medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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 and enables the rapid response to emerging health threats

Jakob Steinfeldt[1,2,3,14,15*], Benjamin Wild[4,*], Thore Buergel[4,15*], Maik Pietzner[3,5,11], Julius Upmeier zu Belzen[4], Andre Vauvelle[6], Stefan Hegselmann[7,8], Spiros Denaxas[6,9,10], Harry Hemingway[6,10], Claudia Langenberg[3,5,11], Ulf Landmesser[1,2,12,13,14,§], John Deanfield[15,§], Roland Eils[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 September 26, 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

The COVID-19 pandemic exposed, with few exceptions, a global deficiency in delivering systematic, data-driven guidance to protect citizens and coordinate vaccination programs. At the same time, medical histories are routinely recorded in most healthcare systems and are instantly available for risk assessment. Here, we demonstrate the utility of medical history in determining the risk for 1,883 diseases across clinical specialties and facilitating the rapid response to emerging health threats at the example of COVID-19. We developed a neural network to learn disease-specific risk states from routinely collected health records of 502,460 UK Biobank participants, demonstrating risk stratification for nearly all conditions, and validated this model on 229,830 individuals from the All of US cohort. When integrated into Cox Proportional Hazard Models, we observed significant discriminative improvements over basic demographic predictors for 1,774 (94.3%). After transferring the unmodified risk models to the All of US cohort, the discriminate improvements were replicated for 1,347 (89.8%) of 1,500 investigated endpoints, demonstrating model generalizability across healthcare systems and historically underrepresented groups. We then show that these risk states can be used to identify individuals vulnerable to severe COVID-19 and mortality. Our study demonstrates the currently underused potential of medical history to rapidly respond to emerging health threats by systematically estimating risk for thousands of diseases at once at minimal cost.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 early phase of the COVID-19 pandemic exposed a global deficiency in delivering systematic, data-driven guidance for individual patients and healthcare providers with critical implications for pandemic preparedness. 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]. Importantly, in the early phases of emerging pandemics such as COVID-19, it is necessary to allocate sparse resources, but risk scores to identify vulnerable subpopulations are not available due to the lack of available data.

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, insurance, 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 to guide clinical decision-making with etiological[14–17], diagnostic[18,19], and prognostic research[15,16,20–22]. Existing efforts often extract

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

and leverage known clinical predictors with new methodologies[19], augment them with additionally extracted data modalities such as clinical notes[23], 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[15,24], cardiovascular conditions[23,25] such as heart failure[26] and atrial fibrillation[27,28]. In contrast, phenome-wide association studies (PheWAS) quantifying the associations of genetic variants with comprehensive phenotypic traits are emerging in genetic epidemiology[29,30]. While approaches have been developed to extract information from longitudinal health records[31,32], 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[33] 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 that are not currently preventable, but the early diagnosis has been shown to substantially slow down the 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).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

We evaluated our approach by integrating the endpoint-specific risk states estimated by the neural network in Cox Proportional Hazard models[34], investigating the phenome-wide predictive potential over basic demographic predictors, selected comorbidities and established modifiable risk factors, 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[35], we show that our models can generalize across healthcare systems and populations, including communities historically underrepresented in biomedical research.

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. We then show that the risk states of pneumonia, sepsis & all-cause death can be used to calculate a combined severity risk score using primary and secondary care records available before the global spread of the COVID-19 pandemic. 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[33,36], 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 were followed for a median of 12.6 years, resulting in ~6.2M overall person-years on 1,883 phenome-wide endpoints[37] with ≥ 100 incident events (> 0.02% of

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

individuals having 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[35,38], 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 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 (Supplementary 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[21], they are often associated with high incident event rates (exemplified by the mortality rate in Supplementary 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 (Supplementary Figure 1d).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

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 (with 88.4M parameters) to predict the phenome-wide onset of 1,883 endpoints (Supplementary Table 1) simultaneously (Figure 1a). For comparison, we also include additional comparisons with a linear baseline (with 29.4M parameters, Supplementary Figure 2), demonstrating superior performance at a minimal increase of complexity.

To ensure that our findings are generalizable and transferable, we spatially validate our models in 22 recruitment centers (Figure 1b) across England, Wales, and Scotland. We developed 22 models, each trained on individuals from 21 recruitment centers at recruitment, randomly split into training and validation sets (Figure 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-specificity were respected in all downstream analyses. After development, the models were externally validated in the All of Us cohort[35].

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.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 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 (Figure 2A). At the example of hypertension, we found that fewer than 10% of the individuals had an incident disease diagnosis in the observation window if they were estimated to be in the bottom risk percentile of the medical history, compared to more than 60% if they were estimated to be in the top risk percentile. Subsequently, the incident event ratio between the top and bottom deciles was ~5.23. 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 1,341 of 1,883 endpoints (71.2%), 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 ~ 11.3), ischemic heart disease (Ratio ~ 23.5), or chronic obstructive pulmonary disease (Ratio ~ 65.4). For 230 (12.2%) of the 1,883 conditions, including abdominal aortic aneurysm (Ratio ~ 163.4), 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 542 (28.8%) endpoints, the separation between high and low-risk individuals was smaller (Ratio < 10), which included hypertension (Ratio ~ 5.2) and anemia (Ratio ~ 6.7), often diagnosed earlier in life or precursors for future comorbidities. Notably, the ratios were > 1 for all but one of the 1,883 investigated endpoints, even though all models were developed in spatially segregated assessment centers. To illustrate how high-risk individuals differ from the moderate cases, we also provide additional ratios comparing the top 10% to

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

individuals in the median 20% of the population. The complete list of all endpoints and corresponding statistics can be found in Supplementary 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 exemplify the potential of our approach, among individuals in the top risk decile for heart failure, 8,018 (16.06%) experienced an event, in contrast to 178 (0.35%) individuals in the bottom decile, resulting in a risk ratio of 46.35 (Figure 2A, B, Supplementary Table 4). Consequently, those at high risk of heart failure could be prioritized for echocardiographic screening and, if necessary, prescribed effective guideline-directed medical therapy. Similarly, individuals with a high risk of developing COPD - where the top 10% face over 65 times the risk compared to the bottom 10% - may be considered for spirometry, an approach already established in the CAPTURE trial[39]. If confirmed, they could benefit from interventions such as long-acting bronchodilators. As a third example, a high-risk estimate for less common diseases, such as multiple sclerosis (risk ratio ~ 8.3), could further support referring individuals to a specialist and potentially shorten the often extensive patient journey before a final diagnosis is reached.

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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

and widely unused potential for the systematic risk estimation of disease onset in the general population.

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, selected comorbidities from the Charlson Comorbidity Index[40], and established modifiable risk factors from the AHA ASCVD pooled cohort equation[3]. 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 Supplementary 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 1,774 (94.2%) of the 1,883 investigated endpoints (Figure 3, Supplementary Table 5). For many of these endpoints, the discriminative improvements were considerable (Delta C-Index Q25%: 0.094, Q50: 0.116, Q75: 0.141). We found significant improvements for 23 of the highlighted subset of 24 endpoints (indicated in Figure 2A), with the largest increases for the prediction of back pain (Delta C-Index: +0.238 (CI 0.236, 0.241)), suicide attempts (Delta C-Index: +0.224 (CI 0.213, 0.235)), psoriasis (Delta C-Index: +0.171 (CI 0.161, 0.178)), all-cause mortality (Delta C-Index: +0.171 (0.169, 0.174)) and chronic obstructive pulmonary disease (Delta C-Index: +0.156 (0.151, 0.159)). In contrast, we did not find significant improvements in the prediction of 86 (4.6%) of the 1,883 endpoints, including, e.g., Parkinson's disease (Delta C-Index: -0.006 (CI -0.013, 0)) or even deteriorations in the

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

prediction of 23 (1.2%) of the endpoints, including neoplasm like cervical cancer (Delta C-Index: -0.025 (-0.059, -0.004)) and gastrointestinal diseases as chronic hepatitis (Delta C-Index: -0.032 (-0.064, -0.007)).

We also present a comparison between our approach and the Charlson Comorbidity Index's[40] predictive performance, both of which can be automated. Additionally, we compare our method to the well-established ASCVD predictors, which are widely accessible but require an additional blood draw. Notably, incorporating the comorbidities from the Charlson Comorbidity Index enhances the discriminative capacity beyond age and sex; however, adding medical history proves to be significantly more effective in improving performance (Supplementary Figure 3, Supplementary Table 5). Likewise, while supplementing ASCVD predictors to age and sex augments the performance for most endpoints, it remains inferior to the combination of age, sex, and medical history alone. Incorporating the medical history alongside the comorbidities or ASCVD predictors further improves the predictive performance for the vast majority of endpoints (AgeSex+Comorbidities augmented by the MedicalHistory: +1726/1883 (91.7%), ASCVD+MedicalHistory: +1727/1883 (91.7%), demonstrating complementary nature of these information sources.

For illustration, we also present individual phenome-wide risk profiles (Figure 3c, Supplementary Figure 4a+b and 5a+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 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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

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 (Supplementary Figure 5b), 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.

Importantly, the model performance is robust to the removal of recent information, indicating that the model effectively incorporates both the individuals' long-term medical history and recent interactions with the healthcare system in order to predict future disease onset (Supplementary Figure 6). We provide Shapley attributions[41] for the most important records (Fig 3d, Suppl. Figure 4c, Suppl. Figure 5c) and all records for the 24 highlighted endpoints (Supplementary Table 9) in the study population, enhancing the interpretability of our findings.

These findings indicate that health records contain substantial predictive information over established predictors for the majority of disease endpoints from across clinical specialties.

Predictive models can generalize 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 generalizable information[42–44].

To better understand 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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

over the baseline model (age and biological sex only) for 1,347 (85.9%) of the 1,568 investigated endpoints with at least 100 incident events (Figure 4a, Supplementary Table 8), replicating 1347/1500 (89.8%) of all significant improvements in the UK Biobank (Figure 4b, Supplementary Table 8). Generally, larger improvements in the UK Biobank were replicated in the All of Us cohort. It's noteworthy that smaller improvements in the UK Biobank often corresponded to proportionately larger improvements in All of Us, while larger improvements in the UK Biobank were attenuated in All of Us (Figure 4c).

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.

To illustrate these improvements further, we replicated significant improvements for all of the 24 a priori selected endpoints, with improvements ranging from modest for hypertension (Delta C-Index: +0.021 (0.016, 0.024)) and Parkinson's disease (Delta C-Index: +0.035 (0.021, 0.05)) to substantial for, e.g., All-Cause Death (Delta C-Index: +0.116 (0.104, 0.127), Pulmonary embolism (Delta C-Index: +0.125 (0.112, 0.137)), and Cardiac arrest (Delta C-Index: +0.176 (0.146, 0.206)) (Figure 4f, g and Supplementary Table 8). Only for a subset of 54 (3.44%) significantly improved endpoints in the UK Biobank, the discriminative performance in All Of Us deteriorated significantly upon transferring the pre-trained medical history risk model and integrating the information beyond age and biological sex alone,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

including hepatitis (Delta C-Index: -0.226 (-0.251, -0.2)), substance abuse (Delta C-Index: -0.037 (-0.05, -0.026)) and osteoporosis (Delta C-Index: -0.015 (-0.021, -0.008)).

Taken together, our findings suggest that predictive models based on medical history can generalize across health systems and are robust to 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 immediate clinical utility, we focused on the primary prevention of cardiovascular disease and the response to newly emerging health threats at the example of COVID-19.

Risk scores are well established in the primary prevention of cardiovascular events and have been recommended to guide preventive lipid-lowering interventions[45]. While cardiovascular predictors are accessible at a low cost, dedicated visits and resources from 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[46], 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 (Figure 5a, Suppl. Table 7): we found comparable C-Indices with differences +0.001 (CI -0.002, 0.005) for ischemic stroke, +0.002 (CI 0.002, 0.005) for ischemic heart disease and +0.006 (CI 0.003, 0.009) for myocardial infarction compared with the comprehensive QRISK3 score. It is noteworthy that these discriminative improvements are

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

substantially better for later-stage diseases, including heart failure (+0.018 (CI 0.015, 0.021)), cardiac arrest (+0.05 (CI 0.042, 0.059)), and all-cause mortality (+0.13 (CI 0.128, 0.132)) when prior health records are considered.

To further illustrate potential utility, we look at newly emerging pathogenic health threats, where rapid and reliable risk stratification is required to protect high-risk groups and prioritize preventive interventions. We investigated how our phenome-wide risk states could have been used in the context of 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 have helped 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) 0.597 (CI 0.591, 0.604) → C-Index (age + COVID-19 severity risk score) 0.647 (CI 0.641, 0.654); Fatal: C-Index (age) 0.720 (CI 0.710, 0.731) → C-Index (age + COVID-19 severity risk score) 0.780 (CI 0.772, 0.789). These discriminative improvements translate into higher cumulative incidence in the Top 5% population compared to age alone (Suppl. Figure 6C, age (left), COVID-19 severity score (right), severe COVID-19 (top),

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

fatal COVID-19 (bottom)): In the top 5% of the age-based risk group (~ 79 (IQR 77, 81) years old), 0.42% (CI 0.34%, 0.5%, n=105) have been hospitalized, 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 hospitalized, 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 0.64% (CI 0.54%, 0.74%, n=160) had been hospitalized, and 0.32% (0.25%, 0.39%, n=80) had died, while by the end of the second wave, 1.74% (CI 1.57%, 1.9%, n=436) had been hospitalized and 0.68% (0.58%, 0.79%, n=172) had died.

In summary, our findings illustrate the clinical utility of medical history for primary prevention of cardiovascular diseases and the rapid response to emerging health threats.

Discussion

Current clinical practice lacks systematic, data-driven guidance for individuals and care providers. Our study demonstrated for the first time that medical history can systematically inform on phenome-wide risk across clinical specialties, as shown in the British UK Biobank cohort. Subsequently, we show that these risk states can be repurposed to identify individuals vulnerable to severe COVID-19 and mortality. Importantly, we found significant improvements in the discriminated performance for the vast majority of disease endpoints, of which almost 90% could be replicated in the US All of US cohort. Our results indicated utility beyond age, sex, selected comorbidities, and established cardiovascular risk factors commonly considered in clinical practice for preventable diseases, treatable diseases, and

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

diseases without existing risk stratification tools. We anticipate that our approach has the potential to facilitate population health at scale.

Designed for outpatient settings and focused on patients without acute complaints, our approach identifies incident disease onset from early (e.g., hypertension) and later (e.g., bypass surgery) health system contacts. We identified three primary scenarios of potential utility: Firstly, medical history can be exploited in diseases that are preventable with effective interventions, such as the prescription of lipid-lowering medication for primary prevention of coronary heart disease[45]. Lowering LDL cholesterol in 10,000 individuals at increased risk by 2mmol/L with atorvastatin 40mg daily (~2€ per month) for 5 years is estimated to prevent 500 vascular events, reducing the individual relative risk by more than a third[47,48]. Secondly, in conditions that are not preventable anymore individuals can benefit from early detection and treatment, like in type 2 diabetes or systolic heart failure. 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[49]. For a 55-year-old male, this translated into an estimated 8.3 additional years free from cardiovascular death or readmission for heart failure. Lastly, in cases where outcomes are neither preventable nor treatable, estimates of prospective individual risk may be of high importance for personal decisions or the 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[50,51]. Multiple studies have shown that palliative care services can improve patients' symptoms and life quality and may even increase survival[52]. Overall, our approach could facilitate the identification of high-risk

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

populations for specific screening programs, potentially improving the value of national health programs.

Importantly, our approach, based on routine health records, shows large discriminative improvements for the majority of diseases compared with conventionally tested biomarkers[53–55] and can generalize across diverse health systems, populations, and ethnicities. However, we also see that including the medical history over age and sex deteriorated the performance for a subset of 1.2% (UK Biobank) and 4.9% (All Of Us cohort), respectively. Three central challenges remain: First, health records, being products of interactions with the medical system, are subject to biological, procedural, and socio-economic biases[56], as well as being dependent on the evolving nature of medical knowledge and policies. Second, research cohorts often comprise healthier individuals with lower disease prevalence than the general population[57], potentially leading to an underestimation of absolute risks. The third challenge concerns ensuring the interpretability of our approach on such complex data. Our approach provided unique insights into how the model used patients' medical history to make risk predictions. The Shapley value attributions highlighted features the model found most informative for inference on both individual and population levels. These attributions are reflective of the model's decision-making process, and while they aligned with our clinical understanding, they should not replace clinician judgment or other forms of evidence. As we refine and deploy this approach, we must remain vigilant in evaluating its performance and understanding the interpretational limitations. 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 better understand how to overcome these challenges.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

Our study presents the first systematic approach to simultaneous risk stratification for thousands of diseases across clinical specialties based on readily available medical history. These risk states can then be used to rapidly respond to emerging health threats such as COVID-19. Our findings demonstrate the potential to link clinical practice with already collected data 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 September 26, 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 analyzed 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[37,58], 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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

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).

To externally validate our risk states, we investigate individuals from the All of Us cohort[35], containing information on 229,830 individuals of diverse descent and from minorities historically underrepresented in biomedical research[38]. 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[59]. 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[37,58] 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.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

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 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[60,61]. Imputation models were fitted on the training sets and applied to the respective validation and test sets. Continuous variables were standardized; 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[62] 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[63], dropout[64], and leaky ReLU activations. The last latent representation serves as a regulariser as it incentives the extraction of robust features for multiple diseases. For comparison, we also benchmarked the linear version of our model with 29.4M instead of 88.4M parameters (see Suppl. Figure 2). The model subsequently computes the log partial hazard (the risk state) for each endpoint with an adapted proportional hazard loss[62], 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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

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[65] with a 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[66] and PyTorch-lightning 1.5.5 (for code availability, see below). The training of a single model on an NVIDIA A100 GPU node for 18 epochs required approximately 11 hours, equating to the emission of approximately 1.08 kg CO2 eq, 4.36 km driven by an average ICE car or 0.54 kgs of coal burned as calculated by the mlco2 calculator[67]. The external validation of these models, conducted within the All of Us cloud computing environment and including data preprocessing, inference, and evaluation, incurred a total compute cost of approximately 150 USD.

Downstream analysis and performance comparisons

We fitted Cox proportional hazards models[34] (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[68] was calculated

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

with the lifelines package[69] by bootstrapping both the aggregated test set and individual assessment centers. The C-Index is a measure of rank correlation that quantifies the agreement between predicted and observed outcomes. It ranges between 0.5 (no better than random prediction) to 1 (perfect prediction). Statistical inferences about model differences were based on the distribution of bootstrapped differences in the C-Index; models were considered different whenever the Bonferroni-corrected 95% CI of the difference did not overlap cross zero, to account for multiple testing. CPH models were fitted with the CoxPHFitter from the Python package lifelines[69] 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.

Response to emerging health threats

We retrained our models using data limited to records until the end of December 2019, keeping the setting (in particular time zero for training) unchanged. Using these updated models, we then predicted the risk states using all data available at the end of 2019, just as the first cases of COVID-19 were reported. We then manually selected specific risk states associated with pneumonia, sepsis, and all-cause mortality to create an unweighted COVID-19 severity risk score. This risk score was subsequently tested against age for the identification of incident severe and fatal COVID-19 cases.

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 used the models to predict the disease-specific risk states for all individuals. Subsequently, we predicted absolute risks with the CPH models developed in the

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

UK Biobank. Finally, we calculated the mean of the predictions from the models for each individual and disease. For baseline comparison with Age and Sex, we fitted 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 17,236 individuals unseen to the model during development[41]. Please refer to Supplementary Table 9 for the aggregated attributions from individuals without prior events. Shapley values in the table are provided in two forms: averaged (so called local attributions to quantify importance for affected individuals) and summed (global attributions to quantify importance for population ranking). The average Shapley attributions, presented in the main text and figures, closely reflect our understanding of importance for affected individuals.

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 September 26, 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 through 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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

the 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, and 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 the 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 September 26, 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

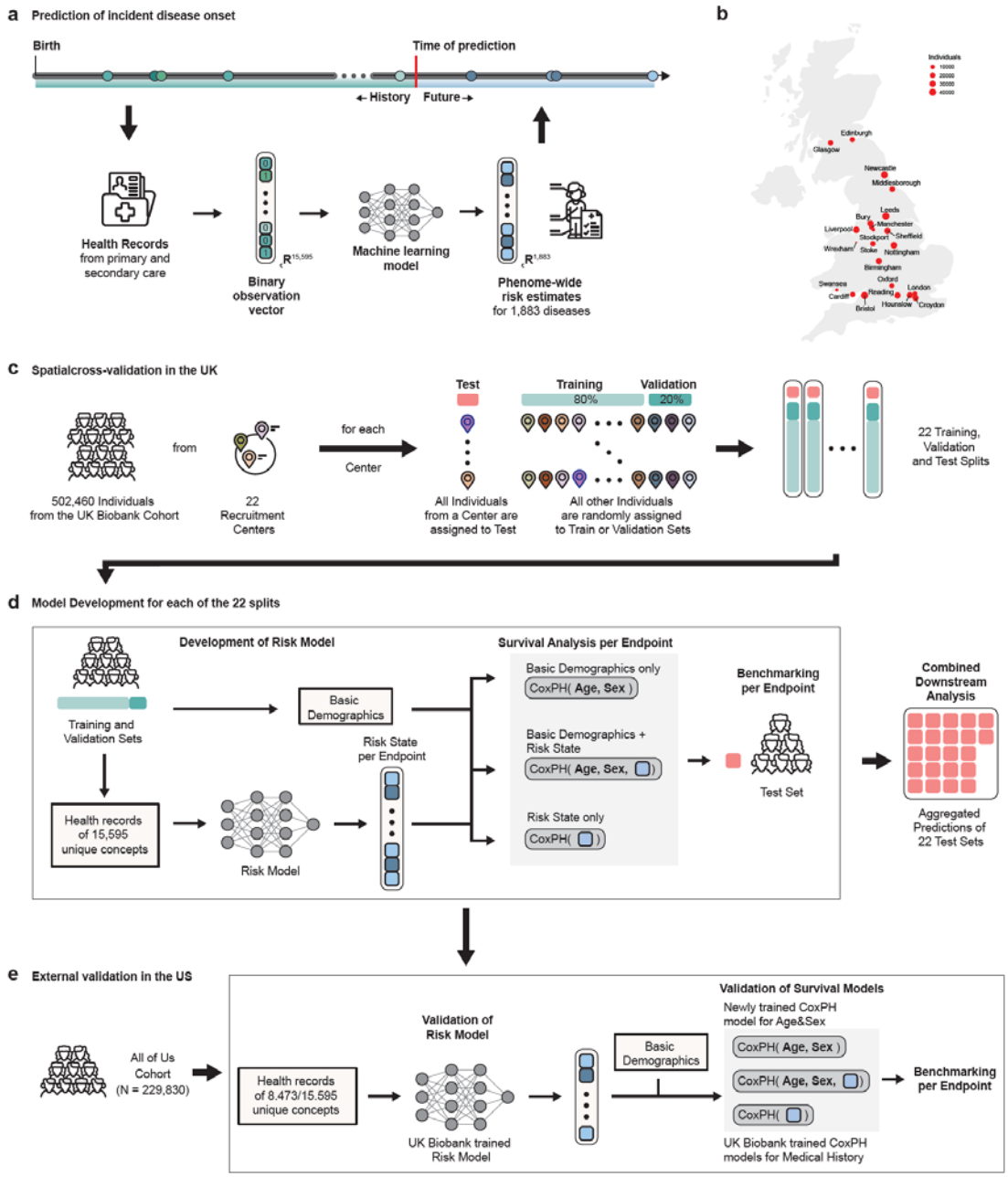
Table 1: The study population.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

Figures

Figure 1: Overview of 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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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.

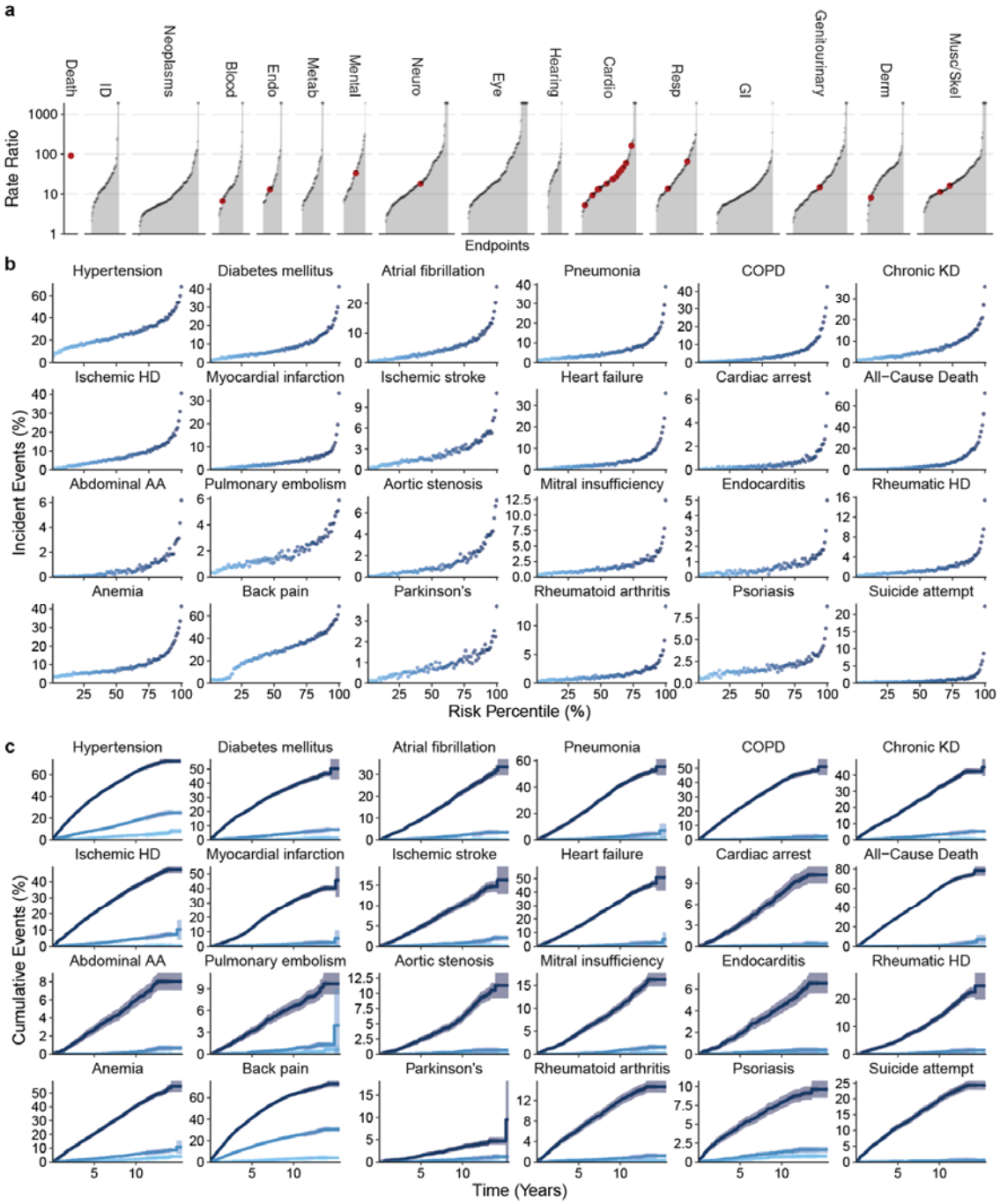
It is made available under a CC-BY 4.0 International license .

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 September 26, 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 but one of the 1,883 investigated endpoints. Red dots indicate 24 selected endpoints detailed in Fig 2b. To illustrate, 1,198 (2.39%) individuals in the top risk decile for cardiac arrest experienced an

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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.

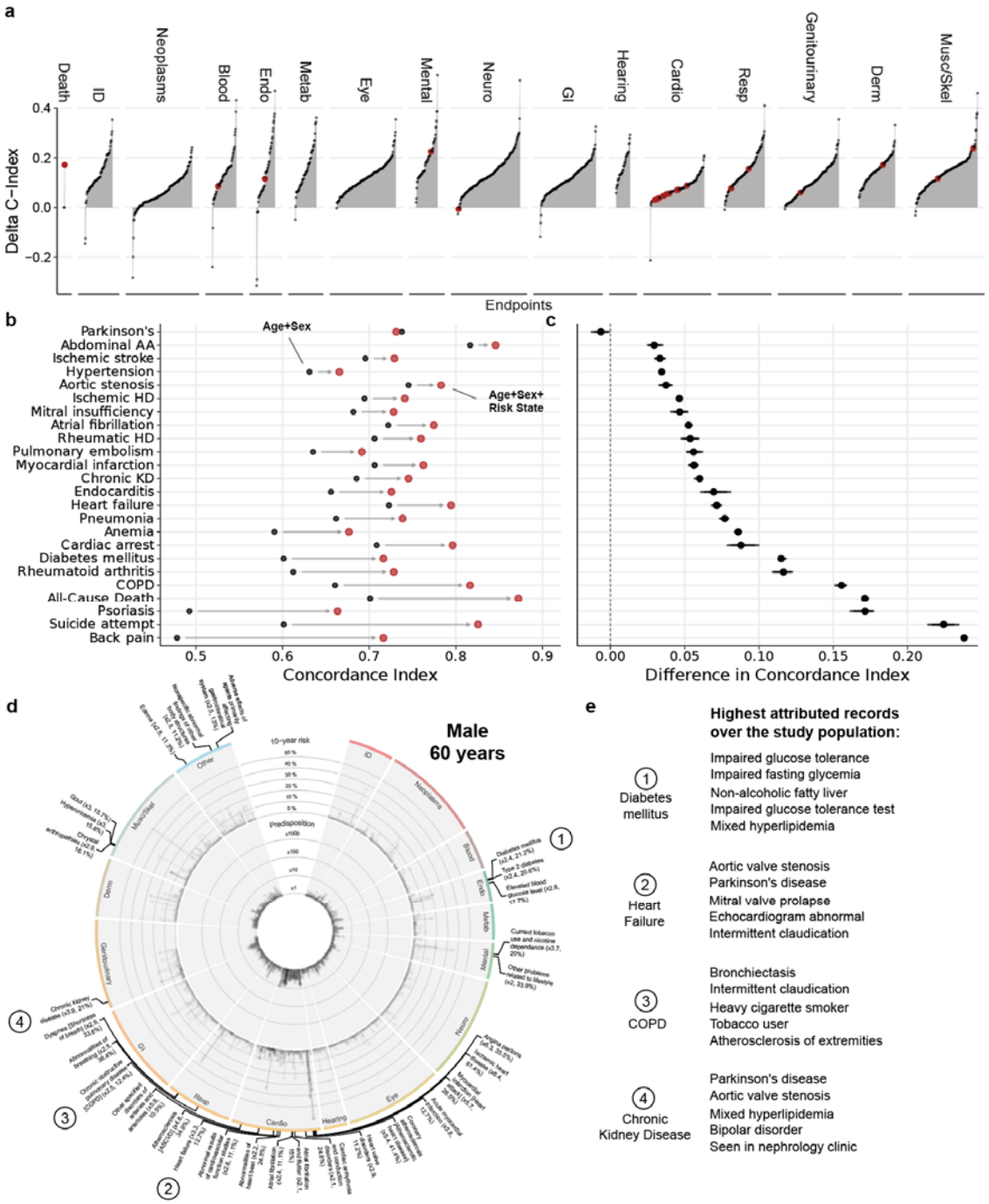
It is made available under a CC-BY 4.0 International license .

event compared with only 30 (0.06%) in the bottom decile, with a risk ratio of 39.93. **b)** Incident event rates for each medical history risk percentile (if medical history was available) for a selection of 24 endpoints. **c)** Cumulative event rates for the Top 1%, median, and Bottom 1% of risk percentiles in b) 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 September 26, 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+MedicalHistory for all 1,883 endpoints. We found significant improvements over the baseline model (Age+Sex, age, and biological sex only)

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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.

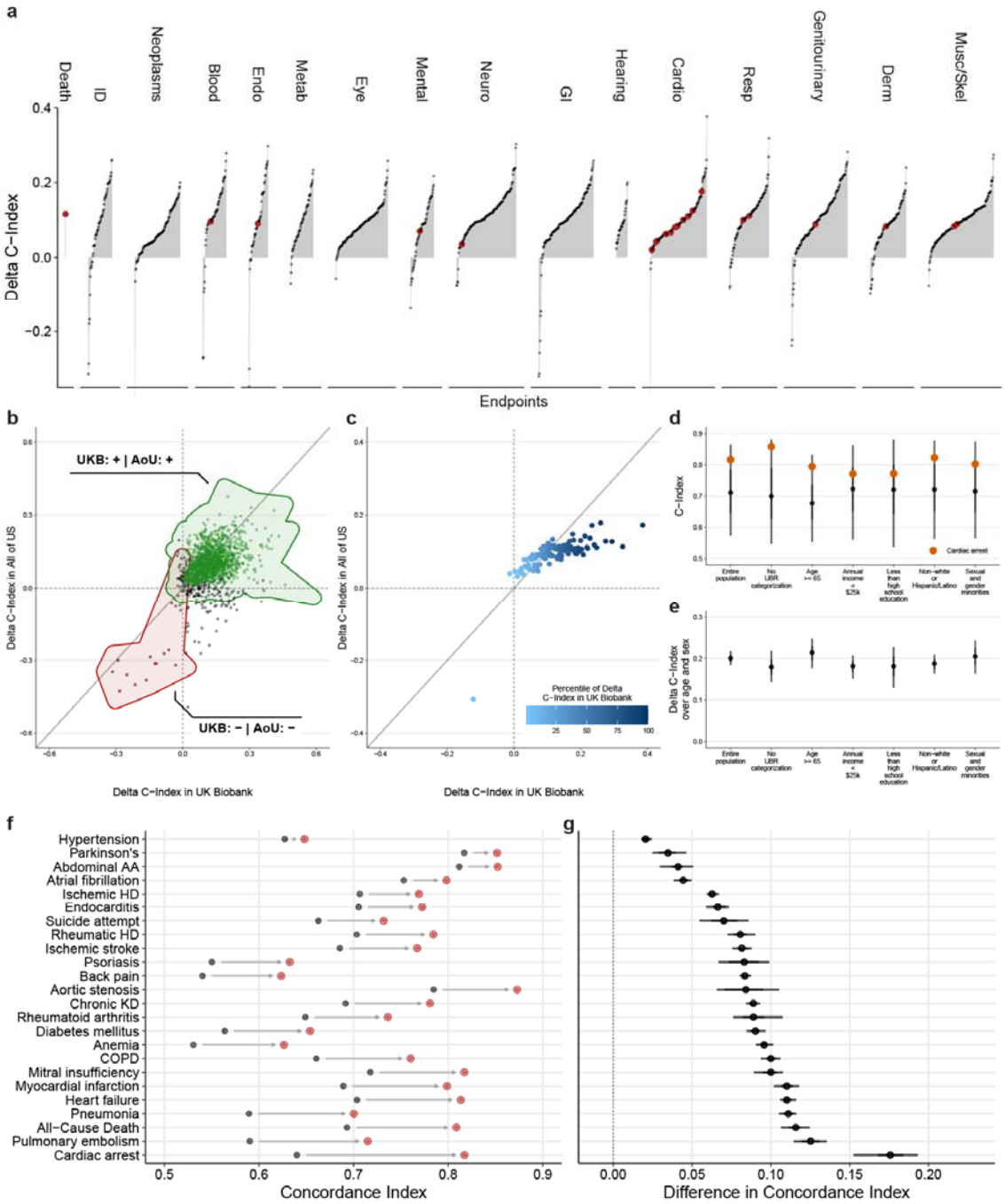
It is made available under a CC-BY 4.0 International license .

for 1774 (94.2%) 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 Bonferroni-corrected 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 September 26, 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 healthcare 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 1.347 (85.9%) of the 1.568 investigated

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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.

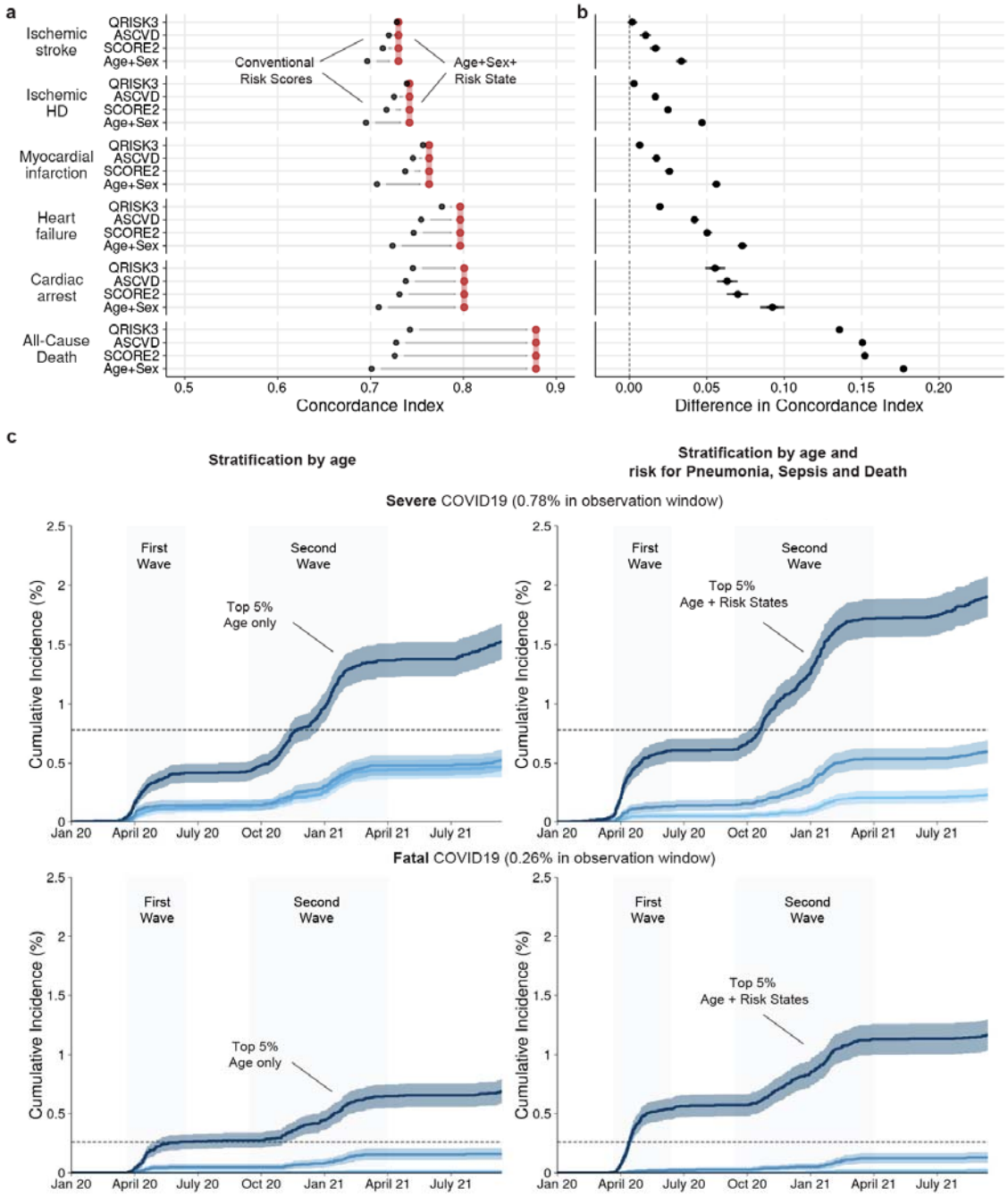
It is made available under a CC-BY 4.0 International license .

endpoints. **b)** Direct comparison of the absolute C-Index in the UK Biobank (x-axis) and the All Of Us cohort (y-axis). Significant improvements can be replicated for 1347 (89.8%) of 1500 endpoints in the All Of Us cohort. **c)** Comparison of mean delta C-Index per delta percentile (derived from the UK Biobank from the 1.568 endpoints available in All Of Us). Improvements in the All Of Us cohort are consistent with the UK Biobank cohort: Small improvements in the UK Biobank tend to be larger in All Of Us, while large improvements in the UK Biobank tend to be attenuated in All Of Us. **d)** Distribution of C-Indices for the 1.568 investigated endpoints stratified by communities historically underrepresented in biomedical research (UPD)[70]. **e)** For the same groups, confidence intervals for the additive performance as measured by the C-Index compared to the baseline model. **f)** 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 points) for a selection of 24 endpoints. **g)** The differences in C-index for the same models.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 5: Predictions can support cardiovascular disease prevention and the response 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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

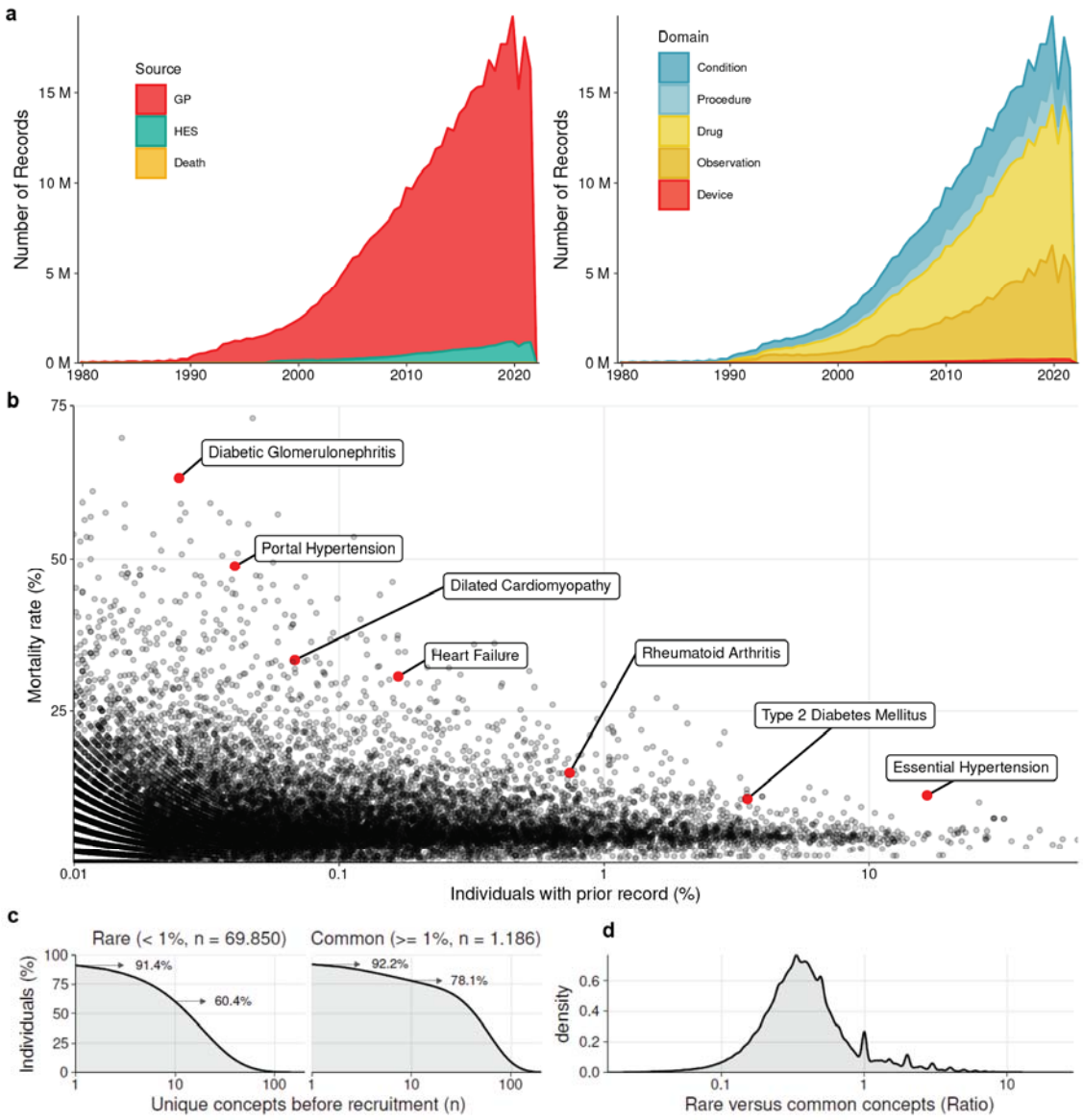
(Age+Sex, SCORE2, ASCVD, and QRISK as indicated) and the risk model based on Age+Sex+RiskState in terms of C-index. Dots indicate medians and whiskers extend to the Bonferroni-corrected 95% confidence interval for a distribution bootstrapped over 100 iterations. **c)** Estimated cumulative event trajectories of severe (with hospitalization) and fatal (death registry) COVID-19 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 September 26, 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.

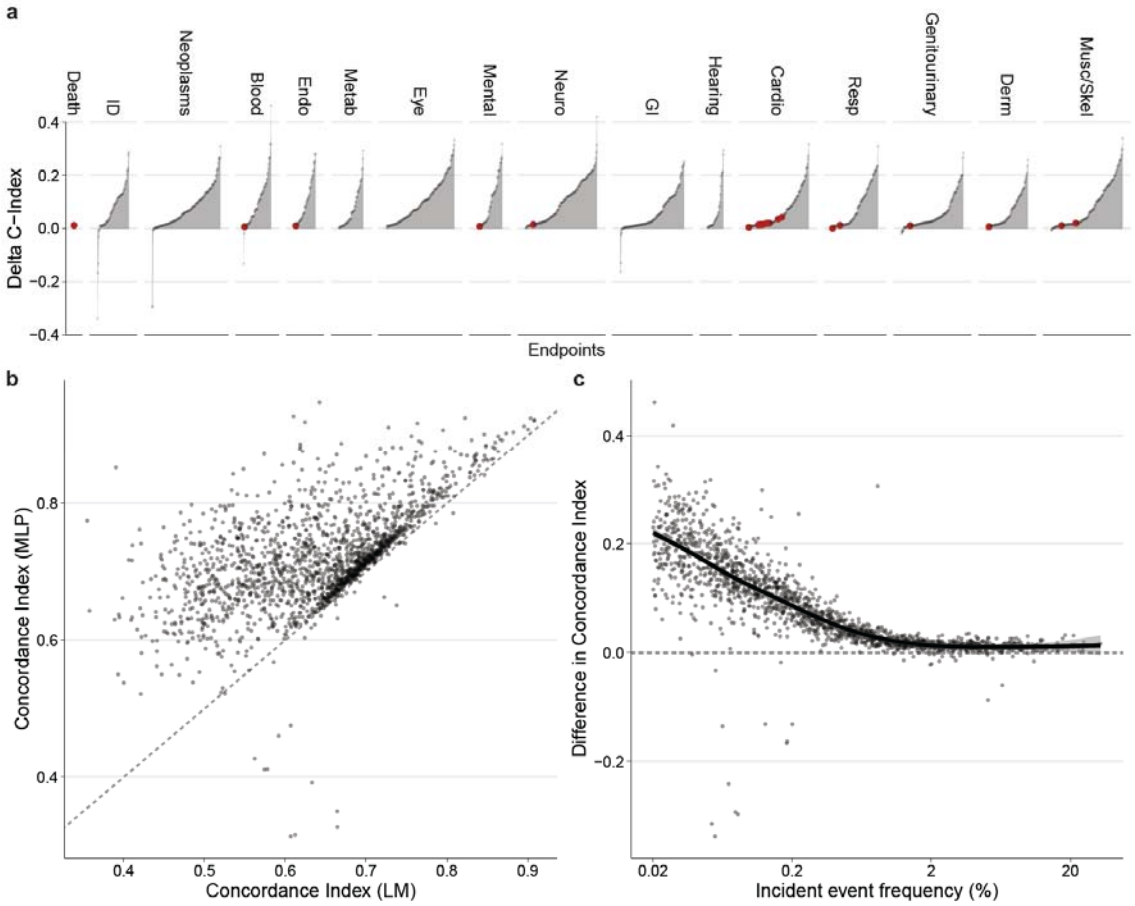
medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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: Comparison of our approach with the linear baseline



a) Differences in discriminatory performance quantified by the C-Index between the linear baseline risk model (LM) and our approach with a multi-layer perceptron (MLP) for all 1,883 endpoints. We found significant improvements for 1815 (96.3%), nonsignificant differences for 42 (2.2%), and significantly worse results for 26 (1.4%) of the 1,883 investigated endpoints. **b)** Direct comparison of the absolute C-Index for the linear model (LM, x-axis) and our approach (MLP, y-axis). **c)** Improvement of our approach (y-axis) by the incident event frequency of the endpoint. The improvements over the linear model are largest for rarer and smallest for common endpoints.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

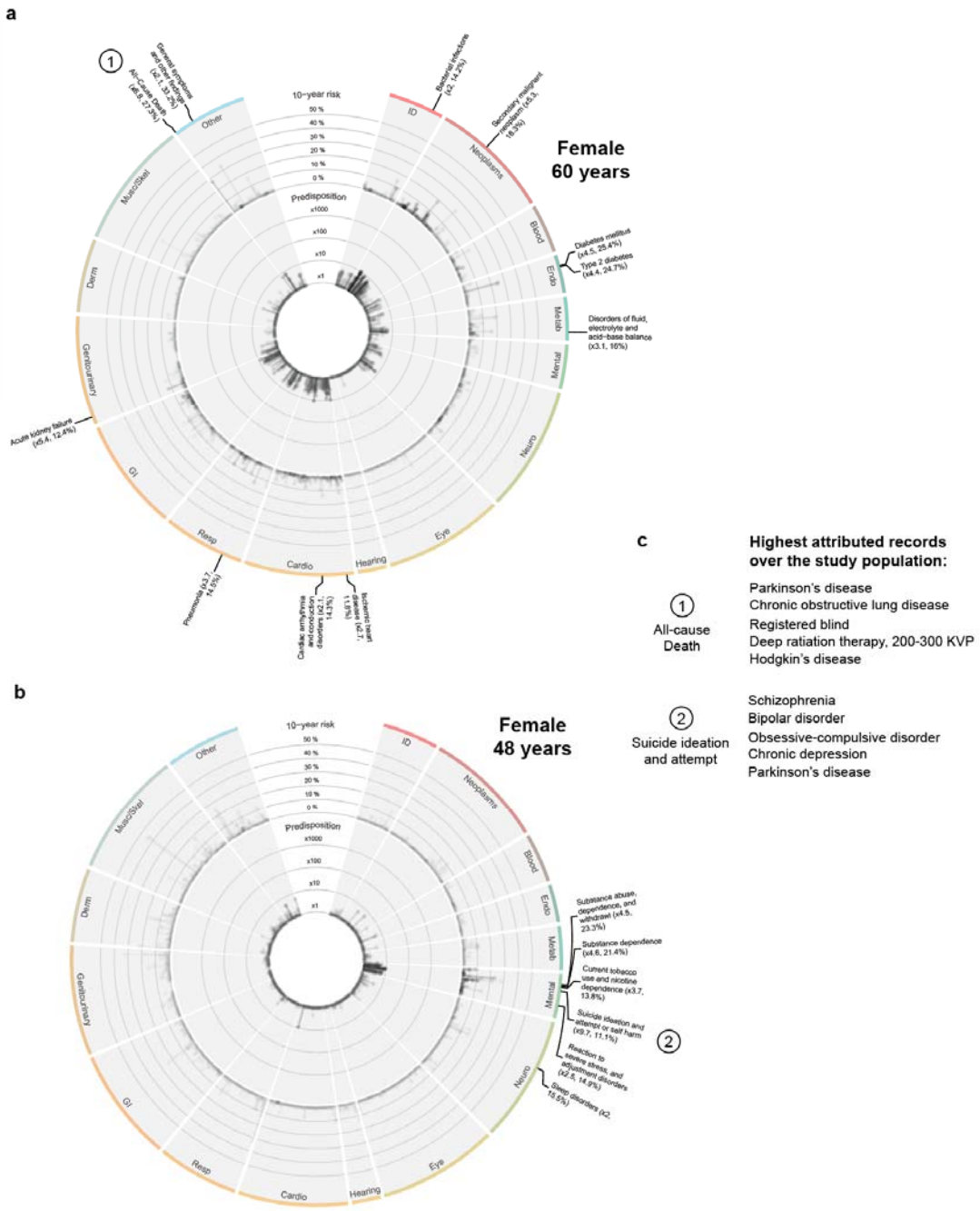
Supplementary Figure 3: Comparison with established comorbidities and risk factors

a) Differences in discriminatory performance quantified by the C-Index between CPH models trained on Age+Sex+Comorbidities from the Charlson Comorbidity Index (CCI) compared with the added medical history for all 1,883 endpoints. We found significant improvements for 1726 (91.7%) of the 1,883 investigated endpoints. **a)** Differences between CPH models trained on ASCVD risk factors only compared with the added medical history for all 1,883 endpoints. We found significant improvements for 1727 (91.7%) of the 1,883 investigated endpoints. Red dots indicate the 24 highlighted endpoints.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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: 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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

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 September 26, 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 5: 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

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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.

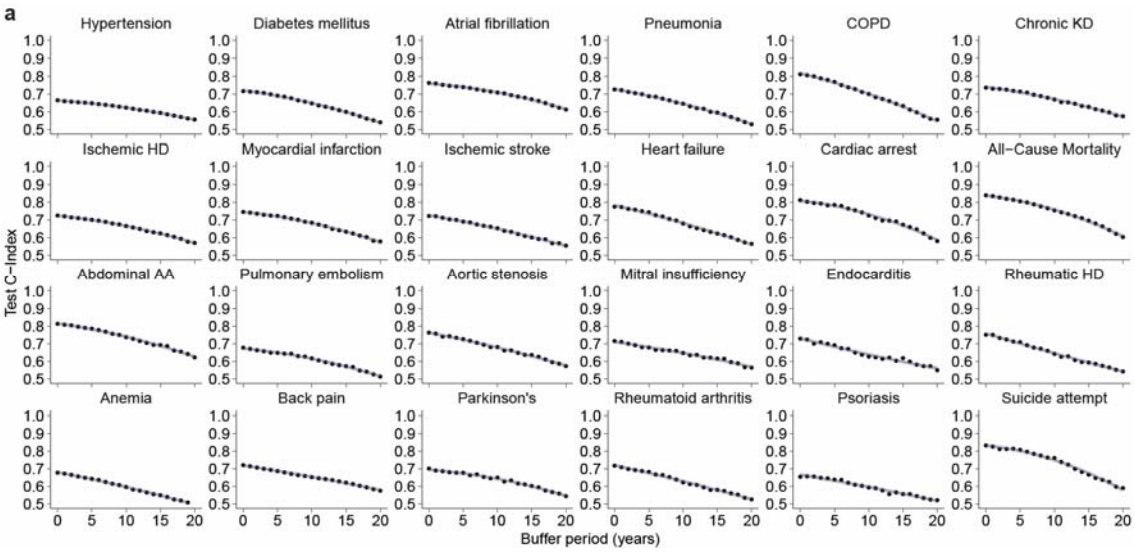
It is made available under a CC-BY 4.0 International license .

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 September 26, 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 6: Stability of model performance with the removal of recent information



Discrimination performance of the medical history, measured in terms of absolute C-Indices, subjected to progressive data removal deprivation. For each time point, a model is trained and evaluated using a dataset from which records documented within a specified time frame prior to recruitment - denoted on the 'Buffer Period (Years)' axis - have been removed. The performance gradually declines in performance across all endpoints, indicating that the model effectively incorporates both the individuals' long-term medical history and recent interactions with the healthcare system in order to predict future disease onset.

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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. *et al.* The CAIDE Dementia Risk Score App: The development of an evidence-based mobile application to predict the risk of dementia. *Alzheimers. Dement.* **1**, 328–333 (2015).
2. Lindström, J. & Tuomilehto, J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* **26**, 725–731 (2003).
3. Goff David C. *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* **129**, S49–S73 (2014).
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* **357**, j2099 (2017).
5. Steyerberg, E. W. *et al.* Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* **10**, e1001381 (2013).
6. Hampton, J. R., Harrison, M. J., Mitchell, J. R., Prichard, J. S. & Seymour, C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. *Br. Med. J.* **2**, 486–489 (1975).
7. Danish eHealth Portal. Danish eHealth Portal. *Danish eHealth Portal* https://www.sundhed.dk/borger/service/om-sundheddk/om-organisationen/ehealth-in-denmark/background/ (2001).
8. e-Health Record. e-Health Record. *e-Health Record* https://e-estonia.com/solutions/healthcare/e-health-records/ (2005).
9. Clalit Research Institute. Clalit Health Services. *Clalit Health Services* (2010).
10. National Electronic Health Record. National Electronic Health Record. *National Electronic Health Record* https://www.ihis.com.sg/nehr/about-nehr (2011).
11. My Health Record. My Health Record. *My Health Record* https://www.myhealthrecord.gov.au/ (2016).
12. Wood, A. *et al.* Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* **373**, n826 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

13. Rush, R. Taking Note. *N. Engl. J. Med.* **381**, 9 (2019).

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* **9**, 3000113 (2021).

15. Langham, J. *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)* 3036–3042 (2021). doi:10.1109/BIBM52615.2021.9669363.

16. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* **1**, 1–10 (2018).

17. Appelbaum, L. *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* **143**, 19–30 (2021).

18. Kronzer, V. L. *et al.* Investigating the impact of disease and health record duration on the eMERGE algorithm for rheumatoid arthritis. *J. Am. Med. Inform. Assoc.* **27**, 601–605 (2020).

19. Sekelj, S. *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.* **28**, 598–605 (2021).

20. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* **6**, 26094 (2016).

21. Estiri, H. *et al.* Predicting COVID-19 mortality with electronic medical records. *NPJ Digit Med* **4**, 15 (2021).

22. Wu, J., Nadarajah, R., Raveendra, K., Cowan, J. C. & Gale, C. P. 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* **24**, (2022).

23. Bagheri, A. *et al.* Multimodal Learning for Cardiovascular Risk Prediction using EHR Data. *arXiv [cs.LG]* (2020).

24. Ben Miled, Z. *et al.* Predicting dementia with routine care EMR data. *Artif. Intell. Med.* **102**,

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

101771 (2020).

25. Zhao, J. *et al.* Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Sci. Rep.* **9**, 717 (2019).
26. Jin, B. *et al.* Predicting the Risk of Heart Failure With EHR Sequential Data Modeling. *IEEE Access* **6**, 9256–9261 (undefined 2018).
27. Hill, N. R. *et al.* Predicting atrial fibrillation in primary care using machine learning. *PLoS One* **14**, e0224582 (2019).
28. Tiwari, P. *et al.* Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. *JAMA Netw Open* **3**, e1919396 (2020).
29. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
30. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **17**, 129–145 (2016).
31. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine* **4**, 1–13 (2021).
32. Li, Y. *et al.* BEHRT: Transformer for Electronic Health Records. *Sci. Rep.* **10**, 1–12 (2020).
33. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
34. Cox, D. R. Regression Models and Life-Tables. *J. R. Stat. Soc. Series B Stat. Methodol.* **34**, 187–202 (1972).
35. All of Us Research Program Investigators *et al.* The 'All of Us' Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
36. Sudlow, C. *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.* **12**, e1001779 (2015).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

37. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325 (2019).

38. Ramirez, A. H. *et al.* The All of Us Research Program: Data quality, utility, and diversity. *Patterns (N Y)* **3**, 100570 (2022).

39. Martinez, F. J. *et al.* A New Approach for Identifying Patients with Undiagnosed Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **195**, 748–756 (2017).

40. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).

41. Castro, J., Gómez, D. & Tejada, J. Polynomial calculation of the Shapley value based on sampling. *Comput. Oper. Res.* **36**, 1726–1730 (2009).

42. Finlayson, S. G. *et al.* The Clinician and Dataset Shift in Artificial Intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).

43. Wong, A. *et al.* External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern. Med.* **181**, 1065–1070 (2021).

44. Guo, L. L. *et al.* Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci. Rep.* **12**, 2726 (2022).

45. National Institute for health and Care Excellence (NICE). Cardiovascular disease: risk assessment and reduction, including lipid modification. https://www.nice.org.uk/guidance/cg181 (2014).

46. 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.* **42**, 2439–2454 (2021).

47. Collins, R. *et al.* Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet* **388**, 2532–2561 (2016).

48. Ference, B. A. *et al.* Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

European Atherosclerosis Society Consensus Panel. *Eur. Heart J.* **38**, 2459–2472 (2017).

49. Vaduganathan, M. *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* **396**, 121–128 (2020).
50. Adelson, K. *et al.* Standardized Criteria for Palliative Care Consultation on a Solid Tumor Oncology Service Reduces Downstream Health Care Use. *J. Oncol. Pract.* **13**, e431–e440 (2017).
51. Weissman, D. E. & Meier, D. E. 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.* **14**, 17–23 (2011).
52. Centeno, C. & Arias-Casais, N. Global palliative care: from need to action. *The Lancet. Global health* vol. 7 e815–e816 (2019).
53. de Lemos, J. A. *et al.* Multimodality Strategy for Cardiovascular Risk Assessment: Performance in 2 Population-Based Cohorts. *Circulation* **135**, 2119–2132 (2017).
54. Steinfeldt, J. *et al.* Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. *Lancet Digit Health* **4**, e84–e94 (2022).
55. Buergel, T. *et al.* Metabolomic profiles predict individual multidisease outcomes. *Nat. Med.* (2022) doi:10.1038/s41591-022-01980-3.
56. Vayena, E. Value from health data: European opportunity to catalyse progress in digital health. *Lancet* **397**, 652–653 (2021).
57. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
58. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* **12**, e0175508 (2017).

medRxiv preprint doi: https://doi.org/10.1101/2023.03.10.23286918; this version posted September 26, 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 .

59. Moons, K. G. M. *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* **162**, W1–73 (2015).

60. Stekhoven, D. J. & Bühlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).

61. miceforest. *PyPI* https://pypi.org/project/miceforest/.

62. Katzman, J. L. *et al. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network*. https://arxiv.org/pdf/1606.00931.pdf (2017).

63. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer Normalization. *arXiv [stat.ML]* (2016).

64. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

65. Kingma, D. P. & Ba, J. L. 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.

66. Paszke, A. *et al.* Automatic differentiation in PyTorch. *Advances in Neural Information Processing Systems 30* 1–4 (2017).

67. Machine Learning CO2 impact calculator. https://mlco2.github.io/impact/.

68. Harrell, F. E., Jr, Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).

69. lifelines 0.25.8. https://lifelines.readthedocs.io/en/latest/ (2021).

70. How does All of Us assess diversity? What communities does All of Us consider 'underrepresented in biomedical research?' https://www.researchallofus.org/faq/how-does-all-of-us-assess-diversity-what-communities-does-all-of-us-consider-underrepresented-in-biomedical-research/.