low resource scenarios.

2.6 Conclusion

Existing research suggests there is growing interest in applying transformers to longitudinal health data to improve patient predictions. In particular, use of MTL, longitudinal patient modelling, and transfer learning, have been identified as promising areas for trans-

formers to improve healthcare predictions. However, as identified in the literature, there are still a number of limitations with regard to transformers and longitudinal health data for cancer pathways modelling.

First, MTL has demonstrated benefits with longitudinal health data for patient predictions, however existing results show inconsistency and lack exploration for cancer in longitudinal health data. Second, transformers are being increasingly used for longitudinal patient modelling. However, variability in the patient features and modelling approaches has resulted in lack of agreement and clarity in how best to encode or represent this data. Furthermore, majority of existing work (including for cancer pathways) has been developed for EHR data, leaving claims data understudied. Third, existing research has highlighted that pretrained transformers and transfer learning can be beneficial for health predictions. However, there is wide discrepancy in approaches used and there remains a deficit of applications on claims data and for cancer. These gaps indicate a need for further approaches in these areas using transformers with longitudinal health data in the cancer context.

As described in the following chapters, this research aims to address the aforementioned limitations by: implementing and evaluating MTL with transformers in longitudinal health data to predict multiple cancer outcomes; developing a hierarchical transformer tailored to modelling longitudinal claims data that integrates dynamic and static features for cancer-related predictions; and providing a comprehensive investigation into pretraining strategies for claims data and applying the model to fine-tune chronic disease predictions for cancer patients.

Chapter 3

Multi-Outcome Prediction for Cancer Patients with Transformer-Based Multi-Task Learning

This chapter focuses on the problem of cancer being a multi-outcome disease, with many outcomes requiring consideration for optimal cancer care. It investigates the potential to extend transformers with multi-task learning to model multiple related outcomes in longitudinal health data.

Publication Declaration

This chapter is based on the following published paper: *Gerrard, L., Peng, X., Clarke, A., Schlegel, C., Jiang, J.: Predicting Outcomes for Cancer Patients with Transformer-Based Multi-task Learning. In: Australasian Joint Conference on Artificial Intelligence. pp. 381-392. Springer (2022).* First published in vol 13151, pages 381 - 392, 2022, by Springer Nature. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-97546-3_31. Reproduced with permission from Springer Nature. This paper was published as part of the Australasian Joint Conference on Artificial Intelligence (AJCAI) 2021. Where appropriate, the paper has been amended to ensure consistency in structure, removal of redundant information, and alignment with thesis formatting requirements.

Author Contributions

LG and XP developed the research idea and designed the experiments. LG defined the scope of the study, performed exploratory data analysis and preprocessed the data. LG implemented and evaluated the core model and performed the experiments. LG analysed and interpreted the results. LG wrote the initial version of the paper, and revised and edited the paper in response to co-author and reviewer feedback. XP provided technical guidance and contributed to code development and review. XP assisted with experimental

setup for baseline comparisons and with data visualisation. XP reviewed the paper for accuracy and quality and provided critical input. AC, CS, and JJ reviewed the paper and provided feedback.

Author Declaration

I confirm that the description above accurately reflects my contribution as an author to the publication listed.

Signatures

Leah Gerrard (LG): Production Note: Signature removed prior to publication.

Dr Xueping Peng (XP): Production Note: Signature removed prior to publication.

Dr Allison Clarke (AC): Production Note: Signature removed prior to publication.

Clement Schlegel (CS): Production Note: Signature removed prior to publication.

A/Prof Jing Jiang (JJ): Production Note: Signature removed prior to publication.

3.1 Introduction

Predicting outcomes for cancer patients is an ongoing challenge due to the complexity of individuals with cancer and growing treatment options [75]. Cancer patients often experience frequent hospital admissions due to cancer symptoms and treatment, resulting in possible delays in therapy, reduced quality of life and increased financial burden [62, 124]. Recent use of EHR data has facilitated the development of models to predict hospital readmission (usually within 30 days) and future diagnosis for admitted patients [128, 146]. These outcomes are of particular importance to cancer patients who have shown an increased risk of hospital readmission [92], and are often burdened with a range of symptoms and side effects which can require hospitalisation [145]. Being able to accurately predict these outcomes for cancer patients provides opportunity for personalised treatment and care planning to improve patient outcomes.

Although EHR data has shown benefits in predicting outcomes for cancer patients, current models largely use statistical and shallow machine learning approaches [165, 69], despite the recent success of deep learning. Deep learning models have been widely used

in cancer prediction and prognosis with image data [68, 114], and have also been used with EHR data for a range of non-cancer health applications [128, 41, 168, 169]. Given the longitudinal nature of EHR data, many studies have leveraged RNN approaches and their variants, which are capable of handling sequential relationships. However, the emergence of transformer models has provided new state-of-the-art performance in language and health-related tasks [52, 186]. While there are some examples of these models using EHR data, this research is still developing, and is particularly limited for cancer use cases.

A further limitation of existing cancer-related EHR studies, is that they are primarily designed for STL, where only a single outcome is predicted (e.g. readmission or future diagnosis). Growing evidence indicates that MTL, where multiple related outcomes are predicted simultaneously, can outperform STL in prediction tasks [68, 79]. Since there are often common diagnoses for hospital readmissions [86], and up to 30% of patients are readmitted for the same diagnosis as their prior admission [27], there may be benefits to using MTL to jointly predict these patient outcomes.

To overcome the above limitations, we propose a **Trans**former-based **M**ulti-**T**ask model (called TransMT), to predict two outcomes for hospitalised patients with cancer; future diagnosis and hospital readmission. This model uses EHR data to capture inherent diagnosis and sequential visit dependencies, and applies the two predictive tasks to simultaneously learn common low-level representations and task-specific knowledge. The main contributions of this work are:

- An end-to-end multi-task transformer model that outperforms single-task and RNN baselines in predicting future diagnosis and hospital readmission.
- An experimental study conducted on two public health datasets applying the proposed model to predict outcomes for cancer patients.

This is, to the best of our knowledge, the first application of this type of model to a cancer-related prediction problem.

The remainder of this chapter is organised as follows: Section 3.2 briefly reviews the related work on transformers and multi-task learning. Section 3.3 describes the proposed model. Section 3.4 presents the experiments and results from two cancer cohorts,

and Section 3.5 concludes the chapter by summarising the research and presenting future directions.

3.2 Related Work

Transformers for EHR

The transformer models most closely related to this study include BEHRT [119] and Med-BERT [188]. These models are both BERT-based approaches that use EHR data to develop generalised pretrained models that can be further fine-tuned for single prediction tasks. BEHRT uses a specialised code mapping (Caliber[1] codes) to group diagnosis codes, however this approach is not well-utilised in EHR deep learning applications. BEHRT also utilises special tokens from language modelling that may lead to information loss in EHR data [188]. While Med-BERT does not use these language tokens, it does include input from multiple diagnosis classifications (ICD-9 and ICD-10), and also uses a potentially unreliable diagnosis ranking strategy. Both of these aspects may have implications for learned diagnosis relationships in EHR data.

To address these issues, the model proposed in this study adopts a common diagnosis grouping approach (see Section 3.4); maps all diagnoses to a single classification; and does not use diagnosis priority rankings. Furthermore, the purpose of the proposed approach differs to that of the above existing models, as it extends transformers to enable multi-task prediction, as opposed to developing a pretrained model. Finally, our proposed model is also related to the work of Lahlou et al. [109], who developed a transformer model for predicting 30-day hospital readmissions using Medicare claims data. However, this paper is focused only on readmission prediction and does not explore MTL.

Multi-Task Learning

MTL aims to improve predictive performance by jointly training multiple related predictive tasks. It has several advantages over STL including reduced overfitting, improved generalisation and increased training sample size [198]. This makes MTL particularly relevant for health data which typically has higher dimensions and noise, but smaller size,

[1] https://www.hdruk.ac.uk/case-studies/caliber/

than other data types [139]. The last few years have seen several studies exploring MTL for cancer applications with image data, such as for breast and lung cancer [68, 114]. MTL has also been used in general EHR studies and has shown benefits over STL for predicting mortality and diagnosis [79, 121]. None of these studies however, included both hospital readmission and future diagnosis as prediction tasks in MTL.

Of the EHR-related MTL research, McDermott et al. [139] compared the performance of RNNs and transformers. Interestingly, this paper found that an RNN-based model outperformed a transformer, and MTL resulted in poorer predictions than STL. However, it is difficult to compare the two models, as they provided a reduced number of samples to the transformer and did not use position encoding. They also included a large number of predictive tasks, some of which may not have been related, which may explain the MTL results. Nevertheless, this work indicates a need for further exploration of MTL with transformers for healthcare predictive tasks.

3.3 Methodology

This section describes the methodology of the proposed model. It starts with notations of important concepts, followed by an overview of the model. Then, the individual components of the model are explained.

3.3.1 Notations

We denote the set of diagnosis codes from the EHR data as $c_1, c_2, \ldots, c_{|\mathbb{C}|} \in \mathbb{C}$ and $|\mathbb{C}|$ is the number of unique diagnosis codes. A patient's EHR record can be represented by a sequence of visits $\boldsymbol{P} = \langle V_1, \ldots, V_t, \ldots, V_T \rangle$, where T is the visit number in the patient record. Each visit V_t consists of a subset of diagnosis codes ($V_t \subseteq \mathbb{C}$). For demonstration purposes, all algorithms are presented for a single patient's record. Table 3.1. summarises the notations used throughout the chapter.

3.3.2 Model Overview

As illustrated in Fig 3.1., TransMT is trained in an end-to-end fashion. First, an embedding layer encodes categorical diagnosis codes to dense numerical vectors. Then, an attention pooling layer compresses a set of diagnosis code embeddings from the visit into

Table 3.1 : Notations for TransMT.

Notation	Description
$\mathbb{C}$	Set of unique diagnosis codes in the dataset
$\|\mathbb{C}\|$	The number of unique diagnosis codes
c_i	$c_i \in \mathbb{C}$, the i-th diagnosis code in $\mathbb{C}, i = 1, \ldots, \|\mathbb{C}\|$
V^t	The t-th visit of the patient, $V^t \subseteq \mathbb{C}$
$\boldsymbol{P}$	The patient record, $\boldsymbol{P} = \langle V^1, \ldots, V^t, \ldots, V^T \rangle$
$\boldsymbol{E}_{i,:}$	Basic embedding vector of diagnosis code c_i
d	The dimension of the diagnosis code embedding

a single context-aware vector representation. Next, the position embeddings are added to the learned visit vectors, and normalized outputs are fed into the prediction tasks. The structure of the two prediction tasks are identical, using a Transformer to learn the visit relationships in the patient record. Lastly, a predictive model is used to predict the outcomes in the final hospital visit simultaneously.

3.3.3 Common Representations

Embedding Layer.

Visit $\boldsymbol{V}^t$ is denoted by a set of diagnosis codes $\boldsymbol{X}^t = [\boldsymbol{x}_{t1}, \boldsymbol{x}_{t2}, ..., \boldsymbol{x}_{tn}]$, where n is the number of diagnosis codes in the visit, and $\boldsymbol{x}_{ti}$ could be a one-hot vector whose dimension length equals the number of unique diagnosis codes $|\mathbb{C}|$. An embedding layer is applied to $\boldsymbol{X}^t$ and transforms all discrete diagnosis codes to a set of low-dimensional dense vector representations $\boldsymbol{E}^t = [\boldsymbol{e}_{t1}, \boldsymbol{e}_{t2}, ..., \boldsymbol{e}_{tn}]$ with $\boldsymbol{e}_{ti} \in \mathbb{R}^d$. This process can be formally written as $\boldsymbol{E}^t = \boldsymbol{W}^{(e)}\boldsymbol{X}^t$, where the diagnosis code embedding weight matrix $\boldsymbol{W}^{(e)} \in \mathbb{R}^{|\mathbb{C}| \times d}$ can be fine-tuned during model training.

Attention Pooling.

Attention Pooling [120, 28] explores the importance of each diagnosis within a visit and compresses a set of diagnosis code embeddings into a single context-aware vector representation. For simplicity, we take the embedding output $\boldsymbol{E}^t$ of the t-th visit $\boldsymbol{V}^t$ as an

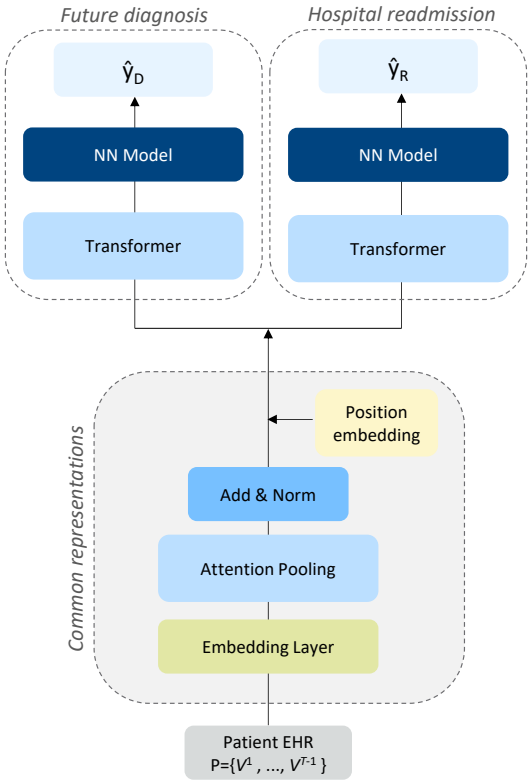

Figure 3.1 : The proposed TransMT model.

example. Formally, it is written as:

$$g(\boldsymbol{E}^t_{i,:}) = \boldsymbol{w}^T \sigma(\boldsymbol{W}^{(1)} \boldsymbol{E}^t_{i,:} + b^{(1)}) + b, \tag{3.1}$$

where $\boldsymbol{E}^t_{i,:}$ is the i-th row of $\boldsymbol{E}^t$ ($1 \leq i \leq n$), σ is ReLU function and $\boldsymbol{w}, \boldsymbol{W}^{(1)}$, $\boldsymbol{b}^{(1)}, \boldsymbol{b}$ are learnable parameters. The probability distribution is formalized as

$$\boldsymbol{\alpha}_t = \texttt{softmax}([g(\boldsymbol{E}^t_{i,:})]^n_{i=1}). \tag{3.2}$$

The final output $\boldsymbol{v}_t$ of the attention pooling is the weighted average of sampling a code according to its importance, i.e.,

$$\boldsymbol{v}_t = \sum_{i=1}^{n} \boldsymbol{\alpha}_t \odot [\boldsymbol{E}^t_{i,:}]^n_{i=1}, \tag{3.3}$$

where $\boldsymbol{v}_t \in \mathbb{R}^d$ $(1 \leq t \leq (T-1))$ represents the t-th visit in the patient record. The patient record in low-dimensional dense vector representations is denoted as $\boldsymbol{J} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t, \ldots, \boldsymbol{v}_{(T-1)}]$.

Position Embedding and Normalization.

To incorporate information on the order of visits in a patient record, we embed each temporal position t to obtain a position embedding matrix $\boldsymbol{P}_e \in \mathbb{R}^{(T-1)\times d}$, then fuse the embeddings into a vector representation. The structure of the position embedding is identical to the embedding layer. The output of the Add&Norm layer is denoted as follows,

$$\boldsymbol{O} = \texttt{LayerNorm}(\boldsymbol{J} + \boldsymbol{P}_e). \tag{3.4}$$

3.3.4 Transformer Model

The Transformer excels in learning relationships between elements in a sequence. We leverage this by using the `Transformer` module to capture the inherent dependencies between patients' sequential visits, which is calculated as follows:

$$\{\boldsymbol{V}^1, \ldots, \boldsymbol{V}^{T-1}\} = \texttt{Transformer}(\{\boldsymbol{O}^1, \ldots, \boldsymbol{O}^{T-1}\}) \tag{3.5}$$

Given the success of BERT-based approaches, we implement the `Transformer` identical to BERT [232] and [52], which has two sub-layers. The first is a multi-head attention mechanism (explained below), and the second is a position wise fully connected feed-forward network. A residual connection is employed around each of the two sub-layers, followed by layer normalization. While BERT uses Masked Language Modelling for pretraining, we focus on the objective of multi-task learning and hence the model is trained directly on the two prediction tasks.

Multi-head Attention.

The multi-head attention mechanism relies on self-attention, where all of the keys, values and queries come from the same place. The self-attention operates on a query $\boldsymbol{Q}$,

a key $\boldsymbol{K}$ and a value $\boldsymbol{V}$:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V} \tag{3.6}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are $n \times d$ matrices, n denotes the number of diagnoses in a visit in a patient record, d denotes the embedding dimension.

The multi-head attention mechanism obtains h (i.e. one per head) different representations of $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$, computes self-attention for each representation, and concatenates the results. This can be expressed as follow:

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V) \tag{3.7}$$

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_1, ..., head_h)\boldsymbol{W}^O \tag{3.8}$$

where the projections are parameter matrices $\boldsymbol{W}_i^Q \in \mathbb{R}^{d \times d_k}$, $\boldsymbol{W}_i^K \in \mathbb{R}^{d \times d_k}$, $\boldsymbol{W}_i^V \in \mathbb{R}^{d \times d_v}$ and $\boldsymbol{W}^O \in \mathbb{R}^{hd_v \times d}$, $d_k = d_v = d/h$.

3.3.5 Prediction via Multi-task Learning

MTL is employed to jointly learn two prediction tasks for the final patient visit. Given a patient's record $\boldsymbol{P} = \{V^1, V^2, \ldots, V^{T-1}\}$, we simultaneously predict future diagnosis and hospital readmission.

Future Diagnosis Prediction.

This is a multi-class classification problem with the objective being to predict the principal diagnosis code (the diagnosis code representing the main reason for hospitalisation) in the last visit V_T, which can be expressed as follows:

$$\hat{\boldsymbol{y}}_D = \texttt{Softmax}(\boldsymbol{W}^D \boldsymbol{p}_D + \boldsymbol{b}^D), \tag{3.9}$$

$$\mathcal{L}_{\mathcal{D}} = \frac{-1}{T-1} \sum_{t=1}^{T-1} \left(\boldsymbol{y}_D{}^{\mathrm{T}} \log \hat{\boldsymbol{y}}_D + (1 - \boldsymbol{y}_D)^{\mathrm{T}} \log (1 - \hat{\boldsymbol{y}}_D) \right), \tag{3.10}$$

where $\boldsymbol{p}^D \in \mathbb{R}^d$ is the output of the attention pooling layer followed by `Transformer` to denote the representation of the patient record in the future diagnosis prediction task,

and $\boldsymbol{W}^D \in \mathbb{R}^{|\mathcal{C}| \times d}$ and $\boldsymbol{b}^D \in \mathbb{R}^{|\mathcal{C}|}$ are learnable parameters.

Hospital Readmission Prediction.

This is a binary classification problem to predict whether the patient's final visit is within 30 days of their previous hospital admission. The implementation is identical to the future diagnosis prediction task, which can be expressed as follows:

$$\hat{\boldsymbol{y}}_R = \texttt{Softmax}(\boldsymbol{W}^R \boldsymbol{p}_R + \boldsymbol{b}^R), \tag{3.11}$$

$$\mathcal{L}_{\mathcal{R}} = \frac{-1}{T-1} \sum_{t=1}^{T-1} \left(\boldsymbol{y}_R{}^{\mathrm{T}} \log \hat{\boldsymbol{y}}_R + (1 - \boldsymbol{y}_R)^{\mathrm{T}} \log \left(1 - \hat{\boldsymbol{y}}_R\right) \right), \tag{3.12}$$

where $\boldsymbol{v}_R \in \mathbb{R}^d$ is the output of attention pooling layer followed by `Transformer` to denote the representation of the patient record in the hospital readmission prediction task, and $\boldsymbol{W}^R \in \mathbb{R}^{|\mathcal{C}| \times d}$ and $\boldsymbol{b}^R \in \mathbb{R}^{|\mathcal{C}|}$ are the learnable parameters.

Objective Function.

To allow shared information between the two predictive tasks, we jointly train the tasks using the weighted loss function below, where λ is a learnable parameter.

$$\mathcal{L} = \lambda \cdot \mathcal{L}_D + (1 - \lambda) \cdot \mathcal{L}_R. \tag{3.13}$$

3.4 Experiments

3.4.1 Datasets

We conduct experiments on two public real-world health datasets, the MIMIC-III (v1.4) [96] and MIMIC-IV (v0.4) [94] datasets. Given these datasets are widely used for deep learning EHR studies, we do not explain them in detail and instead refer readers to the official sources for comprehensive information on these datasets [96, 94].

Table 3.2 : Statistics of the MIMIC datasets for the two cancer cohorts.

Dataset	**MIMIC-III**	**MIMIC-IV**
# of patients	2,070	20,953
# of visits	5,552	79,177
Avg. # of visits per patient	2.68	3.78
# of unique ICD-9 codes	3,228	6,939
Avg. # of ICD-9 codes per visit	13.67	13.47
Max # of ICD-9 codes per visit	39	57
# of CCS categories	166	270

3.4.2 Data Preprocessing

Cancer Cohort Selection.

Patients were included in analysis if they had at least one cancer-related ICD-9 diagnosis code (140-239) (as used elsewhere [197]). Patients with less than 2 visits were excluded and only the most recent 21 visits were used where applicable. All diagnosis codes were mapped to ICD-9 codes, as this is the common classification among MIMIC-III and IV [96, 95]. A summary of the statistical information for the final cancer cohorts is included in Table 3.2.

Clinical Classifications Software (CCS) Mapping.

To reduce the number of diagnosis code categories (and hence dimensions) for model training, we used the CCS categorisation scheme[2] to group diagnosis codes into 285 mutually exclusive categories and use these categories for the future diagnosis prediction task. This scheme has been used elsewhere for similar applications [168, 169].

3.4.3 Experimental Setup

Baseline Approaches.

We compare the performance of our proposed model against three baselines: two common RNN-based approaches (RETAIN and Dipole) and a single-task transformer model.

[2] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf

RETAIN [41] uses a reversed RNN with attention to capture relationships between diagnoses to perform heart failure prediction. **Dipole** [128] uses a bidirectional RNN and three attention mechanisms (location-based, general, concatenation-based) to predict future diagnosis information. We used the location-based Dipole as the baseline method. **STL model** is the proposed transformer model trained separately for single-task prediction.

Evaluation Metrics.

F1 score and AUC score are used as evaluation measures. **F1 score** is the harmonic mean of precision and recall. An F1 score close to 1 indicates high precision and recall. **AUC score** uses the Receiver Operating Characteristic (ROC) curve, which shows the relationship between true and false positive rates. The AUC demonstrates how well a model can distinguish between binary classes. An AUC score of 0.5 indicates a model no better than chance, with 1 indicating a perfect model.

Implementation Details.

We implement all models with Pytorch 1.4.0 and run models on NVIDIA TITAN X with 12GB RAM. For training, we use Adadelta [261], one of the most common optimisers for developing prediction models with EHR data [30], including for training RETAIN and Dipole models [41, 128]. Reflecting the differences in dataset and cohort size, and to promote generalisation [100], we use a minibatch of 16 and 32 patients on MIMIC-III and MIMIC-IV, respectively. These batch sizes are also in line with similar work on multi-task learning with MIMIC-III data and transformers on EHR [79, 116].

We randomly split data into training, validation and test sets in ratios of 80%, 10% and 10%, respectively. The validation set is used to determine the best parameter values for training , where MIMIC-III is trained for 20 iterations and MIMIC-IV for 10 iterations, to balance computational efficiency with performance. Drop-out strategies are used for all approaches, adopting the commonly used drop-out rate for transformers of 0.1 [232, 188]. We set the dimension $d = 200$, as moderate embedding dimensions are sufficient to capture complex temporal dependencies [41, 119, 118, 218], and scale the transformer architecture relative to the data complexity, using two transformer layers for the proposed model and STL baseline for MIMIC-III, and three transformer layers for MIMIC-IV. All

models were run three times to calculate the mean and standard deviation of performance metrics.

3.4.4 Results

F1 Score for Two Prediction Tasks.

Table 3.3. shows the mean F1 scores and standard deviations of the TransMT model compared with baselines for the predictive tasks in the two MIMIC datasets. The results show that the proposed TransMT model outperforms all baselines on both MIMIC-III and MIMIC-IV datasets. This demonstrates the benefits of MTL to jointly learn the future diagnosis and hospital readmission prediction tasks. The results also show that the single-task baseline (STL) outperformed the two RNN attention-based models, RETAIN and Dipole, on both data sets, indicating the benefit of Transformers in modelling the sequential relationships in EHR data.

Table 3.3 : Performance comparison of prediction tasks.

Dataset	**Model**	**F1 Score (%)**	
		Future diagnosis	**Hospital readmission**
MIMIC-III	RETAIN	12.48 $\pm$ 2.35	9.65 $\pm$ 5.67
	Dipole	10.39 $\pm$ 0.49	19.14 $\pm$ 3.85
	STL	16.19 $\pm$ 1.05	24.27 $\pm$ 4.46
	TransMT	**16.56** $\pm$ 1.34	**24.70** $\pm$ 3.95
MIMIC-IV	RETAIN	7.60 $\pm$ 0.54	35.12 $\pm$ 1.64
	Dipole	6.99 $\pm$ 0.23	38.64 $\pm$ 6.03
	STL	9.04 $\pm$ 0.19	38.97 $\pm$ 2.17
	TransMT	**9.19** $\pm$ 0.11	**41.62** $\pm$ 2.26

AUC Score for Hospital Readmission Prediction.

Fig 3.2. depicts the AUC curve for all models on the hospital readmission task, showing that TransMT outperforms baseline models in both MIMIC datasets. In contrast to MIMIC-III data where all models show similar performance, the TransMT model outperforms the best RNN baseline by almost 15% on MIMIC-IV data. This indicates the benefit of TransMT for hospital readmission prediction, particularly with the larger and more complex MIMIC-IV data.

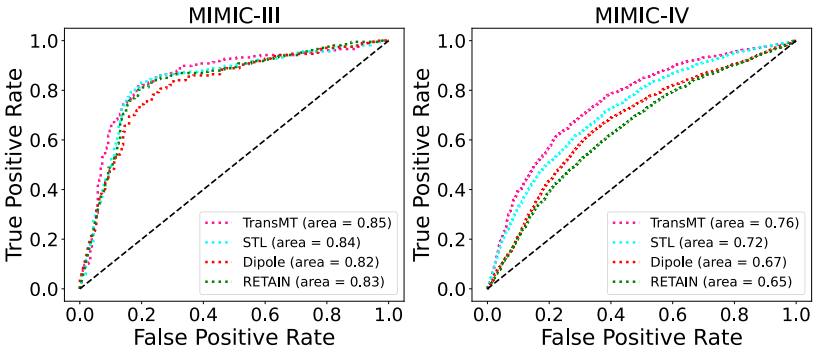

Figure 3.2 : AUC of hospital readmission on two datasets.

Relationship between Diagnosis and Readmission.

We also explored the relationship between future diagnosis and hospital readmission. Table 3.4 shows the proportion of patients with the same principal diagnosis and CCS category in their final visit compared with their previous visit, by hospital readmission status. Patients who were readmitted within 30 days were more likely to have the same diagnosis and CCS category in their final visit than those not readmitted within 30 days. This suggests a relationship between the two prediction tasks, and supports the results that the multi-task model produces better predictions.

Table 3.4 : Relationship between future diagnosis and readmission. Same diagnosis and CCS category refers to the proportion the same as the patient's previous visit.

Dataset	**Readmitted**	**Same Diagnosis (%)**	**Same CCS (%)**
MIMIC-III	Yes	17.76	25.68
	No	8.28	15.14
MIMIC-IV	Yes	3.94	6.71
	No	2.97	5.44

3.5 Conclusion

In this chapter, we proposed the model TransMT, which captures sequential relationships between patient visits to predict future diagnosis and hospital readmission for cancer patients. As demonstrated by the experimental results, TransMT produces bet-

ter predictions than single-task and RNN-based approaches, indicating the potential for transformer-based prediction models with EHR data to facilitate cancer-related research. Given the MIMIC datasets only capture a subset of patients in EHRs, a future direction of this work is to apply this model to other cancer datasets where patients with specific cancers and treatments can be identified. Furthermore, this work only includes diagnosis codes as input to the model, and further work could include addition of other patient information (e.g. medications and demographics) and use of more recent diagnostic classifications such as ICD-10. Finally, experimentation of MTL with other methods, such as pretraining approaches, will help identify where MTL can be most useful in healthcare applications to further improve the performance of risk prediction models.

Chapter 4

Modelling Longitudinal Cancer Pathways in Claims Data with Hierarchical Transformer

This chapter focuses on the problem of effectively modelling longitudinal cancer pathways within claims data using transformers. It investigates how to model the short- and long-term relations within claims data while capturing the dynamic (temporal) and static (demographic and cancer characteristics) patient features.

Publication Declaration

This chapter is based on the following published paper: *Gerrard, L., Peng, X., Clarke, A., Long, G.: Multi-level Transformer for Cancer Outcome Prediction in Large-Scale Claims Data. In: International Conference on Advanced Data Mining and Applications. pp. 63-78. Springer (2023).* First published in vol 14178, pages 63 - 78, 2023, by Springer Nature. The final authenticated version is available online at: https://doi.org/10.1007/978-3-031-46671-7_5. Reproduced with permission from Springer Nature. This paper was published as part of the International Conference on Advanced Data Mining and Applications (ADMA) 2023. Where appropriate, the paper has been amended to ensure consistency in structure, removal of redundant information, and alignment with thesis formatting requirements.

Author Contributions

LG developed the research idea and designed the experimental study. LG collected and managed the datasets used, carried out the data preprocessing, and implemented the models, including training and evaluation. LG analysed and interpreted the results. LG wrote the initial version of the paper, and revised and edited the paper in response to co-author and reviewer feedback. XP provided advice for experimental and model architecture design. XP contributed to code development and refinement. AC provided advice on the

research direction, co-ordinated expert advice for data understanding, and suggested improvements to the paper. GL provided project oversight and contributed to feedback and refinements on the paper.

Author Declaration

I confirm that the description above accurately reflects my contribution as an author to the publication listed.

Signatures

Leah Gerrard (LG): Production Note: Signature removed prior to publication.

Dr Xueping Peng (XP): Production Note: Signature removed prior to publication.

Dr Allison Clarke (AC): Production Note: Signature removed prior to publication.

A/Prof Guodong Long (GL): Production Note: Signature removed prior to publication.

4.1 Introduction

Anticipating likely outcomes for patients initiating chemotherapy is crucial for providing optimal cancer care. Chemotherapy is a frequently used cancer treatment and can offer benefits such as lowering risk of cancer recurrence and improving survival outcomes [58]. However, chemotherapy also contributes to the range of symptoms and side effects patients often experience during treatment, such as complications and adverse events. Two important considerations for cancer patients are likely survival outcomes and cardiovascular (heart) disease risk. Existing research has indicated there are few tools for mortality risk prior to chemotherapy and accurate predictions could be useful to clinicians and patients to inform discussions and decisions [58]. In addition, evidence suggests that patients may be at increased risk of cardiovascular diseases following cancer treatment, and models to predict this risk are important for supporting treatment plans and preventive interventions [9]. Prediction models also offer the potential to inform cancer-related health policy through patient risk stratification.

Recent application of transformers on longitudinal health data have focused on BERT [52], which has demonstrated capability in modelling temporal relations in patient med-

ical codes (e.g., diagnoses, medications and procedures) and has outperformed other sequential methods such as RNNs [118, 119, 188]. However, existing research has primarily focused on the use of EHR data, with only a limited number of applications to claims data [263, 262]. This is critical due to the different temporal structure of claims data, which does not naturally fall into a pattern of visits like EHR. Furthermore, there are still open questions on how to best encode the structure of medical data and collectively model it in BERT-based approaches [188, 262].

In addition to the above, there is a growing trend of leveraging patient information outside of medical codes for developing health-specific transformer-based models. This has largely included patient demographic information, such as age and gender, and some clinical features (e.g., lab tests) [83, 142, 116, 66]. While there is evidence to suggest inclusion of additional patient features by multimodal learning offers improvements for model predictions, many existing approaches do not evaluate the benefit of including such information, and often only incorporate a small number of features. Further exploration and evaluation of patient features with transformer-based models are required.

To address these challenges, we propose a hierarchical **M**ulti-**L**evel **T**ransformer that leverages **Claims** data (**Claims-MLT**) and information from a cancer registry to predict outcomes for breast and colorectal cancer patients. We evaluate our approach on a real-world cancer dataset from Australia and demonstrate the benefit of our approach compared to baselines. To summarise, the main contributions of this work are:

- An end-to-end multi-level transformer that is tailored to claims data and predicts survival and heart disease diagnosis for cancer patients at the point of chemotherapy initiation.
- We propose the use of a dual feature encoder block, where each block consists of a transformer encoder and attention pooling, to learn hierarchical context-aware vector representations for low-level claims item relationships within a month and patient-level relationships from sequential claims patterns.
- Fusion of patient features from a cancer registry with the patients' claim representation, to capture the important static demographic and clinical information at cancer

diagnosis.

- An experimental study on real-world cancer data, demonstrating Claims-MLT outperforms all comparative methods.

This is, to the best of our knowledge, the first transformer-based model to predict cancer outcomes using temporal claims and static patient demographic and clinical data. The remainder of this chapter is organised as follows: Section 4.2 briefly reviews the related work on transformers for EHR and claims data. Section 4.3 presents the model framework and approach. Section 4.4 describes the data and experimental results. Section 4.5 concludes the chapter.

4.2 Related Work

Transformer-based Models for EHR and Claims Data

Compared to the vast development of successful BERT-based approaches on EHR data [119, 188, 163, 199, 116], there has been substantially less application to claims data. Of the claims-based transformers, the most relevant to this study are Claim-PT [263] and TMAE [262]. Developed by the same authors, these approaches obtain medical visit representations using a max-pooling layer to capture the most important features. The pooled outputs are then fed into a Transformer encoder to provide a patient representation, which is then used for prediction tasks, such as the survival and asthma exacerbation predictions with Claim-PT.

While the above applications to claims data demonstrate the superiority of the transformer compared to RNN models, we note a number of existing limitations and challenges. First, as indicated earlier, approaches need to consider the differing temporal structure of claims data compared to EHR, and it is unclear how to develop an optimal data representation for claims data. Claim-PT and TMAE define medical visits as a claim type (inpatient, outpatient, pharmacy) comprised of medical codes, and appear to aggregate medical visits by service date. However, this separates healthcare services that may occur closely in time. Approaches using time windows to split patient sequences are common for deep learning in healthcare [151], and use of longer time periods (i.e., month)

has demonstrated benefits in predicting patient outcomes [146]. Second, attention pooling may be a more appropriate pooling approach for maintaining important features than max-pooling [59]. Third, like other EHR models (i.e., BEHRT, Med-BERT), Claim-PT and TMAE employ only a single Transformer encoder to learn the relationships between medical information.

The hierarchical BEHRT model Hi-BEHRT [118] has demonstrated benefits in using a two-level transformer to capture associations in longer patient EHR sequences. Our approach leverages the ideas of attention pooling and hierarchical transformers to enhance claims data modelling. We also represent a patient's claim history as a sequence of months to model clinical events in close proximity while maintaining temporal information. This can be thought of as providing both a micro (visit level) and macro (patient level) perspective of patient pathways.

Transformers with Additional Patient Information

This work also relates to transformer models that have integrated additional patient features such as demographics into predictions. [83] evaluated several data types for predicting recurrence of colorectal cancer, including tabular (demographics, tumour characteristics, and treatment parameters) and time-series (test results) data. They found the use of multiple data types offered improved model performance, however, did not include medical codes as features. [142] proposed a transformer model to predict depression for patients with breast cancer. This model included medical codes, patient age and gender, and clinical notes, however, only clinical notes were evaluated in terms of impact on predictions. [116] included age, gender, medications, clinical measures, and location information in their ExMed-BERT model.

Other work has also explored inclusion of demographics such as age and gender, patient geographical information, and/or temporal observations (lab tests, vitals), in addition to medical codes [66, 242]. Much of this work, however, fails to evaluate the impact of the additional patient features, making their benefit to model performance unclear. In this work, we provide an ablation study to determine the effect of patient demographic and clinical information in predicting cancer outcomes.

4.3 Methodology

This section describes the methodology of the proposed approach. It firstly provides an overview of the model and then details the individual model components.

4.3.1 Model Overview

An overview of the proposed Claims-MLT model is shown in Fig 4.1, which is trained end-to-end and can be viewed as four parts. In the first part, the *claims input embedding*, an embedding layer is used to encode all claims items to dense numerical vectors. Next, sequences of claims items within a month travel through a *feature encoder block*, which contains a transformer encoder (based on BERT) and attention pooling. The output of the transformer encoder is a hidden vector for each claims item in the input sequence, which are then compressed via the attention pooling layer into a fixed context-aware representation for each month. To differentiate between months of claims data for each patient, position embeddings are added to the month representations, and a second feature encoder block is used to learn a fixed-size representation of a patient's claim history. Static patient features are then fused with the patient claims representation to provide key demographic and cancer clinical information, which forms the final *patient representation*. Finally, fully connected and sigmoid layers enable the *cancer outcome prediction* tasks, which are binary classification tasks and trained independently.

4.3.2 Claims Item Embedding

A patient's claims history can be represented as a sequence of months $[M^1, M^2, \ldots, M^T]$ where T is the number of months in the patient's history and each month is temporally related. A month M^t contains a subset of claims codes $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]$, where n is the number of unique claims codes in the dataset, and includes medical services, prescriptions, and diagnoses, and $\boldsymbol{x}_i$ is a one-hot vector of dimension size n. An embedding layer transforms discrete claims in $\boldsymbol{X}$ to low-dimensional dense vector representations $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_n]$. This can be formally written as $\boldsymbol{E} = \boldsymbol{W}^{(e)}\boldsymbol{X}$, where the claim code embedding weight matrix $\boldsymbol{W}^{(e)}$ is fine-tuned during training.

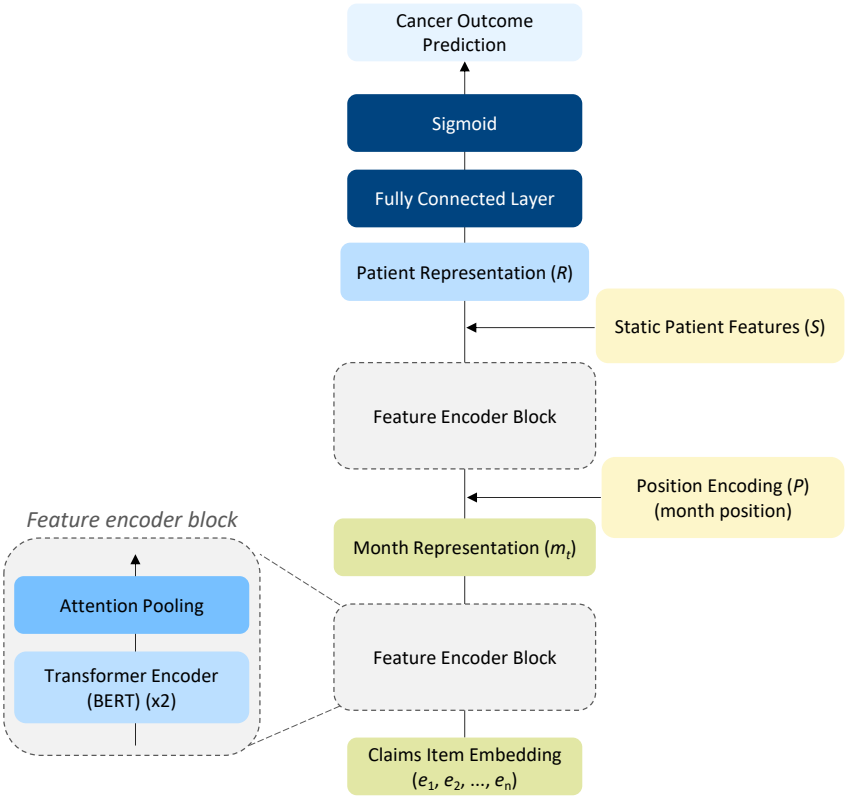

Figure 4.1 : The proposed Claims-MLT model.

4.3.3 Feature Encoder Block

To capture the inherent dependencies between claims items in patient histories, we adopt a multi-level Transformer. This includes dual feature encoder blocks, which each contain a Transformer encoder and attention pooling.

Transformer Encoder.

We leverage the successful BERT [52] model for implementing the Transformer encoder, which contains two sub-layers: a multi-head attention mechanism, which relies on self-attention, and a position-wise fully connected feed-forward network. Residual connections and layer normalisation surround each sub-layer. We refer readers to the original papers on BERT [52] and the Transformer [232] for additional detail.

To capture the sequential context, the Transformer encoder takes as input sequences of items and outputs a hidden vector for each input. In the first feature encoder block, sequences of claims items within a month serve as input. Similar to the original BERT, we use a [CLS] token at the beginning of each sequence and [SEP] tokens after each type of claims data. Not every month contains all types of claims data, hence a sequence can contain one or more claims data types. Due to no strict ordering of diagnosis codes within a hospital visit, and the presence of medical services and prescriptions occurring on the same date, we do not use position embeddings for claims items within a monthly sequence. We represent the output of the Transformer encoder for sequences of claims items as:

$$[\boldsymbol{h_1}, \boldsymbol{h_2}, \ldots, \boldsymbol{h_k}] = \texttt{TransformerEnc}([\boldsymbol{e_1}, \boldsymbol{e_2} \ldots, \boldsymbol{e_k}]) \tag{4.1}$$

where k is the number of claims items in a month, and $\boldsymbol{h}_i (1 \leq i \leq k)$ represents the hidden vector of the i-th claims item in the sequence.

Attention Pooling.

Attention pooling compresses a set of inputs into a fixed vector representation, retaining the most important information during pooling to capture the input context [59]. For claims items within a month sequence, we take the hidden representation output

$\boldsymbol{H}^t = [\boldsymbol{h}_1^t, \boldsymbol{h}_2^t, \ldots, \boldsymbol{h}_k^t]$ of the t-th month as an example:

$$g(\boldsymbol{h}_i^t) = \boldsymbol{w}^T \sigma(\boldsymbol{W}^{(1)} \boldsymbol{h}_i^t + b^{(1)}) + b, \tag{4.2}$$

where $\boldsymbol{h}_i^t$ $(1 \leq i \leq k)$ is the i-th row of $\boldsymbol{H}^t$, σ is ReLU function and $\boldsymbol{w}$, $\boldsymbol{W}^{(1)}$, $\boldsymbol{b}^{(1)}$, $\boldsymbol{b}$ are learnable parameters. The probability distribution can be represented as:

$$\boldsymbol{\alpha}_t = \texttt{softmax}([g(\boldsymbol{h}_i^t)]_{i=1}^k). \tag{4.3}$$

The final output $\boldsymbol{m}_t$ of the attention pooling is the weighted average of sampling a claims item based on its contribution to the input sequence, i.e.,

$$\boldsymbol{m}_t = \sum_{i=1}^{k} \boldsymbol{\alpha}_t \odot [\boldsymbol{h}_i^t]_{i=1}^k, \tag{4.4}$$

where $\boldsymbol{m}_t \in \mathbb{R}^d$ $(1 \leq t \leq T)$ represents the t-th month in the patient claims history, which in the low-dimensional dense vector representations is denoted as $\boldsymbol{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_t, \ldots, \boldsymbol{m}_T]$.

4.3.4 Patient Representation

The patient representation aims to capture the temporal relationships in a patient's claim history and fuse the output with additional features to include demographic and clinical information. This involves the use of position encoding, a second feature encoder bock and fusion of static patient features as described below.

Position Encoding.

Position encoding incorporates information on the order of each month in a patient's claim history. A position embedding matrix $\boldsymbol{P}_e$, which embeds each temporal position, is fused with the patient claims history using an Addition and Normalisation layer. The updated vector representation is denoted as:

$$\boldsymbol{J} = \texttt{LayerNorm}(\boldsymbol{M} + \boldsymbol{P}_e). \tag{4.5}$$

Second Feature Encoder Block.

The updated vector representations containing position information are used as input to the second feature encoder block. This is equivalent to the process described above, in which the Transformer encoder outputs hidden vector representations of the inputs, i.e., $[\boldsymbol{o_1}, \boldsymbol{o_2}, \ldots, \boldsymbol{o_k}] = \texttt{TransformerEnc}([\boldsymbol{j_1}, \boldsymbol{j_2} \ldots, \boldsymbol{j_k}]$, and attention pooling is used to obtain a fixed-length vector representation for the patient's claim history. This can be represented as $\boldsymbol{O} = \texttt{AttentionPooling}([\boldsymbol{o_1}, \boldsymbol{o_2} \ldots, \boldsymbol{o_k}])$, where $\boldsymbol{O}$ signifies the pooled patient claims history representation that captures relationships in cancer and other healthcare activities over time (i.e., between months).

Static Patient Features.

We extract tabular static patient features to include demographics and clinical information in the proposed model. We leverage features primarily from the cancer registry, which includes information relevant to cancer diagnosis, and is therefore important in cancer care and population analysis. To transform the static features into the same dimension as the patient claim representation, we pass the static features through a two-layer neural network, with ReLU activation function after each layer. The output is then fused with the patient's claim history to provide the final patient representation, merging temporal healthcare activities with important static demographic and cancer-related features. The final patient representation used for predictions, where $\boldsymbol{S}$ represents the static patient features, is denoted as:

$$\boldsymbol{R} = \boldsymbol{O} + \boldsymbol{S}. \tag{4.6}$$

4.3.5 Cancer Outcome Prediction

Given a patient's final representation $\boldsymbol{R}$, the multi-level Transformer is trained end-to-end to learn the survival and heart disease prediction tasks independently.

Survival Prediction.

The objective of this task is to predict patient mortality/death within one year (365 days) of chemotherapy initiation, determined using time between death date and chemother-

apy initiation date. If no death date is present, a patient is considered to have survived the study period.

Heart Disease Prediction.

The objective of this task is to predict whether a patient will have a heart disease diagnosis recorded in hospital within one year (365 days) of chemotherapy initiation, determined using time between hospital admission date and chemotherapy initiation date. Heart disease-related admissions were identified using the following ICD-10 codes (I10-I11, I13, I15, I20-I25, I27, I34-I36, I42, I482, I50, U821-823).

Both prediction tasks are binary classification tasks, with average cross-entropy loss computed as follows:

$$\mathcal{L}_{\mathcal{R}} = -\frac{1}{N}\sum_{t=1}^{N}\left(\boldsymbol{y}_R{}^{\mathrm{T}}\log\hat{\boldsymbol{y}}_R + (1-\boldsymbol{y}_R)^{\mathrm{T}}\log\left(1-\hat{\boldsymbol{y}}_R\right)\right), \tag{4.7}$$

where N is the number of data points and $\boldsymbol{y}_R$ is the ground truth of the prediction task (i.e. survival or heart disease diagnosis) and takes the value 0 or 1, and $\hat{\boldsymbol{y}}_R$ is the probability for the N data point.

4.4 Experiments

In this section, we introduce the cancer dataset and describe the data preparation and cancer cohort criteria. We then present the experimental setup and results.

4.4.1 Dataset

We conduct experiments on a real-world cancer dataset that contains data on individuals diagnosed with breast and/or colorectal cancer (ICD-10 codes C50 and C18-21, respectively) in Victoria, Australia, between 2008 and 2019[1]. The data contains linked Commonwealth and Victorian state datasets including the MBS, PBS, Victorian Admitted Episodes Dataset, and Victorian Cancer Registry, which captures patient-level medical

[1] Use of the data has approval from the Australian Institute of Health and Welfare Ethics Committee (EO2015/4/219).

services, medicine prescriptions, hospital diagnoses, and demographics and clinical information, respectively.

4.4.2 Data Preparation

Claims Data.

Claims data from 1 July 2012 onwards was included for analysis[2] and prepared as follows:

- MBS items were mapped to current items using publicly available mapping data[3]. Bulk-billing incentive and modifier items were excluded from analysis. MBS subgroup codes were used to reduce the number of categories. The number of services was derived by summing each patient's service count by subgroup code and date of service. Services with a sum of zero or less were excluded.
- PBS items were mapped to chemical substance Anatomical Therapeutic Chemical (ATC) fifth level (i.e., ATC5) codes, with items unable to be mapped excluded from analysis.
- Diagnosis codes from hospital admissions were refined to three-digit codes to reduce the number of categories. Diagnoses in the same hospital visit (identified using admission date and maximum separation date) were combined. Admissions without corresponding diagnosis information were excluded from analysis.
- MBS subgroup, ATC codes, and diagnoses were aggregated by month and combined to produce the input data. To reduce the length of the input sequences, only the first code was kept for claims items that appeared multiple times in the same month.

---

[2] From this date there was increased capture of medicine dispensing in the PBS through inclusion of under co-payment data and chemotherapy medicines supplied in public and private hospitals as part of the Efficient Funding of Chemotherapy arrangements [140]. For consistency, this date range was also used for MBS and hospital admission data.

[3] http://www.mbsonline.gov.au/

Cancer Registry Data.

A summary of the static patient features included from the cancer registry data can be found in Table 4.1. Data was one-hot encoded prior to use in the analysis pipeline.

Table 4.1 : Patient static features.

Feature	Description
Age	Age group at diagnosis (<40, 10-year age groups to 90+).
Gender	As recorded in the cancer registry at diagnosis.
Tumour grade	Describes grade of tumour and level of differentiation.
Cancer stage	Derived cancer stage (stage I – IV, or missing). Later stage cancers are more advanced.
Region	Victorian region where the patient usually resided at diagnosis.
SEIFA quintile	Quintile of Patient Socio-Economic Indexes for Areas (SEIFA). Index of Relative Socioeconomic Disadvantage (IRSD) based on usual residence at diagnosis.
Screen detected	Whether the cancer was detected through national cancer screening programs. Only partial data for colorectal cancer. Does not include cancers diagnosed in private or other screening. Missing data filled with integer.
Country of birth	Born in Australia, overseas, or unallocated.
Diagnosis year	Calendar year of cancer diagnosis.

4.4.3 Cohort Definitions

Patients with breast and/or colorectal cancer between 1 July 2008 and 31 December 2019 formed the basis of the cancer cohort. Following exclusions[4], chemotherapy initiation date was identified using a patient's first supply of an *antineoplastic agent* (ATC code *L01*). Cohorts were split by cancer types and prediction tasks. Patients with a death date within the study period were excluded from the heart disease diagnosis task but included in survival predictions. Table 4.2 provides the summary statistics of the cancer cohorts for the prediction tasks.

[4] The following exclusions applied: individuals with linkage errors or data quality issues (defined as identifiers linked to multiple dates of birth); records where a cancer diagnosis was reported by death certificate only; secondary cancer diagnoses; and cancer diagnoses prior to 1 July 2012 (for consistency with claims data). Patients who did not initiate chemotherapy or had a death date on or prior to chemotherapy initiation were excluded.

Table 4.2 : Statistics of the cancer data

	Survival		Heart Disease	
Detail	**Breast**	**Colorectal**	**Breast**	**Colorectal**
# of unique patients	15,211	10,746	13,759	8,709
# of unique medical services	250	233	230	217
# of unique prescriptions	663	690	646	658
# of unique diagnoses	1249	1271	1203	1200
Total # of unique claims codes	2162	2194	2079	2075
Proportion positive class	0.032	0.128	0.168	0.332

4.4.4 Experimental Setup

Baselines.

We compare our proposed model to the following baselines:

- **Single level Transformer (Claims-SLT)**. This is equivalent to Claims-MLT but does not include the transformer encoder in the first feature encoder block. It uses only attention pooling on the embedded claims inputs to obtain month representations. This is similar to the existing single-level transformer encoders used for claims [263] and EHR data [119, 118].

- **Shallow machine learning models**. We use the **Decision Tree (DT)** and **Random Forest (RF)** for shallow machine learning approaches, as they have been frequently used for predicting cancer-related outcomes, including survival [253, 265, 98], cancer prognosis and prediction [249, 135], and treatment outcomes [43]. These models are trained on claims items and static patient features, however do not consider the temporal relationships between claims items. For these approaches, claims data is transformed to a single vector of size n (number of unique claims codes in the dataset), with values being 1 if a claims item is present in the patient's claim history, else 0. This approach has been leveraged previously for comparison between deep learning and shallow machine learning approaches in health-related predictions [116]. The claims data is then concatenated with the static patient features.

- **Claims- and static-only models**. To investigate the impact of static patient information, we also implement a claims-only multi-level Transformer (Claims-only-MLT) and claims- and static-only DT models.

Evaluation Metrics.

The following performance metrics are recorded for models: precision, recall, F1 score, and accuracy. We use the **F1 score as the primary evaluation measure**. The F1 score is the harmonic mean of precision and recall, where precision equals true positives / (true positives+false positives); and recall equals true positives / (true positives+false negatives).

Implementation Detail.

We implement all coding and analysis pipelines with Python 3.7.6 and run models on an Intel(R) Xeon(R) CPU E5-4650 v2 @ 2.40GHz with 65GB RAM and 16 CPUs. We use Pytorch 1.10.2+cu102 for transformer models and Scikit-learn is used for the training of DT and RF models. For model training, we randomly split data into training, validation, and test sets in ratios of 80%, 10% and 10% respectively. To avoid data leakage, only data in the months prior to the chemotherapy initiation date was used as input for model training. The input data was also limited to the most recent 18 months of claims data for each patient, to enhance computational efficiency and avoid overemphasis on distant non-cancer historical information. We use up-sampling of the minority (positive) class to obtain balanced training sets. For RF models, we set the number of trees to 40 for model training. All models were run three times to calculate mean and standard deviation of performance metrics on the test set. For reproducibility, seeds 7, 32, and 42 were used for experiments.

For transformer models, we follow our previous work [72] and similar studies [128, 41] using Adadelta [261] for gradient descent optimisation. We use a minibatch of 32 patients and a learning rate of 0.01, with experimentation also performed with various batch sizes (4, 32, 256), learning rates (0.0001, 0.001, 0.01, 0.1), and optimisation methods (Adam, Adadelta), selecting parameters based on F1 scores from the validation dataset. We use a hidden dimension of 128, consistent with use of small dimension sizes for trans-

former models on EHR [72, 119, 188, 118]. For each Transformer encoder, we used two stacked encoder layers (as used in previous work [72]) with 4 attention heads, to reduce complexity due to computational limitations. To avoid overfitting, drop-out strategies are used for all approaches (dropout rate = 0.1).

4.4.5 Results

Survival Prediction Results.

To evaluate the proposed model on survival prediction, we compare performance metrics in mean and standard deviation (in brackets) against the baselines (Table 4.3). As demonstrated by the F1 scores, the proposed Claims-MLT model outperforms shallow machine learning approaches (i.e., DT and RF) and the single-level transformer model (Claims-SLT). This indicates the benefits of Claims-MLT to model temporal claims data leveraging dual Transformer encoder blocks.

We also observe that the way in which claims items in patient histories are represented is critical for deep learning models, as the Claims-SLT model is also outperformed by the DT in predicting survival for colorectal cancer.

Finally, while we note the higher accuracy scores of the shallow machine learning approaches, the imbalanced data makes this an improper measure of model performance [256]. The class imbalance also contributes to the difficulty of the prediction task, and exploration of approaches to deal with imbalanced data, or small datasets (such as transfer learning), may offer improvements. Further work training the survival model on the total cancer cohort rather than only those that undergo chemotherapy could also improve predictions.

Heart Disease Prediction Results.

We also compare model performance of Claims-MLT with baselines for the heart disease prediction task (Table 4.4). Like the survival task, Claims-MLT provides better predictions than shallow machine learning and single-level transformer baselines. This further supports the advantages of leveraging month- and patient-level claims item relationships with dual Transformer encoders for predicting cancer outcomes.

Table 4.3 : Model performance for survival prediction.

	Precision	Recall	F1	Accuracy
Breast Cancer				
Claims-MLT	0.120 (0.015)	0.583 (0.021)	**0.208** (0.019)	0.858 (0.019)
Claims-SLT	0.119 (0.008)	0.604 (0.042)	0.199 (0.013)	0.845 (0.006)
DT	0.167 (0.014)	0.208 (0.021)	0.185 (0.017)	0.942 (0.001)
RF	0.333 (0.577)	0.007 (0.012)	0.014 (0.024)	0.968 (0.001)
Colorectal Cancer				
Claims-MLT	0.234 (0.054)	0.587 (0.081)	**0.329** (0.045)	0.680 (0.115)
Claims-SLT	0.210 (0.028)	0.681 (0.099)	0.319 (0.030)	0.624 (0.067)
DT	0.300 (0.020)	0.346 (0.044)	0.321 (0.031)	0.813 (0.003)
RF	0.618 (0.071)	0.097 (0.008)	0.168 (0.014)	0.876 (0.003)

Compared to survival prediction, we see better model performance (in terms of F1 scores) for the heart disease task, indicating the cancer data may be better suited for prediction of future disease. This may be due to the historical patient context available within claims data, including past diagnoses and medications that may be related to cardiovascular or other associated conditions. Given the noted link between cancer treatment and future heart disease [9], this finding may be relevant for the development of cardiovascular-specific risk prediction models for cancer patients.

Table 4.4 : Model performance for heart disease prediction.

	Precision	Recall	F1	Accuracy
Breast Cancer				
Claims-MLT	0.474 (0.012)	0.807 (0.020)	**0.597** (0.004)	0.816 (0.008)
Claims-SLT	0.463 (0.006)	0.815 (0.021)	0.591 (0.002)	0.810 (0.004)
DT	0.437 (0.010)	0.476 (0.018)	0.456 (0.009)	0.808 (0.005)
RF	0.668 (0.034)	0.368 (0.028)	0.474 (0.030)	0.862 (0.007)
Colorectal Cancer				
Claims-MLT	0.600 (0.014)	0.746 (0.004)	**0.666** (0.007)	0.750 (0.009)
Claims-SLT	0.600 (0.016)	0.762 (0.036)	0.656 (0.007)	0.747 (0.007)
DT	0.554 (0.020)	0.564 (0.009)	0.559 (0.015)	0.703 (0.013)
RF	0.712 (0.016)	0.642 (0.043)	0.658 (0.005)	0.788 (0.006)

Effect of Static Patient Information.

To explore the impact of patient demographic and clinical information, we conduct an ablation study and compare performance metrics for models with claims- and static-only data (Fig 4.2). When comparing the claims-only transformer to Claims-MLT, we find inclusion of static patient information provides better model performance for both tasks and cancer types.

Similar results are also seen across cancers in the survival task for DT models. However, for the heart disease prediction task, the DT model with claims and static patient information outperforms the claims-only DT for colorectal cancer, but not breast cancer. This difference from the transformer models may reflect the distinct static feature integration approaches, where fusion is used for Claims-MLT compared to the concatenation approach for DTs.

The result could also be due to cancer-specific differences in prediction labels for the patient demographic and clinical features. For example, colorectal cancer has a higher proportion of patients with a future heart disease diagnosis for all age groups and cancer stages, and hence these features may be less useful in making predictions for breast cancer patients with the DT. Future work exploring individual demographic and clinical features would be helpful to further understand their impacts on performance, however we note the broad benefit of including these static patient features, particularly with the Transformer model.

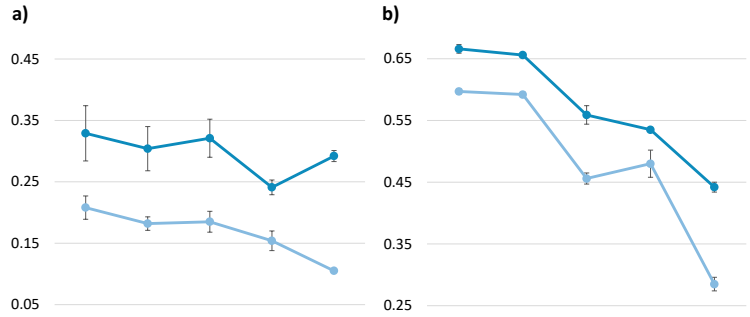

Figure 4.2 : Impact of static patient features on model performance (F1 scores) for breast cancer (light blue) and colorectal cancer (dark blue) on a) survival and b) heart disease diagnosis prediction. Results from left to right in each figure include Claims-MLT, Claims-only MLT, DT, DT (claims only) and DT (static only). Error bars shown for F1 scores.

4.5 Conclusion

In this chapter, we present a multi-level transformer designed to model complex relationships between claims data and integrate additional demographic and clinical features to predict cancer outcomes. Compared to a single-level transformer and traditional machine learning approaches, results indicate the proposed approach provides improved survival and heart disease diagnosis predictions for breast and colorectal cancer patients initiating chemotherapy. There are several avenues for future research, including alternative claims data processing approaches, consideration of machine learning techniques such as transfer learning to further leverage claims item relationships, additional exploration of patient features, and expansion of models to alternative data and prediction tasks. This work demonstrates the potential of leveraging patients' claim histories and static features with transformers to predict cancer outcomes, and offers future opportunities to improve predictive models for cancer, and more broadly in healthcare.

Chapter 5

Pretrained Transformer for Claims Data to Predict Chronic Conditions for Cancer Patients with Limited Data

This chapter focuses on the problem of predicting outcomes for cancer patients with limited data or labels. It investigates the benefits of transfer learning and pretrained transformers in the prediction of chronic conditions for those with cancer or mental health conditions using claims data, which does not contain comprehensive diagnostic information.

Publication Declaration

This chapter is based on the following published paper: *Gerrard, L., Peng, X., Clarke, A., Long, G.: Claimsformer: Pretrained Transformer for Administrative Claims Data to Predict Chronic Conditions. In: Australasian Joint Conference on Artificial Intelligence. pp. 348-362. Springer (2024).* First published in vol 15443, pages 348 - 362, 2024, by Springer Nature. The final authenticated version is available online at: https://doi.org/10.1007/978-981-96-0351-0_26. Reproduced with permission from Springer Nature. This paper was published as part of Australasian Joint Conference on Artificial Intelligence (AJCAI) 2024. Where appropriate, the paper has been amended to ensure consistency in structure, removal of redundant information, and alignment with thesis formatting requirements.

Author Contributions

LG and XP developed the research idea. LG designed the study, prepared the data and labels, and performed the experiments, including model pretraining and evaluation. LG analysed and interpreted the results. LG wrote the initial version of the paper, including supplementary materials, and revised and edited the paper in response to co-author and

reviewer feedback. XP provided advice for experimental design and data preprocessing, and assisted with running experiments (for computational efficiency). XP also supported code review and refinement. AC contributed to the research question and direction, provided feedback on the paper, and facilitated the review of the paper by data experts. GL provided overall project oversight and contributed to method and paper improvements.

Author Declaration

I confirm that the description above accurately reflects my contribution as an author to the publication listed.

Signatures

Leah Gerrard (LG): Production Note: Signature removed prior to publication.

Dr Xueping Peng (XP): Production Note: Signature removed prior to publication.

Dr Allison Clarke (AC): Production Note: Signature removed prior to publication.

A/Prof Guodong Long (GL): Production Note: Signature removed prior to publication.

5.1 Introduction

Chronic conditions can substantially influence health outcomes and are thus critical considerations in policy-making, clinical decisions, and research studies. For example, chronic condition status can affect eligibility or recommendation for healthcare services such as COVID-19 vaccines[1], affect survival outcomes for those with cancer or mental health conditions [113, 137], and help control for confounding factors in health research [26]. Therefore, understanding chronic condition status is essential for optimising population health outcomes and advancing research studies.

While some health data sources, such as hospital records, contain diagnoses, administrative claims data in Australia, including the MBS and PBS, does not have diagnostic information. Consequently, this data must be linked with other sources to obtain diagnoses. The complex governance structures involved in linking such data often results in

[1] https://immunisationhandbook.health.gov.au/resources/tables/table-conditions-for-which-covid-19-vaccination-can-be-considered

incomplete condition information for the population [228], leading to data and label insufficiencies. This is a common issue in healthcare due to data and privacy protections [124], the need for expert domain knowledge [245], and the time-consuming nature of data labelling [246]. Transfer learning has emerged as a promising approach to address the data and label insufficiency problems. This method involves a two-stage process. The first stage is pretraining on a large, general dataset to obtain a model that captures the underlying structure of the data [78, 188]. In the second stage, this pretrained model is fine-tuned on specific tasks with smaller, labelled datasets [270].

In recent years, pretraining has been widely adopted in natural language processing, primarily using the Transformer [232] architecture (e.g., BERT [52] and GPT [179]). These models have since been adapted for healthcare data, to model longitudinal patient trajectories [167, 166, 169, 168], improve predictions for cancer and mental health populations [70, 72, 142], and develop pretrained models [119, 188, 199]. However, most existing applications have focused on EHR data, not claims data, which is arguably an ideal data source for transfer learning due to its near population-level coverage. Additionally, existing methods lack comprehensiveness as they often rely solely on BERT-based approaches, single pretraining strategies, or limited downstream tasks.

To address the aforementioned challenges, we propose the **Claimsformer**, a trans**former** tailored for administrative **Claims** data, to learn generalised patterns and relationships from Australian medical services and prescriptions. To summarise, the main contributions of this work are:

- A comprehensive investigation into pretraining on claims data, which, to the best of our knowledge, is the first application and proof-of-concept demonstration of pretrained transformers on Australian claims histories.
- Evaluation of five pretraining strategies on multimorbidity prediction, to identify the optimal pretraining strategy for the Claimsformer model.
- An experimental study demonstrating the effectiveness of the Claimsformer in predicting chronic conditions for two cohorts: cancer and mental health.

The remainder of this chapter is organised as follows: Section 5.2 briefly reviews the

related work on pretrained Transformers for EHR and claims data. Section 5.3 presents the model detail. Section 5.4 describes the data and experimental results. Section 5.5 concludes the chapter.

5.2 Related Work

Pretrained Models for EHR and Claims Data

While there has been recent growth in the number of pretrained models for longitudinal health data, these have primarily targeted EHR data [119, 188, 118, 199, 240], with claims data receiving less attention. In addition, existing approaches have primarily focused on the BERT-based pretraining objective of MLM, tending not to evaluate other pretraining strategies. Applications to claims data include those by Prakash et al. [176] who developed RareBERT, an enhancement of Med-BERT to detect rare diseases from claims data, and ExMed-BERT, by Lentzen at al. [116], who also extended Med-BERT by pretraining on an EHR and insurance claim dataset. While these existing studies highlight the potential of transformer-based models for claims data, they only use a limited number of downstream tasks to validate the benefit of pretraining, failing to assess broader generalisability. Furthermore, they implement a limited number of pretraining strategies, leaving the optimal approach for claims data unclear. This motivates the exploration of different pretraining strategies and fine-tuning on several downstream tasks in this research.

5.3 Methodology

5.3.1 Model Overview

An overview of the proposed Claimsformer is shown in Fig 5.1. It consists of a two-stage transfer learning approach: pretraining and fine-tuning, which both follow the Transformer model architecture. In the pretraining stage, the embedded claims inputs travel through the transformer block. We implement a number of pretraining strategies to identify the optimal approach for learning generalised knowledge in claims histories. In the fine-tuning stage, the pretrained Claimsformer is used to predict several chronic conditions for cancer and mental health cohorts with limited labelled data. All fine-tuning

tasks are trained independently by adding a task-specific prediction head, initialising with the parameters from pretraining, and optimising all parameters end-to-end.

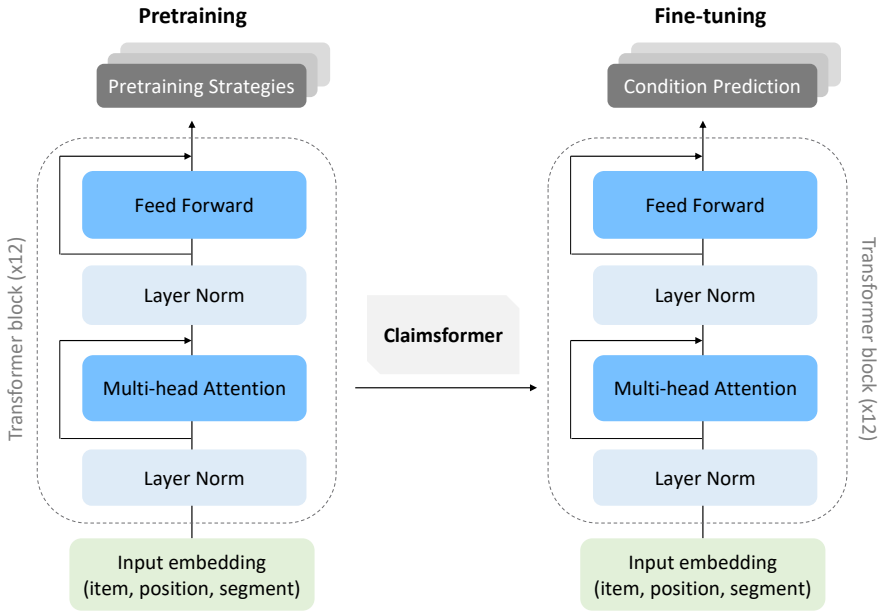

Figure 5.1 : The proposed Claimsformer model.

5.3.2 Input Embedding

Inspired by BEHRT [119], we prepare the data similarly to this model. However unlike EHR data, which can be split into a series of hospital visits, claims data cannot be logically separated in this way. We therefore group claims items into calendar months. This allows for flexibility during short timeframes (i.e., within months), where the specific date (or order) of a claim is less relevant, while still capturing temporal relations over longer timeframes (i.e., between months), which may be indicative of persistent or progressive healthcare intervention related to cancer or other chronic condition.

We denote the set of unique items from the claims data as $C = [c_0, c_1, \ldots, c_n]$, where c_i represents the i-th claims item and n is the number of unique claims items. A person's claims sequence is represented as $X = [M^0, M^1, \ldots, M^T]$, where M^j is the j-th month and T is the total number of months ordered in time. Each month M^j contains a list of claims items from C which can include medical services, medication prescriptions, or

both. This is represented as $M^j = [c_0^j, c_1^j, \ldots, c_k^j]$, where k is the number of claims items in the month.

To ensure all claims sequences are the same length, we pad inputs on the right using the [PAD] token, based on persons with max k (in each month) and T. Like BEHRT [119], we include the [CLS] token to indicate the beginning of the person sequence and the [SEP] token to separate the months. The person claims sequence thus becomes X = [CLS, M^0, SEP, M^1, …, M^T, SEP], which can also be considered the person claims history and is depicted in Fig 5.2.

An embedding layer transforms the person claims sequence X to a low-dimensional dense vector representation of dimension d. As shown in Fig 5.2, we also include information on position and segment to capture the temporal and alternating month relations. The final embedding E is a sum of the three embeddings as depicted below:

$$E = \texttt{LayerNorm}(X_e + P_e + S_e), \tag{5.1}$$

where X_e, P_e and S_e are the respective claims item, position, and segment embeddings for a person's claims history.

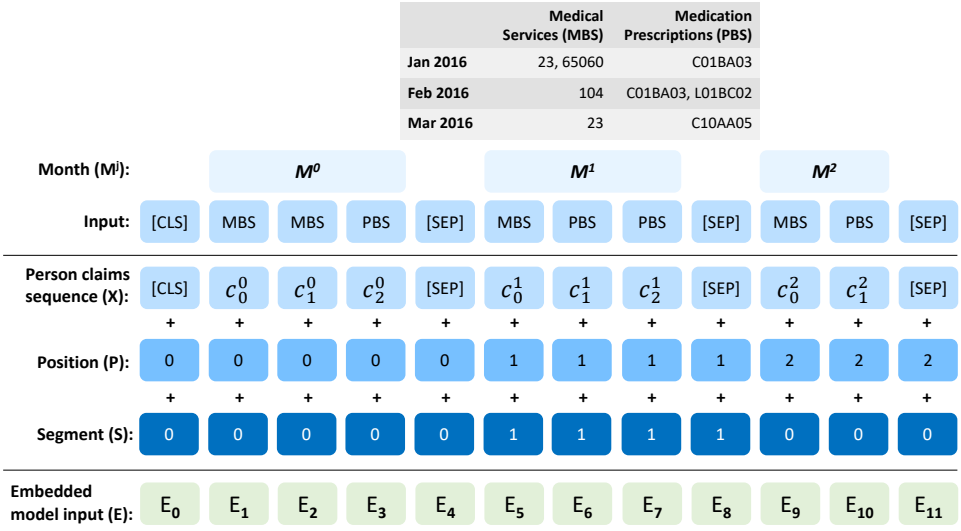

Figure 5.2 : Input from a hypothetical person showing how the Claimsformer model sees the claims data.

5.3.3 Transformer Block

As in our previous work [70], we leverage the Transformer to capture the temporal dependencies within person claims histories. This includes its two sublayers of 1) multi-head attention and 2) a position-wise fully connected feed-forward network. However, we include layer normalisation and residual connections at the beginning of each sublayer (as done in GPT-2 [180]), given the demonstrated benefits in stabilising training and supporting deep models [248, 237]. We further follow the architecture of GPT [179] and GPT-2 [180], by applying 12 x transformer blocks (see Fig 5.1). The summed input embedding serves as input to the transformer block, which outputs a hidden representation for each input. This representation relies on the self-attention mechanism to capture the context of the input within the claims history. This can be represented for the claims items as follows:

$$[h_1, h_2, \ldots, h_t] = \texttt{Transformer\_block}([e_1, e_2, \ldots, e_t]), \tag{5.2}$$

where $e_i (1 \leq i \leq t)$ denotes the claims item embeddings, $h_i (1 \leq i \leq t)$ represents the hidden vectors of the claims items, and t is the number of claims items in the person's claims history. As the transformer is now a well-established deep learning model, we refer the reader to the original Transformer [232], and GPT [179, 180] papers for more information.

5.3.4 Model Pretraining

We detail below the pretraining objectives to learn the general relations in claims data.

Objective 1: Masked Claims Modelling.

The purpose of Masked Claims Modelling (MCM) is to predict the masked claims items based on the surrounding context (i.e., the person's claim sequence). For month M^j, this can be thought of as $M^j = [c_0^j, c_1^j, \ldots, [MASK], c_k^j]$, where implementation of the [MASK] token and training is based on the BERT [52] MLM task, which has been successful for pretrained models such as BEHRT [119] and Med-BERT [188]. In contrast to BERT which uses sentence pairs as input, we follow the approach commonly used for EHR data [119, 188, 199] and provide the entire person's sequence as input, allowing the

transformer access to all of a person's claims history for bidirectional context. We mask a maximum of 20 items per sequence (i.e., claims history).

Objective 2: Future Claims Items Prediction.

The objective of this task is to predict the future claims items based on the preceding context. This approach leverages masked self-attention (from GPT models [179, 180]), restricting the model to attend only to past claims items. Unlike GPT, we do not make predictions in an autoregressive fashion, instead the model uses the representation of the [CLS] token to predict the probability distribution of all claims items being a future item.

We explore two scenarios for the future claims items prediction objective:

- **Last Month Prediction (LMP)**: predict the claims items in the *last* month of the person's claims sequence based on all previous months, i.e., input $X = [M^0, M^1, \ldots, M^{T-1}]$, and ground truth M^T.
- **Next Month Prediction (NMP)**: predict the claims items in the *next* month of the person's claims sequence, where the claims history is split into months, i.e., $\{[M^0], [M^0, M^1], \ldots, [M^{0:T-1}]\}$, and predictions are made sequentially for $M^{1:T}$.

Objective 3: Age Prediction.

We include age prediction as an auxillary task for pretraining, where the objective is to predict a person's age group based on their claims history. Given the link between age and chronic condition prevalence [136], we hypothesised that jointly learning claims and age relations could provide additional context for predicting conditions. For this objective, we process the age feature into 5-year categories until 85 years or older who are grouped together into the same category (e.g., 65-69, ..., 85+).

We choose to implement this as a pretraining objective rather than provide age as input to the model (as has been done for EHR [119, 199]), as data must be sufficiently longitudinal for age to change. Inclusion as an input feature may be worthwhile if including several years of claims histories, which is an area for future work. Similar to the future claims items prediction task, this objective uses the [CLS] token as the representation of the patient input sequence to predict age group.

Pretraining Strategies.

Based on the three objectives above, we implement five pretraining strategies for the Claimsformer, summarised in Table 5.1. These pretraining strategies, implemented either independently, or in combined with other objectives, are compared to determine which options can provide better representations for patient claims data.

Table 5.1 : Summary of the pretraining strategies implemented for the Claimsformer model. Size refers to the number of model parameters (M=million).

Strategy	**Objective/s**	**Size**
MCM	MCM only.	38.04M
MCM+Age	Jointly on MCM and age group prediction.	38.04M
MCM+LMP	Jointly on MCM and LMP.	40.06M
LMP	Future claims items in the *last* month.	37.77M
NMP	Future claims items in the *next* month.	37.77M

Pretraining Loss Function.

As the pretraining objectives are all multi-class classification tasks, they are trained using multi-class cross entropy loss, which for a single instance is calculated as below:

$$\mathcal{L} = -\sum_{i=1}^{n} y \log(\hat{y}), \tag{5.3}$$

where n is the number of classes (i.e., number of claims items or age groups), y is the true label and $\hat{y}$ is the softmax probability for the i-th class. For the pretraining strategies that are jointly trained on multiple objectives (i.e., MCM+Age and MCM+LMP), we implement the loss as 4*MCM+Age and 6*MCM+LMP, respectively. While each strategy uses different objectives for pretraining, we use a uniform set of hyperparameters inspired by the GPT [179] architecture.

5.3.5 Fine-tuning on Downstream Condition Tasks

We conduct fine-tuning tasks for two purposes: 1) to compare the pretraining strategies and identify the optimal approach for the Claimsformer model, and 2) to evaluate

the Claimsformer against other model baselines. We predict multimorbidity and single conditions for our two purposes as described below.

Multimorbidity.

Multimorbidity is defined as two or more coexisting chronic conditions in the same individual [217]. In this work, the objective of the multimorbidity task is to predict the presence of two of more of the following chronic conditions: diabetes (Diab), high cholesterol (HC), hypertension (HT), heart, stroke and vascular disease (HSVD), back pain (BP).

Single Condition Prediction.

The objective of this task is to predict the presence (positive label) or absense (negative label) of a chronic condition. We use the same conditions from the multimorbidity task, however the model is trained to predict each condition independently. Because we make predictions for cancer and mental health cohorts, we also predict mental health (MH) conditions for those with cancer (Can), and vice versa.

Fine-tuning Loss Function.

All fine-tuning tasks are binary classification tasks trained via average cross-entropy loss computed as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{t=1}^{N}\left[y\log\left(\hat{y}\right) + (1-y)\log\left(1-\hat{y}\right)\right], \tag{5.4}$$

where N is the number of data points, y is the true label of the fine-tuning task (i.e. multimorbidity or single condition label) taking a value 0 or 1, and $\hat{y}$ is the probability for the N data point.

5.4 Experiments

5.4.1 Dataset

We conduct experiments on real-world Australian health data, the Person Level Integrated Data Asset (PLIDA)[2], as part of the Healthy Mind, Healthy Body Project[3]. PLIDA uses a number of administrative databases to create a near-population level Person Linkage Spine. The dataset includes medical services from the MBS, medication prescriptions from the PBS, health conditions from the National Health Survey (NHS) 2017-18, and the combined demographics database.

5.4.2 Data Preprocessing

Base Cohort.

Individuals from the Medicare Consumer Directory who had a single PLIDA linkage identifier (we use version 4 of the spine) and were present in the combined demographics database were considered for the base cohort. Age information was derived from the combined demographics database. Exclusions were applied for those with a recorded death date, missing birth year, or aged over 100 years at 1 Jan 2017. This gave a base cohort of 27.63 million.

MBS and PBS Data.

Bulk-billing and incentive items were excluded from the MBS data and number of services calculated by summing the number of services per person, per item, per date of service. Only instances where sums were greater than zero were kept. Sizes of the MBS and PBS cohorts were 23.10 million and 21.30 million, respectively. To further prepare the MBS and PBS datasets for analysis, we follow the below preprocessing steps:

- Items from the PBS were mapped to chemical substance ATC5 codes to obtain

[2] https://www.abs.gov.au/about/data-services/data-integration/integrated-data/person-level-integrated-data-asset-plida

[3] The project has ethics approval from the Australian Institute of Health and Welfare Ethics Committee (EO2019/3/1075). The data was accessed securely through the Australian Bureau of Statistics DataLab infrastructure.

drug level categories. Items that were unable to be mapped to ATC5 codes were excluded.

- MBS and PBS datasets were grouped by person, claims item, and month of service, with duplicate items within a month removed to avoid focus on recurring items and reduce the size of the input data.
- MBS and PBS datasets were combined, grouping by person to obtain the overall monthly services. Not all months contain both MBS and PBS items.

Pretraining Cohort.

The combined MBS and PBS claims data contains abundant health-related information but does not contain condition level labels. Given the known relationship between age and chronic condition prevalence [136], we limited the pretraining cohort to individuals aged 65 to 100 years, reducing the base cohort to 4.42 million.

We then randomly selected a 100,000 sample of individuals from this cohort for pretraining. We merge the sample with the combined claims data, and exclude those without at least five months of data, to ensure there is sufficient temporal information in claims histories for pretraining.

We use combined claims data (MBS and PBS data) from 2016 to mid-2017 for pretraining. After data processing, the pretraining cohort is 78,929 persons with 3939 unique claims items. While we use a relatively small pretraining cohort, this could easily be scaled up and/or broadened in future work.

Fine-tuning Cohorts.

Individuals in the NHS 17-18 who were also present in the base cohort and combined MBS and PBS claims data were considered for fine-tuning. We use conditions reported in the NHS 17-18 as fine-tuning labels and use definitions from the ABS NHS 17-18 User Guide[4]. We make multimorbidity predictions for three NHS cohorts, and single condition predictions for two NHS cohorts aged 65+: those with cancer, and those with mental health conditions. Tables 5.2 and 5.3 provide the statistics of the fine-tuning cohorts.

[4] https://www.abs.gov.au/ausstats/abs@.nsf/mf/4363.0

Table 5.2 : Statistics of the fine-tuning cohorts for multimorbidity prediction.

	NHS	**NHS 65+**	**NHS 65+ w/ Can or MH**
cohort size	17,962	3,692	1,675
multimorbidity (%)	14.88	39.90	45.73

Table 5.3 : Statistics of the fine-tuning cohorts for single condition prediction.

	Cancer cohort	**Mental health cohort**
cohort size	1,127	782
diabetes (%)	16.42	19.69
high cholesterol (%)	29.10	32.86
hypertension (%)	47.83	48.08
HSVD (%)	24.31	25.70
back pain (%)	31.77	39.26
cancer (%)	-	29.92
mental health conditions (%)	20.76	-

5.4.3 Experimental Setup

Baselines.

We compare the Claimsformer with a naive transformer (Naive T) that is not initialised with pretrained weights, and three non-deep learning models for tabular data: Gradient Boosting Classifier (GB), Logistic Regression (LR) and RF. Because these models cannot handle sequential inputs, we use standard one-hot encoding of claims items as input.

Evaluation Metrics.

We use average precision score and top-k accuracy for the pretraining objectives. For fine-tuning, we use the **F1 score** as the primary evaluation measure, however we also report AUC due to the frequent use of this metric for comparing similar models in EHR data [119, 188, 199, 240]. In addition, we record the precision, recall and accuracy of models.

Implementation Details.

We implement all coding and models with Python 3.10.9 and run models on a NVIDIA-SMI 531.14 Tesla T4 Graphics Processing Unit (GPU) with 16GB memory. We use PyTorch 2.2.1 cu12.1 for transformer models and Scikit-learn for the non-deep learning baselines.

We use all data for pretraining due to self-supervised learning. For fine-tuning, data is split into train, validation and test in ratios of 80:10:10. Upsampling of the minority (positive) class is used to provide balanced training datasets. This includes the impact of fine-tuning data size study where the proportion positive class is between 44.6 and 49.2% due to upsampling already applied.

Results shown are based on the test set, of which validation performance is used to identify the best model. For the convergence study, we identified the best validation performance with the epoch ranges (e.g., 1, 1-2, ..., 1-5) and report the results from the test set. For reproducibility purposes, we set the random seed, torch manual seed and numpy random seed for all experiments.

For transformer models, we largely follow the architecture of GPT [179], with a few

modifications. We reduce the batch size and hidden size to fit GPU memory requirements. We increase the number of attention heads to capture more complex sequential patterns, and also increase the learning rate, with attention head and learning rate values consistent with recent transformer models on EHR [107, 255]. We also leverage Adam [104] as the optimiser, as this is frequently used for pretrained transformers in healthcare [188, 118, 244]. We reuse the same hyperparameter settings for fine-tuning, except ReLU is used exclusively and we reduce the number of epochs (further details below). A summary of the hyperparameters can be found in Table 5.4. The number of unique claims items for fine-tuning is determined by adding any new unique claims items in the fine-tuning data to the pretraining vocabulary. Fine-tuning models thus had 4173 unique claims items and 37.77 Million parameters.

All models were run three times to calculate evaluation metrics on the test set. For fine-tuning, we train transformer-based models for a small number of epochs (five epochs) as is standard for pretrained models [52, 179].

Table 5.4 : Hyperparameters for pretraining and fine-tuning.

Hyperparameter	**Pretraining**	**Fine-tuning**
batch size	32	32
learning rate	0.0001	0.0001
number of epochs	50	5
hidden size	512	512
intermediate size	2048	2048
number attention heads	16	16
number layers	12	12
gradient accumulation steps	4	4
dropout	0.1	0.1
hidden activation function	GELU, ReLU for MLM	ReLU
optimiser	Adam [104]	Adam [104]

5.4.4 Results

Comparison of Pretraining Strategies on Multimorbidity Prediction.

To determine the optimal pretraining strategy for the Claimsformer, we evaluated the approaches on the multimorbidity prediction task. We provide the F1 and AUC metrics

in Table 5.5, and report all other metrics in Table A.2 of Appendix A.

We observe that the MCM pretraining strategy leads to the best multimorbidity prediction for two of the three cohorts examined. However, for the cohort aged 65+ with cancer or mental health conditions, pretraining with the MCM+Age strategy outperforms the other methods. This suggests that MCM is the optimal approach for learning relations in claims data, with age information offering additional benefit in some instances.

The difference in multimorbidity prediction between the MCM and MCM+Age strategies may be explained in part by the behaviour during pretraining, where the MCM precision is best when trained via MCM alone, and performance is slightly harmed when introducing a second pretraining objective (See Fig A.1 in Appendix A). Based on the multimorbidity results, we selected the Claimsformer (CF) pretrained via MCM and its age variation (CF+Age) for further evaluation in the single condition tasks.

Table 5.5 : Comparison of pretraining strategies on multimorbidity (MM) prediction for the NHS cohort (MM), NHS cohort 65+ (MM 65+) and the NHS cohort 65+ with cancer or mental health conditions (MM 65+ w/ Can or MH). Performance metrics are average (standard deviation).

Strategy	MM		MM 65+		MM 65+ w/ Can or MH	
	F1	AUC	F1	AUC	F1	AUC
MCM	**72.5** (2.3)	**80.0** (1.0)	**75.3** (1.5)	**79.6** (1.4)	81.9 (4.3)	84.6 (3.1)
MCM+Age	65.6 (2.5)	75.4 (1.5)	72.3 (3.4)	78.2 (2.1)	**86.7** (3.8)	**88.3** (3.0)
MCM+LMP	67.0 (3.1)	76.0 (2.0)	68.9 (5.4)	76.3 (3.1)	77.5 (2.5)	81.6 (1.6)
LMP	67.8 (1.8)	76.4 (1.4)	66.4 (2.7)	74.8 (1.5)	79.3 (5.1)	82.6 (3.4)
NMP	67.6 (3.1)	75.9 (1.8)	69.4 (5.1)	76.6 (3.1)	80.9 (3.2)	84.0 (2.3)

Evaluation of Claimsformer on Single Condition Predictions.

We report the task-specific condition predictions (F1 and AUC scores) for the cancer and mental health cohorts in Tables 5.6 and 5.7, respectively. We include the standard deviations and other model performance metrics in Tables A.3 - A.6 of Appendix A. We observe three main findings.

Firstly, the CF or CF+Age variation provides the best performance (in terms of F1 scores) in predicting all conditions for the mental health cohort, and all but one condition

for the cancer cohort. We find inclusion of age information in pretraining is beneficial for predicting some conditions, but only in the mental health cohort, and hence CF without age appears to be the more robust and generalisable model.

Secondly, the non-deep learning models can provide competitive performance for some conditions, supporting the inclusion of these models as baselines. For example, the GB classifier outperforms all comparative approaches in predicting high cholesterol for the cancer cohort. However, for this task, CF remains the superior model in the mental health cohort (which has a smaller cohort size), and is capable of outperforming the GB model for the cancer cohort at a smaller number of epochs (see ***Convergence Analysis*** and Fig. 5.3). We also observe that a non-deep learning model can outperform the naive Transformer in some instances (e.g., HSVD, BP). However, this is likely due to Naive T being highly parameterised and even with dropout methods, may overfit the fine-tuning data compared to pretrained models [188].

Finally, we find that model pretraining provides better condition predictions than training a Transformer from scratch. This is shown by the Naive T model unable to outperform (in terms of F1 scores) both the CF and CF+Age models across all conditions. However, Naive T does particularly well in predicting diabetes, and we further explore low-resource settings for this task (see ***Impact of Fine-tuning Data Size***). Overall, the results support the effectiveness of the Claimsformer, and the use of pretraining on claims histories to enhance chronic condition predictions.

Table 5.6 : Average model performance metrics on single condition prediction for the cancer cohort.

Model	Diab		HC		HT		HSVD		BP		MH	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
CF	**89.3**	90.4	53.9	69.6	**92.6**	**93.2**	**67.4**	**76.0**	**57.9**	**69.9**	**45.8**	**65.4**
CF+Age	87.1	88.6	39.8	62.3	89.1	90.4	56.7	70.4	49.3	66.3	26.0	57.4
Naive T	84.5	**93.0**	52.2	66.7	83.1	82.4	56.5	71.1	19.5	50.7	41.1	62.7
GB	77.8	89.5	**62.7**	**73.2**	78.7	81.7	65.6	74.8	41.1	62.1	41.7	63.2
LR	84.8	86.8	46.8	64.8	73.6	78.8	58.5	70.8	52.8	67.5	40.0	62.5
RF	71.9	78.0	3.9	51.0	73.4	79.0	36.9	61.3	15.3	54.2	0.0	50.0

Table 5.7 : Average model performance metrics on single condition prediction for the mental health cohort.

Model	Diab		HC		HT		HSVD		BP		Can	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
CF	93.3	93.8	**66.7**	**75.3**	80.8	80.7	**63.7**	**74.7**	**54.3**	**68.8**	**56.4**	**69.0**
CF+Age	**94.3**	**94.8**	45.3	64.7	**84.4**	**86.1**	62.0	72.7	37.3	60.8	31.5	59.4
Naive T	93.5	94.6	53.7	60.2	43.3	50.0	50.6	64.1	0.0	50.0	15.5	50.0
GB	80.0	85.9	37.5	61.5	74.7	79.8	60.0	71.4	29.2	58.6	29.5	58.1
LR	85.7	87.5	61.5	72.1	71.2	77.6	50.0	66.7	40.0	61.9	50.0	66.7
RF	71.9	78.1	14.3	53.8	63.4	73.3	19.9	55.6	22.8	56.5	17.5	54.8

Convergence Analysis

To further investigate the benefits of pretraining, we explored whether the Claimsformer converges earlier than the naive Transformer during fine-tuning. We show the results for a subset of conditions in Fig. 5.3. As demonstrated by the results, CF exhibits higher performance at earlier epochs compared to Naive T. In addition, we observe that the CF performance can stagnate and even deteriorate with more epochs, such as for the high cholesterol prediction in the cancer cohort (which achieves an F1 score of 63.4% in the first epoch). This suggests that additional epochs may lead to model overfitting and thus a smaller number of epochs, or other training strategies (e.g., early stopping), may be beneficial.

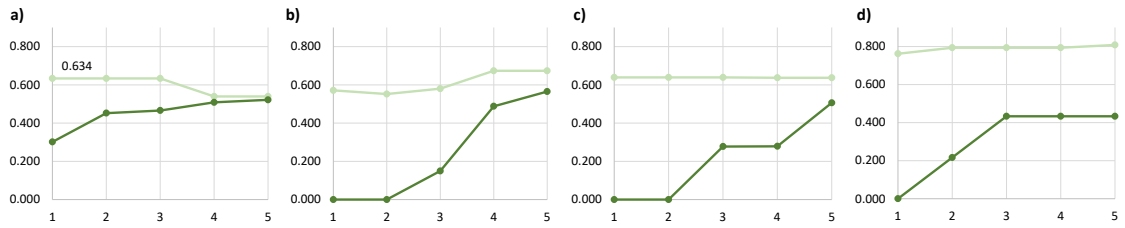

Figure 5.3 : Convergence analysis showing average F1 scores within five epochs for the Claimsformer (light green) and naive Transformer (dark green) for the cancer cohort in predicting a) high cholesterol and b) HSVD, and for the mental health cohort in predicting c) high cholesterol and d) hypertension.

Impact of Fine-tuning Data Size

Given the competitive performance of Naive T compared to CF in predicting diabetes, we conducted additional experiments varying the proportion of the fine-tuning data, and hence examine low-resource scenarios. As shown in Table 5.8, CF obtains better F1 scores for all fine-tuning data sizes and better AUC metrics for all but one sample. It is also able to provide good performance with only 20% of fine-tuning data. These results confirm the benefits of pretraining and the advantage in instances where labels are particularly limited.

Table 5.8 : Model performance metrics for diabetes prediction on various fine-tuning data sizes. Results are presented as average (standard deviation).

	Cancer Cohort				Mental Health Cohort			
Data size	CF		Naive T		CF		Naive T	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
10%	**67.3** (4.2)	**84.4** (2.1)	29.8 (1.9)	53.3 (5.3)	**78.0** (1.5)	**90.3** (2.5)	33.7 (0.0)	50.0 (0.0)
20%	**80.8** (4.0)	**89.6** (2.1)	28.6 (0.0)	50.0 (0.0)	**89.7** (6.3)	**92.7** (1.8)	22.5 (19.5)	50.0 (0.0)
50%	**88.1** (1.4)	**91.4** (0.3)	78.2 (3.4)	89.0 (1.8)	**94.5** (4.2)	**95.6** (4.5)	64.6 (24.6)	80.5 (15.1)
80%	**88.4** (0.4)	90.2 (1.2)	86.7 (3.6)	**93.7** (1.9)	**94.5** (2.0)	**94.8** (1.8)	88.6 (6.2)	91.2 (5.9)

5.5 Conclusion

In this work, we propose the Claimsformer, a pretrained transformer tailored to claims data. We compare various strategies for pretraining, identifying MCM as the optimal approach for learning generalised relations in claims histories. Through experimental results we show the Claimsformer can offer improved chronic condition predictions for cancer and mental health cohorts, earlier model convergence, and effectiveness in low-resource settings, validating the feasibility and usefulness of pretraining with claims data. Future work includes upscaling the pretrained model through a larger population sample, additional historical context or broader population, as well as exploring alternative pretraining (such as generative pretraining) and downstream tasks. This work presents the first demonstration of the potential for transfer learning with Australian claims data, and lays the foundations for further development of such models.

Chapter 6

Discussion and Conclusion

This chapter discusses the key findings of this research and their relevance within the broader context of healthcare. It examines the implications of the research, considers the potential limitations, and suggests future studies. Given the interdisciplinary nature of this research, this chapter also reflects on the ongoing challenges of applying deep learning models in healthcare to highlight additional considerations and opportunities for further work.

6.1 Summary of Contributions

This research aimed to develop more effective transformer-based models with longitudinal health data for complex cancer pathways modelling. In supporting this purpose, several transformers are proposed tailored to longitudinal health data, leveraging MTL, longitudinal patient modelling, and transfer learning, to overcome existing challenges in cancer. These contributions can be summarised as follows:

- **Contribution 1:** The first multi-task Transformer focused on EHR medical code data that specifically models the following related outcomes for hospitalised cancer patients: future diagnosis and readmission. This work confirmed the importance of considering related outcomes for cancer patients who have shown an increased risk of hospital readmission compared to the general population [92], and are often burdened with symptoms and side effects as a result of their cancer and/or treatment [145]. By adapting transformers to incorporate MTL for predicting multiple cancer outcomes, we achieve Objective 1 and address Research Problem (RP) 1.

- **Contribution 2:** Adaptation of the Transformer to claims data, through the first implementation of a hierarchical transformer leveraging month and history multi-level

claims item relations to predict outcomes for cancer patients initiating chemotherapy. The model also integrates demographic and clinical features from a cancer registry to evaluate and reaffirm the value of this additional static patient information in predicting cancer outcomes. By showing the effectiveness of the Transformer specifically designed for the structure of claims data and cancer characteristics, we achieve Objective 2, and address RP2.

- **Contribution 3:** An Australian-first application and proof-of-concept demonstration of pretraining on claims data, through developing a pretrained transformer specifically tailored to model relations between medical services and prescriptions. The pretrained model is evaluated on chronic condition predictions, which are critical in cancer and healthcare given the known links between chronic conditions and survival outcomes [137, 113]. By demonstrating the benefit of pretraining and transfer learning for cancer tasks with limited data, we achieve Objective 3 and address RP3.

Together, these findings demonstrate the effectiveness, flexibility, and data-efficiency of a longitudinal health transformer for cancer pathways modelling and risk prediction. This aligns with existing research on the benefits of transformers for complex sequential and temporal longitudinal health data, supporting the superiority of the Transformer architecture for healthcare information. Furthermore, this advances existing research in the areas of oncology and Australian claims data, by developing domain-specific and tailored transformers to this context and data. The progressive nature of these contributions provides a foundation for a comprehensive cancer modelling framework, and demonstrates the feasibility of developing large-scale and generalised solutions for oncology and healthcare.

6.2 Research Implications

This research has theoretical, methodological, practical, and public health implications. In this section, these implications and the beneficiaries, which include researchers, health professionals, government, the healthcare system, and cancer patients, are discussed.

Theoretical

This research reaffirms and extends the theoretical understanding of transformer architectures for longitudinal health data. Firstly, by showing the superiority of the Transformer compared to shallow machine learning and recurrent architectures, this reinforces existing theory on the benefits of the self-attention mechanism and capability of the Transformer to capture complex sequential relations in longitudinal health data [232]. This deepens the rationale for using transformers over other models for longitudinal health data, particularly where healthcare events are time-ordered and long-term such as for oncology. Secondly, by demonstrating the benefits of modelling long-range sequential patterns with the Transformer, this validates existing views about the inherent temporal relations within healthcare histories [119] and that capturing these patterns is essential for cancer predictive modelling. Thirdly, by showing the benefits when the Transformer is equipped with demographic and cancer clinical characteristics, this further contributes to knowledge on feature importance for cancer-related predictions. This supports the value of combining dynamic and static patient information and the ability of the Transformer to encode this information for richer patient representations.

Additionally, by successfully adapting transformers to claims data and demonstrating their potential for improved predictions, this extends knowledge about their applicability beyond EHR, supporting their robustness in alternative health datasets. Finally, by demonstrating the ability to transfer knowledge from claims data, this contributes to the current theory on self-supervised pretraining in health data [188]. The findings in this research suggest that patient pathways within longitudinal claims data contain meaningful information that together with transformer-based models can be captured and transferred to other tasks. This supports the flexible and transferable nature of transformers outside of natural language and reaffirms their utility in oncology and healthcare.

Methodological

This research has methodological implications for the AI and medical research community, contributing to areas of both computer science and bioinformatics. By demonstrating opportunities to improve transformer-based models with longitudinal health data

for cancer, this advances the development of context-aware, data-driven transformers in oncology. By tailoring models to the differing structure of claims data, specifically Australian claims data, this provides the first methodological insight into how to effectively encode and represent this data with transformers. In addition, this research guides the training and evaluation of large-scale claims data by providing a comprehensive evaluation of pretraining strategies, and examining fine-tuning performance, convergence and low-resource scenarios. By developing an approach for the successful pretraining of claims data with transformers, this has implications for the advancement of domain-specific foundation models and generative systems, as these similarly learn generalised relations and utilise transfer learning [233, 178]. The use of the same underlying model architecture (the Transformer) means that the models in this research can be easily scaled to create larger foundation models, or adapted to pretrain for generative tasks. This can ultimately drive more scalable, data-efficient, generalisable, and comprehensive longitudinal health transformers.

This research also has implications for future predictive models in oncology and more broadly for healthcare. The methods proposed enhance predictions for several key cancer-related outcomes including survival, hospital readmission, as well as predicting current and future disease. This informs the development of subsequent cancer prediction models focused on these and similar tasks. This research also provides an approach for combining dynamic and static healthcare features, to better model sequential healthcare behaviour while including cancer clinical characteristics. This also guides future deep learning-based risk models in oncology. Finally, while the methods in this work are focused on cancer, they are easily transferable to other health conditions, such as chronic disease opening avenues for widening the application of transformers in healthcare.

Ultimately, the methods in this research support the development of more effective models for cancer pathways modelling, contributing to, and advancing, the growing application of transformers in healthcare. This work will help inform the design of future models better tailored to claims data, and encourages the use of transformers for further AI-based innovations for oncology.

Practical

This research has broad implications for practice in both clinical and government settings. Below, a description is provided for each opportunity, followed by comment on the potential facilitators and barriers for adoption.

Support care coordination. The ability of the Transformer to analyse longitudinal health data and provide accurate predictions can enhance the understanding of risks, diagnosis of disease, and patient monitoring [51]. In particular, the use of multi-task models (such as TransMT) that can simultaneously predict outcomes (future disease) and complications (readmission) can provide more comprehensive risk profiles. This allows clinicians to consider multiple competing risks at once, which can help provide prioritised, coordinated, and responsive patient care [38]. Multi-task learning can also be applied to enhance care coordination for those with chronic conditions, by addressing the misalignment between siloed care delivery (disease-specific) and the needs of the patient (patient-centric and coordinated), enhancing care planning, reducing treatment delays, and avoiding adverse events [103, 38]. These models also reduce the interpretation time for clinicians (one multi-task model instead of several individual models), facilitating more timely and efficient insights, contributing to enhanced health care delivery and care coordination for patients [209].

More personalised care. Leveraging large volumes of patient data enables transformers to learn trends and patterns to better provide individualised predictions [51, 148]. This can facilitate healthcare and oncological management through personalised treatment recommendations, interventions, and follow-up care [4, 38, 164, 8]. For instance, the Claims-MLT model developed to predict survival for cancer patients can provide a personalised prognosis to help inform suitability for treatment or palliative care. Similarly, the prediction of future heart disease risk can help understand likelihood of complications or adverse events and predict response to treatment to inform individualised monitoring. These types of models can also help determine suitability for alternative care models, such as Victoria's home-based cancer care [234]. Furthermore, they can enable ongoing monitoring through use of temporal trends to update risks over time, providing dynamic predictions that can continually inform decision making and further enable personalised

care [38].

Earlier preventive interventions. Capturing dependencies in long-range healthcare data assists the Transformer in predicting patient risk, which can accelerate the presence of signals in patient pathways [45]. For chronic conditions, this could include subtle disease markers that facilitate identification of those at increased or growing risk of disease, that may even emerge before symptoms appear [189]. The Claimsformer model developed in this research could be used in clinical settings to warn or flag individuals at risk of developing chronic conditions. This potential is further enhanced by using a pretrained transformer, as the already present generalised understanding of the data structure allows rapid recognition of patterns, demonstrated by the reduced data requirements and quick convergence when fine-tuned ([71] and in Chapter 5). Earlier disease detection enhances ability for proactive and timely interventions to improve patient outcomes and reduce costs [162, 189]. This could occur by reducing impacts of complications, delaying or avoiding onset of chronic disease or multimorbidity; or identifying early disease progression (e.g., cancer recurrence). This is critical for cancer patients, as chronic disease and recurring cancer is linked to poorer outcomes [60, 93].

Improved access to data and knowledge. Analysing large, complex, patient data can be extremely challenging for an individual clinician [7]. Transformers can make this information more accessible by processing and providing insights that would otherwise be potentially unknown [51]. The transformers developed in this research thus have potential to contribute to, or extend, the medical and specialist knowledge of clinicians, in the areas of cancer and chronic disease [45]. By providing improved access to data and knowledge, this has potential to inform clinical decision making to benefit patient outcomes [51]. It also reduces reliance on individual clinician experience, while enabling clinicians to focus on delivering care rather than navigating data [162].

Healthcare analysis efficiency. The proposed models in this research are purpose built for government administrative claims data, and are designed to operate within the secure environments co-located with the data. This facilitates the application and development of transformers on healthcare data by providing already implemented methodologies and practical examples for those working within government. Additionally, the creation of the

pretrained Claimsformer model provides opportunity for a general-purpose transformer that can be applied to numerous health-related prediction tasks. This reduces the need to train a specialised model from scratch for every condition or task and instead leverage the power of generalised pretraining to reuse data [184]. This can help accelerate the use of transformers, and AI, more broadly within government, and provide efficiency gains in health data preparation, preprocessing, and modelling to predict health-relevant outcomes.

Support evidence generation. For governments, which in Australia hold large amounts of longitudinal health data, models that can learn from this vast information and provide insights can support government operations by enhancing the evidence base for policy [51]. For instance, the Claimsformer model can be used by government to predict chronic conditions and infer disease for populations where labelled data is not available. This could contribute to existing reporting and initiatives related to chronic disease (such as the National Strategic Framework for Chronic Conditions [14]) or facilitate further research in understanding impacts of chronic conditions on outcomes. In addition, prediction of elements and outcomes along the cancer care pathway (such as in the Claims-MLT model) can reveal population trends or patterns related to optimal care, by leveraging demographic, health service, and treatment information. This can support the development of policy by facilitating evidence-based data-driven decisions [51, 184].

Near real-time decision making. The ability to analyse information across healthcare with minimal feature engineering enables more rapid analysis and real-time monitoring in government, which can facilitate planning and prompt decision making [164, 8]. There is particular opportunity to support public health surveillance, as transformers can overcome the traditional time-consuming manual data collection and analysis through automation to quickly predict outcomes and issue warnings [164, 247]. This also provides opportunity to better monitor risk factor trends for non-communicable diseases by speeding up the flow of information [164]. By leveraging the Claimsformer model to infer chronic conditions across the population, this could allow near real-time monitoring of uptake of vaccines (e.g., influenza), new treatments, or prescription medication use. This can provide insights that enable more timely interventions (such as communication strategies) or that help with the adoption and implementation of new policies [184].

While the opportunities above indicates promise for this research to improve practice, the adoption of these pathways is not straightforward in the Australian healthcare system. Adoption of all pathways is made easier through continual investment in large, linked data sources that span across healthcare and offer national populations, as there is a need for high-quality and high-quantity data for transformer models [189, 4]. Adoption of predictive models by clinicians is facilitated by easy integration into existing workflows, where models can augment clinician judgement (e.g., through early identification or risk flagging) to support care decisions and reduce clinician burden (through insights that are not easily determined or accessible elsewhere) [148]. Additional longitudinal research using transformers can help bridge the gap between research and adoption by demonstrating the potential scalability and real-world effectiveness [45]. Further training and education, both in clinical and government sectors, will facilitate adoption by enhancing AI literacy and addressing existing workforce skill and knowledge gaps [189, 162]. Additionally, investment into digital infrastructures, particularly into sufficient and scalable compute availability, will aid the analysis of data and model development [164].

In contrast, adoption in clinical practice is made challenging by fragmented data and systems, where data is not available or accessible to provide trends or predictive insights at the right time [4, 45]. Clinicians also need to be willing to engage with AI modes, with existing evidence suggesting lack of clinician involvement and hesitancy to adopt AI [209, 4]. Barriers limiting adoption in both clinical and government practice include model explainability or interpretability issues, potential bias in model outcomes, data protection and privacy, ethical concerns, and costs [209, 45, 164, 51, 189]. In addition, an evolving and unclear AI regulatory landscape [45, 51, 12] can slow implementation across clinical and government sectors, and so adoption of this research may not necessarily be easier in government despite being developed on its infrastructure. Nevertheless, focusing on ways to grow the enablers and address the barriers will facilitate the adoption of this research into practice. Further discussion on the considerations for real-world implementation can be found in Section 6.5.

Public Health

Finally, this research has public health implications at the population and system level for patients, policymakers, and governments. These implications are outlined below, followed by a brief description on which opportunities may be most and least promising for implementation.

Inform resource allocation. Transformers can accurately forecast healthcare utilisation and demand using historical data, current trends, and external factors, to help optimise resource allocation for policymakers and governments [103, 164]. Transformer models predicting cancer-related outcomes at scale can help plan for workforce capacity – e.g., location of specialist cancer nurses; health infrastructure demand – e.g., likely use of Symptom and Urgent Review Clinics (SURCs) in Victoria by cancer patients; and need for community services – e.g., for Victorian home-based cancer care or outpatient chemotherapy [234]. This can directly inform allocation of resources to where they are anticipated to be most needed. This is a key goal of the Australian Government's Australian Cancer Plan (ACP) [29], which emphasises the need to understand and anticipate workforce supply and demand to identify gaps and strategies to address them. This will be essential for longevity and efficiency of the Australian healthcare system, given the ageing population, increasing demand for healthcare services, and rising cost of health [45, 184].

Targeted health policy. Population-level risk stratification using long-term trajectories can help identify characteristics of cohorts at high or persistent risk of poorer health outcomes. For instance, using transformers to predict those at risk of developing cancer can reveal insights to help design or improve government-funded screening programs [219]. Likewise, understanding populations at high risk of cancer progression or recurrence can inform targeted interventions via new health policies or scaled up policies to improve outcomes and reduce mortality rates [184, 189]. Additionally, identifying subgroups at high risk of chronic conditions, such as by using the Claimsformer, is a key opportunity to better target interventions to reduce the burden of disease [189, 162]. By facilitating more tailored health policy, these types of models can also support the goals of national and jurisdictional cancer plans [29, 234, 235], which aim to develop new and personalised

cancer screening programs, improve cancer survival outcomes, increase early access to palliative care, and better support patients in survivorship.

Reduce health inequalities. Health inequalities remain a challenge for healthcare systems [162]. By integrating healthcare service use with sociodemographic information and outcomes (such as in the Claims-MLT model), transformers can model trajectories to identify health disparities that appear or persist across populations. These insights can support the development of equity-focused policies or programs to enhance diagnosis, improve access, or tailor interventions [184, 162]. Given it is widely acknowledged for cancer that outcomes differ by sociodemographic groups (such as age and remoteness) [39, 137], it is critical to develop approaches that can inform more equitable care and outcomes. This is emphasised in the ACP [29], which includes achieving equity in cancer outcomes as a key measure of success, demonstrating the importance of investigating and reducing health disparities for priority populations.

Inform models of care. By considering the sequences and transitions of events in patient trajectories, transformers can capture complex temporal interactions and provide detailed insights across the entire cancer care continuum. This enables transformers to optimise patient care pathways through better data integration and predictive modelling that can identify which pathways deliver the best outcomes [38]. For cancer, this could inform existing care standards, such as the OCPs[1], by estimating each phase of care in the pathway and identifying high-value care patterns. This could, in turn, support the development and incentivisation of new models of care that are outcome rather than service driven, and enable the provision of the right care, to the right patient, at the right time [206]. This offers opportunity to help standardise care across populations and settings, reducing unwarranted variations, and drive more efficient healthcare [162, 4]. This can also support objectives of the ACP [29], which encourages innovative and evidence-based models for providing optimal care, including incorporating AI and predictive modelling to complement the OCPs, and to enhance cancer care planning and treatment approaches. The ACP [29] also stresses the importance of new models for cancer survivorship to reduce cancer recurrence, manage chronic disease, and monitor the long-term impacts of cancer. The

[1] https://www.cancervic.org.au/get-support/for-health-professionals/optimal-care-pathways

transformers in this research thus offer opportunity to investigate cancer pathways with longitudinal health data to inform models of care that can support the healthcare system to move towards more comprehensive and value-based care [206].

Support health policy evaluation. Using longitudinal health data (i.e., diagnostic intervals, treatment cycles, and outcomes), transformer models can support more nuanced health policy evaluation through leveraging entire patient trajectories for testing under different scenarios. This can include estimating the potential impact of interventions or policies, and model trade-offs prior to implementation [164]. This could facilitate new or updated cancer screening programs, with AI-based simulations already providing insights into the impacts of prevention programs, screening expansions, and smoking policies on cancer [219]. Transformers can also evaluate and monitor impacts of strategies post-implementation, including how they affect risks of population groups [184], which would be beneficial if evaluating policies for chronic conditions. This provides opportunity for policymakers to better understand and measure (or monitor) impacts of interventions, and enable more strategic and dynamic policy.

Of the above opportunities, perhaps the most fruitful is enhanced risk stratification that can enable more targeted health interventions, particularly for long-term and complex conditions. This is already supported through integration of population-level data, and this research has demonstrated the potential of the Transformer to analyse long-term trends and patterns to predict outcomes related to cancer and chronic disease. Furthermore, this topic has high political interest and appetite, as demonstrated by initiatives like the ACP and the National Strategic Framework for Chronic Conditions, and due to the rising prevalence of chronic disease and health system impacts [45, 221]. However, despite these supporting factors, integration of such methods does face challenges like fragmented data (across the Australian healthcare ecosystem – e.g., general practice, hospital, Commonwealth), data sharing reluctance and risks, explainability limitations, and bias and fairness issues [45, 51, 189, 257].

Other promising directions include informing models of care and resource allocation, as these are also supported by existing data. Additionally, these opportunities will benefit from leveraging longitudinal trajectories for precise predictions related to patient

pathways or demand forecasting. However, implementing models of care is made challenging by the multi-layered responsibilities within the Australian healthcare system (i.e., Australian Government, jurisdictions) [55]. Furthermore, the nature of government budget cycles can result in misalignment between resource allocation and insight generation from complex data [203].

Despite their importance, the most challenging to implement are health policy evaluation and reducing health disparities. This is primarily due to data constraints or lack of representative data; challenges related to causal inference; political sensitivity; and risk of bias amplification that can perpetuate existing inequalities [45, 162]. However, it is also challenging to apply transformers and AI to these areas due to the complex and interdisciplinary nature of developing policy and existing digital health inequalities [184, 4]. Finally, like the practical implications, regulatory, ethical, and data privacy concerns are also prominent issues [4, 12]. While the transformers developed in this work demonstrate the opportunity to leverage these models to improve public health, it is important to consider the above barriers to help bridge the gap between research and implementation. More details on the challenges and opportunities of implementing transformer models in healthcare is discussed in Section 6.5.

6.3 Limitations

While this research provides valuable insights into transformers on longitudinal health data for cancer, and has considerable real-world implications, it is important to also acknowledge the limitations.

Dataset Size

The datasets used in this research are relatively small for deep learning models compared to that of natural language. Experiments on MIMIC-III and MIMIC-IV EHR data (in Chapter 3) utilised 2,000 and 20,000 patients, respectively. For the experiments on claims data, this ranged from 8,700 to 15,200 for the Claims-SLT model (in Chapter 4) to approximately 79,000 for the pretrained Claimsformer (in Chapter 5). These datasets sizes are largely the result of cancer- and population-specific cohorts. However, the use of small datasets may limit the ability of models to fully capture the complex relations in

longitudinal health data and can contribute to model overfitting and poor generalisability. To help address model overfitting, and data and label insufficiencies, strategies are used in this research including dropout and up-sampling, and model pretraining, respectively. It is also argued that while the datasets used may be small, the use of several diverse data sources in this research assists in supporting the evidence of the superiority of the Transformer with longitudinal health data.

External Validation

While the studies in this research did use internal validation for evaluating model performance, this research did not investigate external validation. This means that for the experiments in this research, data for training, validation, and testing were derived from the same data source. The reason for solely relying on internal validation was primarily due to privacy and data sharing restrictions which prevented the movement of data or models outside of the secure environments. Although this ensures the protection of the data, it does limit the ability for external validation, and hence examination of generalisation. It is therefore, unclear, how these models may perform on independent datasets from different institutions, time periods or cohorts. However, the benefit of using real-world data in this research (as opposed to synthetic), is that it provides a demonstration of how models perform on actual health data.

Generalisability

As mentioned above, the use of limited sized datasets can reduce model generalisability. This means that the model may not generalise well in other settings that differ from the training conditions. In addition, the focus on cancer- and population-specific cohorts means that the model may have poor generalisation to other cancer types, demographic groups, or health conditions. The development of the Claimsformer in Chapter 5 does provide some indication on the generalisability of a pretrained model with claims data, demonstrating benefits from the learned claims relations for predicting chronic conditions. However, further development of this pretrained model, as well as exploration on additional datasets would provide further detail on the generalisability of the transformer-based models in this research.

Hyperparameter Testing

This research primarily leveraged hyperparameters from similar models implemented on longitudinal health data and did not perform detailed experiments on tuning hyperparameters. Leveraging the hyperparameters of BEHRT [119] and Med-BERT [188] models has been common practice [188, 118, 116, 199, 149]. However, a lack of hyperparameter testing means there could be more appropriate alternatives that provide better model performance. Regardless, detailed hyperparameter testing was considered infeasible in this research either due to computational limitations (as described below), or due to a focus on other model aspects (e.g., various pretraining strategies in the Claimsformer model), where inclusion of hyperparameter tuning would have substantially increased the time, computation, and cost required for experiments.

Computing Resources

Given the studies on claims data in this research were restricted to secure data environments, this research was also computationally constrained to these settings. This was primarily an issue for the Claims-MLT model developed in Chapter 4, as this model was restricted to Central Processing Units (CPUs), limiting the ability to train transformers efficiently. Compared to the Claimsformer in Chapter 5 which was trained using similar claims data and the same batch size, but with a GPU, the Claims-MLT model had to use a much simpler model architecture, as it was infeasible to train using the Claimsformer architecture (Claims-MLT: 4 attention heads, 2 layers, and a hidden dimension of 128 vs Claimsformer: 16 attention heads, 12 layers, and hidden dimension of 512). In addition, the long training times limited the flexibility for experimentation and iteration. Collectively, this may have impacted the capture of the complex relations within longitudinal health data, and potentially reduced model performance.

6.4 Future Directions

Building on the developments in this research, there are a number of possible directions for future work. The most promising opportunity is working towards broader solutions for longitudinal health data through development of foundation models and generative capability. This can be achieved by further expanding and improving upon the

approaches in this research and addressing the previously noted limitations. There is also promise in exploring alternative transformer architectures, incorporating model explainability, and extending models to new contexts and diseases. These directions intend to provide more accurate and comprehensive models that can be leveraged for cancer and other conditions, further developing AI-based technologies for modern day healthcare data and challenges. More detail on these directions is provided below.

Foundation Model for Claims Data

Following the demonstrated feasibility of transfer learning in this research, a natural next step is to further improve pretrained transformers for claims data. This will advance the generalisability and data-efficiency of models, facilitating the development of a universal foundation model that can leverage the intricate relations between claims items for a wide variety of tasks. The following opportunities are proposed for the pretraining and fine-tuning of transformers:

- **Expand pretraining data.** One promising direction is to explore opportunities to expand and upscale the proposed pretrained transformer developed in Chapter 5 (the Claimsformer). This can be achieved in various ways such as pretraining on a larger population sample from the 65+ aged group used for pretraining (approximately 4 million individuals available), or using a broader population with more diverse age groups (not just 65+). Alternatively, pretraining could use additional historical years of patient data, to capture more longitudinal claims item relations. Furthermore, the pretrained model could be expanded to include diagnosis codes, in addition to health services and medications. This would provide additional health context that would likely be helpful for understanding generalised patterns. All of these opportunities however, would need to be balanced against available training infrastructure and computational requirements as larger data will require additional training time. Investment and use of flexible Cloud-based solutions will assist this research direction and help overcome computational limitations. This is particularly relevant for government who hold vasts volumes of national claims data that could be leveraged for model pretraining to enhance risk stratification and prediction at scale.

- **Pretraining tasks and configurations.** Another opportunity is to explore alternative tasks and configurations for pretraining. While this research investigated a series of pretraining strategies, there are additional avenues that could be explored, such as implementing the Med-BERT [188] PLOS task, or predicting the type of claims item (e.g., health service, medication, or diagnosis). Further research could also look at alternative configurations for model pretraining. For example, in developing the Carefully Optimised and Rigourously Evaluated BEHRT (CORE-BEHRT), Odgaard et al. [160] report that there may be more optimal MLM ratios for pretraining. In addition, future work could focus on generative pretraining, rather than using BERT-based approaches. This could include pretraining similar to the earlier mentioned TransformEHR [255] model, which showed benefits in cancer prediction, but could leverage more information from longitudinal health data and requires evaluation on claims data and against other pretraining strategies.

- **Alternative input and patient representations.** While this research demonstrates effective longitudinal patient representation of claims data for cancer-related predictions, there are a number of alternative ways the input data could be could be embedded or encoded that may provide better representations. For example, in this work claims items were grouped into months for data modelling to retain longer-range temporal relations while allowing flexibility in short-term relations. This is more similar to how one may logically consider healthcare service use, as events might not always appear in exactly the same order. However, claims items do have more detailed date information (i.e., date of healthcare service or medication supply) that could be explored. This research also aggregated claims item codes for modelling (e.g. for healthcare services, medications and diagnoses) to reduce computational demand. However lower level (i.e., less aggregated) claims item codes could be considered and tested if the data is large enough and computation is sufficient. Outside of the claims item codes, there is also potential to further integrate patient demographics and clinical features. For example, age could be included as its own embedding in the input representation (as done for BEHRT [119]), to reflect changes in age over time and may be helpful in learning relations with diseases. Improvements in the data input representation can further tailor transformers to claims

data and provide opportunity to improve predictions.

- **Further multi-task learning.** Based on the demonstrated benefits of transformer-based MTL for cancer-related predictions, another future direction is to leverage MTL for additional cancer outcomes, particularly with claims data. For example, MTL could be used to predict the outcomes investigated for chemotherapy patients (in Chapter 4) (i.e., survival and future heart disease risk). It could also be expanded to include simultaneous prediction of other outcomes such as emergency department presentations, hospital admissions, mental health conditions, and/or medication complications for cancer patients. Similarly, MTL could be integrated into the fine-tuning tasks for pretrained models, such as for the prediction of multiple chronic conditions in Chapter 5. Given that existing evidence suggests MTL can work well for prediction of multiple conditions [129], and that multimorbidity is common for those with chronic conditions, it is feasible there may be relations here that would benefit predictions.

- **Zero-shot settings**. Another area of exploration for evaluating pretrained models is in zero-shot settings, where no fine-tuning is provided to models and their performance on downstream tasks is measured. Evidence suggests that pretrained models do not necessarily require model fine-tuning for good performance. Recent exploration with transformers for medical event [194] and health trajectory prediction [195] has suggested that models can be used effectively in zero-shot settings for health-related predictions. This offers a promising opportunity to provide more efficient pretrained models.

- **Hyperparameter tuning.** Finally, as opposed to re-using the hyperparameters from similar transformer-based EHR models (such as BEHRT [119] or Med-BERT [188]), detailed hyperparameter testing could be explored for models to determine if there are more suitable parameters for model performance. This could include the use of methods such as grid search [118], random search [216], or Bayesian Optimisation [119], all of which have already been used for AI models in healthcare. Given the additional benefit of hyperparameter testing to help inform model stability, reliability, and robustness under different conditions [11, 183], this is recom-

mended for future studies. As mentioned previously, more powerful computational resources and infrastructure will help support this research direction.

Alternative Transformer Architectures

Given the explosive nature of this field in the last few years, there have been many methods developed using alternative transformer architectures that could be leveraged for future work. These architectures may be able to provide more effective models for cancer pathways modelling by tackling the characteristics of these pathways (e.g., temporal, long sequences) or improving prediction of key cancer outcomes. By examining what methods have been successfully implemented in EHR data, this can help guide potential approaches for claims data. Below is a list of a few promising developments that could serve as inspiration:

- **Long sequences.** One opportunity is to target the issue of long sequences that can exist in longitudinal health data, as transformers can only accept a certain sequence length (512 or less). Recently, Ma et al. [129] used the Longformer (originally proposed by Beltagy et al. [21]) to model multimodal EHR data encoding temporal and sequential information with temporal cross-attention. The Longformer uses sliding window attention and scales linearly to sequence length to enable the processing of longer sequence lengths [21]. Similarly, Chauhan et al. [34] used cross-attention to model lengthy sequences with transformers in EHR data in their model called Perceiver. The model performed as well as baselines but was about nine times more efficient in terms of computations. Given the continued growth of longitudinal health data, dealing with lengthy sequences is an important area for future work.

- **Time-to-event, survival, and causal modelling.** There are also opportunities for improved time-to-event and causal modelling related to cancer. Researchers from Stanford proposed a transformer-based time-to-event foundation model called Many Outcome Time Oriented Representations (MOTOR) [222] that used EHR and claims data. This model focused on estimating the time until an event occurs and showed better predictions for cancer than other deep learning survival models. In related

work, researchers from the University of Oxford, some of which were involved in development of the BEHRT [119] model, have recently proposed a survival modelling variant of BEHRT called TRisk [185], for predicting all-cause mortality. The authors make a number of modifications to BEHRT for survival modelling, and validate the TRisk model in predicting 36-month mortality for heart failure patients. When analysing risk factors, it was found that cancer diagnoses, even 10 years earlier, provided substantial predictive value, emphasising the impacts of cancer on outcomes for heart failure patients. While this study focused on those with heart failure, it it conceivable it could perform well in predicting survival for those with cancer. Several of the authors from this study were also involved in development of Targeted-BEHRT (or T-BEHRT) [187], designed to investigate causal modelling of the effect of antihypertensive medicines on cancer risk in EHR data. According to this research, T-BEHRT provided better estimates of relative risk than baselines (from ground truth), providing another research avenue for potential exploration.

- **Graph transformers.** Another architecture that could be explored is the Graph Transformer. Poulain and Beheshti [174] proposed GT-BERT, which used temporal embeddings from a graph transformer with BERT for robust patient representations. The model showed improvements on mortality and length of stay prediction over BEHRT [119], CEHR-BEHRT [163], and their previously developed model based on Generative Adversarial Networks (GANs) called CEHR-GAN-BERT [175].

Model Explainability/Interpretability

Given the high stakes nature of cancer (and healthcare), and the potential for decision making to alter patient lives and outcomes, interpretable and explainable models are vital for healthcare. Interpretability and explainability are frequently used interchangeably to describe the ability to interpret or explain model predictions [127]. The study of this is often referred to as eXplainable AI (XAI). There has been growing and varied use of XAI methods with transformer-based models on longitudinal health data. For example, Lahlou et al. [109] used the attention mechanism in the BRLTM model to provide model interpretability. Lentzen et al. [116] used integrated gradients to investigate feature importance in the ExMed-BERT model. Other research has used post-hoc interpretability with meth-

ods such as SHapely Additive exPlanations (SHAP) [54]. These approaches have been commonly used across other transformer models in healthcare [132, 127]. Extending the proposed methods in this research by integrating XAI methods can assist in providing additional transparency and increased trust in deep learning models, particularly among clinicians, and hence are another opportunity for future work.

Application to New Contexts and Diseases

Applying the proposed models in this research to new contexts and diseases can help determine whether the benefits of transformers also carry to other settings. In terms of new contexts, this can include new prediction tasks, populations, or data. There are a number of outcomes that are of interest for cancer patients that were not explored in this research - such as complications or side effects, cancer progression, cancer recurrence, or development of subsequent cancers. Transformers may be able to improve the prediction of these outcomes, particularly via fine-tuning. Recent work by He et al. [81] showed that adjusting the fine-tuning task to better align with the pretrained model objectives can improve prediction of pancreatic cancer. This suggests there are potential benefits in optimising fine-tuning tasks to enhance downstream predictions. In addition, there are also several other tasks that could be explored for evaluating pretrained transformers, such as survival prediction, future hospitalisation risk, or patient deterioration. The use of multiple and varied tasks for fine-tuning would help to understand how a pretrained model performs with other objectives, informing potential foundation or generative models. GenAI can also open up further opportunity for the investigation and testing of multiple scenarios or hypotheses with respect to cancer pathways and care [193], and to better identify complex or rare conditions [178].

Transformers can also be used for exploring predictions for the same task but within different prediction windows. Literature has suggested there is lack of agreement on prediction windows for cancer-related outcomes and recommends that various time windows be investigated [147]. It is also important that transformers be investigated for different cancer populations or cancer types. For example, models could be explored on the priority populations identified in the Australian Cancer Plan [29], or for other common cancers outside of breast and bowel cancer which were explored in Chapter 4. Further-

more, transformers could be evaluated in alternative datasets to enable external validation and address this earlier noted limitation. Exploration with new data (particularly whole-of-population data) also offers opportunity to address other cancer challenges, such as identifying those at increased risk of developing cancer, who may benefit from additional screening interventions.

Finally, while this research focused on the prediction of cancer-related outcomes, other long-term diseases including chronic conditions face many of the same challenges as cancer (i.e., complex, longitudinal, heterogeneous pathways). The transformer-based strategies proposed in this thesis can be leveraged for other conditions to improve the modelling of these pathways and patient predictions.

6.5 Further Challenges and Opportunities

While the above future directions reflect opportunities to improve the proposed transformer models in this research for longitudinal health data, in order to realise the benefits of deep learning approaches, they ultimately need to be applied in the real world. This requires several considerations that are beyond the immediate future directions of this research, but are nonetheless essential for obtaining impact. This section discusses key existing challenges and opportunities related to application of deep learning models to healthcare. It aims to provide guidance and recommendations for researchers and stakeholders to enhance the development and implementation of AI-based predictive models for cancer and healthcare more broadly.

Data Availability

In order to develop deep learning models for real-world use, data needs to be available. Many models, including those in this research, are typically trained and evaluated on retrospective data, that varies in its currency (how recent it is) and size. Cancer registries in Australia (which capture cancer incidence) currently experience long lag times of up to 24 months before official reporting [5]. This is primarily due to quality control and verification processes to confirm cancer diagnoses, however it consequently results in outdated data, trends, and modelling. Ideally, cancer data needs to be current and available for analysis to enable real-time or near-time predictive capability. In addition, cancer data

preferably needs to be large to maximise the benefits of using deep learning models, and have diversity in its populations and cancer types, as this will facilitate more versatile AI models in oncology [23].

It is therefore important to seek opportunities for further linkage of cancer data that is both current and comprehensive. For example, the creation of national linked cancer data in Australia would provide larger and more diverse data for exploring cancer pathways and developing predictive models. This would help overcome the dataset size limitations noted for this research. Furthermore, the inclusion of more detailed clinical data (such as general practice or primary care data) has potential to provide more comprehensive information related to cancer care and can also help overcome the lack of specificity in administrative claims data (particularly for chronic conditions) [35, 63]. Exploration of international data sources with similar healthcare systems to Australia could also potentially support these objectives, as the aggregation of national and international data offers benefits for the development of a global set of predictive variables for cancer studies [173].

In lieu of the above, methods can continue to utilise existing cancer or linked health datasets to demonstrate their potential, and further explore methods such as transfer learning in instances where datasets are small. Data such as PLIDA (used in Chapter 5) or the National Health Data Hub (NHDH)[2] could be further leveraged within Australia to better understand general health and patient pathways at the national level. These data assets can also be linked to cancer-specific data such as cancer registry or patient reported outcomes to enhance cancer research. In addition, recent work has found that data collected up to three months post cancer diagnosis is representative of diagnoses verified 24 months later [5]. This could be explored as a potential approach for early identification of cancer cases to support the development of more up-to-date modelling.

Another avenue that does not require linked data but can still leverage large-scale health information is federated learning. Federated learning provides a machine learning-based framework to learn shared models across sites without requiring direct access to the data [124]. This enables model training on larger and more diverse data and helps overcome data privacy and sharing challenges [124, 1]. Federated learning has been shown

[2] https://www.aihw.gov.au/reports-data/nhdh

to improve the prediction of patient outcomes such as early cancer diagnosis for breast, lung, and colon cancer, and hospitalisation due to heart disease [124, 1]. Federated learning has also been combined with foundation models to create federated foundation models [123]. This has great promise for more personalised care as foundation models that are trained with federated learning can reveal new biomarkers and therapeutic agents [117]. Federated learning thus enhances the application of foundation models and offers scalable solutions that can address growing foundation models with new data and guide more efficient learning [254, 117]. Federated learning also offers opportunity for a more collaborative learning process through leveraging AI Agents, by integrating federated foundation models and collective intelligence in federated intelligence [123]. This supports the development of an intelligent collaborative platform that could share knowledge across organisations or institutions to enhance capability in instances where data sharing is challenging or not feasible [123]. This opportunity could therefore be explored to aid in development of AI models across the healthcare system.

Ethics and Regulation

When developing deep learning models in healthcare, it is essential to consider the ethical and regulatory challenges and landscape. The use of AI models pose a number of ethical issues - including awareness of use, consent to AI use, bias and fairness, transparency, accountability etc. Protecting the privacy and security of healthcare data is also critical for AI model development and use [257], with specific guidance released in Australia on privacy and developing and training generative AI models[3]. Australia's AI Ethics Principles[4] detail the key ethics principles to consider to ensure that AI is safe, secure, and reliable. These ethics principles have been the basis for much of the guidance in Australia in the last few years to support the responsible use of AI. For example, the principles align with the Voluntary AI Safety Standard[5], which guides safe and responsible use of AI in Australia through ten AI guardrails. Similarly, the Pilot Assurance Framework[6] facilitates

[3] https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-developing-and-training-generative-ai-models

[4] https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles/australias-ai-ethics-principles

[5] https://www.industry.gov.au/publications/voluntary-ai-safety-standard

[6] https://www.digital.gov.au/policy/ai/pilot-ai-assurance-framework

impact assessment of AI-based opportunities against the ethics principles. The assurance framework should also be applied together with the Policy for responsible use of AI in government and the recently released Technical Standards for government's use of AI[7].

As models are developed and implemented into healthcare, they should comply with relevant guidance, principles, and regulations. For example, if foundation models are classified as Software-as-a-Medical-Device [243], they would need to consider Australia's regulation of medical device software including AI[8]. Those working across AI in healthcare in Australia will need to be aware of, and understand, obligations, as AI regulation and standards in healthcare is an evolving landscape [257]. There are already indications that additional regulation is needed for AI, as documented in the Safe and Responsible AI in Health Care - Legislation and Regulation Review[9]. It will therefore be important to maintain awareness and currency in this expanding space. While regulation does offer opportunity for more clarity and standardisation of AI models in healthcare, it needs to be balanced to support innovation.

One opportunity to offer a more balanced approach could be to adjust requirements based on progress of AI system development. This could include less stringent requirements on exploratory or proof-of-concept AI approaches to encourage experimentation. Then, once AI methods have been evaluated and determined to offer potential benefits, additional requirements could be incorporated that ensure compliance before real-world use or larger roll-out. This would facilitate the ability to more rapidly and efficiently assess potential AI methods while not increasing risk, which aligns with existing AI policies in Australia that encourage early-stage experimentation to embrace AI opportunities and accelerate AI adoption[10]. In addition, targeted pilot projects and studies with longitudinal health data are critical to evaluate potential real-world effectiveness, sustainability, and scalability of AI systems in healthcare [45, 57]. Enhanced AI experimentation can also help encourage implementation of AI research into practice. Literature suggests that majority of existing AI research in healthcare is at the preclinical stage and does not have

[7] https://www.digital.gov.au/policy/ai

[8] https://www.tga.gov.au/news/news/tga-ai-review-outcomes-report-published

[9] https://www.tga.gov.au/sites/default/files/2025-07/safe-and-responsible-artificial-intelligence-in-health-care-legislation-and-regulation-review-final-report.pdf

[10] https://www.digital.gov.au/policy/ai

real-world validation, limiting its likelihood for impact [45]. This emphasises the need for exploration projects and small-scale implementation to demonstrate likely benefits before investment into wide-scale application or deployment.

Transparency and Trust

Transparency and trust of predictive models are critical in healthcare and are an important part of ethical AI. Trustworthiness and reliability of deep learning models is a concern [127], particularly given their ”black-box” nature. Literature indicates that oncologists and pathologists have scepticism about the insights from AI models [257]. Healthcare professionals also have concerns related to legal liability and accountability in the instance of clinical errors where AI is used [82]. This is a challenging issue as it has been argued that placing sole responsibility on clinicians for AI-based decisions may not be appropriate, as clinicians do not have control over an AI model or its recommendations [143]. Nevertheless, health practitioners must continue to meet professional obligations when using AI in Australia[11], necessitating the need for transparency and trust in AI for healthcare

Providing explainable models through XAI approaches can help provide model transparency, build trust, and identify potential model flaws [132]. It can also help identify predictors for outcomes that can be compared to existing knowledge or may reveal new unknown risk factors. A recent study using the Transformer with XAI for lung cancer prediction indicated that their explainability approach discovered a number of symptoms, co-morbid conditions, and procedures related to lung cancer diagnosis [236]. However, research has indicated there is no agreed definition of 'explainability', no standard best practice approaches, and existing methods each have limitations [132]. For example, attention weights may not necessarily reflect feature importance, and SHAP can be computationally infeasible [127]. While there are a range of other approaches such as counterfactual or perturbation explanations, or visualisation via heatmaps, the XAI field is still immature in healthcare [127]. It will be important to continue observation of XAI development, and as the field continues to grow, it is expected that explainability for transformers will also improve.

[11] https://www.ahpra.gov.au/Resources/Artificial-Intelligence-in-healthcare.aspx

In addition, it will be essential to work with the end user (such as clinicians and policymakers) to understand their needs and develop their AI literacy in terms of elements related to XAI of deep learning models. This would align with recommendations for explainability to be based both on the model and the requirements of the end user (clinicians, researcher, policymaker) to understand the prediction rationale [127]. Seneviratne et al. [204] suggest that a key issue related to trust and safety is not necessary interpretability (which is often blamed for lack of trust), but is rather the lack of evidence to demonstrate the safety and effectiveness in the real-world. The authors draw a comparison to medicines, where the biological mechanism may not be fully understood, but studies have verified the safety and efficacy. This highlights the importance of considering the needs and requirements of the decision maker in determining appropriate explainability and fostering trust.

Real-World Validation and Evaluation

Throughout model experimentation and development, real-world validation and evaluation should be considered. Validation refers to the process of determining whether a model is adequate for its intended use and is a broader definition of validation that encompasses model bias, fairness, and generalisability [201]. Models that are intended for real-world use in healthcare (such as for clinical decision support or population health management) need to be assessed for bias and fairness to ensure they do not carry through existing biases from the training data [257]. Left unaddressed, this can lead to inaccurate predictions, which could exacerbate healthcare disparities for gender groups or underrepresented racial, ethnic, socioeconomic, or remote populations [257]. This has already been demonstrated in claims data where a model over-predicted readmission across race and gender subgroups [109].

Resolving issues of bias and fairness is not as simple as removing features from models. This was shown in a study that investigated racial and ethnic bias in prediction models for colorectal cancer recurrence, which found greater predictive performance and algorithm fairness when these features were included in models [102]. It also found that removing information related to ethnicity led to poorer predictions. Evaluating and addressing bias and fairness is thus crucial for deep learning predictive models. Hazanzadeh

et al. [80] provide a detailed overview of bias in healthcare applications and suggest mitigations across all aspects of the model lifecycle. Cross and colleagues [49] also provide a detailed overview of bias in medical AI, summarising the implications and suggested solutions.

As mentioned previously, generalisability of models is also important. Models need to be validated in diverse data and across multiple institutions to ensure robustness and consistency across clinical settings [257, 23]. Models also should be compared to, and validated against, the existing methods in current practice to understand whether models can provide performance on par, or above, the status quo. For cancer, this could include comparing against an existing clinically utilised cancer prognostic tools. Existing research has suggested that much of the AI research in oncology is experimental, without clinical validation [130]. This indicates a gap in our understanding of how AI can actually impact patient outcomes [44, 125]. Therefore, real-world studies such as controlled clinical trials are needed to examine and evaluate the use of AI in healthcare and better understand its potential real-world benefits in terms of clinically useful metrics [208, 125]. Extensive validation will assist in building confidence and credibility into predictive models, which will enhance their likelihood for real-world use.

Deployment and Monitoring

In addition to validation and evaluation, models need to be capable of being deployed in applicable settings and monitored over time. It is not straightforward to integrate models into existing clinical workflows and systems, as this requires infrastructure, experts, and processes that can provide actionable outputs of models to end users (e.g., clinicians, policymakers) for consideration [125]. Furthermore, performance of a prediction model can deteriorate over time due to data drift (or data shift), where the training data differs from the data the model is applied to in the real world [200]. In healthcare, this can occur due to a range of reasons such as changes in populations, disease epidemiology, or care practices or policies [111, 200]. Due to healthcare data typically being smaller and noisier than other data used for deep learning, as well as the risk poor models could pose to health outcomes, is it essential to implement monitoring for models to ensure continuous effectiveness and safety [65].

To facilitate model monitoring, examining the data, evaluating susceptibility to data drift, and implementing procedures for automatic model retraining can be investigated. For example, monitoring can look at whether there are changes to the predictor variable, the input variables, and/or the relationships between the two [65]. For managing data drift, transfer learning can be leveraged via pretrained or foundation models. Recent research in EHR data has suggested that transformer and recurrent foundation models are more robust than logistic regression against temporal distribution shift [74]. This study also found that foundation models based on transformers continued to show improvements as pretraining data increased. In further support of this, a recent study used transfer learning to help maintain model performance during monitoring. Lam and colleagues [111] developed and deployed a prediction model for respiratory failure for patients in the hospital ICU. They used a predetermined change control plan for model monitoring, requiring a model to fine-tune when performance dropped below a predetermined threshold. This provides another example of the potential for transfer learning to support model robustness and performance.

Despite these use cases, the literature has suggested that methods for updating clinical models are only in their infancy, and there are limited studies evaluating model monitoring in medical data [111]. In addition, there are challenges specific to clinical AI models such as long-term gaps between predictions and outcomes (e.g., mortality or cancer recurrence) and AI-induced confounding, where clinicians change their care plan based on the model prediction [65]. Further research is therefore needed in this space to help inform and design appropriate monitoring techniques, and ensure models remain effective and reliable after deployment.

Translation and Implementation Science

Finally, it is important to consider translation and implementation science, as these can support the generation of impact from deep learning models in healthcare. While much research has demonstrated the benefits of deep learning, implementation into practice has, in general, lagged behind other fields [131]. Both the use and likely benefits of deep learning in healthcare is still not well understood [193]. Leveraging frameworks related to implementation science can facilitate translation and help realise the potential of AI.

Implementation science is a field of research that is focused on the development and testing of strategies that help translate evidence into practice [87]. In healthcare, it aims to enhance the adoption and implementation of evidence into clinical and public health practice [87].

One well-known and utilised framework is the Consolidated Framework for Implementation Research (CFIR), which can be used to understand the barriers and facilitators of an innovation. Many of the reported barriers align with those discussed earlier in this section (i.e., data, trust, ethics etc) [42], however studies leveraging frameworks from implementation science have identified additional barriers and facilitators. For example, Swart et al. [225] used CFIR to develop an implementation science approach to enhance models for radiotherapy. The study found that knowledge and education of AI concepts, trust in AI systems, a multidisciplinary approach, and stakeholder involvement were facilitators for implementation. Preti et al. [177] also used CFIR in a systematic review of studies related to implementation of machine learning in healthcare. They found that organisational culture and early engagement of stakeholders was important for implementation of machine learning models.

Other research has suggested tailored frameworks or the use of other frameworks for implementation of AI into healthcare. For example, Adnan et al. [2] proposed their own framework, called the Health xAI Implementation Framework, to support implementation of AI in healthcare. Reddy [193] proposes the use of two alternative frameworks to help transition healthcare to leverage AI. This paper suggests that factors such as perception, attitudes, and intent to use can affect implementation, along with social influence from regulations, ethical principles, and overall healthcare policy [193]. Clinician and patient perception of what is AI has also been identified in another study [131]. Furthermore, a recent study exploring the concerns of medical AI for health care professionals suggested that extensive education, regulation, and separation of roles is necessary to address current resistance [13]. All these noted barriers and facilitators indicate the complexity of applying AI in healthcare and the need for implementation science.

Despite the reference of varying frameworks, what all of the above have in common is the importance of transdisciplinary collaboration and involvement. Collaboration be-

tween machine learning developers and stakeholders has been identified as key for implementation, and stakeholders should be involved from initiation and throughout model development [131]. It is therefore encouraged that AI researchers, government, healthcare stakeholders, and implementation experts, engage in collaborations to help bridge the translational gap between AI and healthcare. While this research contributes to this space through cross-disciplinary stakeholder collaboration, continued industry-based research, and a deeper focus on implementation science are important next steps [190, 122, 257, 8].

6.6 Concluding Remarks

This research focused on the development of more effective transformers for cancer pathways modelling in longitudinal health data. Through experimental studies, the advantages of leveraging multi-task learning, longitudinal patient modelling of claims data, and transfer learning are demonstrated with transformers to address the challenges of modelling cancer pathways. The approaches proposed not only enhance the accuracy of predicting cancer-related outcomes, but provide more flexible and data-efficient models for a more holistic understanding of cancer in longitudinal health data. This supports the development of a comprehensive cancer pathways modelling framework that can benefit the evolution towards large-scale generalised foundation models and generative systems to further improve healthcare. Overall, this work confirms the value of a longitudinal health transformer in capturing the complex relationships between patient and healthcare system, offering a promising basis for advancing the modelling of cancer and other chronic diseases to improve patient care and outcomes.

Appendix A

Supplementary Material

Chapter 5: Pretrained Transformer for Claims Data to Predict Chronic Conditions for Cancer Patients with Limited Data

A.1 Results of the Pretraining Objectives

We report the results of the pretraining objectives for each of the strategies in Table A.1, and the training performance for the MCM objective in Fig. A.1. While we note the best MCM precision may appear low (being 49.4%), the score is not far below the precision of the BEHRT model (54.6%) on EHR data with 1916 possible diagnoses [199]. Our method must predict from 3939 possible claims items, making it an even more complex pretraining task. In addition, the performance of the MCM objective during pretraining (see Fig. A.1) continues to increase over the 50 epochs, suggesting that further performance gains may be possible with additional model training. For the pretraining strategies that use the future claims items objectives, we find that the LMP-based models achieve higher accuracy than training via NMP, likely reflecting the simpler prediction task.

Table A.1 : Results of the objectives for each pretraining strategy, presented as average precision or accuracy @ top k.

Pretraining	Prec.	Acc. @ top k for NML/LMP					Prec. for age
Strategy	for MCM	k=1	k=5	k=10	k=15	k=20	group prediction
MCM	**49.4**	-	-	-	-	-	-
MCM+Age	48.7	-	-	-	-	-	**95.7**
MCM+LMP	47.4	35.0	83.1	96.1	98.8	99.5	-
LMP	-	**36.1**	**84.3**	**96.8**	**99.2**	**99.7**	-
NMP	-	26.2	71.4	88.0	93.6	96.0	-

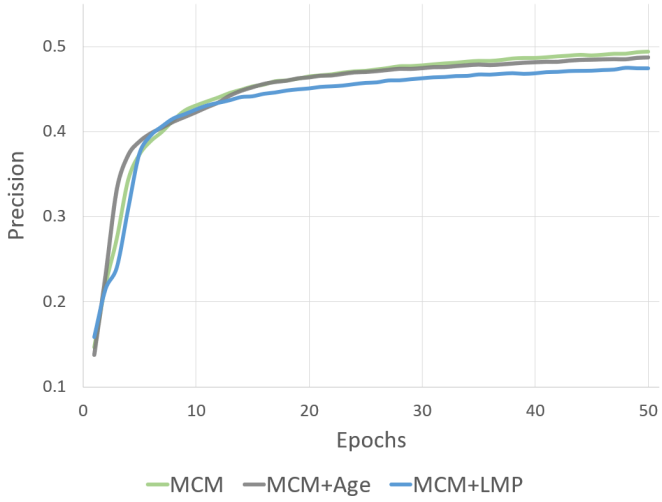

Figure A.1 : Performance during training for the MCM Pretraining Objective.

A.2 Additional Performance Metrics for Pretraining Strategies

The precision, recall, and accuracy for the multimorbidity tasks are reported in Table A.2.

Table A.2 : Precision, recall and accuracy metrics for the multimorbidity (MM) task on the NHS cohort (MM), NHS cohort 65+ (MM 65+) and the NHS cohort 65+ with cancer or mental health conditions (MM 65+ w/ Can or MH). Performance metrics are average (standard deviation).

Pretraining	MM			MM 65+			MM 65+ w/ Can or MH		
Strategy	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.
MCM	88.4 (3.6)	61.4 (1.6)	93.1 (0.6)	81.8 (10.0)	70.5 (4.4)	81.4 (2.5)	96.9 (3.8)	71.4 (9.0)	85.7 (2.6)
MCM+Age	88.6 (0.6)	52.1 (3.0)	91.8 (0.5)	94.4 (2.1)	58.8 (5.4)	82.1 (1.5)	97.5 (4.3)	78.3 (6.1)	89.1 (2.8)
MCM+LMP	91.7 (1.3)	52.9 (4.1)	92.3 (0.5)	98.4 (0.6)	53.1 (6.4)	80.9 (2.5)	99.3 (1.2)	63.6 (3.4)	83.1 (1.5)
LMP	92.8 (3.7)	53.5 (3.0)	92.4 (0.3)	98.3 (1.5)	50.2 (3.4)	79.7 (1.2)	94.3 (5.7)	69.2 (11.1)	83.7 (2.8)
NMP	97.0 (1.1)	52.0 (3.5)	92.6 (0.6)	99.6 (0.6)	53.4 (6.4)	81.3 (2.5)	100.0 (0.0)	67.9 (4.6)	85.3 (2.1)

A.3 Standard Deviations for Single Condition Predictions

Tables A.3 and A.4 show the standard deviation of the F1 and AUC performance metrics for the single condition prediction tasks.

Table A.3 : Standard deviations of F1 and AUC metrics on single condition predictions for the cancer cohort. Results are average.

Model	Diab		HC		HT		HSVD		BP		MH	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
CF	(1.8)	(1.5)	(16.4)	(7.9)	(4.9)	(4.3)	(4.3)	(3.4)	(6.7)	(3.8)	(6.4)	(3.1)
CF+Age	(2.0)	(1.6)	(8.6)	(3.2)	(6.9)	(5.4)	(14.8)	(7.1)	(3.3)	(1.6)	(7.3)	(2.6)
Naive T	(8.9)	(2.2)	(11.1)	(7.4)	(4.7)	(5.3)	(7.9)	(4.9)	(25.4)	(1.2)	(1.6)	(1.1)
GB	(2.8)	(1.8)	(0.8)	(0.5)	(2.6)	(1.9)	(2.3)	(1.2)	(0.5)	(0.4)	(3.0)	(1.2)
LR	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
RF	(2.5)	(1.5)	(3.4)	(0.9)	(3.1)	(1.9)	(2.7)	(1.0)	(4.8)	(1.4)	(0.0)	(0.0)

Table A.4 : Standard deviations of F1 and AUC metrics on single condition predictions for the mental health cohort. Results are average.

Model	Diab		HC		HT		HSVD		BP		Can	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
CF	(0.0)	(0.0)	(11.7)	(7.2)	(3.4)	(5.9)	(8.3)	(5.9)	(8.7)	(4.1)	(9.6)	(5.7)
CF+Age	(5.2)	(4.8)	(7.1)	(3.0)	(9.5)	(8.1)	(9.4)	(5.1)	(3.5)	(2.4)	(11.1)	(3.9)
Naive T	(3.3)	(2.1)	(4.7)	(8.9)	(37.5)	(0.0)	(8.3)	(8.8)	(0.0)	(0.0)	(26.9)	(0.0)
GB	(0.0)	(0.0)	(0.0)	(0.0)	(1.2)	(0.8)	(0.0)	(0.0)	(7.1)	(2.4)	(3.3)	(1.2)
LR	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
RF	(5.1)	(3.1)	(0.0)	(0.0)	(2.8)	(1.5)	(4.4)	(1.3)	(5.1)	(1.7)	(8.2)	(2.4)

A.4 Additional Performance Metrics for Single Condition Predictions

Tables A.5 and A.6 show the precision, recall and accuracy metrics for the single condition prediction tasks.

Table A.5 : Precision, recall and accuracy metrics on single condition predictions for the cancer cohort. Performance metrics are average (standard deviation).

Model	Diab			HC			HT		
	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.
CF	100.0 (0.0)	80.7 (3.1)	96.8 (0.5)	78.3 (19.3)	47.5 (23.6)	78.8 (1.8)	100.0 (0.0)	86.4 (8.8)	93.5 (4.2)
CF+Age	100.0 (0.0)	77.2 (3.0)	96.2 (0.5)	88.8 (9.8)	26.3 (7.6)	77.3 (1.4)	100.0 (0.0)	80.9 (10.9)	90.9 (5.2)
Naive T	80.0 (17.4)	91.2 (3.0)	94.1 (3.6)	48.8 (4.0)	57.6 (19.9)	70.5 (2.3)	75.8 (5.8)	92.0 (2.8)	82.0 (5.4)
GB	71.0 (2.5)	86.0 (3.1)	91.8 (1.0)	75.8 (1.3)	53.5 (1.7)	81.4 (0.0)	92.5 (0.4)	68.5 (3.7)	82.3 (1.8)
LR	100.0 (0.0)	73.7 (0.0)	95.6 (0.0)	78.6 (0.0)	33.3 (0.0)	77.9 (0.0)	97.0 (0.0)	59.3 (0.0)	79.6 (0.0)
RF	100.0 (0.0)	56.1 (3.1)	92.7 (0.5)	66.7 (57.7)	2.0 (1.7)	71.4 (0.5)	100.0 (0.0)	58.0 (3.9)	79.9 (1.8)
	HSVD			BP			MH		
	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.
CF	93.9 (10.5)	53.6 (9.5)	87.4 (0.5)	76.4 (12.5)	47.2 (7.4)	78.2 (3.6)	71.8 (28.0)	37.5 (11.0)	81.6 (4.6)
CF+Age	95.6 (3.9)	41.7 (14.9)	85.1 (3.1)	95.2 (8.3)	33.3 (2.8)	78.2 (1.3)	91.7 (14.4)	15.3 (4.8)	81.9 (1.3)
Naive T	66.4 (11.9)	52.4 (16.1)	80.7 (3.5)	32.9 (33.4)	35.2 (56.2)	56.3 (21.2)	39.3 (2.4)	43.1 (2.4)	74.0 (1.3)
GB	91.5 (3.5)	51.2 (2.1)	86.8 (0.9)	79.0 (3.7)	27.8 (0.0)	74.6 (0.5)	100.0 (0.0)	26.4 (2.4)	84.5 (0.5)
LR	92.3 (0.0)	42.9 (0.0)	85.1 (0.0)	82.4 (0.0)	38.9 (0.0)	77.9 (0.0)	100.0 (0.0)	25.0 (0.0)	84.2 (0.0)
RF	100.0 (0.0)	22.6 (2.1)	84.3 (6.3)	100.0 (0.0)	8.3 (2.8)	70.8 (0.9)	0.0 (0.0)	0.0 (0.0)	78.9 (0.0)

Table A.6 : Precision, recall and accuracy metrics on single condition predictions for the mental health cohort. Performance metrics are average (standard deviation).

Model	Diab			HC			HT		
	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.
CF	100.0 (0.0)	87.5 (0.0)	97.5 (0.0)	85.4 (6.0)	55.1 (13.5)	82.3 (5.1)	81.5 (17.2)	83.3 (10.9)	80.6 (6.5)
CF+Age	100.0 (0.0)	89.6 (9.6)	97.9 (1.9)	100.0 (0.0)	29.5 (5.9)	76.8 (1.9)	94.4 (9.6)	76.3 (9.1)	86.5 (8.0)
Naive T	97.8 (3.9)	89.6 (3.6)	97.5 (1.3)	36.3 (14.2)	83.3 (25.6)	52.3 (18.0)	32.1 (27.8)	66.7 (57.7)	49.4 (2.2)
GB	85.7 (0.0)	75.0 (0.0)	92.4 (0.0)	100.0 (0.0)	23.1 (0.0)	74.7 (0.0)	100.0 (0.0)	59.6 (1.5)	80.6 (0.8)
LR	100.0 (0.0)	75.0 (0.0)	94.9 (0.0)	92.3 (0.0)	46.2 (0.0)	81.0 (0.0)	100.0 (0.0)	55.3 (0.0)	78.5 (0.0)
RF	100.0 (0.0)	56.2 (6.3)	91.1 (1.3)	100.0 (0.0)	7.7 (0.0)	69.6 (0.0)	100.0 (0.0)	46.5 (3.1)	74.2 (1.4)
	HSVD			BP			Can		
	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.	Pre.	Rec.	Acc.
CF	79.9 (18.2)	55.6 (14.5)	83.8 (3.3)	100.0 (0.0)	37.6 (8.2)	75.5 (3.2)	74.8 (22.8)	45.8 (4.2)	78.1 (6.3)
CF+Age	96.7 (5.8)	46.0 (9.9)	85.4 (2.8)	90.0 (17.3)	23.7 (1.8)	68.8 (2.7)	95.8 (7.2)	19.5 (8.7)	75.1 (1.9)
Naive T	46.4 (17.4)	68.2 (27.9)	62.1 (24.8)	0.0 (0.0)	0.0 (0.0)	60.8 (0.0)	10.1 (17.6)	33.3 (57.7)	56.5 (22.6)
GB	100.0 (0.0)	42.9 (0.0)	85.0 (0.0)	100.0 (0.0)	17.2 (4.9)	67.5 (1.9)	81.1 (1.9)	18.1 (2.4)	73.8 (0.8)
LR	100.0 (0.0)	33.3 (0.0)	82.5 (0.0)	88.9 (0.0)	25.8 (0.0)	69.6 (0.0)	100.0 (0.0)	33.3 (0.0)	79.7 (0.0)
RF	100.0 (0.0)	11.1 (2.8)	76.6 (0.8)	100.0 (0.0)	12.9 (3.2)	65.8 (1.3)	100.0 (0.0)	9.7 (4.8)	72.6 (1.4)

Bibliography

[1] Abbas, S.R., Abbas, Z., Zahir, A., Lee, S.W.: Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with iot integration. In: Healthcare. vol. 12, p. 2587. MDPI (2024)

[2] Adnan, H.S., Shidani, A., Clifton, L., Bankhead, C.R., Perera-Salazar, R.: Implementation framework for AI deployment at scale in healthcare systems. iScience **28**(5), 112406 (May 2025)

[3] Afshar, N., English, D., Blakely, T., Thursfield, V., Farrugia, H., Giles, G., Milne, R.: Differences in cancer survival by area-level socio-economic disadvantage: A population-based study using cancer registry data. PLOS ONE **15**, e0228551 (01 2020). https://doi.org/10.1371/journal.pone.0228551

[4] Ahmed, M.I., Spooner, B., Isherwood, J., Lane, M., Orrock, E., Dennison, A.: A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus **15**(10), e46454 (Oct 2023)

[5] Ahmed, M., Walton, R., Creighton, N., Gugusheff, J., Saini, N., Moritz, P., Roder, D.: Innovative use of australian cancer registry data for early detection of the effects of epidemics and other mass disruptions on cancer incidence. Cancer Epidemiol. **91**(102608) (Aug 2024)

[6] Alboaneen, D., Alqarni, R., Alqahtani, S., Alrashidi, M., Alhuda, R., Alyahyan, E., Alshammari, T.: Predicting colorectal cancer using machine and deep learning algorithms: Challenges and opportunities. Big Data Cogn. Comput. **7**(2), 74 (Apr 2023)

[7] Ali, O., Abdelbaki, W., Shrestha, A., Elbasi, E., Alryalat, M.A.A., Dwivedi, Y.K.: A systematic literature review of artificial intelligence in the healthcare sector:

Benefits, challenges, methodologies, and functionalities. Journal of Innovation Knowledge **8**(1) (2023). https://doi.org/https://doi.org/10.1016/j.jik.2023.100333

[8] Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A., Al Yami, M.S., Al Harbi, S., Albekairy, A.M.: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med. Educ. **23**(1), 689 (Sep 2023)

[9] Altena, R., Hubbert, L., Kiani, N., Wengström, Y., Bergh, J., Hedayati, E.: Evidence-based prediction and prevention of cardiovascular morbidity in adults treated for cancer. Cardio-Oncology **7** (05 2021)

[10] Amirahmadi, A., Ohlsson, M., Etminani, K.: Deep learning prediction models based on EHR trajectories: A systematic review. J. Biomed. Inform. **144**(104430), 104430 (Aug 2023)

[11] Arnold, C., Biedebach, L., Küpfer, A., Neunhoeffer, M.: The role of hyperparameters in machine learning models and how to tune them. Polit. Sci. Res. Meth. **12**(4), 841–848 (Oct 2024)

[12] Artsi, Y., Sorin, V., Glicksberg, B.S., Korfiatis, P., Freeman, R., Nadkarni, G.N., Klang, E.: Challenges of implementing LLMs in clinical practice: Perspectives. J. Clin. Med. **14**(17), 6169 (Sep 2025)

[13] Arvai, N., Katonai, G., Mesko, B.: Health care professionals' concerns about medical AI and psychological barriers and strategies for successful implementation: Scoping review. J. Med. Internet Res. **27** (Apr 2025)

[14] Australian Health Ministers' Advisory Council: National strategic framework for chronic conditions. https://www.health.gov.au/resources/publications/national-strategic-framework-for-chronic-conditions?language=en (2017)

[15] Australian Institute of Health and Welfare: Cancer data in australia. https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/contents/overview (2024)

[16] Ayala Solares, J.R., Diletta Raimondi, F.E., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., Pinho Gomes, A.C., Payberah, A.H., Zottoli, M., Nazarzadeh, M., Conrad, N., Rahimi, K., Salimi-Khorshidi, G.: Deep learning for electronic health records: A comparative review of multiple deep neural architectures. Journal of Biomedical Informatics **101**, 103337 (2020)

[17] Bajwa, J., Munir, U., Nori, A., Williams, B.: Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthcare Journal **8**, e188–e194 (07 2021). https://doi.org/10.7861/fhj.2021-0095

[18] Balagopalan, A., Eyre, B., Rudzicz, F., Novikova, J.: To BERT or not to BERT: comparing speech and language-based approaches for alzheimer's disease detection. CoRR **abs/2008.01551** (2020), https://arxiv.org/abs/2008.01551

[19] Bandyopadhyay, A., Albashayreh, A., Zeinali, N., Fan, W., Gilbertson-White, S.: Using real-world electronic health record data to predict the development of 12 cancer-related symptoms in the context of multimorbidity. JAMIA Open **7**(3), ooae082 (09 2024). https://doi.org/10.1093/jamiaopen/ooae082

[20] Bayoudh, K., Knani, R., Hamdaoui, F., Abdellatif, M.: A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer (06 2021). https://doi.org/10.1007/s00371-021-02166-7

[21] Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. CoRR **abs/2004.05150** (2020), https://arxiv.org/abs/2004.05150

[22] Bengio, Y., Courville, A.C., Vincent, P.: Representation learning: A review and new perspectives. CoRR **abs/1206.5538** (2012), http://arxiv.org/abs/1206.5538

[23] Bilal, M., Hamza, A., Malik, N.: Nlp for analyzing electronic health records and clinical notes in cancer research: A review. Journal of Pain and Symptom Management **69**(5), e374–e394 (May 2025). https://doi.org/10.1016/j.jpainsymman.2025.01.019

[24] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R.B., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N.S., Chen, A.S., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L.E., Goel, K., Goodman, N.D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M.S., Krishna, R., Kuditipudi, R., et al.: On the opportunities and risks of foundation models. CoRR **abs/2108.07258** (2021), https://arxiv.org/abs/2108.07258

[25] Borg, S.J., Borg, D.N., Foster, M.M., Bell, R., Bowley, J., Geraghty, T.: Use and cost of medicare benefits schedule and pharmaceutical benefits scheme services following inpatient rehabilitation for acquired disability in australia. Aust. Health Rev. **47**(2), 165–174 (Apr 2023)

[26] Brookhart, M., Stürmer, T., Glynn, R., Rassen, J., Schneeweiss, S.: Confounding control in healthcare database research. Medical care **48**, S114–20 (2010)

[27] Brunner-La Rocca, H.P., Peden, C., Soong, J., Holman, P., Bogdanovskaya, M., Barclay, L.: Reasons for readmission after hospital discharge in patients with chronic diseases—information from an international dataset. PLOS ONE **15**, e0233457 (06 2020)

[28] Cai, X., Gao, J., Ngiam, K.Y., Ooi, B.C., Zhang, Y., Yuan, X.: Medical concept embedding with time-aware attention. In: IJCAI. pp. 3984–3990 (2018)

[29] Cancer Australia: Australian cancer plan. https://www.australiancancerplan.gov.au/ (2025)

[30] Carrasco-Ribelles, L.A., Llanes-Jurado, J., Gallego-Moll, C., Cabrera-Bean, M., Monteagudo-Zaragoza, M., Violán, C., Zabaleta-del Olmo, E.: Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review. Journal of the American Medical Informatics

Association **30**(12), 2072–2082 (09 2023). https://doi.org/10.1093/jamia/ocad168, https://doi.org/10.1093/jamia/ocad168

[31] Cascarano, A., Mur-Petit, J., Hernández-González, J., Camacho, M., de Toro Eadie, N., Gkontra, P., Chadeau-Hyam, M., Vitrià, J., Lekadir, K.: Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artif. Intell. Rev. **56**(S2), 1711–1771 (Nov 2023)

[32] Sainz de Cea, M.V., Diedrich, K., Bakalo, R., Ness, L., Richmond, D.: Multi-task learning for detection and classification of cancer in screening mammography. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 241–250. Springer International Publishing, Cham (2020)

[33] Chan, T.H., Yin, G., Bae, K., Yu, L.: Multi-task heterogeneous graph learning on electronic health records (2024), https://arxiv.org/abs/2408.07569

[34] Chauhan, V.K., Thakur, A., O'Donoghue, O., Rohanian, O., Molaei, S., Clifton, D.A.: Continuous patient state attention model for addressing irregularity in electronic health records. BMC Med. Inform. Decis. Mak. **24**(1), 117 (May 2024)

[35] Cheah, R., Canaway, R., Hallinan, C.M., de Mendonça, L., Manski-Nankervis, J.A.: Using primary care data for research: What are the issues and potential solutions? Aust. J. Gen. Pract. **53**(6), 408–411 (Jun 2024)

[36] Chen, H.Y., Wang, H.M., Lin, C.H., Yang, R., Lee, C.C.: Lung cancer prediction using electronic claims records: A transformer-based approach. IEEE J. Biomed. Health Inform. **27**(12), 6062–6073 (Dec 2023)

[37] Chen, Y., Esmaeilzadeh, P.: Generative AI in medical practice: In-depth exploration of privacy and security challenges. J. Med. Internet Res. **26**, e53008 (Mar 2024)

[38] Chen, Y., Lehmann, C.U., Malin, B.: Digital information ecosystems in modern care coordination and patient care pathways and the challenges and opportunities

for AI solutions. J. Med. Internet Res. **26**, e60258 (Dec 2024)

[39] Cheng, C.H., Shi, S.S.: Artificial intelligence in cancer: applications, challenges, and future perspectives. Mol. Cancer **24**(1), 274 (Oct 2025)

[40] Choi, E., Bahadori, M.T., Sun, J.: Doctor AI: predicting clinical events via recurrent neural networks. CoRR **abs/1511.05942** (2015), http://arxiv.org/abs/1511.05942

[41] Choi, E., Bahadori, M.T., Sun, J., Kulas, J., Schuetz, A., Stewart, W.: Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In: NeurIPS. pp. 3504–3512 (2016)

[42] Chomutare, T., Tejedor, M., Svenning, T.O., Marco-Ruiz, L., Tayefi, M., Lind, K., Godtliebsen, F., Moen, A., Ismail, L., Makhlysheva, A., Ngo, P.D.: Artificial intelligence implementation in healthcare: A theory-based scoping review of barriers and facilitators. Int. J. Environ. Res. Public Health **19**(23) (Dec 2022)

[43] Chu, C., Lee, N., Adeoye, J., homson, P., Choi, S.W.: Machine learning and treatment outcome prediction for oral cancer. Journal of Oral Pathology & Medicine **49** (08 2020)

[44] Chua, I.S., Gaziel-Yablowitz, M., Korach, Z.T., Kehl, K.L., Levitan, N.A., Arriaga, Y.E., Jackson, G.P., Bates, D.W., Hassett, M.: Artificial intelligence in oncology: Path to implementation. Cancer Med. **10**(12), 4138–4149 (Jun 2021)

[45] Chustecki, M.: Benefits and risks of AI in health care: Narrative review. Interact. J. Med. Res. **13**, e53616 (Nov 2024)

[46] Considine, J., Fox, K., Plunkett, D., Mecner, M., O Reilly, M., Darzins, P.: Factors associated with unplanned readmissions in a major australian health service. Aust. Health Rev. **43**(1), 1–9 (Feb 2019)

[47] Coupland, H., Scheidwasser, N., Katsiferis, A., Davies, M., Flaxman, S., Hulvej Rod, N., Mishra, S., Bhatt, S., Unwin, H.J.T.: Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. BMC Public Health **25**(1), 1520 (Apr 2025)

[48] Crawshaw, M.: Multi-task learning with deep neural networks: A survey (2020), https://arxiv.org/abs/2009.09796

[49] Cross, J.L., Choma, M.A., Onofrey, J.A.: Bias in medical AI: Implications for clinical decision-making. PLOS Digit. Health **3**(11) (Nov 2024)

[50] Denaxas, S., Gonzalez-Izquierdo, A., Direk, K., Fitzpatrick, N.K., Fatemifar, G., Banerjee, A., Dobson, R.J.B., Howe, L.J., Kuan, V., Lumbers, R.T., Pasea, L., Patel, R.S., Shah, A.D., Hingorani, A.D., Sudlow, C., Hemingway, H.: UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J. Am. Med. Inform. Assoc. **26**(12), 1545–1559 (Dec 2019)

[51] Denecke, K., May, R., Rivera-Romero, O.: Transformer models in healthcare: A survey and thematic analysis of potentials, shortcomings and risks. J. Med. Syst. **48**(1), 23 (Feb 2024)

[52] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)

[53] Dharmarajan, K., Hsieh, A.F., Lin, Z., Bueno, H., Ross, J.S., Horwitz, L.I., Barreto-Filho, J.A., Kim, N., Bernheim, S.M., Suter, L.G., Drye, E.E., Krumholz, H.M.: Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. JAMA **309**(4), 355–363 (Jan 2013)

[54] Dick, K., Kaczmarek, E., Ducharme, R., Bowie, A.C., Dingwall-Harvey, A.L.J., Howley, H., Hawken, S., Walker, M.C., Armour, C.M.: Transformer-based deep learning ensemble framework predicts autism spectrum disorder using health administrative and birth registry data. Sci. Rep. **15**(1), 11816 (Apr 2025)

[55] Dixit, S.K., Sambasivan, M.: A review of the australian healthcare system: A policy perspective. SAGE Open Med. **6**, 2050312118769211 (Apr 2018)

[56] Ebbehoj, A., Thunbo, M.Ø., Andersen, O.E., Glindtvad, M.V., Hulman, A.: Transfer learning for non-image data in clinical research: A scoping review. PLOS

Digit. Health **1**(2), e0000014 (Feb 2022)

[57] El Arab, R.A., Abu-Mahfouz, M.S., Abuadas, F.H., Alzghoul, H., Almari, M., Ghannam, A., Seweid, M.M.: Bridging the gap: From ai success in clinical trials to real-world healthcare implementation—a narrative review. In: Healthcare. vol. 13, p. 701. MDPI (2025)

[58] Elfiky, A.A., Pany, M.J., Parikh, R.B., Obermeyer, Z.: Development and Application of a Machine Learning Approach to Assess Short-term Mortality Risk Among Patients With Cancer Starting Chemotherapy. JAMA Network Open **1**(3) (07 2018)

[59] Er, M.J., Zhang, Y., Wang, N., Pratama, M.: Attention pooling-based convolutional neural network for sentence modelling. Information Sciences **373**, 388–403 (2016)

[60] Eton, D.T., Anderson, R.T., Cohn, W.F., Kennedy, E.M., St Sauver, J.L., Bucknell, B.J., Ruddy, K.J.: Risk factors for poor health-related quality of life in cancer survivors with multiple chronic conditions: exploring the role of treatment burden as a mediator. Patient Relat. Outcome Meas. **10**, 89–99 (Mar 2019)

[61] Evans, R.S.: Electronic health records: Then, now, and in the future. IMIA Yearbook **25** (05 2016). https://doi.org/10.15265/IYS-2016-s006

[62] Fadol, A., Estrella, J., Shelton, V., Zaghian, M., Vanbenschop, D., Counts, V., Mendoza, T., Rubio, D., Johnston, P.: A quality improvement approach to reducing hospital readmissions in patients with cancer and heart failure. Cardio-Oncology **5**, 5 (06 2019)

[63] Fahridin, S., Agarwal, N., Bracken, K., Law, S., Morton, R.L.: The use of linked administrative data in australian randomised controlled trials: A scoping review. Clin. Trials **21**(4), 516–525 (Aug 2024)

[64] Falzone, L., Salomone, S., Libra, M.: Evolution of cancer pharmacological treatments at the turn of the third millennium. Frontiers in Pharmacology **9** (2018). https://doi.org/10.3389/fphar.2018.01300

[65] Feng, J., Phillips, R.V., Malenica, I., Bishara, A., Hubbard, A.E., Celi, L.A., Pirracchio, R.: Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit. Med. **5**(1), 66 (May 2022)

[66] Fouladvand, S., Talbert, J.C., Dwoskin, L.P., Bush, H., Meadows, A.L., Peterson, L.E., Kavuluru, R., Chen, J.: Predicting opioid use disorder from longitudinal healthcare data using multi-stream transformer. CoRR **abs/2103.08800** (2021)

[67] Fountzilas, E., Pearce, T., Baysal, M.A., Chakraborty, A., Tsimberidou, A.M.: Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digit. Med. **8**(1), 75 (Jan 2025)

[68] Gao, R., Li, L., Tang, Y., Antic, S.L., Paulson, A.B., Huo, Y., Sandler, K.L., Massion, P.P., Landman, B.A.: Deep multi-task prediction of lung cancer and cancer-free progression from censored heterogenous clinical imaging. CoRR (2019)

[69] Gensheimer, M., Henry, A., Wood, D., Hastie, T., Aggarwal, S., Dudley, S., Pradhan, P., Banerjee, I., Cho, E., Ramchandran, K., Pollom, E., Koong, A., Rubin, D., Chang, D.: Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. Annals of Oncology **29** (10 2018)

[70] Gerrard, L., Peng, X., Clarke, A., Long, G.: Multi-level transformer for cancer outcome prediction in large-scale claims data. In: International Conference on Advanced Data Mining and Applications. pp. 63–78. Springer (2023)

[71] Gerrard, L., Peng, X., Clarke, A., Long, G.: Claimsformer: Pretrained transformer for administrative claims data to predict chronic conditions. In: Australasian Joint Conference on Artificial Intelligence. pp. 348–362. Springer (2024)

[72] Gerrard, L., Peng, X., Clarke, A., Schlegel, C., Jiang, J.: Predicting outcomes for cancer patients with transformer-based multi-task learning. In: Australasian Joint Conference on Artificial Intelligence. pp. 381–392. Springer (2022)

[73] Glatzer, M., Panje, Cedric, M., Siren, C., Cihoric, N., Putora, Paul, M.: Decision making criteria in oncology. Oncology **98**, 1–9 (09 2018). https://doi.org/10.1159/000492272

[74] Guo, L.L., Steinberg, E., Fleming, S.L., Posada, J., Lemmon, J., Pfohl, S.R., Shah, N., Fries, J., Sung, L.: EHR foundation models improve robustness in the presence of temporal distribution shift. Sci. Rep. **13**(1), 3767 (2023)

[75] Gupta, S., Tran, T., Luo, W., Phung, D., Kennedy, R.L., Broad, A., Campbell, D., Kipp, D., Singh, M., Khasraw, M., Matheson, L., Ashley, D.M., Venkatesh, S.: Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. BMJ Open **4**(3) (2014)

[76] Hama, T., Alsaleh, M.M., Allery, F., Choi, J.W., Tomlinson, C., Wu, H., Lai, A., Pontikos, N., Thygesen, J.H.: Enhancing patient outcome prediction through deep learning with sequential diagnosis codes from structured electronic health record data: Systematic review. J Med Internet Res **27**, e57358 (Mar 2025). https://doi.org/10.2196/57358, https://www.jmir.org/2025/1/e57358

[77] Han, S., Sohn, T.J., Ng, B.P., Park, C.: Predicting unplanned readmission due to cardiovascular disease in hospitalized patients with cancer: a machine learning approach. Sci. Rep. **13**(1), 13491 (Aug 2023)

[78] Han, W., Pang, B., Wu, Y.N.: Robust transfer learning with pretrained language models through adapters. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 854–861 (2021)

[79] Harutyunyan, H., Khachatrian, H., Kale, D.C., Ver Steeg, G., Galstyan, A.: Multitask learning and benchmarking with clinical time series data. Scientific Data **6**(1) (Jun 2019)

[80] Hasanzadeh, F., Josephson, C., Waters, G., Adedinsewo, D., Azizi, Z., White, J.: Bias recognition and mitigation strategies in artificial intelligence healthcare

applications. NPJ digital medicine **8**, 154 (03 2025). https://doi.org/10.1038/s41746-025-01503-7

[81] He, J., Rasmy, L., Zhi, D., Tao, C.: Advancing pancreatic cancer prediction with a next visit token prediction head on top of Med-BERT. Cancers (Basel) **17**(3) (Feb 2025)

[82] Henzler, D., Schmidt, S., Koçar, A., Herdegen, S., Lindinger, G.L., Maris, M.T., Bak, M.A.R., Willems, D.L., Tan, H.L., Lauerer, M., Nagel, E., Hindricks, G., Dagres, N., Konopka, M.J.: Healthcare professionals' perspectives on artificial intelligence in patient care: a systematic review of hindering and facilitating factors on different levels. BMC Health Serv. Res. **25**(1) (May 2025)

[83] Ho, D., Tan, I.B.H., Motani, M.: Predictive models for colorectal cancer recurrence using multi-modal healthcare data. In: Proceedings of CHIL. p. 204–213. ACM (2021)

[84] Hong, J., Niedzwiecki, D., Palta, M., Tenenbaum, J.: Predicting emergency visits and hospital admissions during radiation and chemoradiation: An internally validated pretreatment machine learning algorithm. JCO Clinical Cancer Informatics **2**, 1–11 (08 2018). https://doi.org/10.1200/CCI.18.00037

[85] Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., Azim, M.A.: Transfer learning: a friendly introduction. J. Big Data **9**(1), 102 (Oct 2022)

[86] Hughes, L., Witham, M.: Causes and correlates of 30 day and 180 day readmission following discharge from a medicine for the elderly rehabilitation unit. BMC Geriatrics **18** (08 2018)

[87] Huguet, N., Chen, J., Parikh, R.B., Marino, M., Flocke, S.A., Likumahuwa-Ackman, S., Bekelman, J., DeVoe, J.E.: Applying machine learning techniques to implementation science. Online J. Public Health Inform. **16** (Apr 2024)

[88] Hur, K., Oh, J., Kim, J., Kim, J., Lee, M.J., Cho, E., Moon, S.E., Kim, Y.H., Atallah, L., Choi, E.: Genhpf: General healthcare predictive framework for

multi-task multi-source learning. IEEE Journal of Biomedical and Health Informatics **28**(1), 502–513 (Jan 2024). https://doi.org/10.1109/jbhi.2023.3327951

[89] Hwang, S., Urbanowicz, R., Lynch, S., Vernon, T., Bresz, K., Giraldo, C., Kennedy, E., Leabhart, M., Bleacher, T., Ripchinski, M.R., Mowery, D.L., Oyer, R.A.: Toward predicting 30-day readmission among oncology patients: Identifying timely and actionable risk factors. JCO Clin. Cancer Inform. **7** (Feb 2023)

[90] Islam, K., Anggondowati, T., Deviany, P., Ryan, J., Fetrick, A., Bagenda, D., Copur, M., Tolentino, A., Vaziri, I., McKean, H., Dunder, S., Gray, J., Huang, C., Ganti, A.: Patient preferences of chemotherapy treatment options and tolerance of chemotherapy side effects in advanced stage lung cancer. BMC Cancer **19** (08 2019). https://doi.org/10.1186/s12885-019-6054-x

[91] Iyer, P.G., Sachdeva, K., Leggett, C.L., Codipilly, D.C., Abbas, H., Anderson, K., Kisiel, J.B., Asfahan, S., Awasthi, S., Anand, P., Kumar M, P., Singh, S.P., Shukla, S., Bade, S., Mahto, C., Singh, N., Yadav, S., Padhye, C.: Development of electronic health record-based machine learning models to predict barrett's esophagus and esophageal adenocarcinoma risk. Clin. Transl. Gastroenterol. **14**(10), e00637 (Oct 2023)

[92] Ji, H., Abushomar, H., Chen, X.K., Qian, C., Gerson, D.: All-cause readmission to acute care for cancer patients. Healthcare quarterly (Toronto, Ont.) **15**, 14–6 (07 2012)

[93] Jiang, C., Deng, L., Karr, M.A., Wen, Y., Wang, Q., Perimbeti, S., Shapiro, C.L., Han, X.: Chronic comorbid conditions among adult cancer survivors in the united states: Results from the national health interview survey, 2002-2018. Cancer **128**(4), 828–838 (Feb 2022)

[94] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: Mimic-iv (version 0.4). PhysioNet (2020)

[95] Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., Lehman, L.W.H., Celi, L.A., Mark, R.G.: MIMIC-IV, a freely accessible electronic health record dataset. Sci. Data **10**(1), 1 (Jan 2023)

[96] Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**, 160035 (2016)

[97] Jung, A.W., Holm, P.C., Gaurav, K., Hjaltelin, J.X., Placido, D., Mortensen, L.H., Birney, E., Brunak, S.R., Gerstung, M.: Multi-cancer risk stratification based on national health data: a retrospective modelling and validation study. Lancet Digit. Health **6**(6), e396–e406 (Jun 2024)

[98] Jung, J.O., Crnovrsanin, N., Wirsik, N., Nienhueser, H., Peters, L., Popp, F., Schulze, A., Wagner, M., Müller-Stich, B., Büchler, M., Schmidt, T.: Machine learning for optimized individual survival prediction in resectable upper gastrointestinal cancer. Journal of Cancer Research and Clinical Oncology **149** (05 2022)

[99] Kann, B.H., Hosny, A., Aerts, H.J.W.L.: Artificial intelligence for clinical oncology. Cancer Cell **39**(7), 916–927 (Jul 2021)

[100] Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization gap and sharp minima (2017), https://arxiv.org/abs/1609.04836

[101] Khan, S.H., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. CoRR **abs/2101.01169** (2021), https://arxiv.org/abs/2101.01169

[102] Khor, S., Haupt, E.C., Hahn, E.E., Lyons, L.J.L., Shankaran, V., Bansal, A.: Racial and ethnic bias in risk prediction models for colorectal cancer recurrence when race and ethnicity are omitted as predictors. JAMA Netw. Open **6**(6), e2318495 (Jun 2023)

[103] Khosravi, M., Zare, Z., Mojtabaeian, S.M., Izadi, R.: Artificial intelligence and decision-making in healthcare: A thematic analysis of a systematic review of reviews. Health Serv. Res. Manag. Epidemiol. **11**, 23333928241234863 (Jan 2024)

[104] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)

[105] Kodialam, R.S., Boiarsky, R., Lim, J., Dixit, N., Sai, A., Sontag, D.: Deep contextual clinical prediction with reverse distillation (2020), https://arxiv.org/abs/2007.05611

[106] Kolla, L., Parikh, R.B.: Uses and limitations of artificial intelligence for oncology. Cancer **130**(12), 2101–2107 (Jun 2024)

[107] Kraljevic, Z., Bean, D., Shek, A., Bendayan, R., Hemingway, H., Yeung, J.A., Deng, A., Baston, A., Ross, J., Idowu, E., Teo, J.T., Dobson, R.J.: Foresight – generative pretrained transformer (gpt) for modelling of patient timelines using ehrs (2023), https://arxiv.org/abs/2212.08072

[108] Kuan, V., Denaxas, S., Gonzalez-Izquierdo, A., Direk, K., Bhatti, O., Husain, S., Sutaria, S., Hingorani, M., Nitsch, D., Parisinos, C.A., Lumbers, R.T., Mathur, R., Sofat, R., Casas, J.P., Wong, I.C.K., Hemingway, H., Hingorani, A.D.: A chronological map of 308 physical and mental health conditions from 4 million individuals in the english national health service. Lancet Digit. Health **1**(2), e63–e77 (Jun 2019)

[109] Lahlou, C., Crayton, A., Trier, C., Willett, E.: Explainable health risk predictor with transformer-based medicare claim encoder (2021), https://arxiv.org/abs/2105.09428

[110] Lal, A., McCaffrey, N., Gold, L., Roder, D., Buckley, E.: Variations in utilisation of colorectal cancer services in south australia indicated by MBS/PBS benefits: a benefit incidence analysis. Aust. N. Z. J. Public Health **46**(2), 237–242 (Apr 2022)

[111] Lam, J.Y., Lu, X., Shashikumar, S.P., Lee, Y.S., Miller, M., Pour, H., Boussina, A.E., Pearce, A.K., Malhotra, A., Nemati, S.: Development, deployment, and

continuous monitoring of a machine learning model to predict respiratory failure in critically ill patients. JAMIA Open **7**(4), ooae141 (12 2024). https://doi.org/10.1093/jamiaopen/ooae141

[112] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations (2020), https://arxiv.org/abs/1909.11942

[113] Lawrence, D., Hancock, K.J., Kisely, S.: The gap in life expectancy from preventable physical illness in psychiatric patients in western australia: retrospective analysis of population based registers. BMJ **346** (2013)

[114] Le, T.L.T., Thome, N., Bernard, S., Bismuth, V., Patoureaux, F.: Multitask classification and segmentation for cancer diagnosis in mammography (2019), https://arxiv.org/abs/1909.05397

[115] Lee, J., Panagiotelis, A., Cairns, R., Wheate, N.J.: An analysis of the trends in the usage of pharmaceutical benefits scheme-subsidised cancer drugs in australia from 2012 to 2022. J. Cancer Res. Clin. Oncol. **150**(8), 375 (Jul 2024)

[116] Lentzen, M., Linden, T., Veeranki, S., Madan, S., Kramer, D., Leodolter, W., Fröhlich, H.: A transformer-based model trained on large scale claims data for prediction of severe covid-19 disease progression. IEEE Journal of Biomedical and Health Informatics **27**(9), 4548–4558 (2023)

[117] Li, X., Peng, L., Wang, Y.P., Zhang, W.: Open challenges and opportunities in federated foundation models towards biomedical healthcare. BioData Mining **18**(1), 2 (2025)

[118] Li, Y., Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., Lukasiewicz, T., Rahimi, K.: Hi-BEHRT: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. IEEE J. Biomed. Health Inform. **27**(2), 1106–1117 (Feb 2023)

[119] Li, Y., Rao, S., Solares, J.R.A., Hassaïne, A., Canoy, D., Zhu, Y., Rahimi, K., Khorshidi, G.S.: BEHRT: transformer for electronic health records. CoRR

abs/1907.09538 (2019)

[120] Lin, Z., Feng, M., dos Santos, C.N., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding (2017), https://arxiv.org/abs/1703.03130

[121] Liu, L., Liu, Z., Wu, H., Wang, Z., Shen, J., Song, Y., Zhang, M.: Multi-task learning via adaptation to similar tasks for mortality prediction of diverse rare diseases. In: AMIA Annual Symposium Proceedings. vol. 2020, p. 763. American Medical Informatics Association (2020)

[122] Liz-López, H., de Sojo-Hernández, Á.A., D'Antonio-Maceiras, S., Díaz-Martínez, M.A., Camacho, D.: Deep learning innovations in the detection of lung cancer: Advances, trends, and open challenges. Cognit. Comput. **17**(2) (Apr 2025)

[123] Long, G.: The rise of federated intelligence: from federated foundation models toward collective intelligence. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 8547–8552 (2024)

[124] Long, G., Shen, T., Tan, Y., Gerrard, L., Clarke, A., Jiang, J.: Federated learning for privacy-preserving open innovation future on digital health. In: Humanity driven AI: productivity, well-being, sustainability and partnership, pp. 113–133. Springer (2021)

[125] Lotter, W., Hassett, M.J., Schultz, N., Kehl, K.L., Van Allen, E.M., Cerami, E.: Artificial intelligence in oncology: Current landscape, challenges, and future directions. Cancer Discov. **14**(5), 711–726 (May 2024)

[126] Luo, J., Ye, M., Xiao, C., Ma, F.: Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 647–656. KDD '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3394486.3403107

[127] Lyu, D., Wang, X., Chen, Y., Wang, F.: Language model and its interpretability in biomedicine: A scoping review. iScience **27**(4), 109334 (Apr 2024)

[128] Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., Gao, J.: Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: SIGKDD. pp. 1903–1911. ACM (Aug 2017)

[129] Ma, Y., Kolla, S., Kaliraman, D., Nolan, V., Hu, Z., Guan, Z., Ren, Y., Armfield, B., Ozrazgat-Baslanti, T., Loftus, T.J., Rashidi, P., Bihorac, A., Shickel, B.: Temporal cross-attention for dynamic embedding and tokenization of multimodal electronic health records (2024), https://arxiv.org/abs/2403.04012

[130] Macheka, S., Ng, P.Y., Ginsburg, O., Hope, A., Sullivan, R., Aggarwal, A.: Prospective evaluation of artificial intelligence (AI) applications for use in cancer pathways following diagnosis: a systematic review. BMJ Oncol. **3**(1), e000255 (May 2024)

[131] MacKay, C., Klement, W., Vanberkel, P., Lamond, N., Urquhart, R., Rigby, M.: A framework for implementing machine learning in healthcare based on the concepts of preconditions and postconditions. Healthc. Anal. (N. Y.) **3**(100155) (Nov 2023)

[132] Madan, S., Lentzen, M., Brandt, J., Rueckert, D., Hofmann-Apitius, M., Fröhlich, H.: Transformer models in biomedicine. BMC Med. Inform. Decis. Mak. **24**(1), 214 (Jul 2024)

[133] Mahesh, N., Devishamani, C.S., Raghu, K., Mahalingam, M., Bysani, P., Chakravarthy, A.V., Raman, R.: Advancing healthcare: the role and impact of AI and foundation models. Am. J. Transl. Res. **16**(6), 2166–2179 (Jun 2024)

[134] Maleki Varnosfaderani, S., Forouzanfar, M.: The role of AI in hospitals and clinics: Transforming healthcare in the 21st century. Bioengineering (Basel) **11**(4) (Mar 2024)

[135] Manikandan, P., Durga, U., Ponnuraja, C.: An integrative machine learning framework for classifying seer breast cancer. Scientific Reports **13** (2023)

[136] Maresova, P., Javanmardi, E., Barakovic, S., Barakovic Husic, J., Tomsone, S., Krejcar, O., Kuca, K.: Consequences of chronic diseases and other limitations associated with old age – a scoping review. BMC Public Health **19** (2019)

[137] te Marvelde, L., McNair, P., Whitfield, K., Autier, P., Boyle, P., Sullivan, R., Thomas, R.: Alignment with indices of a care pathway is associated with improved survival. EClinicalMedicine **15** (2019)

[138] Mbaye, N.M., Danziger, M., Toussaint, A., Dumas, E., Guerin, J., Hamy-Petit, A.S., Reyal, F., Rosen-Zvi, M., Azencott, C.A.: Multimodal behrt: Transformers for multimodal electronic health records to predict breast cancer prognosis. medRxiv (2024). https://doi.org/10.1101/2024.09.18.24312984

[139] McDermott, M.B.A., Nestor, B., Kim, E., Zhang, W., Goldenberg, A., Szolovits, P., Ghassemi, M.: A comprehensive evaluation of multi-task learning and multi-task pre-training on ehr time-series data (2020), https://arxiv.org/abs/2007.10185

[140] Mellish, L., Karanges, E.A., Litchfield, M., Schaffer, A.L., Blanch, B., Daniels, B., Segrave, A., Pearson, S.A.: The australian pharmaceutical benefits scheme data collection: a practical guide for researchers. BMC Research Notes **8** (2015)

[141] Meng, M., Gu, B., Bi, L., Song, S., Feng, D.D., Kim, J.: Deepmts: Deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment pet/ct (2021), https://arxiv.org/abs/2109.07711

[142] Meng, Y., Speier, W., Ong, M.K., Arnold, C.W.: Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE Journal of Biomedical and Health Informatics **25**(8), 3121–3129 (Aug 2021)

[143] Mennella, C., Maniscalco, U., De Pietro, G., Esposito, M.: Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon **10**(4) (Feb 2024)

[144] Meropol, N.J., Donegan, J., Rich, A.S.: Progress in the application of machine learning algorithms to cancer research and care. JAMA Netw. Open **4**(7), e2116063 (Jul 2021)

[145] Miaskowski, C., Dunn, L., Ritchie, C., Paul, S., Cooper, B., Aouizerat, B., Alexander, K., Skerman, H., Yates, P.: Latent class analysis reveals distinct subgroups of patients based on symptom occurrence and demographic and clinical characteristics. Journal of pain and symptom management **50 1**, 28–37 (2015)

[146] Min, X., Yu, B., Wang, F.: Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: A case study on copd. Scientific Reports **9** (2019)

[147] Moglia, V., Johnson, O., de Kamps, M., Cook, G., Smith, L.: Artificial intelligence methods applied to longitudinal data from electronic health records for prediction of cancer: a scoping review. BMC Medical Research Methodology **25**(1), 24 (2025). https://doi.org/10.1186/s12874-025-02473-w

[148] Mohamed, A., AlAleeli, R., Shaalan, K.: Advancing predictive healthcare: A systematic review of transformer models in electronic health records. Computers **14**(4), 148 (Apr 2025)

[149] Moore, A., Orset, B., Yassaee, A., Irving, B., Morelli, D.: Healthrecordbert (herbert): Leveraging transformers on electronic health records for chronic kidney disease risk stratification. ACM Trans. Comput. Healthcare **5**(3) (Sep 2024). https://doi.org/10.1145/3665899

[150] Morgan, D.J., Pineles, L., Owczarzak, J., Magder, L., Scherer, L., Brown, J.P., Pfeiffer, C., Terndrup, C., Leykum, L., Feldstein, D., Foy, A., Stevens, D., Koch, C., Masnick, M., Weisenberg, S., Korenstein, D.: Clinician Conceptualization of the Benefits of Treatments for Individual Patients. JAMA Network Open **4**(7) (07 2021). https://doi.org/10.1001/jamanetworkopen.2021.19747

[151] Morid, M.A., Sheng, O.R.L., Dunbar, J.: Time series prediction using deep learning methods in healthcare. ACM Trans. Manage. Inf. Syst. **14**(1) (jan 2023)

[152] Morishima, T., Matsumoto, Y., Koeda, N., Shimada, H., Maruhama, T., Matsuki, D., Nakata, K., Ito, Y., Tabuchi, T., Miyashiro, I.: Impact of comorbidities on

survival in gastric, colorectal, and lung cancer patients. Journal of Epidemiology **29** (07 2018). https://doi.org/10.2188/jea.JE20170241

[153] Moulaei, K., Yadegari, A., Baharestani, M., Farzanbakhsh, S., Sabet, B., Reza Afrash, M.: Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. Int. J. Med. Inform. **188**(105474), 105474 (Aug 2024)

[154] Nasarudin, N.A., Al Jasmi, F., Sinnott, R.O., Zaki, N., Al Ashwal, H., Mohamed, E.A., Mohamad, M.S.: A review of deep learning models and online healthcare databases for electronic health records and their use for health prediction. Artif. Intell. Rev. **57**(9) (Aug 2024)

[155] Nebbia, G., Arefan, D., Zuley, M., Sumkin, J., Wu, S.: Multi-task learning to incorporate clinical knowledge into deep learning for breast cancer diagnosis. In: Mazurowski, M.A., Drukker, K. (eds.) Medical Imaging 2021: Computer-Aided Diagnosis. vol. 11597, pp. 236 – 241. International Society for Optics and Photonics, SPIE (2021). https://doi.org/10.1117/12.2582285

[156] Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., Khezeli, K., Rashidi, P.: Transformers and large language models in healthcare: A review. Artificial Intelligence in Medicine **154**, 102900 (2024). https://doi.org/https://doi.org/10.1016/j.artmed.2024.102900

[157] Nguyen, P., Tran, T., Wickramasinghe, N., Venkatesh, S.: Deepr: A convolutional net for medical records (2016), https://arxiv.org/abs/1607.07519

[158] Nori, N., Kashima, H., Yamashita, K., Kunisawa, S., Imanaka, Y.: Learning implicit tasks for patient-specific risk modeling in icu. Proceedings of the AAAI Conference on Artificial Intelligence **31** (Feb 2017), https://ojs.aaai.org/index.php/AAAI/article/view/10766

[159] Nors, J., Iversen, L.H., Erichsen, R., Gotschalck, K.A., Andersen, C.L.: Incidence of recurrence and time to recurrence in stage I to III colorectal cancer: A

nationwide danish cohort study. JAMA Oncol. **10**(1), 54–62 (Jan 2024)

[160] Odgaard, M., Klein, K.V., Thysen, S.M., Jimenez-Solem, E., Sillesen, M., Nielsen, M.: Core-behrt: A carefully optimized and rigorously evaluated behrt (2024), https://arxiv.org/abs/2404.15201

[161] Olawade, D.B., David-Olawade, A.C., Wada, O.Z., Asaolu, A.J., Adereni, T., Ling, J.: Artificial intelligence in healthcare delivery: Prospects and pitfalls. Journal of Medicine, Surgery, and Public Health **3**, 100108 (2024). https://doi.org/https://doi.org/10.1016/j.glmedi.2024.100108, https://www.sciencedirect.com/science/article/pii/S2949916X24000616

[162] Osonuga, A., Osonuga, A.A., Fidelis, S.C., Osonuga, G.C., Juckes, J., Olawade, D.B.: Bridging the digital divide: artificial intelligence as a catalyst for health equity in primary care settings. Int. J. Med. Inform. **204**(106051), 106051 (Dec 2025)

[163] Pang, C., Jiang, X., Kalluri, K.S., Spotnitz, M., Chen, R., Perotte, A., Natarajan, K.: Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks (2021), https://arxiv.org/abs/2111.08585

[164] Panteli, D., Adib, K., Buttigieg, S., Goiana-da Silva, F., Ladewig, K., Azzopardi-Muscat, N., Figueras, J., Novillo-Ortiz, D., McKee, M.: Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions. Lancet Public Health **10**(5) (May 2025)

[165] Parikh, R.B., Manz, C., Chivers, C., Regli, S.H., Braun, J., Draugelis, M.E., Schuchter, L.M., Shulman, L.N., Navathe, A.S., Patel, M.S., O'Connor, N.R.: Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. JAMA Network Open **2**(10) (10 2019)

[166] Peng, X., Long, G., Shen, T., Wang, S., Jiang, J.: Self-attention enhanced patient journey understanding in healthcare system. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020. pp. 719–735. Springer (2021)

[167] Peng, X., Long, G., Shen, T., Wang, S., Jiang, J.: Sequential diagnosis prediction with transformer and ontological representation. In: IEEE International Conference on Data Mining. pp. 489–498. IEEE (2021)

[168] Peng, X., Long, G., Shen, T., Wang, S., Jiang, J., Blumenstein, M.: Temporal self-attention network for medical concept embedding. In: IEEE international conference on data mining. pp. 498–507. IEEE (2019)

[169] Peng, X., Long, G., Shen, T., Wang, S., Jiang, J., Zhang, C.: Bitenet: bidirectional temporal encoder network to predict medical outcomes. In: IEEE International Conference on Data Mining. pp. 412–421. IEEE (2020)

[170] Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. CoRR **abs/1906.05474** (2019), http://arxiv.org/abs/1906.05474

[171] Pérez-Núñez, J.R., Rodríguez, C., Vásquez-Serpa, L.J., Navarro, C.: The challenge of deep learning for the prevention and automatic diagnosis of breast cancer: A systematic review. Diagnostics (Basel) **14**(24), 2896 (Dec 2024)

[172] Pham, T., Yin, C., Mehta, L., Zhang, X., Zhang, P.: Cardiac complication risk profiling for cancer survivors via multi-view multi-task learning. CoRR **abs/2109.12276** (2021), https://arxiv.org/abs/2109.12276

[173] Placido, D., Yuan, B., Hjaltelin, J.X., Zheng, C., Haue, A.D., Chmura, P.J., Yuan, C., Kim, J., Umeton, R., Antell, G., Chowdhury, A., Franz, A., Brais, L., Andrews, E., Marks, D.S., Regev, A., Ayandeh, S., Brophy, M.T., Do, N.V., Kraft, P., Wolpin, B.M., Rosenthal, M.H., Fillmore, N.R., Brunak, S., Sander, C.: A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat. Med. **29**(5), 1113–1122 (May 2023)

[174] Poulain, R., Beheshti, R.: Graph transformers on EHRs: Better representation improves downstream performance. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=pe0Vdv7rsL

[175] Poulain, R., Gupta, M., Beheshti, R.: Few-shot learning with semi-supervised transformers for electronic health records. In: Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M., Yeung, S. (eds.) Proceedings of the 7th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 182, pp. 853–873. PMLR (05–06 Aug 2022), https://proceedings.mlr.press/v182/poulain22a.html

[176] Prakash, P., Chilukuri, S., Ranade, N., Viswanathan, S.: Rarebert: Transformer architecture for rare disease patient identification using administrative claims. Proceedings of the AAAI Conference on Artificial Intelligence **35**(1), 453–460 (2021)

[177] Preti, L.M., Ardito, V., Compagni, A., Petracca, F., Cappellaro, G.: Implementation of machine learning applications in health care organizations: Systematic review of empirical studies. J. Med. Internet Res. **26** (Nov 2024)

[178] Rabbani, S.A., El-Tanani, M., Sharma, S., Rabbani, S.S., El-Tanani, Y., Kumar, R., Saini, M.: Generative artificial intelligence in healthcare: Applications, implementation challenges, and future directions. BioMedInformatics **5**(3), 37 (Jul 2025)

[179] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (2018), openAI Blog

[180] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2019), openAI Blog

[181] Rai, H.M., Yoo, J.: A comprehensive analysis of recent advancements in cancer detection using machine learning and deep learning models for improved diagnostics. J. Cancer Res. Clin. Oncol. **149**(15), 14365–14408 (Nov 2023)

[182] Rai, H.M., Yoo, J., Dashkevych, S.: Transformative advances in AI for precise cancer detection: A comprehensive review of non-invasive techniques. Arch. Comput. Methods Eng. **32**(4), 2467–2548 (May 2025)

[183] Raiaan, M.A.K., Sakib, S., Fahad, N.M., Mamun, A.A., Rahman, M.A., Shatabda, S., Mukta, M.S.H.: A systematic review of hyperparameter optimization techniques in convolutional neural networks. Decision Analytics Journal **11**(100470), 100470 (Jun 2024)

[184] Ramezani, M., Takian, A., Bakhtiari, A., Rabiee, H.R., Ghazanfari, S., Mostafavi, H.: The application of artificial intelligence in health policy: a scoping review. BMC Health Serv. Res. **23** (Dec 2023)

[185] Rao, S., Ahmed, N., Salimi-Khorshidi, G., Yau, C., Su, H., Conrad, N., Asselbergs, F.W., Woodward, M., Jackson, R., Cleland, J.G., Rahimi, K.: A transformer-based survival model for prediction of all-cause mortality in heart failure patients: a multi-cohort study (2025), https://arxiv.org/abs/2503.12317

[186] Rao, S., Li, Y., Ramakrishnan, R., Hassaïne, A., Canoy, D., Cleland, J., Lukasiewicz, T., Khorshidi, G.S., Rahimi, K.: An explainable transformer-based deep learning model for the prediction of incident heart failure (2021), https://arxiv.org/abs/2101.11359

[187] Rao, S., Mamouei, M., Salimi-Khorshidi, G., Li, Y., Ramakrishnan, R., Hassaine, A., Canoy, D., Rahimi, K.: Targeted-BEHRT: Deep learning for observational causal inference on longitudinal electronic health records. IEEE Trans. Neural Netw. Learn. Syst. **35**(4), 5027–5038 (Apr 2024)

[188] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**(1), 1–13 (2021)

[189] Rathi, K., Sharma, S., Barnwal, A.: Detecting the undetected: Machine learning in early disease diagnosis. Basic Clin. Pharmacol. Toxicol. **137**(4), e70104 (Oct 2025)

[190] Ravera, F., Gilardi, N., Ballestrero, A., Zoppoli, G.: Applications, challenges and future directions of artificial intelligence in cardio-oncology. Eur. J. Clin. Invest. **55 Suppl 1**(S1), e14370 (Apr 2025)

[191] Razavian, N., Marcus, J., Sontag, D.: Multi-task prediction of disease onsets from longitudinal lab tests (2016). https://doi.org/10.48550/ARXIV.1608.00647

[192] Read, A.J., Zhou, W., Saini, S.D., Zhu, J., Waljee, A.K.: Prediction of gastrointestinal tract cancers using longitudinal electronic health record data. Cancers (Basel) **15**(5) (Feb 2023)

[193] Reddy, S.: Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implement. Sci. **19**(1), 27 (Mar 2024)

[194] Redekop, E., Wang, Z., Kulkarni, R., Pleasure, M., Chin, A., Hassanzadeh, H.R., Hill, B.L., Emami, M., Speier, W., Arnold, C.W.: Zero-shot medical event prediction using a generative pre-trained transformer on electronic health records (2025), https://arxiv.org/abs/2503.05893

[195] Renc, P., Jia, Y., Samir, A.E., Was, J., Li, Q., Bates, D.W., Sitek, A.: Zero shot health trajectory prediction using transformer. NPJ Digit. Med. **7**(1), 256 (Sep 2024)

[196] Rodriguez-Acevedo, A.J., Chan, R.J., Olsen, C.M., Pandeya, N., Whiteman, D.C., Gordon, L.G.: Out-of-pocket medical expenses compared across five years for patients with one of five common cancers in australia. BMC Cancer **21**(1) (Sep 2021)

[197] Ruck, J., Canner, J., Smith, T., Johnston, F.: Use of inpatient palliative care by type of malignancy. Journal of Palliative Medicine **21** (06 2018)

[198] Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)

[199] Rupp, M., Peter, O., Pattipaka, T.: ExBEHRT: Extended Transformer for Electronic Health Records, p. 73–84. Springer Nature Switzerland (2023)

[200] Sahiner, B., Chen, W., Samala, R.K., Petrick, N.: Data drift in medical machine learning: implications and potential remedies. British Journal of Radiology **96**(1150) (03 2023). https://doi.org/10.1259/bjr.20220878

[201] Schröder, T., Schulz, M.: Monitoring machine learning models: a categorization of challenges and methods. Data Science and Management **5**(3), 105–116 (2022). https://doi.org/https://doi.org/10.1016/j.dsm.2022.07.004

[202] Sebastian, A.M., Peter, D.: Artificial intelligence in cancer research: Trends, challenges and future directions. Life (Basel) **12**(12), 1991 (Nov 2022)

[203] Semahegn, A., Manyazewal, T., Hanlon, C., Getachew, E., Fekadu, B., Assefa, E., Kassa, M., Hopkins, M., Woldehanna, T., Davey, G., Fekadu, A.: Challenges for research uptake for health policymaking and practice in low- and middle-income countries: a scoping review. Health Res. Policy Syst. **21**(1), 131 (Dec 2023)

[204] Seneviratne, M.G., Shah, N.H., Chu, L.: Bridging the implementation gap of machine learning in healthcare. BMJ Innov. **6**(2), 45–47 (Apr 2020)

[205] Seyfried, T.M., Huysentruyt, L.: On the origin of cancer metastasis. Critical reviews in oncogenesis **18**, 43–73 (01 2013). https://doi.org/10.1615/CritRevOncog.v18.i1-2.40

[206] Shah, R., Bozic, K.J., Jayakumar, P.: Artificial intelligence in value-based health care. HSS J. p. 15563316251340074 (May 2025)

[207] Shahid, M.S., Imran, A.: Breast cancer detection using deep learning techniques: challenges and future directions. Multimed. Tools Appl. (Jan 2025)

[208] Shahraki-Mohammadi, A., Aliabadi, A., Karimi, A.: Clinical application of artificial intelligence in cancer treatment: A systematic literature review. Health Scope **14**(2) (Apr 2025)

[209] Shamszare, H., Choudhury, A.: Clinicians' perceptions of artificial intelligence: Focus on workload, risk, trust, clinical decision making, and clinical integration. Healthcare (Basel) **11**(16), 2308 (Aug 2023)

[210] Shao, Y., Cheng, Y., Nelson, S.J., Kokkinos, P., Zamrini, E.Y., Ahmed, A., Zeng-Treitler, Q.: Hybrid value-aware transformer architecture for joint learning from longitudinal and non-longitudinal clinical data. Journal of Personalized Medicine **13**(7) (2023). https://doi.org/10.3390/jpm13071070

[211] Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Multi-task prediction of clinical outcomes in the intensive care unit using flexible multimodal transformers (2021), https://arxiv.org/abs/2111.05431

[212] Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. IEEE J Biomed Health Inform **22**(5), 1589–1604 (2018)

[213] Si, Y., Roberts, K.: Deep patient representation of clinical notes via multi-task learning for mortality prediction. AMIA Summits on Translational Science Proceedings **2019**, 779 (2019)

[214] Siebra, C.A., Kurpicz-Briki, M., Wac, K.: Transformers in health: a systematic review on architectures for longitudinal data analysis. Artif. Intell. Rev. **57**(2) (Feb 2024)

[215] Singh, A.P., Balogh, E.P., Carlson, R.W., Huizinga, M.M., Malin, B.A., Melamed, A., Meropol, N.J., Pisano, E.D., Winn, R.A., Yabroff, K.R., Shulman, L.N.: Re-envisioning electronic health records to optimize patient-centered cancer care, quality, surveillance, and research. JCO Oncol. Pract. **21**(2), 128–135 (Feb 2025)

[216] Singh, J., Sandhu, J.K., Kumar, Y.: Metaheuristic-based hyperparameter optimization for multi-disease detection and diagnosis in machine learning. Serv. Oriented Comput. Appl. **18**(2), 163–182 (Jun 2024)

[217] Skou, S., Mair, F., Fortin, M., Guthrie, B., Nunes, B., Miranda, J., Boyd, C., Pati, S., Mtenga, S., Smith, S.: Multimorbidity. Nature Reviews Disease Primers **8**(1), 48 (2022)

[218] Song, H., Rajan, D., Thiagarajan, J.J., Spanias, A.: Attend and diagnose: Clinical time series analysis using attention models. In: AAAI (2018)

[219] Srivastav, A.K., Singh, A., Singh, S., Rivers, B., Lillard, J.W., Singh, R.: Revolutionizing oncology through ai: Addressing cancer disparities by improving screening, treatment, and survival outcomes via integration of social determinants of health. Cancers **17**(17) (2025). https://doi.org/10.3390/cancers17172866, https://www.mdpi.com/2072-6694/17/17/2866

[220] Stabellini, N., Nazha, A., Agrawal, N., Huhn, M., Shanahan, J., Hamerschlak, N., Waite, K., Barnholtz-Sloan, J.S., Montero, A.J.: Thirty-day unplanned hospital readmissions in patients with cancer and the impact of social determinants of health: A machine learning approach. JCO Clin. Cancer Inform. **7** (Jul 2023)

[221] Stacherl, B., Sauzet, O.: Chronic disease onset and wellbeing development: longitudinal analysis and the role of healthcare access. Eur. J. Public Health **34**(1), 29–34 (Feb 2024)

[222] Steinberg, E., Fries, J., Xu, Y., Shah, N.: Motor: A time-to-event foundation model for structured medical records (2023), https://arxiv.org/abs/2301.03150

[223] Summaira, J., Li, X., Shoib, A.M., Li, S., Abdul, J.: Recent advances and trends in multimodal deep learning: A review. CoRR **abs/2105.11087** (2021), https://arxiv.org/abs/2105.11087

[224] Suresh, H., Gong, J.J., Guttag, J.V.: Learning tasks for multitask learning: Heterogenous patient populations in the icu. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018)

[225] Swart, R., Boersma, L., Fijten, R., van Elmpt, W., Cremers, P., Jacobs, M.J.G.: Implementation strategy for artificial intelligence in radiotherapy: Can implementation science help? JCO Clin. Cancer Inform. **8**(8) (Dec 2024)

[226] Swinckels, L., Bennis, F.C., Ziesemer, K.A., Scheerman, J.F.M., Bijwaard, H., de Keijzer, A., Bruers, J.J.: The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: Scoping review. J. Med. Internet Res. **26** (Aug 2024)

[227] Timilsina, M., Buosi, S., Razzaq, M.A., Haque, R., Judge, C., Curry, E.: Harmonizing foundation models in healthcare: A comprehensive survey of their roles, relationships, and impact in artificial intelligence's advancing terrain. Comput. Biol. Med. **189**(109925), 109925 (May 2025)

[228] Topp, S.M., Thompson, F., Johnston, K., Smith, D., Edelman, A., Whittaker, M., Rouen, C., Moodley, N., McDonald, M., Barker, R., Larkins, S.: Democratising data to address health system inequities in australia. BMJ Global Health **8**(5) (2023)

[229] Tran, K.A., Kondrashova, O., Bradley, A., Williams, E.D., Pearson, J.V., Waddell, N.: Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. **13**(1), 152 (Sep 2021)

[230] Tu, H., Wen, C.P., Tsai, S.P., Chow, W.H., Wen, C., Ye, Y., Zhao, H., Tsai, M.K., Huang, M., Dinney, C.P., Tsao, C.K., Wu, X.: Cancer risk associated with chronic diseases and disease markers: prospective cohort study. BMJ **360**, k134 (Jan 2018)

[231] Tufail, A.B., Ma, Y.K., Kaabar, M.K.A., Martínez, F., Junejo, A.R., Ullah, I., Khan, R.: Deep learning in cancer diagnosis and prognosis prediction: A minireview on challenges, recent trends, and future directions. Computational and Mathematical Methods in Medicine **2021**(1) (2021). https://doi.org/https://doi.org/10.1155/2021/9025470

[232] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)

[233] Vehi, J., Mujahid, O., Beneyto, A., Contreras, I.: Generative artificial intelligence in diabetes healthcare. iScience **28**(8), 113051 (Aug 2025)

[234] Victorian Department of Health: Victorian cancer plan 2020-2024: Improving cancer outcomes for all victorians. https://www.health.vic.gov.au/sites/default/files/migrated/files/collections/research-and-reports/v/

victorian-cancer-plan-2020-2024_improving-cancer-outcomes-for-all-victorians.pdf (2020)

[235] Victorian Department of Health: Victorian cancer plan 2024-2028: Optimal and equitable cancer outcomes for all victorians. https://www.health.vic.gov.au/sites/default/files/2024-09/victorian-cancer-plan-2024-28_0.pdf (2024)

[236] Wang, L., Yin, Y., Glampson, B., Peach, R., Barahona, M., Delaney, B.C., Mayer, E.K.: Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data. EBioMedicine **110**(105442) (Dec 2024)

[237] Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D.F., Chao, L.S.: Learning deep transformer models for machine translation (2019), https://arxiv.org/abs/1906.01787

[238] Wang, R., Weng, Y., Zhou, Z., Chen, L., Hao, H., Jing, W.: Multi-objective ensemble deep learning using electronic health records to predict outcomes after lung cancer radiotherapy. Physics in Medicine & Biology **64** (11 2019). https://doi.org/10.1088/1361-6560/ab555e

[239] Wang, X., Wang, F., Hu, J., Sorrentino, R.: Exploring joint disease risk prediction. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium **2014**, 1180–7 (11 2014)

[240] Wang, X., Luo, J., Wang, J., Yin, Z., Cui, S., Zhong, Y., Wang, Y., Ma, F.: Hierarchical pretraining on multimodal electronic health records (2023)

[241] Wang, Y., Long, G., Peng, X., Clarke, A., Stevenson, R., Gerrard, L.: Interactive deep metric learning for healthcare cohort discovery. In: Australasian Conference on Data Mining. pp. 208–221. Springer (2019)

[242] Wang, Y., Guan, Z., Hou, W., Wang, F.: TRACE: early detection of chronic kidney disease onset with transformer-enhanced feature embedding. CoRR **abs/2012.03729** (2020)

[243] Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M.A., Fries, J., Shah, N.H.: The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit. Med. **6**(1), 135 (Jul 2023)

[244] Xian, S., Grabowska, M., Kullo, I., Luo, Y., Smoller, J., Walunas, T., Wei, W.Q., Jarvik, G., Mooney, S., Crosslin, D.: Transformer patient embedding using electronic health records enables patient stratification and progression analysis. NPJ digital medicine **8**, 521 (08 2025). https://doi.org/10.1038/s41746-025-01872-z

[245] Xiao, C., Choi, E., Sun, J.: Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association **25**(10), 1419–1428 (2018)

[246] Xie, F., Yuan, H., Ning, Y., Ong, M.E.H., Feng, M., Hsu, W., Chakraborty, B., Liu, N.: Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. Journal of Biomedical Informatics **126**, 103980 (2022)

[247] Xie, M., Jiang, J., Shen, T., Wang, Y., Gerrard, L., Clarke, A.: A green pipeline for out-of-domain public sentiment analysis. In: International Conference on Advanced Data Mining and Applications. pp. 190–202. Springer (2022)

[248] Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.Y.: On layer normalization in the transformer architecture (2020), https://arxiv.org/abs/2002.04745

[249] Xu, C., Wang, J., Zheng, T., Cao, Y., Fan, Y.: Prediction of prognosis and survival of patients with gastric cancer by weighted improved random forest model. Archives of Medical Science **18** (04 2021)

[250] Xu, D., Shi, Y., Tsang, I.W., Ong, Y., Gong, C., Shen, X.: A survey on multi-output learning. CoRR **abs/1901.00248** (2019), http://arxiv.org/abs/1901.00248

[251] Xu, J., Xi, X., Chen, J., Sheng, V.S., Ma, J., Cui, Z.: A survey of deep learning for electronic health records. Applied Sciences **12**(22) (2022). https://doi.org/10.3390/app122211709, https://www.mdpi.com/2076-3417/12/22/11709

[252] Yang, O., Zhang, Y., To, Y.H., M J IJzerman, M., Liu, J., Gibbs, P., Trapani, K., Pearson, S.A., Franchini, F., PRIMCAT Group: Effects of clinical and socioeconomic factors on medicare and patient costs for colorectal cancer in australia: a retrospective multivariate regression analysis. BMJ Open **14**(12), e081483 (Dec 2024)

[253] Yang, Y., Xu, L., Sun, L., Zhang, P., Farid, S.S.: Machine learning application in personalised lung cancer recurrence and survivability prediction. Computational and Structural Biotechnology Journal **20**, 1811–1820 (2022)

[254] Yang, Y., Long, G., Lu, Q., Zhu, L., Jiang, J., Zhang, C.: Federated low-rank adaptation for foundation models: A survey. arXiv preprint arXiv:2505.13502 (2025)

[255] Yang, Z., Mitra, A., Liu, W., Berlowitz, D., Yu, H.: Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nature Communications **14** (11 2023). https://doi.org/10.1038/s41467-023-43715-z

[256] Yanminsun, Wong, A., Kamel, M.S.: Classification of imbalanced data: a review. International Journal of Pattern Recognition and Artificial Intelligence **23** (11 2011)

[257] Yao, I.Z., Dong, M., Hwang, W.Y.: Deep learning applications in clinical cancer detection: A review of implementation challenges and solutions. Mayo Clinic Proceedings: Digital Health p. 100253 (2025). https://doi.org/https://doi.org/10.1016/j.mcpdig.2025.100253

[258] Yi, H., Qin, Z., Lao, Q., Xu, W., Jiang, Z., Wang, D., Zhang, S., Li, K.: Towards general purpose medical ai: Continual learning medical foundation model (2023),

https://arxiv.org/abs/2303.06580

[259] Yu, J., Dai, Y., Liu, X., Huang, J., Shen, Y., Zhang, K., Zhou, R., Adhikarla, E., Ye, W., Liu, Y., Kong, Z., Zhang, K., Yin, Y., Namboodiri, V., Davison, B.D., Moore, J.H., Chen, Y.: Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras (2024), https://arxiv.org/abs/2404.18961

[260] Yuan, Q., Cai, T., Hong, C., Du, M., Johnson, B.E., Lanuti, M., Cai, T., Christiani, D.C.: Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. JAMA Netw. Open **4**(7), e2114723 (Jul 2021)

[261] Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)

[262] Zeng, X., Lin, S., Liu, C.: Transformer-based unsupervised patient representation learning based on medical claims for risk stratification and analysis (2021), https://arxiv.org/abs/2106.12658

[263] Zeng, X., Linwood, S.L., Liu, C.: Pretrained transformer framework on pediatric claims data for population specific tasks. Sci. Rep. **12**(1), 3651 (Mar 2022)

[264] Zhang, C., Zhang, C., Wang, Q., Li, Z., Lin, J., Wang, H.: Differences in Stage of Cancer at Diagnosis, Treatment, and Survival by Race and Ethnicity Among Leading Cancer Types. JAMA Network Open **3**(4) (04 2020). https://doi.org/10.1001/jamanetworkopen.2020.2950

[265] Zhang, I., Hart, G., Qin, B., Deng, J.: Long-term survival and second malignant tumor prediction in pediatric, adolescent, and young adult cancer survivors using random survival forests: a seer analysis. Scientific Reports **13** (02 2023)

[266] Zhang, Y., Yang, Q.: An overview of multi-task learning. Natl. Sci. Rev. **5**(1), 30–43 (Jan 2018)

[267] Zhang, Y., Yang, Q.: A survey on multi-task learning (2021), https://arxiv.org/abs/1707.08114

[268] Zhang, Y., Hong, J.: Challenges of deep learning in cancers. Technol. Cancer Res. Treat. **22** (Jan 2023)

[269] Zhu, W., Xie, L., Han, J., Guo, X.: The application of deep learning in cancer prognosis prediction. Cancers **12**, 603 (03 2020). https://doi.org/10.3390/cancers12030603

[270] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. CoRR **abs/1911.02685** (2019), http://arxiv.org/abs/1911.02685

[271] Zhuang, L., Park, S.H., Skates, S.J., Prosper, A.E., Aberle, D.R., Hsu, W.: Advancing precision oncology through modeling of longitudinal and multimodal data. IEEE Reviews in Biomedical Engineering p. 1–19 (2025). https://doi.org/10.1109/rbme.2025.3577587