






An Explainable Transformer-Based Deep Learning Model for the Prediction of Incident Heart Failure

Shishir Rao , Yikuan Li , Rema Ramakrishnan , Abdelaali Hassaine , Dexter Canoy, John Cleland , Thomas Lukasiewicz, Gholamreza Salimi-Khorshidi, and Kazem Rahimi

Abstract—Predicting the incidence of complex chronic conditions such as heart failure is challenging. Deep learning models applied to rich electronic health records may improve prediction but remain unexplainable hampering their wider use in medical practice. We aimed to develop a deep-learning framework for accurate and yet explainable prediction of 6-month incident heart failure (HF). Using 100,071 patients from longitudinal linked electronic health records across the U.K., we applied a novel Transformer-based risk model using all community and hospital diagnoses and medications contextualized within the age and calendar year for each patient's clinical encounter. Feature importance was investigated with an ablation analysis to compare model performance when alternatively removing features and by comparing the variability of temporal representations. A post-hoc perturbation technique

was conducted to propagate the changes in the input to the outcome for feature contribution analyses. Our model achieved 0.93 area under the receiver operator curve and 0.69 area under the precision-recall curve on internal 5-fold cross validation and outperformed existing deep learning models. Ablation analysis indicated medication is important for predicting HF risk, calendar year is more important than chronological age, which was further reinforced by temporal variability analysis. Contribution analyses identified risk factors that are closely related to HF. Many of them were consistent with existing knowledge from clinical and epidemiological research but several new associations were revealed which had not been considered in expert-driven risk prediction models. In conclusion, the results highlight that our deep learning model, in addition high predictive performance, can inform data-driven risk factor identification.

Manuscript received July 21, 2021; revised December 17, 2021; accepted January 25, 2022. Date of publication February 7, 2022; date of current version July 4, 2022. The work of Yikuan Li and Kazem Rahimi was supported by British Heart Foundation (BHF) under Grant FS/PhD/21/29110. The work of Dexter Canoy and Kazem Rahimi was supported by BHF under Grant PG/18/65/33872. The work of John Cleland was supported in part by the National Institute of Health Research (NIHR) under Grant NIHRDH-NIHR130487 and in part by BHF under Grant RE/18/6/34217. The work of Thomas Lukasiewicz was supported in part by the Alan Turing Institute (ATI) under EPSRC Grant EP/N510129/1, in part by the AXA Research Fund, and in part by EU TAILOR Grant. The work of Kazem Rahimi was also supported in part by the UKRI's Global Challenges Research Fund (GCRF) under Grant ES/P0110551/1, in part by the Oxford NIHR Biomedical Research Centre, and in part by the Oxford Martin School (OMS), University of Oxford. (Shishir Rao and Yikuan Li contributed equally to this work.) (Corresponding author: Kazem Rahimi.)

Shishir Rao, Yikuan Li, Abdelaali Hassaine, and Gholamreza Salimi-Khorshidi are with the Nuffield Department of Women's and Reproductive Health, University of Oxford, OX1 2JD Oxford, U.K. (e-mail: shishir.rao@wrh.ox.ac.uk; yikuan.li@wrh.ox.ac.uk; abdelalaali.hassaine@wrh.ox.ac.uk; reza.khorshidi@wrh.ox.ac.uk).

Dexter Canoy and Kazem Rahimi are with the Nuffield Department of Women's and Reproductive Health, University of Oxford, OX1 2JD Oxford, U.K., and also with the NIHR Oxford Biomedical Research Centre, Oxford University Hospitals National Health Service (NHS) Foundation Trust, OX3 9DU Oxford, U.K. (e-mail: dexter.canoy@wrh.ox.ac.uk; kazem.rahimi@wrh.ox.ac.uk).

Rema Ramakrishnan is with the National Perinatal Epidemiology Unit, University of Oxford, OX1 2JD Oxford, U.K. (e-mail: rema.ramakrishnan@npeu.ox.ac.uk).

John Cleland is with the Robertson Centre for Biostatistics, University of Glasgow, G12 8QQ Glasgow, U.K. (e-mail: john.cleland@glasgow.ac.uk).

Thomas Lukasiewicz is with the Department of Computer Science, University of Oxford, OX1 2JD Oxford, U.K. (e-mail: thomas.lukasiewicz@cs.ox.ac.uk).

Digital Object Identifier 10.1109/JBHI.2022.3148820

Index Terms—Electronic health records, heart failure, research design.

I. INTRODUCTION

HEART failure (HF) remains a major cause of morbidity, mortality, and economic burden [1]. Despite recent evidence suggesting improvements in the quality of clinical care that patients with HF receive, and favourable trends in prognosis [2], the incidence of HF has changed little [3]. Indeed, as a consequence of population growth and ageing, the absolute burden of HF has been increasing, with incidence rates similar to the four most common causes of cancer combined [3]. These observations reinforce the need for fuller implementation of existing strategies for HF prevention and further investigations into risk factors. Several statistical models have been developed to predict risk of incident HF; however, the predictive performance of these models has been largely unsatisfactory [1].

The growing availability of comprehensive clinical datasets, such as linked electronic health records (EHR) with extensive clinical information from a large number of individuals, together with advances in machine learning, offer new opportunities for developing more robust risk-prediction models than conventional statistical approaches [4], [5]. Such data-driven approaches can also potentially discover new associations that are less dependent on expert knowledge. However, empirical evidence of robust prediction of complex chronic conditions, such as HF is limited. Prominent deep learning (DL) architectures have shown modest performance in large-scale, complex EHR datasets [6]–[8] for risk prediction of various conditions

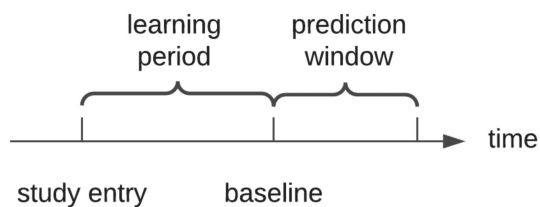


Fig. 1. Illustration of the risk prediction task. For a hypothetical patient, study entry is the start of patient medical history. And baseline marks the beginning of the “follow-up” or prediction window.

including HF. Due to their high level of abstraction, these DL models have typically had poor “explainability” or ability to demonstrate results in a language understandable by humans. This has limited their trustfulness and contribution to risk factor discoveries and wider clinical adoption [9]. Recent research has shown progress in explaining DL models in the fields of natural language and computer vision and methods such as saliency map [10] and feature perturbation [11] have gained wide popularity. However, explainable DL with rich EHR is still in its nascency; hence, tailoring known methods to improve model explainability in the medical context is crucial.

In this study, we applied a state-of-the-art sequential deep learning model for predicting incident HF using temporal, multi-modal EHR. In addition to comparing our model to state-of-the-art DL models, we investigated explainability of our Transformer model from three perspectives: ablation study, temporal variability analysis, and post-hoc perturbation analysis to deliver more explainable risk prediction. The ablation study highlighted the importance of modalities in EHR to HF risk prediction. Temporal variability analysis of the two forms of temporality we encode in the model – age in months and calendar year – demonstrated that the model found calendar year as an informative modality. The perturbation analysis lastly provided post-hoc explainability to better understand risk and preventative factors for incident HF.

II. METHODS

A. Dataset

We used U.K. Clinical Practice Research Datalink (CPRD), one of the largest de-identified longitudinal population based-EHR databases nationally representative in terms of age, sex, and ethnicity [12]. It contains primary care data from general practices (GP) since 1985, and links to secondary care and other administrative databases [4], [12] (e.g., Hospital Episode Statistics [13]).

B. Introduction to Incident HF Risk Prediction

In this work, we focused on predicting the risk of incident HF using longitudinal EHR. As shown in Fig. 1, for each patient, we used all available information before the baseline for learning to predict the risk of incident HF within a six-month window after the baseline. The incident HF was defined as the first recorded HF diagnosis code (adopted from Caliber code repository [14]) for each patient in the EHR. For patients with at least one diagnosis of HF, the baseline was defined as a randomly sampled timestamp up to six months before the first record of HF; our process ensured no HF diagnoses in the learning period. For each patient without HF, the baseline was a randomly selected

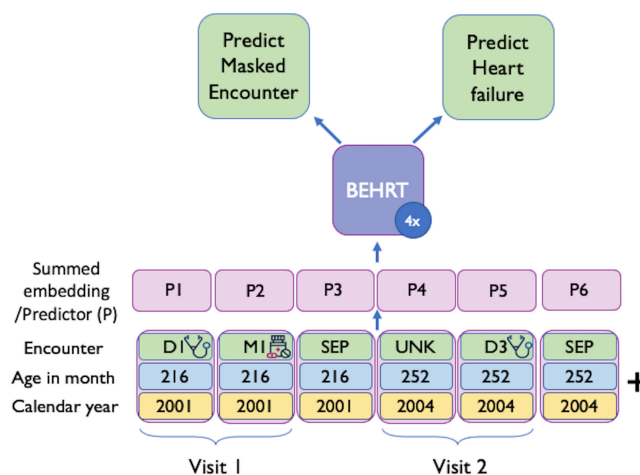


Fig. 2. Diagram of BEHRT for MLM and risk prediction. BEHRT utilizes encounters (diagnoses, medications), age in months, and calendar year annotations for EHR modelling. Predictors are represented as summed embeddings. “UNK”, “D#”/“M#”, and “SEP” represent unknown code, diagnoses/medications, and visit separations, respectively.

time stamp. The study entry was the date of the earliest available information of a patient, and it was the date of GP registration in our case.

C. Cohort Selection

GP records between 1985 and 2015 that met certain quality standards for research (as assessed by CPRD) with full data-linkage with secondary care and from patients who are at least 16 years old were included in our study. Afterwards, patients with at least five visits were included for general representation pre-training. This cohort included 1,609,024 patients in total and is referred to as dataset A. Additionally, to develop models for predicting incident HF, we selected a subset of dataset A with richer medical information. More specifically, we kept patients with i) at least 10 visits to their GP or hospital, ii) at least three years of records, and iii) and at least 10 unique codes recorded. This led to the selection of a cohort of 100,071 patients, with 13,050 (13%) cases of incident HF, henceforth referred to as dataset B.

D. Model Architecture and Details of Model Training

We extended the BEHRT architecture reported by Li *et al.* [15] inspired by the Transformer [16] to model the classification objective. In brief, BEHRT captures disease/medication associations within their temporal context to bolster predictive performance. BEHRT works robustly with large-scale, sequential data and outperforms other traditional and DL models on subsequent visit prediction tasks (Appendix section A “Details for BEHRT model”) [15].

We included encounter (disease/medication), age, and calendar year as input information. Each of the three modalities are represented by a trainable embedding matrix [15], which is a two-dimensional matrix with each instance as a vector. Each encounter and its respective age and calendar year layers are summed to form a single predictor in the model (Fig. 2).

The model was implemented using PyTorch [17]. We applied Bayesian optimization [18] for model hyperparameter tuning on the number of layers, hidden size and intermediate size. After

20 iterations of searching the parameter space, we chose the optimal hyperparameters for the model, with number of layers: 4, hidden size: 120, number of attention heads: 6, and intermediate size: 108. We pre-trained BEHRT's weights using the Masked Language Modelling [19] pre-training task using the dataset A: we randomly masked some encounters in the medical history of the patient and predicted the masked encounters. This task is unsupervised and undertaken to let the model gain a general understanding of the predictors and their temporal relationship in the longitudinal data. After pre-training the model, we implemented the model for the heart failure prediction task on dataset B.

We also replicated two state-of-the-art DL models, DeepR [7] and RETAINEX [20]. Both models have been benchmarked as superior EHR prediction models and we apply them on our dataset and compare them directly with the BEHRT model on incident HF risk prediction.

E. Ablation Study

Ablation study is a commonly used approach to provide explanations for the feature importance in modelling. In this work, we conducted ablation study to assess importance of different data modalities (diagnoses (D), medications (M), age (A), and calendar year (Y)) by alternatively removing each of them. More specifically, we devised six experiments with the following combination of modalities as input into the models: D, DA, DAY, DM, DMA, and DMAY (letters corresponding to modalities), and assessed how each of the additional modalities can influence the model performance.

F. Analysis of Temporal Variability Using Embedding Similarity

Temporal variability is intrinsic to EHR because the population, the disease pattern, and many other properties can change over time. Considering BEHRT modelled medical trajectory in the context of time-related modalities (i.e., age and calendar year), we would expect the learned representation of age and calendar year can capture such information. In this work, we applied "cosine distance", a commonly used distance metric, to measure the similarity between the representation of different ages and calendar years, respectively. The larger the difference in the learned representations over a unit of time (age/year), the more substantial the temporal variability.

G. Model Explanation Using Post-hoc Perturbation

We aimed to develop ways of quantifying encounter contributions to the prediction of incident HF, as a way of making models explainable. For this, we extended a perturbation technique inspired by work in language modelling [11] on summed, predictor embeddings to represent the disease/medication. The fundamental concept is to measure change in predictive probability after perturbing the input in order to indicate the contribution of predictors. If large perturbations of predictors minimally change outcome probability, then predictors are unimportant for prediction. However, if minimal perturbation greatly changes outcome probability, the respective predictors are highly important. In this work, we proposed an asymmetric loss function to prioritize encounters that enhanced HF/non-HF predictions. HF and non-HF represent HF positive and negative, respectively. Algorithm 1 demonstrates the general algorithm for conducting perturbation analyses and more details can be found in Appendix section "Details of perturbation methods".

Algorithm 1 Perturbation Algorithm.

M : Model, N : Number of encounters, X : input encounter embeddings $X \in \mathcal{R}^{N \times K}$, \tilde{X} : perturbed input embeddings, Y : indicator of HF patient, L : loss function, S : contribution score.

PERTURBATION ()

Initialise $\epsilon \leftarrow [\epsilon_1, \epsilon_2, \dots, \epsilon_N]$ and $\epsilon_{1i} \sim \mathcal{N}(0, \sigma_i^2 I)$

While (not converged)

$\tilde{X} \leftarrow X + \epsilon$

$O \leftarrow M(X), \tilde{O} \leftarrow M(\tilde{X})$

$Loss \leftarrow L(O, \tilde{O}, Y, \tilde{X})$

update ϵ

$S \leftarrow transform(\epsilon)$

Return S

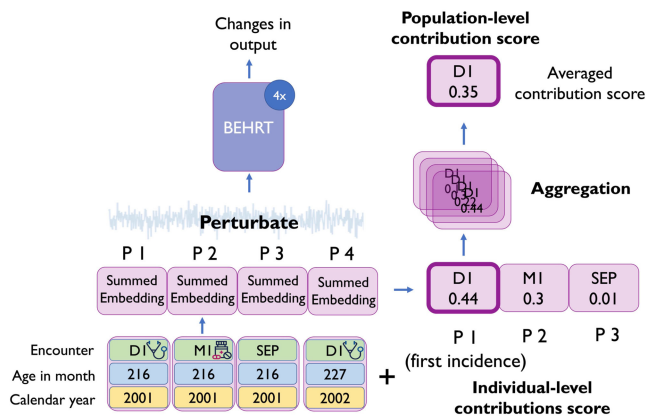


Fig. 3. Illustration for conducting population-level encounter contribution. We use trainable noise to understand contribution of each predictor to output prediction. We aggregate individual-level contribution scores for a disease/medication (e.g., disease code: D1) to form population-level metrics of the same.

The perturbation approach is a local surrogate method that can only quantify the contribution of encounters at the individual level. However, by aggregating the individual-level encounter (i.e., diagnosis or medication) contribution across the population, we can analyze an encounter's contribution at the population level. In this study, we focused on the contribution of the first incidence (if repetition was applicable) of a disease/medication in the context of age and calendar year for each patient (as shown in Fig. 3).

H. Model Evaluation and Perturbation Analysis

60%, 20%, and 20% of the patients in Dataset B were selected as training, tuning, and held-out cohorts, respectively. The training and tuning cohorts were used for hyper-parameter tuning, and the combined tuning and held-out cohort was used for conducting perturbation analysis. Furthermore, to evaluate the model performance in a more robust approach, we applied five-fold cross-validation on Dataset B and reported model performance using area under the receiver operating characteristic (AUROC) and precision-recall curve (AUPRC) with 95% confidence interval (CI) over five folds.

The perturbation analysis investigated the association between an encounter code and HF using relative contribution (RC) with 95% CI [21]. It is calculated by dividing average contribution (Fig. 3) of the encounter code in HF patients by

TABLE I
CHARACTERISTICS OF PATIENTS IN THE TRAINING, TEST, AND VALIDATION SET

	Training	Test	Validation
Total number of patients (%)	60,043 (100)	20,014 (100)	20,014 (100)
Number of incident cases of heart failure (%)	7,853 (13.1)	2,621 (13.1)	2,576 (12.9)
Women (%)	35,094 (58.4)	11,634 (58.1)	11,603 (58.0)
Men (%)	24,949 (41.6)	8,380 (41.9)	8,411 (42.0)
Median follow-up duration (year)	9	9	10
Median age (year); Interquartile Range	70; (59,79)	70; (59,79)	70; (59,79)
Diabetes Mellitus (%)	12,348 (20.5)	4,130 (20.6)	4,128 (20.6)
Hypertension (%)	39,427 (65.6)	13,096 (65.4)	13,237 (66.1)
Rheumatoid arthritis (%)	1,936 (3.2)	673 (3.4)	679 (3.4)
Atrial fibrillation and flutter (%)	15,826 (26.3)	5,179 (25.9)	5,252 (26.2)
Myocardial infarction (%)	5,588 (9.3)	1,851 (9.2)	1,839 (9.2)
Chronic obstructive pulmonary disease (COPD) (%)	8,364 (13.9)	2,770 (13.8)	2,763 (13.8)
Ischemic stroke (%)	3,022 (5.0)	1,037 (5.2)	1,065 (5.3)

average contribution of the encounter code in non-HF patients. $RC > 1.0$ and < 1.0 implies the encounter code is positively associated with HF and non-HF respectively.

We applied the perturbation analysis on patients with relatively confident predictions (predictive probability larger than 0.8 and smaller than 0.2) and focused on encounter codes that were trained sufficiently (with at least 5% prevalence in the cohorts). Additionally, we specifically investigated some established risk factors for HF [22], [23] and then looked into other encounters that were positively or negatively associated with HF. To understand differential contribution to HF by age and calendar year, we conducted contribution analyses stratified by age groups: (50–60], (60–65], (65–70], (70–75], (75–80] and calendar year groups: [1990–1995], (1995–2000], (2000–2005], (2005–2010] when clinical events were first recorded. For stratification, we clarify the “(” and “[/|” represent the exclusion and the inclusion, respectively.

III. RESULTS

A. Dataset Preparation

Our dataset included diagnostic codes (299 Caliber codes [24]) and medications (426 codes) as well as patient age in months and calendar year. Both disease and medication codes with unknown mapping were mapped to an “UNKNOWN” category. Of 100,071 patients for incident HF prediction (Dataset B), the median age in years at baseline was 70; 1st and 3rd quartile: (59, 79), 52% were men, 65% had history of hypertension, 9% a prior myocardial infarction, and 5.1% an ischemic stroke. The median follow-up duration (date of first record to baseline) was 9 years. More details for training, test, and external validation set, as well as code and HF phenotyping processes can be found in Table I and Appendix section “Details of disease and medication phenotyping”.

TABLE II
MODEL PERFORMANCE

Model	AUROC (95% CI)	AUPRC (95% CI)
BEHRT	0.93 (0.926, 0.934)	0.69 (0.667, 0.713)
RETAINEX	0.90 (0.893, 0.901)	0.62 (0.596, 0.636)
DeepR	0.91 (0.901, 0.913)	0.61 (0.577, 0.633)

CI: Confidence Interval; bold model is best performing.

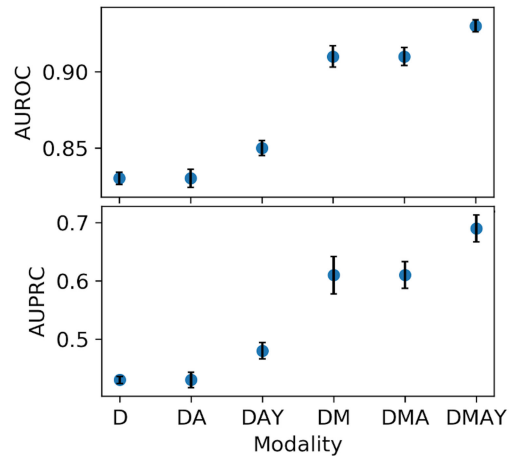


Fig. 4. Ablation study with inclusion of diagnoses (D), medications (M), age (A), year (Y). Model performance metrics are AUROC and AUPRC with 95% confidence intervals.

B. Model Performance

The BEHRT model, with all four modalities (DMAY), showed best performance on five-fold internal cross-validation with an AUROC of 0.93 (0.926, 0.934) and AUPRC of 0.69 (0.667, 0.713) (Table II). BEHRT showed noticeable improvement in predictive ability – approximately 2%/7% absolute improvement in AUROC/AUPRC compared to RETAINEX and DeepR.

C. Analysis of Ablation Study

Fig. 4 shows the results of the ablation study. It demonstrates that the BEHRT model with the inclusion of medication data outperforms the model with just diagnoses in both AUROC and AUPRC. Furthermore, utilisation of calendar year achieves substantial improvements in terms of AUROC/ AUPRC than the model with age solely, thus, indicating calendar year to be more informative for contextualizing predictors than chronological age.

D. Analysis of Temporal Variability

Fig. 5(a) and 5(b) show two cosine similarity [25] matrices for each pair of instances in embeddings of calendar year and age. We chose four diseases with high occurrence in the dataset to ensure these embeddings were well-trained.

Calendar year showed substantial dissimilarity across years (from 0 to 0.8). By contrast, the dissimilarity as a consequence of variation in age in months was less pronounced (from 0 to 0.3). In other words, representation of diseases among individuals was more sensitive to variation in year in which they were recorded

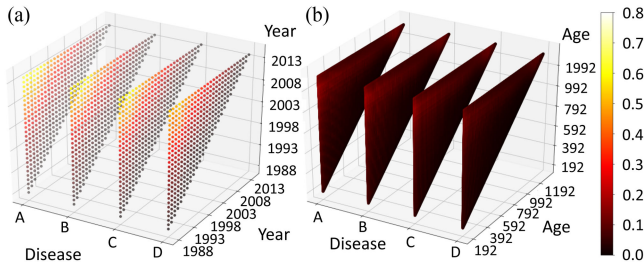


Fig. 5. Age and year embedding analysis. We show cosine similarity measurement for summed embedding of diseases within different year (a) and age groups (b). A: depression, B: peripheral arterial disease, C: anxiety disorders, D: hypo or hyperthyroidism; age axis represents age in months from 16 to 100 years in months; year axis represents year from 1988 to 2014. Lighter colours indicate higher dissimilarity and darker colours, lower dissimilarity.

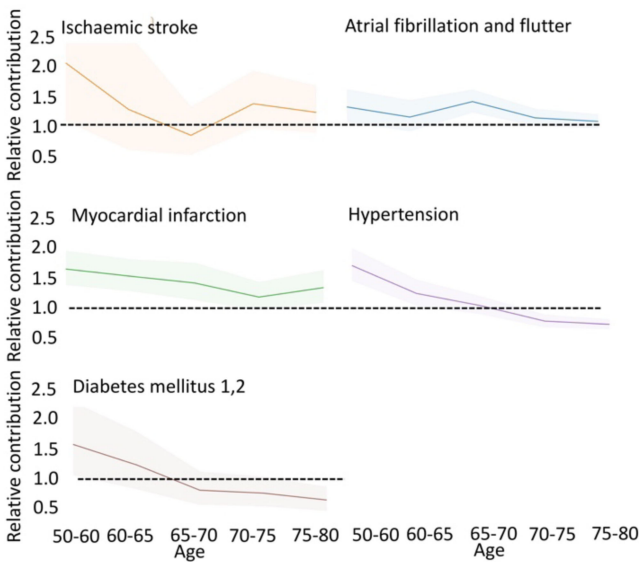


Fig. 6. Age-stratified RC analyses for established risk factors. X and y axes represent age groups and RC (mean; 95% CI). The black dotted line denotes 1.0 RC.

than to the age of the patient. This suggests that calendar year shows stronger temporal variability, and the latent information of diseases in the context of calendar year can differ across different years, and hence, was more informative for the incident HF prediction than chronological age.

E. Contribution Analysis

We first investigated if BEHRT could naturally capture established risk factors [23] through the proposed RC metric. We derived RC for diseases: hypertension, atrial fibrillation and flutter, myocardial infarction, diabetes (I,II), ischaemic stroke. For these HF risk factors, the average RC was >1 both in the general (Table III) as well as age-stratified analyses (Fig. 6/Table IV) demonstrating that the BEHRT (DMAY) model associated these diseases with HF. Furthermore, the RC was generally higher for those aged 50–60 years and lower in older ages, implying little contribution of the risk factors individually to HF in older ages (consistent with evidence from past epidemiological studies [26], [27]).

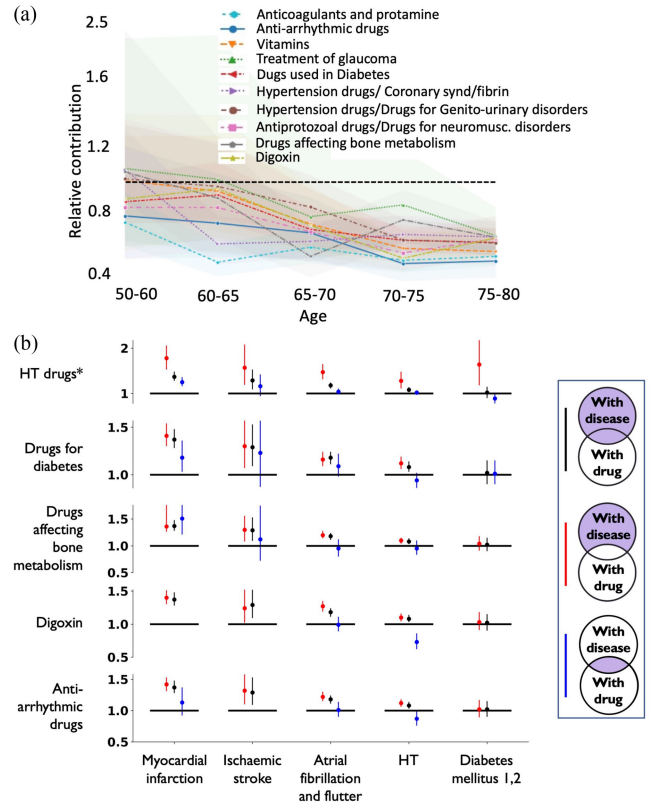


Fig. 7. Lowest age-stratified RC for medications and contextuality of medications and risk factors analysis. (a), Lowest 10 RC of medications. x and y axis represent age groups and RC (mean; 95% CI). The black dotted line denotes 1.0 RC. (b), RC (mean; 95% CI) for established risk factors stratified by treatment status. Black forest plot: general population of people with disease (column); red forest plot: those with disease (column) not treated with medication (row); blue forest plot: those with disease (column) and treated with medication (row). Some graphs fail to have treated/untreated forest plot due to insufficient size in subgroup (eg diabetes; drugs for diabetes).

Age-stratified RC was <1 for hypertension and diabetes in some age groups (Fig. 6), which was unexpected given that these are established risk factors for HF. We hypothesized that treatments might be contextually associated with these diseases in older age patients and thus bias the association of these diseases.

Taking hypertension as an example: if hypertension is commonly treated with antihypertensives, and antihypertensives are associated with non-HF, our perturbation analyses would yield that both antihypertensives and by contextualisation, hypertension are associated with non-HF. To test our hypothesis, we first investigated disease prevalence and found that 73% of patients >65 with hypertension are treated with antihypertensives while 70% of all diabetic patients are treated with medications for diabetes. Now understanding that these diseases frequently contextualise with their respective treatments in older ages, we investigated if these treatments associate strongly with non-HF (RC <1). We found that indeed medications such as antihypertensives, digoxin, and drugs for diabetes, all established treatments of validated risk factors, were largely associated with non-HF (Fig. 7(a)/Table VIII) [28].

While Fig. 7(a) demonstrates that treatments of known risk factors are associated with non-HF, to better disentangle the

relationship between risk and treated risk, we conducted stratified analyses on untreated and treated patients. For example, to understand the association of treated and untreated hypertension to HF as opposed to general hypertension, we derive RC for diseases in the subgroup of patients with hypertension treated and untreated with antihypertensives. In Fig. 7(b), with respect to general population (black lines), in treated patients (blue lines), there is a general decrease in the RC for each disease (16 of 19 cases) with lesser association to HF. Also, while RCs of risk factors were generally attenuated in treated subgroups compared to untreated subgroups, RCs were most attenuated in subgroups treated with antihypertensives, digoxin, and medications for diabetes. In some cases, RC was not calculated for the untreated or treated subgroup because it failed to have enough patients for substantive calculation. Through these experiments, we show that BEHRT naturally captures untreated risk factors associate strongly with HF while treated risk factors are mitigated in risk due to treatment and thus, appropriately associate lesser with HF.

After validating our model’s ability to capture known risk factors and the interplay between risk and treatments of risk, we investigated other novel risk factors independently derived by the model. We found diseases like bacterial diseases, lower respiratory tract infections, myocardial infarction and pleural effusion, and medications such as “corticosteroid /antibacterial drugs”, “bronchodilators” and “acne and rosacea drugs” were all positively associated with HF (Tables V and VI). Furthermore, the age-stratified RC analysis for the ten (i.e., highest RC) diagnoses (Fig. 8(a)) and medications (Fig. 8(b)) most strongly associated with HF showed consistent pattern as the analysis of the validated risk factors implying limited discriminatory contribution of these predictors individually in older people. However, for some predictors (e.g., left bundle branch block), CIs were too wide to allow firm conclusions about any differential RC by age (Tables VII and VIII).

Additionally, analogous to the relationship between contextualised treatments of risks and risks themselves, many of the medications that showed a high RC to HF (Fig. 8(b)) are contextualised treatments for paired diagnoses (Fig. 8(a)). This implies that the model identified diagnosis and treatment pairs that were at least contemporaneous and often causally associated. For example, dermatitis may be treated with corticosteroids, which may be linked to cardiovascular risk [29] as may be depression, and therefore its treatments [30]. And lower respiratory tract infection, asthma and chronic obstructive lung disease are often treated with “cough preparations”, “bronchodilators” and “antibacterial drugs” [31], [32]. This could signal delayed diagnosis of HF due to misattribution of HF symptoms to respiratory diseases [32]. Or it might be that direct effects of drugs such as non-steroidal anti-inflammatory drugs (NSAIDs) are at least in part responsible for causing HF [33].

We saw in both ablation and temporal variability analyses (Results section D) that calendar year substantially effects prediction, hence, we wish to further investigate the differential RC by calendar year. For presentation, we have analysed two case studies of <1 RC medications depicted in Fig. 7(a): digoxin and treatments for glaucoma.

Fig. 9(a) demonstrates the two medications stratified by year. While digoxin was consistently associated with non-HF across year groups, the results were more heterogenous for treatment of glaucoma. We hypothesized that while the prescription medicine

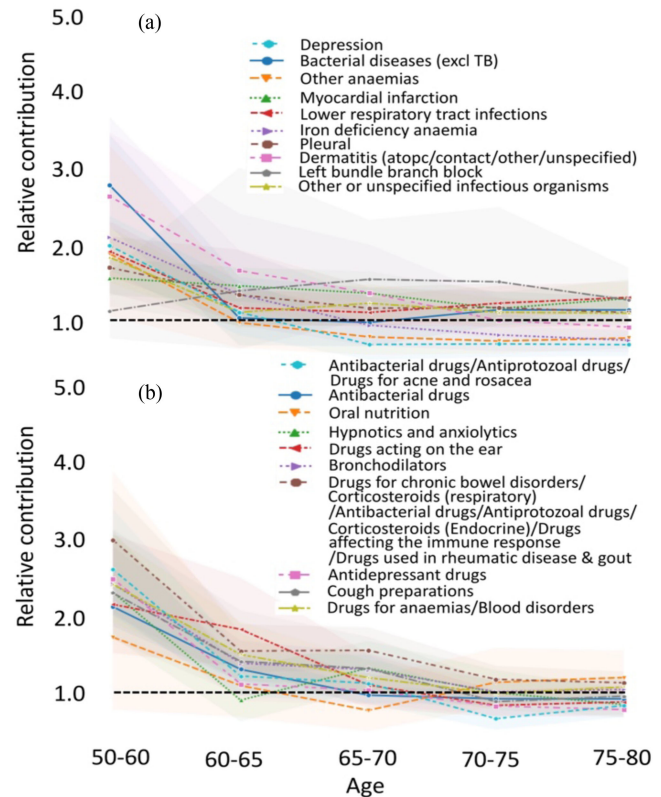


Fig. 8. Age-stratified RC analyses for top 10 diseases (a) and medications (b) identified by model. x and y axis represent age groups and RC (mean; 95% CI), respectively. RC equals to 1 (black dotted line) implies equal contribution to both HF and non-HF predictions.

had constant BNF coding across the years, the underlying drug composition might have changed around the year, 2000. So, we analyse the number of times different glaucoma drugs and digoxin were first prescribed in patients from Dataset A between the years 1990 and 2010 in Fig. 9(b) (not counting repeat prescriptions).

We found that throughout the 1990’s, timolol (beta blocker) was a common topical treatment for glaucoma [34]. With the introduction of new medications in the 2000’s, the use of ophthalmic timolol started to decline [35].

Our RC analysis in Fig. 9(a) shows that BEHRT implicitly demonstrates the ability to capture this change in prescription of the particular glaucoma treatment. Specifically, BEHRT identifies that the prevalent treatment, timolol, [36]prescribed prior to 2000 highly associated with HF ($RC > 1$). Timolol has known cardiovascular side-effects such as bradycardia with potential to exacerbate HF [36]. Following 2000, BEHRT identifies that the prevalent glaucoma treatment – namely, prostaglandin analogues – has <1 RC. Prostaglandins and analogues such as prostaglandin I_2 [37], [38] and others [39] have known vasodilating properties with the potential of reducing cardiovascular risk, although large-scale randomized trials to investigate preventative effects are currently lacking [37], [38].

In stark contrast, prescription of digoxin wanes following 2005 (Fig. 9(b)). However, RC for this positive inotropic drug remains stable in any year strata further lending support to the hypothesis that digoxin could play a role in prevention of HF.

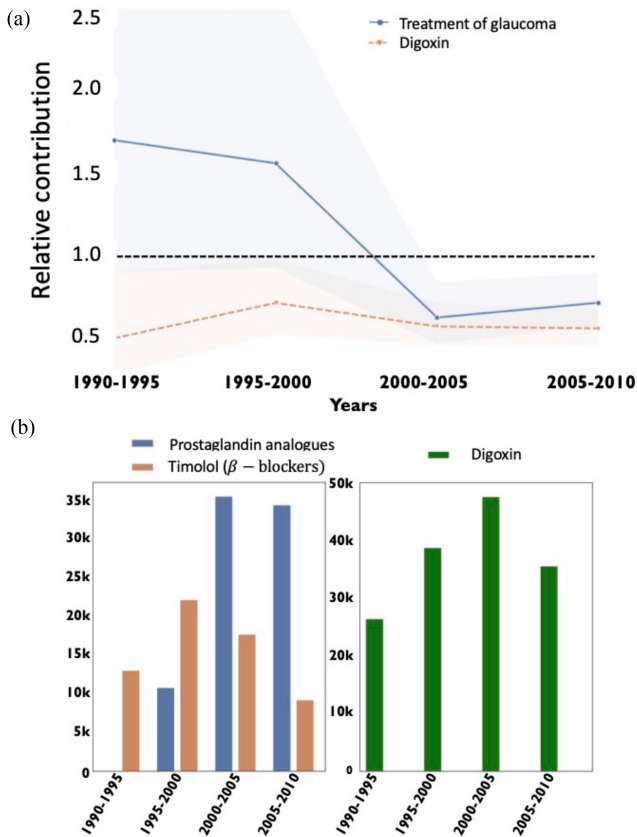


Fig. 9. Year-stratified RC of medications. (a), RC (mean; 95% CI) of three medications to HF prediction stratified by year. x and y axes represent year group and RC respectively; to the black line denotes the 1.0 RC. (b), frequency of drugs by components in different year groups in Dataset (a). X/y represent the year group/counts of first-time drug (component) prescriptions to patients respectively. Individual drug components are represented with bars in different colour.

IV. DISCUSSION

By incorporating large-scale routine EHR data, we developed and validated a model for incident HF prediction. It showed superior performance compared to state-of-the-art models, RETAINEX and DeepR. Compared to age, calendar year substantially improved predictive performance while contribution analysis demonstrated that diseases and medications were strong predictors. Our explainable framework also confirmed the relative importance of established risk factors [23] and provided insights into medications that might negatively/positively contribute to HF prediction.

Our work has several novelties. We included many potentially predictive variables not previously included in epidemiological studies. Although age is usually incorporated as a risk factor for risk prediction [6], [20], our ablation analysis found that incorporating calendar year provided additional and stronger information for accurate prediction of incident HF. A potential explanation was the temporal variability caused by the changes in medicine (e.g., the change in disease pattern, policy, or the availability of and the use of treatments) was substantial in the changes of the calendar year. Therefore, the calendar year was an expressive feature that can capture such temporal changes and reflected by the dissimilarity of year representations as shown

in the temporal variability analysis. It is further supported by our perturbation analysis stratified by year. This showed that medications over different years made quantitatively different contributions to disease prediction, which would be missed if temporal context was not included in the models. Changes in such predictors over time or more subtle changes in disease patterns for instance due to advances in technologies leading to more accurate and frequent diagnosis are well known to clinicians. BEHRT enables incorporation of such information for better prediction.

With regards to encounter contribution, we discovered that BEHRT can independently capture the interplay between risk and treated risk. While BEHRT generally captured risk factors appropriately, in some age groups, we saw some risk factors associating with non-HF. A cursory analysis might lead to false conclusions that for instance, hypertension decreases the risk of HF in older age. However, this conclusion is biased by indication; the correct interpretation is that the medication serves as a proxy for hypertension, which has a strong effect on HF incidence. Our analysis of risk and treated risk generally demonstrates while risk factors associate with HF, treated risk attenuates that association – conclusions consistent with medical understanding of HF-risk.

Additionally, through both age and year stratified analysis, BEHRT demonstrates medications with potentially preventative effects on HF. In the case of digoxin and prostaglandin analogues, the stable $RC < 1$ in both age and year stratification signals potentially preventative effects of these drugs. However, as with standard statistical models, a causal interpretation should be made with great caution. Rather, our method generates hypotheses, which depending on the totality of evidence from this work and other sources, should provide the impetus for additional confirmatory studies.

Our study has some limitations. First, the phenotyping method for diagnoses maps codes to 299 disease categories[40] losing information in the original granularity of the disease encoding and potentially biased by an expert’s preferences. Second, during cohort selection, we kept patients with sufficient records to make robust predictions potentially compromising model generalizability for prediction in low-risk groups who have fewer clinical encounters. Additionally, model transferability needs further investigation since only CPRD is used in our study.

V. CONCLUSION

We developed a superior model for incident HF prediction using routine EHR providing a promising avenue for research into prediction of other complex conditions. Incorporating BEHRT into routine EHR could alert clinicians to those at risk for more targeted preventive care or recruitment into clinical trials. In addition, we highlight a data-driven approach for identification of potential risk factors that generate new hypotheses requiring causal exploration. We note there are several medications which contribute negatively to HF prediction. Not only are many used to treat established risk factors of HF, but others have not been tested for such an indication and might provide a starting point for drug repurposing studies. The model and analysis could be applied to more deeply phenotyped populations for discovery of new disease mechanisms and patterns in other complex conditions.

TABLE III

RELATIVE CONTRIBUTION SCORES FOR MEDICALLY VALIDATED RISK FACTORS. IN THE COLUMNS, WE HAVE HEART FAILURE (HF) PATIENT’S AVERAGE CONTRIBUTION SCORES, NON-HEART FAILURE (NON-HF) PATIENTS AVERAGE CONTRIBUTION SCORES, RELEVANT STANDARD DEVIATION MEASURES, AND RELATIVE CONTRIBUTION AND 95% CONFIDENCE INTERVAL FOR THESE RISK FACTORS

Disease	RC	CI	
		LB	UB
Myocardial infarction	1.37	1.28	1.48
Ischaemic stroke	1.29	1.09	1.53
Atrial fibrillation and flutter	1.18	1.11	1.24
Hypertension	1.08	1.03	1.14
Diabetes mellitus 1,2	1.02	0.9	1.15

TABLE IV

AGE-STRATIFIED, RELATIVE CONTRIBUTION AND 95% CONFIDENCE INTERVALS FOR MEDICALLY VALIDATED RISK FACTORS. IN THE COLUMNS, WE HAVE RC AND CI FOR FIVE AGE CATEGORIES

Disease	50-60	60-65	65-70	70-75	75-80
Atrial fibrillation and flutter	1.23; (1.02, 1.47)	1.08; (0.89, 1.32)	1.3; (1.15, 1.47)	1.07; (0.96, 1.2)	1.02; (0.93, 1.12)
Diabetes mellitus 1,2	1.48; (1.05, 2.07)	1.19; (0.86, 1.65)	0.84; (0.64, 1.1)	0.8; (0.62, 1.03)	0.7; (0.56, 0.89)
Hypertension	1.59; (1.39, 1.83)	1.21; (1.05, 1.4)	1.04; (0.93, 1.17)	0.83; (0.75, 0.92)	0.79; (0.73, 0.86)
Ischaemic stroke	1.86; (1.05, 3.3)	1.21; (0.64, 2.27)	0.84; (0.57, 1.24)	1.29; (0.95, 1.75)	1.17; (0.88, 1.54)
Myocardial infarction	1.55; (1.33, 1.8)	1.45; (1.25, 1.68)	1.36; (1.13, 1.63)	1.16; (0.99, 1.37)	1.29; (1.09, 1.53)

APPENDIX

A. Details for BEHRT Model

In NLP literature and BERT, [19] words in sentences are considered “tokens” and sentences are separated from one another with a separation element. Similarly, we conceptualized medical events such as diagnoses and medications in a doctor/hospital visit as encounters (or tokens) and separate visits by a separation element (“SEP”). Similar to the original BERT model, we implemented an annotation ordering the sequential medical history data. Furthermore, we added layers of information that involve age and calendar year of each encounter. Thus, the total input comprised of three layers of information for each and every encounter: the encounter itself (diagnoses and/or medications), age, and calendar year.

B. Details of Perturbation Methods

Additionally to Guan *et al.*’s method [11], we developed an asymmetric loss function to prioritize learning perturbations in addition to the information entropy-based loss term. Shown in the equations below is the full description of the loss function

TABLE V

RELATIVE CONTRIBUTION SCORES FOR DISEASES THAT OCCURRED IN AT LEAST 5% OF THE POPULATION

Disease	RC	CI	
		LB	UB
Dermatitis (atopic/contact/other/unspecified)	1.76	1.56	1.98
Bacterial diseases (excluding tuberculosis)	1.56	1.42	1.71
Other or unspecified infectious organisms	1.47	1.37	1.58
Lower respiratory tract infections	1.47	1.37	1.57
Depression	1.45	1.29	1.63
Iron deficiency anaemia	1.4	1.22	1.61
Pleural effusion	1.38	1.24	1.53
Myocardial infarction	1.37	1.28	1.48
Left bundle branch block	1.34	1.13	1.60
Other anaemias	1.33	1.17	1.50
Hypo or hyperthyroidism	1.26	1.13	1.41
Atrial fibrillation and flutter	1.18	1.11	1.24
Acute kidney injury	1.15	1.01	1.31
Hypertension	1.08	1.03	1.14
Hearing loss	1.03	0.86	1.22
Urinary tract infections	1.00	0.88	1.12
Enthesopathies & synovial disorders	0.9	0.79	1.02
Stroke not otherwise specified (NOS)	0.89	0.79	0.99
Chronic obstructive pulmonary disease (COPD)	0.87	0.80	0.95
Cataract	0.84	0.75	0.94
Hyperplasia of prostate	0.82	0.72	0.95
Stable angina	0.81	0.75	0.88
Osteoarthritis (excluding spine)	0.75	0.68	0.84
Diabetic ophthalmic complications	0.7	0.62	0.79

conducted over one patient’s medical history.

$$alpha(y, x, s) = \begin{cases} \beta_1, & \text{if } y = 1, M(\tilde{x}) - s \geq 0 \\ \beta_2, & \text{if } y = 0, M(\tilde{x}) - s \leq 0 \\ \beta_3, & \text{otherwise} \end{cases} \quad (1)$$

$$L(\sigma) = alpha(y, x, s) \times E_{\epsilon} \| (M(x) - s) \|^2 \lambda \sum_{i=1}^n H(\tilde{x} | s) |_{\epsilon_i \sim \mathcal{N}(0, \sigma_i^2 I)} \quad (2)$$

where \tilde{x} represents perturbed input encounter embeddings (original represented by x), n represents number of encounters in this patient’s medical history, s represents output state of original input (without perturbation), $M(\tilde{x})$ is the output state of perturbed input, β_1 and β_2 are weight hyperparameters ($\beta_1 < \beta_2$); if $\beta_1 = \beta_2$, then loss function is symmetric, y is heart failure label, $E_{\epsilon} \| M(\tilde{x}) - s \|^2$ represents mean squared error described in Guan *et al.* [11], $alpha(y, \tilde{x}, s)$ means asymmetric weight function, $\lambda \sum_{i=1}^n H(\tilde{x} | s) |_{\epsilon_i \sim \mathcal{N}(0, \sigma_i^2 I)}$ information entropy based loss function described in Guan *et al.* [11].

To describe (1) and (2) in words, for heart failure patients, we prioritize perturbations that increases the outcome probability than those that decreases it; and we prioritize the opposite for non-heart failure patients. To do so, we penalize with the $alpha(y, x, s)$ constant. Asymmetric losses are often used in scenarios where an error in one direction (perhaps positive) is more costly than an error in the opposite direction[12], [24].

This perturbation method delivers learned $\epsilon = [\epsilon_1, \epsilon_2, \dots, \epsilon_n]$ with ϵ_1 per predictor i – the trained, allowable variance for the predictor, with maximum variance defined by a user defined hyperparameter (set to 0.5 in our work). To assess contribution of predictor, we transform ϵ_1 to $0.5 - \epsilon_1$ to reflect the inverse relationship: the lower the ϵ_1 , the higher contribution to heart failure prediction and vice-versa. This is shown as the

TABLE VI
RELATIVE CONTRIBUTION SCORES FOR MEDICATIONS THAT OCCURRED IN AT LEAST 5% OF THE POPULATION

Medication	RC	CI	
		LB	UB
Chronic bowel disorders/ corticosteroids (respiratory) etc	1.71	1.59	1.84
Anaemias and other blood disorders	1.53	1.42	1.66
Bronchodilators	1.52	1.43	1.63
Cough preparations	1.45	1.28	1.64
Oral nutrition	1.41	1.68	1.68
Antibacterial drugs/ antiprotozoal drugs/ acne and rosacea	1.41	1.25	1.58
Hypnotics and anxiolytics	1.39	1.22	1.57
Antidepressant drugs	1.38	1.28	1.50
Drugs acting on the ear	1.36	1.2	1.54
Antibacterial drugs	1.32	1.26	1.39
Antibacterial drugs/ acne and rosacea	1.31	1.2	1.44
Drugs used in nausea and vertigo	1.27	1.16	1.40
Soft-tissue disorders and topical pain relief	1.24	1.13	1.36
Drugs acting on the oropharynx	1.22	1.07	1.41
Hypnotics and anxiolytics/ general anaesthesia	1.22	1.06	1.40
Analgesics/ analgesics	1.21	1.13	1.29
Corticosteroids (respiratory)	1.20	1.09	1.32
Drugs used in rheumatic diseases and gout	1.19	1.11	1.28
Anti-infective eye preparations	1.17	1.05	1.05
Laxatives	1.16	1.08	1.25
Anti-infective skin preparations	1.14	1.01	1.29
Vaccines and antisera	1.13	1.06	1.20
Antihistamine, hyposensitivity and allergic emergent	1.11	1.01	1.22
Dyspepsia and gastro-oesophageal reflux disease	1.09	0.98	1.20
Antiplatelet drugs	1.08	1.02	1.14
Emollient and barrier preparations	1.08	1.00	1.16
Sex hormones	1.07	0.92	1.25
Nitrates, calcium-channel blocker and other antianginal drugs	1.06	1.01	1.12
Hypnotics and anxiolytics/ antiepileptic drugs/ etc	1.06	0.94	1.20
Miscellaneous ophthalmic preparations	1.05	0.92	1.19
Topical corticosteroids	1.05	0.98	1.11
Local preparations for anal and rectal disorders	1.03	0.9	1.17
Corticosteroids and other anti-inflammatory preparations	1.02	0.88	1.17
Drugs acting on the nose	1.01	0.93	1.11
Acute diarrhoea	1.01	0.88	1.15
Topical corticosteroids/ anti-infective skin preparations	0.99	0.86	1.13
Analgesics	0.98	0.93	1.03
Hypertension and heart failure	0.95	0.91	1.00
Drugs used in psychoses and Related disorders/ drugs used in nausea and vertigo	0.95	0.82	1.11
Diuretics	0.93	0.88	0.98
Acute diarrhoea/ cough preparations/ analgesics	0.92	0.82	1.03
Bronchodilators/ corticosteroids (respiratory)	0.92	0.83	1.01
Lipid-regulating drugs	0.91	0.87	0.97
Analgesics/ drugs used in rheumatic diseases and gout	0.91	0.79	1.05
Antidepressant drugs/ analgesics/ analgesics	0.91	0.82	1.01
Beta-adrenoceptor blocking drugs	0.87	0.79	0.95
Antisecretory drugs and mucosal protectants	0.85	0.81	0.90
Diuretics/ minerals	0.85	0.77	0.93
Drugs for Genito-urinary disorders	0.81	0.73	0.91
Thyroid and antithyroid drugs	0.81	0.73	0.9
Treatment of glaucoma	0.77	0.64	0.92
Hypertension and heart failure/ drugs for genito-urinary disorders	0.71	0.65	0.78
Drugs used in diabetes	0.71	0.66	0.77
Vitamins	0.7	0.63	0.79
Drugs affecting bone metabolism	0.7	0.78	0.78
Antiprotozoal drugs/ drugs used in neuromuscular disorders	0.67	0.58	0.77
Beta-adrenoceptor blocking drugs/ hypertension drugs/etc	0.67	0.61	0.73
Positive inotropic drugs	0.61	0.54	0.69
Anti-arrhythmic drugs	0.56	0.48	0.65
Anticoagulants and protamine	0.53	0.48	0.57

TABLE VII
AGE-STRATIFIED, RELATIVE CONTRIBUTION AND 95% CONFIDENCE INTERVALS FOR DISEASES THAT OCCURRED IN AT LEAST 5% OF THE TEST AND VALIDATION DATASET. IN THE COLUMNS, WE HAVE RC AND CI FOR FIVE AGE CATEGORIES WITH LOWER BOUND AND UPPER BOUND

Disease	50-60	60-65	65-70	70-75	75-80
Bacterial diseases (excluding tuberculosis)	2.71; (2.2, 3.35)	1.04; (0.7, 1.56)	0.99; (0.69, 1.42)	1.15; (0.95, 1.38)	1.14; (1, 1.3)
Depression	1.95; (1.64,2.3 3)	1.12; (0.82,1 .51)	0.71; (0.52,0 .98)	0.72; (0.55,0 .94)	0.71; (0.56,0. 89)
Dermatitis (atopic/contact/ other/unspecified)	2.58; (1.96, 3.38)	1.64; (1.14,2 .36)	1.36; (0.94,1 .97)	1.01; (0.79,1 .3)	0.93; (0.77,1. 12)
Left bundle branch block	1.13; (0.79, 1.62)	1.39; (0.65,2 .93)	1.53; (1.03,2 .44)	1.5; (0.92,2 .44)	1.27; (0.96,1. 67)
Lower respiratory tract infections	1.88; (1.57, 2.26)	1.17; (0.97,1 .41)	1.11; (0.93,1 .33)	1.23; (1.04,1 .45)	1.3; (1.13,1. 5)
Myocardial infarction	1.55; (1.33, 1.8)	1.45; (1.25,1 .68)	1.36; (1.13,1 .63)	1.16; (0.99,1 .37)	1.29; (1.09,1. 53)
Other anaemias	1.85; (1.07, 3.21)	0.98; (0.67,1 .45)	0.81; (0.58,1 .12)	0.75; (0.59,0 .96)	0.79; (0.66,0. 94)
Other or unspecified infectious organisms	1.81; (1.49, 2.18)	1.11; (0.91,1 .36)	1.23; (1.03,1 .47)	1.11; (0.96,1 .3)	1.1; (0.97,1. 25)
Pleural effusion	1.68; (1.36, 2.07)	1.33; (0.94,1 .9)	1.17; (0.95,1 .43)	1.18; (0.94,1 .48)	1.11; (0.94,1. 31)
Iron deficiency anaemia	2.06; (1.19, 3.57)	1.34; (0.95,1 .89)	0.95; (0.69,1 .31)	0.83; (0.62,1 .1)	0.76; (0.61,0. 94)

transform() function in Algorithm 1. As seen in Fig. 3, we first establish patient-level contribution, or $0.5 - \epsilon 1$ for a particular encounter (disease/medication) in time.

C. Details of Disease and Medication Phenotyping

In our study, we used all diagnostic codes and medications at each encounter as well as patient age in months and calendar year. Encounters represent each individual's time-stamped recording of a diagnosis or medication. For medication specifically, we included all available prescription records as new encounters in the dataset. Because of how CPRD records medications in six-week increments, medications make up the largest number of encounters.

In the U.K., primary care and secondary care use different coding systems for diagnosis (i.e., Read Code [41] in GP and 10th revision of International Statistical Classification of Diseases and Related Health Problems [ICD-10] Code [42] in Hospital Episode Statistics). We mapped these codes using a dictionary given by CPRD. This led to 56,624 unique codes, which were then mapped to 299 clinically meaningful disease categories, using Caliber, a previously published and clinically-validated phenotyping method. [14] Medication codes were classified using the British National Formulary hierarchical coding format. [43] Medication data in CPRD indicate medication prescription as opposed to medication retrieval or dispensation. We used codes at the section level prevalent in the population leading to 426 unique medication group codes.

TABLE VIII

AGE-STRATIFIED, RELATIVE CONTRIBUTION, 95% CONFIDENCE INTERVALS FOR MEDICATIONS THAT OCCURRED IN AT LEAST 5% OF THE TEST, VALIDATION DATASET. IN THE COLUMNS, WE HAVE RC, CI FOR FIVE AGE CATEGORIES WITH LOWER BOUND, UPPER BOUND

Medications	50-60	60-65	65-70	70-75	75-80
Antibacterial drugs	2.07; (1.82, 2.36)	1.26; (1.1, 1.44)	0.92; (0.81, 1.04)	0.87; (0.79, 0.96)	0.87; (0.81, 0.95)
Anaemias, other blood disorders	2.36; (1.87, 2.99)	1.45; (1.12, 1.88)	1.15; (0.96, 1.38)	0.94; (0.8, 1.11)	1.03; (0.9, 1.17)
Antibacterial drugs/ antiprotozoal drugs/ acne, rosacea	2.56; (1.96, 3.4)	1.17; (0.92, 1.49)	1.08; (0.8, 1.44)	0.61; (0.47, 0.81)	0.79; (0.64, 0.97)
Antidepressant drugs	2.44; (2.01, 2.96)	1.07; (0.84, 1.35)	0.98; (0.8, 1.21)	0.78; (0.66, 0.93)	0.73; (0.63, 0.85)
Bronchodilators	2.37; (1.97, 2.85)	1.34; (1.1, 1.63)	1.25; (1.07, 1.47)	0.96; (0.84, 1.11)	0.99; (0.88, 1.12)
Chronic bowel disorders/ corticosteroids / etc	2.94; (2.41, 3.6)	1.5; (1.21, 1.84)	1.51; (1.25, 1.81)	1.13; (0.97, 1.3)	1.08; (0.95, 1.23)
Cough preparations	2.25; (1.65, 3.08)	1.36; (0.92, 2.02)	1.27; (0.96, 1.68)	0.84; (0.64, 1.09)	0.91; (0.73, 1.13)
Drugs acting on the ear	2.1; (1.47, 3.01)	1.78; (1.29, 2.45)	1.05; (0.76, 1.46)	0.79; (0.63, 0.99)	0.83; (0.66, 1.04)
Hypnotics, anxiolytics	2.26; (1.67, 3.06)	0.86; (0.58, 1.27)	1.27; (0.99, 1.63)	0.95; (0.73, 1.25)	0.8; (0.63, 1.01)
Oral nutrition	1.68; (0.73, 3.83)	1.05; (0.63, 1.73)	0.72; (0.46, 1.15)	1.09; (0.77, 1.53)	1.15; (0.88, 1.5)
Anti-arrhythmic drugs	0.73; (0.46, 1.18)	0.69; (0.47, 1.0)	0.62; (0.46, 0.85)	0.43; (0.31, 0.58)	0.44; (0.34, 0.58)
Anticoagulants, protamine	0.69; (0.54, 0.88)	0.44; (0.35, 0.55)	0.53; (0.43, 0.66)	0.45; (0.37, 0.54)	0.47; (0.41, 0.55)
Diabetes drugs	0.82; (0.68, 0.99)	0.87; (0.7, 1.08)	0.65; (0.53, 0.79)	0.58; (0.5, 0.67)	0.56; (0.49, 0.65)
Hypertension/Genito-urinary disorders	0.97; (0.72, 1.32)	0.92; (0.71, 1.2)	0.79; (0.63, 0.99)	0.58; (0.47, 0.72)	0.56; (0.48, 0.65)
Positive inotropic drugs	0.84; (0.51, 1.39)	0.91; (0.6, 1.38)	0.67; (0.49, 0.93)	0.46; (0.38, 0.57)	0.59; (0.47, 0.74)
Treatment of glaucoma	1.04; (0.57, 1.87)	0.97; (0.5, 1.89)	0.73; (0.44, 1.2)	0.8; (0.56, 1.15)	0.61; (0.47, 0.79)
Vitamins	0.96; (0.74, 1.24)	0.89; (0.64, 1.25)	0.68; (0.51, 0.9)	0.52; (0.43, 0.64)	0.51; (0.42, 0.61)
Beta-adrenoceptor blocking drugs/ hypertension drugs/etc	1.03; (0.82, 1.29)	0.55; (0.42, 0.73)	0.57; (0.46, 0.7)	0.61; (0.51, 0.74)	0.6; (0.51, 0.7)
Drugs affecting bone metabolism	1.02; (0.54, 1.92)	0.85; (0.58, 1.24)	0.47; (0.34, 0.66)	0.71; (0.56, 0.9)	0.6; (0.51, 0.7)
Antiprotozoal drugs/ neuromuscular disorders	0.79; (0.46, 1.35)	0.79; (0.51, 1.21)	0.63; (0.44, 0.91)	0.5; (0.38, 0.64)	0.59; (0.47, 0.73)
Analgesics	1.62; (1.37, 1.9)	0.93; (0.79, 1.09)	0.78; (0.68, 0.89)	0.69; (0.62, 0.77)	0.66; (0.6, 0.72)

Both disease and medication codes with unknown mapping were mapped to an “UNKNOWN” category. The “heart failure” phenotype is defined by Caliber to be a collection of Read and ICD-10 codes. We looked at the diagnoses codes strictly (as opposed to codes of historical diagnoses). The codes are found: <https://www.caliberresearch.org/portal> under the “heart failure” section using incident codes from primary (Read) and secondary care (ICD-10).

VI. DATA AVAILABILITY

The data acquired for this study is available from Clinical Practice Research Datalink (CPRD).¹ To access the data, we refer readers to the website² where it explains: “Access to data from CPRD is subject to a full licence agreement containing detailed terms and conditions of use. Anonymised patient datasets can be extracted for researchers against specific study specifications, following protocol approval from the Independent Scientific Advisory Committee (ISAC).” Therefore, data that supports the findings of this study was used under license, and not publicly available.

ACKNOWLEDGMENT

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the authors and not necessarily those of the OMS, the BHF, the GCRF, the NIHR, the ATI, the AXA Research Fund, TAILOR, or the Department of Health and Social Care.

REFERENCES

- [1] B. W. Sahle, A. J. Owen, K. L. Chin, and C. M. Reid, “Risk prediction models for incident heart failure: A systematic review of methodology and model performance,” *J. Cardiac Failure*, vol. 23, no. 9, pp. 680–687, Sep. 2017, doi: [10.1016/j.cardfail.2017.03.005](https://doi.org/10.1016/j.cardfail.2017.03.005).
- [2] E. R. C. Millett and G. Salimi-Khorshidi, “Temporal trends and patterns in mortality after incident heart failure a longitudinal analysis of 86 000 individuals,” *JAMA Cardiol.*, vol. 4, pp. 1102–1111, 2019, doi: [10.1001/jamacardio.2019.3593](https://doi.org/10.1001/jamacardio.2019.3593).
- [3] N. Conrad et al., “Temporal trends and patterns in heart failure incidence: A population-based study of 4 million individuals,” *Lancet*, vol. 391, no. 10120, pp. 572–580, 2018, doi: [10.1016/S0140-6736\(17\)32520-5](https://doi.org/10.1016/S0140-6736(17)32520-5).
- [4] F. Rahimian et al., “Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records,” *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002695, doi: [10.1371/journal.pmed.1002695](https://doi.org/10.1371/journal.pmed.1002695).
- [5] K. W. Johnson et al., “Artificial intelligence in cardiology,” *J. Amer. College Cardiol.*, vol. 71, no. 23, pp. 2668–2679, 2018, doi: [10.1016/j.jacc.2018.03.521](https://doi.org/10.1016/j.jacc.2018.03.521).
- [6] J. R. A. Solares et al., “Deep learning for electronic health records: A comparative review of multiple deep neural architectures,” *J. Biomed. Informat.*, vol. 101, 2020, Art. no. 103337. [Online]. Available: <https://doi.org/10.1016/j.jbi.2019.103337>
- [7] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, “DeepR: A convolutional net for medical records,” *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 22–30, Jan. 2017, doi: [10.1109/JBHI.2016.2633963](https://doi.org/10.1109/JBHI.2016.2633963).
- [8] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, “RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism,” in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3512–3520.
- [9] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, vol. 13-17, pp. 1135–1144, doi: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778).

¹[Online]. Available: <https://www.cprd.com/Data>

²[Online]. Available: <https://www.cprd.com/primary-care>

- [10] D. Smilov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017. [Online]. Available: <http://arxiv.org/abs/1706.03825>
- [11] C. Guan, X. Wang, Q. Zhang, R. Chen, D. He, and X. Xie, "Towards a deep and unified understanding of deep neural models in {NLP}," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, vol. 97, pp. 2454–2463. [Online]. Available: <http://proceedings.mlr.press/v97/guan19a.html>
- [12] E. Herrett *et al.*, "Data resource profile: Clinical practice research datalink (CPRD)," *Int. J. Epidemiol.*, vol. 44, no. 3, pp. 827–836, 2015, doi: [10.1093/ije/dyv098](https://doi.org/10.1093/ije/dyv098).
- [13] A. Herbert, L. Wijlaars, A. Zylbersztejn, D. Cromwell, and P. Hardelid, "Data resource profile: Hospital episode statistics admitted patient care (HES APC)," *Int. J. Epidemiol.*, vol. 46, no. 4, pp. 1093–1093, 2017, doi: [10.1093/ije/dyx015](https://doi.org/10.1093/ije/dyx015).
- [14] V. Kuan *et al.*, "A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service," *Lancet Digit. Health*, vol. 1, no. 2, pp. e63–e77, 2019, doi: [10.1016/s2589-7500\(19\)30012-3](https://doi.org/10.1016/s2589-7500(19)30012-3).
- [15] Y. Li *et al.*, "BEHRT: Transformer for electronic health records," *Sci. Rep.*, vol. 10, no. 1, 2020, Art. no. 7155, doi: [10.1038/s41598-020-62922-y](https://doi.org/10.1038/s41598-020-62922-y).
- [16] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Informat. Process. Syst.*, no. Nips, 2017. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- [17] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
- [18] P. I. Frazier, "Bayesian Optimization," 2018 *INFORMS Ann. Meet.*, no. Section 5, pp. 1–22, 2018, doi: [10.1287/educ.2018.0188](https://doi.org/10.1287/educ.2018.0188).
- [19] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. 2019 Conf. North*, pp. 4171–4186, 2019, doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
- [20] B. C. Kwon *et al.*, "RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 299–309, Jan. 2019, doi: [10.1109/TVCG.2018.2865027](https://doi.org/10.1109/TVCG.2018.2865027).
- [21] J. O. Friedrich, N. K. J. Adhikari, and J. Beyene, "The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: A simulation study," *BMC Med. Res. Methodol.*, vol. 8, 2008, Art. no. 32, doi: [10.1186/1471-2288-8-32](https://doi.org/10.1186/1471-2288-8-32).
- [22] T. J. Cahill and R. K. Kharbanda, "Heart failure after myocardial infarction in the era of primary percutaneous coronary intervention: Mechanisms, incidence and identification of patients at risk," *World J. Cardiol.*, vol. 9, pp. 407–415, 2017, doi: [10.4330/wjc.v9.i5.407](https://doi.org/10.4330/wjc.v9.i5.407).
- [23] J. Hippisley-Cox and C. Coupland, "Development and validation of risk prediction equations to estimate future risk of heart failure in patients with diabetes: A prospective cohort study," *BMJ Open*, vol. 5, no. 9, Sep. 2015, Art. no. e008503, doi: [10.1136/bmjopen-2015-008503](https://doi.org/10.1136/bmjopen-2015-008503).
- [24] V. Kuan *et al.*, "A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service," *Lancet Digit. Health*, vol. 1, no. 2, pp. e63–e77, 2019, doi: [10.1016/S2589-7500\(19\)30012-3](https://doi.org/10.1016/S2589-7500(19)30012-3).
- [25] T. von der Brück and M. Pouly, "Text similarity estimation based on word embeddings and matrix norms for targeted marketing," in *Proc. 2019 Conf. North Amer. Chapt. Associat. Computat. Linguist.: Human Lang. Technol., Volume 1 (Long and Short Papers)*, 2019, pp. 1827–1836.
- [26] J. Tromp *et al.*, "Age dependent associations of risk factors with heart failure: Pooled population based cohort study," *BMJ*, vol. 372, 2021, Art. no. 461, doi: [10.1136/bmj.n461](https://doi.org/10.1136/bmj.n461).
- [27] S. J. Jacobsen, D. S. Freedman, R. G. Hoffmann, H. W. Gruchow, A. J. Anderson, and J. J. Barboriak, "Cholesterol and coronary artery disease: Age as an effect modifier," *J. Clin. Epidemiol.*, vol. 45, no. 10, pp. 1053–1059, Oct. 1992, doi: [10.1016/0895-4356\(92\)90145-d](https://doi.org/10.1016/0895-4356(92)90145-d).
- [28] T. J. Campbell and P. S. MacDonald, "Digoxin in heart failure and cardiac arrhythmias," *Med. J. Aust.*, vol. 179, no. 2, pp. 98–102, 2003.
- [29] D. E. Sholter and P. W. Armstrong, "Adverse effects of corticosteroids on the cardiovascular system," *Can. J. Cardiol.*, vol. 16, no. 4, pp. 505–511, 2000.
- [30] R. D. Goodwin, K. W. Davidson, and K. Keyes, "Mental disorders and cardiovascular disease among adults in the United States," *J. Psychiatr. Res.*, vol. 43, no. 3, pp. 239–246, Jan. 2009, doi: [10.1016/j.jpsychires.2008.05.006](https://doi.org/10.1016/j.jpsychires.2008.05.006).
- [31] C. C. Butler *et al.*, "Treatment of acute cough/lower respiratory tract infection by antibiotic class and associated outcomes: A 13 European country observational study in primary care," *J. Antimicrobial Chemotherapy*, vol. 65, no. 11, pp. 2472–2478, Nov. 2010, doi: [10.1093/jac/dkq336](https://doi.org/10.1093/jac/dkq336).
- [32] C. Macie, K. Wooldrage, J. Manfreda, and N. Anthonisen, "Cardiovascular morbidity and the use of inhaled bronchodilators," *Int. J. Chronic Obstructive Pulmonary Dis.*, vol. 3, no. 1, pp. 163–169, 2008, doi: [10.2147/copd.s1516](https://doi.org/10.2147/copd.s1516).
- [33] G. S. Bleumink, J. Feenstra, M. C. J. M. Sturkenboom, and B. H. C. Stricker, "Nonsteroidal anti-inflammatory drugs and heart failure," *Drugs*, vol. 63, no. 6, pp. 525–534, 2003, doi: [10.2165/00003495-200363060-00001](https://doi.org/10.2165/00003495-200363060-00001).
- [34] J. Mäenpää and O. Pelkonen, "Cardiac safety of ophthalmic timolol," *Expert Opin. Drug Saf.*, vol. 15, pp. 1549–1561, 2016, doi: [10.1080/14740338.2016.1225718](https://doi.org/10.1080/14740338.2016.1225718).
- [35] P. Harasymowycz *et al.*, "Medical management of glaucoma in the 21st century from a Canadian perspective," *J. Ophthalmol.*, vol. 2016, 2016, Art. no. 6509809, doi: [10.1155/2016/6509809](https://doi.org/10.1155/2016/6509809).
- [36] W. T. Abraham, "β-blockers: The new standard of therapy for mild heart failure," *Arch. Intern. Med.*, vol. 160, pp. 1237–1247, 2000, doi: [10.1001/archinte.160.9.1237](https://doi.org/10.1001/archinte.160.9.1237).
- [37] M. Lièvre, S. Morand, B. Besse, J. N. Fiessinger, and J. P. Boissel, "Oral beraprost sodium, a prostaglandin I₂ analogue, for intermittent claudication: A double-blind, randomized, multicenter controlled trial," *Circulation*, vol. 102, pp. 426–431, 2000, doi: [10.1161/01.CIR.102.4.426](https://doi.org/10.1161/01.CIR.102.4.426).
- [38] E. R. Mohler, W. R. Hiatt, J. W. Olin, M. Wade, R. Jeffs, and A. T. Hirsch, "Treatment of intermittent claudication with beraprost sodium, an orally active prostaglandin I₂ analogue: Double-blinded, randomized, controlled trial," *J. Amer. Coll. Cardiol.*, vol. 41, pp. 1679–1686, 2003, doi: [10.1016/S0735-1097\(03\)00299-7](https://doi.org/10.1016/S0735-1097(03)00299-7).
- [39] H. I. Pass and H. W. Pogrebnik, "Potential uses of prostaglandin E1 analog for cardiovascular disease," *J. Thoracic Cardiovasc. Surg.*, vol. 108, pp. P789–P790, 1994, doi: [10.1016/s0022-5223\(94\)70312-4](https://doi.org/10.1016/s0022-5223(94)70312-4).
- [40] V. Kuan *et al.*, "A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service," *Lancet Digit. Health*, vol. 1, no. 2, pp. e63–e77, 2019, doi: [10.1016/S2589-7500\(19\)30012-3](https://doi.org/10.1016/S2589-7500(19)30012-3).
- [41] T. Benson, "The history of the read codes: The inaugural James read memorial lecture 2011," *Informat. Primary Care*, vol. 19, no. 3, pp. 173–182, 2012, doi: [10.14236/jhi.v19i3.811](https://doi.org/10.14236/jhi.v19i3.811).
- [42] W. H. Organization, *ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision*. 1st ed., Washington, DC, USA: Pan American Health Organization, 2004, Spanish version.
- [43] H. J. Curtis and B. Goldacre, "OpenPrescribing: Normalised data and software tool to research trends in English NHS primary care prescribing 1998–2016," *BMJ Open*, vol. 8, 2018, Art. no. e019921, doi: [10.1136/bmjopen-2017-019921](https://doi.org/10.1136/bmjopen-2017-019921).