medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Medical history predicts phenome-wide disease onset

Jakob Steinfeldt MD[1,2,3,14,15*], Benjamin Wild PhD[4,*], Thore Buergel MSc[4,15*], Maik Pietzner PhD[3,5,11], Julius Upmeier zu Belzen MSc[4], Andre Vauvelle MSc[6], Stefan Hegselmann MSc[7,8], Professor Spiros Denaxas PhD[6,9,10], Professor Harry Hemingway MD[6,10], Professor Claudia Langenberg MD, PhD[3,5,11], Professor Ulf Landmesser MD[1,2,3,12,13,14,§], Professor John Deanfield MD, FMedSci[15,§], Professor Roland Eils PhD[4,16,§,†]

* these authors contributed equally, shared first authorship

§ shared senior authorship

† corresponding author

Affiliation

[1]Department of Cardiology, Angiology and Intensive Care Medicine, Deutsches Herzzentrum der Charité (DHZC), Berlin, Germany,

[2]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Klinik/Centrum, Charitéplatz 1, 10117 Berlin, Germany.

[3]Computational Medicine, Berlin Institute of Health (BIH), Charite - University Medicine Berlin, Berlin, Germany,

[4]Center for Digital Health, Berlin Institute of Health (BIH), Charite - University Medicine Berlin, Berlin, Germany,

[5]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK,

[6]Institute of Health Informatics, University College London, London, UK,

[7]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Massachusetts, USA,

[8]Pattern Recognition and Image Analysis Lab, University of Münster, Münster, Germany,

[9]Data Science Centre, British Heart Foundation, London, UK,

[10]Health Data Research UK, London, UK,

[11]Precision Health University Research Institute, Queen Mary University of London and Barts NHS Trust, London, UK,

[12]Berlin Institute of Health (BIH), Charite - University Medicine Berlin, Berlin, Germany,

[13]DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, Berlin, Berlin, Germany,

[14]Friede Springer Cardiovascular Prevention Center@Charite, Charite - University Medicine Berlin, Berlin, Germany,

[15]Institute of Cardiovascular Sciences, University College London, London, UK,

[16]Health Data Science Unit, Heidelberg University Hospital and BioQuant, Heidelberg, Germany.

Note: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Abstract

Background: Current medicine falls short at providing systematic data-driven guidance to individuals and care providers. While an individual's medical history is the foundation for every medical decision in clinical practice and is routinely recorded in most health systems, the predictive potential and utility for most human diseases is largely unknown.
Methods: We explored the potential of the medical history to inform on the phenome-wide risk of onset for 1,883 disease endpoints across clinical specialties. Specifically, we developed a neural network to learn disease-specific risk states from routinely collected health records of 502,460 individuals from the British UK Biobank and validated this model in the US-American All of US cohort with 229,830 individuals. In addition, we illustrated the potential in 24 selected conditions, including type 2 diabetes, hypertension, coronary heart disease, heart failure, and diseases not formerly considered predictable from health records, such as rheumatoid arthritis and endocarditis.
Results: We show that the medical history stratifies the risk of onset for all investigated conditions across clinical specialties. For 10-year risk prediction, the medical history provided significant improvements over basic demographic predictors for 1,800 (95.6%) of the 1,883 investigated endpoints in the UK Biobank cohort. After transferring the unmodified risk models to the independent All of US cohort, we found improvements for 1,310 (83.5%) of 1,568 endpoints, demonstrating generalizability across healthcare systems and historically underrepresented groups. Finally, we found predictive information comparable with current guideline-recommended scores for the primary prevention of cardiovascular diseases and illustrated how the risk scores could facilitate rapid response to emerging pathogenic health threats.
Conclusion: Our study demonstrates the great potential of leveraging the medical history to provide comprehensive phenome-wide risk estimation at minimal cost. We anticipate that this approach has the potential to disrupt medical practice and decision-making, from early disease diagnosis, slowing of disease progression to interventions against preventable diseases.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Introduction

The assessment of an individual's risk for future disease is central to guiding preventive interventions, early detection of disease, and the initiation of treatments. However, bespoke risk scores are only available for a few common diseases[1–4], leaving healthcare providers and individuals with little to no guidance on the majority of relevant diseases. Even for diseases with established risk scores, little consensus exists on which score to use and associated physical or laboratory measurements to obtain, leading to highly fragmented practice in routine care[5].

At the same time, most medical decisions on diagnosis, treatment, and prevention of diseases are fundamentally based on an individual's medical history[6]. With the widespread digitalization, this information is routinely collected by healthcare providers, insurances, and governmental organizations at a population scale in the form of electronic health records[7–12]. These readily accessible records, which include diseases, medications, and procedures, are potentially informative about future risk trajectories, but their potential to improve medical decision-making is limited by the human ability to process and understand vast amounts of data[13].

To date, routine health records have been used for etiological[14–17], diagnostic[18], and prognostic research[15,16,19–21]. Existing efforts often extract and leverage known clinical predictors with new methodologies[18], augment them with additionally extracted data modalities such as clinical notes[22], or aim to identify novel predictors among the recorded concepts[14–17]. Prior work on the prediction of disease onset has mainly focused on single diseases, including dementia[15,23], cardiovascular conditions[22,24] such as heart failure[25] and atrial fibrillation[26,27]. In contrast, phenome-wide association studies (PheWAS) quantifying the associations of genetic

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

variants with comprehensive phenotypic traits are emerging in genetic epidemiology[28,29]. While approaches have been developed to extract information from longitudinal health records[30,31], no studies have investigated the predictive potential and potential utility over the entire human phenome. Consequently, the predictive information in routinely collected health records and its potential to systematically guide medical decision-making is largely unexplored.

Here we examined the predictive potential of an individual's entire medical history and propose a systematic approach for phenome-wide risk stratification. We developed, trained, and validated a neural network, specifically a multi-layer-perceptron, in the UK Biobank cohort[32] to simultaneously estimate disease risk from routinely collected health records for 1,883 endpoints across clinical specialties. These endpoints include preventable diseases (e.g. coronary heart disease), diseases which are not currently preventable, but early diagnosis has been shown to substantially slow down progression and development of complications (e.g. heart failure) and outcomes which are currently neither entirely preventable nor treatable (e.g. death). They also include both diseases with risk prediction models recommended in guidelines and used in practice (e.g. cardiovascular diseases or breast cancer) as well as diseases without current risk prediction models (e.g. psoriasis and rheumatoid arthritis).

We evaluated our approach by integrating the endpoint-specific risk states estimated by the neural network in Cox Proportional Hazard models[33], investigating the phenome-wide predictive potential over basic demographic predictors, and illustrating how phenome-wide risk stratification could benefit individuals by providing risk estimates, facilitating early disease diagnosis, and guiding preventive interventions. Furthermore, by externally validating in the All Of Us cohort[34], we show that our models can generalize across healthcare systems and populations, including communities historically underrepresented in biomedical research.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

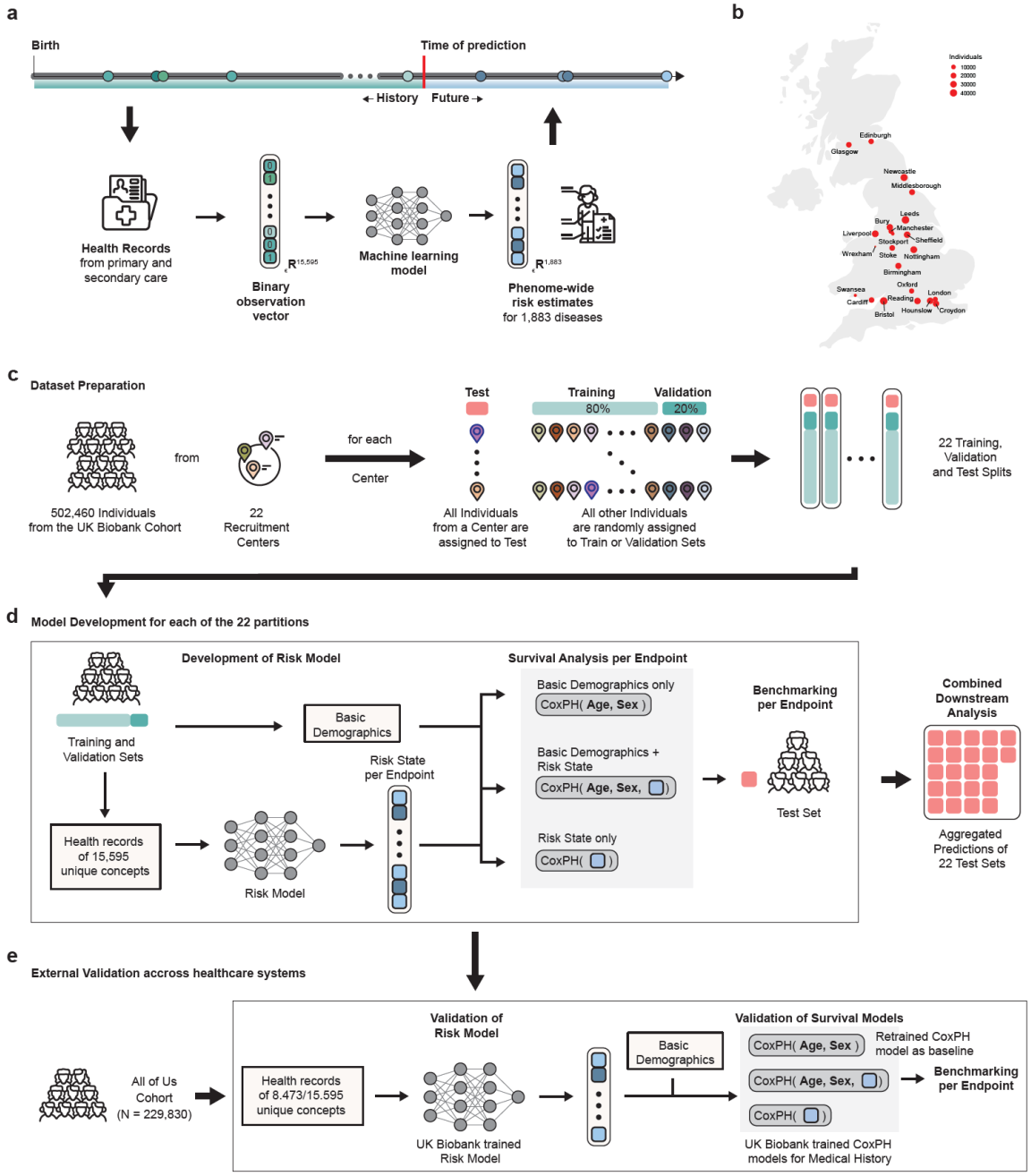


Figure 1: Overview on the study: a) The medical history captures encounters with primary and secondary care, including diagnoses, medications, and procedures (ideally) from birth. Here we train a multi-layer perceptron on data before recruitment to predict phenome-wide incident disease onset for 1,883 endpoints. **b)** Location and size of the 22 assessment centers of the UK Biobank cohort across England, Wales, and Scotland. **c)** To learn risk states from individual medical histories, the UK Biobank population was partitioned by their respective assessment center at recruitment. **d)** For each of the 22 partitions, the Risk Model was trained to predict phenome-wide incident disease onset for 1,883 endpoints. Subsequently, for each endpoint, Cox proportional hazard (CPH) models were developed on the risk states in combination with sets of commonly available predictors to model disease risk. Predictions of the CPH model on the test set were aggregated for downstream analysis. **e)** External validation in the All of US cohort. After mapping to the OMOP vocabulary, we transferred the trained risk model to the All of US cohort and calculated the risk state for all endpoints. To validate these risk states, we compared the unchanged CPH models developed in the UK Biobank with refitted CPH models for age and sex.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Finally, we assessed the potential of our approach to aid risk stratification for the primary prevention of cardiovascular disease and to respond to emerging health threats at the example of COVID-19. Our results demonstrate the currently unused potential of routine health records to guide medical practice by providing comprehensive phenome-wide risk estimates.

Results

Characteristics of the study population and integration of routine health records

This study is based on the UK Biobank cohort[32,35], a longitudinal population cohort of 502,460 relatively healthy individuals of primarily British descent, with a median age of 58 (IQR 50, 63) years, 54.4% biological females, 11% current smokers, and a median BMI of 26.7 (IQR 24.1, 29.9) at recruitment (Table 1 for detailed information). Individuals recruited between 2006 and 2010 and were followed for a median of 12.6 years resulting in ~6.2M overall person-years on 1,883 phenome-wide endpoints[36] with ≥ 100 incident events (> 0.02% of individuals have the event in the observation time). We externally validated our findings in individuals from the All of Us cohort, a longitudinal cohort of 229,830 individuals with linked health records recruited from all over the United States. Individuals in the All of Us cohort are of diverse descent, with 46% of reportedly non-white ethnicity and 78% of groups historically underrepresented in biomedical research[34,37], and have a median age of 54 (IQR 38, 65) years with 61.1% biological females (see Table 1 for detailed information). Individuals were recruited from 2019 on and followed for a median of 3.5 years, resulting in ~787,300 person-years on 1,568 endpoints.

Central to this study is the prior medical history, defined as the entirety of routine health records before recruitment. Before further analysis, we mapped all health records to the OMOP

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

vocabulary. While most records originate from primary care and, to a lesser extent, secondary care (Suppl. Fig 1a), the predominant record domains are drugs and observations, followed by conditions, procedures, and devices (Supp. Fig 1b). Interestingly, while rare medical concepts (with a record in < 1% of individuals in the study population) are not commonly included in prediction models[20], they are often associated with high incident event rates (exemplified by the mortality rate in Suppl. Figure 1c) compared to common concepts (a record present in >= 1% of the study population). For example, the concept code for "portal hypertension" (OMOP 34742003) is only recorded in 0.04% (203) of individuals at recruitment, but 48.7% (99 individuals) will die over the course of the observation period. Importantly, there are many distinct rare concepts, and thus 91.7% of individuals have at least one rare record before recruitment, compared with 92.5% for common records. In addition, 60.7% of individuals have ≥ 10 rare records compared with 78.4% for common records, and individuals have only slightly fewer rare than common records (Suppl. Figure 1d).

After excluding very rare concepts (< 0.01%, less than 50 individuals with the record in this study), we integrated the remaining 15,595 unique concepts (Supplementary Table 2) with a multi-task multi-layer perceptron to predict the phenome-wide onset of 1,883 endpoints (Supplementary Table 1) simultaneously (Figure 1a).

To ensure that our findings are generalizable and transferable, we spatially validate our models in 22 recruitment centers (Fig 1b) across England, Wales, and Scotland. We developed 22 models, each trained on individuals from 21 recruitment centers, randomly split into training and validation sets (Fig 1c). We subsequently tested the models on individuals from the additional recruitment center unseen for model development for internal spatial validation. After checkpoint selection on the validation data sets and obtaining the selected models' final predictions on the individual test sets, the test set predictions were aggregated for downstream analysis (Figure 1d). Subsequently, disease-specific exclusions of prior events and sex-

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

specificity were respected in all downstream analyses. After development, the models were externally validated in the All of Us cohort[34].

Routine health records stratify phenome-wide disease onset

Central to the utility of any predictor is its potential to stratify risk. The better the stratification of low and high-risk individuals, the more effective targeted interventions and disease diagnoses are.

To investigate whether health records can be used to identify high-risk individuals, we assessed the relationship between the risk states estimated by the neural network for each endpoint and the risk of future disease (Figure 2). For illustration, we first aggregated the incident events over the percentiles of the risk states for each endpoint and subsequently calculated ratios between the top and bottom 10% of risk states over the entire phenome (Fig 2A). Importantly, we found differences in the event rates, reflecting a stratification of high and low-risk individuals for almost all endpoints covering a broad range of disease categories and etiologies: For 1,404 of 1,883 endpoints (74.6%), we observed >10-times as many events for individuals in the top 10% of the predicted risk states compared to the bottom 10%. For instance, these endpoints included rheumatoid arthritis (Ratio ~ 12.5), coronary heart disease (Ratio ~ 23), or chronic obstructive pulmonary disease (Ratio ~ 63). For 286 (15.1%) of the 1,883 conditions, including abdominal aortic aneurysm (Ratio ~ 212), and all-cause mortality (Ratio ~ 107), more than 100 times the number of individuals in the top 10% of predicted risk states had incident events compared to the bottom 10%. For 479 (25.4%) endpoints, the separation between high and low-risk individuals was smaller (Ratio < 10), which included hypertension (Ratio ~ 5.5) and anaemia (Ratio ~ 6.9), often diagnosed earlier in life or precursors for future comorbidities. Notably, the ratios were > 1 for all 1,883 investigated endpoints, even though all models were

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

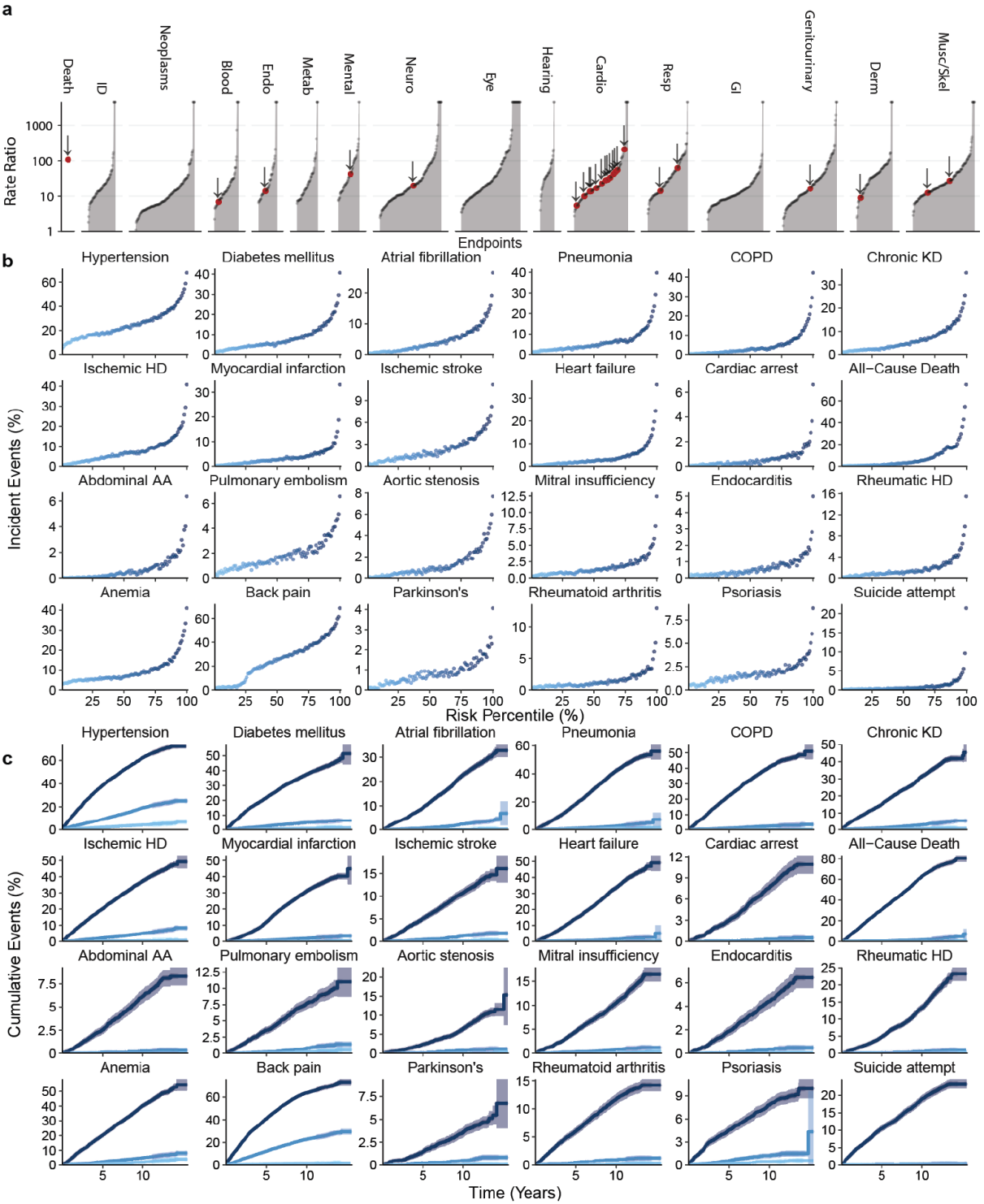


Figure 2: Routine health records stratify phenome-wide disease onset: a) Ratio of incident events in the Top 10% compared with the Bottom 10% of the estimated risk states. Event rates in the Top 10% are higher than in the Bottom 10% for all 1,883 investigated endpoints. Red dots indicate 24 selected endpoints detailed in Fig 2B. To illustrate, 1,238 (2.49%) individuals in the top risk decile for cardiac arrest experienced an event compared with only 29 (0.06%) in the bottom decile, with a risk ratio of 42.69. **b)** Incident event rates for a selection of 24 endpoints. **c)** Cumulative event rates for the Top 1%, median, and Bottom 1% of risk percentiles over 15ys. Statistical measures were derived from 502.460 individuals. Individuals with prevalent diseases were excluded from the endpoints-specific analysis.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

developed in spatially segregated assessment centers. We found a small positive correlation between the number of incident events and the rate ratio of an endpoint, i.e., the model performs slightly better on average for more common diseases (Pearson's correlation coefficient log(incident events), log(rate ratio): 0.117 CI (0.082, 0.152)). The complete list of all endpoints and corresponding statistics can be found in Suppl. Table 4.

In addition to the phenome-wide analysis of 1,883 endpoints, we also provide detailed associations between the risk percentiles and incident event ratios (Fig 2b) as well as cumulative event rates for up to 15 years (Fig 2c) of follow-up for the top, median, and bottom percentiles for a subset of 24 selected endpoints. This set was selected to comprise actionable endpoints and common diseases with significant societal burdens, specific cardiovascular conditions with pharmacological and surgical interventions, as well as endpoints without established tools to stratify risk to date. To illustrate the potential, 1,238 (2.49%) individuals in the top risk decile for cardiac arrest experienced an event compared with only 29 (0.06%) in the bottom decile, with a risk ratio of 42.69 (Fig. 2A, B). In the top 1% percentile, 332 (6.61%) of the 5021 individuals experienced an event 15 years after recruitment, while no event was recorded for the 5,022 individuals in the bottom 1% risk percentile (Fig. 2B). Thus, high-risk individuals could be, for instance, considered to receive a preventive implantable cardioverter-defibrillator (ICD)[38].

In summary, the disease-specific states stratify the risk of onset for all 1,883 investigated endpoints across clinical specialties. This indicates that routine health records provide a large and widely unused potential for the systematic risk estimation of disease onset in the general population.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Discriminative performance indicates potential utility

While routine health records can stratify incident event rates, this does not prove utility. To test whether the risk state derived from the routine health records could provide utility and information beyond ubiquitously available predictors, we investigated the predictive information over age and biological sex. We modeled the risk of disease onset using Cox Proportional-Hazards (CPH) models for all 1,883 endpoints, which allowed us to estimate adjusted hazard ratios (denoted as HR in Suppl. Table 6) and 10-year discriminative improvements (indicated as Delta C-index in Figure 3a).

We found significant improvements over the baseline model (age and biological sex only) for 1,800 (95.6%) of the 1,883 investigated endpoints (Figure 3, Supplementary Table 5). For several of these endpoints, the discriminative improvements were considerable (Delta C-Index Q25%: 0.099, Q50: 0.119, Q75: 0.139). We found significant improvements for 23 of the highlighted subset of 24 endpoints (indicated in Fig 2A), with the largest increases for the prediction of suicide attempts (C-Index: 0.608 (CI 0.602, 0.615) → 0.831 (CI 0.826, 0.837)), back pain (C-Index: 0.519 (CI 0.517, 0.521) → 0.72 (CI 0.719, 0.722)), all-cause mortality (C-Index 0.701 (CI 0.699, 0.703) → 0.878 (CI 0.877, 0.88)) and chronic obstructive pulmonary disease (C-Index 0.662 (0.66, 0.666) → 0.818 (CI 0.815, 0.82)). In contrast, we did not find improvements in the prediction of other endpoints, e.g., Parkinson's disease (C-Index: 0.738 (CI 0.732, 0.745) → 0.737 (CI 0.731, 0.743)).

For illustration, we also present individual phenome-wide risk profiles (Figure 3c, Suppl. Figure 2a+b, Suppl. Figure 3a+b). The risk profiles varied substantially in the predispositions relative to the age and sex reference (the inner circle, see methods for details) and the absolute 10-year risk estimates (the outer circle). The first individual (Figure 3c), a 60-year-old man, is

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

predicted to be at a particularly high 10-year risk of metabolic, cardiovascular, respiratory, and genitourinary conditions, including diabetes mellitus (19.4%), heart failure (22%), COPD (14.9%), and chronic kidney disease (16.8%). Increased risk of neoplastic, dermatological, and musculoskeletal conditions was not predicted by the prior health records of this individual. In contrast, another individual, a 48-year-old woman (Suppl. Figure 3b), is not estimated at increased cardiovascular risk but conversely to have almost 10x the risk for suicide ideation and attempt or self-harm compared to the reference group.

We provide the highest attributed records in the study population for selected endpoints (Fig 3d, Suppl. Figure 2c, Suppl. Figure 3c) and the full attributions for all 24 highlighted endpoints (Supplementary Table 9).

These findings indicate that health records contain substantial predictive information beyond basic demographic predictors for a wide range of endpoints from across clinical specialities.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

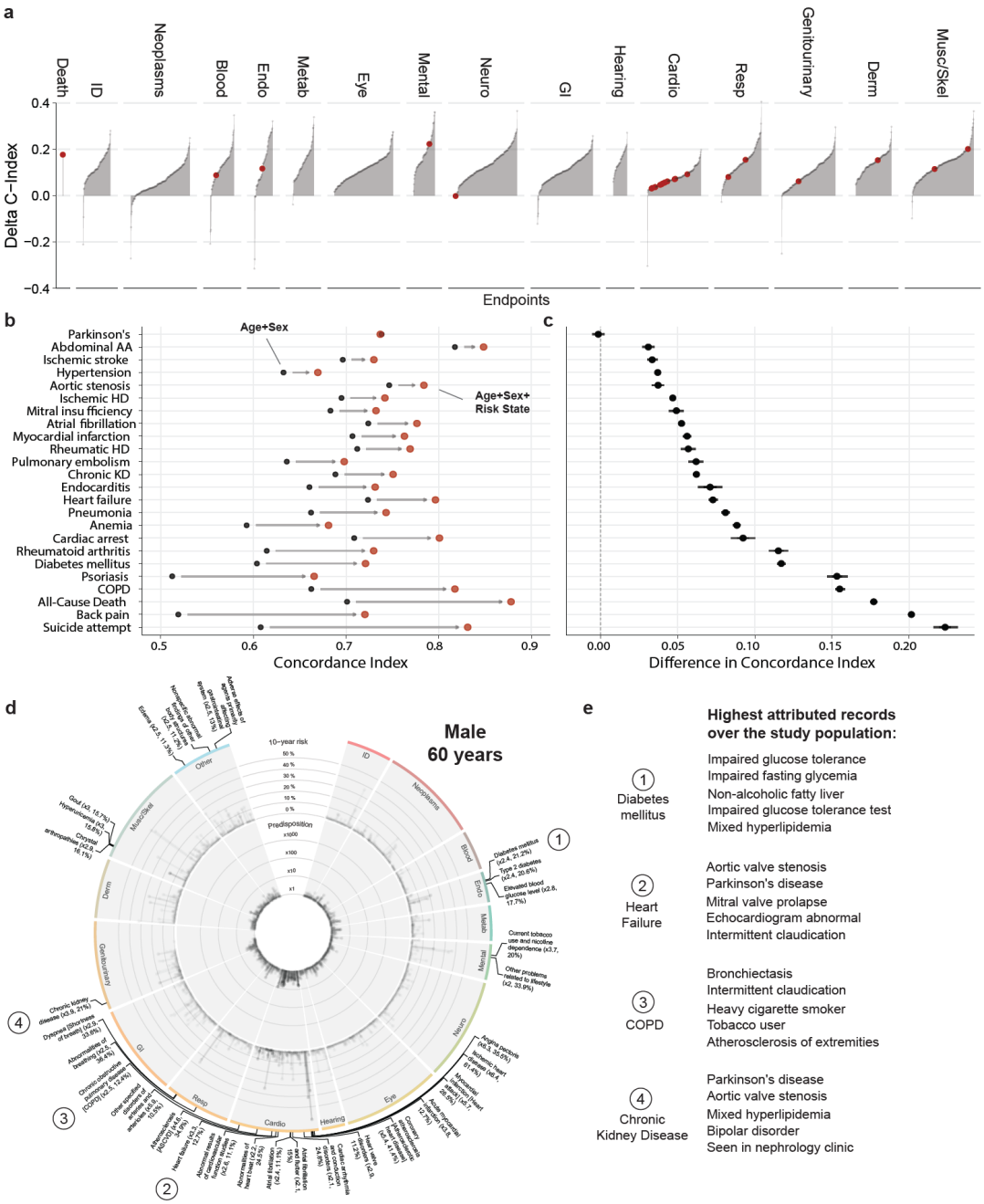


Figure 3: Discriminative performance indicates potential utility: a) Differences in discriminatory performance quantified by the C-Index between CPH models trained on Age+Sex and Age+Sex+RiskState for all 1,883 endpoints. We find significant improvements over the baseline model (Age+Sex, age, and biological sex only) for 1800 (95.6%) of the 1,883 investigated endpoints. Red dots indicate selected endpoints in Fig. 3b. **b)** Absolute discriminatory performance in terms of C-Index comparing the baseline (Age+Sex, black point) with the added routine health records risk state (Age+Sex+RiskState, red point) for a selection of 24 endpoints. **c)** The direct C-index differences for the same models. Dots indicate medians and whiskers extend to the 95% confidence interval for a distribution bootstrapped over 100 iterations. **d)** Example of individual predicted phenome-wide risk profile. Predisposition (10-year risk estimated by Age+Sex+RiskState compared to risk estimated by Age+Sex alone) is displayed in the inner circle, and absolute 10-year risk estimated by Age+Sex+RiskState can be found in the outer circle. Labels indicate endpoints with a high individual predisposition (> 2 times higher than the Age+Sex-based reference estimate) and absolute 10-year risk > 10%. **e)** Top 5 highest attributed records for selected endpoints.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Predictive models can generalise across healthcare systems and populations

While our findings indicate potential utility in the UK Biobank, health records vary substantially across healthcare systems and over time due to differences in medical and coding practices ("distribution shift") and underlying differences in the populations. Thus, predictive models can fail to learn robust and generalisable information[39–41].

To understand better the generalisability across different healthcare systems, we predicted risk states and absolute risk estimates for all individuals in the All of Us cohort with linked medical records (N=229,830; see Table 1). Importantly, we found significant improvements over the baseline model (age and biological sex only) for 1,310 (83.5%) of the 1,568 investigated endpoints with at least 100 incident events (Figure 4A, Supplementary Table 8). We furthermore found significant improvements for all of the 24 selected endpoints, with improvements ranging from modest for hypertension (C-Index: 0.627 (CI 0.623, 0.632) → 0.643 (CI 0.639, 0.648)) and Parkinson's disease (C-Index: 0.817 (CI 0.800, 0.834) → 0.849 (CI 0.832, 0.0.863)) to substantial for, e.g., All-Cause Death (C-Index: 0.693 (CI 0.685, 0.704) → 0.807 (CI 0.799, 0.815)), Pulmonary embolism (C-Index: 0.590 (CI 0.578, 0.603) → 0.711 (CI 0.701, 0.723)), and Cardiac arrest (C-Index: 0.641 (CI 0.625, 0.659) → 0.818 (CI 0.804, 0.831)) (Figure 4B,C). For a subset of 65 (3.93%) endpoints, the discriminative performance deteriorated significantly with the addition of health record information over age and biological sex alone.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

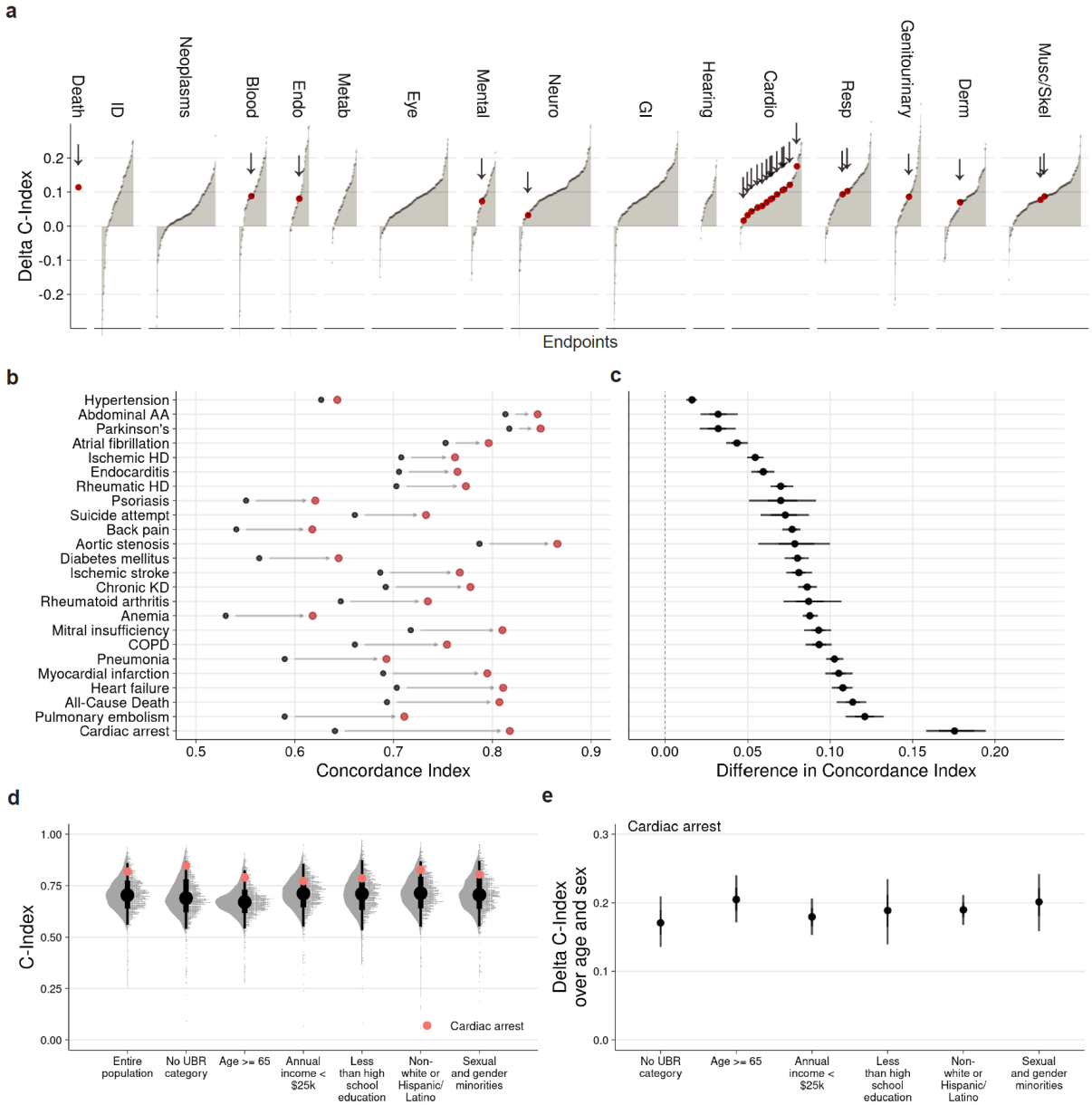


Figure 4: Predictive models can generalize across health care systems and populations: a) External validation of the differences in discriminatory performance quantified by the C-Index between CPH models trained on age and biological sex and age, biological sex and the risk state for 1.568 endpoints in the All of Us cohort. We find significant improvements over the baseline model (age and biological sex only) for 1.310 (83.5%) of the 1.658 investigated endpoints. **b)** Absolute discriminatory performance in terms of C-Index comparing the baseline (age and biological sex, black point) with the added routine health records risk state (red point) for a selection of 24 endpoints. **c)** The differences in C-index for the same models. **d)** Distribution of C-Indices for the 1.658 investigated endpoints stratified by communities historically underrepresented in biomedical research (UPD). **e)** For the same groups, confidence intervals for the additive performance as measured by the C-Index compared to the baseline model.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

As the risk states were largely derived from white, middle-aged, and generally affluent and healthy individuals from the UK, it was critical to validate the discriminative performance in diverse and historically underserved and underrepresented groups and ethnicities. Generally, we found comparable discriminative performances (Figure 4D) and substantial benefits over basic demographic predictors (example of cardiac arrest in Fig 4E) across all investigated groups.

Taken together, our findings suggest that predictive models based on the medical history can generalise across health systems and diverse populations.

Predictions can support cardiovascular disease prevention and the response to emerging health threats.

While comprehensive phenome-wide risk profiles provide opportunities to guide medical decision-making, not all of the predictions are actionable. To illustrate the immediately actionable potential, we focused on the primary prevention of cardiovascular disease and the response to newly emerging health threats.

Risk scores are well established in the primary prevention of cardiovascular events and have been recommended to guide preventive lipid-lowering interventions[42]. While cardiovascular predictors are accessible at a low cost, dedicated visits to healthcare providers for physical and laboratory measurements are required. Therefore, we compared our phenome-wide risk score, based only on age, sex, and routine health records, to models based on established cardiovascular risk scores, the SCORE2[43], the ASCVD[3], and the British QRISK3[4] score. Interestingly, the discriminative performance of our phenome-wide model is competitive with the established cardiovascular risk scores for all investigated cardiovascular endpoints (Suppl.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Figure 4a, Suppl. Table 7): we found comparable C-Indices with differences +0.002 (CI -0.001, 0.005) for ischemic stroke, +0.003 (CI 0.001, 0.005) for ischemic heart disease and +0.007 (CI 0.004, 0.009) for myocardial infarction compared with the comprehensive QRISK3 score. It is noteworthy, that these discriminative improvements are substantially better for later-stage diseases, including heart failure (+0.02 (CI 0.017, 0.022)), cardiac arrest (+0.055 (CI 0.049, 0.062)), and all-cause mortality (+0.136 (CI 0.134, 0.138)), when prior health records are considered.

With newly emerging pathogenic health threats, rapid and reliable risk stratification is required to protect high-risk groups and prioritise preventive interventions. We investigated how our phenome-wide risk states could be repurposed for COVID-19, a respiratory infection with pneumonia and sepsis as common, life-threatening complications of severe cases. We repurposed the risk states for pneumonia, sepsis, and all-cause mortality to calculate a combined COVID-19 severity risk score using information available at the end of 2019 before the global spread of the COVID-19 pandemic (see Methods for details). The COVID-19 severity risk score resembles the risk for developing severe or fatal COVID-19 and illustrates how health records could help to identify individuals at high risk and to prioritize individuals in initial vaccination campaigns better. Augmenting age with the COVID-19 severity risk score, we found substantially improved discriminative performance for both severe and fatal COVID-19 outcomes (Severe: C-Index (age) 0.598 (CI 0.590, 0.605) → C-Index (age + COVID-19 severity risk score) 0.649 (CI 0.642, 0.655); Fatal: C-Index (age) 0.720 (CI 0.710, 0.729) → C-Index (age + COVID-19 severity risk score) 0.783 (CI 0.775, 0.792). These discriminative improvements translate into higher cumulative incidence in the Top 5% population compared to age alone (Suppl. Figure 4C, age (left), COVID-19 severity score (right), severe COVID-19 (top), fatal COVID-19 (bottom)): In the top 5% of the age-based risk

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

group (~ 79 (IQR 77, 81) years old), 0.42% (CI 0.34%, 0.5%, n=105) have been hospitalised, and 0.26% (CI 0.2%, 0.33%, n=66) had died by the end of the first wave. By the end of the second wave, around 0.96% (CI 0.83%, 1.08%, n=240) had been hospitalised and 0.44% (0.36%, 0.52%, n=111) had died. In contrast, for individuals in the top 5% of the COVID-19 severity risk score, by the end of the first wave, around 0.61% (CI 0.51%, 0.71%, n=153) had been hospitalised, and 0.54% (0.45%, 0.63%, n=136) had died, while by the end of the second wave, 1.24% (CI 1.1%, 1.38%, n=312) had been hospitalised and 0.83% (0.72%, 0.94%, n=208) had died.

In summary, our findings indicate that medical history facilitates both the primary prevention of cardiovascular diseases and the rapid response to emerging health threats.

Discussion

Current clinical practice falls short of providing systematic data-driven guidance to individuals and care providers. In this study, we demonstrated for the first time the potential of the medical history to systematically inform on phenome-wide risk across clinical specialties. Our results indicated utility beyond conventional predictors, for preventable diseases, treatable diseases and diseases without existing risk stratification tools. We anticipate that this approach has the potential to disrupt medical practice and facilitate population health at scale.

There are three main scenarios of potential utility: First, the disease is preventable and effective interventions exist and can be recommended early to individuals at high risk, e.g., in the case of lipid-lowering medication for primary prevention of coronary heart disease[42]. Here, for example lowering LDL cholesterol in 10,000 individuals at increased risk by 2mmol/L with atorvastatin 40mg daily (~2€ per month) for 5 years would prevent 500 vascular events,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

reducing the individual relative risk by more than a third[44,45]. Second, while most diseases are not currently preventable, early detection has been shown to slow substantially the progression and development of adverse events of other conditions, e.g., optimal medical therapy in individuals with type 2 diabetes[46] or systolic heart failure[47]. In individuals with heart failure with reduced ejection fraction a comprehensive treatment regime (including ARNI, beta blockers, MRA and SGLT2 inhibitors) compared to a conventional regime (ACEi or ARB and beta blockers) reduced the hospital admissions for heart failure by more than two thirds, all-cause mortality by almost half[48]. For a 55-year old male, this translated into an estimated 8.3 additional years free from cardiovascular death or readmission for heart failure. Finally, even if an outcome is not preventable or treatable, estimates of prospective individual risk may be of high importance for personal decisions or planning of advanced care, e.g., a high short-term mortality could identify patients in need of transitioning from curative to palliative strategies for optimal care[49,50]. Multiple studies have shown that palliative care services can improve patients symptoms, life quality and may even increase survival[51]. In conclusion, our approach could facilitate the identification and targeting of high risk populations for specific screening programs, and thus has the potential to improve the value of national health programmes.

In addition, our findings indicate that predictive models based on routine health records can generalise across diverse health systems, populations, and ethnicities. Surprisingly, despite the vast differences in the records from the U.S. health system in the All of Us cohort and the records from England, Wales, and Scotland in the UK Biobank, the models could be successfully transferred without further modification or retraining. In contrast to the UK Biobank, it is highly diverse and emphasises enrollment of groups historically underrepresented in biomedical research. Notably, records in the All of Us cohort originate from a highly fragmented healthcare system with vastly different coding standards and patterns

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

compared to the UK. Nonetheless, two central challenges remain to be considered before the application of the described approach in routine care: First, despite our promising initial findings, health records are recorded as a consequence of interactions with the medical system. As such, health records are subject to biological, procedural, and socio-economic biases[52] as well as conditional on the ever-changing nature of medical knowledge and policies. Closely connected, our findings indicate that not all diagnoses are captured explicitly in the health records but can still be inferred from the prescribed drugs (e.g., 66 individuals have a record of *Calcipotriol medication,* but no prior diagnosis of Psoriasis, and the model utilises this information to predict a high likelihood of a future diagnostic code) or procedures (e.g., the presence of a record of *Impaired glucose tolerance test* in 30 individuals without recorded diagnosis increases the models' predicted likelihood of Diabetes mellitus) for some individuals. While this indicates successful identification of disease phenotypes, not excluding these individuals could lead to an overestimation of the discriminative performance for incident events and could limit the actionable potential for these individuals. Second, as individuals in research cohorts are often healthier and have lower disease prevalence than the general populations[53], absolute risks are expected to be underestimated. However, downstream recommendations critically depend on the choice of absolute decision thresholds. Ultimately, if routine health records are to be used for risk prediction, robust governance rules to protect individuals, such as opt-out and usage reports, need to be implemented. With many national initiatives emerging to curate routine health records for millions of individuals in the general population, future studies will allow us to understand better how to overcome these challenges.

Our study presents the first systematic approach to simultaneous risk stratification for thousands of diseases across clinical specialties. It is based on the medical history, which is available in real-time in many healthcare systems at no additional cost. Our findings show the

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

potential to disrupt medical practice, leveraging data as a central element to inform and guide preventive interventions, early diagnosis, and treatment of disease.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Methods

Data source and definitions of predictors and endpoints

To derive risk states, we analysed data from the UK Biobank cohort. Participants were enrolled from 2006 to 2010 in 22 recruitment centers across England, Scotland, and Wales; the follow-up is ongoing and records until the 24th of September 2021 are included in this analysis. The UK Biobank cohort comprises 273.353 women and 229.107 men aged between 37-73 years at the time of their assessment visit. Participants are linked to routinely collected records from primary care (GP), hospital records (HES, PEDW and SMR), and death registries (ONS), providing longitudinal information on diagnosis, procedures, and prescriptions for the entire cohort from Scotland, Wales, and England. Routine health records were mapped to the OMOP CDM and represented as a 71.036-dimensional binary vector, indicating whether a concept has been recorded at least once in an individual prior to recruitment. A subset of 15.595 unique concepts, all found in at least 50 individuals, was chosen for model development. Endpoints were defined as the set of PheCodes X[36,54], and after the exclusion of very rare endpoints (recorded in < 100 individuals), 1,883 PheCodes X endpoints were included in the development of the models. Due to the adult population, congenital, developmental, and neonatal endpoints were excluded. For each endpoint, subsequently, time-to-event outcomes were extracted, defined by the first occurrence after recruitment in primary care, hospital or death records. Detailed information on the predictors and endpoints is provided in Supplementary Table 1+2.

While all individuals in the UK Biobank were used to integrate the routine health records, develop the model, and estimate phenome-wide log partial hazards, individuals were excluded from endpoint-specific downstream analysis if they were already diagnosed with a disease (defined by a prior record of the respective endpoint) or are generally not eligible for the specific endpoint (females were excluded from the risk estimation for prostate cancer).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

To externally validate our risk states, we investigate individuals from the All of Us cohort[34], containing information on 229,830 individuals of diverse descent and from minorities historically underrepresented in biomedical research[37]. Because we only use the All of Us cohort for validation, we evaluate the predictive performance for the subset of 1,568 endpoints with at least 100 incident events in the All of Us cohort.

The study adhered to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement for reporting[55]. The completed checklist can be found in the Supplementary Information.

Extraction and preparation of the routine health records

To extract the routine health records of each individual, we first aggregated the linked primary care, hospital records, and mortality records and mapped the aggregated records to the OMOP CDM (mostly SNOMED and RxNorm). Specifically, we used mapping tables provided by the UK Biobank, the OHDSI community, and SNOMED International to map concepts from the provider and country-specific non-standard vocabularies to OMOP standard vocabularies. We restricted the analysis to the domains “Observation”, “Condition”, “Procedure”, “Drug” and “Device”. To reduce the complexity we did not include any laboratory measures. The PheCode X endpoints[36,54] were derived from either mapping directly from ICD-10 (hospital and death records) or mapping from SNOMED to ICD-10 (using the official mapping table) and subsequently to Phecodes X.

Spatial validation and data preprocessing

For model development and testing, we split the data set into 22 spatially separated partitions based on the location of the assessment center at recruitment. We analyzed the data in 22-fold

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

nested cross-validation, setting aside one of the spatially separated partitions as a test set, aggregating the remaining partitions, and randomly selecting 10% of the aggregated data for the validation set. Within each of the 22 cross-validation loops, the individual test set (i.e. the spatially separated partition) remained untouched throughout model development, and the validation set was used to validate the fitting progress and checkpoint selection. All 22 obtained models were then evaluated on their respective test sets. We assumed missing data occurred randomly and performed multiple imputations using chained equations with gradient boosting machines[56,57]. Imputation models were fitted on the training sets and applied to the respective validation and test sets. Continuous variables were standardised; Categorical variables were one-hot encoded.

Development of the phenome-wide risk model

The risk model is a multi-task neural network that uses the binary representations of an individual's prior health records before recruitment to simultaneously predict log partial hazards[58] for a set of 1,883 endpoints. The model consists of three fully connected linear layers with 4,096 hidden units, each with layer normalisation[59], dropout[60], and leaky ReLU activations. The last latent representation serves as a regulariser as it incentives the extraction of robust features for multiple diseases. The model subsequently computes the log partial hazard (the risk state) for each endpoint with an adapted proportional hazards loss[58], resulting in a 1,883-dimensional output representation. The individual losses are averaged and then summed to derive the final loss of the model. We subsequently tuned hyperparameters (via Bayesian Optimization) on train and validation splits over a constrained parameter space, tuning batch size, learning rate, weight decay, number of nodes in the layers of the endpoint heads, number of hidden layers, dropout rates, and size of the output vector of the shared network. The final models were trained with batch size 512 using the Adam optimiser[61] with a

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

learning rate of 0.0006 and weight decay of 0.3, and early stopping tracking of the performance on the validation set. We implemented the model in Python 3.9 using PyTorch 1.11[62] and PyTorch-lightning 1.5.5 (for code availability, see below).

Downstream analysis and performance comparisons

We fitted Cox proportional hazards models[33] (CPH) to derive absolute risk predictions from the endpoint-specific risk states for the individual endpoints. For each endpoint, we developed models with distinct covariate sets: For all endpoints, we investigated age, biological sex, and the risk states from the health records. For cardiovascular endpoints, we additionally investigated predictors from established and guideline-recommended scores for the primary prevention of cardiovascular diseases, the SCORE2, ASCVD, and QRISK3. Model development was repeated independently for each assessment center thus, for each cross-validation split, models were trained on the respective train set, and checkpoints were selected on the respective validation set. For the final evaluation, test set predictions from the spatially separate recruitment centers were aggregated. Harell's C-Index was calculated with the lifelines package[63] by bootstrapping both the aggregated test set and individual assessment centers. Statistical inferences about model differences were based on the distribution of bootstrapped differences in the C-Index; models were considered different whenever the 95% CI of the difference did not overlap cross zero. CPH models were fitted with the CoxPHFitter from the python package lifelines[63] with default parameters and a step size of 0.5, 0.1, or 0.01 to facilitate model convergence. Confidence intervals for all statistical analyses were calculated over 1,000 bootstrapping iterations.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Independent validation in the All Of Us cohort

After mapping the linked health records from All Of Us to the OMOP vocabulary, we transferred the neural networks developed in the UK Biobank to the All Of Us research environment. We then inferred the models to predict the disease-specific risk states for all individuals. Subsequently, we predicted absolute risks with the CPH models developed in the UK Biobank. For baseline comparison with Age and Sex, we developed new CPH models in the All Of Us cohort.

Calculation of record attributions

To determine which records are most important on an individual level, we calculated attributions for the selection of 24 endpoints based on Shapley values. For computational efficiency, we approximated Shapley values via sampling for only 18432 individuals unseen to the model during development[64]. Please see Supplementary Table 9 for the calculated attributions for individuals with and without prior events.

Data availability

UK Biobank data, including all linked routine health records, are publicly available to bona fide researchers upon application at http://www.ukbiobank.ac.uk/using-the-resource/. In this study, primary care data was used following the COPI regulations. The All Of Us cohort data were provided by the All of Us Research Program by permission that can be sought by scientists and the public alike. Currently, however, data access requires affiliation with a US institution. All patient data used throughout this study has been subject to patient consent as covered by the UK Biobank and All Of Us. Detailed information on the predictors and endpoints is presented in Supplementary Tables 1-3.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Code availability

All code developed and used throughout this study has been made open source and is available on GitHub. The code to train the medical history model can be found here: github.com/nebw/medhist, while the code to run analysis on trained models can be found here: github.com/JakobSteinfeldt/MedicalHistoryPhenomeWide.

Acknowledgments

We would like to acknowledge the support of the UK Biobank and the All of Us Research Program in providing access to their respective datasets. This research has been conducted using data from the UK Biobank (application number 51157) and the All of Us Research Program (by S.H. UserID 5703). Both studies have received ethical approval from their respective institutional review boards and have obtained informed consent from participants. We are grateful to the participants who generously contributed their time and data to make this research possible. This project has been funded by the Charité - Universitätsmedizin Berlin and the Einstein Foundation Berlin, throught the Einstein BIH Visiting Fellowship awarded to J.D. The study has been supported by the BMBF-funded Medical Informatics Initiative (HiGHmed, 01ZZ1802A - 01ZZ1802Z) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 437531118 – SFB 1470.

Author contributions

J.S., B.W., T.B., M.P., H.H., C.L., U.L., J.D., and R.E. conceived and designed the project. J.S., B.W., and T.B. implemented models, conducted experiments, and performed data analysis. J.U. and A.V. supported the analysis. S.H. performed the external validation. M.P., S.D., H.H., and C.L. provided methodological support and contributed to the discussion of the

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

results. J.S., B.W., T.B., U.L., J.D. and R.E. wrote and prepared the manuscript. All authors read, revised, and approved the manuscript.

Competing interests

U.L. received grants from Bayer, Novartis, Amgen, consulting fees from Bayer, Sanofi, Amgen, Novartis, Daichy Sankyo, and honoraria from Novartis, Sanofi, Bayer, Amgen, Daichy Sankyo. J.D. received consulting fees from GENinCode UK Ltd, honoraria from Amgen, Boehringer Ingelheim, Merck, Pfizer, Aegerion, Novartis, Sanofi, Takeda, Novo Nordisk, Bayer, and is chief medical advisor to Our Future Health. R.E. received honoraria from Sanofi and consulting fees from Boehringer Ingelheim. All other authors do declare no competing interests.

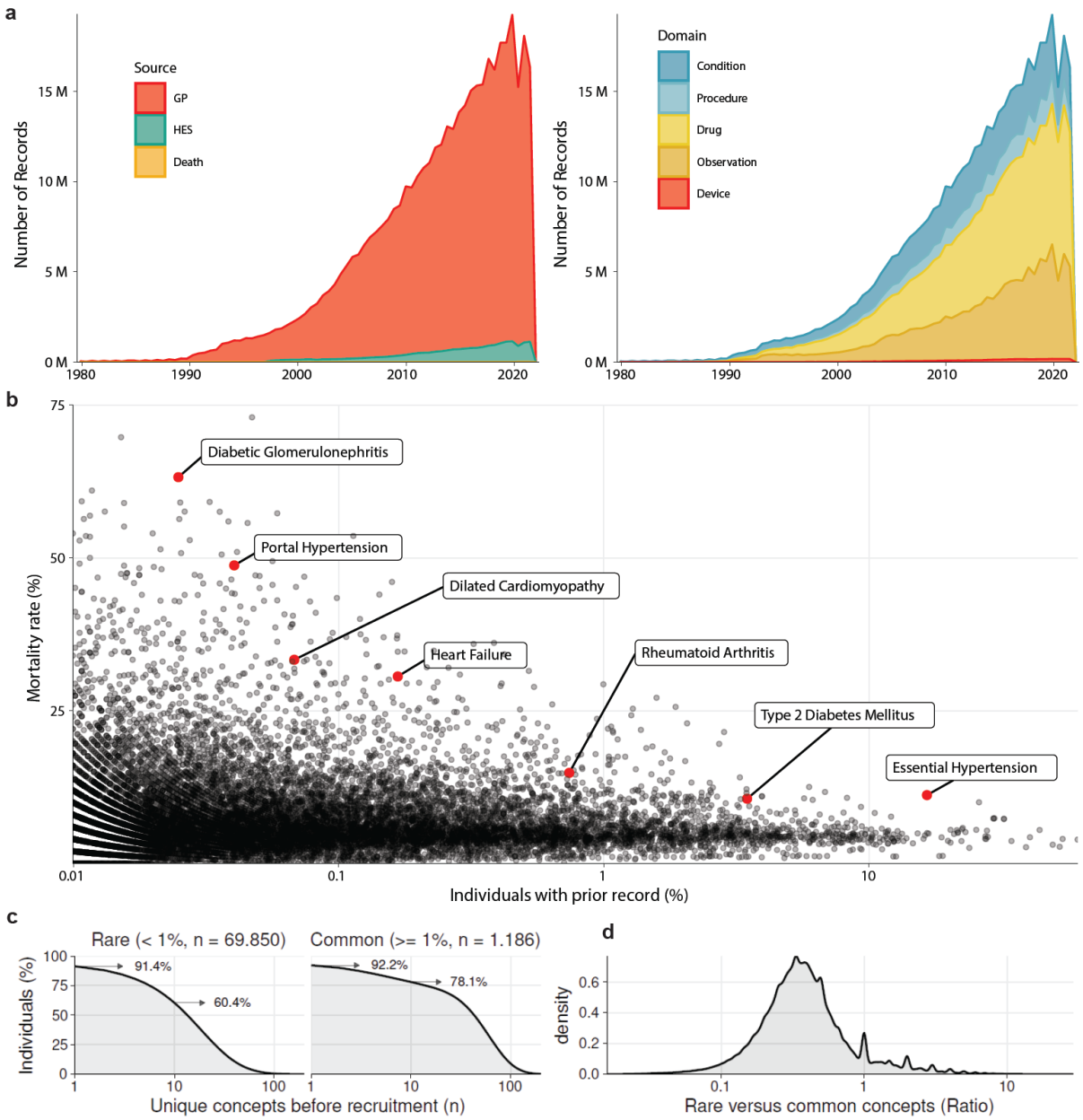
medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Tables

Table 1: The study population.

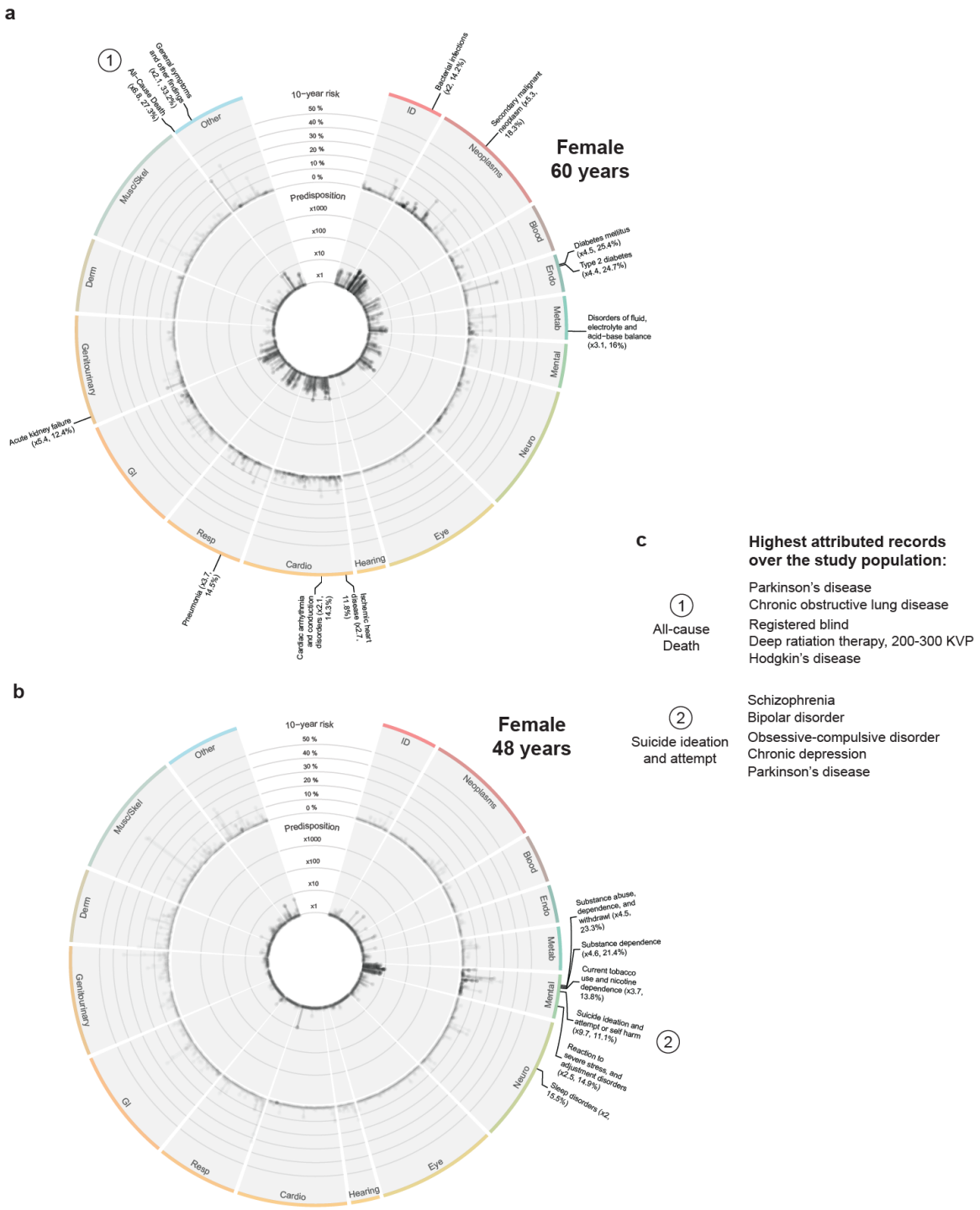
medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Supplementary Figures



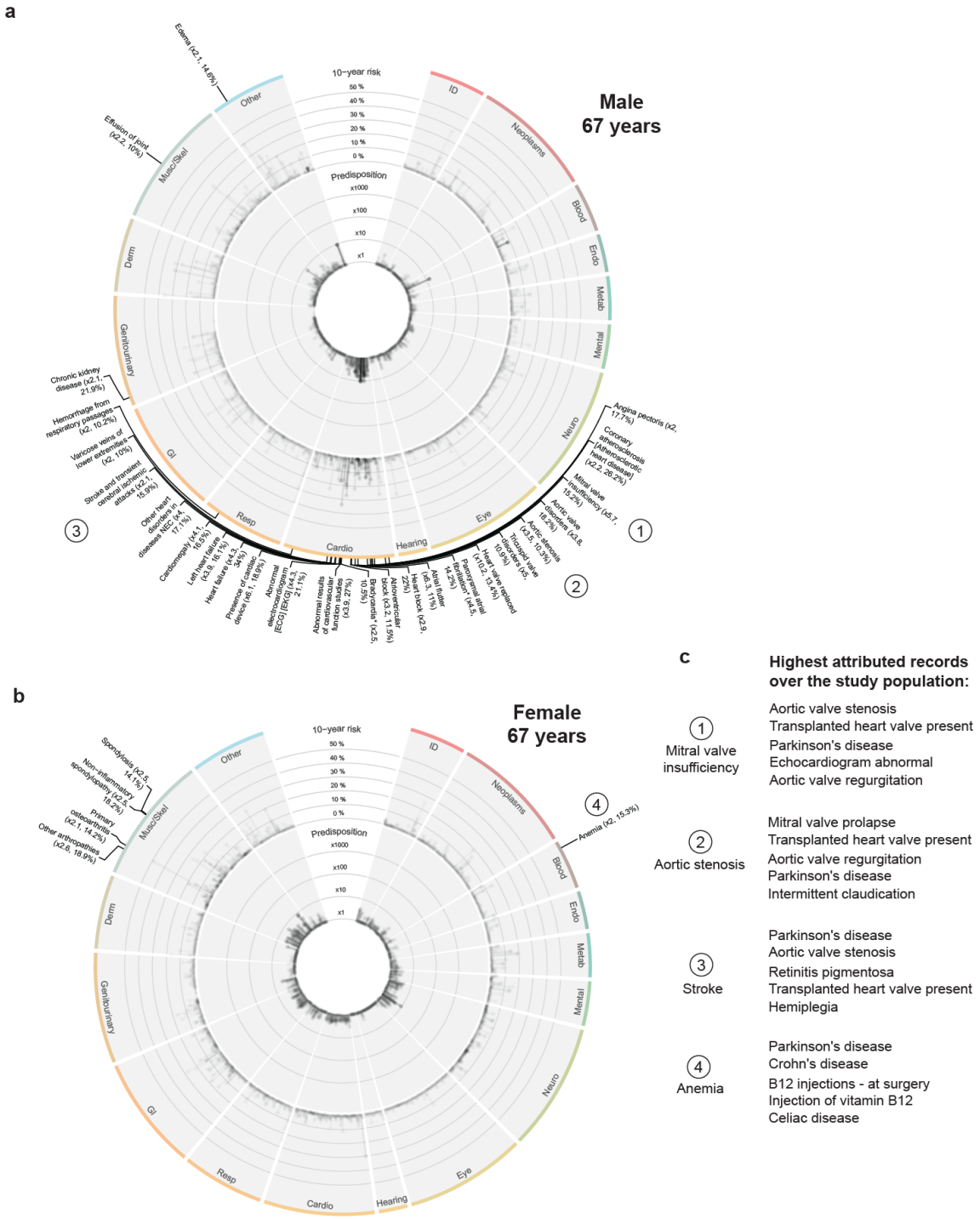
Supplementary Figure 1: Characterisation of routine health records: **a)** Yearly counts of health records stratified by GP, hospital, and death records. **c)** Yearly counts of health records stratified by record domain. **c)** Mortality rate conditional on prior records. Highlighted are high-risk records with gradually increasing frequency. **d)** Percentage of individuals with prior rare or common records. **e)** Ratio of rare and common records per individual.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



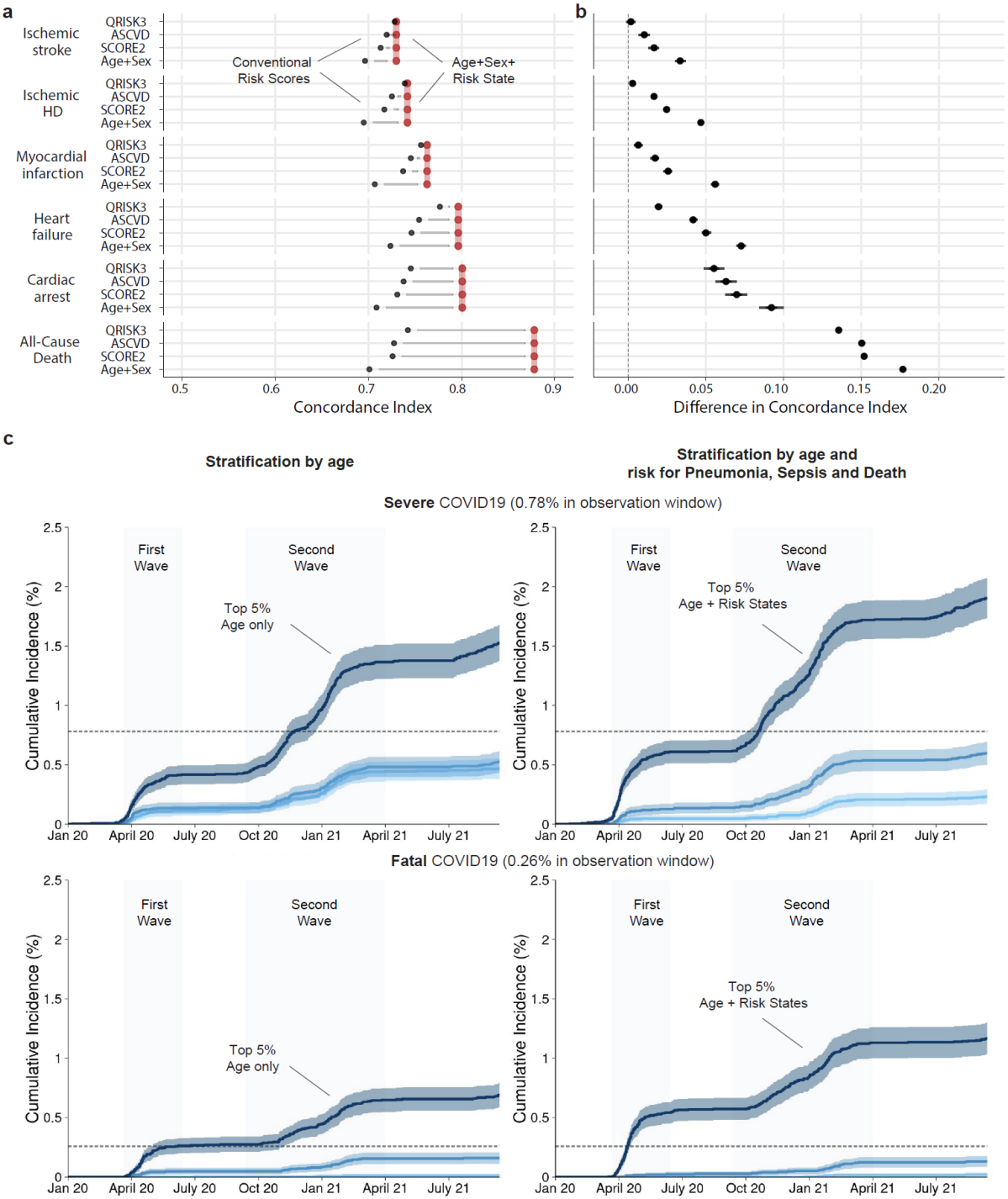
Supplementary Figure 2: Individual predicted phenome-wide risk profiles: a+b) Example of individual predicted phenome-wide risk profile for a 60-year-old (a) and a 48-year-old female (b). Predisposition (10-year risk estimated by Age+Sex+RiskState compared to risk estimated by Age+Sex alone) is displayed in the inner circle, and absolute 10-year risk estimated by Age+Sex+RiskState can be found in the outer circle. Labels indicate endpoints with a high individual predisposition (> 2 times higher than the Age+Sex-based reference estimate) and absolute 10-year risk > 10%. e) Top 5 highest attributed records for selected endpoints.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



Supplementary Figure 3: Individual predicted phenome-wide risk profiles: a+b) Example of individual predicted phenome-wide risk profile for a 67-year-old male (a) and a 67-year-old female (b). Predisposition (10-year risk estimated by Age+Sex+RiskState compared to risk estimated by Age+Sex alone) is displayed in the inner circle, and absolute 10-year risk estimated by Age+Sex+RiskState can be found in the outer circle. Labels indicate endpoints with a high individual predisposition (> 2 times higher than the Age+Sex-based reference estimate) and absolute 10-year risk > 10%. e) Top 5 highest attributed records for selected endpoints.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .



Supplementary Figure 4: Predictions can support cardiovascular disease prevention and the reaction to emerging health threats: a) Discriminatory performances in terms of absolute C-Indices comparing risk scores (Age+Sex, SCORE2, ASCVD, and QRISK as indicated, black point) with the risk model based on Age+Sex+RiskState (red segment). **b)** Direct differences between risk scores (Age+Sex, SCORE2, ASCVD, and QRISK as indicated) and the risk model based on Age+Sex+RiskState in terms of C-index. Dots indicate medians and whiskers extend to the 95% confidence interval for a distribution bootstrapped over 100 iterations. **c)** Estimated cumulative event trajectories of severe (with hospitalization) and fatal (death registry) COVID19 outcomes stratified by the Top, Median and Bottom 5% based on age (left) or risk states of pneumonia, sepsis and all-cause mortality as estimated by Kaplan-Meier analysis.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

References

1 Sindi S, Calov E, Fokkens J, *et al.* The CAIDE Dementia Risk Score App: The development of an evidence-based mobile application to predict the risk of dementia. *Alzheimers Dement* 2015; **1**: 328–33.

2 Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* 2003; **26**: 725–31.

3 Goff David C., Lloyd-Jones Donald M., Bennett Glen, *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* 2014; **129**: S49–73.

4 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017; **357**: j2099.

5 Steyerberg EW, Moons KGM, van der Windt DA, *et al.* Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; **10**: e1001381.

6 Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. *Br Med J* 1975; **2**: 486–9.

7 Danish eHealth Portal. Danish eHealth Portal. Danish eHealth Portal. 2001. https://www.sundhed.dk/borger/service/om-sundheddk/om-organisationen/ehealth-in-denmark/background/.

8 e-Health Record. e-Health Record. e-Health Record. 2005. https://e-estonia.com/solutions/healthcare/e-health-records/.

9 Clalit Research Institute. Clalit Health Services. Clalit Health Services. 2010. http://clalitresearch.org/about-us/our-data/ (accessed 2010).

10 National Electronic Health Record. National Electronic Health Record. National Electronic Health Record. 2011. https://www.ihis.com.sg/nehr/about-nehr.

11 My Health Record. My Health Record. My Health Record. 2016. https://www.myhealthrecord.gov.au/.

12 Wood A, Denholm R, Hollings S, *et al.* Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021; **373**: n826.

13 Rush R. Taking Note. *N Engl J Med* 2019; **381**: 9.

14 Tsang G, Zhou S-M, Xie X. Modeling Large Sparse Data for Feature Selection: Hospital Admission Predictions of the Dementia Patients Using Primary Care Electronic Health Records. *IEEE J Transl Eng Health Med* 2021; **9**: 3000113.

15 Langham J, Stamate D, Wu CA, *et al.* Predicting risk of dementia with machine learning and survival models using routine primary care records. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2021: 3036–42.

16 Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* 2018; **1**: 1–10.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

17 Appelbaum L, Cambronero JP, Stevens JP, *et al.* Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study. *Eur J Cancer* 2021; **143**: 19–30.

18 Sekelj S, Sandler B, Johnston E, *et al.* Detecting undiagnosed atrial fibrillation in UK primary care: Validation of a machine learning prediction algorithm in a retrospective cohort study. *Eur J Prev Cardiol* 2021; **28**: 598–605.

19 Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* 2016; **6**: 26094.

20 Estiri H, Strasser ZH, Klann JG, Naseri P, Wagholikar KB, Murphy SN. Predicting COVID-19 mortality with electronic medical records. *NPJ Digit Med* 2021; **4**: 15.

21 Wu J, Nadarajah R, Raveendra K, Cowan JC, Gale CP. FIND-AF: a widely applicable artificial intelligence algorithm to target systematic screening for atrial fibrillation in older individuals through primary care electronic health records. *Europace* 2022; **24**. DOI:10.1093/europace/euac053.565.

22 Bagheri A, Groenhof TKJ, Veldhuis WB, de Jong PA, Asselbergs FW, Oberski DL. Multimodal Learning for Cardiovascular Risk Prediction using EHR Data. arXiv [cs.LG]. 2020; published online Aug 27. http://arxiv.org/abs/2008.11979.

23 Ben Miled Z, Haas K, Black CM, *et al.* Predicting dementia with routine care EMR data. *Artif Intell Med* 2020; **102**: 101771.

24 Zhao J, Feng Q, Wu P, *et al.* Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Sci Rep* 2019; **9**: 717.

25 Jin B, Che C, Liu Z, Zhang S, Yin X, Wei X. Predicting the Risk of Heart Failure With EHR Sequential Data Modeling. *IEEE Access* undefined 2018; **6**: 9256–61.

26 Hill NR, Ayoubkhani D, McEwan P, *et al.* Predicting atrial fibrillation in primary care using machine learning. *PLoS One* 2019; **14**: e0224582.

27 Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. *JAMA Netw Open* 2020; **3**: e1919396.

28 Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013; **31**: 1102–10.

29 Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat Rev Genet* 2016; **17**: 129–45.

30 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* 2021; **4**: 1–13.

31 Li Y, Rao S, Solares JRA, *et al.* BEHRT: Transformer for Electronic Health Records. *Sci Rep* 2020; **10**: 1–12.

32 Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018; **562**: 203–9.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

33 Cox DR. Regression Models and Life-Tables. *J R Stat Soc Series B Stat Methodol* 1972; **34**: 187–202.

34 All of Us Research Program Investigators, Denny JC, Rutter JL, *et al.* The 'All of Us' Research Program. *N Engl J Med* 2019; **381**: 668–76.

35 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med* 2015; **12**: e1001779.

36 Wu P, Gifford A, Meng X, *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* 2019; **7**: e14325.

37 Ramirez AH, Sulieman L, Schlueter DJ, *et al.* The All of Us Research Program: Data quality, utility, and diversity. *Patterns (N Y)* 2022; **3**: 100570.

38 McDonagh TA, Metra M, Adamo M, *et al.* 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J* 2021; published online Aug 27. DOI:10.1093/eurheartj/ehab368.

39 Finlayson SG, Subbaswamy A, Singh K, *et al.* The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med* 2021; **385**: 283–6.

40 Wong A, Otles E, Donnelly JP, *et al.* External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in hospitalized Patients. *JAMA Intern Med* 2021; **181**: 1065–70.

41 Guo LL, Pfohl SR, Fries J, *et al.* Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci Rep* 2022; **12**: 2726.

42 National Institute for health and Care Excellence (NICE). Cardiovascular disease: risk assessment and reduction, including lipid modification. 2014; published online July 18. https://www.nice.org.uk/guidance/cg181 (accessed Sept 16, 2022).

43 SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J* 2021; **42**: 2439–54.

44 Collins R, Reith C, Emberson J, *et al.* Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet* 2016; **388**: 2532–61.

45 Ference BA, Ginsberg HN, Graham I, *et al.* Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. *Eur Heart J* 2017; **38**: 2459–72.

46 Pozzilli P, Strollo R, Bonora E. One size does not fit all glycemic targets for type 2 diabetes. *J Diabetes Investig* 2014; **5**: 134–41.

47 Fonarow GC, Yancy CW, Hernandez AF, Peterson ED, Spertus JA, Heidenreich PA. Potential impact of optimal implementation of evidence-based heart failure therapies on mortality. *Am Heart J* 2011; **161**: 1024–30.e3.

48 Vaduganathan M, Claggett BL, Jhund PS, *et al.* Estimating lifetime benefits of comprehensive disease-modifying pharmacological therapies in patients with heart failure with reduced ejection fraction: a comparative analysis of three randomised controlled trials. *Lancet* 2020; **396**: 121–8.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted March 16, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

49 Adelson K, Paris J, Horton JR, *et al.* Standardized Criteria for Palliative Care Consultation on a Solid Tumor Oncology Service Reduces Downstream Health Care Use. *J Oncol Pract* 2017; **13**: e431–40.

50 Weissman DE, Meier DE. Identifying patients in need of a palliative care assessment in the hospital setting: a consensus report from the Center to Advance Palliative Care. *J Palliat Med* 2011; **14**: 17–23.

51 Centeno C, Arias-Casais N. Global palliative care: from need to action. Lancet Glob Health. 2019; **7**: e815–6.

52 Vayena E. Value from health data: European opportunity to catalyse progress in digital health. *Lancet* 2021; **397**: 652–3.

53 Fry A, Littlejohns TJ, Sudlow C, *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017; **186**: 1026–34.

54 Wei W-Q, Bastarache LA, Carroll RJ, *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 2017; **12**: e0175508.

55 Moons KGM, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; **162**: W1–73.

56 Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012; **28**: 112–8.

57 miceforest. PyPI. https://pypi.org/project/miceforest/ (accessed July 6, 2022).

58 Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. 2017 https://arxiv.org/pdf/1606.00931.pdf.

59 Ba JL, Kiros JR, Hinton GE. Layer Normalization. arXiv [stat.ML]. 2016; published online July 21. http://arxiv.org/abs/1607.06450.

60 Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* 2014; **15**: 1929–58.

61 Kingma DP, Ba JL. Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations 2015* 2015. DOI:http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503.

62 Paszke A, Chanan G, Lin Z, *et al.* Automatic differentiation in PyTorch. *Advances in Neural Information Processing Systems 30* 2017; : 1–4.

63 lifelines 0.25.8. 2021. https://lifelines.readthedocs.io/en/latest/ (accessed Feb 3, 2021).

64 Castro J, Gómez D, Tejada J. Polynomial calculation of the Shapley value based on sampling. *Comput Oper Res* 2009; **36**: 1726–30.