



Ph.D. Thesis

Kiril Vadimovic Klein

Scalable Causal Inference on Electronic Health Records Using Transformers.

Advisor: Mads Nielsen

This thesis has been submitted to the Ph.D. School of The Faculty of Science, University of Copenhagen on 30.11.2025.

Preface

This thesis is the product of more than three years of research carried out under the supervision of Prof. Mads Nielsen at the Department of Computer Science, University of Copenhagen. The work was conducted within the PHAIR (Pharmacovigilance by AI Real-time Analyses) project and was supported by the Innovation Fund Denmark (DK-IFD 1061-00077B) and the Pioneer Centre for AI (DNRF grant number P1).

Chapter 1 introduces the background and motivation. Chapter 2 provides context for the manuscript presented in chapter 3. Chapter 4 is a brief intermezzo on causal inference which prepares the reader for chapters 5 and 6. Finally, chapter 7 summarizes the main findings and outlines directions for future research.

The intended reader is familiar with deep learning and transformer-based models, but does not necessarily have deep expertise in medical data or healthcare. The required concepts from causal inference are introduced as needed.

One of the included manuscripts (chapter 5) is written for an epidemiologic audience. It aims to demystify transformer-based models, and in a familiar target-trial emulation setting, show how they can be integrated into standard pharmacoepidemiologic workflows. The thesis also has a translational aim, to make modern representation learning methods accessible and usable for practising pharmacoepidemiologists.

Abstract

Electronic health records (EHRs) and national health registries contain longitudinal data on millions of patients, enabling broader and faster drug-safety surveillance than traditional, hypothesis-driven studies. Transformer models serve as a flexible tool for learning representations from complex clinical sequences. However, their use in pharmacoepidemiology is limited by open questions about EHR-specific model design, integration with causal analyses, and scalability from a single, well-specified hypothesis to untargeted screening.

This thesis makes three contributions. First, it revisits BERT-style models for structured EHR data and systematically optimises data representation, model architecture, and training procedures, resulting in CORE-BEHRT. The model is evaluated across a broad panel of clinical prediction tasks, providing an empirically grounded baseline and practical design guidelines for EHR transformers.

Second, it develops a framework that combines transformers with target trial emulation in a way that is accessible to epidemiologists and clinicians. Using semaglutide as a case study, it emulates a cardiovascular outcomes trial with observational data using transformer-based propensity scores and inverse probability weighting. The work uses familiar causal estimands to lower the barrier for adoption of representation learning in routine pharmacoepidemiologic practice.

Third, it scales the same machinery to an untargeted causal screen of semaglutide, estimating two-year effects on hundreds of ATC-coded medication outcomes and highlighting patterns with potential safety relevance.

Together, these contributions show how transformers can serve as flexible propensity-score models within classical causal designs, supporting both carefully specified trials and large-scale, assumption-aware drug safety surveillance in real-world data.

Dansk Resumé

Elektroniske patientjournaler og nationale sundhedsregistre (EHR) indeholder data, der følger millioner af patienter over mange år, hvilket muliggør en bredere og hurtigere overvågning af lægemiddelsikkerhed end traditionelle, hypotese-baserede studier. Transformer-modeller er et fleksibelt værktøj til at lære repræsentationer fra komplekse kliniske sekvenser. Deres anvendelse i farmakoepidemiologi er imidlertid begrænset af uafklarede spørgsmål om EHR-specifikt modeludformning, integration med kausale analyser og skalerbarhed fra en enkelt, veldefineret hypotese til ikke-målet screening.

Denne afhandling bidrager med tre ting. Første genbesøger den BERT-lignende modeller til strukturerede EHR-data og optimerer systematisk datarepræsentation, modelarkitektur og træningsprocedurer, hvilket resulterer i CORE-BEHRT. Modellen evalueres på tværs af et bredt panel af kliniske forudsigelsesopgaver, og giver en empirisk funderet baseline samt praktiske designretningslinjer for EHR-transformere.

Dernæst udvikler det en ramme, der kombinerer transformere med target-trial imulering på en måde, der er tilgængelig for epidemiologer og klinikere. Ved hjælp af semaglutide som case-studie simulerer arbejdet et kardiovaskulært outcome-forsøg baseret på observationsdata ved at anvende transformer-baserede tilbøjelighedsscores og invers sandsynlighedsvægtning. Der udnyttes velkendte kausale estimander for at sænke barrieren for at integrere repræsentationslæring i den rutinemæssige farmakoepidemiologiske praksis.

Til sidst skalerer det samme metode til en ikke-målet kausal screening af semaglutid, hvor det estimerer toårige effekter på hundredvis af ATC-kodede medicinresultater og fremhæver mønstre med potentiel sikkerhedsrelevans.

Tilsammen viser disse bidrag, hvordan transformere kan fungere som fleksible tilbøjelighedsscoremodeller inden for klassiske kausale designs og dermed understøtte både nøje specificerede forsøg og antagelsessensitiv lægemiddelsikkerhedsovervågning i virkelige data.

List of Publications

The work presented in this thesis includes the following manuscripts.

1. Mikkel Odgaard, Kiril Vadimovic Klein, Sanne Møller Thysen, Espen Jimenez-Solem, Martin Sillesen, and Mads Nielsen. Core-behrt: A carefully optimized and rigorously evaluated behrt. In *Proceedings of Machine Learning Research*, volume 252, pages 1–33. PMLR, 2024.
2. Kiril Vadimovic Klein, Josefine Windfeld Mathiasen, Espen Jimenez-Solem, Thomas Leth Jensen, Kasper Bonnesen, Martin Hylleholdt Sillesen, Mads Nielsen, and Sanne Møller Thysen. An introduction to causal transformer models in epidemiological studies. Manuscript in preparation, 2025.
3. Kiril Vadimovic Klein, Maria Montgomery, Mikkel Odgaard, Sanne Møller Thysen, Thomas Jensen, Espen Jimenez-Solem, Martin Sillesen, and Mads Nielsen. Untargeted causal drug effect estimation: Populationwise effects of semaglutide. Manuscript in preparation, 2025.

Table of Contents

Preface	i
Abstract	ii
Dansk Resumé	iii
List of Publications	iv
1 Introduction	1
1.1 Motivation and Background	1
1.2 Objectives and Research Questions	2
2 Transformers for EHR	4
2.1 Why Transformers for Structured EHR Data	4
2.2 CORE-BEHRT in Context	6
3 CORE-BEHRT: A Carefully Optimized and Rigorously Evaluated BEHRT	8
4 Intermezzo: Causal Inference for EHR	42
4.1 Identification and Target-Trial Design	42
4.2 Confounding, Pre-treatment Variables, and Proxies	43
4.3 Nuisance Modelling with Transformer Representations	44
4.4 Estimators Used in Practice	45
4.5 Links to the Applied Chapters	45
5 An Introduction to Causal Transformer Models in Epidemiological Studies	47
6 Untargeted causal drug effect estimation: populationwise effects of Semaglutide	77
7 Summary and Discussion	108
7.1 Recap	108
7.2 Summary of Main Contributions	108
7.3 Limitations	110
7.4 Future Directions	111
Bibliography	112

Chapter 1

Introduction

1.1 Motivation and Background

Therapeutic drugs are central to modern healthcare and have substantially improved survival rates and quality of life in various medical conditions. Yet, every prescription carries risks. While benefits are intended, harms may only become visible when drugs are used widely, over long periods of time, and in diverse patient populations.

Adverse Drug Reactions Adverse drug reactions (ADRs) remain common and clinically important. A classic meta-analysis of prospective hospital studies estimated that about 6.7% of hospitalized patients experience a serious ADR and 0.3% die as a result, placing ADRs among the leading causes of death in the United States [4]. More recent work confirms that ADRs are a frequent cause of unplanned hospital admissions and healthcare expenses [5]. Spontaneous reporting systems, which remain the backbone of global pharmacovigilance, suffer from severe underreporting. Even for severe reactions, a widely cited systematic study calculated a median underreporting rate of approximately 94–95% across settings [6, 7]. Because of this, many drug-related side effects are discovered late or remain undetected.

Randomised Controlled Trials Randomised controlled trials (RCTs) are seen as the gold standard for demonstrating efficacy and short-term safety. However, they have the following important limitations. Trial populations are highly selected, and often exclude older, multimorbid, pregnant, or otherwise vulnerable patients. Follow-up is limited, and sample sizes can be insufficient to detect rare or delayed negative effects [8, 9]. Post-marketing monitoring and observational studies are therefore essential complements to RCTs, especially for the above-mentioned scenarios [10, 11, 12].

Historical drug safety crises Historical drug safety crises, most notably thalidomide, exposed the consequences of insufficient surveillance and led to major reforms that strengthened the FDA’s authority over drug safety and efficacy [13, 14]. The 2004 withdrawal of the COX-2 inhibitor rofecoxib (Vioxx), after widespread use and emerging evidence of increased

cardiovascular risk, again highlighted the limitations of post-marketing surveillance and the need for earlier, more proactive safety monitoring [15, 16].

The COVID-19 pandemic provided a more recent example. Vaccines were developed rapidly and were quickly authorised through accelerated approvals, before long-term safety data could be obtained [17]. Regulatory guidance from WHO and the European Medicines Agency, therefore, stressed the need for robust vaccine safety surveillance based on real-world data [18, 19]. In practice, large-scale pharmacovigilance using spontaneous reports and electronic health records detected rare adverse events, such as thrombosis and myocarditis, that have not been seen in pre-marketing clinical trials [20, 21]. These experiences underscored the need for timely, continuous surveillance with methods that are both fast and reliable.

Pharmacoepidemiology then and now Traditional pharmacoepidemiology provides the conceptual and methodological foundation for such work through rigorously designed observational studies in large administrative or clinical databases [22, 23]. However, specifying cohorts, and especially identifying confounders, for each drug-outcome pair, is resource-intensive and does not readily scale to the thousands of interactions that require monitoring. At the same time, routinely collected data sources, such as electronic health records (EHRs) and medical registries, offer rich, longitudinal information that could be used more automatically for signal detection and prioritisation. Current approaches, however, only partially use the granularity and longitudinal structure of these data [24, 25].

Opportunities in Pharmacovigilance Recent advances in deep learning and transformer architectures for structured and unstructured health data open new possibilities to automate parts of this process [26, 27, 28, 29, 30]. Methodological work on signal detection with routinely collected data, including longitudinal exposure-outcome models and sequential monitoring strategies, suggests that more automated, real-time surveillance is technically feasible [31, 32, 33].

This thesis sits at the intersection of these developments and asks:

Can we use modern representation learning and causal inference to make drug safety surveillance more systematic, timely, and scalable?

1.2 Objectives and Research Questions

The PHAIR project (Pharmacovigilance by AI Real-time Analyses), of which this work is a part of, is an attempt to bring these strands together in a concrete national setting [34]. PHAIR combines Danish health registers and hospital EHR data with deep learning and epidemiological methods to develop a system for earlier and broader detection of drug side effects. Within this broader effort, the present thesis focuses on transformer-based models for EHR representation and risk prediction and on their integration with causal inference.

The goal is to take a step toward scalable, real-time drug safety monitoring. More concretely, it pursues the following objectives:

- Learn robust sequence representations from EHR using transformers.
- Emulate a target trial of semaglutide.
- Perform large-scale causal screening of the effects of semaglutide.

In addition to these primary objectives, the thesis has a pedagogical aim. Chapter 5 provides a practice-oriented introduction to transformer models for epidemiological studies, framed around a concrete target-trial emulation. It aims to lower the barrier for epidemiologists and clinicians to use transformers as the basis for propensity scores in observational studies.

Chapter 2

Transformers for EHR

2.1 Why Transformers for Structured EHR Data

To understand the role of CORE-BEHRT in this thesis, let’s explore why the transformer architecture is an attractive choice for structured EHR data.

2.1.1 EHR as Event Sequences

Electronic health records are generated as sequences of encounters and events, not as a single, well-behaved table. Diagnoses, medications, procedures, and laboratory tests occur at irregular times, in changing combinations, and under coding practices that evolve across units and years. Information is spread across visits and often across care settings. A single snapshot at admission, or a small set of hand-crafted summary features, captures only a fraction of this complexity. For many questions of interest in this thesis, such as long-term risk prediction, identification of high-risk subgroups, or adjustment for complex confounders, both the order and the timing of events are central.

2.1.2 Why Transformers are a Natural Candidate.

Early predictive models on EHR data compressed patient histories into fixed-length feature vectors, for example, by counting codes in predefined time windows [35, 36]. While this can work for some tasks, it largely ignores temporal structure and makes it difficult to reuse the same representation across tasks. Recurrent neural networks and related sequence models improved on this by processing ordered sequences of events or visits, but they often struggle with long-range interactions [37, 38, 39]. They were also usually tuned for one outcome at a time, which limits their value as general-purpose building blocks.

Transformers offer a different perspective. Self-attention allows the model to relate events across the entire history, irrespective of how far apart they are in time, and to work naturally with variable-length sequences. With a suitable choice of tokens and a way to encode time, this makes it possible to learn patient representations that integrate information from diagnoses, medications, procedures, and other concepts over extended periods [40, 41, 42].

For this thesis, an important advantage is that a single underlying model can support several roles. It can be pretrained on generic objectives, extended with risk prediction heads for specific outcomes, and later provide embeddings that feed into causal analyses, without redesigning the representation for each new task.

2.1.3 What makes EHR different from language

At the same time, applying transformers to structured EHR data is not simply a matter of reusing methods from natural language processing. Key differences are outlined in MedBERT [41] and in subsequent transformer adaptations to structured EHRs such as BEHRT [40], and CEHR-BERT [42]. Within a hospital visit, many events do not have a natural order in the way that words do in a sentence, and are better viewed as sets rather than sequences [43]. Timestamps reflect clinical rhythms, seasonal patterns, and changes in coding as much as disease progression, and long patient histories often require hierarchical or temporally aware architectures to be modeled efficiently [44, 42]. Different concept types, such as diagnoses and medications, play distinct roles in clinical reasoning. As a result, practical design choices matter. How we group events into tokens, how we indicate visits, how we encode gaps between events, how much history we include, and how we parameterise the model all have a direct impact on what information is available to the attention mechanism and how efficiently it can be used.

2.1.4 Design uncertainty at the time of CORE-BEHRT

When CORE-BEHRT was developed, several EHR-specific transformer models had already been proposed. They made different, and often incompatible, choices about vocabularies, visit markers, temporal encodings, pooling strategies, and training procedures. There was no shared recipe and little guidance on which aspects of the representation were crucial, which architectural refinements were worth their additional complexity, and how stable any conclusions were across endpoints and time periods. For this PhD project, that uncertainty had two consequences. First, it made it difficult to justify using transformer-based representations as a foundation for later causal work without a clearer understanding of their strengths and limitations. Second, there was a risk that apparent performance gains could be driven by details of cohort construction or leakage in the evaluation, rather than by genuinely better use of longitudinal information.

2.1.5 Limits to keep in mind.

Throughout, it is also important to remember the limitations of EHR data that no modelling choice can fully resolve. EHR primarily records encounters, not health. Mild disease and many side effects never appear, which leaves room for residual confounding [45, 46]. Coding is influenced by administrative and financial incentives as well as clinical practice, so diagnostic labels do not always map cleanly to underlying biology [45, 47]. Omissions and temporal drift are common, which makes out-of-time evaluation and periodic retraining essential when

these models are used in practice [48, 49]. These considerations motivate several of the design decisions in CORE-BEHRT and will reappear in later chapters when we use the resulting embeddings and risk scores within causal inference pipelines.

2.2 CORE-BEHRT in Context

In "CORE-BEHRT: A Carefully Optimized and Rigorously Evaluated BEHRT", we revisited BERT-style models for structured EHR and asked a pragmatic question: how much of the progress in EHR transformers comes from genuine innovations and how much is because of under-tuned baselines and ad-hoc choices for data representation? Starting from a BEHRT-like [40] approach, we incrementally optimized three aspects: (i) data representation (ii) model components (iii) training and fine-tuning. While the optimization was performed on three generic tasks, it generalized to 25 diverse clinical prediction tasks with consistent AUROC improvements over the original BEHRT configuration. By showing significant gains on 17 out of 25 tasks and robustness in a prospective evaluation, we established a solid encoder-style baseline for predictions on EHR data.

At the time of publication, most structured-EHR "foundation models" were encoder-only and trained with a masked-language modelling objective: BEHRT [40], Med-BERT [41], CEHRT-BERT [42], and others mentioned in CORE-BEHRT [1]. They differed in vocabulary, temporal encodings, which data sources were included, and how models were evaluated. CORE-BEHRT's main contributions is therefore not a radically new architecture, but a systematic disentangling of these design choices. Notably, we observed the largest difference being made by the inclusion of additional data sources and by better encoding of time. It is, however, important to remember the finite context length of the transformer. When using small context-lengths such as 512 or 1024, inclusion of additional data sources could harm performance, because truncation would mean removal of other events. As an extreme example, imagine using vital signs and naively inserting them into the sequence. Even with modern architectures that can handle much longer sequences than what was used in CORE-BEHRT, this could quickly fill up the context window and remove events that are further in the past but still important. Therefore, one has to consider what is clinically important and use smart truncation strategies when including more data sources.

Overall, we showed that seemingly small choices can have a meaningful impact, but also how many small choices can accumulate in a large performance difference. CORE-BEHRT [1] provides a solid, rigorously evaluated reference point for subsequent work on structured EHR transformers, against which newer models and objectives can be compared, and has since been benchmarked again in comparison to other models [50].

Since the publication of CORE-BEHRT, the surrounding ecosystem has evolved along several fronts. First, there has been a move towards standardization in data formats for longitudinal EHRs. At the forefront, there is the medical-event data standard, a lightweight schema that represents patient histories as time-stamped event streams. It targets interoperability and reproducibility, allowing to exchange of deep learning models easily and facilitates streamlined benchmarking [51, 52]. CORE-BEHRT's own EHR representation aligns closely

with MEDS. Building on this, the BONSAI framework [53] now provides an updated, modular implementation of CORE-BEHRT, ingesting MEDS directly and handling preprocessing and censoring in a standardized and easily configurable way. By default, it uses the faster ModernBERT [54] backbone while still following CORE-BERT’s design principles.

Finally, there is shift toward GPT-like, autoregressive next-event modelling of EHR. Models such as Foresight [55], CEHR-GPT [56], and Ethos [57] adapt a generative approach. Classifying sequences with generative models is more involved, as it requires generating multiple future trajectories to compute over these possible futures. For instance, one can count how often a certain disease occurs. We therefore stick to encoder-only architectures, which we believe are more suitable for classification tasks.

In this landscape, CORE-BEHRT can be positioned as the optimized encoder-style baseline. It clarifies which representation and training decisions are robustly beneficial for structured EHR transformers, and it provides a rigorous evaluation template (multiple heterogeneous endpoints, temporal generalization, and careful handling of censoring) that remains relevant even as the field moves toward longer-context, next-token, and chat-style foundation models. Subsequent tools, such as BONSAI, bring these ideas into the MEDS ecosystem and onto more modern backbones, making CORE-BEHRT’s contributions directly reusable in contemporary EHR modeling pipelines. This sets the stage for using transformer representations in the causal analyses in Chapters 5 and 6.

Chapter 3

CORE-BEHRT: A Carefully Optimized and Rigorously Evaluated BEHRT

The work presented in this chapter has been published at the Machine Learning for Healthcare conference.

CORE-BEHRT: A Carefully Optimized and Rigorously Evaluated BEHRT

Mikkel Odgaard*¹

MIOD@DI.KU.DK

Kiril Klein*¹

KIKL@DI.KU.DK

Sanne Møller Thyssen²

SANNE.MARIE.THYSSEN.01@REGIONH.DK

Espen Jimenez-Solem^{2,3,4}

ESPEN.JIMENEZ.SOLEM@REGIONH.DK

Martin Sillesen⁵

MARTIN.HYLLEHOLT.SILLESEN@REGIONH.DK

Mads Nielsen¹

MADSN@DI.KU.DK

**Equal contribution, order determined by coin toss*

¹*Department of Computer Science, University of Copenhagen*

²*Department of Clinical Pharmacology, Copenhagen University Hospital - Bispebjerg and Frederiksberg*

³*Copenhagen Phase IV Unit (Phase4CPH), Department of Clinical Pharmacology, Center for Clinical Research and Prevention, Copenhagen University Hospital - Bispebjerg and Frederiksberg*

⁴*Department of Clinical Medicine, University of Copenhagen*

⁵*Department of Organ Surgery and Transplantation, Copenhagen University Hospital - Rigshospitalet*

Abstract

The widespread adoption of Electronic Health Records (EHR) has significantly increased the amount of available healthcare data. This has allowed models inspired by Natural Language Processing (NLP) and Computer Vision, which scale exceptionally well, to be used in EHR research. Particularly, BERT-based models have surged in popularity following the release of BEHRT and Med-BERT. Subsequent models have largely built on these foundations despite the fundamental design choices of these pioneering models remaining underexplored. Through incremental optimization, we study BERT-based EHR modeling and isolate the sources of improvement for key design choices, giving us insights into the effect of data representation, individual technical components, and training procedure. Evaluating this across a set of generic tasks (death, pain treatment, and general infection), we showed that improving data representation can increase the average downstream performance from 0.785 to 0.797 AUROC ($p < 10^{-7}$), primarily when including medication and timestamps. Improving the architecture and training protocol on top of this increased average downstream performance to 0.801 AUROC ($p < 10^{-7}$). We then demonstrated the consistency of our optimization through a rigorous evaluation across 25 diverse clinical prediction tasks. We observed significant performance increases in 17 out of 25 tasks and improvements in 24 tasks, highlighting the generalizability of our results. Our findings provide a strong foundation for future work and aim to increase the trustworthiness of BERT-based EHR models.

The code is available on <https://github.com/mikkelfo/CORE-BEHRT>

1. Introduction

In recent years, leveraging Electronic Health Records (EHR) to predict patient outcomes has gained significant interest. Initially, progress was gradual due to a select number of available machine learning methods, such as logistic regression and random forest (MJ et al., 1997; Khalilia et al., 2011). Although models like XGBoost (Chen and Guestrin, 2016) show impressive performance on tabular data, the field has focused largely on adopting deep learning models from natural language processing (NLP) and computer vision. These models, coupled with the drastic increase in available EHR data (Slawomirski et al., 2023), allow sequential processing of the data, taking into account its temporal aspect.

This has been exemplified by the use of recurrent (Choi et al., 2015, 2016a,b) and convolutional (Nguyen et al., 2016) neural networks on EHR data. The introduction of the transformer architecture (Vaswani et al., 2017), known for its versatility across different data modalities and exceptional scaling, significantly impacted EHR analysis. Specifically, building on the seminal work of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), the field has seen the emergence of transformer-based models tailored to EHR data (Li et al., 2020; Rasmy et al., 2021). These early models inspired a wave of subsequent research introducing a variety of modifications touching on input data representation, model architecture, downstream tasks, and evaluation metrics as seen in works like BRLTM (Y et al., 2021) and Hi-BEHRT (Li et al., 2022) or use ICU-specific data (Zhang et al., 2022). This heterogeneity complicates direct comparisons and makes it untractable to pinpoint the sources of improvement.

The field of NLP quickly recognized that when optimally configured, base transformer models could outperform later variants (Yang et al., 2019). A prime example is RoBERTa, which showed that an optimally configured base BERT model could achieve superior results (Liu et al., 2019). We suspect a similar situation exists in the EHR domain, where initial approaches like BEHRT and Med-BERT may not be fully optimized.

Despite the parallels with NLP, it is also important to acknowledge the distinct characteristics of EHR data, notably its strong temporal aspect, irregular time intervals, and the distinct distribution of sequence lengths (Rasmy et al., 2021). These disparities necessitate a comprehensive evaluation of the components that have shown success in NLP in the context of EHR. Furthermore, design choices take a weighted importance in healthcare compared to NLP, as even minor performance improvements can significantly impact patient outcomes.

In contrast to fields like radiology, where deep learning models are increasingly finding their way into clinical practice (Yasaka and Abe, 2018; Montagnon et al., 2020; Lauritzen et al., 2023), there is a lack of trust towards EHR-based models, a cornerstone for clinical adoption (Asan et al., 2020). Therefore, by providing insights into these models, we increase their trustworthiness and make a step towards the integration of BERT-based EHR models into clinical workflows.

Specifically, our study examines the fundamental design choices in data handling, modeling, and training processes for applying BERT-based models to EHR data, drawing inspiration from NLP optimizations. By examining model performance on a broad and diverse set of tasks, we provide robust and extensive evidence to solidify our recommendations.

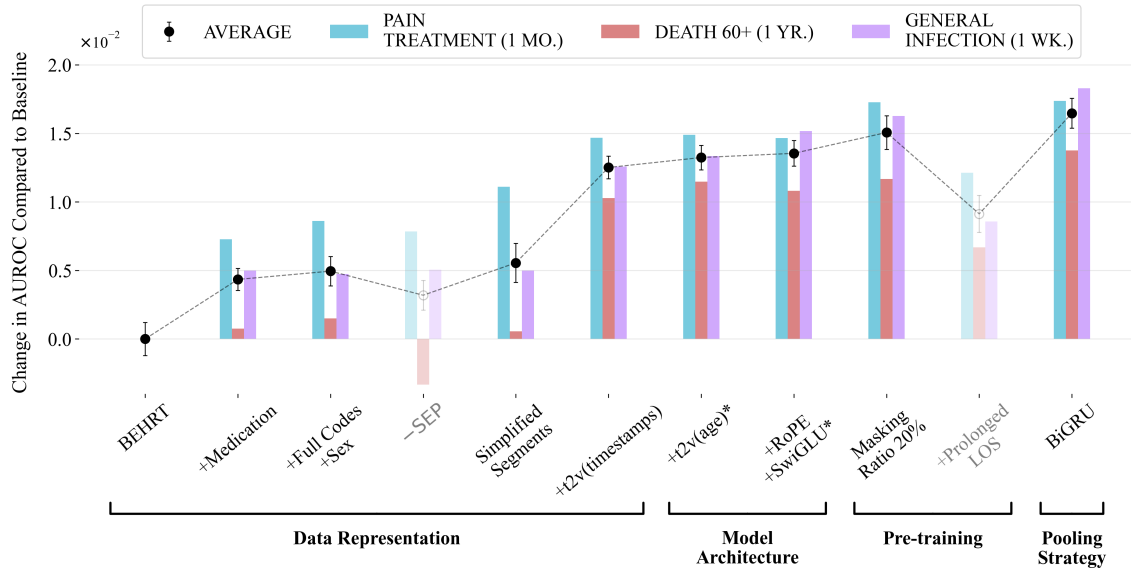


Figure 1: This figure depicts the incremental optimization of the configurations. Non-selected configurations are depicted in a lighter shade, indicating they were considered but not chosen based on their impact on the average Area Under the Receiver Operating Characteristic (AUROC), across three prediction tasks, whose prediction windows are denoted in parentheses. Models were trained in five-fold cross-validation and evaluated on one test set per task. An extension of these experiments for masking ratios and pooling strategies is demonstrated in [Appendix K](#).

* Model performance metrics are averaged not only across cross-validation folds but also over multiple pre-training runs to mitigate instabilities in convergence.

Furthermore, we investigate the variability in model performance to ensure that the models deployed in healthcare are optimized responsibly to inform life-critical decisions.

Generalizable Insights about Machine Learning in the Context of Healthcare

Weak and outdated foundations, coupled with the heterogeneity in newer models, pose challenges in making direct comparisons of BERT-based EHR models. This creates a lack of trust, a major hurdle to clinical adoption, and restricts these models' use to research settings. We address these issues by conducting a careful optimization rigorously evaluated across a diverse set of tasks to increase insights into and confidence towards BERT-based EHR models. Our main contributions consist of:

- Key insights in the data representation, model architecture, and training protocol of EHR modeling through careful optimization
- A comprehensive and rigorous evaluation across 25 diverse clinical prediction tasks showcasing the consistency of the optimizations

- A strongly optimized and comprehensively evaluated foundation serving as a stepping stone for wider adoption of BERT-based EHR models

2. Related Work

This section examines BERT’s transformative impact on EHR analysis, illustrating its application through notable examples. We also extend our discussion to the developments related to BERT in the NLP field.

2.1. BERT in EHR

BEHRT (Li et al., 2020), serving as a proof-of-concept, was pre-trained on EHR data from 1.6 million patients, using sequences of patient events, along with age and visit details, and employed Caliber codes mapping ICD-10 (International Classification of Diseases 10th Revision) codes onto phenotypes. The model was evaluated in disease prediction tasks, showcasing superior performance over traditional recurrent neural network approaches. Their model used BERT’s (Devlin et al., 2018) self-supervised masked language modeling (MLM) pre-training (PT) technique.

Another early example was G-BERT (Shang et al., 2019), which introduced a graph-based approach that leveraged GNNs to represent the hierarchical structure of the ontologies of diagnosis and medication codes. G-BERT was trained on MIMIC-III (Johnson et al., 2016), a small but public dataset consisting of 40k patients from a critical care unit in the US. They evaluated using medication recommendation, a multilabel classification task recommending medications based on diagnosis codes.

To fully leverage the advantages of transformers in large-scale PT, Med-BERT (Rasmy et al., 2021) scaled up the PT dataset to encompass over 28 million patients. It introduced a novel PT task, aimed at predicting extended hospital stays, and expanded its vocabulary to cover all ICD-9 and ICD-10 codes. Med-BERT’s design also diverged from its predecessors by omitting age information and separator tokens and experimenting with various classification heads, among which the Bi-GRU (Schuster and Paliwal, 1997) turned out to be the most effective.

Since then, BEHRT and Med-BERT have been the foundation of many subsequent models. For example, Hi-BEHRT (Li et al., 2022) incorporated hierarchical attention mechanisms to better handle patients with extensive medical histories. CEHR-BERT (Pang et al., 2021) integrated time intervals between patient visits to address the critical aspect of temporal information in EHRs, offering a more nuanced view of patient histories. Targeted-BEHRT (Rao et al., 2022) performed causal inference on EHR data reproducing associations found in randomized controlled trials. BRLTM (Y et al., 2021) included medications, procedures, gender, and topics from clinical notes for depression and chronic disease prediction. RareBERT (Prakash et al., 2021) extended Med-BERT with an adaptive loss for class imbalance, type embeddings, and temporal reference embeddings to better handle rare diseases.

2.2. Transformers in NLP

Despite the advancements on EHR data, the journey of transformer-based models in health-care mirrors the early challenges faced by foundational models like BERT (Devlin et al., 2018). Initial iterations were soon recognized to be underoptimized, brought to light by replication studies such as RoBERTa (Liu et al., 2019).

Their study showed that the original BERT model was under-trained and questioned the usefulness of its Next Sentence Prediction task. They advocated for dynamic over static masking, including longer sequences, and larger batch sizes. They highlighted the importance of key design choices by showcasing better performance than later variants like XLNet (Yang et al., 2019).

We argue that this critique extends to the realm of BERT-based EHR models, which exhibit a wide range of configurations. The diversity in vocabularies, datasets, tasks, and incorporated clinical information underscores the challenge of isolating the effect of individual model components. This heterogeneity complicates the assessment of progress within the field, blurring the lines between genuine innovation and incremental modifications.

Lastly, transformer-based models have seen rapid development, primarily driven by the rise of Large Language Models (LLMs). Innovations include new activation functions, (Shazeer, 2020), improved position embeddings (Su et al., 2023), and compute efficient fine-tuning methods (Dettmers et al., 2023). Furthermore, comprehensive advancements such as the transformer++ recipe have been integrated into modern LLMs like PaLM (Chowdhery et al., 2022), and Llama (Touvron et al., 2023), marking a significant shift from earlier architectures like BERT.

3. Methods

We will introduce our base model and the modifications tested, separated into a data presentation and a technical part. We test a few representative components in different areas, covering as much ground as possible while keeping the computational demand within our budget. The data representation part investigates the impact of representing commonly available background EHR data with little to no implementation overhead. The technical part investigates the impact of improved architecture, embeddings, and training schemes, primarily relevant to machine learning practitioners. Modifications are introduced incrementally, evaluating their impact on the performance of three generic downstream tasks and including them when there is an increase in average Area Under the Receiver Operating Characteristic (AUROC). An overview of this procedure is shown in Figure 2. Next, we describe the optimization procedure before defining the various downstream tasks and close with the hyperparameters of our architecture and optimization.

3.1. Base Model

We used a slightly adapted BEHRT as our starting point. BEHRT’s input data consist of medical events with CLS (classification) and SEP (separator) tokens, age in years, and visits denoted by "positions" and binary visit encodings (segments).

In BEHRT, the events consist of level four ICD10 codes alongside Read codes (Booth, 1994), mapped to 301 Caliber codes (V et al., 2019) representing disease phenotypes. Lack-

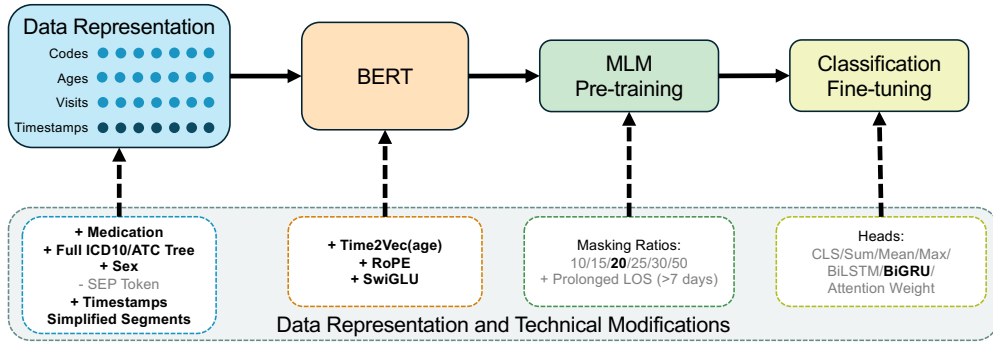


Figure 2: Overview of the iterative optimization procedure, highlighting adapted settings in bold. Symbols denote either addition (+) or removal (-) of a setting with no symbol indicating a direct change. We augmented the input data with medication codes, patient sex, event timestamps, full International Classification of Diseases (ICD-10), and Anatomical Therapeutic Chemical (ATC) codes. Improvements to the model include integrating time2vec embeddings, Rotary Position Embeddings (RoPE), and the SwiGLU activation function. For pre-training (PT), we investigated various masking ratios and Med-BERT’s secondary PT task: predicting prolonged length of stay in the hospital. We investigated various pooling strategies for fine-tuning (FT), finalizing on Bidirectional Gated Recurrent Units (BiGRU) based on its superior results.

ing access to Read codes, we utilized level four ICD10 codes. Similarly, we employed level four ATC (Anatomical Therapeutic Chemical) codes for experiments involving medications. We believe this also strengthens the generalization of our experiments, as ICD and ATC are more commonly used than mappings such as Read and Caliber codes.

We adopted Med-BERT’s leaner configuration for hyperparameters to maintain manageable model dimensions, as BEHRT could use more layers and heads due to its smaller vocabulary size. Further details are provided in [subsection 3.6](#).

Henceforth, we will refer to our modified baseline configuration simply as BEHRT for ease of discussion.

3.2. Data Representation

We investigate different key areas of data representation by including more data sources, extending existing data, and exploring visit representation. This approach provides valuable insight into the aspects of data representation that are crucial for performance gains.

Medication Codes We adopted the BEHRT model to include medication codes, recognizing their potential to improve predictive modeling (Wei et al., 2016). Although the inclusion of medication codes, due to their similar coding system to diagnoses, is not unconventional, it has yet to become standard practice. We consider medication a vital addition to the model, as diagnosis codes are often incomplete or inadequate (Horsky et al., 2017).

For simplicity and computational feasibility, our current model includes only the medication codes, akin to how we handle diagnoses, without considering dosage information. Although our dataset also contains lab tests, vital signs, and procedures, with significantly higher event counts (672.2 million, 605.9 million, and 300.3 million events, respectively), we focus solely on the 227.6 million medication events due to their direct relevance to treatment decisions.

Full-Depth Codes and Background Sentence We propose including more fine-grained data by using the full-depth medical and diagnostic codes and including the patient’s sex. It is common for EHR models to limit the input to higher-level codes, such as in BEHRT (Li et al., 2020). This can be beneficial as high-level representation decreases the vocabulary size which in turn decreases model size. It may also reduce noise to represent information on a coarse level only. Despite these advantages, it can also remove critical information and might especially be detrimental when rare conditions are of interest. BEHRT mapped their codes to 301 Caliber categories, far less than the number of ICD10 codes. When utilizing level four codes, our model works with a vocabulary size of 2,475, whereas using the full codes expands the vocabulary to 17,469. This highlights the balance between model complexity and the granularity of information captured.

To include the patient’s sex in the model, we prepend a token representing male or female to their sequence. We foresee that other relevant demographic variables could also be included to form a ‘background sentence’.

Separator Tokens Separator (SEP) tokens are designed to provide an additional signal to separate visits. Since visit numbers are given explicitly to the model, Med-BERT hypothesized that removing SEP tokens should not diminish performance. Furthermore, for sequences longer than the truncation length, removing SEP tokens allows for including more medical events in the sequence.

Simplified Segments We believe that BEHRT has redundant visit separation (segment) information, including separator tokens, visit numbers, and binary visit encodings. We hypothesize that the model can infer the binary visit encodings through visit numbers and differentiate between visits in this way. We remove the binary visit embedding and change BEHRT’s initialization of the visit number embedding to a standard one rather than their cosine and sinus version.

Timestamps We suggest the inclusion of medical event timestamps to enhance our model, moving beyond BEHRT’s sole reliance on age for a notion of time. Timestamps allow for the precise tracking of short-term changes and are crucial for accurately predicting outcomes that depend on fine time resolution. Unlike the patient’s age, which offers a *relative* sense of time, timestamps provide an *absolute* timeline, enabling the model to account for societal shifts, including pandemics, changes in treatment protocols, and seasonal trends. This feature is particularly useful for detecting shifts in clinical EHR coding practices and adapting to variable baseline risks in conditions like influenza, which intensifies in winter.

Furthermore, external factors can affect hospital interactions, as highlighted during the COVID pandemic when many avoided hospital visits, even for severe conditions (Nourazari et al., 2021). This shift can alter the interpretation of EHR sequences. We can account for these temporal variations in EHR semantics by incorporating event timestamps, ensuring

our model remains accurate over time. Timestamps are input into the model as the time difference in hours from a specific reference point (January 26, 2020).

3.3. Technical Modifications

Here, we introduce the technical components selected from three key technical areas: model architecture, pre-training, and finetuning strategies. The components have been chosen due to their presence and success in the EHR or NLP literature.

Time2Vec Like (Pang et al., 2021), we suggest substituting BEHRT’s conventional age embedding with a Time2Vec (Kazemi et al., 2019) embedding. BEHRT’s current model employs embeddings that rely on discrete age in years, which can be problematic, especially in pediatric cases where changes occur on shorter timescales. These embeddings may also struggle to discern the relationship between different ages. Time2Vec models this relationship explicitly, using differently parameterized periodic functions, by giving similar representations to similar ages. Adequately understanding the relationships between ages necessitates exposure to a wide range of examples for each age. To overcome these limitations and to accommodate the timestamps mentioned previously, we propose the adoption of Time2Vec. We leave the details of Time2vec and our implementation of it in [Appendix A](#).

Improved Transformer Recipe Numerous refinements have been suggested since the original transformer paper, and we incorporate two such advancements from the improved transformer recipe, dubbed transformer++, that are particularly effective in boosting performance: Rotary Position Embedding (RoPE) (Su et al., 2023) and the SwiGLU activation function (Shazeer, 2020).

RoPE replaces the standard positional embedding by encoding the absolute position through a rotation matrix and incorporating relative position in self-attention. The idea is to rotate the embeddings with an angle corresponding to the positional index of the event which is applied at every key and query multiplication in each attention head.

The SwiGLU activation function is a combination of Swish (Ramachandran et al., 2017), also called Sigmoid Linear Unit (SiLU), and GLU (Dauphin et al., 2017). It was shown to improve performance (Shazeer, 2020) and is defined as:

$$\text{SwiGLU}(x) = (\text{Swish}(xW_1) \otimes xW_2)W_3, \quad (1)$$

$$\text{Swish}(x) = x * \text{sigmoid}(x) \quad (2)$$

Pre-training Objective The training scheme of BEHRT plays a major role in the performance, as the strength of these models comes largely from PT, specifically Masked Language Modelling (MLM). Here, BEHRT copied the setup from BERT, masking 15% of the tokens — referred to as the masking ratio (MR) — using the 80-10-10 strategy (80% mask, 10% random word, 10% unchanged). We have yet to see any evidence-based reason for picking 15% in EHR, and other modalities (images, audio, and videos) have much higher MRs (Dosovitskiy et al., 2021; He et al., 2021; Baevski et al., 2020; Tong et al., 2022). It has also been seen that larger models tend to favor larger MRs (Wettig et al., 2023).

Furthermore, we tested the second binary classification PT objective introduced by Med-BERT (Rasmy et al., 2021), in which a prediction is made on whether the patient has a prolonged length of stay in the hospital (> 7 days) for each patient. This task was run alongside MLM.

Pooling Strategies Lastly, we test different pooling strategies for the downstream tasks. It is common to use the CLS token, i.e., the first token in the sequence, as BERT introduced it as a special token specifically for classification. We investigate three additional simple pooling strategies, sum, mean, and max, which aggregate information across the sequence dimension. We also test a weighted variation of mean pooling, which uses an attention layer with a key and query to find the weights of tokens. Lastly, we test two bidirectional Recurrent Neural Networks (RNNs), BiGRU and BiLSTM, as this has been shown to increase performance (Rasmy et al., 2021).

3.4. Optimization Procedure

After introducing the various modifications, we now outline the optimization procedure in this section, as detailed in Figure 2. Our approach begins with improving the data representation, followed by model improvements, and concludes with optimizing the PT process and fine-tuning (FT) pooling strategy.

Each setting is introduced incrementally in the optimization procedure. We begin with the baseline model (BEHRT) and first introduce medication codes. If this modification enhances performance, it is incorporated into the model. Next, we test the second modification (Full ICD10/ATC Tree) on the updated model, which now includes the baseline and the improvement. This iterative process continues for each subsequent modification, with each new change tested on a model that incorporates all previous beneficial changes.

For evaluation, we use three generic downstream tasks for generalizable performance: one-year death prediction for patients older than 60 y.o., one-month pain treatment prediction, and one-week infection prediction; the corresponding definitions are given in Appendix M, consisting of diagnosis and medication codes as well as free text. Death encapsulates the final outcome and can signal the overall health status, encompassing various conditions and factors affecting patient well-being. Pain treatment is very frequent and covers a wide range of causes. It can be both mild and severe, indicative of many underlying conditions, and thus quite general. Infection points to a compromised immune system, which also tells us about the general condition of a patient.

Our three tasks constitute a generalizable evaluation by covering a broad range of health conditions. To compare the performance between models, we mainly considered the area under the receiver operating curve (AUROC) but also reported the area under the precision-recall curve (AUPRC) in Appendix L. For each downstream task, we perform five-fold cross-validation (CV) evaluating on a test set and compute the mean and standard deviation. The downstream task was a binary classification and was constructed as follows. For patients with positive outcomes, we used the outcome of interest as the index date. We then excluded all events within a specified prediction window leading up to the index date, with the window tailored to each downstream task. Disease definitions and prediction windows are given by expert clinicians, ensuring that the model sees no outcome-related information during prediction, preventing any information leakage. For patients with negative outcomes, we assign random censoring dates matching the distribution of the positive cases to avoid information leakage introduced by censoring. This makes the censoring identical for both cases and controls. To account for the imbalance in the dataset, we sampled with the inverse square root of the frequency of each class (Mahajan et al., 2018; Mikolov et al., 2013).

3.5. Assessing Model Generalizability

This section outlines our methodology for evaluating the model’s performance across a diverse array of health conditions. Our goal was to ensure the generalizability of our conclusions by covering the prediction of a wide spectrum of diseases and conditions (used interchangeably from now on), showcasing the optimized model’s adaptability and precision compared to the baseline model in various clinical contexts.

Our comprehensive evaluation encompassed conditions with varying prevalence, from commonly prescribed treatments like pain medications to less frequent conditions such as pancreatic cancer. We considered how conditions develop over time ranging from chronic diseases like diabetes to acute events like major bleeding.

We also covered conditions with different associations to overall health and underlying factors. For instance, the strong correlation between stroke and overall health (Arboix, 2015) contrasts with the more complex interplay of genetic and environmental factors in schizophrenia (Stilo and Murray, 2019; Henriksen et al., 2017), which presents a subtler EHR data signature. We also covered conditions with a limited range of causes, such as anaphylaxis, which primarily involves drug-related causes, and those that can have a wide range of underlying conditions, such as sleep disorder. The conditions are listed in Appendix I together with prediction window and age groups, which were chosen based on relevance and considering the age distributions provided in Appendix J.

This structured approach to task selection is designed to demonstrate the generalizability of our model in identifying a wide range of health outcomes, underscoring its potential utility in diverse clinical scenarios without having to revisit the model choices for every new downstream task. We construct each downstream task as outlined in the previous subsection 3.4, reusing the definitions from Appendix M, and perform hypothesis testing as described in Appendix B.

In addition, we performed experiments using out-of-time evaluation to mimic real-world scenarios more closely. Here, we look at temporally separated outcomes for train, validation, and test sets. A detailed description is provided in Appendix G. This method allows the model to predict the future and assess its response to temporal shifts.

3.6. Hyperparameters

To ensure a fair comparison, we standardized the hyperparameters across all experiments. We aimed to facilitate stable model convergence across all models while keeping manageable training durations. We utilized a base model architecture consisting of six layers and six attention heads with a hidden size of 192 and an intermediate size of 64. Sequences were truncated to 512 tokens, and a batch size of 256 for PT and 512 for FT was used. Learning rates were set to 10^{-3} for PT and $2 \cdot 10^{-5}$ for FT using a linear warmup of five epochs. Models were trained for 50 epochs during optimization and 150 epochs for the generalization experiments with an early stopping of five epochs. A more detailed overview can be found in Appendix D.

Table 1: Statistics for the different train-val sets for death (60+), pain treatment, and general infection for the optimization experiment. We provide the prediction window and mean age at censoring as well as the prevalence of positive patients.

Condition (Age Group)	Prediction Window	Mean Age	Patients [thousands]	Prevalence [%]
Death (60+)	1 yr.	75	55.5	25.6
Pain Treatment	1 mo.	44	107.3	52.7
General Infection	7 d.	49	106.4	57.7

4. Cohort

The data consists of EHRs from 1,825 million patients, prospectively recorded during the period from 2016 to the end of 2022 in the Capital Region of Denmark. We consider only patients with at least one code eligible, which results in 1,805 million patients. For FT, we deem at least one code to be a proper input for binary classification but use patients with ≥ 3 codes for PT to ensure a proper input for the MLM objective task. An overview of the selection pipeline can be seen in [Appendix H](#).

4.1. Cohort Selection

For the optimization experiments, we randomly selected 500 thousand patients, split 80/20 into train and validation sets, and reserved 260 thousand for CV. The specific statistics for each CV task can be seen in [Table 1](#). This subgroup was used as part of the PT group for the final generalizability evaluation, where we used a total of 1,202 million patients, with the remaining 702 thousand used for CV.

We observed a discrepancy in prevalence rates between the optimization and generalization experiments. This could have been caused by the assignment of the subgroup to the generalization PT group, as the subgroup was filtered, which may have left fewer positive cases for the generalization evaluation.

To test generalizability across a wide range of tasks while managing computational demands, we maintained a manageable cohort size for the main experiments. However, we also performed "maximizing performance" experiments, where we extended the cohorts, including PT patients in the FT set for select tasks. Details of these tasks and the expanded cohort sizes are provided in [Appendix F](#).

The out-of-time evaluation, aimed at mimicking a real-world setting, was constructed by censoring training data on 01/01/2020. Without performing patient splits and keeping patients with at least one event we obtain 1,033,465 PT patients. The time separation is described in detail in [Appendix G](#) and patient numbers are provided in [Table 3](#).

For model selection, we require each patient in the PT cohort to have at least three diagnoses to ensure a fair comparison between the diagnosis only and the diagnosis+medication models. For the final evaluation, we require at least three of any (diagnosis or medication) codes, to accurately reflect the increased cohort when including medication.

For each condition, we defined a relevant age and/or sex and a prediction window that

we use for censoring, given in [Table 1](#) for the optimization experiments and in [Appendix I](#) for the generalization assessment. The full age distributions for the cohorts before age group selection are given in [Appendix J](#). We consider patients to be positive when having at least one of the codes given in [Appendix M](#) and negative otherwise. We address imbalance and potential information leakage using the methods described in [subsection 3.4](#).

4.2. Data Extraction

We extract all medical events, with an event consisting of three main variables: concept, timestamp, and admission ID. These events undergo minimal data processing through a simple three-stage process of replacing missing ICD10 codes with textual descriptions, attempting to impute missing values, only when confident, before excluding any events with missing info or ages outside a normal range. We leave a more detailed description in [Appendix C](#).

4.3. Feature Choices

After processing, we end up with a simple data format that describes each medical event with the three main variables: concept, timestamp, and admission ID. The concept of each event is the medical code (diagnosis or medication) and is the main input to the model. We include the timestamp of the event and use it to calculate the patient’s age. The admission ID, together with the timestamp, is used to calculate the visit numbers.

5. Results on Real Data

5.1. Evaluation Approach/Study Design

We first evaluated different components of data representation and technical modifications by pre-training for 50 epochs using a cohort of 500 thousand patients before doing a five-fold CV on death, pain treatment, and infection prediction. We iteratively added components to the pipeline if they improved the average AUROC across these three downstream tasks.

Subsequently, we compared the baseline (BEHRT) against a model with only improved data representation (BEHRT_{+D}) and a model with both improved data representation and technical components (CORE-BEHRT). We evaluated the performance on diverse tasks to ensure the generalizability of the performance gains.

Before setting up our experiments, we investigated the variance in PT and FT by running the base model multiple times. We found that both yielded stable performances with standard deviations of $6.6 \cdot 10^{-4}$ and $3.2 \cdot 10^{-4}$ for PT validation loss and FT validation AUROC, respectively. We investigated the source of variance in FT by conducting a leave-2-out CV; see [Appendix E](#) for details. The analysis revealed that the mean difference within test sets is $7.9 \cdot 10^{-4}$ and pairwise difference for the means across test sets is $4.1 \cdot 10^{-3}$, showcasing that variation is mainly attributed to the test set. These results gave us sufficient confidence to justify a single PT, followed by a five-fold CV. To ensure robust evaluation, we intentionally used larger-than-usual test sets.

5.2. Optimization Experiments

Through iterative enhancements in data representation, model architecture, and training as visualized in [Figure 1](#), our optimization experiments have systematically advanced AUROC across multiple clinical prediction tasks.

Data Representation The proposed data changes outline two main contributions, medication, and timestamps. Both yielded a clear and significant increase in AUROC, whereas full-length codes and sex showed a smaller increase. This makes it clear that adding deeper and more precise data has a strong positive impact on the model.

However, it must be mentioned that adding medication introduced 258 new codes, which increased average sequence lengths by 143.3 codes and training time by five times compared to baseline. Changing to full-depth codes and adding sex introduced almost 15 thousand new codes, which increased the model size by 3.7 times up to 7.933 million parameters. This is due to the increased size of the embedding layer, which increases training time, as these are trainable parameters but do not affect inference time.

Notably, removing SEP tokens worsened performance, primarily for death, while the simplified segments yielded improvements. This reveals that the model mainly uses the SEP tokens to differentiate between visits and relies on explicit embeddings.

Technical Components The two technical components for the model, Time2Vec for age and the improved transformer recipe, yield smaller AUROC gains. We noticed during PT that RoPE, which is part of the improved transformer recipe, yielded a huge boost in MLM performance (33% $\rightarrow$ 58% for top-1 and 82% $\rightarrow$ 91% for top-10 accuracy). This large improvement primarily comes from an improved ability to predict masked medication codes (35% $\rightarrow$ 61%) with a smaller improvement for diagnosis codes (20% $\rightarrow$ 22%).

Masking Ratios We found that an MR of 20% was optimal, showing higher AUROCs for each of the three tasks compared to the base MR of 15%, as seen in [Appendix K](#). At the same time, we observed that changes in masking ratios did not result in substantially different average AUROCs, and in general, the impact was inconsistent across the downstream tasks.

Pooling Strategies The largest improvement in pooling strategies comes from the Bi-GRU head, but we show in [Appendix K](#) that most techniques performed better than the standard CLS token. Even simple aggregation functions, such as mean and max pooling, which use no additional trainable parameters, showed general improvements compared to CLS.

Table 2: Model performance for BEHRT (baseline), BEHRT_{+D} (BEHRT with optimized data representation) and CORE-BEHRT for multiple conditions alongside average age at censoring, patient numbers, and prevalence for positive patients in the test set. If applicable, the selected age group and sex are given in parentheses.

Condition (Age Group/Sex)	Prediction Window	Mean Age	Patients [thousands]/ Prevalence[%]	AUROC [%] BEHRT/BEHRT _{+D} / CORE-BEHRT		
Cancer						
Basal Cell Carcinoma (70+)	1 yr.	80	20.6 / 4.7	64.4	65.8**	67.2**
Breast (60+/f.)	3 mo.	74	13.5 / 4.4	57.0	57.6	62.4
Colon (60+)	3 mo.	75	35.5 / 1.3	61.4	63.6**	62.8
Lung (60+)	3 mo.	74	35.0 / 2.3	66.5	67.9*	68.4**
Pancreatic (60+)	3 mo.	74	37.7 / 0.5	52.9	57.5	56.5
Cardiovascular						
Arrhythmia (60+)	6 mo.	76	35.5 / 10.7	68.9	71.5***	71.6***
Myocardial Infarction (60+)	14 d.	75	40.4 / 2.6	65.7	68.3***	68.7***
Stroke (60+)	14 d.	75	30.2 / 1.3	64.5	69.4**	70.2***
Chronic						
Alzheimer’s Disease (70+)	3 mo.	82	23.1 / 3.7	73.9	76.5**	77.0***
Diabetes	6 mo.	65	107.4 / 4.9	83.5	84.8***	84.8***
Osteonecrosis (70+)	6 mo.	77	25.7 / 0.2	82.5	84.6	84.2
Death (60+)	1 yr.	80	35.0 / 22.7	86.6	87.5***	87.7***†
Emergencies						
Anaphylaxis	24 hr.	32	169.1 / 0.2	71.0	71.4	71.9
Diarrhea	3 d.	59	158.1 / 1.4	85.6	86.5**	87.5**†
Major Bleeding	24 hr.	59	162.6 / 4.6	78.2	78.9**	79.7***†
Infections						
General Infections	7 d.	47	141.0 / 42.8	78.2	79.7***	79.9***†
Pneumonia	14 d.	61	146.6 / 5.5	82.3	83.0***	83.2***†
Psychiatric						
Depression (15 – 30)	3 mo.	45	142.7 / 1.8	71.3	71.8**	71.9**
Schizophrenia (20 – 30)	3 mo.	24	12.7 / 1.2	83.8	85.8**	86.0**
Pain treatment	1 mo.	47	133.8 / 42.0	80.8	82.1***	82.2***††
Pregnancy-Related						
Preterm birth	at pregnancy	30	11.1 / 13.7	53.4	54.5	59.5
Fetal Growth Restriction	at pregnancy	31	11.1 / 6.9	56.9	61.6***	61.7*†††
Placental insufficiency	at pregnancy	32	11.1 / 0.5	58.8	56.3	58.6
Large fetus	at pregnancy	31	11.1 / 2.4	57.9	60.8	61.6
Others						
Sleep disorder (20 – 40)	3 mo.	30	34.0 / 4.1	74.8	75.9**	75.9**

*/**/*** Model is significantly better than BEHRT with $p < 0.05/0.01/0.001$

†/††/††† Model is significantly better than BEHRT_{+D} with $p < 0.05/0.01/0.001$

Table 3: Results obtained mimicking a real-world setting, performing out-of-time evaluation, with train data being censored at 01/2020 and the test set contains outcomes from 09/2021 onwards. For comparison, we provide the difference in mean AUROC to the generalization results from Table 2.

Condition (Age Group/Sex)	Patients train/test [thousands]	Prevalence in test [%]	AUROC [%]	Δ AUROC [%]
Breast Cancer (60+/f.)	72.0/262.9	0.9	64.4	+2.0
Stroke (60+)	174.3/516.8	0.2	67.1	-3.1
Diabetes	593.9/1,657.1	0.9	85.6	+0.8
Death (60+)	281.3/441.0	9.3	87.4	-0.3
Schizophrenia (20 - 30)	59.0/219.9	0.1	87.7	+1.7

5.3. Generalization Results

Our analysis revealed that the CORE-BEHRT model outperformed the baseline in all evaluated scenarios, as detailed in Table 2. When compared to BEHRT_{+D}, significant improvements were observed in only seven downstream tasks, notably including death, pain treatment, and general infection prediction. Despite this, the CORE-BEHRT model consistently demonstrated increased average AUROC scores across the board with exceptions in only five tasks, with two of them being equal in performance.

It is worth noting that tasks where significance was not achieved for both BEHRT_{+D} and CORE-BEHRT, compared to baseline, were predominantly among the more challenging ones, characterized by lower AUROC scores. Specifically, six out of eight tasks lacking significant improvements had AUROC scores below 0.65.

The impact of increasing the CV set can be seen in Appendix F. Breast cancer and preterm birth has notable AUROC increases, while stroke and schizophrenia remain largely unchanged despite the significant increases in patient numbers. This suggests a performance saturation at lower patient counts, supported by their high baseline AUROCs.

5.4. Out-of-Time Evaluation

When mimicking a real-world setting by performing an out-of-time evaluation, the size of the train and test sets increases dramatically and the prevalence decreases. Most tasks see a substantial performance increase, while death sees a small decrease and stroke a larger decrease as detailed in Table 3. Overall, this showcases the model’s ability to handle temporal shifts well.

6. Discussion

We have carefully optimized a BEHRT on three generic tasks, followed by a rigorous evaluation across a broad spectrum of clinical prediction tasks, showcasing the generality of our

findings. This not only lays a solid foundation for future research but also sets a benchmark for the meticulous evaluation necessary to develop effective healthcare models.

The discrepancy in performance gains between the MLM and our downstream EHR tasks, observed with the improved transformer recipe, could be explained by the fact that EHR tasks are not as closely tied to the PT objective as in NLP. EHR tasks that are more closely related to MLM might minimize the discrepancy. This might also explain our observation that different masking ratios impact different tasks in an individual and inconsistent manner. Developing a more tailored PT objective, that closely aligns with the nature of EHR downstream tasks, might alleviate this discrepancy altogether.

The generalization results show a clear trend. Improving data representation consistently, except for placental insufficiency, improves performance. However, expanding the data representation can be associated with increased model size and training times. Currently, EHR models are small compared to other domains, so increases are still manageable with a low compute budget. The technical components give the models an additional performance boost across all but four tasks and reach significance in seven of them. The technical components also never perform significantly worse and do not affect training times much, making them a solid addition to the optimized model.

Increasing the FT set size yielded limited improvements for some tasks, suggesting performance saturation beyond a certain patient number. This saturation occurs faster with higher base scores, as seen with schizophrenia. Thus, adding more patients may not always help; instead, focusing on data representation and technical components could be more effective.

The out-of-time evaluation shows the model’s robustness to temporal shifts, maintaining high scores even with unseen time periods, likely due to the time2vec embedding. We believe this method is appropriate, as it mirrors real-world scenarios where models are used without constant retraining. The performance decrease on stroke might be caused by the larger prediction window with a minimum of three months, while the generalization experiment was performed with a 14-day prediction window.

Here, we highlight the performance of our model on key examples to illustrate its comparative strengths. In stroke prediction, our model achieves state-of-the-art results with an AUROC of 70.7%, which exceeds the 66.9% AUROC reported in recent studies (Teoh, 2018). It also shows comparable performance in preterm birth prediction with an AUROC of 64% versus the 65% reported in the literature (AlSaad et al., 2022), and in breast cancer prediction, where it achieves an AUROC of 63.9% nearly matching the 64.8% in existing studies (Wu et al., 2017). Notably, we used narrower age windows in our work for stroke and breast cancer and employed fewer modalities for preterm birth. These examples illustrate our model’s strengths, even under more constrained conditions.

Our optimization experiments were conducted on three distinct tasks, introducing some potential bias. The three out of seven tasks where CORE-BEHRT significantly outperforms BEHRT_{+D} (BEHRT with optimized data representation) constituted the optimization tasks. This indicates a possibility for bias during the optimization, but we designed the tasks to be general and diverse to mitigate this effect.

Limitations Due to the sensitive nature of EHR data, it can be problematic to replicate setups exactly as public large-scale EHR datasets are rare. We thus used BEHRT with hyperparameters adapted to our dataset, which is also not publicly available. We also did not convert our codes into Caliber codes, which we believe hold greater value since ICD10 and ATC codes are more widely employed.

We found that the data representation and especially the inclusion of medication greatly impacted performance. Despite this, we only investigated the effect of medication codes and ignored sources like lab tests, vital signs, and procedure codes. These data sources are not as easy to integrate as medications and introduce many more events. This made it not only hard to integrate all data sources but also computationally infeasible for us to run a comprehensive optimization and evaluation.

A key limitation of our work is the data size and collection period. Our cohort consists of approximately 1.8 million patients, significantly smaller than the dataset used for MedBERT and tiny compared to typical NLP data sets. The data spans only a seven-year window, limiting the number of diagnosis and medication codes, as well as the feasible prediction windows. Additionally, EHR data quality varies as delays in code registration or missing codes can occur, potentially leading to incorrect patient labeling. Although we partially address this by including free text descriptions, the issue of missing codes remains. Longer prediction windows would be preferable to avoid potential information leakage, especially for slowly developing conditions such as cancer. However, extending these windows proves challenging due to the short data collection period.

Our study focuses only on BERT-based EHR models, ignoring the classical models like Linear Regression, Random Forest, and Gradient Boosted Trees (XGBoost). While BERT-based EHR models have proven to beat simpler models, they have yet to be tested vs. XGBoost. We simply focus on BERT-based models as we believe these models have the most potential due to their versatility and scalability, as evident in NLP and computer vision. These models also have the potential to be multimodal in the future, as seen with LLMs, making them a strong candidate model for the future.

We show that improved data representation significantly boosts performance, with technical components providing an additional improvement. Although we only tested some modifications, we hope they act as representative examples and provide valuable insights into their respective areas. The consistent effects of these modifications remain unexplained, highlighting an important area for future research to uncover and analyze the underlying patterns.

References

- Rawan AlSaad, Qutaibah Malluhi, and Sabri Boughorbel. Predictptb: an interpretable preterm birth prediction model using attention-based recurrent neural networks. *BioData Mining*, 15(1):6, 2022.
- Adrià Arboix. Cardiovascular risk factors for acute stroke: Risk profiles in the different subtypes of ischemic stroke. *World Journal of Clinical Cases: WJCC*, 3(5):418, 2015.

- Onur Asan, Alparslan Emrah Bayrak, and Avishek Choudhury. Artificial intelligence and human trust in healthcare: Focus on clinicians. *J Med Internet Res.*, 2020.
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.
- Nick Booth. What are the read codes? *Health libraries review*, 11(3):177–182, 1994.
- Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16. ACM, August 2016. doi: 10.1145/2939672.2939785. URL <http://dx.doi.org/10.1145/2939672.2939785>.
- E. Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association : JAMIA*, 24:361 – 370, 2016a. URL <https://api.semanticscholar.org/CorpusID:3763233>.
- Edward Choi, Mohammad Taha Bahadori, and Jimeng Sun. Doctor AI: predicting clinical events via recurrent neural networks. *CoRR*, abs/1511.05942, 2015.
- Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: interpretable predictive model in healthcare using reverse time attention mechanism. *CoRR*, abs/1608.05745, 2016b.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.
- Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks, 2017.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.
- Mads G Henriksen, Julie Nordgaard, and Lennart B Jansson. Genetics of schizophrenia: overview of methods, findings and limitations. *Frontiers in human neuroscience*, 11:322, 2017.
- Jan Horsky, Elizabeth A Drucker, and Harley Z Ramelson. Accuracy and completeness of clinical coding using icd-10 for ambulatory visits. In *AMIA annual symposium proceedings*, volume 2017, page 912. American Medical Informatics Association, 2017.
- Johnson, A., Pollard, T., Shen, and L. et al. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 2016.
- Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time, 2019.
- Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11:1–13, 2011.
- Andreas D Lauritzen, My C von Euler-Chelpin, Elsebeth Lynge, Ilse Vejborg, Mads Nielsen, Nico Karssemeijer, and Martin Lillholm. Assessing breast cancer risk by combining ai for lesion detection and mammographic texture. *Radiology*, 308(2):e230227, 2023.
- Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.
- Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 2022.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
- Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. *CoRR*, abs/1805.00932, 2018. URL <http://arxiv.org/abs/1805.00932>.

- Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. URL <http://arxiv.org/abs/1310.4546>.
- Fine MJ, Auble TE, Yealy DM, Hanusa BH, Weissfeld LA, Singer DE, Coley CM, Marrie TJ, and Kapoor WN. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med.*, 1997.
- Emmanuel Montagnon, Milena Cerny, Alexandre Cadrin-Chênevert, Vincent Hamilton, Thomas Derennes, André Ilinca, Franck Vandenbroucke-Menu, Simon Turcotte, Samuel Kadoury, and An Tang. Deep learning workflow in radiology: a primer. *Insights into imaging*, 11:1–15, 2020.
- Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. \ mathtt {Deepr} : a convolutional net for medical records. *IEEE journal of biomedical and health informatics*, 21(1):22–30, 2016.
- Sara Nourazari, Samuel R Davis, Rachel Granovsky, Randolph Austin, Dean J Straff, Joshua W Joseph, and Leon D Sanchez. Decreased hospital admissions through emergency departments during the covid-19 pandemic. *The American journal of emergency medicine*, 42:203–210, 2021.
- Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR, 2021.
- P. Prakash, S. Chilukuri, N. Ranade, and S. Viswanathan. Rarebert: Transformer architecture for rare disease patient identification using administrative claims. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
- Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017.
- Shishir Rao, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Yikuan Li, Rema Ramakrishnan, Abdelaali Hassaine, Dexter Canoy, and Kazem Rahimi. Targeted-behrt: Deep learning for observational causal inference on longitudinal electronic health records. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.
- Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
- Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*, 2019.
- Noam Shazeer. Glu variants improve transformer, 2020.

- Luke Slawomirski, Luca Lindner, Katherine de Bienassis, Philip Haywood, Tiago Cravo Oliveira Hashiguchi, Melanie Steentjes, and Jillian Oderkirk. Progress on implementing and using electronic health record systems: Developments in oecd countries as of 2021. *OECD Health Working Papers*, 2023.
- Simona A Stilo and Robin M Murray. Non-genetic factors in schizophrenia. *Current psychiatry reports*, 21:1–10, 2019.
- Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.
- Douglas Teoh. Towards stroke prediction using electronic health records. *BMC medical informatics and decision making*, 18:1–11, 2018.
- Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- Kuan V, Denaxas S, Gonzalez-Izquierdo A, Direk K, Bhatti O, Husain S, Sutaria S, Hingorani M, Nitsch D, Parisinos CA, Lumbers RT, Mathur R, Sofat R, Casas JP, Wong ICK, Hemingway H, and Hingorani AD. A chronological map of 308 physical and mental health conditions from 4 million individuals in the english national health service. *Lancet Digit Health*, 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Wei-Qi Wei, Pedro L Teixeira, Huan Mo, Robert M Cronin, Jeremy L Warner, and Joshua C Denny. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association*, 23(e1):e20–e27, 2016.
- Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15
- Yirong Wu, Elizabeth S Burnside, Jennifer Cox, Jun Fan, Ming Yuan, Jie Yin, Peggy Peissig, Alexander Cobian, David Page, and Mark Craven. Breast cancer risk prediction using electronic health records. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 224–228. IEEE, 2017.
- Meng Y, Speier W, Ong MK, and Arnold CW. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE Journal of Biomedical and Health Informatics*, 2021.

- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.
- Koichiro Yasaka and Osamu Abe. Deep learning and artificial intelligence in radiology: Current applications and future directions. *PLoS medicine*, 15(11):e1002707, 2018.
- Tianran Zhang, Muhao Chen, and Alex A.T. Bui. Adadiag: Adversarial domain adaptation of diagnostic prediction with clinical event sequences,. *Journal of Biomedical Informatics*, 2022.

A. Time2vec Implementation

Given a scalar input τ Time2Vec (t2v) produces a vector of size k as follows:

$$\text{t2v}(\tau)[i] = \begin{cases} \omega_i \tau + \phi_i, & \text{if } i = 0. \\ \mathcal{F}(\omega_i \tau + \phi_i), & \text{if } 1 \leq i \leq k, \end{cases} \quad (3)$$

where $\text{t2v}(\tau)[i]$ is the i^{th} element of $\text{t2v}(\tau)$, $\mathcal{F}$ is the periodic function, where we choose cosine, ω_i and ϕ_i are learnable frequency and phase-shift parameters. Time2vec uses fewer parameters than a standard embedding and can easily be implemented as a drop-in replacement.

We noticed a certain degree of instability when using time2vec out of the box, but observed that scaling τ and clipping the 0^{th} dimension helped with this issue and stabilized the performance. After a rough hyperparameter search, we found stability when scaling age with a factor of 10^{-2} , constraining it from 0 to around 1, and scaling absolute time with a factor of 10^{-4} , converting a year of absolute time (8760 hours) to ≈ 1 . Furthermore, the 0^{th} dimension was clipped to $[-100, 100]$ to avoid large embedding values for very early events.

B. P-value Computation

To assess the statistical significance between the base and best models’ performances, we calculated one-sided p-values using a t-test, under the null hypothesis that the model is not superior to either BEHRT or BEHRT_{+D}, if applicable. The means and standard deviations, based on five observations each, allowed for precise standard error estimation and subsequent t-statistic computation for each disease category. We apply the Benjamini-Hochberg procedure to account for the false discovery rate.

C. Data processing

Code-level replacement Some events had missing ICD10 codes, but also an accurate textual description of the diagnosis. Rather than excluding the events with missing ICD10 codes, we replace the missing value with the textual description. We account for these replacements when defining conditions, as each diagnosis code can now have multiple tokens to represent it.

Admission-level Imputation Missing admission IDs and timestamps can sometimes be inferred by using surrounding events. Specifically, we try to impute admission IDs and timestamps, *only* in the places where this can be done confidently.

Missing admission IDs can be inferred when a missing admission ID exists between two identical admissions, e.g. [A None A]. Here, it is highly likely that the missing admission is part of the A admissions. Any missing admission ID that cannot be inferred is replaced with a unique identifier and treated as a standalone admission.

Some missing timestamps can be imputed if they are part of an admission with at least one other entry. Here, the missing timestamp can be estimated as most admissions lie within a similar time frame. The first, average, or last timestamp can be used, where we choose the last so as not to interfere with the ordering of the other codes.

Event-level Exclusion Any event still containing missing values or events where the patient’s age is outside a normal range ($0 \leq \text{age} \leq 120$) is removed.

D. Hyperparameters

Architecture Specification As part of our architecture choices, we adopted settings from Med-BERT with six layers, six attention heads, and a hidden dimension size of 192. These were chosen due to their efficient performance on larger datasets. The intermediate size is 64, with a context length of 512, considering variations in vocabulary sizes which affect model size.

Optimization and Training Parameters We used an epsilon value of 10^{-6} for the AdamW optimizer. The learning rates were set to 10^{-3} for PT and $2 \cdot 10^{-5}$ for FT, with a linear warmup phase of five epochs. Optimization experiments were limited to 50 epochs with an early stopping mechanism based on validation loss for PT and AUROC for FT. For the generalization experiments, the number of training epochs was increased to 150.

Computational Resources and Training Duration Training was conducted using Nvidia A100 GPUs with 80 GiB of VRAM. Batch sizes were set at 256 for PT and 512 for FT, ensuring maximal GPU utilization for our largest model versions. The duration for PT one model on 400,000 patients ranged from 2.5 to 16.5 hours, depending on model size.

E. Leave-2-Out Cross Validation

The main objective of the leave-2-out CV experiment was to analyze the sources of variance in model performance. The dataset was divided into five distinct folds, facilitating a comprehensive assessment through multiple test runs. During each iteration of the experiment, two of the five folds were selected: one served as the validation set and the other as the test set. This configuration allowed each test set to be evaluated by two different models, which were trained on overlapping subsets of the remaining three folds. The experiment was structured to execute a total of 10 unique runs, each providing insights into model behavior under different training and validation scenarios. For each test set, we computed the performance differences between the two models and then averaged these differences to quantify the variability within the model evaluations. Furthermore, this average was compared against the mean of the pairwise differences in the mean performances across different test sets. This comparison aimed to distinguish between the intrinsic variability of the model and variability attributable to the composition of the test sets.

F. Maximizing Performance

Table 4: Results obtained with CORE-BEHRT model for an increased cross-validation set on select conditions. AUROC is given with standard deviation computed from the five folds. The same test sets as described in Table 2 were used.

Condition (Age Group/Sex)	Patients in CV [thousands]	Prevalence [%]	AUROC [%]
Breast Cancer (60+/f.)	124.3	5.2	63.9 $\pm$ 0.3
Preterm Birth	100.0	13.7	63.7 $\pm$ 0.2
Schizophrenia (20 – 30)	89.2	1.4	85.9 $\pm$ 0.6
Stroke (60+)	276.2	1.5	70.7 $\pm$ 0.5

The impact of increasing the CV set on model performance varies by condition, as shown in Table 2 and Table 4. For breast cancer and preterm birth, AUROC increases by 2.4% and 7.1%, respectively. However, AUROCs for stroke and schizophrenia remain largely unchanged despite a 3.6-fold and 2.5-fold increase in patient numbers. This suggests performance saturation at lower patient counts, given their high baseline AUROCs of 70% and 86%.

G. Out of Time Evaluation

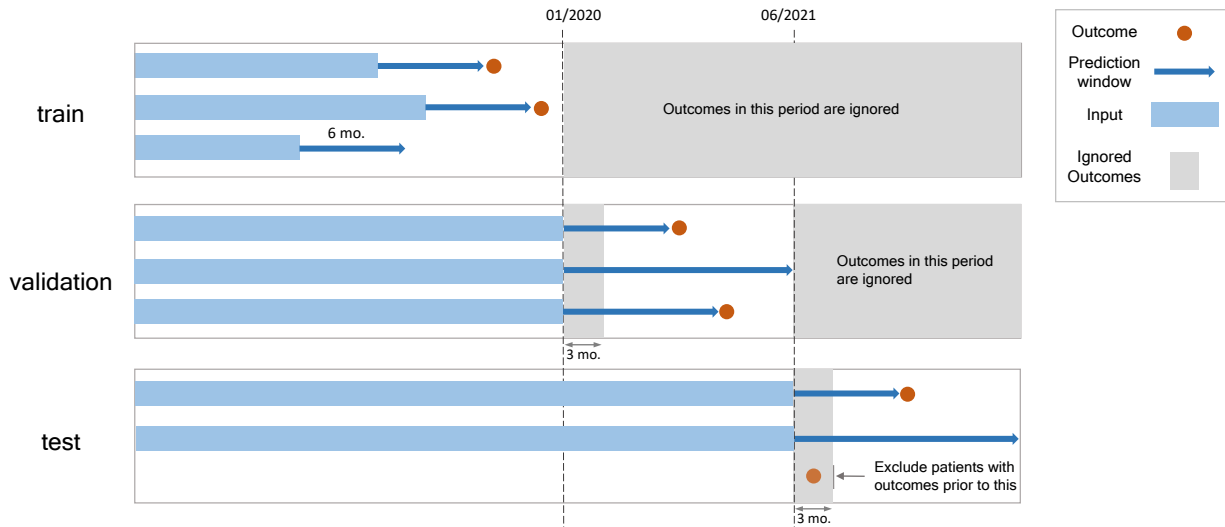


Figure 3: This figure illustrates the out-of-time fine-tuning and evaluation process. Training used data only up to 01/2020. The remaining three years were split for validation and testing: validation labels covered outcomes from 03/2020 to 06/2021, and testing labels covered outcomes from 09/2021 onward. Patients with outcomes before these dates in their respective sets were excluded.

H. Cohort Selection

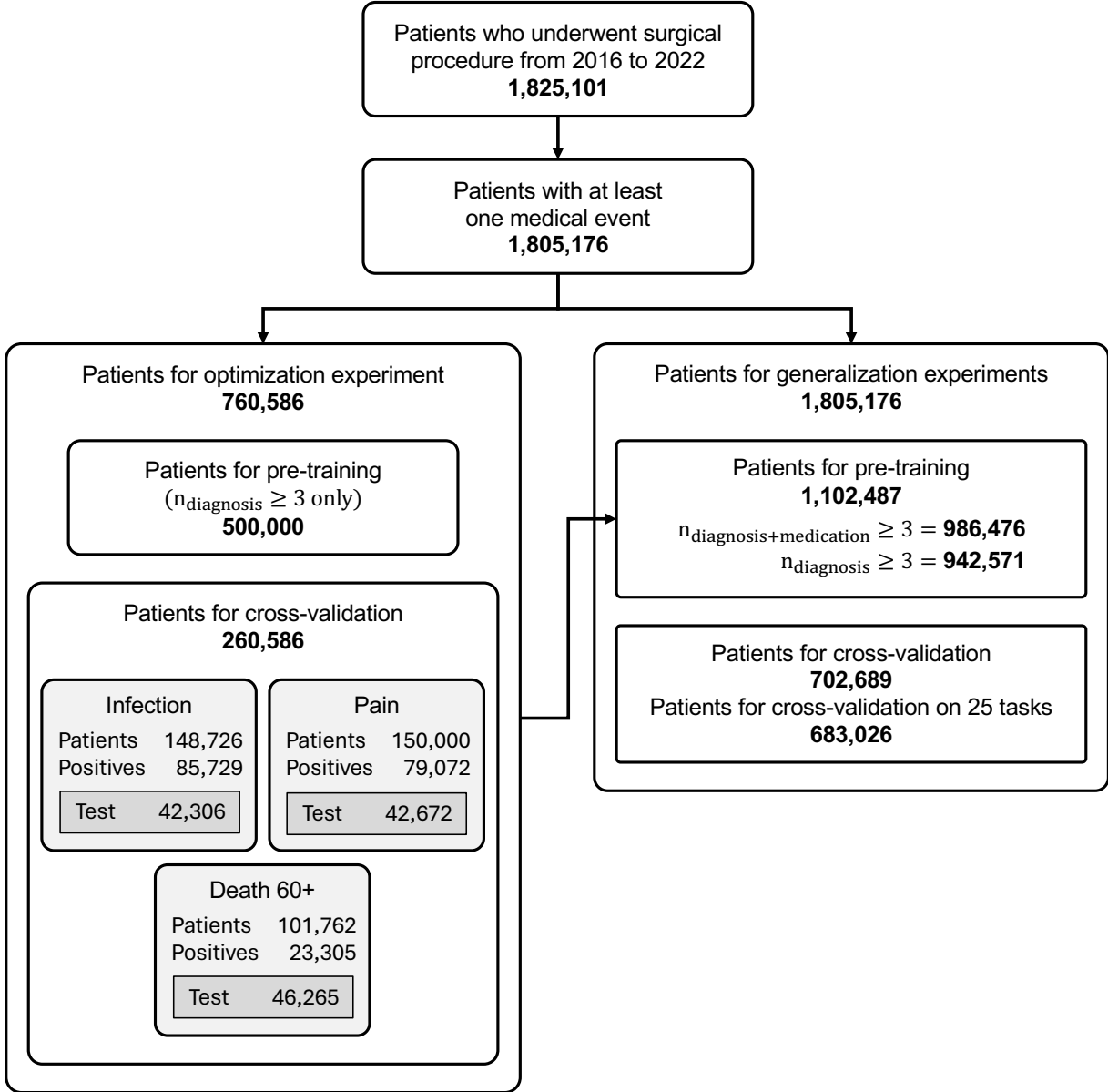


Figure 4: Selection pipeline for model selection and final evaluation

I. Generalization Cohort Statistics

Table 5: Statistics for all 25 CV sets for the generalization experiment. sequence length is N events per patient sequence (post-censoring). Trajectory length* is in years;

Condition	Prediction Window	Age mean $\pm$ std	Sequence Length Q2(Q1-Q3)	Trajectory Length Q2(Q1-Q3)	Patients [thousands]/ Prevalence[%]
Alzheimer's Disease	3 mo.	79.1 $\pm$ 7.1	27 (3-161)	2.5 (1.3-3.9)	46.2 / 3.7
Anaphylaxis	24 hr.	41.6 $\pm$ 24.7	4 (1-28)	2.2 (1.0-3.5)	338.1 / 0.2
Arrhythmia	6 mo.	73.6 $\pm$ 9.2	9 (2-85)	2.4 (1.4-3.7)	70.9 / 10.7
Basal Cell Carcinoma	1 yr.	79.4 $\pm$ 7.2	22 (3-147)	3.0 (1.9-4.3)	41.2 / 4.6
Breast Cancer	3 mo.	74.4 $\pm$ 9.6	11 (2-102)	2.3 (1.2-3.8)	27.1 / 4.5
Colon Cancer	3 mo.	73.7 $\pm$ 9.3	14 (2-110)	2.4 (1.3-3.9)	71.0 / 1.3
Lung Cancer	3 mo.	73.7 $\pm$ 9.2	14 (2-110)	2.4 (1.3-3.9)	69.9 / 2.1
Pancreatic Cancer	3 mo.	73.7 $\pm$ 9.3	15 (3-117)	2.5 (1.3-3.8)	75.5 / 0.4
Death	1 yr.	73.7 $\pm$ 9.1	11 (2-89)	3.0 (2.0-4.3)	70.1 / 22.5
Depression	3 mo.	22.5 $\pm$ 4.3	2 (1-7)	2.1 (1.1-3.4)	53.5 / 2.7
Diabetes	6 mo.	41.9 $\pm$ 24.7	3 (1-17)	2.0 (1.1-3.1)	214.8 / 5.0
Diarrhea	3 d.	41.5 $\pm$ 24.7	4 (1-26)	1.9 (0.9-3.1)	316.3 / 1.3
Infection	7 d.	41.2 $\pm$ 24.7	2 (1-7)	1.4 (0.6-2.7)	281.9 / 42.6
Major Bleeding	24 hr.	41.7 $\pm$ 24.7	4 (1-26)	2.0 (0.9-3.4)	325.1 / 4.7
Myocardial Infarction	14 d.	73.7 $\pm$ 9.3	16 (3-118)	2.3 (1.1-3.7)	80.8 / 2.7
Osteonecrosis	6 mo.	79.4 $\pm$ 7.2	32 (4-176)	3.0 (1.8-4.4)	51.4 / 0.2
Pain Treatment	1 mo.	40.8 $\pm$ 24.5	2 (1-5)	1.4 (0.6-2.5)	267.5 / 42.0
Pneumonia	14 d.	41.7 $\pm$ 24.8	4 (1-24)	1.8 (0.8-3.3)	293.2 / 5.5
Large Fetus	at pregnancy	30.4 $\pm$ 5.1	3 (2-7)	0.5 (0.0-2.1)	22.2 / 2.3
Placental Insufficiency	at pregnancy	30.4 $\pm$ 5.0	3 (2-7)	0.4 (0.0-2.1)	22.2 / 80.6
Fetal Growth Disorder	at pregnancy	30.4 $\pm$ 5.1	3 (2-7)	0.4 (0.0-2.1)	22.2 / 6.6
Preterm Birth	at pregnancy	30.4 $\pm$ 5.0	3 (2-7)	0.4 (0.0-2.1)	22.2 / 13.7
Schizophrenia	3 mo.	24.7 $\pm$ 2.9	2 (1-7)	1.6 (0.8-2.9)	25.5 / 1.2
Sleep Disorder	3 mo.	29.3 $\pm$ 5.6	2 (1-11)	1.8 (1.0-2.9)	67.9 / 4.3
Stroke	14 d.	73.5 $\pm$ 9.1	12 (2-102)	1.9 (0.9-3.5)	60.4 / 1.4

* From first encounter to censoring

J. Age Distribution for Generalization Cohort

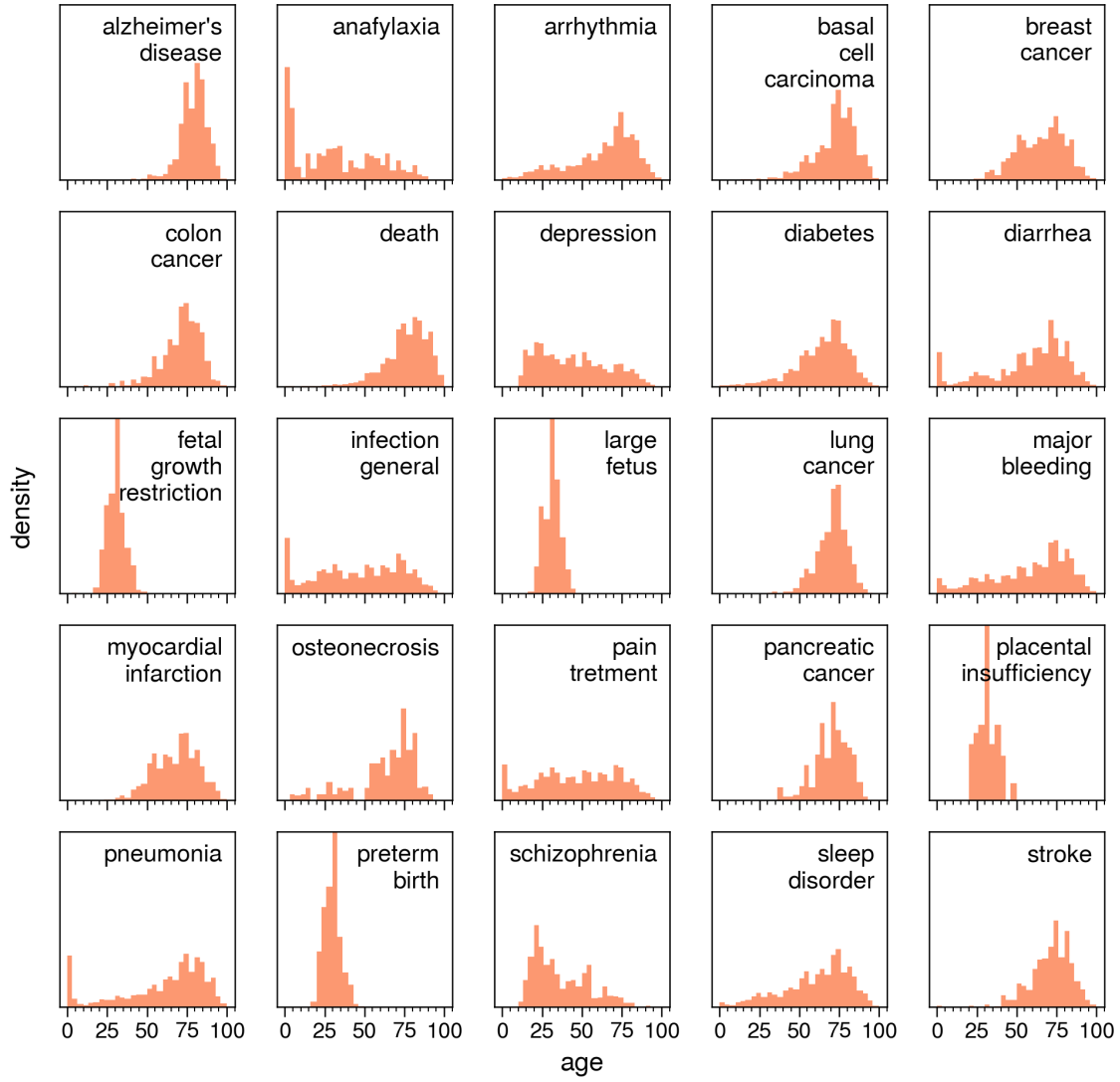


Figure 5: Age distribution for positive patients in the cross-validation set, before age group selection, for each of the 25 downstream tasks.

K. Results of Generalization Experiments

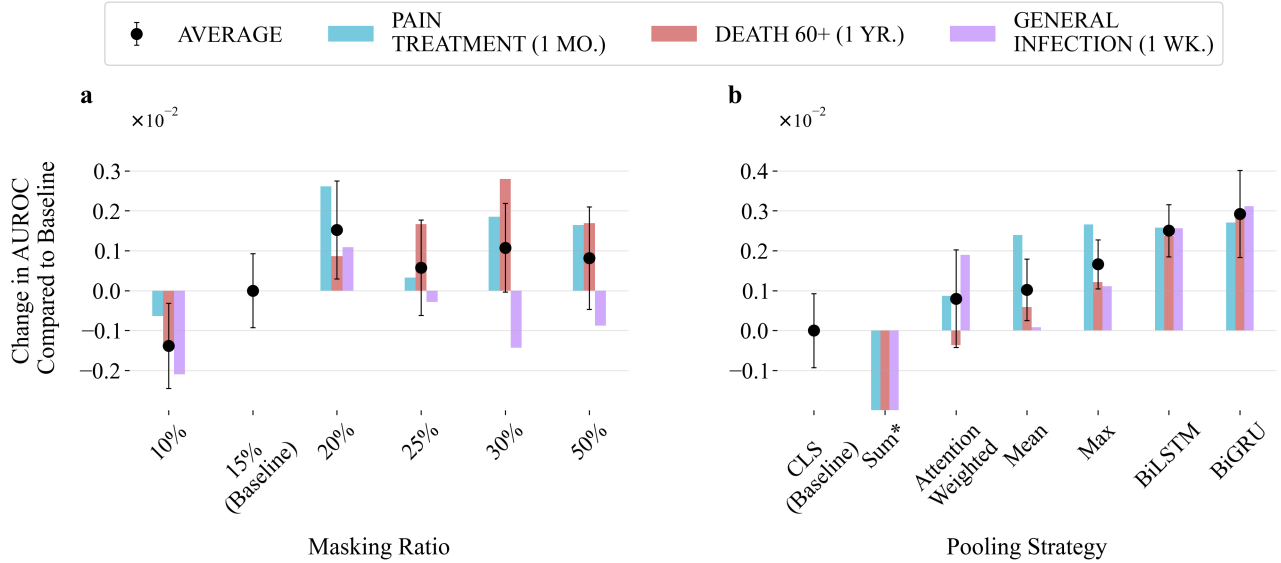


Figure 6: Ablation studies analyzing the effects of different masking ratios (MR) and pooling strategies on predictive performance. In (a) we compare the change in AUROC across varying MRs using CORE-BEHRT with 15% MR and CLS (classification) token pooling as a baseline. While an MR of 20% shows the highest average AUROC improvement, certain conditions such as death prediction benefit from a higher MR of 30%. In (b) we evaluate various pooling strategies against CORE-BEHRT with CLS pooling as a baseline. The Bidirectional Gated Recurrent Unit (BiGRU) pooling strategy emerges as superior based on the average AUROC improvement across conditions.

* This change resulted in a change of average AUROC of $-1.5 \cdot 10^{-2}$, which is not displayed to maintain readability of other differences.

L. Results of Optimization Experiments

Table 6: Test scores for initial experiments given as AUROC and AUPRC with mean and standard deviations obtained from the five models trained in CV for three prediction tasks pain treatment, death, and general infection. Our intermediate models are highlighted in bold. Adapted settings are shown with an asterisk.

Model	AUROC (%)			AUPRC (%)		
	Pain	Death	Infection	Pain	Death	Infection
Data representation						
Med-BERT	76.53±0.07	79.23±0.13	72.67±0.11	79.10±0.05	52.77±0.21	77.92±0.13
*BEHRT	77.01±0.13	84.54±0.15	73.87±0.07	79.51±0.17	60.27±0.35	78.84±0.06
*+Medication	77.74±0.08	84.62±0.06	74.37±0.10	80.28±0.07	60.91±0.23	79.50±0.09
*+FV+Sex	77.87±0.06	84.69±0.17	74.34±0.04	80.44±0.06	61.24±0.37	79.39±0.06
-SEP	77.80±0.02	84.21±0.17	74.37±0.07	80.30±0.05	61.13±0.39	79.34±0.06
*Simplified Segments	78.12±0.04	84.60±0.16	74.37±0.19	80.50±0.04	61.06±0.29	79.30±0.19
*+t2v(timestamps)	78.48±0.11	85.57±0.11	75.12±0.07	80.99±0.06	63.41±0.30	80.29±0.10
BEHRT_{+D}						
Technical components						
*+t2v(age)	78.50±0.05	85.69±0.11	75.20±0.06	80.97±0.04	63.67±0.27	80.33±0.10
*+RoPE+SwiGLU	78.48±0.05	85.62±0.13	75.38±0.08	80.81±0.07	63.56±0.42	80.60±0.13
Masking Ratio 10%	78.41±0.06	85.48±0.13	75.21±0.12	80.82±0.07	63.44±0.23	80.27±0.17
*Masking Ratio 20%	78.74±0.12	85.77±0.08	75.57±0.13	81.08±0.09	63.89±0.23	80.62±0.10
Masking Ratio 25%	78.64±0.06	85.94±0.15	75.24±0.11	80.96±0.08	64.22±0.19	80.44±0.09
Masking Ratio 30%	78.56±0.06	85.74±0.19	75.41±0.11	80.89±0.03	63.94±0.42	80.42±0.14
Masking Ratio 50%	78.23±0.02	85.21±0.20	74.72±0.12	80.56±0.02	63.28±0.32	79.78±0.12
MR 20% + PLOS	78.66±0.13	85.68±0.08	75.69±0.10	80.94±0.15	63.44±0.45	80.65±0.13
Max Pooling	78.71±0.04	85.74±0.06	75.50±0.09	81.12±0.03	64.13±0.15	80.66±0.09
Mean Pooling	78.66±0.06	85.94±0.14	75.69±0.11	81.12±0.03	63.94±0.33	80.89±0.11
Sum Pooling	78.16±0.15	83.45±1.10	73.37±1.04	80.24±0.21	56.87±2.03	78.05±0.90
Attention Weighted	78.84±0.08	85.86±0.08	75.74±0.10	81.22±0.09	63.63±0.13	80.84±0.11
BiLSTM	78.74±0.06	85.74±0.08	75.54±0.06	81.18±0.09	64.22±0.20	80.62±0.09
*BiGRU	78.69±0.13	85.94±0.14	75.70±0.09	81.22±0.11	63.94±0.33	80.89±0.09
CORE-BEHRT						

M. Definitions of Conditions

Table 7: Table of conditions and their corresponding codes as given in the SKS-browser, sorted alphabetically by condition.

Condition	Codes
Alzheimer's Disease	DG30, DF00
Anaphylaxis	DT782, DT788, DT789, DT882, DT886, DT881
Arrhythmia	DI47, DI49, DR00, MC01A, MC01B, Abnorm hjerterytme
Basal Cell Carcinoma	DC44, anden hudkræft exclude: Unspecified Cancer
Breast Cancer	DC50, brystkræft exclude: Unspecified Cancer
Colon Cancer	DC18, kræft i tyktarmen exclude: Unspecified Cancer
Depression	Depression, DF32, DF33
Diabetes	MA10BA02, MA10BD07, MA10BD08, MA10BD10, MA10BD11, MA10BD13, MA10BD19, MA10BD21, MA10BD24, MA10BH01, MA10BH02, MA10BH03, MA10BH04, MA10BH05, MA10A, MA10BB, MA10BK, MA10BJ, DE10, DE11, diabetes
Diarrhea	DK529, DK521, MA07D
Fetal Growth Restriction	DO365
Fetal Macrosomia	DO366
Lung Cancer	DC34, Kræft i bronkier eller lunge med metastaser, Kræft i lungens, kræft i lunge exclude: Unspecified Cancer
Major bleeding	DD62, DI60, DI61, DI62, DI690, DI691, DJ942, DK250, DK252, DK254, DK256, DK260, DK262, DK264, DK276, DK270, DK272, DK274, DK276, DK280, DK282, DK284, DK286, DK290, DK298A, DK625, DK638C, DK920, DK921, DK922, DN02, DN93, DR04, DR31, DS064, DS065, DS066
Myocardial Infarction	DI21, Komplikationer i efterforløbet af akut myokardieinfarkt
Osteonecrosis	DM87
Pain Treatment	MN02A, MN02B, MM02A
Pancreatic Cancer	DC25, Kræft i caput pancreatis med metastaser, Kræft i cauda pancreatis, Kræft i corpus pancreatis exclude: Unspecified Cancer
Placental Insufficiency	DO365A
Pregnancy	DZ340, DZ348, DZ349
Preterm Birth	DO42, DO47, DO60
Schizophrenia	DF20
Sleep Disorder	DF51, DG47, MN05CF, MN05CH
Stroke	DI64, Slagtilfælde uden oplysning om blødning eller infarkt

Table 7 (continued)

Condition	Codes
Unspecified Cancer	Kræftsygdom UNS med metastaser, Kræft UNS overgribende flere lokalisationer med metastaser
Infection	DA, DB, DD738C, DD738D, DD762, DE060B, DG00-DG07, DG531, DG630, DG734, DG940, DI00-DI02, DI30, DI320, DI321, DI33, DI38, DI39, DI400-DI412, DI430, DI520, DI521, DK112, DK113, DK12, DK140, DK230, DK770, DK858E, DK930, DM00, DM01, DM03, DM461-DM465, DM490-DM493, DM600, DM630, DM632, DM650, DM651, DM680, DM710, DM711, DM726, DM730, DM731, DM86, DM900-DM902, DN61, DN70-DN77, MJ01 DN080, DN109A-DN109C, DN12, DN136, DN151, DN160, DN288D, DN288E, DN288F, DN290, DN291, DN30, DN330, DN340, DN341, DN342A, DN342B, DN390, DN412, DDN45, DN481, DN482, DN49, DN511, DN512 DA0, DK35, DDK37, DK57, DK65, DK67, MJ01, DH60, DH610, DH620, DH621, DH623, DH65, DH66, DH670, DH671, DH68, DH70, DH730, DH750, DH830, DH940 DA46, DL00-DL08, DL303, DL738H, DL88 DA541, DB43, DD733, DE060A, DE236A, DE321, DG06, DG07, DH000A, DH050A, DH440A, DH600, DJ340A, DJ36, DJ383D, DJ387G, DJ390, DJ391, DJ398A, DJ851, DJ852, DJ853, DK113, DK122, DK130A, DK140A, DK209A, DK353A, DK353B, DK570, DK572, DK574, DK578, DK61, DK630, DK650, DK750, DK810A, DK858A, DL02, DL050, DL059, DM608A, DM868A, DM869A, DN151, DN340, DN412, DN450, DN482, DN492A, DN619A, DN619B, DN700A, DN700B, DN710A, DN730A, DN730B, DN732A, DN732B, DN733A, DN735A, DN738A, DN738C, DN751, DN764, DN768A, 'infektiøse', 'infektion', exclude: DA06, DA07, DA33, DA630, DA65, DA66, DA67, DA68, DA69, DA881, DB079D, DB18, DB35, DB36, DB37, DB38, DB39, DB4-DB7, DB80-DB83, DB85-DB92, DB94, DI300, DI521C, DK112B, DM031A, DM036, DM863, DM864, DM865, DM866 DN701, DN702, DN711, DN729B, DN729E, DN731, DN734, DN736, DN738B, DN75, DN750, DN758, DN758B, DN759, DN761A, DN761C, DN761E, DN763A, DN765, DN766 DN301, DN302, DN304, DN308, DN459B, DN481D DA06, DA07, DK573C, DK573D, DK574, DK574A, DK575, DK578, DK579, DK658A, DK658C, DK658E, DK658F, DK658G, DK658H DH604, DH604A, DH605, DH605B, DH608, DH608A, DH652, DH653, DH654, DH654C, DH661, DH662, DH663, DH681, DH701, DH708 DJ340E, DJ340F, DJ340G, DJ340H, DL889, DL889C, DA541B, DB430, DB438, DB439, DK570B, DK570C, DK572B, DK572C, DK574A, DK650M, DK650N, DK650O, DK650P

Chapter 4

Intermezzo: Causal Inference for EHR

This chapter links the representation work in chapter 3 to the two causal applications. It is intentionally concise and practical. Rather than providing a full introduction to causal inference, we (i) state the causal estimands and assumptions (ii) explain how design choices and transformer-based embeddings make these assumptions plausible.

4.1 Identification and Target-Trial Design

We denote the (potential) outcome under treatment $A \in \{0, 1\}$ with $Y(a)$ and target the ATE [58]

$$\tau = \mathbb{E}[Y(1) - Y(0)].$$

Intuitively, τ answers the “what if” question: *What would the average outcome have been if everyone had received treatment, minus what it would have been if no one had?*[59] In a randomised controlled trial (RCT), randomisation guarantees that any difference in outcomes between treatment and control can be interpreted causally. In observational data, we do not control treatment assignment. Instead, we try to emulate an RCT by adjusting for variables that jointly predict treatment and outcome, typically through the propensity score. Identification of τ from observational data requires three standard conditions [60, 61, 58]:

1. **Consistency:** observed and potential outcome under observed treatment are the same and the treatment of one patient has no effect on another patients outcome.
2. **Conditional exchangeability:** $Y(a) \perp A \mid X$ for $a \in \{0, 1\}$; i.e. there is no unmeasured confounding after conditioning on pre-treatment covariates X .
3. **Positivity:** $0 < \Pr(A=a \mid X=x) < 1$ for all x with $\Pr(X=x) > 0$.

We try to make these assumptions as plausible as possible by following a symmetric approach for exposed and unexposed patients:

- **New-user design:** we focus on new initiators of semaglutide and corresponding non-users and define time zero at treatment initiation (or an assigned pseudo-index date for non-initiators).

- **Non-exposed control group:** we apply inclusion/exclusion criteria symmetrically.
- **Pre-specified index and grace period:** all adjustment variables are constructed from a fixed pre-index window. Grace-period rules are applied identically in both arms.
- **As-treated follow-up and synchronized censoring:** follow-up starts after a short lag to reduce reverse causation, and censoring rules are mirrored between exposed and unexposed to avoid immortal-time bias and differential follow-up.

Crucially, *all* covariates used for adjustment are constructed from *strictly pre-index* histories. This reduces the risk of controlling for mediators or post-treatment colliders and leaves residual bias mainly driven by truly unmeasured or poorly proxied confounders rather than by avoidable design flaws.

Semaglutide context. For the SUSTAIN-6 emulation, we mirror the RCT contrast (semaglutide vs. placebo) by using a pre-specified active comparator in EHR, aligning eligibility criteria where feasible, and summarising composite major adverse cardiovascular events over a fixed horizon [62]. The work in chapter 6 reuses the same indexing and censoring logic but extends it to hundreds of downstream medication outcomes with formal multiplicity control.

4.2 Confounding, Pre-treatment Variables, and Proxies

The main threats to identification in EHR-based studies are **unobserved confounding** and violations of positivity. Many clinically important variables (frailty, lifestyle, adherence, socioeconomic status) are not directly recorded. Our strategy is therefore to:

1. Include a *broad* set of pre-treatment events (diagnoses, medications, procedures, labs, contact types) rather than hand-picking a small set of confounders
2. Let flexible models use these histories to construct proxy-adjusted summaries of patient state.

In practice, it is very hard to assert that a pre-treatment variable is *not* linked to either treatment choice or outcome. We therefore avoid trying to separate pre-treatment variables into “confounders” and “non-confounders” by hand. Instead, we work under two heuristics:

- **Adjustment by proxy.** Under-coded or unmeasured causes of treatment and outcome (e.g. disease severity, adherence, lifestyle) often leave patterns in related events: characteristic lab trajectories, diagnostic clusters, or combinations of medications. By adjusting for rich pre-index histories, we hope to capture these latent factors indirectly, even when they are not explicitly labelled in the data.

- **Pure instruments are rare.** Many pre-treatment events that look like “instruments” at first sight (for example, a particular hospital’s preferred protocol) can still influence outcomes through intermediate pathways. We assume that variables with *zero* causal relation to the outcome are uncommon in real EHR data, so the main risk of including many pre-treatment variables is variance inflation rather than large additional bias. This motivates careful regularisation and weighting diagnostics, but not aggressive manual exclusion of candidate covariates.

Two caveats are worth noting. First, EHR coding is delayed and incomplete: some diagnoses are recorded only after treatment initiation, so even “pre-treatment” windows may miss relevant information. Second, prescription and redemption times do not always align with true treatment start. We address these issues with conservative censoring lags and by treating this chapter’s assumptions as working models rather than guarantees.

4.3 Nuisance Modelling with Transformer Representations

4.3.1 Embeddings as high-dimensional confounder summaries

Pretraining on longitudinal EHR sequences (e.g. via masked-language modelling) encourages the encoder to capture co-occurrence and temporal structure across diagnoses, medications, procedures, labs, and encounters. The resulting patient representation $\phi(X)$ summarises each history into a dense vector that can be used as input to treatment and outcome models. When proxies for latent confounders are strong, adjusting for functions of $\phi(X)$ can reduce residual confounding even if some causes are never explicitly labelled [63, 40]. This follows a similar logic to high-dimensional propensity score approaches, which algorithmically construct large sets of claims-based proxy covariates for confounding adjustment [64].

This does not remove the need for the identification assumptions in section 4.1: if important determinants of treatment and outcome are completely absent from recorded history, no representation can recover them. But it does allow us to move from dozens of hand-crafted covariates to rich, automatically learned summaries that are more likely to contain useful proxies.

4.3.2 Confounder-focused training

To further align the representations with the causal task, we fine-tune the shared encoder *jointly* on treatment and outcomes:

- one head predicts semaglutide initiation $\Pr(A=1 \mid X)$ (propensity model), and
- additional heads predict the risk of each outcome $\Pr(Y_k \mid A, X)$.

Gradients from all heads update the same encoder, so features that help predict *both* exposure and outcomes are reinforced. Intuitively, such features behave like confounders. This is only a heuristic, as we cannot force the network to ignore all instrument-like patterns, but it gives a more principled way of focusing model capacity on clinically relevant structure than training a propensity model in isolation.

We distinguish between:

- **Identification failures** (e.g. unrecorded confounders, interference), which no amount of modelling can fix; and
- **Estimation misspecification**, which can be reduced with high-capacity, regularised learners for the propensity score

4.4 Estimators Used in Practice

In all applied chapters of this thesis, causal effects are estimated using **inverse-probability weighting** (IPW) based on transformer-derived propensity scores. Let $\hat{g}(X)$ denote the estimated propensity score.

IPW and overlap weights. We construct stabilised IPW weights from $\hat{g}(X)$ and use them in weighted analyses to estimate per-arm risks, risk differences, and risk ratios. We trim extreme propensity scores as well as scores that are outside the overlap region $\hat{g}(X)$ or use *overlap weights*, which target the clinical overlap population and improve stability in the tails [65, 66].

Beyond IPW. G-computation and doubly robust estimators (AIPW/TMLE) are natural extensions when reliable outcome models are available and can offer improved robustness and efficiency [67, 68, 69]. However, they are *not* used in the main analyses reported in Chapters 5 and 6, so we do not describe them in detail here.

4.5 Links to the Applied Chapters

This intermezzo underpins the two causal applications in the second part of the thesis:

- **SUSTAIN-6 emulation (chapter 5).** The target-trial framework in section 4.1, the pre-treatment proxy strategy in section 4.2, and transformer-based propensity estimation in section 4.3 are instantiated for a single, clinically focused question: The cardiovascular effects of semaglutide in high-risk patients with type 2 diabetes.
- **Untargeted Screening for Effects and Side-effects (chapter 6)** The same identification strategy and nuisance modelling are reused, but scaled up to hundreds of medication outcomes. IPW with calibrated transformer-based propensity scores.

The message of this chapter is deliberately modest: transformers do not solve causal inference. They give us flexible nuisance models and high-dimensional proxies. Combined with transparent design and restriction to pre-treatment histories, they make large-scale confounding control in EHR data more tractable, but not assumption-free.

Chapter 5

An Introduction to Causal Transformer Models in Epidemiological Studies

The following manuscript is being prepared for a submission to the "Education Corner" of the International Journal of Epidemiology.

An Introduction to Causal Transformer Models in Epidemiological Studies

Kiril Vadimovic Klein¹

KIKL@DI.KU.DK

Josefine Windfeld Mathiasen²

JOSEFINE.WINDFELD-MATHIASEN@REGIONH.DK

Espen Jimenez-Solem^{2,3,4}

ESPEN.JIMENEZ.SOLEM@REGIONH.DK

Thomas Leth Jensen²

THOMAS.LETH.JENSEN@REGIONH.DK

Kasper Bonnesen^{2,6}

KASPER.BONNESEN@REGIONH.DK

Martin Hylleholdt Sillesen⁵

MARTIN.HYLLEHOLT.SILLESEN@REGIONH.DK

Mads Nielsen¹

MADSN@DI.KU.DK

Sanne Møller Thysen²

SANNE.THYSEN@REGIONH.DK

¹*Department of Computer Science, University of Copenhagen*

²*Department of Clinical Pharmacology, Copenhagen University Hospital - Bispebjerg and Frederiksberg*

³*Copenhagen Phase IV Unit (Phase4CPH), Department of Clinical Pharmacology, Center for Clinical Research and Prevention, Copenhagen University Hospital - Bispebjerg and Frederiksberg*

⁴*Department of Clinical Medicine, University of Copenhagen*

⁵*Department of Organ Surgery and Transplantation, Copenhagen University Hospital - Rigshospitalet*

⁶*Department of Clinical Medicine, Aarhus University*

1. Abstract

Transformer models are increasingly used to analyse longitudinal electronic health records (EHRs), but the underlying concepts and their practical application often appear inaccessible to epidemiologists and clinicians. This paper provides a clear and practice-oriented introduction to transformer models, explains how they can be integrated with established causal inference methods, and illustrates their use in health outcomes research. We outline the main model components in straightforward terms and describe strengths and limitations relevant to epidemiological analyses.

To demonstrate the method in a realistic setting, we conducted a target trial emulation of the SUSTAIN-6 cardiovascular outcomes trial [Marso et al. \(2016\)](#) using EHR data from two Danish regions. Eligibility criteria were operationalised using the PICOTS framework, semaglutide initiators were compared with non-exposed individuals assigned index dates based on the exposure-date distribution of treated patients, and transformer-based propensity scores were used with inverse probability weighting. Five outcomes were evaluated: all-cause death, a composite of death, myocardial infarction, and stroke, heart-failure hospital admission, non-fatal myocardial infarction, and non-fatal stroke.

A total of 31,150 individuals (11,454 exposed/ 19,696 unexposed) entered the emulated cohort. Semaglutide was associated with reduced all-cause mortality and with a beneficial effect on the composite outcome. In contrast, no clear associations were seen for non-fatal myocardial infarction or non-fatal stroke, indicating that the composite effect in our emulation was driven primarily by mortality rather than non-fatal events.

By providing an accessible introduction and a worked example, this paper aims to lower the barrier to adopting transformer models in epidemiology and to highlight their potential for applications such as adverse drug reaction signal detection and subgroup-specific safety assessment.

2. Introduction

Recent advances in machine learning have created new opportunities for analysing complex, longitudinal health data. Among these, transformer models have attracted increasing attention because of their ability to identify temporal and contextual patterns in electronic health records (EHRs) and other real-world data sources [Li et al. \(2020\)](#); [Rasmy et al. \(2021\)](#); [Guo et al. \(2023\)](#). Their use in health research is growing rapidly, extending from diagnostic prediction to treatment response modelling [Hakeem et al. \(2022\)](#); [Gu and Yu \(2025\)](#).

In healthcare, well-established epidemiological methods remain essential for evaluating treatment effectiveness, drug safety, and health policy evaluations [Schneeweiss et al. \(2023\)](#); [Narita et al. \(2023\)](#). However, these approaches often require manual identification and adjustment for confounders, which can condense nuanced longitudinal EHR data into a limited number of summary variables. While existing high-dimensional approaches [Schneeweiss et al. \(2009\)](#) partially address this limitation by expanding the set of covariates, they still struggle to fully exploit the temporal structure and complexity inherent in patient health trajectories.

Transformer-based approaches offer a complementary way to address this challenge by learning directly from patient trajectories without requiring extensive feature engineering [Yang et al. \(2023\)](#); [Kraljevic et al. \(2024\)](#). Rather than replacing classical epidemiological methods, they can strengthen them by providing more detailed representations of patient histories and uncovering complex associations that may inform causal analyses.

In this paper, we aim to provide an intuitive introduction to transformer models and illustrate their application through an applied example: a target trial emulation of the SUSTAIN-6 trial using real-world data. Our aim is to make this emerging methodology accessible to epidemiologists and clinicians by demonstrating both its potential and its limitations in causal inference and health outcomes research.

What are transformers?

Transformer models are a type of machine learning method developed to analyse sequences of data. Transformers were originally designed for natural language processing [Vaswani et al. \(2017\)](#), but are now increasingly used in health research to model patient histories based on EHRs [Kraljevic et al. \(2024\)](#); [Li et al. \(2020\)](#); [Odgaard et al. \(2024\)](#); [Rasmy et al. \(2021\)](#). Unlike traditional models that rely on manually selected variables, transformers can learn directly from raw, time-ordered data. The central innovation of the transformer is the attention mechanism, which allows the model to weigh the importance of each event in a sequence, such as diagnoses, prescriptions, or test results, when making predictions. This enables the model to capture complex, time-dependent relationships across a patient’s clinical history. Each clinical event is represented as a ”token”, similar to a word in a sentence, and the model learns how these events relate to one another over time. Transformers are typically pre-trained on large datasets to learn general patterns, and then fine-tuned for specific tasks such as risk prediction, drug safety assessment, or treatment effect estimation. Transformers are not a replacement for classical epidemiological methods, but offer a complementary approach that makes use of high-dimensional, time-ordered health data in a more flexible and automated manner.

3. Modelling EHR complexity with transformer models

Transformer models (transformers) are a type of deep learning algorithm for processing sequences of information, such as EHR data. They are increasingly used in healthcare research because they can model complex patient histories in a way that captures relationships between events over time, without the need for hand-crafted rules [Rasmy et al. \(2021\)](#); [Li et al. \(2020\)](#); [Odgaard et al. \(2024\)](#). Earlier machine learning models were also able to model longitudinal data. However, these suffered from unstable convergence and slow training times, making them unsuitable for large datasets [Choi et al. \(2016\)](#); [Pascanu et al. \(2013\)](#). Since the introduction of transformers in 2017 [Vaswani et al. \(2017\)](#), they have become a core technology in natural language processing, powering applications like ChatGPT [Achiam et al. \(2023\)](#). EHR data exhibit a close structural similarity to natural language, where each medical event can be seen as a word and each patient as a document [Rasmy et al. \(2021\)](#). Given this similarity, transformers offer a clinically meaningful way to model patient histories.

Before any prediction task can be performed, the model first converts each clinical event into a numerical representation known as an embedding. This process maps diagnoses, medications, laboratory results, patient age, and timing information into a shared vector space, allowing the model to treat heterogeneous clinical information in a unified format. The aim of this representation learning is to capture the underlying factors that jointly influence treatment assignment and outcomes. In practice, this means that potential confounders become encoded in the model’s internal representation rather than being specified manually, which allows the transformer to use the full patient trajectory when estimating risks or treatment effects.

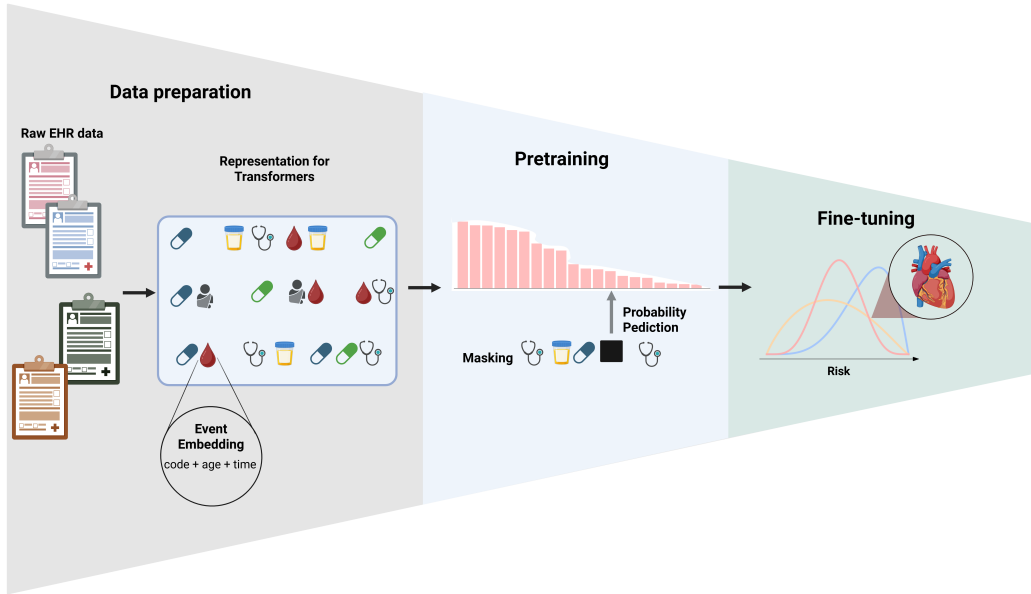


Figure 1: An overview of the stages involved in obtaining risk predictions using a transformer. Raw EHR records are transformed into sequences of clinical events, where each event embedding combines code, patient age, and time (Further details can be found in Data preparation). The model is pretrained with masked-language-modelling to learn context-aware representations from event co-occurrence and temporal structure (Pretraining). The pretrained encoder is then fine-tuned on cohort-specific tasks, such as disease risk prediction, using labelled outcomes (Fine-tuning).

3.1. Attention Mechanism

At the core of the transformer is the attention mechanism, which functions as a dynamic relevance map over the patient’s history. When analyzing a specific event, such as heart failure, the model considers all other events and assigns a weight reflecting its importance of each event in relation to heart failure. For example, a prior diagnosis of hypertension might receive a high weight, while a flu infection might receive a low one. The model’s understanding of heart failure is then updated based on these weights, focusing on relevant events while assigning low priority to the rest of the patient’s history. Crucially, timing is part of the signal because the model encodes when events occurred using representations of temporal relationships. This allows it to learn patterns where recent events are more informative, while still recognizing that distant but critical events can remain highly relevant. This selective focus enables transformers to handle common EHR characteristics such as irregular timing, repeated events, and long periods between events [Wornow et al. \(2025\)](#).

4. Transformer workflow from raw data to clinical predictions

Applying a transformer model to EHR data follows a three-step procedure: data preparation, pre-training, and fine-tuning (See Figure 1).

4.1. Data preparation

Transformers enable the natural representation of EHRs as sequences, preserving temporal relationships with minimal preprocessing. Each patient becomes a sequence of medical events (diagnoses, prescriptions, laboratory tests, etc.), where each event is characterized by a code and timestamp. Events are simply arranged chronologically and fed into the transformer. It is often beneficial for the model to include additional features, such as age at each event. This straightforward approach allows transformers to directly process the inherent sequential structure of medical records.

Technical details of the practical application of transformers on EHR data

While the basic approach requires minimal preprocessing, some practical considerations may arise. Modern transformers for EHRs handle sequences with lengths of up to 2048 tokens [Wornow et al. \(2025\)](#); [Renc et al. \(2024\)](#). However, patients with extensive medical histories may require sequence truncation, leading to loss of information. Common strategies for this include removing the oldest events or prioritizing by event type (e.g., removing visit information before procedures). Repeated high-frequency events, such as frequently logged hospital medications, can be consolidated into a single representation. This prevents the patient record from being overloaded with near-identical entries and allows the model to focus on the clinically important information rather than technical repetitions. More recent methods can handle longer sequences [Beltagy et al. \(2020\)](#); [Dai et al. \(2019\)](#), but truncation may still be required for patients with too many events. Representation of numerical values (laboratory results) remains an active area of research. However, the fundamental advantage remains: We can feed EHRs in their natural sequential form into the transformer.

4.2. Pre-training

The goal of pre-training is to learn a general understanding of the EHR language, which can then be used for specific clinical tasks. During pre-training, the model learns by playing a “fill-in-the-blanks” game with healthcare data. It is given a patient’s entire medical history, but with several events randomly hidden or “masked”. To achieve this, the model needs to develop an understanding of the medical codes and their relations. For example, it has to discover that if it sees Metformin and an HbA1c test, the missing diagnosis is likely to be type 2 diabetes.

The result of this learning process is that the model builds its internal map of medicine. On this map, every clinical concept (like a diagnosis or medication) has a location, represented by a numerical “embedding”. Related concepts like Metformin and Diabetes are placed close together, while unrelated ones like Asthma and knee surgery are far apart. An example of this can be seen in BEHRT [Li et al. \(2020\)](#). This elegant approach, where the model learns from the data itself without supervision or manually curated labels, is formally known as self-supervised learning.

What makes pre-training so attractive is that a model pre-trained once can be reused across many downstream tasks and specialized with only a few hundred examples while showing better performance than training from scratch [Guo et al. \(2023, 2024\)](#); [Wornow et al.](#)

(2023). More data and bigger models lead to a better understanding of medical patterns and greater benefits from pre-training [Kaplan et al. \(2020\)](#). Therefore, pre-training should use as much patient data as possible.

This has led EHR modelling research towards “foundation models”, which are deep neural networks pre-trained on massive, multimodal datasets combining structured data with medical images [Dai et al. \(2025\)](#); [Huo et al. \(2024\)](#), clinical notes [Lyu et al. \(2022\)](#), and omics data [Ren et al. \(2024\)](#). Foundation models promise to replicate the success of large language models by developing a comprehensive understanding of the data with capabilities far beyond what small, specialized models can achieve.

Technical details on pre-training

To train on large datasets, pre-training should not require manually provided labels. A common pre-training technique for sequences is masked-language modelling (MLM) [Odgaard et al. \(2024\)](#); [Rasmy et al. \(2021\)](#); [Li et al. \(2020\)](#) where events are randomly removed from a patient’s sequences and the model learns to predict the missing events from context. More recent methods use next-word prediction (NWP) [Renc et al. \(2024\)](#); [Yang et al. \(2023\)](#); [Kraljevic et al. \(2024\)](#), predicting the most likely next event from all previous events.

The choice between MLM and NWP depends on the downstream task. MLM develops a diagnostic understanding of a patient’s current state, while NWP focuses on prognostic understanding by modelling future trajectories. During next word predictions, the model predicts the next event given all previous events. E.g., we give it the first event, and it tries to predict the second, then we give it the first two, and it tries to predict the third, etc. Both approaches enable self-supervised learning that scales to massive datasets without manual annotation.

4.3. Fine-tuning

The goal of fine-tuning is to adapt the pre-trained model to answer specific clinical questions. Having learned the general meaning of EHR data, the model can now be efficiently specialized for particular tasks. For example, to predict 3-year heart failure risk, the model is trained on pairs of patient history (censored up to an index date) and the label of whether the heart failure occurred.

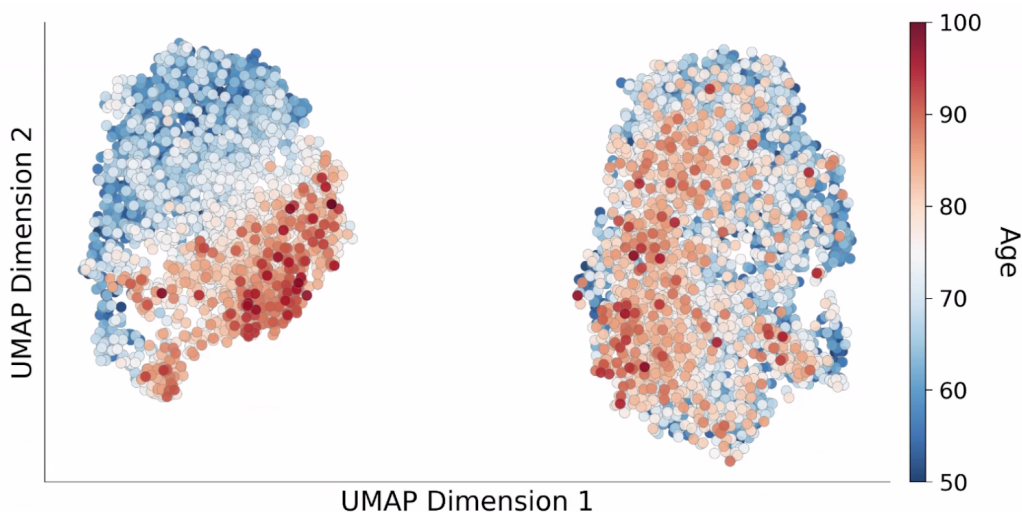


Figure 2: Patient embeddings obtained from a fine-tuned transformer model are visualized in two dimensions using a dimensionality reduction technique called UMAP [McInnes et al. \(2018\)](#). In this representation, each patient is represented by one point, colored by their age.

An important aspect of this process is to specialize the model without losing the knowledge acquired during pre-training. The parameters of the core transformer are only slightly adjusted, while a new lightweight “prediction head” is added on top of the model. This component’s sole purpose is to “interpret” the patient’s history provided by the transformer for the current task. This approach ensures that the model’s general understanding of EHRs remains intact while becoming an expert on a new task.

To illustrate what the model learns in this process, Figure 2 shows a two-dimensional projection of the patient embeddings after fine-tuning for semaglutide treatment, coloured by age at the index date. We observe that the model learns a meaningful representation with regards to age, separating patients by age. This showcases that the model organises patients along clinically meaningful dimensions rather than arbitrary artefacts. In other words, fine-tuning not only adapts the model to the specific outcome but also preserves and sharpens latent features (such as age and related comorbidity burden) that are central for risk stratification.

4.4. Integrating transformers with causal inference

With the fundamentals of transformers established, we now turn to how these models can be integrated into causal effect estimation on observational data. In traditional epidemiological analyses, confounding is often addressed using propensity-score (PS)-based methods such as PS matching or inverse probability weighting (IPW). These approaches depend on correctly specifying the PS model, and bias can occur if the model does not adequately capture the relationships between confounders, treatment, and outcome [Waernbaum and Pazzagli \(2023\)](#); [Narita et al. \(2023\)](#). To improve robustness, doubly robust estimators such as Targeted Maximum Likelihood Estimation (TMLE) [Van der Laan et al. \(2011\)](#) were developed to jointly model the treatment assignment and the outcome, and thus be more resilient to this issue.

Transformers can enhance causal inference in two main ways. First, they can serve as drop-in replacements for traditional PS models. Instead of requiring investigators to specify which variables and interactions to include, transformers learn a representation of the full patient history in a high-dimensional vector space. In this representation, the model identifies patterns that are simultaneously predictive of treatment assignment and of the outcome, thereby capturing information that functions as potential confounders. This reduces the risk of misspecification bias, because the model does not depend on manually chosen covariates or functional forms. However, as with all observational methods, this approach still relies on relevant confounders being present and recorded in the data. Second, transformers can be combined with doubly-robust methods like TMLE [Rao et al. \(2024\)](#). In these approaches, the same model architecture is used to learn both the PS and the outcome model, which further encourages the internal representation to focus on features that relate to both exposure and outcome. This joint learning can help the model capture confounding structures more effectively, resulting in more accurate and stable estimates of treatment effects [Rao et al. \(2024\)](#).

How transformers account for confounders

Transformers do not identify confounders one by one. Instead, they learn patterns in the full patient history that are linked to both the treatment a patient receives and the outcome that follows. During pre-training, the model learns general relationships between clinical events across many patients. During fine-tuning, the model is trained on the specific treatment and outcome of interest. Across these two stages, the model embeds shared patterns into a single representation. These shared patterns function as confounders because they influence both the exposure and the outcome. Because this representation captures these joint predictors automatically, confounders do not need to be specified manually. This learned representation can then be used within standard epidemiological methods such as IPW or TMLE, providing more complete adjustment and reducing misspecification bias. However, transformers do not distinguish between confounders, mediators, or colliders. They treat all events in the patient history as inputs. To avoid adjusting for mediators, the PS model must only include pre-exposure information, while the outcome model may include post-exposure events if appropriate for the estimand. In practice, this means structuring the EHR sequence so that the PS model receives only events recorded before treatment initiation. Correct handling of mediators therefore depends on study design and preprocessing, not on the transformer architecture itself.

5. Emulating the SUSTAIN-6 trial using transformer models

SUSTAIN-6 [Marso et al. \(2016\)](#) was a randomised, placebo-controlled trial that evaluated cardiovascular effects of semaglutide in adults with type 2 diabetes at high cardiovascular risk. Semaglutide is a GLP-1 receptor agonist approved for the treatment of diabetes [Agency \(a\)](#) and, more recently, for weight management in individuals with obesity [Agency \(b\)](#). The original trial compared once-weekly semaglutide with placebo over a two-year follow-up period. To illustrate how transformers can be used with healthcare data, we conducted a

target trial emulation (TTE) of the SUSTAIN-6 trial using Danish EHR data from 2,218,028 patients in the Capital Region and Region Zealand. Following the PICOTS framework, we applied the SUSTAIN-6 eligibility criteria where feasible (see [subsection A.5](#)). The population included adults with type 2 diabetes who met the operationalised inclusion and exclusion criteria, see tables [S5](#), [S6](#) and [S7](#). The intervention was initiation of semaglutide (ATC A10BJ02), and the comparator was a cohort of non-exposed adults assigned index dates based on the exposure-date distribution of semaglutide initiators. Five outcomes were examined: all-cause death, composite death/myocardial infarction/stroke, heart-failure hospital admission, non-fatal myocardial infarction, and non-fatal stroke. Major Adverse Cardiovascular Events (primary outcome of SUSTAIN-6) could not be investigated as the cause of death could not be reliably identified. Timing, lookback windows, and follow-up definitions followed SUSTAIN-6 where possible (see tables [S5](#), [S6](#) and [S7](#)). The cohort flow is shown in [Figure S1](#). We adjusted for confounding using inverse probability of treatment weighting (IPW), with propensity scores estimated using a transformer. For more details, see [subsubsection A.2.4](#).

After applying all eligibility criteria, a total of 31,150 (11,454 cases and 19,696 controls) individuals were included in the TTE cohort. All effects are reported as risk differences (RD) with 95% confidence intervals in the format RD [CI_{lower}; CI_{upper}]. The IPW-adjusted estimates showed a reduction in all-cause mortality (-0.040 [-0.044 ; -0.035]) and in the composite outcome of death, myocardial infarction, and stroke (-0.038 [-0.045 ; -0.032]). For hospital admission with heart failure (0.000 [-0.001 ; 0.001]), non-fatal myocardial infarction (-0.001 [-0.004 ; 0.002]), and non-fatal stroke (-0.002 [-0.005 ; 0.002]), the risk differences were small with confidence intervals spanning 0, consistent with no clear effect.

SUSTAIN-6 demonstrated that semaglutide reduced the primary composite cardiovascular endpoint (cardiovascular death, non-fatal myocardial infarction, or non-fatal stroke) compared with placebo (RD -0.022 [-0.041 ; -0.003]). Using published SUSTAIN-6 data, we estimated risk differences for hospitalisation with heart failure (0.003 [-0.009 ; 0.015]), all-cause death (0.001 [-0.012 ; 0.014]), non-fatal myocardial infarction (-0.010 [-0.023 ; 0.002]), and non-fatal stroke (-0.010 [-0.020 ; 0.000]), which were all close to zero with relatively wide confidence intervals. Our TTE reproduced the beneficial effect on the composite outcome but did not detect clear effects on non-fatal myocardial infarction or non-fatal stroke. The reduction in all-cause mortality observed in our TTE is notable, as all-cause mortality was not a prespecified endpoint in SUSTAIN-6 [Marso et al. \(2016\)](#).

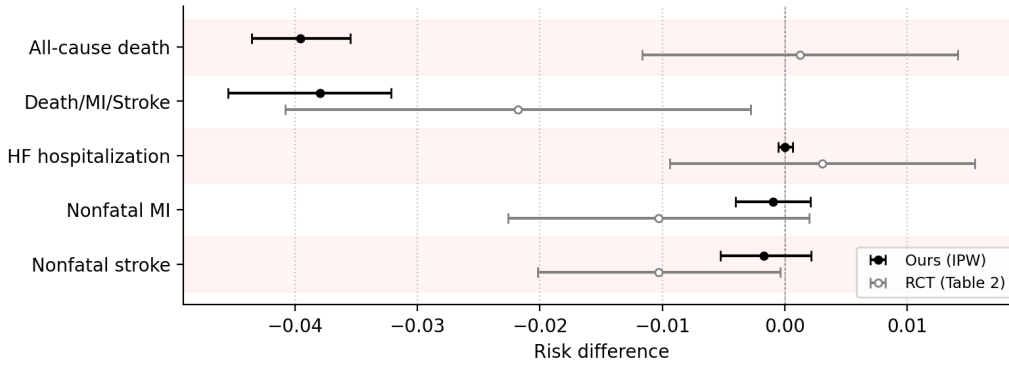


Figure 3: Adjusted risk difference for the target trial emulation using IPW, with propensity score estimated with a transformer model.

Our findings are consistent with emerging real-world evidence. A recent large-scale EHR-based emulation of SUSTAIN-6 also found reductions in all-cause mortality when semaglutide was compared with sitagliptin as a replacement for placebo [Krüger et al. \(2025\)](#). Taken together, these results show that transformer-based TTEs can reproduce key findings from the original trial, while also capturing mortality patterns that may become apparent in broader real-world populations.

6. Strengths and limitations of transformers

Transformers offer important opportunities for working with health data, but they also present specific limitations that must be considered. In the following, we outline the key strengths and challenges relevant to epidemiological applications.

6.1. Strengths

Minimal Preprocessing Transformers require minimal preprocessing of EHR data, as they use the complete patient history and the temporal sequence of events. This avoids collapsing longitudinal information into static variables and reduces the need for imputation or hand-crafted data transformation.

Explainability Although deep learning models, and transformers in particular, are often seen as black-box models, sophisticated techniques exist to explain their predictions. Notable examples include SHAP values [Lundberg and Lee \(2017\)](#) and gradient-based approaches [Chefer et al. \(2021\)](#), both of which provide attributions of the output score to each input value, thereby clarifying how much each input value contributed to the prediction. This provides researchers and clinicians with insights into which aspects of the data the model focuses on. Furthermore, the attention scores themselves learned by the transformer during pre-training and fine-tuning can serve as a tool for interpretability, showing which tokens the model attends to and how tokens attend to each other [Li et al. \(2020\)](#); [Rasmy et al. \(2021\)](#). These tools have the potential to increase trustworthiness in the model’s predictions and lead to more widespread adaptation. However, to make these truly valuable in the clinical setting, a close collaboration between data scientists and clinicians is needed.

Missingness of Data Because transformers can take sequences of events, missingness is handled as a natural characteristic of the data. The model does not expect every feature to be present for every patient. Missing values or unmeasured events simply appear as absent tokens, and the model focuses on observed parts of the history. This is in contrast to many traditional methods that require explicit imputation prior to analysis.

Capturing complex confounding structures Because transformers learn representations that encode patterns predictive of both treatment and outcomes, they can help capture complex combinations of confounding variables that may be difficult to specify manually. This can reduce misspecification bias when used within frameworks such as PS modelling or doubly robust estimators.

Reuse across tasks A single pre-trained transformer can be fine-tuned for multiple downstream tasks, allowing the same model to support outcome prediction, risk stratification and causal modelling. This can reduce the long-term computational burden once the initial model has been trained.

6.2. Limitations

Data and Computational Requirements The biggest strength of transformers is also one of their weaknesses. They are notoriously “data-hungry” [Rao et al. \(2025\)](#); [Guo et al. \(2023\)](#) and thus need sufficiently large datasets (at least hundreds of thousands of patients) to perform well. Training a transformer on a substantial dataset, for example, 400,000 patients, requires around 36 hours on specialized hardware [Renc et al. \(2024\)](#) and can cost around 100 USD when rented from a data center. However, this expense can be justified in a high-stakes healthcare setting and can be further mitigated by training a large model once and adapting it across many tasks and datasets. Reuse of pre-trained models can mitigate this, but may not fully remove the challenge.

Portability and generalisability Models trained in one healthcare system may not transfer seamlessly to another because of differences in coding practices, clinical workflows, and patient populations. Privacy concerns may prevent the sharing of model parameters, and distribution shifts can reduce performance. While emerging methods aim to improve model efficiency [Wan et al. \(2023\)](#), external validation remains essential. Ultimately, the choice between a large transformer and a small specialized model trained from scratch requires a pragmatic cost-benefit analysis, weighting performance gains (favour large transformer model) against computational cost, available dataset size, computational requirements (favours targeted model), and clinical needs.

Causal Validity While transformers are powerful at modelling complex sequences, they do not defy the fundamental rules of causal inference. They cannot account for unmeasured confounding and must be combined with appropriate study design choices. Moreover, transformers do not distinguish between confounders, mediators, or colliders. If post-exposure information is included inadvertently (for example, in a PS model), this can introduce bias. Proper temporal structuring of the input and transparency about causal assumptions are, therefore, critical.

Potential overfitting and stability concerns In high-dimensional settings with limited sample size, transformers may overfit or produce unstable PS. Regularisation, appropriate validation strategies, and comparison with simpler models remain important safeguards, especially when estimating causal effects.

Interpretation and clinical integration While explainability tools exist [Lundberg and Lee \(2017\)](#); [Chefer et al. \(2021\)](#); [Li et al. \(2020\)](#); [Rasmy et al. \(2021\)](#), interpreting transformer outputs in a clinically meaningful way remains challenging. Without careful collaboration between clinicians and data scientists, there is a risk of overstating the certainty of model-derived causal estimates.

7. Potential future use cases

In this paper, we have outlined the key ideas behind transformer models and shown how they can be used within familiar epidemiological workflows. Although these methods may initially seem distant from traditional approaches, transformers are already being applied in a wide range of clinical prediction tasks [Placido et al. \(2023\)](#); [Kraljevic et al. \(2024\)](#); [Zisser and Aran \(2024\)](#); [Wang and Sun \(2022\)](#); [Rupp et al. \(2023\)](#); [Yang et al. \(2023\)](#) and are beginning to be adapted for causal inference as well [Rao et al. \(2024\)](#). Their ability to use the full patient history and handle complex, irregular data makes them a flexible tool for modern health research. Building on this foundation, there are several areas where transformers may offer particular promise in the future. Below, we highlight two examples.

Signal detection for adverse drug reactions Signal detection aims to identify patterns in healthcare data that suggest a possible link between a drug and an adverse drug reaction (ADR). Traditionally, this has relied heavily on spontaneous reports from healthcare professionals or patients, which can be slow to accumulate and may not capture all relevant cases. Rich, routinely collected healthcare data offer a broader and more timely picture but can be difficult to analyse because they are high-dimensional, irregular, and recorded across multiple clinical settings. As described above, transformers are well-suited to analyse large and complex datasets and identify patterns, not obvious to researchers, clinicians, or regulators. By learning relationships across diagnoses, prescriptions, procedures, and timing, transformers may help identify and prioritise potential ADRs for further evaluation. Established epidemiological methods could be used to validate signals identified by transformer models to ensure that findings are interpreted in a clinical context. In this way, transformer-based approaches can complement existing methods, potentially identifying important safety issues earlier and more comprehensively.

Stratified safety information Patients differ widely in their likelihood of experiencing both the benefits and the risks of a treatment, yet most regulatory documents report only a single risk estimate per ADR. This obscures clinically important heterogeneity. Age, sex, comorbidities, and the use of other medicines are all relevant modifiers, but conventional methods often struggle to represent such variation at scale without extensive manual modelling.

The Transformer can help identify groups of patients who may have higher or lower risks by learning patterns in the full clinical history that are associated with treatment response or adverse outcomes. This may reveal subgroup-specific risk profiles, such as higher

susceptibility to certain ADRs among older adults with specific comorbidity patterns. Such insights could support more tailored prescribing decisions, help refine clinical guidelines, and improve shared decision-making. Used carefully and alongside traditional methods, transformers may therefore contribute to a more nuanced understanding of treatment risks across diverse patient populations.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- European Medicines Agency. Ozempic (semaglutide) EPAR. <https://www.ema.europa.eu/en/medicines/human/EPAR/ozempic>, a. European Medicines Agency, accessed 22/11/2025.
- European Medicines Agency. Wegovy (semaglutide) EPAR. <https://www.ema.europa.eu/en/medicines/human/EPAR/wegovy>, b. European Medicines Agency, accessed 22/11/2025.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. In *arXiv preprint*, 2020.
- Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.
- Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: interpretable predictive model in healthcare using reverse time attention mechanism. *CoRR*, abs/1608.05745, 2016.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019. doi: 10.18653/v1/P19-1285.
- Zihang Dai, Ye Chen, et al. CLIMB: Data foundations for large scale multimodal clinical foundation models. *arXiv preprint*, 2025.
- Zheng Gu and Junxian Yu. Transformer-based model for drug design and therapy response prediction. *Chemistry & Biology Drug Design*, 106(5):e70200, 2025. doi: 10.1111/cbdd.70200.
- Lin Lawrence Guo, Ethan Steinberg, Scott Lanyon Fleming, Jose Posada, Joshua Lemmon, Stephen R. Pfohl, Nigam Shah, Jason Fries, and Lillian Sung. EHR foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*, 13(3767):1–10, 2023. doi: 10.1038/s41598-023-30820-8.

- Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, et al. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*, 7(171):1–11, 2024. doi: 10.1038/s41746-024-01166-w.
- Haris Hakeem, Wei Feng, Zhibin Chen, Jiun Choong, Martin J. Brodie, Terence J. O’Brien, and Patrick Kwan. Development and validation of a deep learning model for predicting treatment response in patients with newly diagnosed epilepsy. *JAMA Neurology*, 79(10): 986–996, 2022. doi: 10.1001/jamaneurol.2022.2514.
- Zepeng Huo, Jason A. Fries, Alejandro Lozano, Jeya Maria Jose Valanarasu, Ethan Steinberg, Louis Blankemeier, Akshay S. Chaudhari, Curtis Langlotz, and Nigam H. Shah. Time-to-event pretraining for 3D medical imaging. *arXiv preprint*, 2024.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Zeljko Kraljevic, Daniel Bean, Anthony Shek, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024. doi: 10.1016/S2589-7500(24)00025-6.
- Nils Krüger, Sebastian Schneeweiss, Rishi J Desai, Sushama Kattinakere Sreedhara, Anna R Kehoe, Kenshiro Fuse, Georg Hahn, Heribert Schunkert, and Shirley V Wang. Cardiovascular outcomes of semaglutide and tirzepatide for patients with type 2 diabetes in clinical practice. *Nature Medicine*, pages 1–1, 2025.
- Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pages 623–631. PMLR, 2017.
- Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.
- Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. *AMIA Annual Symposium Proceedings*, 2022:719–728, 2022.
- Steven P. Marso, Stephen C. Bain, Agostino Consoli, et al. Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine*, 375(19): 1834–1844, 2016. doi: 10.1056/NEJMoa1607141.
- Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint*, 2018.

- Kaori Narita, Juan de Dios Tena, and Claudio Detotto. Causal inference with observational data: A tutorial on propensity score analysis. *The Leadership Quarterly*, 34(3):101678, 2023. doi: 10.1016/j.leaqua.2023.101678.
- Mikkel Odgaard, Kiril V. Klein, Sanne Thysen, Espen Jimenez-Solem, Martin Sillesen, and Mads Nielsen. CORE-BEHRT: A carefully optimized and rigorously evaluated BEHRT, 2024.
- Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.
- Davide Placido, Bo Yuan, Jessica X. Hjaltelin, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature Medicine*, 29(5):1113–1122, 2023. doi: 10.1038/s41591-023-02332-5.
- Shishir Rao, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Yikuan Li, Rema Ramakrishnan, Abdelaali Hassaine, Dexter Canoy, and Kazem Rahimi. Targeted-BEHRT: Deep learning for observational causal inference on longitudinal electronic health records. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5027–5038, 2024. doi: 10.1109/TNNLS.2022.3183864.
- Shishir Rao, Nouman Ahmed, Gholamreza Salimi-Khorshidi, et al. A transformer-based survival model for prediction of all-cause mortality in heart failure patients: a multi-cohort study. *arXiv preprint*, 2025.
- Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.
- Yuchen Ren, Wenwei Han, Qianyu Zhang, et al. COMET: Benchmark for comprehensive biological multi-omics evaluation tasks and language models. *arXiv preprint*, 2024.
- Pawel Renc, Yugang Jia, Anthony E. Samir, et al. Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7(256):1–11, 2024. doi: 10.1038/s41746-024-01235-0.
- Maurice Rupp, Oriane Peter, and Thirupathi Pattipaka. ExBEHRT: Extended transformer for electronic health records to predict disease subtypes and progressions. *Lecture Notes in Computer Science*, 2023. doi: 10.1007/978-3-031-39539-0_7. Also available as arXiv:2303.12364.
- Sebastian Schneeweiss, Jeremy A. Rassen, Robert J. Glynn, Jerry Avorn, Helen Mogun, and M. Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512–522, 2009. doi: 10.1097/EDE.0b013e3181a663cc.
- Sebastian Schneeweiss, Maria C. Schneeweiss, et al. Concepts of designing and implementing pharmacoepidemiology studies on the safety of systemic treatments in dermatology practice. *JID Innovations*, 3(6):100226, 2023. doi: 10.1016/j.xjidi.2023.100226.

- Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Ingeborg Waernbaum and Lucio Pazzagli. Model misspecification and bias for inverse probability weighting estimators of average causal effects. *Biometrical Journal*, 65(2): 2100118, 2023. doi: 10.1002/bimj.202100118.
- Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*, 2023.
- Zifeng Wang and Jimeng Sun. SurvTRACE: Transformers for survival analysis with competing events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2022. doi: 10.1145/3535508.3545521.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.
- Michael Wornow, Yizhe Xu, Rahul Thapa, et al. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(135): 1–9, 2023. doi: 10.1038/s41746-023-00879-8.
- Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H. Shah. Context clues: Evaluating long context models for clinical prediction tasks on EHR data. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2412.16178.
- Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. TransformEHR: transformer-based encoder–decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 14(7550):1–14, 2023. doi: 10.1038/s41467-023-43715-z.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.
- Moshe Zisser and Dvir Aran. Transformer-based time-to-event prediction for chronic kidney disease deterioration. *Journal of the American Medical Informatics Association*, 31(4): 980–990, 2024. doi: 10.1093/jamia/ocae025.

A. Supplementary material

A.1. Word list

Attention: A mechanism that assigns scores to pairs of tokens to determine how much each token should focus on other tokens during processing.

Embedding: The process of converting categorical or numerical input features into dense numeric vectors that capture meaningful relationships between them for use in a transformer model.

Fine-tuning: The process of adapting a pre-trained transformer model to a specific task by continuing training on a smaller, labelled dataset relevant to that task.

Pre-training: The initial phase where a transformer model learns general patterns from a large dataset, building a general foundation before being fine-tuned for a specific task.

Portability: A transformer model’s ability to be applied to new datasets or settings with minimal retraining while maintaining good performance.

Token: A single unit of input data. In EHR data, a token typically represents a clinical event such as a diagnosis, a prescription, a lab test, or a hospital admission.

Transformers: A shorthand for transformer models.

Vector: In transformer models, a numeric representation of an input element, such as a clinical event, that encodes its relevant features for the model to process.

A.2. Trial Emulation

A.2.1. OVERVIEW

We emulated the SUSTAIN-6 trial using routinely collected EHR data from Region Hovedstaden and Region Sjælland, linked to national registries. A new-user, non-exposed comparator design was used, with transformer-based propensity scores and stabilised inverse-probability weighting. Eligibility criteria, follow-up, and censoring rules were aligned with SUSTAIN-6 where feasible. Further details are provided in [subsection A.5](#).

A.2.2. ELIGIBILITY (TRIAL ANALOGOUS)

Inclusion criteria required type 2 diabetes, $\text{HbA1c} \geq 7\%$, and either age ≥ 50 years with clinical cardiovascular disease (CVD) or age ≥ 60 years with evidence of subclinical CVD, while being on ≤ 2 anti-diabetic drugs (or basal/pre-mixed insulin) [Marso et al. \(2016\)](#). Exclusions mirrored SUSTAIN-6, where measurable (e.g., recent acute cardiovascular events, pancreatitis, dialysis, malignancy) and were omitted where not observable in EHR (e.g., NYHA class IV, planned revascularization, trial feasibility/hypersensitivity/illicit drug use/awaiting transplant). All inclusion and exclusion criteria are provided in tables [S5](#), [Table S6](#) and [Table S7](#). Cohort attrition from the source population to the final analysis sample is shown in [Figure S1](#).

A.2.3. INDEXING, EXPOSURE AND FOLLOW-UP

Index dates. In the RCT, the index date is the date of treatment/placebo initiation. To mimic this, we defined index dates as follows:

1. Exposed patients (semaglutide initiators).

- (a) The index date is the date of the first recorded semaglutide treatment.
- (b) Patients must have at least two years between the index date and the end of data recording (10 September 2024); otherwise, they are excluded.
- (c) Trial-analogous inclusion and exclusion criteria are applied using events occurring up to (and including) the index date.

2. Unexposed patients (non-initiators).

- (a) Each unexposed patient is assigned to a pool of exposed patients matched on birth year (tolerance ± 3 years).
- (b) An index date is randomly sampled from the index dates of patients in the corresponding pool and assigned to the unexposed patient.
- (c) If the sampled index date occurs after the unexposed patient’s date of death, a new index date is redrawn up to two further times.
- (d) If no valid index date is obtained after three attempts, the unexposed patient is discarded.
- (e) Trial-analogous inclusion and exclusion criteria are then applied using events up to the assigned index date.

This procedure yields clinically and temporally plausible index dates for both exposed and unexposed patients.

Exposure Strategy As-treated: on-treatment until a 90-day gap in orders.

Follow up Follow-up started 30 days after the index date to exclude outcomes with delayed coding that may reflect pre-existing events. For each patient, follow-up ended at the earliest of:

- two years after the index date,
- treatment discontinuation (90-day gap in semaglutide orders, as above),
- death.

To mitigate bias from differential follow-up, censoring was pair-synchronized using the link established at the index date: if an exposed patient’s follow-up ended due to discontinuation or death, the follow-up of their linked unexposed counterpart was censored at the same time. For non-fatal endpoints with potential reporting delays, the observation window was extended by 14 days after death.

A.2.4. CONFOUNDING CONTROL AND ESTIMATION

We estimated the probability of semaglutide initiation (propensity score) from pre-index longitudinal EHR data using a ModernBERT [Warner et al. \(2024\)](#) transformer model. Similar to [Rao et al. \(2024\)](#), we obtained confounder embeddings by training the model jointly on treatment and outcome, so that the shared patient representations capture patterns predictive of both exposure and risk. Technical details of the transformer architecture, pre-training, fine-tuning, cross-validation, and calibration are provided in [subsection A.7](#), with hyperparameters listed in [Table S8](#) (pre-training) and [Table S9](#) (fine-tuning).

The primary estimand was the average treatment effect (ATE) under an as-treated strategy. From the estimated propensity scores, we constructed stabilised inverse probability of treatment weights and truncated them at the 0.1th and 99.9th percentiles. Weighted and unweighted covariate balance is reported in [Table S2](#) and [Table S3](#). Discrimination and calibration of the propensity-score model (AUROC and calibration curves) are shown in [Figure S2](#).

A.3. Raw Data

Table S1: High-level overview of data sources and their content.

Data category	Description
Patient demographics	Basic patient information, including birth and death dates, gender, and identifiers used to link all other data sources.
Hospital admissions/encounters	Information on inpatient and outpatient contacts, admission events, timestamps, patient movements, and clinical departments.
Diagnoses	Diagnosis codes (e.g., ICD-10/SKS), including timestamps and links to clinical encounters.
Procedures	Surgical and non-surgical procedure codes (SKS), with timestamps and encounter associations.
Medications	Medication records from hospitals and national registries, including ATC codes, drug names, administration events, and packaging information.
Laboratory results	Laboratory test records with timestamps, numeric results, and test identifiers.

A.4. Cohort Selection

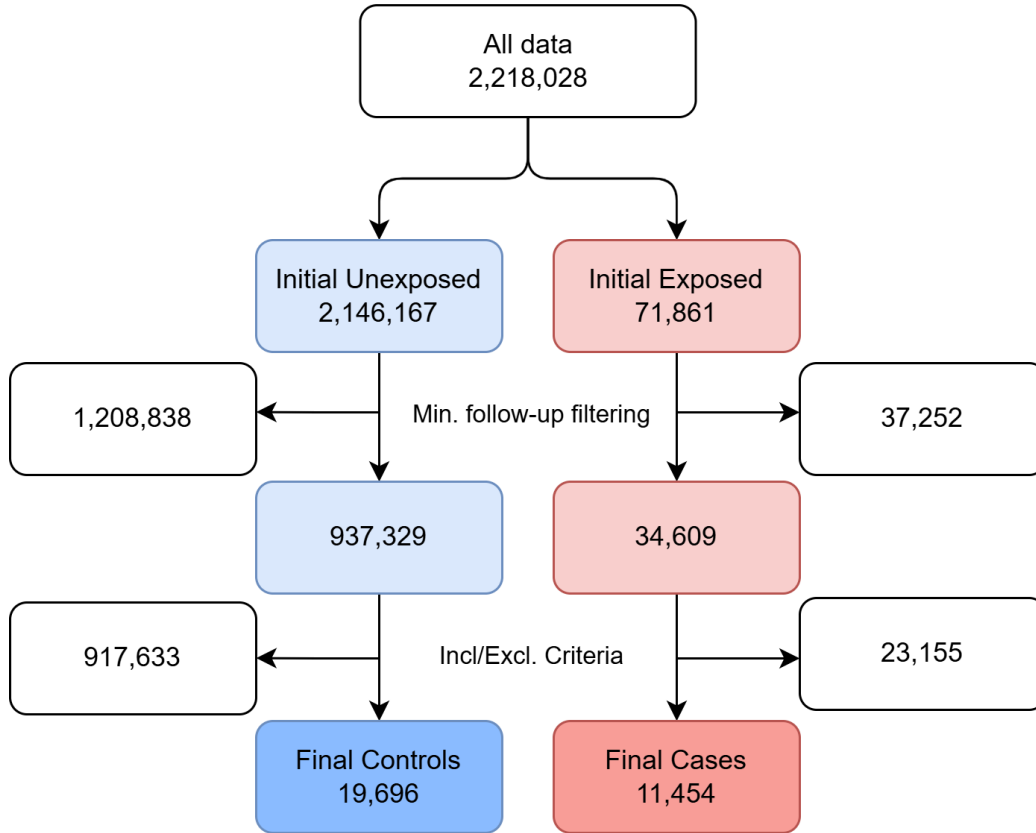


Figure S1: Flowchart showing patient inclusion. Starting with all patients, we select the exposed, those with at least one Semaglutide treatment. Their index date is determined by the date of first exposure. Patients with less than 2 years between the index date and end-of-time (2024/09/10) are excluded. Subsequently, the inclusion and exclusion criteria are applied.

Table S2: Baseline characteristics by exposure status (part 1). Values are n (%) unless otherwise specified; continuous variables are mean (SD).

	Unweighted		Weighted
	Control	Exposed	Control
ACE inhibitors	2,795 (14.6)	1,464 (13.0)	4,318 (14.2)
Age at index date, years	71.52 (± 10.11)	65.81 (± 9.01)	69.54 (± 10.14)
ARBs	2,886 (15.1)	1,722 (15.3)	4,532 (14.9)
Anticoagulants	2,356 (12.3)	1,016 (9.1)	3,248 (10.7)
Antiplatelets	3,555 (18.5)	1,805 (16.1)	5,343 (17.6)
Aortic aneurysm dilation	253 (1.3)	106 (0.9)	359 (1.2)
Aortic regurgitation/stenosis	641 (3.3)	243 (2.2)	883 (2.9)
Arterial claudication	368 (1.9)	192 (1.7)	573 (1.9)
Atrial fibrillation/flutter	2,055 (10.7)	893 (8.0)	2,855 (9.4)
Beta blockers	3,963 (20.7)	2,050 (18.3)	6,089 (20.0)
Bradycardia	213 (1.1)	80 (0.7)	295 (1.0)
Calcium channel blockers	3,274 (17.1)	1,783 (15.9)	5,103 (16.8)
Canagliflozin	45 (0.2)	59 (0.5)	101 (0.3)
CCS with stable angina pectoris	1,663 (8.7)	1,026 (9.1)	2,637 (8.7)
Chronic kidney disease	1,197 (6.2)	983 (8.8)	2,063 (6.8)
Dapagliflozin	686 (3.6)	1,009 (9.0)	1,505 (4.9)
Deep vein thrombosis	245 (1.3)	159 (1.4)	373 (1.2)
Duration of diabetes ^a	13,452 (70.2)	8,925 (79.5)	22,037 (72.5)
DPP-4 inhibitors	732 (3.8)	831 (7.4)	1,348 (4.4)
Diuretics	3,653 (19.1)	1,938 (17.3)	5,418 (17.8)
Empagliflozin	784 (4.1)	1,241 (11.1)	1,861 (6.1)
eGFR, ml/min/1.73m ²	67.09 (± 18.40)	68.28 (± 18.52)	67.81 (± 18.35)
Glucocorticoids	536 (2.8)	255 (2.3)	760 (2.5)
Heart failure	1,389 (7.2)	607 (5.4)	2,104 (6.9)
HDL, mmol/L	1.31 (± 0.41)	1.18 (± 0.33)	1.30 (± 0.40)

Abbreviations: ACE, angiotensin-converting enzyme; ARB, angiotensin II receptor blocker; CCS, chronic coronary syndrome; DPP-4, dipeptidyl peptidase-4; eGFR, estimated glomerular filtration rate; ; HDL, high-density lipoprotein; Ctrl, control; Exp, exposed; W-ctrl, weighted control.

Table S3: Baseline characteristics by exposure status (part 2). Values are n (%) unless otherwise specified; continuous variables are mean (SD).

	Unweighted		Weighted
	Control	Exposed	Control
MI	547 (2.9)	318 (2.8)	905 (3.0)
Stroke	1,019 (5.3)	430 (3.8)	1,483 (4.9)
Hypercholesterolaemia	1,231 (6.4)	783 (7.0)	2,046 (6.7)
Hypertension	11,734 (61.2)	6,577 (58.6)	18,291 (60.1)
Hypertension (ATC)	10,248 (53.5)	5,707 (50.8)	15,968 (52.5)
Hypertension (ICD-10)	4,228 (22.1)	2,478 (22.1)	6,704 (22.0)
Insulin	1,440 (7.5)	1,703 (15.2)	2,918 (9.6)
Ischaemic stroke outcome	781 (4.1)	348 (3.1)	1,134 (3.7)
Liver disease	345 (1.8)	190 (1.7)	591 (1.9)
LDL, mmol/L	2.00 (± 0.89)	1.98 (± 0.90)	2.00 (± 0.89)
Metformin	9,531 (49.7)	6,087 (54.2)	15,394 (50.6)
Mitral regurgitation/stenosis	124 (0.6)	38 (0.3)	172 (0.6)
NSAIDs	1,129 (5.9)	725 (6.5)	1,922 (6.3)
Obesity	549 (2.9)	743 (6.6)	1,006 (3.3)
Opioids	1,722 (9.0)	862 (7.7)	2,620 (8.6)
Pulmonary embolism	167 (0.9)	116 (1.0)	249 (0.8)
Respiratory medications	1,934 (10.1)	1,008 (9.0)	2,958 (9.7)
SGLT2 inhibitors	1,516 (7.9)	2,303 (20.5)	3,470 (11.4)
Statins	7,085 (37.0)	3,991 (35.6)	11,344 (37.3)
Sulfonylureas	554 (2.9)	433 (3.9)	1,024 (3.4)
Triglycerides, mmol/L	1.93 (± 1.18)	2.56 (± 1.83)	1.97 (± 1.22)
T2D	17,875 (93.3)	11,075 (98.7)	28,596 (94.0)
Unstable angina pectoris	185 (1.0)	125 (1.1)	285 (0.9)

Abbreviations: MI, myocardial infarction; LDL, low-density lipoprotein; NSAID, non-steroidal anti-inflammatory drug; SGLT2, sodium-glucose cotransporter-2; T2D, type 2 diabetes; ATC, Anatomical Therapeutic Chemical classification; ICD-10, International Classification of Diseases, 10th revision; Ctrl, control; Exp, exposed; W-ctrl, weighted control.

A.5. PICOTS Framework

Table S4: PICOTS framework for the SUSTAIN-6 target trial emulation (TTE)

Element	Definition in the TTE (SUSTAIN-6 emulation)
P Population	Adults with type 2 diabetes identified in hospital-based EHR data. Eligibility was based on SUSTAIN-6 inclusion and exclusion criteria, operationalised using ICD-10 diagnoses, ATC codes, and laboratory values. Full list in subsection A.6 .
I Intervention	Initiation of semaglutide (ATC A10BJ02) . New users only, defined as no GLP-1 receptor agonist exposure in the prior 90 days.
C Comparator	Non-exposed individuals . Index dates were assigned by drawing from the exposure-date distribution of semaglutide initiators, requiring individuals to be alive at the assigned date, allowing up to three redraws if needed, and excluding individuals without a valid date. All eligibility criteria were applied identically after index assignment.
O Outcomes	Outcomes included: all-cause death, composite death/myocardial infarction/stroke, heart failure hospital admission, non-fatal myocardial infarction , and non-fatal stroke . Cardiovascular death could not be defined reliably, and major cardiovascular events (MACE) was therefore not included. Outcomes were defined using ICD-10 diagnoses and mortality records.
T Timing	Baseline was defined at semaglutide initiation for exposed individuals and at the assigned index date for controls. Lookback windows, exclusion periods and follow-up definitions followed SUSTAIN-6 where feasible. All specific time windows, where applicable, are detailed in subsection A.6 .
S Setting	Routine clinical care recorded in the shared EHR infrastructure of Region Hovedstaden and Region Sjælland (Sundhedsplatformen).

A.6. Alignment of Eligibility Criteria

Table S5: Alignment of SUSTAIN-6 inclusion criteria with TTE implementation (adapted from [Marso et al. \(2016\)](#)).

SUSTAIN-6 criterion	TTE implementation	Comments
Type 2 diabetes	≥ 1 ICD-10 E11 diagnosis or ≥ 2 dispensations of ATC A10B (non-insulin antidiabetics).	Routine-care equivalent definition.
Age ≥ 50 years + clinical CVD	Age ≥ 50 at index and ≥ 1 of: MI (I21–I23), stroke (I63–I66), TIA (G45), or revascularisation (SKS KFN*).	Imaging/stress-test components not available.
Chronic renal impairment	Latest eGFR < 60 ml/min/1.73 m ² before index date.	Direct lab mapping.
Additional CVD criteria (stenosis, CHD symptoms, stress tests, NYHA II–III)	Not implemented.	Requires detailed imaging, stress testing and NYHA class; not coded.
Age ≥ 60 years + subclinical CVD	Age ≥ 60 and ≥ 1 of: albuminuria/proteinuria, hypertension (I10) with LVH where available, LV dysfunction where available, or ABI proxy (peripheral artery disease codes / vascular procedures such as DI739, DI702, KPEE–KPFN).	ABI and LVH/LV dysfunction only partially captured.
Max. two antidiabetic drugs or insulin	Count ATC A10 agents; A10BD combinations count as 2 drugs; NPH/long-acting/premixed insulin allowed.	Approximates trial medication rule.
HbA1c $\geq 7\%$	Latest HbA1c ≥ 53 mmol/mol within 0–90 days before index.	Direct lab threshold.

Table S6: Alignment of SUSTAIN-6 exclusion criteria with TTE implementation (part 1: implemented criteria)(adapted from [Marso et al. \(2016\)](#)).

SUSTAIN-6 criterion	TTE implementation	Comments
Type 1 diabetes	ICD-10 E10 (ever).	Direct ICD mapping.
GLP-1 receptor agonist use (last 90 days)	ATC A10BJ codes (and A10AE56 where relevant) in 0–90-day look-back.	Direct ATC mapping.
DPP-4 inhibitor use (last 30 days)	ATC A10BH and DPP-4 combinations (A10BD07–A10BD13, A10BD18–A10BD22, A10BD24–A10BD25, A10BD27) in 0–30-day look-back.	Direct ATC mapping.
Insulin except basal/premixed (90 days)	ATC A10AB dispensations in 0–90-day look-back.	Direct mapping.
Acute glycaemic decompensation (90 days)	ICD-10 E11.1, E12.1, E13.1, E14.1 (local: DE111, DE121, DE131, DE141) in 0–90-day look-back.	Direct ICD mapping.
Chronic pancreatitis	ICD-10 K86 with 2-year look-back.	Direct mapping.
Idiopathic acute pancreatitis	ICD-10 K850.	Direct mapping.
Acute coronary / cardiovascular event (90 days)	ICD-10 I21–I22, I24, I63–I66 in 0–90-day look-back.	Direct mapping.
Chronic haemodialysis / peritoneal dialysis	ICD-10 N18.5 and SKS BJFD20, BJFD21, BJFD23–BJFD25, BJFD27.	Direct mapping of dialysis modalities.
End-stage liver disease	ICD-10 K72/K72.1 plus supporting labs (bilirubin, albumin, INR) and/or ascites, encephalopathy, or variceal bleeding codes.	Composite algorithm mirroring trial definition.
Prior liver transplant	SKS KJJC (ever).	Direct procedure code.
Prior solid organ transplant	ICD-10 DZ944, DY830 and SKS KFQ, KGDG, KJJC, KJLE, KKAS.	Direct mapping.
Malignancy within 5 years	ICD-10 C00–C99 (excluding C44) in 5-year look-back.	Direct cancer look-back.
MEN2 / medullary thyroid carcinoma	ICD-10 E31.2 and D09.3.	Direct mapping.
Calcitonin ≥ 50 ng/L	Relevant lab codes (e.g. L/CALCITONIN;P, NPU01439, NPU58815) with value ≥ 50 ng/L.	Direct lab rule.
Recent pregnancy / breast-feeding	ICD-10 O00–O99 in 0–12-month look-back.	Proxy for pregnancy/birth in trial.

Table S7: Alignment of SUSTAIN-6 exclusion criteria with TTE implementation (part 2: not or partly implemented criteria)(adapted from [Marso et al. \(2016\)](#)).

SUSTAIN-6 criterion	TTE implementation	Comments
NYHA class IV heart failure	Not implemented.	NYHA class not coded; only heart failure diagnosis available.
Planned coronary/carotid/peripheral revascularisation	Not implemented.	Procedure scheduling not represented in registries.
Pramlintide use (90 days)	Not implemented.	Drug not marketed / not identifiable in Danish data.
Awaiting solid organ transplant	Not implemented.	“Waiting list” status not reliably coded.
Known hypersensitivity to trial products or related drugs	Not implemented.	Specific hypersensitivity not systematically coded.
Known use of non-prescribed narcotics or illicit drugs	Not implemented.	Use of illicit or non-prescribed narcotics not reliably recorded.
Other factors likely to limit protocol adherence or AE reporting	Not implemented.	Investigator-judgement criterion with no registry analogue.

A.7. Model Training

We used a BERT-like architecture with embeddings adapted to EHR inputs, similar to CoreBEHRT [Odgaard et al. \(2024\)](#) but with a ModernBERT [Warner et al. \(2024\)](#) encoder. The encoder was first pre-trained using masked language modelling on 70% of all available patients (excluding the SUSTAIN-6 emulation cohort) to learn context-aware event representations.

For the propensity-score model, the pre-trained encoder was then fine-tuned on longitudinal pre-index EHR sequences (two-week delay for diagnoses to account for delayed coding) to jointly predict treatment assignment and clinical outcomes, so that the shared patient embeddings capture patterns that are predictive of both exposure and outcome. During joint fine-tuning, we used PCGrad [Yu et al. \(2020\)](#) to align gradients from the exposure and outcome heads.

Fine-tuning was carried out in 5-fold cross-validation. For each patient, the propensity score used in the analysis is the model’s predicted probability of treatment from the fold in which that patient was included in the validation set. Hyperparameters for pre-training and fine-tuning are listed in [Table S8](#) and [Table S9](#), respectively.

We used the beta-calibration method [Kull et al. \(2017\)](#) to calibrate the propensity scores (see [Figure S2](#)).

Table S8: Hyperparameters used for pre-training.

Hyperparameter	Value	Hyperparameter	Value
<i>Dataset</i>		<i>Optimization</i>	
Masking ratio	0.8	Learning rate	0.001
Replace ratio	0.1	Warmup epochs	4
Select ratio	0.2	Batch size	512
Validation ratio	0.2	Max. epochs	50
Vocabulary Size	40,839	Early stopping	3
<i>Model Architecture</i>		Gradient clipping	1.0
Hidden size	256	Scheduler	Constant w/ warmup
Intermediate size	1024		
Max position emb.	1024		
Attention heads	4		
Hidden layers	4		
Total Params.	25.5 M		

Table S9: Hyperparameters used for fine-tuning.

Hyperparameter	Value	Hyperparameter	Value
<i>Head Architecture</i>		<i>Optimization</i>	
Bidirectional	True	Learning rate	4e-4
Bottleneck dim.	128	Weight decay	0
Pooling strategy	GRU	Optimizer ϵ	1e-6
Loss	BCE	Scheduler	Cosine w/ warmup
<i>Training</i>		Warmup epochs	3
Batch size	512	Max. epochs	25
Val. batch size	512	Early stopping	5
Gradient clipping	1.0	PCGrad	True

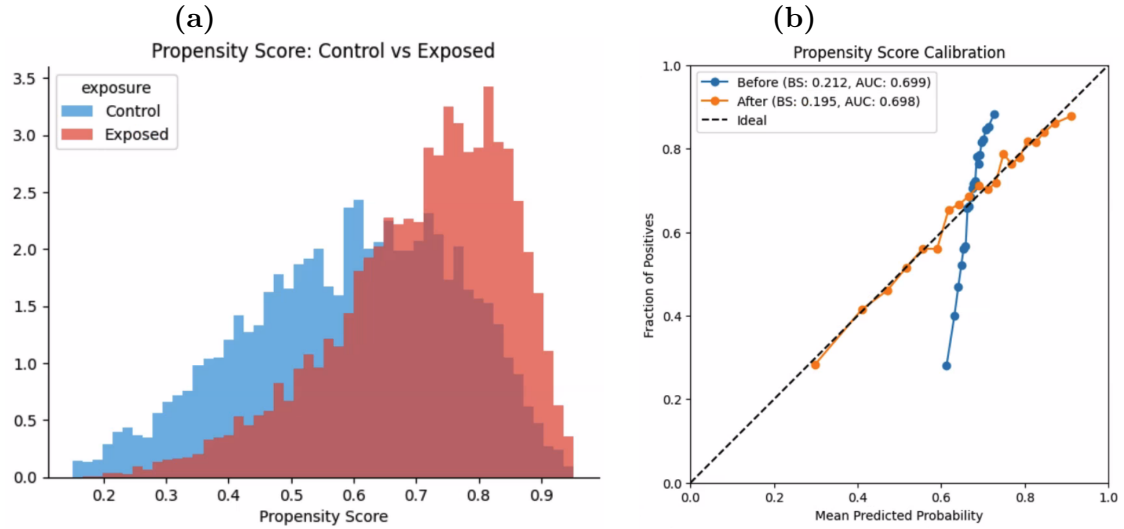


Figure S2: Propensity score distribution after beta-calibration (a) and calibration curve before and after calibrating (b).

Chapter 6

Untargeted causal drug effect estimation: populationwise effects of Semaglutide

The following manuscript is being prepared for submission in a clinical journal.

Untargeted causal drug effect estimation: populationwise effects of Semaglutide

Kiril Klein^{1*}, Maria Montgomery¹, Mikkel Odgaard^{1,2},
Sanne Møller Thysen², Thomas Jensen²,
Espen Jimenez-Solem^{2,3,4}, Martin Sillesen⁵, Mads Nielsen¹

^{1*}Department of Computer Science, University of Copenhagen,
Copenhagen, Denmark.

²Department of Clinical Pharmacology, Copenhagen University
Hospital - Bispebjerg and Frederiksberg, Copenhagen, Denmark.

³Copenhagen Phase IV Unit (Phase4CPH), Department of Clinical
Pharmacology, Center for Clinical Research and Prevention,
Copenhagen University Hospital - Bispebjerg and Frederiksberg,
Copenhagen, Denmark.

⁴Department of Clinical Medicine, University of Copenhagen,
Copenhagen, Denmark.

⁵Department of Organ Surgery and Transplantation, Copenhagen
University Hospital - Rigshospitalet, Copenhagen, Denmark.

*Corresponding author(s). E-mail(s): kikl@di.ku.dk;

Abstract

Background: The momentous uptake of certain drugs challenges conventional post-marketing surveillance, which is often too slow or selective to quantify real-world effects. Semaglutide is a recent example with widespread use and substantial public attention. We developed and applied a scalable data-driven approach to screen for causal effects of drug initiation on subsequent medication use and demonstrated this on semaglutide.

Methods: We developed high-dimensional propensity-weighted estimation using a sequence analysis transformer architecture with confounder embedding. Using electronic health records (EHRs) and linked registers from Eastern Denmark (2,218,028 subjects including 71,861 semaglutide users), we estimated the 2-year effect of semaglutide initiation on subsequent dispensing of all ATC-coded drugs. The model generated patient-level weights to approximate exchangeability between initiators and non-users.

Results: We screened for 328 drugs of which 83 drugs had a significantly increased prescription rate and 10 a decreased one ($p \leq 0.001$, min. prevalence 1%). Both the main and secondary analyses showed a decreased prescription of dementia drugs and drugs affecting bone structure, while there was an increase in urological agents, antidepressants, drugs for acid-related disorders, analgesics and antipyretics.

Conclusion: High-dimensional transformer-based causal screening can recover expected treatment-related changes in medication patterns and highlight potential beneficial and adverse effects at scale. This approach may support early signal identification for widely used therapies and guide targeted epidemiological follow-up studies.

1 Introduction

Despite rigorous pre-approval evaluation, serious adverse drug reactions (ADRs) still come to light after therapies are widely adopted in routine care. Key patient groups, including pregnant women, children, older adults, and individuals with multimorbidity, remain under-represented or excluded from pivotal randomised trials, leaving important gaps in our understanding of safety precisely at the point when clinicians begin prescribing these drugs at scale [1–3]. Consequently, many therapies are used not only off-label by indication but effectively off-label with respect to the populations in which their safety has not been established, leaving important gaps in our understanding of real-world risk at scale [4]. Although regulators increasingly mandate post-approval safety studies, these are frequently delayed, incomplete, or never completed; meanwhile, millions of patients may be exposed before robust safety data emerge [5, 6]. Even when pre-approval trials exist, they are typically underpowered to detect infrequent, delayed, or context-specific harms, and ADRs may be under-ascertained, selectively reported, or monitored only over limited follow-up [7, 8]. Together, these constraints create a predictable deficit in our understanding of drug safety.

Existing post-marketing surveillance systems only partially fill this gap. Spontaneous reporting depends on clinician or patient recognition and reporting of suspected ADRs and is prone to under-reporting, notoriety bias, and confounding by indication [8]. Traditional observational studies tend to focus on a small number of pre-specified safety outcomes for each drug, leaving the much larger space of unanticipated or subtle harms largely unexplored. There is thus a need for methods that enable agnostic, large-scale screening of drug safety signals in real-world data, capable of detecting both the rate of suspected ADRs and entirely unforeseen ADRs across the full spectrum of routine care [9, 10].

Conventional propensity-score approaches, while offering causal estimands, are cumbersome to deploy repeatedly across thousands of drug-outcome pairs and do not scale to truly untargeted discovery [10]. High-dimensional propensity methods can, in principle, adjust for extensive confounding but are vulnerable to model misspecification and extreme weights, particularly when non-confounding instruments are inadvertently included. In contrast, self-controlled designs offer excellent scalability and protection against time-invariant confounding through within-person comparisons but yield incidence rate ratios that do not directly correspond to clinically meaningful causal quantities such as the average treatment effect on the treated [11]. These limitations create a methodological gap: current tools either scale but lack causal interpretability, or are causally interpretable but impractical for systematic, hypothesis-free safety screening.

To address this, we develop a scalable architecture that combines high-dimensional confounding control with the capacity for agnostic safety discovery. An overview of the pipeline is shown in Figure 1. We implement a calibrated transformer-based machine learning model to estimate propensity scores from the full time series of each patient’s health history. By learning vector-space representations that prioritise confounding information, this approach enables efficient, high-dimensional propensity scoring at a population scale and mitigates the instability associated with conventional high-dimensional models.

We then apply this framework in a real-world evidence proof-of-concept study of semaglutide, a glucagon-like peptide-1 receptor agonist (GLP-1 RA) whose rapid uptake has been accompanied by intense clinical and public interest in its safety profile [12]. Leveraging Denmark’s nationwide health registries, we focus on new adult users in Eastern Denmark and treat incident drug redemptions recorded in national prescription registries as proxies for emergent clinical events (i.e., ADRs). In Denmark, community pharmacies are legally required to register all dispensed prescriptions using reimbursement-linked, largely automated bar-code-based systems, yielding near-complete coverage and high coding accuracy for redeemed drugs, in contrast to the more heterogeneous validity of diagnostic coding [13]. Unlike signal-detection efforts based primarily on diagnostic codes, our approach captures pharmacologically managed symptoms and conditions that may never be formally diagnosed, such as treatment-requiring gastrointestinal complaints, rash, or subacute complications that trigger medication changes but not hospitalisation [9]. By estimating inverse probability-weighted risks for all downstream medications and visualising over- and under-represented events in volcano plots (Figure A1), we generate an agnostic, data-driven safety map of semaglutide in routine use. This registry-based screening strategy complements diagnosis-centred pharmacovigilance and illustrates a scalable framework for mapping potential ADRs of any health-care intervention, thereby prioritising candidate safety signals for targeted clinical and epidemiological follow-up [8–10].



Fig. 1: Overview of the untargeted causal screening pipeline. Longitudinal electronic health record trajectories are encoded with a transformer model to obtain confounder-focused patient embeddings. These are used to estimate propensity scores and compute inverse probability-weighted risk ratios for all downstream drug outcomes.

2 Results

We studied the relation between initiation of semaglutide and subsequent prescriptions of other medications at ATC level 4. The primary analysis included individuals aged ≥ 50 years at index (hereafter the “main cohort”), and the secondary analysis included patients with prior cardiovascular disease (the “CVD cohort”). After applying eligibility criteria and index date matching, we obtained 8,557 exposed and 249,315 unexposed patients in the main cohort and 1,837 exposed and 51,409 unexposed in the CVD cohort (see flow diagram [Figure B6](#)). Baseline profiles reflected clear clinical differences between semaglutide users and non-users. Users had higher cardiometabolic risk, including more diabetes-related complications, greater use of statins, antihypertensives, and antiplatelet therapy, and higher HbA1c. The baseline characteristics are provided in [Table 1](#).

2.1 Main result

Across both cohorts, semaglutide treatment was associated with systematic shifts in subsequent medication use. We report risk ratios (RRs) with 95% confidence intervals (CIs). The most pronounced increases were observed within anti-diabetic medications (ATC group A), where several other glucose-lowering therapies showed elevated RRs, including SGLT-2 inhibitors [13.6 (12.8 – 14.4)], biguanides [9.3 (8.9 – 9.8)], insulins (A10AD) [6.3 (5.6 – 7.1)], and sulfonylureas [8.4 (7.8 – 9.1)]. Several cardiovascular drug classes in ATC group C also exhibited RRs close to two, indicating broader adjustments in cardiometabolic management following semaglutide initiation. These group-level trends were visually apparent in the circular summaries of ATC categories ([Figure 2](#)). Conversely, prescriptions of dementia treatments (N06DA) and bisphosphonates (M05BA) were consistently lower among semaglutide users in both the main and CVD cohorts, with RRs of 0.44 (0.39 – 0.48) and 0.45 (0.43 – 0.48), respectively. Higher use was observed for drugs for erectile dysfunction (G04BE) [2.09 (1.97 – 2.22)], antidepressants (N06AX, N06AA) [1.38 (1.32 – 1.43); 1.74 (1.64 – 1.84)], medications for acid-related disorders (A02AA, A02BC) [1.41 (1.36 – 1.45); 1.55 (1.51 – 1.58)] and analgesics (N02BF) [1.78 (1.72 – 1.84)]. These estimates, together with their full confidence intervals, are shown in [Figure 3](#). Complete results of log-risk ratios with CIs are provided in [Figure A4](#) and [Figure A5](#), and log-risk ratios with significance and prevalence can be seen in [Figure A2](#) and [Figure A3](#). Overall, the patterns suggest broader shifts in symptom management and comorbidity treatment following semaglutide initiation, beyond changes in glucose-lowering therapy alone.

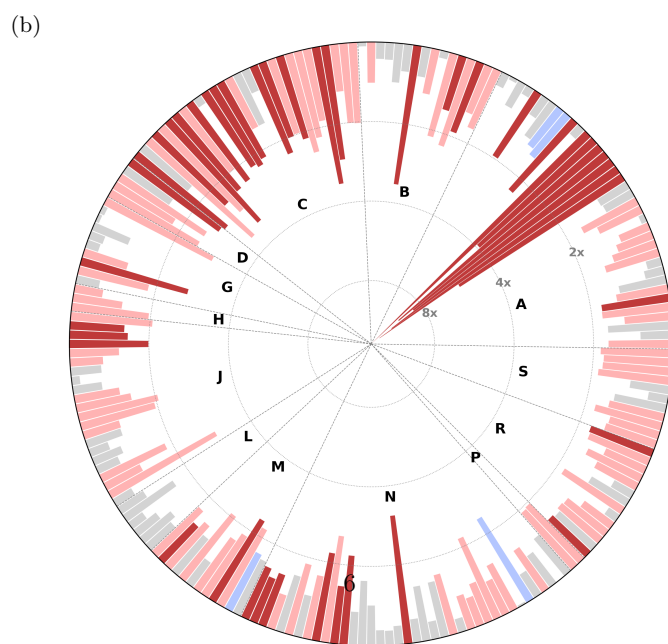
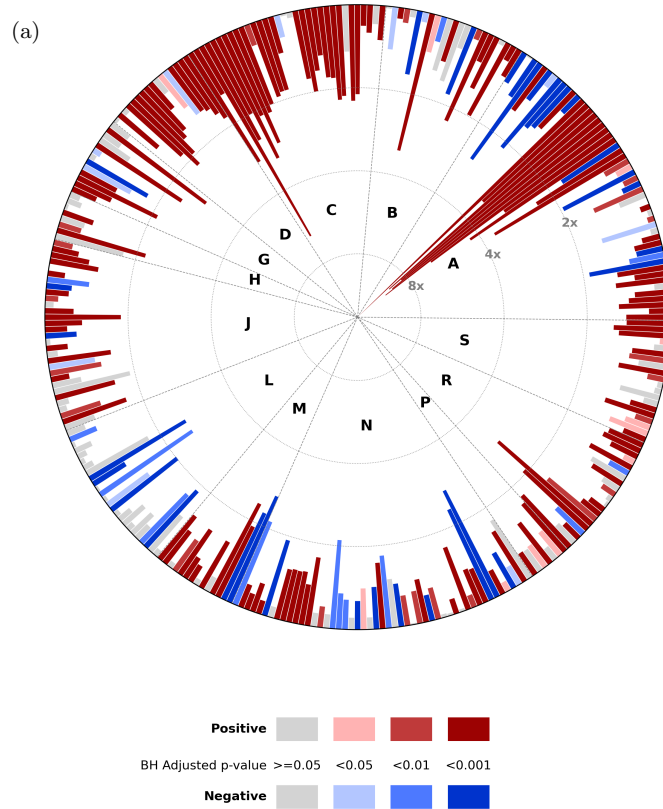


Fig. 2: Overview of semaglutide-related changes in medication redemptions. Each bar represents an ATC level four medication, grouped by ATC level 1 category (A–V) around the circle. Bar height encodes the adjusted log risk ratio, with red bars indicating increased use after semaglutide initiation and blue bars decreased use. Colour intensity reflects the Benjamini–Hochberg-adjusted p-value. (a) Age ≥ 50 years cohort. (b) Cardiovascular disease (CVD) cohort.

Table 1: Baseline characteristics of semaglutide-exposed and unexposed patients in the cardiovascular disease (CVD) and main cohorts. Continuous variables are reported as mean (SD); categorical variables as n (%).

Characteristic	CVD cohort		Main cohort	
	Control	Exposed	Control	Exposed
ALT, U/L	28.33 ($\pm$ 22.91)	34.94 ($\pm$ 24.12)	27.84 ($\pm$ 20.87)	34.28 ($\pm$ 21.82)
ARB, <i>n</i> (%)	6 093 (11.9)	527 (28.7)	29 866 (12.0)	2 364 (27.7)
ACEi, <i>n</i> (%)	5 001 (9.8)	430 (23.4)	22 629 (9.1)	1 884 (22.1)
Age at index date, years	58.62 ($\pm$ 19.27)	63.46 ($\pm$ 11.55)	67.29 ($\pm$ 11.10)	65.69 ($\pm$ 9.14)
Antiplatelet therapy, <i>n</i> (%)	12 489 (24.4)	856 (46.7)	34 920 (14.1)	2 736 (32.1)
Beta blockers, <i>n</i> (%)	8 137 (15.9)	646 (35.2)	33 005 (13.3)	2 477 (29.1)
Chronic kidney disease, <i>n</i> (%)	805 (1.6)	64 (3.5)	2 644 (1.1)	222 (2.6)
Depression, <i>n</i> (%)	8 376 (16.3)	945 (24.8)	3 851 (14.8)	1 912 (22.4)
Diabetic nephropathy, <i>n</i> (%)	193 (0.4)	136 (7.4)	814 (0.3)	616 (7.2)
Diabetic neuropathy, <i>n</i> (%)	258 (0.5)	157 (8.6)	779 (0.3)	646 (7.6)
Diabetic retinopathy, <i>n</i> (%)	436 (0.9)	185 (10.1)	1 565 (0.6)	717 (8.4)
eGFR, mL/min/1.73 m ²	71.21 ($\pm$ 17.11)	68.79 ($\pm$ 19.25)	71.89 ($\pm$ 15.98)	69.32 ($\pm$ 18.60)
Female, <i>n</i> (%)	26 067 (50.9)	666 (36.3)	130 477 (52.6)	3 358 (39.4)
Heart failure, <i>n</i> (%)	2 036 (4.0)	162 (8.8)	5 451 (2.2)	404 (4.7)
HbA1c, mmol/mol	39.16 ($\pm$ 8.92)	63.89 ($\pm$ 18.50)	39.50 ($\pm$ 8.40)	63.60 ($\pm$ 17.71)
Ischemic heart disease, <i>n</i> (%)	6 714 (13.1)	908 (27.7)	10 681 (4.3)	3 439 (40.3)
Ischemic stroke, <i>n</i> (%)	5 215 (10.2)	266 (14.5)	4 888 (2.0)	249 (2.9)
LDL, mmol/L	2.39 ($\pm$ 1.00)	1.88 ($\pm$ 0.89)	2.62 ($\pm$ 1.00)	1.98 ($\pm$ 0.93)
Myocardial infarction, <i>n</i> (%)	3 902 (7.6)	260 (14.2)	3 455 (1.4)	241 (2.8)
Oral anticoagulants, <i>n</i> (%)	4 997 (9.8)	290 (15.8)	21 389 (8.6)	1 134 (13.3)
Statins, <i>n</i> (%)	15 141 (29.5)	1 301 (70.9)	55 922 (22.5)	5 664 (66.4)
Thiazides, <i>n</i> (%)	2 861 (5.6)	215 (11.7)	15 908 (6.4)	1 090 (12.8)

Notes. Continuous variables are presented as mean (SD); categorical variables as *n* (%).

Abbreviations. ALT, alanine aminotransferase; ARB, angiotensin receptor blocker; ACEi, angiotensin-converting enzyme inhibitor; eGFR, estimated glomerular filtration rate; HbA1c, glycated haemoglobin; LDL, low-density lipoprotein; CVD, cardiovascular disease.

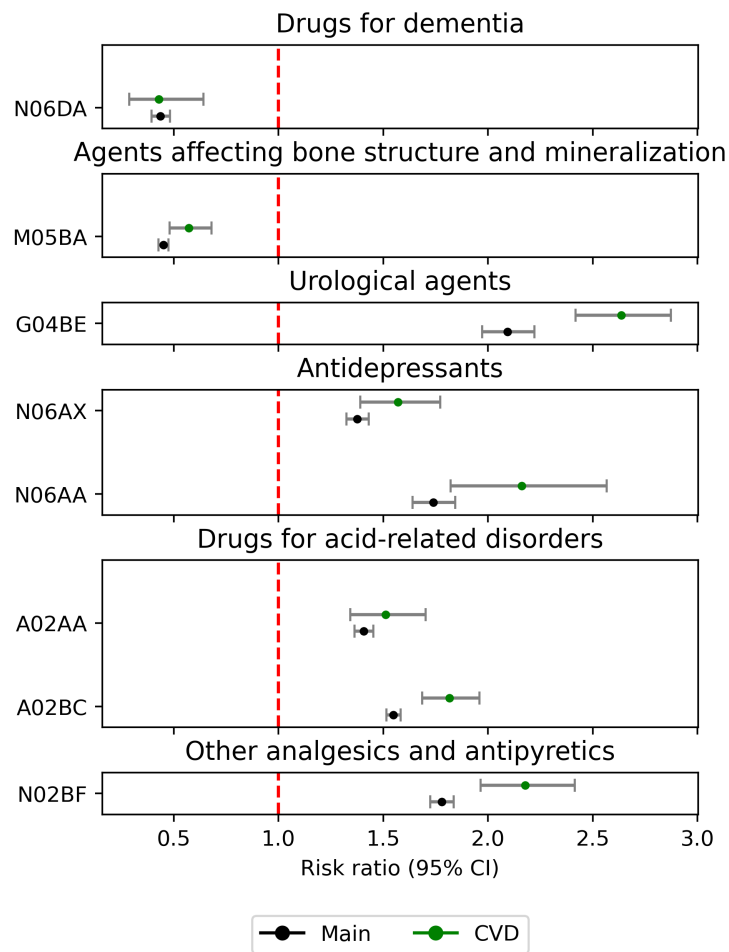


Fig. 3: Forest plot of selected medication outcomes in the main and CVD cohorts. Panels group ATC level 4 codes into broader pharmacological classes. Points show the estimated adjusted risk ratios, and horizontal lines indicate 95% confidence intervals.

3 Discussion

Across several outcomes, our results aligned with emerging evidence on potential ADRs of semaglutide and other GLP-1 receptor agonists. These included decreased risk of medications used for dementia treatment and increased risks of medications related to gastrointestinal symptoms and erectile dysfunction (ED) as shown in [Figure 2](#).

We observed a markedly reduced risk of initiating medications used for dementia (ATC N06D*), consistent with emerging evidence suggesting favourable cognitive outcomes among users of GLP-1 receptor agonists. Large-scale observational studies provide converging support: a US EHR analysis from the TriNetX network reported lower dementia incidence among GLP-1 RA users (HR 0.63; 95% CI, 0.50–0.81) [14], and a Danish trial-and-registry study found reduced rates in both randomised comparisons (HR 0.47 [0.25; 0.86]) and a nationwide cohort with increasing cumulative exposure (HR 0.89 [0.86; 0.93]) [15]. Additional TriNetX-based studies have shown lower risks of incident Alzheimer’s disease and Alzheimer’s related outcomes in association with semaglutide [16, 17], findings that contributed to the initiation of phase 3 trials evaluating its disease-modifying potential [18]. While our prescription-based approach cannot identify mechanisms, the consistent reduction in dementia-related medications is directionally aligned with this body of evidence and supports further evaluation of cognitive outcomes.

Importantly, these observational signals must be interpreted in light of the recent termination of the EVOKE and EVOKE+ phase 3 trials, in which semaglutide did not slow clinical progression of early Alzheimer’s disease despite favourable biomarker changes [19]. The sponsor discontinued the planned extension phase following the negative primary outcome. This divergence between observational associations and trial results underscores the uncertainty around potential neuroprotective effects and highlights the need to disentangle biological mechanisms from treatment-related behavioural or healthcare-system factors that may influence dementia-related prescribing.

Gastrointestinal ADRs are well-recognised with GLP-1 RAs, affecting more than 60% of trial participants and representing a leading cause of treatment discontinuation [20]. In our analysis, semaglutide use was associated with higher initiation of proton pump inhibitors (PPIs), plausibly reflecting treatment for nausea, dyspepsia, or other upper GI symptoms triggered by GLP-1 RA therapy. These findings demonstrate that our untargeted, prescription-based approach reliably captures common ADRs that may not lead to hospital contact but require pharmacological management.

We observed an increased risk of redeeming prescriptions for ED medications (e.g., sildenafil), which may indicate increased management of sexual dysfunction in semaglutide users. This is consistent with findings in non-diabetic men with obesity, where semaglutide was associated with a significantly higher risk of ED or new prescriptions for phosphodiesterase-5 inhibitors (1.47% vs 0.32%; RR: 4.5 [2.3; 9.0]) [21]. Analysis of the FDA Adverse Event Reporting System (FAERS) reports similarly identified ED-related signals for semaglutide, whereas other clinical and mechanistic data suggest that GLP-1 RAs may, in some settings, improve erectile function [22]. Our findings contribute additional real-world evidence indicating that sexual dysfunction merits further targeted studies.

In addition to confirming known or emerging ADRs, our analysis revealed several signals that are either novel or insufficiently characterised in the literature, including antidepressant initiation, gabapentin initiation, and osteoporosis-related medications.

We observed significantly elevated RRs for initiating antidepressants, most prominently in the main cohort. Evidence linking GLP-1 RAs to depression remains mixed: some analyses have found no disproportionality for depression in pharmacovigilance data [23], and a meta-analysis of randomised trials reported no association with suicidal behaviour [24]. However, recent disproportionality analyses in FAERS and WHO global individual case safety report database (VigiBase) identified a signal for semaglutide, though not for liraglutide or tirzepatide, suggesting a potential drug-specific effect [25, 26]. Our findings support this emerging concern and indicate that the relationship between semaglutide and mood disorders warrants dedicated investigation.

Gabapentin redemption was more common among semaglutide-exposed individuals. Because other antiepileptic medications showed reduced RRs, this pattern may relate to indications other than epilepsy (e.g., neuropathic pain), but further targeted studies are needed. Evidence on GLP-1 RAs and diabetic peripheral neuropathy is sparse and inconclusive [27]. Importantly, recent registry-based work has linked semaglutide to non-arteritic anterior ischemic optic neuropathy [28], raising the hypothesis that certain forms of neuropathy may be associated with GLP-1 RA use. Our findings provide additional motivation for targeted evaluation of neuropathic outcomes.

We also identified a markedly reduced risk of initiating medications used to prevent or treat osteoporosis. Evidence in this area remains limited, although weight loss induced by caloric restriction is known to adversely affect bone density [29]. Our finding [0.45 (0.43 – 0.48)] in the main cohort, with similar reductions in the CVD cohort [0.57 (0.48 – 0.68)], highlights an unexpected association that merits further mechanistic and clinical studies.

Overall, our findings show that large-scale, untargeted analysis of prescription data can identify established ADRs of semaglutide while highlighting emerging safety signals that merit closer analyses. The associations with antidepressants, gabapentin use, and osteoporosis-related treatments point to potential biological effects that remain insufficiently understood. As semaglutide use continues to expand across diverse populations, scalable approaches such as ours may help prioritise safety signals for targeted clinical and mechanistic investigation.

3.1 Strengths and Limitations

Methodological

The confounder-focused embeddings have shown excellent efficiency in estimating the effect, both with high sensitivity and computational feasibility. The training of embeddings simultaneously fine-tuned for treatment and multiple (several hundreds of) outcomes will, up to usual limitations of embedding space capacity, model capacity, and training data, encompass all identifiable confounders. It may, however, also include variables that are instruments and thereby diminishing the sensitivity. Therefore, it is a conservative signal detection methodology. Potential further development could focus on eliminating instruments while preserving confounders.

Methodology implementations are compatible with the MEDS standard [30], making application to other data sources seamless the code is publically available at <https://github.com/kirilklein/PHAIR.EHR>.

Clinical

A key clinical strength of this study is the use of a whole, non-selected population, which allows evaluation of semaglutide users across the full clinical spectrum, including older adults, multimorbid patients, and individuals with polypharmacy who are typically underrepresented in trials. This enhances the clinical relevance of the findings and enables the detection of treatment-related changes that may only emerge in routine care. Another strength is the ability to capture both direct and indirect effects of semaglutide use, since changes in prescription patterns may reflect symptom management, comorbidities, or downstream consequences of treatment. By using medication initiation rather than diagnoses, we also capture clinically meaningful events that may never lead to hospital contact but nonetheless require pharmacological management, such as gastrointestinal symptoms, mood changes, or neuropathic pain.

Several limitations should be considered when interpreting the clinical implications. First, all data come from a single country with a unified healthcare system, which provides internal consistency but may limit generalisability to settings with different prescribing practices, access pathways, or thresholds for specialist referral. Second, although changes in medication use can reflect clinically relevant symptoms or disease progression, they cannot distinguish whether these patterns arise from biological effects of semaglutide, from changes in weight, from evolving comorbidities, or from altered interactions with the healthcare system, including increased clinical scrutiny around treatment initiation. Finally, while prescription data are robust and prospectively collected, they do not provide information on the underlying clinical rationale, symptom severity, or patient-reported outcomes, and therefore cannot fully resolve the mechanisms driving the observed associations.

3.2 Conclusions and Future Directions

We showcased how confounder embeddings obtained with pre-trained transformers can be used to obtain causal estimates from large-scale EHR data. Expected effects were recovered, such as increased use of glucose-lowering drugs and gastrointestinal medication. It also identified a potentially protective relation with dementia and osteoporosis. Additional signals were found for ED and antidepressant initiation.

These findings merit a targeted and mechanistic follow-up. Future work should extend this framework to screen all drug types and consider additional sources as outcomes, such as diagnoses, procedures, and laboratory results.

4 Methods

4.1 Study design

We performed an observational, retrospective cohort study comparing semaglutide (at least one occurrence of ATC A10BJ06) initiators with non-initiators assigned a pseudo index date. Eligible exposed patients were required to have at least two years between the index date and the end of database follow-up (09/10/2024). The primary cohort included patients who were at least 50 years old at the index date. The secondary cohort included all patients with a history of cardiovascular disease (CVD), with code definitions provided in [Table B1](#).

For each cohort, unexposed patients were sampled from the pool of patients without any semaglutide exposure. Within strata defined by birth year (± 3 years), we randomly sampled an index date from the exposed patients and assigned it to unexposed patients in the same stratum. If death occurred before the drawn index date, we redrew up to two times and discarded patients with no valid index date. The same inclusion and exclusion criteria were then applied using the assigned index date.

Follow-up started six months after the index date, to reduce bias from pre-existing conditions with delayed coding, and ended at the earliest of two years after index, death, or treatment discontinuation. For each ATC level 4 code, we defined a binary outcome indicating at least one prescription during follow-up. The full cohort construction and matching procedure is illustrated in the flow diagram in [Figure B6](#). Additional censoring rules and time-window definitions (including handling of delayed diagnoses and events around death) are provided in [subsection B.1](#).

4.2 Ethics

This retrospective, observational study used de-identified registry and EHR data from the Capital Region of Denmark and Region Zealand. In accordance with the Danish Act on Ethics Review of Health Research Projects and Health Data Research Projects (the “Committee Act”) and its 2024 amendments (including Sections 5a and 21c), projects based solely on secondary use of existing registry data (“dry” data), without any trial-specific intervention or contact with patients, are classified as health data research projects and do not require informed consent from individual patients or approval by a research ethics committee. Instead, such projects are authorised by the relevant data protection authorities.

4.3 Model Development

Deep learning on EHRs took off with BEHRT [31] and Med-BERT [32] serving as the proof-of-concept of applying transformer models to EHRs based on the successful architectures in natural language processing [33]. CORE-BEHRT [34] iteratively improved this and enabled longer attention windows by using Modern-BERT, and is used for pretraining on sequences of tokenized health data events, creating health embeddings. The methodology was extended, with inspiration from Targeted BERT [35], to a multi-head fine-tuning architecture simultaneously fine-tuning embeddings

for the propensity of treatment, $P(E | X)$, but also for the risk of N outcomes Y_i on which the causal effect of treatment must be estimated, $P(Y_i | X, E)$, $i = 1, \dots, N$.

Pretraining, fine-tuning, and calibration were performed as described in the supplementary material. Uncertainty was estimated using bootstrapping in nested fine-tuning sampling loops. For a more detailed description, see [Appendix E](#).

4.4 Statistical Analysis

Samples in the control group were weighted with the relative propensity score

$$w(X) = \frac{P(E | X)}{1 - P(E | X)}$$

to estimate the treatment effect on the treated. Log-risk ratios were computed with standard deviations and 95% confidence intervals from the nested bootstrap. The effects were characterized by a weighted risk ratio, logit risk ratio significance level (BH), and prevalence (%) of outcome. Logit risk ratios were tested with an unpaired t -test and considered statistically significant when $p < 0.05$, with Benjamini–Hochberg adjustment to control the false discovery rate.

5 Declarations

Acknowledgements. The study was financed through governmental grants DK-IFD 1061-00077B and DNRF P1.

Competing Interest. None relevant for this manuscript.

Data Availability. Data were obtained from Region Zealand and the Capital Region of Denmark and the Danish Health Data Authority. For access to register data please contact Danish Health Data Authority: <https://english.sundhedsdatastyrelsen.dk/health-data-and-registers/research-services>

Code Availability. The following software packages were used in this work. The main pipeline is available at https://github.com/kirilklein/PHAIR_EHR and is largely based on the EHR framework available at <https://github.com/FGA-DIKU/EHR> [36]. Causal estimates were computed using the CausalEstimate library, available at <https://github.com/kirilklein/CausalEstimate>.

Appendix A Results

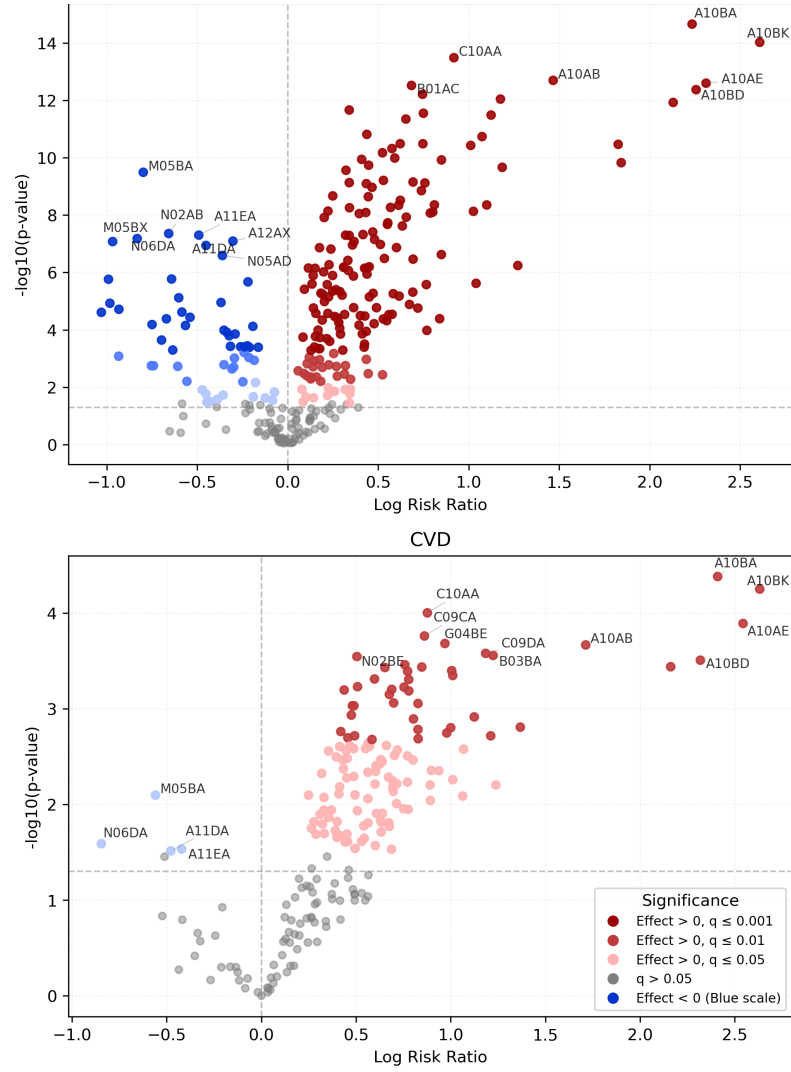


Fig. A1: Volcano plot showing the significance vs effect strength for the main and CVD cohorts.

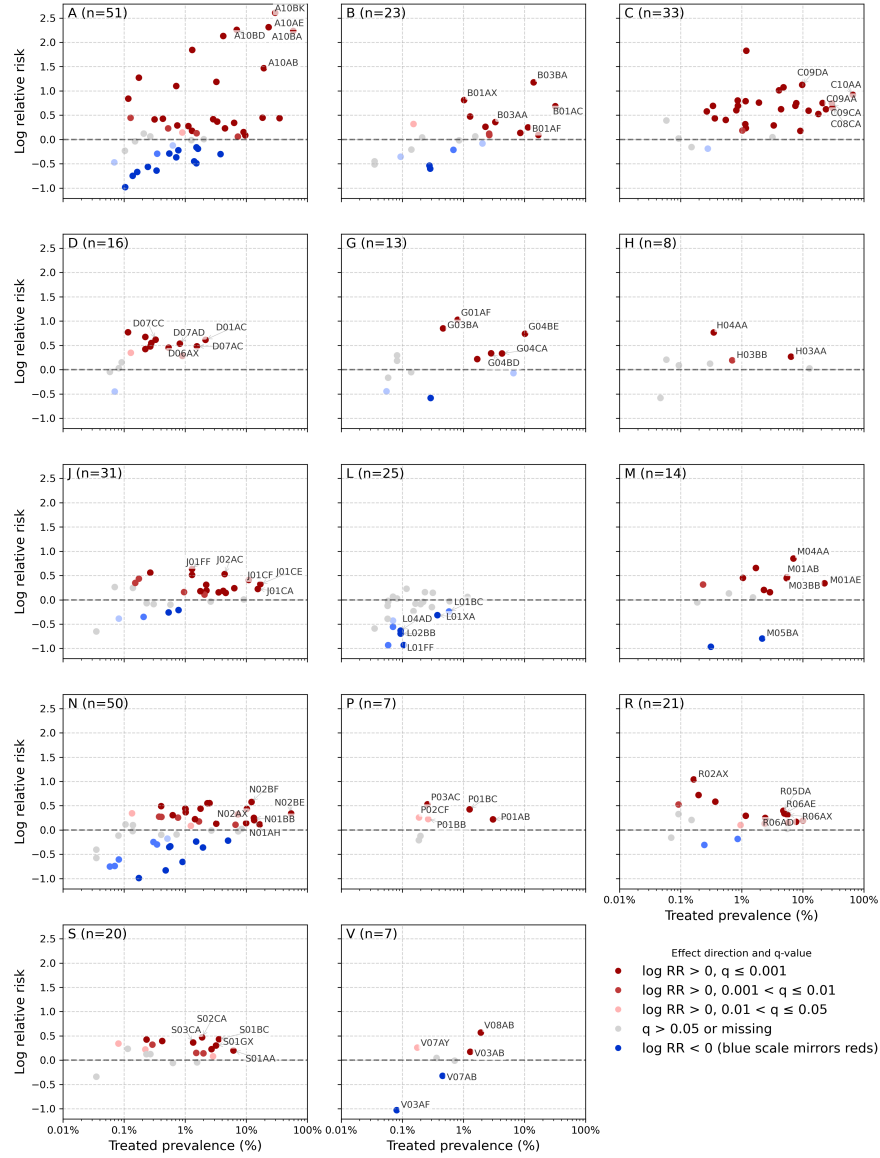


Fig. A2: Adjusted log risk ratios versus treated prevalence for all drug outcomes in the age ≥ 50 years cohort. Panels A–V correspond to ATC level 1 categories. Each point represents a single ATC level 4 medication outcome. Red points denote increased risk ($RR \geq 1$) and blue points reduced risk ($RR \leq 1$) in semaglutide users. Colour intensity indicates the Benjamini–Hochberg-adjusted p-value. Grey indicates insignificant results.

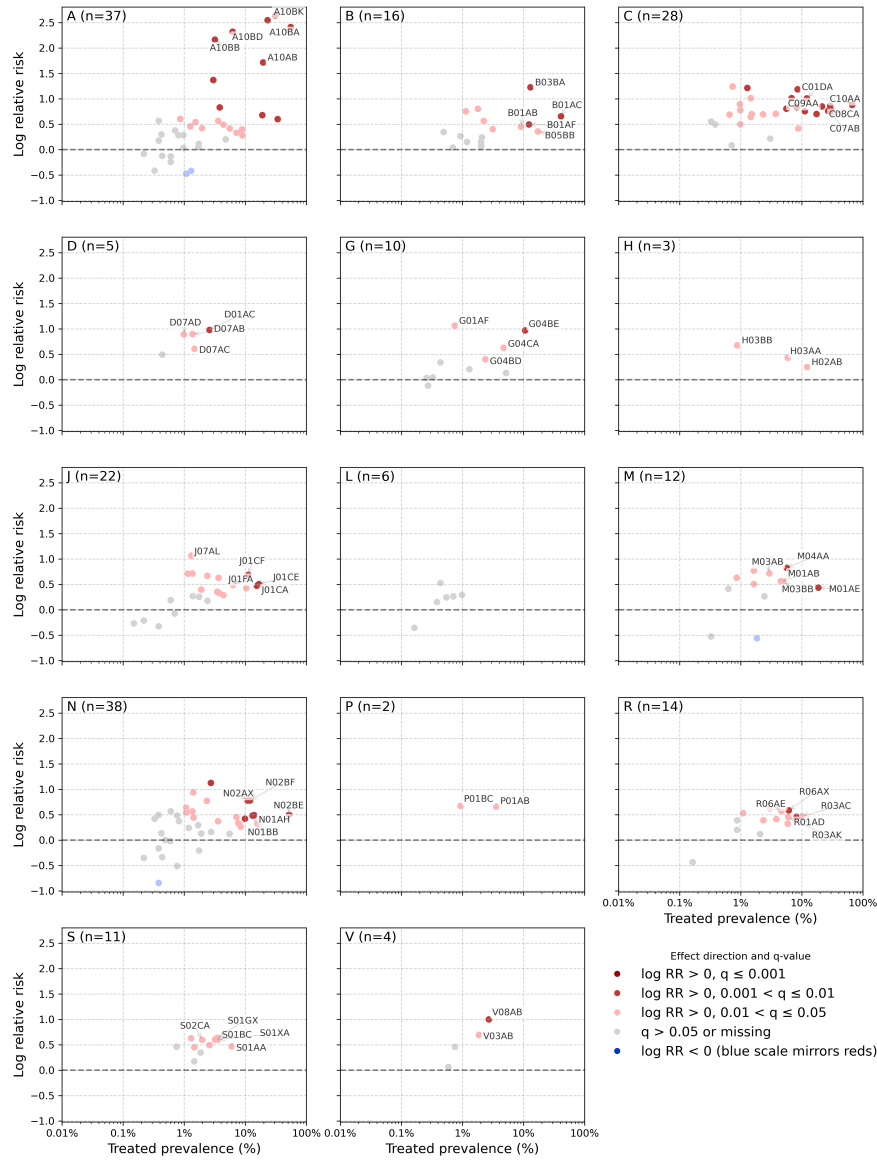


Fig. A3: Adjusted log risk ratios versus treated prevalence for all drug outcomes in the cardiovascular disease (CVD) cohort. Layout, colouring, and interpretation are as in Figure A2.

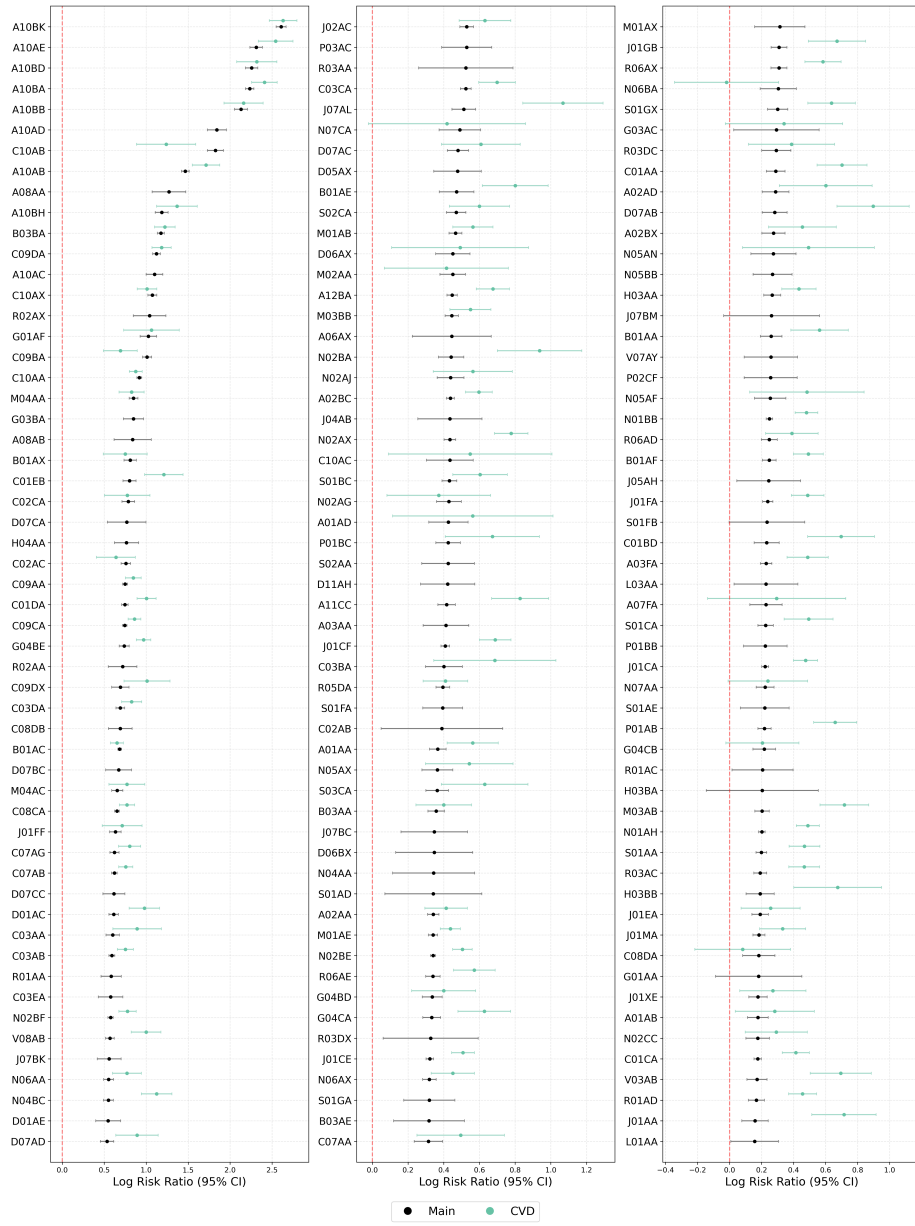


Fig. A4: Forest plots of adjusted log risk ratios (95% CI) for all (included) ATC level 4 medications in the main and secondary cohort, ordered by effect size (Part 1).

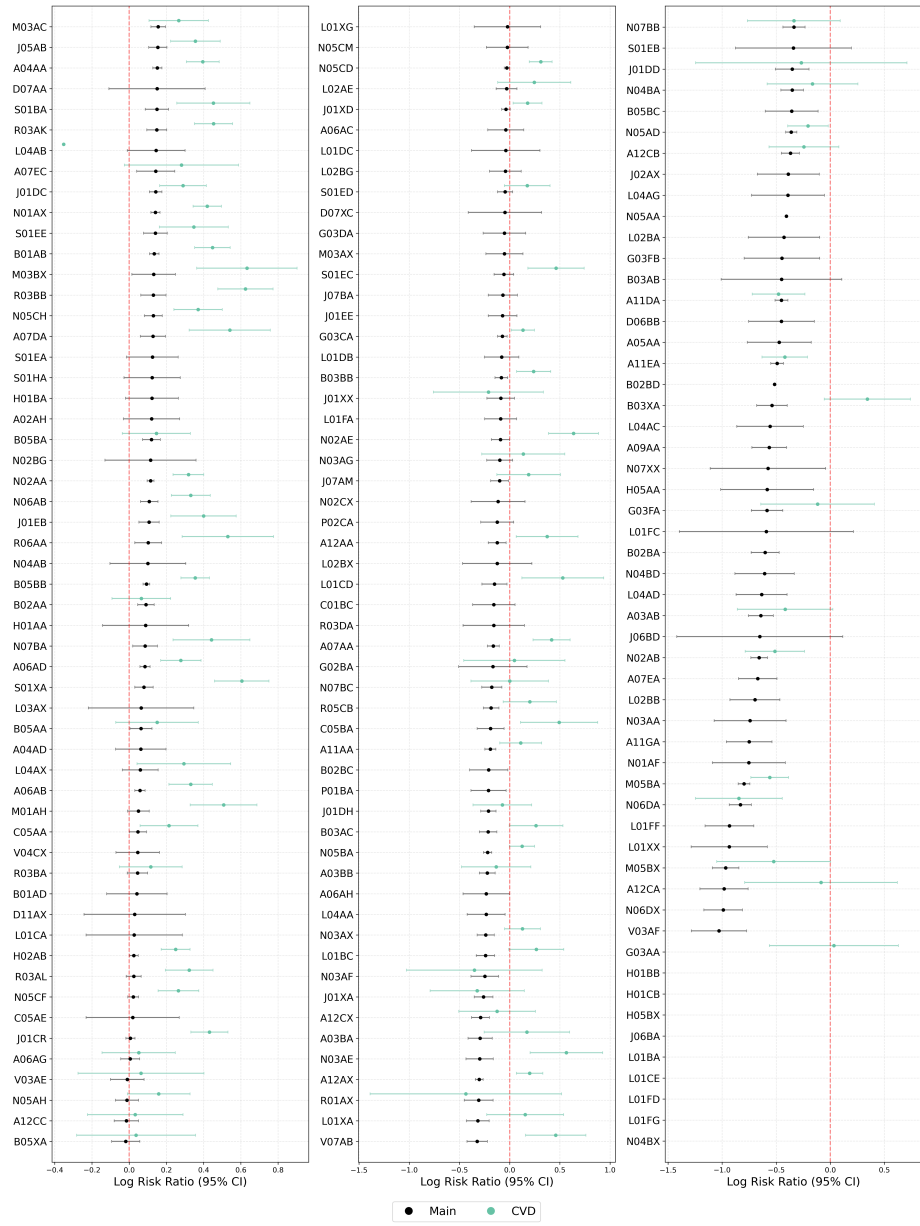


Fig. A5: Forest plots of adjusted log risk ratios (95% CI) for all (included) ATC level 4 medications in the main and secondary cohort, ordered by effect size (Part 2).

Appendix B Study Design

B.1 Censoring and time windows

We applied a set of time-window and censoring rules with four goals: (i) to include baseline covariates that may be coded with delay around the index date, (ii) to delay the start of outcome follow-up so that pre-existing conditions are not misclassified as outcomes, (iii) to define the end of follow-up in a way that treats exposed and unexposed patients symmetrically and reflects treatment discontinuation, and (iv) to capture outcomes that are coded shortly after death.

Pre-index truncation of covariates.

For each patient, we truncated the input sequence 24 hours before the index date. This avoids information leakage of the exposure, for example, if co-occurring drugs are administered immediately before the first semaglutide prescription. This truncation affects the construction of baseline covariates and model inputs, but not the start of person-time.

After applying the censoring, sequences are truncated from the left to fit into the 512-long context window of the transformer.

Post-index windows for baseline covariates and outcomes.

To allow for delayed registration of diagnoses while still treating them as baseline confounders rather than post-treatment mediators, diagnoses were allowed to enter the baseline covariate set up to 336 hours (14 days) after the index date. Diagnoses recorded after this window were treated as potential outcomes or downstream consequences of treatment. To avoid misclassifying pre-existing conditions that were coded late as outcomes, we additionally introduced a post-index run-in period: outcomes were only counted after a run-in period of six months after the index date, and events in this period were not considered outcomes in the effect estimation.

End of follow-up and treatment discontinuation.

The main analysis defined the end of follow-up as the earliest of two years after index, death, or treatment discontinuation. To approximate discontinuation while allowing for some irregularity in prescription refills, we defined a maximum allowed gap of seven days between consecutive semaglutide prescriptions. Follow-up was censored at the first time this gap was exceeded. In addition, if a patient in a (index-date) linked set died or was lost to follow-up before 2 years, we censored the corresponding patients with the same index date to keep follow-up distributions comparable between exposed and unexposed groups.

Events around death.

To account for delayed coding close to the end of life, we considered outcome events occurring up to 14 days after the recorded date of death. This window applies, for example, to prescriptions dispensed shortly before death but registered with a delay.

B.2 Cohort construction and CVD definition

Here, we provide the patient flow chart and the detailed definition of cardiovascular disease (CVD) used for the CVD cohort.

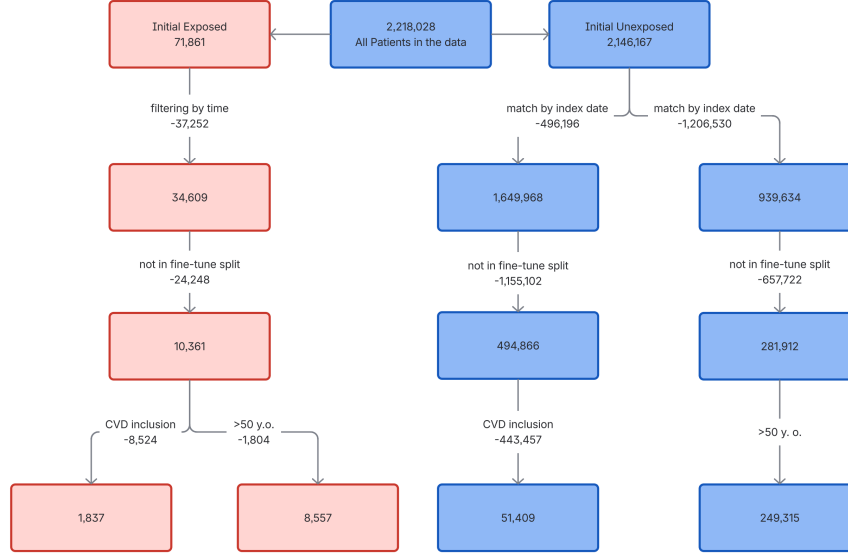


Fig. B6: Flow chart of cohort selection for exposed and unexposed patients. Starting from all patients with available EHR data ($N = 2,218,028$), we first identified exposed patients (left) and unexposed patients (right). Exposed patients were restricted by follow-up time, presence in the tuning split with sufficient data to evaluate eligibility criteria, and then split into two analytic cohorts: (i) patients with prior cardiovascular disease (CVD) and (ii) patients aged ≥ 50 years at index. For each analytic cohort, separate pools of unexposed patients were created, index dates were matched to exposed patients, and the same tuning-split and cohort-specific inclusion criteria were applied to define the final control groups. Final boxes show the numbers of exposed and matched unexposed patients included in each cohort.

Table B1: Definition of cardiovascular disease (CVD) used for cohort selection.

CVD component	Codes / definition
Myocardial infarction	DI2[1-3].*
Stroke	DI6[3-6].*
Transient ischaemic attack	DG45.*
Arterial revascularization	KFN.*
Reduced kidney function	.*EGFR.*1[.,]73.*, eGFR $\leq$ 60 ml/min/1.73 m ²

Appendix C Description of Algorithm

We pre-trained a modernbert [37] model on 70% of patients with a masked language modelling objective similar to [34]. The model was pretrained using masked language modelling as described in [34]. The vocabulary size after remapping rare codes was 40,839. The architecture uses hidden size 256, intermediate size 1024, 4 attention heads, 4 hidden layers, and about 25.5M parameters. We used a learning rate of 0.001 with 4 warmup epochs, batch size 512, and trained up to 50 epochs with early stopping after 3, gradient clipping at 1.0, and a constant-with-warmup scheduler. The model was pre-trained on an A100 GPU for about one day. During fine-tuning, we use a shared BiGRU head as shown in Figure C7, training jointly on Exposure and all the available outcomes. To ensure stability during training, we removed all outcomes with fewer than 100 total positive occurrences. On top of the confounder embedding produced by the BiGRU, we had an MLP with two layers performing binary classification. For the outcomes, we also used the exposure status, concatenating it with the confounder embedding (-1 for negative and 1 for positive).

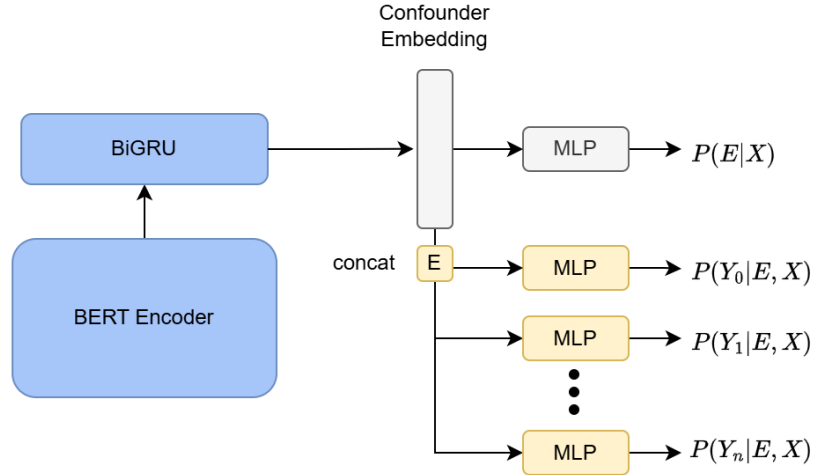


Fig. C7: We use a shared pre-trained ModernBert [37] encoder. The output sequence gets processed by a bi-directional gated recurrent unit to obtain one confounder embedding per subject. A multi-layer perceptron (MLP) then predicts the propensity score. In addition, concatenated with the exposure status, the embedding is passed to one MLP per outcome.

Appendix D Data Preprocessing

We first extract data from tables described in [Table D2](#), including diagnoses, medication, laboratory test results, procedures, and encounter information. Summary statistics on token counts and vocabulary size for each event type are reported in [Table D2](#). We map rare codes (< 50 occurrences) that have a hierarchy to the parent. For labtest, we create a category called lab_rare, while for diagnoses, we use the icd10 hierarchy, and for medication, the ATC hierarchy.

We convert this data to the MEDS format [\[30\]](#). All events with missing timestamps are removed. As in corebeht [\[34\]](#), we prepend gender as a token to the sequence.

We aggregate repeating codes within a one-hour time window. As part of the lab tests, some vital signs are logged repeatedly. As a result they quickly fill up the context window. We thus group them into one event if they are close to each other in time.

D.1 Statistics on total number of tokens and unique codes.

Table D2: Total token counts and number of unique codes by group.

Group	Count (millions)	Unique Codes
Medication	357.0	2,067
Register medication	290.3	1,382
Procedures	105.5	12,474
Contacts	84.8	59
Labs	79.6	8,482
Register procedures	63.6	17,041
Hospital visit	50.8	896
Register diagnoses	42.0	20,507
Diagnoses	15.9	17,103
Other	4.6	6
Total	1094.2	80,017

Appendix E Estimation of Risk Ratios and Confidence Intervals

In order to get reliable uncertainty estimates for the effect measures, we performed a two-step resampling procedure. Since the propensity scores themselves are subject to uncertainty due to finite sample size and stochastic optimization, we explicitly propagate this uncertainty into the effect estimates. Methodological work on variance estimation for propensity-score analyses has emphasised the importance of accounting for propensity-score uncertainty when constructing confidence intervals for treatment effects [38]. Our procedure is designed to take this into account. First, we ran the fine-tuning procedure $K = 10$ times (with different random seeds and reshufflings of the data), obtaining K distinct fitted exposure heads and thus K sets of propensity scores $\{\hat{e}_k(X_i)\}_{i=1}^n$, $k = 1, \dots, K$. For each refit k , we computed inverse-probability weights and the corresponding weighted risks in the treated and control arms for each exposure–outcome pair.

To ensure common support, we removed patients whose propensity scores lay outside the empirical 0.1th and 99.9th percentiles of the propensity-score distribution. Conditional on each set of propensity scores $\hat{e}_k$, we then quantified sampling uncertainty by a non-parametric bootstrap over patients. Specifically, for each refit k we drew B bootstrap samples from the original cohort (sampling patients with replacement), recomputed the weighted risks per arm in each bootstrap sample, and stored the resulting estimates. In our experiments, we used $B = 100$ bootstrap replicates.

Let $\hat{\eta}_1$ and $\hat{\eta}_0$ denote the estimated log-odds (logits) for the treatment and control groups, respectively, with corresponding variances $\text{Var}(\hat{\eta}_1)$ and $\text{Var}(\hat{\eta}_0)$. The point

estimate for the Risk Ratio ($\widehat{RR}$) is calculated by transforming logits to probabilities using the inverse-logit function $\sigma(\cdot)$:

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_0} = \frac{\sigma(\hat{\eta}_1)}{\sigma(\hat{\eta}_0)} = \frac{(1 + e^{-\hat{\eta}_0})}{(1 + e^{-\hat{\eta}_1})} \quad (\text{E1})$$

To construct confidence intervals, we work on the logarithmic scale to ensure symmetry and strictly positive limits. We approximate the standard error of $\log(\widehat{RR})$ using the Delta method. Noting that $\frac{\partial \log p}{\partial \eta} = 1 - p$, the approximate variance is:

$$\text{Var}(\log \widehat{RR}) \approx (1 - \hat{p}_1)^2 \text{Var}(\hat{\eta}_1) + (1 - \hat{p}_0)^2 \text{Var}(\hat{\eta}_0) \quad (\text{E2})$$

where the variances $\text{Var}(\hat{\eta}_1)$ and $\text{Var}(\hat{\eta}_0)$ are estimated from the variability across the $K \times B$ resampled estimates. Finally, the 95% confidence interval for the risk ratio is obtained by exponentiating the bounds computed on the log scale:

$$CI_{95\%} = \exp \left(\log(\widehat{RR}) \pm 1.96 \sqrt{\text{Var}(\log \widehat{RR})} \right) \quad (\text{E3})$$

References

- [1] Blehar, M.C., Spong, C., Grady, C., Goldkind, S.F., Sahin, L., Clayton, J.A.: Enrolling pregnant women: issues in clinical research. *Women's Health Issues* **23**(1), 39–45 (2013) <https://doi.org/10.1016/j.whi.2012.10.003>
- [2] Shields, K.E., Lyster, A.D.: Exclusion of pregnant women from industry-sponsored clinical trials. *Obstetrics & Gynecology* **122**(5), 1077–1081 (2013) <https://doi.org/10.1097/AOG.0b013e3182a9ca67>
- [3] Eke, A.C., Olagunju, A., Momper, J., Penazzato, M., Abrams, E.J., Best, B.M., Capparelli, E.V., Bekker, A., Belew, Y., Kiser, J.J., Struble, K., Taylor, G., Waitt, C., Mirochnick, M., Cressey, T.R., Colbers, A.: Optimizing pharmacology studies in pregnant and lactating women using lessons from hiv: a consensus statement. *Clinical Pharmacology & Therapeutics* **110**(1), 36–48 (2021) <https://doi.org/10.1002/cpt.2048>
- [4] Radley, D.C., Finkelstein, S.N., Stafford, R.S.: Off-label prescribing among office-based physicians. *Archives of Internal Medicine* **166**(9), 1021–1026 (2006)
- [5] Downing, N.S., Shah, N.D., Aminawung, J.A., Pease, A.M., Zeitoun, J.-D., Krumholz, H.M., Ross, J.S.: Postmarket safety events among novel therapeutics approved by the us food and drug administration between 2001 and 2010. *JAMA* **317**(18), 1854–1863 (2017)
- [6] Wallach, J.D., Egilman, A.C., Dhruva, S.S., McCulloch, C.E., Krumholz, H.M., Ross, J.S.: Postmarket studies required by the US food and drug administration: a cross-sectional analysis. *BMJ* **361**, 2031 (2018)
- [7] Golder, S., Loke, Y.K., Wright, K., Norman, G.: Reporting of adverse events in published and unpublished studies of health care interventions: a systematic review. *PLOS Medicine* **13**(9), 1002127 (2016) <https://doi.org/10.1371/journal.pmed.1002127>
- [8] Hazell, L., Shakir, S.A.W.: Under-reporting of adverse drug reactions: a systematic review. *Drug Safety* **29**(5), 385–396 (2006) <https://doi.org/10.2165/00002018-200629050-00003>
- [9] Davis, S.E., Zabolka, L., Desai, R.J., Matheny, M.E., Johnson, K., Marsolo, K., Curtis, L.H.: Use of electronic health record data for drug safety signal identification: a scoping review. *Drug Safety* **46**(8), 725–742 (2023) <https://doi.org/10.1007/s40264-023-01325-0>
- [10] Schuemie, M.J., Ryan, P.B., Pratt, N., *et al.*: Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study. *Journal of the American Medical Informatics Association* **27**(8), 1268–1277 (2020) <https://doi.org/10.1093/jamia/ocaa124>

- [11] Hallas, J., Pottegård, A.: Use of self-controlled designs in pharmacoepidemiology. *Journal of Internal Medicine* **275**(6), 581–589 (2014) <https://doi.org/10.1111/joim.12186>
- [12] Zarei, M., Sabetkasaei, M., Mozafari, M., Zaeri, S.: The expanding role of semaglutide: beyond glycemic control. *Journal of Diabetes and Metabolic Disorders* **24**(2), 160 (2025) <https://doi.org/10.1007/s40200-025-01663-z>
- [13] Jensen, T.B., Andersen, J.T., Jimenez-Solem, E., Lund, M.: How patients in Denmark acquire their medicines: overview, data sources and implications for pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology* **128**(1), 46–51 (2021) <https://doi.org/10.1111/bcpt.13472>
- [14] Lin, H., Tsai, Y., Liao, P., Wei, J.C.: Neurodegeneration and stroke after semaglutide and tirzepatide in patients with diabetes and obesity. *JAMA Network Open* **8**(7), 2521016 (2025) <https://doi.org/10.1001/jamanetworkopen.2025.21016>
- [15] Nørgaard, C.H., *et al.*: Treatment with glucagon-like peptide-1 receptor agonists and incidence of dementia: Data from pooled double-blind randomized controlled trials and nationwide disease and prescription registers. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* **8**(1), 12268 (2022) <https://doi.org/10.1002/trc2.12268>
- [16] Wang, W., Wang, Q., Qi, X., Gurney, M., Perry, G., Volkow, N.D., Davis, P.B., Kaelber, D.C., Xu, R.: Associations of semaglutide with first-time diagnosis of Alzheimer’s disease in patients with type 2 diabetes: target trial emulation using nationwide real-world data in the US. *Alzheimer’s & Dementia* **20**(12), 8661–8672 (2024) <https://doi.org/10.1002/alz.14313>
- [17] De Giorgi, R., Koychev, I., Adler, A., Cowen, P., Harmer, C., Harrison, P.: 12-month neurological and psychiatric outcomes of semaglutide use for type 2 diabetes: a propensity-score matched cohort study. *EClinicalMedicine* **74**, 102726 (2024) <https://doi.org/10.1016/j.eclinm.2024.102726>
- [18] Cummings, J.L., Atri, A., Feldman, H.H., Hansson, O., Sano, M., Knop, F.K., Johannsen, P., León, T., Scheltens, P.: evoke and evoke+: design of two large-scale, double-blind, placebo-controlled, phase 3 studies evaluating efficacy, safety, and tolerability of semaglutide in early-stage symptomatic Alzheimer’s disease. *Alzheimer’s Research & Therapy* **17**(1), 14 (2025) <https://doi.org/10.1186/s13195-024-01666-7>
- [19] Novo Nordisk A/S: Evoke phase 3 trials did not demonstrate a statistically significant reduction in Alzheimer’s disease progression. Company announcement. Accessed 29 November 2025 (2025). <https://www.novonordisk.com/content/nncorp/global/en/news-and-media/news-and-ir-materials/news-details.html?id=916462#>

- [20] Davies, M., Færch, L., Jeppesen, O.K., Pakseresht, A., Pedersen, S.D., Perreault, L., Rosenstock, J., Shimomura, I., Viljoen, A., Wadden, T.A., Lingvay, I.: Semaglutide 2.4 mg once a week in adults with overweight or obesity, and type 2 diabetes (step 2): a randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. *Lancet* **397**(10278), 971–984 (2021) [https://doi.org/10.1016/S0140-6736\(21\)00213-0](https://doi.org/10.1016/S0140-6736(21)00213-0)
- [21] Able, C., *et al.*: Prescribing semaglutide for weight loss in non-diabetic, obese patients is associated with an increased risk of erectile dysfunction: a trinetx database study. *International Journal of Impotence Research* **37**(4), 315–319 (2025) <https://doi.org/10.1038/s41443-024-00895-6>
- [22] Kounatidis, D., *et al.*: The impact of glucagon-like peptide-1 receptor agonists on erectile function: Friend or foe? *Biomolecules* **15**(9), 1284 (2025) <https://doi.org/10.3390/biom15091284>
- [23] Ruggiero, R., Longo, M., *et al.*: Real-world safety comparison of liraglutide and semaglutide in weight management: insights from European pharmacovigilance data. *European Journal of Pharmacology* **1004**, 178004 (2025) <https://doi.org/10.1016/j.ejphar.2025.178004>
- [24] Chen, J., Zhang, Q., Wu, Q., Zhang, X., Xiang, Z., Zhu, S., Dai, T., Si, Y.: Impact of GLP-1 receptor agonists on suicide behavior: a meta-analysis based on randomized controlled trials. *Journal of Diabetes* **17**(9), 70151 (2025) <https://doi.org/10.1111/1753-0407.70151>
- [25] Wang, M., Chen, X., Liu, Z., Li, Z., Zhu, Z., Liu, S., Xiao, S.: Exploring potential associations between GLP-1RAs and depressive disorders: a pharmacovigilance study based on FAERS and Vigibase data. *EClinicalMedicine* **86**, 103385 (2025) <https://doi.org/10.1016/j.eclinm.2025.103385>
- [26] Wang, M., Yang, Z., Yan, M., Liu, S., Xiao, S.: Depression and suicide/self-injury signals for weight loss medications: a disproportionality analysis of semaglutide, liraglutide, and tirzepatide in FAERS database. *Journal of Affective Disorders* **389**, 119670 (2025) <https://doi.org/10.1016/j.jad.2025.119670>
- [27] García-Casares, N., *et al.*: Effects of GLP-1 receptor agonists on neurological complications of diabetes. *Reviews in Endocrine and Metabolic Disorders* **24**(4), 655–672 (2023) <https://doi.org/10.1007/s11154-023-09807-3>
- [28] Simonsen, E., Lund, L.C., Ernst, M.T., Hjellvik, V., Hegedüs, L., Hamann, S., Jørstad, Ø.K., Gulseth, H.L., Karlstad, Ø., Pottgård, A.: Use of semaglutide and risk of non-arteritic anterior ischemic optic neuropathy: a Danish-Norwegian cohort study. *Diabetes, Obesity and Metabolism* **27**(6), 3094–3103 (2025) <https://doi.org/10.1111/dom.16316>

- [29] Herrou, J., *et al.*: Narrative review of effects of glucagon-like peptide-1 receptor agonists on bone health in people living with obesity. *Calcified Tissue International* (2023) <https://doi.org/10.1007/s00223-023-01150-8>
- [30] McDermott, M.B., Xu, J., Bergamaschi, T.S., Jeong, H., Lee, S.A., Oufattole, N., Rockenschaub, P., Stankevičiūtė, K., Steinberg, E., Sun, J., *et al.*: Meds: Building models and tools in a reproducible health ai ecosystem. In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6243–6244 (2025)
- [31] Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: Behrt: transformer for electronic health records. *Scientific reports* **10**(1), 1–12 (2020)
- [32] 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), 86 (2021)
- [33] Lee, J., Toutanova, K.: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **3**(8), 4171–4186 (2018)
- [34] Odgaard, M., Klein, K.V., Thysen, S.M., Jimenez-Solem, E., Sillesen, M., Nielsen, M.: Core-behrt: A carefully optimized and rigorously evaluated behrt. *arXiv preprint arXiv:2404.15201* (2024)
- [35] 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 Transactions on Neural Networks and Learning Systems* (2022)
- [36] Montgomery, M.E., Klein, K., Odgaard, M., Lorenzen, S.S., Sobhaninia, Z.: Bonsai: A framework for processing and analysing electronic health records (ehr) data using transformer-based models. *Journal of Open Source Software* **10**(114), 8869 (2025)
- [37] Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., Poli, I.: Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference (2024). <https://arxiv.org/abs/2412.13663>
- [38] Austin, P.C., Small, D.S.: The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine* **33**(24), 4306–4319 (2014)

Chapter 7

Summary and Discussion

7.1 Recap

At the beginning of this thesis, we set out to answer the following question:

Can we use modern representation learning together with causal inference to make drug safety surveillance more systematic and scalable while respecting pharmacoepidemiological principles?

The thesis proceeds in three steps.

First, we studied how to obtain robust, reusable EHR representations with transformer models and what their limitations are for downstream clinical tasks. Second, we combined these representations with a careful target-trial emulation framework to address a single, clinically anchored question: the cardiovascular effects of semaglutide in high-risk patients with type 2 diabetes. Finally, we scaled this approach to more than 300 drug-outcome pairs in an untargeted semaglutide screening, showing that the transformer-based propensity scores can support high-throughput, hypothesis-generating pharmacovigilance.

Taken together, these parts show that representation learning and causal inference are not competing paradigms but complementary parts of a practical toolkit for observational drug safety research. Representations provide flexible summaries of complex patient histories, while target-trial thinking and cautious causal estimation ensure that these summaries are used within a transparent, interpretable framework.

7.2 Summary of Main Contributions

7.2.1 Transformers for electronic health records

The first main contribution is a systematic exploration of transformer architectures for longitudinal EHR data. Rather than proposing yet another bespoke model, we focused on clarifying design choices that matter in practice: how to tokenise heterogeneous clinical

events, how to encode temporal structure, how to pool information across long trajectories, and how to pretrain in a way that transfers across tasks.

We showed that a carefully configured transformer can serve as a general-purpose EHR encoder, delivering competitive performance across risk prediction tasks without extensive task-specific feature engineering. At the same time, our experiments highlighted important limitations. Representation quality is sensitive to seemingly minor design decisions (such as how to handle repeated events or long-range dependencies), and gains are not uniform across all subgroups and outcomes. This motivates the cautious view taken in the rest of the thesis: transformers can be appropriate PS models, but they are not magical black boxes that automatically solve all modelling problems.

7.2.2 Causal transformer models for a single, clinically anchored question

The second main contribution is a case study in which transformer-based representations are integrated into a classical pharmacoepidemiologic workflow to address a single, well-defined causal question. Using a target-trial emulation design, we specified the population, intervention, comparator, outcomes, follow-up time, and setting in close analogy to the SUSTAIN-6 trial. We then used the pretrained encoder to construct high-dimensional summaries of pre-treatment history and estimated transformer-based propensity scores for semaglutide initiation.

Inverse-probability weighting with stabilised or overlap weights was used to estimate effects for major adverse cardiovascular events. The emulated results are qualitatively consistent with the SUSTAIN-6 findings, while naturally reflecting the different patient mix, follow-up patterns, and outcome definitions in routine care. This chapter demonstrates that transformer-based PS models can be used in target-trial emulations as a drop-in for traditional models. Beyond answering this specific clinical question, the chapter is intentionally written as a bridge between communities: terminology, figures, and diagnostics are chosen to map transformer-based PS modelling onto concepts that are already familiar in pharmacoepidemiology. In that sense, it functions as an educational piece that lowers the barrier for epidemiologists to experiment with transformer models without abandoning established design principles.

7.2.3 Untargeted causal screening for semaglutide

The third main contribution is an untargeted screening analysis of semaglutide across hundreds of downstream medication outcomes. Here we reuse the same basic ingredients—new-user design, symmetric eligibility, pre-treatment transformer embeddings, and PS weighting—but scale up from a single endpoint to an atlas of ATC level-4 medication families.

For each outcome, we estimate transformer-based PS, construct weights, and compute adjusted risk estimates, accompanied by a shared set of diagnostics and multiplicity control. The resulting pattern of associations is clinically plausible: expected gastrointestinal

effects and related treatment changes are prominent; other clusters (e.g. potential effects on bone health or neurocognitive outcomes) emerge as hypothesis-generating signals rather than definitive findings. This chapter illustrates that high-dimensional nuisance modelling can support systematic, large-scale pharmacovigilance.

7.2.4 Reflection on the causal applications

The two causal applications are deliberately complementary. The semaglutide trial emulation offers *depth*: a single, well-specified question with a close RCT benchmark, clear design, and strong emphasis on explaining the role of transformers to an epidemiologic audience. The untargeted screening analysis offers *breadth*: a systematic framework showing that essentially the same machinery scales to hundreds of outcomes and can be used as a front-end for signal generation. This dual role, as both case study and tutorial, reflects a broader goal of the thesis: to build a shared language between machine learning and pharmacoepidemiology around target-trial thinking and transformer-based confounding adjustment.

Even with flexible models, calibration and overlap can remain fragile. Small changes in model specification or calibration can have large effects on weight stability, especially in the tails of the propensity distribution. This motivated the emphasis on regularisation and calibration in the methodological discussion. Interpretation was another challenge. High-dimensional confounding adjustment is opaque, and it remains difficult to communicate to clinicians and regulators exactly which features drive adjustment and how much residual bias may remain. These difficulties do not negate the approach, but they underline that representation learning must be accompanied by careful design, transparent reporting, and humility about what can and cannot be learned from observational data.

7.3 Limitations

Several limitations qualify the findings of this thesis.

First, residual confounding and measurement error are unavoidable in routine care data. Key variables such as lifestyle, frailty, over-the-counter medication use, and disease severity are incompletely captured, if at all. Our strategy of adjusting for pre-treatment histories via transformer embeddings can mitigate some of this through proxies, but it cannot guarantee conditional exchangeability. Delayed and incomplete coding of diagnoses and procedures further complicates the definition of clean pre-treatment windows.

Second, transportability is limited. All analyses are based on Danish EHR and register data, with their specific coding practices, health-care organisation, and prescribing patterns. Both the learned representations and the estimated treatment effects may behave differently in other countries, health-care systems, or time periods. External validation and adaptation will be necessary before using similar models in other settings.

Third, the approach is computationally demanding. Training large transformer models on long EHR sequences and repeatedly fitting nuisance models for many outcomes requires substantial GPU resources and engineering effort. This may slow down iteration and limit

the feasibility of very frequent re-estimation in real-time monitoring contexts. section 7.4 discusses several directions for efficiency improvements, but they remain to be fully implemented and evaluated.

Finally, the empirical scope is intentionally narrow. Semaglutide serves as a rich and clinically important case study, but it represents only one drug class. Likewise, the untargeted screening focuses on a particular family of outcomes (ATC level-4 medication patterns) rather than on the full space of possible clinical endpoints. The methods are designed to generalise beyond this setting, but this has not yet been demonstrated empirically.

7.4 Future Directions

The methodological and empirical work in this thesis suggests several natural extensions.

A first line of work concerns *better regularisation and calibration of PS models*. Developing outcome-aware regularisation schemes, principled ways of attenuating instrument-like signals, and calibration procedures that explicitly trade off predictive performance and weight stability could make propensity-based estimators more robust and interpretable.

A second direction is *scaling to untargeted analysis*. Applying the same framework to other drug classes and to richer outcome families (including diagnosis-based phenotypes and composite safety endpoints) would help map out where transformer-based confounding adjustment is most useful and where traditional, low-dimensional approaches may suffice. In the longer term, integrating these analyses into routine, near-real-time monitoring pipelines is an important goal.

Third, there is considerable room for *efficiency and engineering improvements*. Longer patient histories, more frequent model updates, and broader coverage of exposures and outcomes will require more efficient attention mechanisms, memory-aware training strategies, and lightweight fine-tuning. These directions are outlined in more detail in the dedicated future-work chapter and form a natural bridge between methodological research and deployable systems.

In summary, this thesis takes a first step towards combining modern representation learning with principled causal inference for observational drug safety.

Bibliography

- [1] Mikkel Odgaard, Kiril Vadimovic Klein, Sanne Møller Thysen, Espen Jimenez-Solem, Martin Sillesen, and Mads Nielsen. Core-behrt: A carefully optimized and rigorously evaluated behrt. In *Proceedings of Machine Learning Research*, volume 252, pages 1–33. PMLR, 2024.
- [2] Kiril Vadimovic Klein, Josefine Windfeld Mathiasen, Espen Jimenez-Solem, Thomas Leth Jensen, Kasper Bonnesen, Martin Hylleholdt Sillesen, Mads Nielsen, and Sanne Møller Thysen. An introduction to causal transformer models in epidemiological studies. Manuscript in preparation, 2025.
- [3] Kiril Vadimovic Klein, Maria Montgomery, Mikkel Odgaard, Sanne Møller Thysen, Thomas Jensen, Espen Jimenez-Solem, Martin Sillesen, and Mads Nielsen. Untargeted causal drug effect estimation: Populationwise effects of semaglutide. Manuscript in preparation, 2025.
- [4] Jason Lazarou, Bruce H Pomeranz, and Paul N Corey. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, 279(15):1200–1205, 1998.
- [5] Emma C Davies, Christopher F Green, Stephen Taylor, Paula R Williamson, David R Mottram, and Munir Pirmohamed. Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. *PLoS one*, 4(2):e4439, 2009.
- [6] Lorna Hazell and Saad AW Shakir. Under-reporting of adverse drug reactions. *Drug safety*, 29(5):385–396, 2006.
- [7] JE Campbell, M Gossell-Williams, and MG Lee. A review of pharmacovigilance. *The West Indian Medical Journal*, 63(7):771, 2015.
- [8] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.
- [9] Peter M Ellis, Larissa Long, Courtney H Coschi, and Arani Sathiyapalan. A narrative review of the strengths and limitations of real-world evidence in comparison to randomized clinical trials: What are the opportunities in thoracic oncology for real-world evidence to shine? *Current Oncology*, 32(11):629, 2025.

- [10] Vera Vlahović-Palčevski and Dirk Mentzer. Postmarketing surveillance. In *Pediatric clinical pharmacology*, pages 339–351. Springer, 2011.
- [11] Goran Palčevski, Nataša Skočibušić, and Vera Vlahović-Palčevski. Unlicensed and off-label drug use in hospitalized children in croatia: a cross-sectional survey. *European journal of clinical pharmacology*, 68(7):1073–1077, 2012.
- [12] Miranda R Waggoner and Anne Drapkin Lyerly. Clinical trials in pregnancy and the “shadows of thalidomide”: Revisiting the legacy of frances kelsey. *Contemporary clinical trials*, 119:106806, 2022.
- [13] Jeremy A. Greene and Scott H. Podolsky. Reform, regulation, and pharmaceuticals—the Kefauver–Harris amendments at 50. *New England Journal of Medicine*, 367(16):1481–1483, 2012. doi: 10.1056/NEJMp1210007.
- [14] Michelle Meadows. Promoting safe and effective drugs for 100 years. *FDA Consumer Magazine*, The Centennial Edition, jan 2006. URL <https://www.fda.gov/files/Promoting-Safe-and-Effective-Drugs-for-100-Years-%28download%29.pdf>. Accessed on November 29, 2025.
- [15] Eric J. Topol. Failing the public health—rofecoxib, Merck, and the FDA. *New England Journal of Medicine*, 351(17):1707–1709, 2004. doi: 10.1056/NEJMp048286.
- [16] Peter Jüni, Linda Nartey, Stephan Reichenbach, Rebekka Sterchi, Paul A. Dieppe, and Matthias Egger. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet*, 364(9450):2021–2029, 2004. doi: 10.1016/S0140-6736(04)17514-4.
- [17] Gary D Novack. Expedited regulatory product approval in the time of covid-19. *The Ocular Surface*, 26:345, 2022.
- [18] World Health Organization. Covid-19 vaccine safety surveillance manual. <https://www.who.int/publications/i/item/9789240032781>, 2020. Accessed 30 November 2025.
- [19] European Medicines Agency. Safety of covid-19 vaccines. <https://www.ema.europa.eu/en/human-regulatory-overview/public-health-threats/coronavirus-disease-covid-19/covid-19-medicines/safety-covid-19-vaccines>, 2023. Accessed 30 November 2025.
- [20] Pierre-Philippe Piché-Renaud et al. A narrative review of vaccine pharmacovigilance during mass vaccination campaigns: Focus on myocarditis and pericarditis after covid-19 mrna vaccination. *British Journal of Clinical Pharmacology*, 89(3):967–981, 2023. doi: 10.1111/bcp.15625.
- [21] Julie Durand et al. Safety monitoring of covid-19 vaccines: Perspective from the european medicines agency. *Clinical Pharmacology & Therapeutics*, 113(6):1223–1234, 2023. doi: 10.1002/cpt.2828.

- [22] Anton Pottegård. Core concepts in pharmacoepidemiology: Fundamentals of the cohort and case-control study designs. *Pharmacoepidemiology and Drug Safety*, 31(8):817–826, 2022. doi: 10.1002/pds.5482.
- [23] Rohini K. Hernandez et al. Advancing principled pharmacoepidemiologic research to support regulatory and healthcare decision making: The era of real-world evidence. *Clinical Pharmacology & Therapeutics*, 117(4):927–937, 2025. doi: 10.1002/cpt.3563.
- [24] Astrid Coste, Angel Y. S. Wong, Marleen Bokern, Andrew Bate, and Ian J. Douglas. Methods for drug safety signal detection using routinely collected observational electronic health care data: A systematic review. *Pharmacoepidemiology and Drug Safety*, 32(1):28–43, 2023. doi: 10.1002/pds.5548.
- [25] Sharon E. Davis, Luke Zabotka, Rishi J. Desai, et al. Use of electronic health record data for drug safety signal identification: A scoping review. *Drug Safety*, 46(8):725–742, 2023. doi: 10.1007/s40264-023-01325-0.
- [26] Qiaozhi Hu, Yuxian Chen, Dan Zou, and Ting Xu. Predicting adverse drug event using machine learning based on electronic health records: A systematic review and meta-analysis. *Frontiers in Pharmacology*, 15:1497397, 2024. doi: 10.3389/fphar.2024.1497397.
- [27] Muhammad A. Shamim, Muhammad A. Shamim, Poonam Arora, and Prabhat Dwivedi. Artificial intelligence and big data for pharmacovigilance and patient safety. *Journal of Medical and Surgical Public Health*, 3:100139, 2024.
- [28] Ankit Nagar, Joga Gobburu, and Aloka Chakravarty. Artificial intelligence in pharmacovigilance: Advancing drug safety monitoring and regulatory integration. *Therapeutic Advances in Drug Safety*, 16:20420986251361435, 2025. doi: 10.1177/20420986251361435.
- [29] Vivian Ukamaka Nwokedi, Michael Kayode Odedele, Taofeek Adeshina Yusuff, et al. Developing and validating predictive models for adverse drug reactions using electronic health records (ehrs): A narrative review. *Epidemiology and Health Data Insights*, 1(6):ehdi019, 2025. doi: 10.63946/ehdi/17369.
- [30] Stella Dimitsaki, Pantelis Natsiavas, and Marie-Christine Jaulent. Applying ai to structured real-world data for pharmacovigilance purposes: Scoping review. *Journal of Medical Internet Research*, 26:e57824, 2024. doi: 10.2196/57824.
- [31] Martin Kulldorff, Robert L. Davis, Michael Kolczak, Edwin Lewis, Tracy Lieu, and Richard Platt. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Statistics in Biopharmaceutical Research*, 3(1):63–74, 2011. doi: 10.1080/07474946.2011.539924.

- [32] David Martin et al. Sequential surveillance for drug safety in a regulatory environment. *Pharmacoepidemiology and Drug Safety*, 27(7):707–712, 2018. doi: 10.1002/pds.4407.
- [33] Mia Aakjær, Marie Louise De Bruin, and Morten Andersen. Epidemiological surveillance of drug safety using cumulative sequential analysis in electronic healthcare data. *Basic and Clinical Pharmacology and Toxicology*, 134(1):129–140, 2024. doi: 10.1111/bcpt.13955.
- [34] University of Copenhagen, Bispebjerg and Frederiksberg Hospital, Danish Medicines Agency, and Innovation Fund Denmark. Phair: Pharmacovigilance by ai real-time analyses. <https://di.ku.dk/english/news/2022/faster-knowledge-of-side-effects-via-artificial-intelligence/>, 2022. Accessed 30 November 2025.
- [35] Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, 3(4):277–287, 2015. doi: 10.1089/big.2015.0020.
- [36] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6:26094, 2016. doi: 10.1038/srep26094.
- [37] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pages 301–318, Boston, MA, USA, 2016. PMLR.
- [38] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29, pages 3504–3512, 2016.
- [39] Alvin Rajkomar, Eyal Oren, Kai Chen, et al. Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 1:18, 2018. doi: 10.1038/s41746-018-0029-1.
- [40] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.
- [41] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

- [42] Chao Pang, Xinzhuo Jiang, Krishna S. Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 239–260. PMLR, 2021.
- [43] Tianran Zhang, Muhao Chen, and Alex A. T. Bui. Diagnostic prediction with sequence-of-sets representation learning for clinical events. In Martin Michalowski and Robert Moskovitch, editors, *Artificial Intelligence in Medicine*, volume 12299 of *Lecture Notes in Computer Science*, pages 348–358. Springer, Cham, 2020. doi: 10.1007/978-3-030-59137-3_31.
- [44] Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abde-laali Hassaïne, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-BEHRT: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 27(2):1106–1117, 2023. doi: 10.1109/JBHI.2022.3224727.
- [45] George Hripcsak and David J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013. doi: 10.1136/amiajnl-2012-001145.
- [46] Miguel A. Hernán and James M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016. doi: 10.1093/aje/kwv254.
- [47] Benjamin A. Goldstein, Ann Marie Navar, Michael J. Pencina, and John P. A. Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, 2017. doi: 10.1093/jamia/ocw042.
- [48] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013. doi: 10.1136/amiajnl-2011-000681.
- [49] Lin Lawrence Guo, Stephen R. Pfohl, Jason Alan Fries, et al. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Applied Clinical Informatics*, 12(4):808–815, 2021. doi: 10.1055/s-0041-1735184.
- [50] Chao Pang, Vincent Jeanselme, Young Sang Choi, Xinzhuo Jiang, Zilin Jing, Aparajita Kashyap, Yuta Kobayashi, Yanwei Li, Florent Pollet, Karthik Natarajan, et al. Fomoh: A clinically meaningful foundation model evaluation for structured electronic health records. *arXiv preprint arXiv:2505.16941*, 2025.

- [51] Bert Arnrich, Edward Choi, Jason Alan Fries, Matthew BA McDermott, Jungwoo Oh, Tom Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. Medical event data standard (meds): Facilitating machine learning for health. In *ICLR 2024 Workshop on Learning from Time Series For Health*, pages 03–08, 2024.
- [52] Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jack Galifant, Jason A Fries, Jeffrey N Chiang, Jungwoo Oh, Justin Xu, et al. Meds decentralized, extensible validation (meds-dev) benchmark: Establishing reproducibility and comparability in ml for health. *Machine Learning For Health Conference*, 2024.
- [53] Maria Elkjær Montgomery, Kiril Klein, Mikkel Odgaard, Stephan Sloth Lorenzen, and Zahra Sobhaninia. Bonsai: A framework for processing and analysing electronic health records (ehr) data using transformer-based models. *Journal of Open Source Software*, 10(114):8869, 2025.
- [54] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, 2025.
- [55] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.
- [56] Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S Kalluri, Elise L Minto, Jason Patterson, Linying Zhang, George Hripcsak, Gamze Gürsoy, Noémie Elhadad, et al. Cehr-gpt: Generating electronic health records with chronological patient timelines. *arXiv preprint arXiv:2402.04400*, 2024.
- [57] Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ digital medicine*, 7(1):256, 2024.
- [58] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.
- [59] A Miguel, ROBINS HERNAN, and M James. *Causal inference: what if*. CRC PRESS, 2023.
- [60] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. 2020. URL https://content.sph.harvard.edu/wwwhsph/sites/1268/2024/01/hernanrobins_WhatIf_2jan24.pdf. Free online version (accessed 2025-11).

- [61] Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012. doi: 10.1177/0962280210386207.
- [62] Steven P. Marso et al. Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine*, 375(19):1834–1844, 2016. doi: 10.1056/NEJMoa1607141.
- [63] Shishir Rao, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Yikuan Li, Rema Ramakrishnan, Abdelaali Hassaine, Dexter Canoy, and Kazem Rahimi. Targeted-behrt: Deep learning for observational causal inference on longitudinal electronic health records. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11): 6763–6775, 2022. doi: 10.1109/TNNLS.2022.3183864.
- [64] Sebastian Schneeweiss, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun, and M Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512–522, 2009.
- [65] Fan Li, Kari Morgan, and Alan M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018. doi: 10.1080/01621459.2016.1260466.
- [66] Fan Li, Lynn E. Thomas, and Yang Li. Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188(1):250–257, 2019. doi: 10.1093/aje/kwy201.
- [67] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- [68] Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):11, 2006.
- [69] Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011. doi: 10.1007/978-1-4419-9782-1.