Development and assessment of a machine learning tool for predicting emergency admission in Scotland

James Liley[1,2,*,⊥], Gergo Bohner[1,3,*], Samuel R. Emerson[4], Bilal A. Mateen[1,5], Katie Borland[6], David Carr[6], Scott Heald[6], Samuel D. Oduro[6], Jill Ireland[6], Keith Moffat[6,7], Rachel Porteous[6], Stephen Riddell[6], Nathan Cunningham[1,8], Chris Holmes[1,9], Katrina Payne[1], Sebastian J. Vollmer[1,3], Catalina A. Vallejos[1,2,⊥], and Louis J. M. Aslett[1,4,⊥]

[1]**Alan Turing Institute, London, UK**
[2]**MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK**
[3]**Mathematics institute, University of Warwick, UK**
[4]**Department of Mathematical Sciences, Durham University, Durham, UK**
[5]**Institute of Health Informatics, University College London, London, UK, and Wellcome Trust, London, UK**
[6]**Public Health Scotland (PHS)**
[7]**University of St Andrews, St Andrews, UK**
[8]**Department of Statistics, University of Warwick, UK**
[9]**Department of Statistics, University of Oxford, UK**
[*]**Equal contribution**
[⊥]**Corresponding**

ABSTRACT

Avoiding emergency hospital admission (EA) is advantageous to individual health and the healthcare system. We develop a statistical model estimating risk of EA for most of the Scottish population ($> 4.8M$ individuals) using electronic health records, such as hospital episodes and prescribing activity. We demonstrate good predictive accuracy (AUROC 0.80), calibration and temporal stability. We find strong prediction of respiratory and metabolic EA, show a substantial risk contribution from socioeconomic decile, and highlight an important problem in model updating. Our work constitutes a rare example of a population-scale machine learning score to be deployed in a healthcare setting.

Keywords: Emergency admission, Primary care, Machine learning

INTRODUCTION

Emergency hospital admission (EA) - the event in which an individual is admitted to a hospital in an unplanned way - indicates an unforeseen decline in an individual's health: for instance, loss of control over disease state, breakdown of mental health, infection or trauma. Modern health and social care policies aim to reduce EA by preventing it through proactive strategies [3]. EA can often be avoided by appropriate primary care intervention [22, 26, 9]. Hence, machine learning (ML) approaches to determine EA risk for each patient may be of value to primary care practitioners [16], particularly through identifying moderate-risk patients who are not already prioritised. The recent COVID-19 pandemic has further highlighted the need for rapid identification of populations at risk of needing acute health services, and directing preventive interventions such as shielding.

A range of models have been developed to predict EA from routinely-collected electronic health records (EHR) [25, 21, 30, 7]. However, transferability of predictive models across temporal and geographical settings is limited both due to differing demographics and differences in data availability [30]. Development of models in the setting they will eventually be used is thus typically preferable to reapplication of models trained in different settings. Here, we focus on SPARRA (Scottish Patients At Risk of Re-admission and Admission) — a model developed by the Information Services Division of the National Services Scotland (now incorporated into Public Health Scotland; PHS) to predict EA risk in the Scottish population [1]. To date, three versions of SPARRA have been deployed, with the target population increasing from only people aged 65 and older who had already had an EA in the past 3 years (version 1, 2006) to essentially the entire Scottish population (version 3, 2012). Every month, the predicted risk for each patient is shared with general practitioners (GPs). GPs are given incentives to take action on patients with an estimated EA probability between 40 and 60%. It is assumed those with an estimated EA probability between 61-99% are already known to GPs; however, there may be patients in that score range who GPs are unaware of and who would benefit from review. SPARRA version 3 (v3) scores are also used for planning in Health Boards, Health and Social Care Partnerships and academic research projects.

The SPARRA v3 model was derived from national EHR databases including emergency and elective admissions, day cases, outpatient attendances, accident and emergency (A&E) attendances, and community prescribing records. Most long term conditions are flagged from hospital inpatient records alone (e.g. asthma), while other long term conditions are flagged from hospital inpatient records and/or community prescribing records (e.g. evidence of epilepsy). SPARRA v3 uses separate logistic regressions on three subcohorts of individuals: frail elderly conditions (FEC; individuals aged > 75); long-term conditions (LTC; individuals aged <75 with prior healthcare system contact) and young emergency department (YED; individuals aged <55 who have had at least one A&E attendance in the previous year). If an individual belongs to more than one of these groups, the maximum between the associated scores is reported. SPARRA v3 was fitted once (at its inception) with regression coefficients remaining fixed thereafter.

In this work, we develop SPARRA version 4 (v4), using contemporary supervised and unsupervised ML methods using the same input data sources available for SPARRA v3. Our goal is two-fold: to improve predictive performance and calibration whilst using more contemporary data as input. Our data set comprises around 4.5 million individuals across seven time points (years), with a total of around 30 million individual-time observations. We validate SPARRA v4 using three-fold cross-validation, and evaluate the stability of the model and behaviour of risk scores across time. We examine the predictive model to determine the types of admission most readily predictable, and the effect of age and deprivation (measured by Scottish Index of Multiple Deprivation (SIMD) decile [15]) on risk scores. We use Shapley values [20] to partition individual risk scores: to ascertain the most important overall predictors and to assess potentially non-linear contributions to the overall risk.

By updating an already-deployed model, our work highlights a critical problem in this area [19]. Since SPARRA v3 is already visible to GPs (who may intervene to lower the risk of high-risk patients), SPARRA v3 can alter the observed risk in the training data used for SPARRA v4 and the score can become a 'victim of its own success' [18, 27]. This is potentially hazardous: if some risk factor R confers high v3 scores and the score prompts GP intervention (for instance, enhanced follow-up), then from the point of view of training data for v4, R no longer confers increased risk. Should v4 be used in place of v3, some individuals would therefore have their EA risk underestimated, potentially diverting important anticipatory care away from them. As a solution to this critical issue, we propose to use v4 as an *adjuvant* to v3, rather than a naive model replacement.

Beyond merely developing a new predictive model, an important additional contribution of this work is in demonstrating practicalities in model delivery, as SPARRA v4 will be deployed across all of Scotland. Privacy considerations in use of EHR data required all model fitting to be performed on secure servers, with limited computational power and software availability. We describe the challenges arising in this context and how we adopted reproducible research practices. Finally, to ensure long-term use and maintenance of SPARRA v4, the project necessitated close collaboration between PHS staff and other authors, with regular knowledge transfer between them. Our work therefore provides a wider framework to support future collaboration between PHS and academia, demonstrating and resolving a range of challenges in *in-vivo* ML models in healthcare.

This manuscript was written to conform to the TRIPOD guidelines [10]. See supplementary table 5 for details. All code is publicly available along with all data used to generate figures and tables at `github.com/jamesliley/SPARRAv4`.

RESULTS

Data and model fitting procedures

The samples in our data consist of *(patient,time)* pairs (Figure 1a). The patients included the majority of the Scottish population ($\approx 80\%$) across three time cutoffs: 1 May and 1 December 2016, and 1 May 2017. Our raw data was in the form of patient databases routinely collected by PHS (Figure 1b, Supplementary Table 1). A patient-time pair was included in our study cohort if the patient was alive at the time cutoff, the patient had a healthcare interaction recorded in at least one of the source EHR tables prior to the time cutoff (Supplementary Table 1), and the patient had a valid SPARRA v3 score at the time cutoff (see inclusion criteria for SPARRA v3 [1]). An EA was deemed to have occurred for a given individual and time cutoff if, in the year following the time cutoff, the individual either had a recorded emergency admission to a hospital in Scotland or died. (Supplementary Note 2). Our total dataset comprised 5,829,532 unique individuals, 16,142,279 potential *(patient,time)* pairs, and 443,566,149 recorded interactions with the healthcare system (not including recorded deaths). After exclusions, our cohort comprised 4,870,488 unique individuals and 12,957,648 *(patient,time)* pairs, in which there were 1,163,520 EA events (9%). Demographics (age, sex, SIMD) of individuals in each source table, in the whole dataset, in our study cohort and stratified by EA status are shown in Figure 1b. A breakdown of demographics across these patient cohorts is shown in Figure 1c. Overall, we observe that our study cohort is slightly older, has a higher percentage of females, and is moderately more deprived than the population in general. When stratifying our study cohort between patients with and without EA, we observed that patients with an EA tend to be older and have a lower socioeconomic status.

We transformed source EHR tables to a design matrix (Figure 1a, Online Methods). For each patient, information from up to three years before the time cutoff was considered when building the input predictors, with the exception of long-term condition records, where information from ten years before the time cutoff was used. Most predictors were either counts of previous events (e.g. number of respiratory related prescriptions, number of long-term conditions), amount of time since most recent events (e.g. years since last EA) or binary indicators (e.g. diagnosis of multiple sclerosis). A list and description of derived predictors is provided in Supplementary Table 2. These predictors were augmented with engineered features using a topic model [5] which aggregates community prescription and long-term conditions data into thirty 'topics', roughly coding for classes of conditions or multimorbidity. We included the probability of membership for each of these topics as additional predictors in our model (Online Methods). The only predictors with missing values were SIMD (0.67% missingness) and topic features (0.82% missingness), the latter arising when individuals had hospital records but no filled prescriptions or recorded long-term conditions. Treatment of missing values is described in the Online Methods.

For the prediction task, we considered an ensemble of a range of constituent model types, building model comparison into our validation procedure (see Supplementary Note 3). We used three-fold cross validation, randomly selecting patients to be included in each fold (Figure 1a). In each cross-validation stage, a random one-third of patients were excluded from model fitting and used as a test set. Our cross-validation scheme was designed such that all constituents of the model evaluated on that test set were agnostic to samples in that test set (Online methods, Supplementary Figure 1, Supplementary Note 3).

The ultimate goal for this project is to produce SPARRA v4, an updated risk prediction score that can be deployed by PHS and used by GPs across Scotland (Figure 1a). However, since SPARRA v3 is already universally deployed in Scotland with an unknown degree of use, we needed to take particular care in the process of updating v3 to v4. Namely, if v4 were to naively replace v3, individuals who were intervened on in response to v3 may be assigned inappropriately low scores by v4 [18, 27, 19]. To avoid this problem, we propose to deploy the maximum of v3 and v4 to GPs, which averts this potential danger at the cost of a small decrease in calibration. See Online Methods for further details.

Overall predictive performance

We compared the predictive performance of SPARRA v3 and SPARRA v4 (Figure 2a,2b). In test sets, the SPARRA v4 model was effective at predicting EA, and stronger than SPARRA v3 on the basis of AUROC (SPARRA v4: 0.801; SE $< 3 \times 10^{-4}$, vs SPARRA v3: 0.781 ; SE $< 3 \times 10^{-4}$) and AUPRC (SPARRA v4: 0.402; SE $< 5 \times 10^{-4}$, vs SPARRA v3: 0.359; SE $< 5 \times 10^{-4}$). SPARRA v4 was also better calibrated (Figure 2c).

The apparently subtle differences in performance between v3 and v4 are substantial when considered in terms of real-world consequences. Amongst the 50,000 individuals judged to be at highest EA risk by SPARRA v3, around 4,000 fewer individuals were eventually admitted than were amongst the 50,000 individuals judged to be at highest risk by SPARRA v4 (Figure 2d). If we assume that 20% of admissions are avoidable [6] and assume that avoidable admissions are as predictable as non-avoidable admissions (which may be conservative, as avoidable admissions are more likely to be pre-empted by other medical problems and hence be predictable), then in order to preempt 3,000 avoidable admissions by targeting the highest risk patients under SPARRA v4, we would need to intervene for approximately 1,500 fewer patients than when using v3 in the same way (Figure 2e).

Finally, we evaluated the performance of $\max(v3, v4)$ — the practical solution proposed above to avoid the risks of naive model updating. As seen in Figure 2a,2b, there was a small change in predictive performance (sligthly lower AUROC; slightly

higher AUPRC) and slightly attenuated calibration (Figure 2c). A density plot of v3 and v4 scores is shown in Supplementary figure 5. The performance of v4 may attenuate in a healthcare system where v3 is not in use [27, 18, 19].

Stratified performance of SPARRA v3 and SPARRA v4

We found that the difference in performance (by AUROC) between v3 and v4 was higher in subcohorts of patients at high risk (age > 80) and at low risk (age $20-70$, no previous EA), with a smaller performance differential in the complement of these subcohorts (denoted as moderate risk, Figures 2f, 2g, 2h). In high- and low- risk cohorts, general predictive accuracy (rather than differential) was lower than for the whole cohort (AUROC ≈ 0.7 for high- and low- risk subcohorts), suggesting that the differential performance of v4 was driven by improved performance on difficult-to-predict high- and low- risk samples, with more equivalent performance on easy-to-predict moderate-risk samples. In individuals for which v3 and v4 disagreed ($|v3-v4| > 0.1$), we found that v4 was better-calibrated when $v4 > v3$ and equivocal to $v3$ when $v3 > v4$ (Figure 2i).

To examine differences in performance more closely, we next explored the performance of SPARRA v3 and SPARRA v4 across different patient subcohorts defined by age, SIMD deciles and the three subcohorts defined as part of SPARRA v3 development. We noted that our model outperformed the existing score in terms of AUROC over all subcohorts (Figure 3a-c). We also explored the performance of SPARRA v4 after stratifying different types of EA admissions. We found that our model was able to predict certain medical classes of admission disproportionately well (Figures 3d,3e), namely predicting respiratory and endocrine/metabolic related admissions. As expected, we were less able to predict traumatic admissions, and admissions due to external causes of morbidity (ICD codes V00-Y99 [24]). Supplementary figure 10 shows relative performance by type of admission (injury, referral etc), with slightly better performance for non-injury-related than injury-related admissions.

When further analysing the risk scores predicted by SPARRA v4, we found that our model tends to better predict imminent admissions: individuals with high risk scores were more likely to have an EA near the start of the 1 year outcome period (Figure 3f). This was partly because individuals with a high risk score were disproportionately more likely to have multiple admissions (Figure 3g). However, amongst those with only one admission, individuals with a high SPARRA v4 score were also more likely to have an admission earlier in the year (Figure 3h).

Stability over time

Using the same analysis pipeline as for SPARRA v4, we fitted a model M_0 to an early time cutoff (t_0=1 May 2014), using only one year of data to derive predictors. We then assembled test matrices from data 1 year prior to the time points t_1=1 May 2015, t_2=1 Dec 2015, t_3=1 May 2016, t_4=1 Dec 2016, and t_5=1 May 2017 and applied M_0 to predict EA in the year following each time point. This enabled assessment of how the performance of M_0 changed over time. In order to ensure a fair comparison, we compared performance on subsamples of 1 million individuals from time points $t_1 - t_5$ chosen such that global admission rates matched those at t_0. We found that M_0 performed essentially equally well when used to predict admission following $t_1 - t_5$, with no statistically significant decrease in performance (adjusted p-values > 0.05 or improved performance with time for all comparisons of AUROCs; Supplementary Figures 6a, 6c, 6b).

We then compared how well fixed *predictions* computed at t_0 performed over time. We subsampled individuals with complete data such that admission rates and age distributions at $t_1 - t_5$ matched those at t_0, and assessed how well the initial scores predicted EA subsequent to $t_1 - t_5$. We observed that the fixed scores performed reasonably well even 2-3 years after t_0, although calibration was gradually lost (Figures 4a, 4b, 4c). We also note that the fixed scores obtained at t_0 predicted the event 'any EA between t_1 and a year after t_5' with AUROC 0.84 (SE $< 4 \times 10^{-4}$), higher than that achievable for predicting EA in one year periods. More generally, we observe that scores fitted and calculated at a fixed timepoint had successively lower AUROCs for predicting EA over future periods (Supplementary figure 8).

We then considered how individual scores changed amongst all patients with complete data at all time periods. We found that while quantiles of the distribution of scores at $t_1 - t_5$ increased as the cohort grew older (Figure 4d), the mean risk scores of individuals in the highest centiles of risk at t_0 tended to decrease over time (Figure 4e), suggesting that very high risk scores tend to be transient. We analysed this more closely by comparing bivariate densities of scores (Figure 4f and Supplementary Figure 7). We found that lower scores are more stable than higher scores over subsequent time points, and individuals 'jump' to higher scores (upper left in Figure 4f) more than they drop to lower scores (bottom right). Amongst individuals with a given score at t_1, median score at t_2 tends to be lower, although quantiles of the overall score distribution increase (Figures 4d and 4e).

Predictor importance

We examined the contribution of predictors to risk scores at an individual level by estimating Shapley values [20] for each predictor/individual pair. The predictors with largest mean absolute Shapley value (excluding v3 and topic features) were age, the number of recorded long-term conditions, the number of previous A&E attendances, and the number of previous EAs (Table 5e). Most predictors had non-linear effects (Supplementary Figures 14-20).

We found a non-linear importance of age: the risk contribution from age was high in infancy, dropped rapidly through childhood, remained stable until around age 65, then rose rapidly, stabilising at around age 75 until death (Figure 5a). We also

found a non-linear importance of SIMD [13] (Figure 5b) and number of previous emergency hospital admissions (Supplementary Figure 13). We further investigated the contribution of SIMD by comparing Shapley values between variables. We computed the mean difference in contribution of SIMD to risk score between individuals in SIMD 10 areas (lowest deprivation) and SIMD 1 areas (highest deprivation), and the additional years of age which would contribute an equivalent amount. This was generally around 10-40 additional years (Figure 5c). In terms of raw admission rates, disparity was further apparent: individuals aged 20 in SIMD 1 areas had similar admission rates to individuals aged 70 in SIMD 8-10 areas (Figure 5d).

Model analysis

We assessed the added value of inclusion of topic-model derived features, which summarise more granular information about the previous medical history of a patient with respect to those included in SPARRA v3. For this purpose, we refitted model M_0 with topic-derived features excluded from the predictor matrix (model M_0'). We compared the performance of M_0 and M_0' on a set of test data for time point t_1. On comparison of AUROC curves, M_0 had a slightly better performance than M_0', though the absolute AUROC difference was small (p-value $= 6 \times 10^{-47}$; AUC 0.789 vs 0.788; Supplementary figure 12). Analogously to Figure 2, we also computed the additional number of samples correctly identified as having an EA amongst the top scores by M_0 and M_0', and found that use of M_0 over M_0' increased the number of EAs detected in the top 500,000 scores by around 200. Supplementary table 4 shows the most pertinent features (ICD10 codes and prescription types) for each topic in the model used for prediction on fold 1.

Finally, we explored a breakdown of model performance (AUROCs, AUPRCs, and calibration curves) across the different model constituents included in our ensemble (Table 3, Figures 2, 3, and 4). The performance of the ensemble was slightly better than the best constituent models (XG-boosted trees and random forests). Several constituent models (ANN, GLM, Naive Bayes) had vanishing coefficients in all ensembles, so predictions for these models need not be computed when generating SPARRA v4.

DISCUSSION

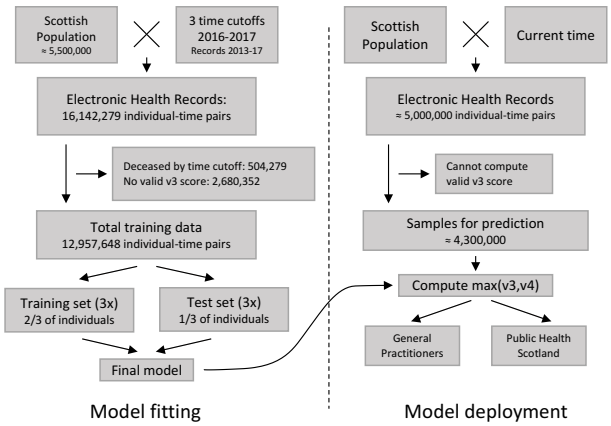
The SPARRA v4 model is a state-of-the-art risk prediction model for EA developed using routinely collected EHR data, optimised for use in the Scottish population. It augments and consistently outperforms the SPARRA v3 model that is currently in clinical use. Although the improvement on model metrics (AUROC, AUPRC) from SPARRA v3 to SPARRA v4 is small, this difference is highly meaningful in terms of absolute performance. More generally, we demonstrate that sophisticated machine learning methods can enable meaningful gains in performance in a real-world setting. In particular, to our knowledge, this is the first use of a topic model to summarise complex multi-morbidity data (captured by historic prescriptions and diagnosis) and an ensemble prediction approach in the context of EA prediction. We highlight a hazard of updating widely-circulated predictive scores, and discuss how machine leaning models can be deployed, an area of growing importance for health data science.

We find a range of epidemiological associations with predictable EA risk. Both the tendency to predict imminent admissions and the relative instability of high risk scores with time indicate that SPARRA v4 is sensitive to short-term periods of high risk, particularly relating to respiratory, metabolic and mental-health related illness. The relationship between deprivation (SIMD) and EA risk is complex; SIMD includes a constituent of general health [13], and individuals in low-SIMD areas are more likely to miss primary care appointments [12] suggesting a lower overall rate of primary care uptake. We do not assert causality for this or any other reported associations. In particular, we cannot make counterfactual predictions, and thus cannot directly recommend any changes to patient management which would lower EA risk. Furthermore, lowering of EA risk does not necessarily entail overall patient benefit. : for instance, long-term oral corticosteroid use in mild asthmatics would reduce EA risk at an unacceptable cost of long-term morbidity. actions taken to stabilise at-risk patients As such, any potential interventions should be carefully assessed by GPs and other members of the clinical team on a case-by-case basis.

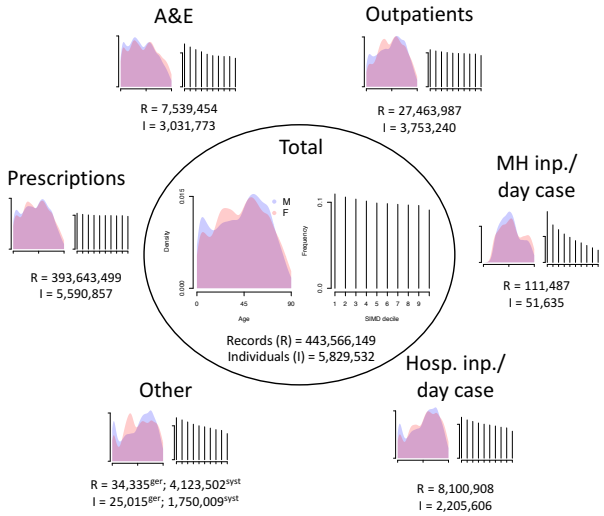
We chose to code our target variable as EA in the subsequent year. This is not a universal choice in EA prediction [25, 30], and targets such as survival time to admission may lead to different model properties and performance (noting that our model M_0 could predict next-three-year EA better than one-year). As well as the advantage of maintaining interpretability (a one-year EA probability was used in earlier versions of SPARRA) this target has the advantage of being independent of seasonal influences on EA frequency. Although a range of models have been proposed to predict EA, it is difficult to transfer information between geographic settings, since they are highly dependent on data collection protocols of healthcare systems [30]. We do not expect that our model will directly be used outside of Scotland. However, the design of our reproducible analysis pipeline is general and could be adapted to other settings. Performance of our model is weaker than that of a related model in England [25], but this is unsurprising given reduced predictor availability in our study. In fact, data in [25] included primary care records and health indicators (body mass index, smoking status, laboratory tests). However, our model has the advantage of using only data routinely collected centrally at a national level by Scottish healthcare authorities (albeit not including records from GP practices), and can thus be computed for the general Scottish population without the need for any additional data collection.

As for any predictive model in this area, the question of of translation into clinical action in a way which optimally benefits the patient is a vital area of ongoing research, and is essential for quantifying the benefit of including such scores in clinical practice. To do this, EA risk scores must be comprehensive, accurate, and interpretable, which we have aimed for in SPARRA v4. The SPARRA v4 model will be deployed in the Scottish NHS, augmenting the SPARRA v3 model. We expect that the SPARRA score will be most useful to patients for whom anticipatory care plans are not already in place, and for whom GPs do not know well. For this reason, well calibrated risk scores are particularly important. We anticipate that the SPARRA v4 tool will contribute to the overall effectiveness of the Scottish NHS, and correspondingly to the health of the Scottish population. We will collaborate to achieve a successful deployment and will carefully consider the feedback from GPs to improve the model and the communication of its results further (e.g. via informative dashboards).

Ongoing work will evaluate the use of different target variable types, new machine learning methods, improved coverage of the Scottish population (SPARRA v4 only covers around 80% of the population and likely excludes healthy individuals with low engagement with the healthcare system), and translation to clinical action. As the COVID-19 pandemic resolves, it will also be important to assess potential effects of dataset shift [28], e.g. due to disproportionate mortality burden in older individuals and long term consequences of COVID-19 infections. We anticipate that robust and reproducible development of risk prediction scores such as SPARRA v4 will play an important role in the design of proactive interventions to manage future risk.



(a) Flow chart

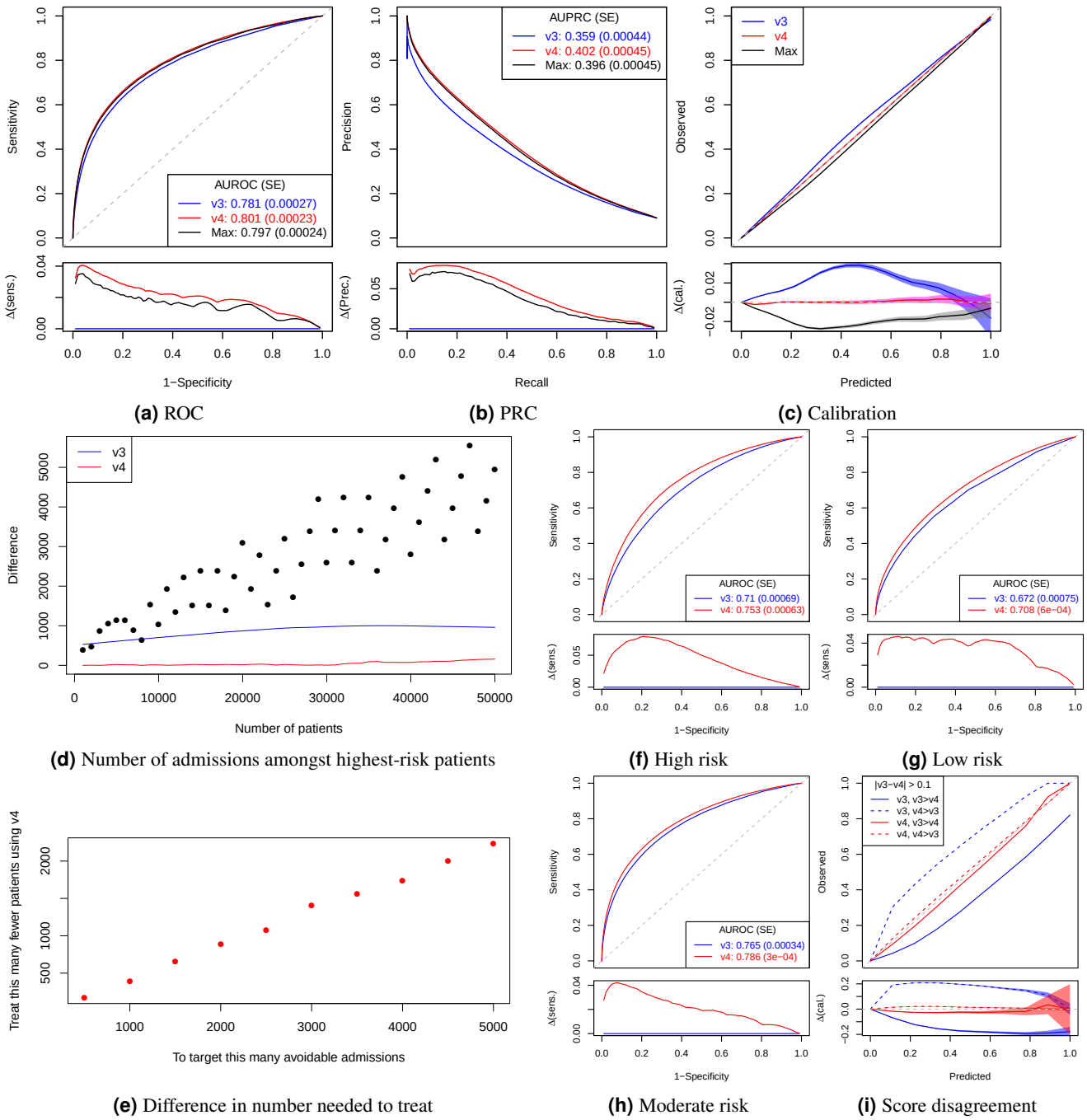


(b) Data sources

	Patient cohort				
Variable	Population	Any EHR	In study	EA	No EA
Sex (%)					
Male	48.5	47.5	45.4	46.3	45.3
Female	51.5	52.5	54.6	53.7	54.7
Age (%)					
0-20	16.9	20.3	19.7	11.7	20.5
20-70	71.2	64.5	64.9	50.4	66.4
70+	11.9	15.2	15.4	37.8	13.2
SIMD decile (%)					
1-5	50.0	50.6	51.6	59.2	50.9
6-10	50.0	48.5	47.7	40.2	48.4
Any LTC (%)	Unknown	15.6	15.9	41.4	13.4

(c) Demographics

Figure 1. Data sources and prediction pipeline. 1a: Flow chart for model fitting and deployment. 1b: Distribution of age and SIMD across different input EHR data sources prior to any exclusions, excluding death records (also see Supplementary Table 1). Numbers show total number of records in each source, and figures show demographics of individuals in that source. All figures are drawn to the same scale. MH: mental health; inp.: inpatient; A&E: accident and emergency; Other: geriatric long stay and System Watch data; ger.: geriatric long stay; syst.: System Watch. 1c: Demographics of individuals in different patient cohorts: the whole Scottish population, those present in at least one input EHR table, our study cohort and our cohort after stratifying by EA status. LTC denotes long-term conditions (e.g. epilepsy). The cohort with 'any EHR' includes deceased and otherwise excluded individuals from our study cohort. Data for the Scottish population is from the 2011 Census [23]. This precedes our cohort by three years.

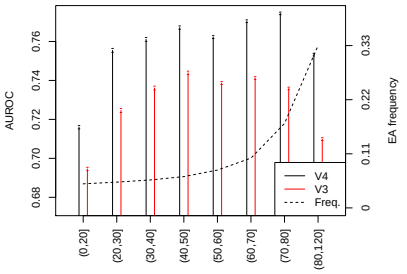


(a) ROC **(b)** PRC **(c)** Calibration

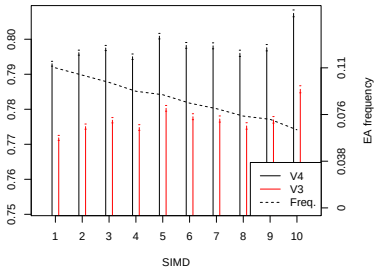
(d) Number of admissions amongst highest-risk patients **(f)** High risk **(g)** Low risk

(e) Difference in number needed to treat **(h)** Moderate risk **(i)** Score disagreement

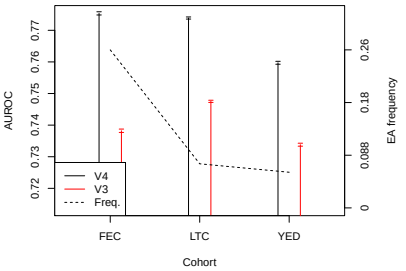
Figure 2. Comparison of SPARRA v3 and v4. Panels 2a, 2b, 2c show ROC, PRC and calibration curve respectively for SPARRA v4 (new score), SPARRA v3 (existing score), and $\max v3, v4$ (deployed score). Lower sub-panels show difference between v4/$\max(v3, v4)$ and v3 in panels 2a and 2b, and difference between curves and the X-Y line in panel 2c. Confidence intervals in panels 2a and 2b are negligible; confidence envelopes in panel 2c are pointwise (that is, for each x-value, not the whole curve). Predicted/true value pairs are combined across cross-validation folds in all plots for simplicity. Panel 2d shows the difference in the number of individuals who had an EA amongst individuals designated highest-risk by v3 and v4. Lines indicate deviation from perfect calibration. The repeating pattern is due to rounding of v3. Panel 2e shows the difference in the number of highest-risk individuals to target to avoid a given number of admissions. Panels 2f,2g and 2h compare ROC curves of SPARRA v4 (new model) and SPARRA v3 (existing model) in patient subcohorts defined by different risk levels. Comparisons of performance over other subcohorts are shown in Supplementary Figures 9. Panel 2i shows calibration curves for v3 and v4 in samples for which $|v3 - v4| > 0.1$.



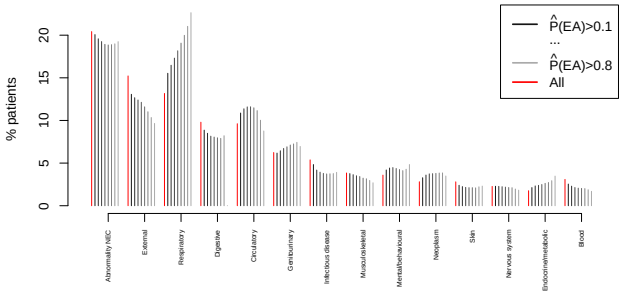
(a) AUROC, EA freq. by age



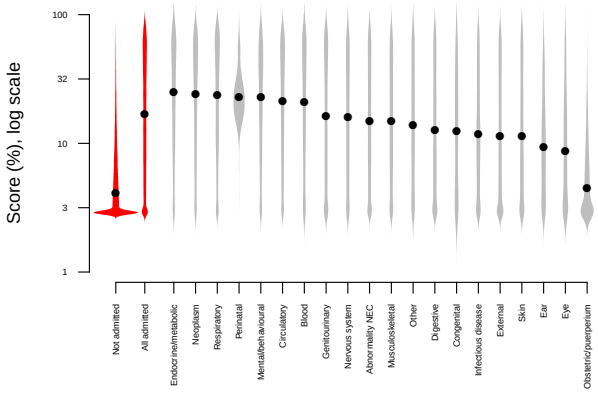
(b) AUROC, EA freq. by SIMD



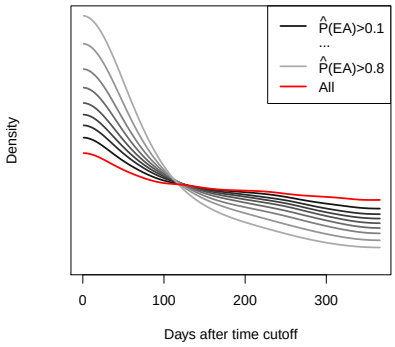
(c) AUROC, EA freq. by v3 submodels



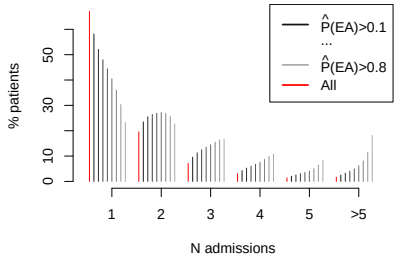
(d) Distribution of admission type for score cutoffs



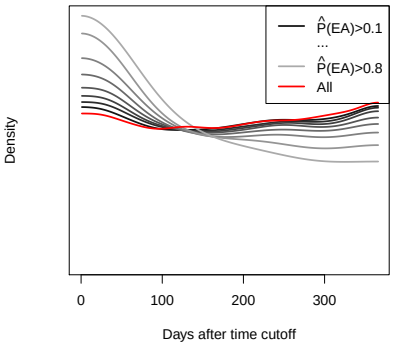
(e) Distribution of scores by admission types



(f) Time to first event

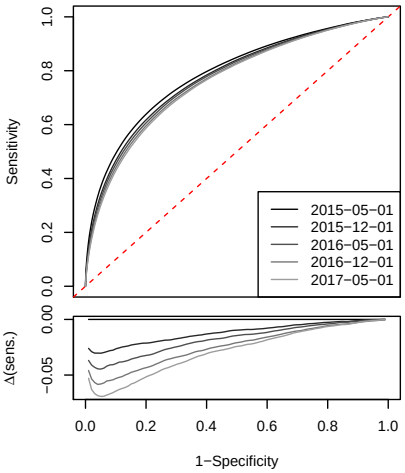


(g) Number of events

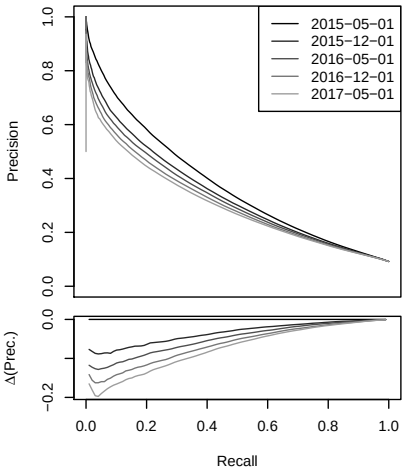


(h) Time to only event

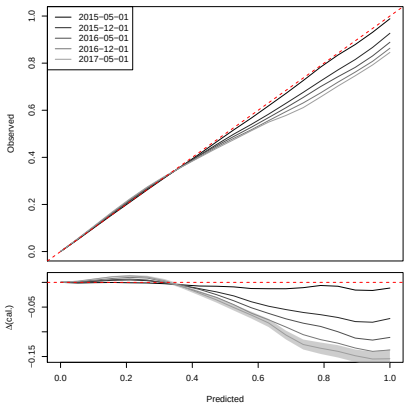
Figure 3. Stratified performance of SPARRA v3 and v4. Panels 3a, 3b and 3c show difference in AUROC between v3 and v4 (red/black) and EA frequency (dashed lines) in subcohorts defined by age, SIMD and the original subcohorts defined during SPARRA v3 development (see Introduction), respectively. Panels 3d and 3e show SPARRA v4 performance by category of disease by ICD10 code (Abnormality NEC: abnormality not elsewhere classified;). In panel 3d, bar height for a given particular x-axis category and a given colour (score threshold) is the proportion of individuals with EA reaching that score threshold admitted for that reason. Percentages sum to 100% across bars of each colour (score category) rather than disease category. A rising profile indicates relatively better prediction of this category. Panel 3e shows distribution of log-scores for individuals admitted in each class, with black points indicating the associated medians. Supplementary figure 11 shows equivalent data to panel 3d as positive-predictive values. Panel 3f shows the distribution of time-to-event after the first time cutoff amongst cohorts of individuals who 1) had an EA in the year following the time cutoff and 2) had a SPARRA v4 score above a given cutoff. Panel 3g shows the discrete distribution of number of events over the same cohorts. Panel 3h shows the distribution of time-to-event amongst cohorts of individuals who were admitted exactly once in the year following the first time cutoff.



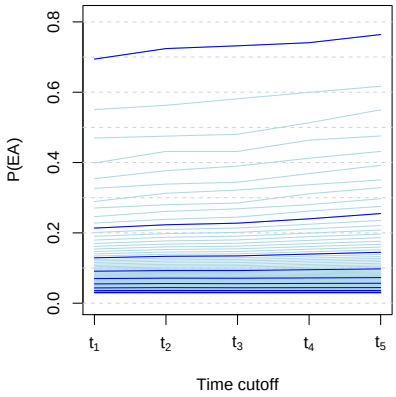
(a) ROC for fixed predictions



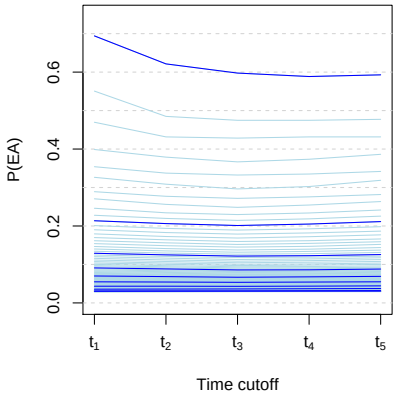
(b) PRC for fixed predictions



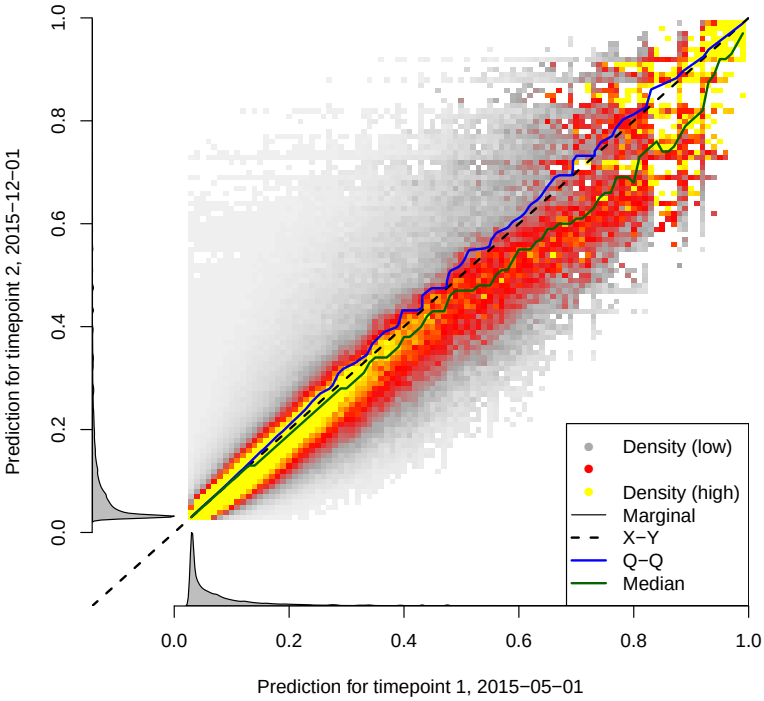
(c) Cal. for fixed predictions



(d) Score distribution quantiles over time

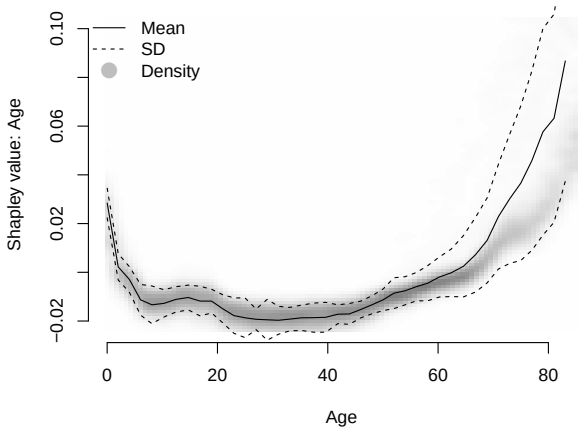


(e) Cohort tracked scores over time

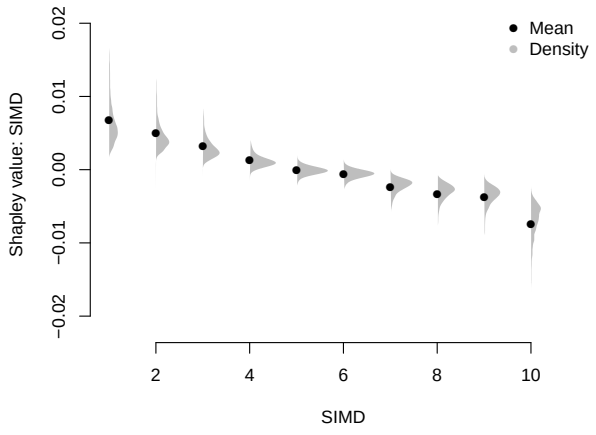


(f) Bivariate density of scores 6 months apart

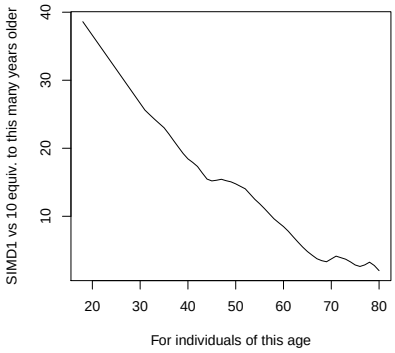
Figure 4. Behaviour of the same predictions over time. Panels 4a, 4b, 4c show performance of static *predictions* made for the early time point at estimating risk for later time points. Lower panels for ROC curves show differences between sensitivity at t_i and at t_1; for PRC curves difference in PPV between t_i and t_1; and for calibration curves the difference between observed and expected EA frequency. Pointwise 95% confidence intervals for calibration curves are shown only for the final time point for clarity. Panel 4d shows the change in distribution centiles (light blue) and deciles (dark blue) of risk scores with time, across all individuals with data available at all time points. Panel 4e considers cohorts of individual by risk centiles (light blue) and deciles (dark blue) at the first time point (2 May 2015) and tracks average risks of these cohorts over time. Panel 4f shows joint density (low to high: white-grey-red-yellow) of individual risk scores at t_1 (2 May 2015) and t_2 (1 Dec 2015). The density is normalised to uniform marginal on the Y axis, then the X axis; true marginal distributions of risk scores are shown alongside in grey.



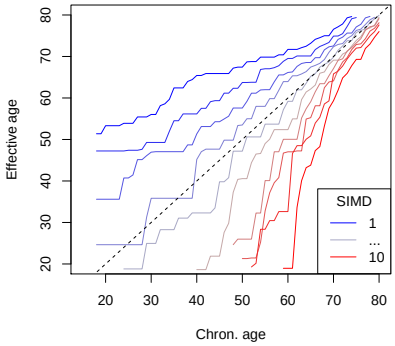
(a)



(b)



(c)



(d)

Variable	Importance
Number of recorded long term conditions	2
Age at time cutoff	2
Days since last emergency inpatient or day case attendance	0.82
Number of previous A&E attendances	0.46
Number of infection related prescriptions	0.39
Sex	0.37
Number of central nervous system related prescriptions	0.37
SIMD decile	0.34
Days since last A&E attendance	0.32
Total number of paid prescription items	0.31
Days since last outpatient attendance	0.3
Total amount paid in prescription costs	0.25
Days since last acute/inpatient or day case attendance	0.23
Number of BNF sections from which a prescription was filled	0.2
Number of elective bed days	0.19
Number of previous outpatient mental health appointments	0.17
Number of antisecretory and mucosal protectant prescriptions	0.15
Number of endocrine related prescriptions	0.12
Number of respiratory corticosteroid prescriptions	0.12
Number of emergency bed days	0.11

(e)

Figure 5. Panels 5a, 5b and 13 show non-linear influence of age, SIMD, and number of previous elective (Elec.) and emergency (Em.) admissions on risk scores. Panel 5c shows the number of additional years of age needed to contribute as much to risk score as a SIMD 1-SIMD 10 difference. Panel 5d shows 'Effective ages' for each age: for an (age,SIMD) pair, the age at mean SIMD with the equivalent EA rate. Table 5e shows the 20 most important variables by mean absolute Shapley value (percentage scale). Importance can be interpreted as the average percent added or subtracted to risk score due to this factor. Also see supplementary table 2 and Supplementary Figures 14-20.

ONLINE METHODS

Feature engineering

A typical entry in the source EHR tables (Supplementary Table 1) recorded a single interaction between an individual and NHS Scotland (e.g. hospitalisation), with each record comprising a unique identifier (ID) for the individual, the date on which the interaction was initiated (admission), the date it ended (discharge), and details of the interaction (diagnoses made, procedures performed). We generated input features by scanning source tables for records corresponding to each ID-time pair. For each individual, we considered all records from up to three years before the time cutoff.

Most input predictors were similar to those used for SPARRAv3 [1]. See Supplementary Table 2 for a full list of predictors. In terms of demographics, we included age, sex, and SIMD decile (the latter serving to use using geographic-level deprivation as a proxy for individual-level deprivation). Other input predictors include counts of previous admissions (e.g. A&E admissions, drug-and-alcohol related admissions), and the amount of time spent in hospital (e.g. emergency bed days). Following findings in [25], where possible, we also included predictors encoding time-since-last-event (e.g. days since last outpatient attendance). From community prescribing data, we used predictors encoding the number of prescriptions for various pre-specified types (e.g. respiratory), as well as the total number of different types of prescriptions, number of paid prescription items, and total amount spent on prescriptions. From data on long-term conditions, we evaluated recorded presence of a long term condition (eg. multiple sclerosis, epilepsy), the number of years since diagnosis of a range of long-term conditions (e.g. asthma), the total number of long-term conditions recorded, and the number of long-term conditions resulting in hospital admissions.

In addition to the predictors listed above, we used a Latent Dirichlet Allocation (LDA)-based topic model [5] to derive more information from prescriptions and long-term condition data. We jointly modelled prescriptions and long-term conditions using 30 topics, considering samples as 'documents' and conditions/prescriptions as 'words'. We retained the map from documents to topic probabilities, and used derived topic probabilities as input variables for our model. Document to topic probability maps were fitted only to the training set of data, and applied to test set data when evaluating accuracy.

Unavailable predictors

A range of other predictors are known to influence EA. In particular, an English study found that marital and smoking status, blood test parameters and family histories were highly predictive [25]. However, these were not included in our data which only contains information that is routinely collected in secondary care. In addition, due to privacy considerations, we did not have access to geographic location data. We would expect that patterns of EA risk would vary between rural and urban areas of Scotland, and this variation may be substantial given the geographic diversity of the region. As well as being unable to use geographic location as a predictor, this precluded the use of a geographically separated test set [30].

Machine learning prediction methods

We had little cause to believe in advance that any particular class of models would be best for this problem. Thus, we considered a range of model types. Rather than selecting and using only the best performing model, we chose an ensemble defined as an optimal linear combination of model outputs (L_1-penalised regression) in a similar way to Van der Laan (2007) [29]. Compared to choosing only one model, this slightly increases the range of predictor functions which are able to be modelled, slightly improves prediction, and reduces reliance on behaviour of a single model type, at the cost of a small number of additional degrees of freedom. We monotonically transformed the ensemble predictor to improve calibration by inverting the empirical calibration function (see Supplementary note 3 for details).

As part of the ensemble, we considered a variety of constituent (base) models. These include an artificial neural network (ANN) [17], two random forests (RF) [17] (one shallow, of depth 20, and one deep, of depth 40), three gradient boosted trees (XGB) [8] (of maximum constituent tree depth 3,4, and 8), a generalised linear model (GLM) [17], a naive Bayes model (NB) [17], and the model used in SPARRA v3 (v3) [1]). Full details of these models are given in the Supplementary Note 1.

Missing values

Since all non-primary care interactions with NHS Scotland are recorded in the input databases, there was no missingness for most predictors. For 'time-since-interaction' type predictors, individual-time pairs for which there was no recorded interaction were coded as the maximum lookback time. There was some missingness in SIMD records ($\sim$0.7%), possibly due to individuals without a fixed address, and in topic features ($\sim$0.8%). In both cases, missingness would be expected to be non-random and possibly informative. We thus managed predictor missingness in different ways across constituent models. We used mean-value imputation in the ANN and GLM models (deriving mean values from training data only), used missingness to inform tree splits (defaults in [17]) in RF, used sample-wise imputation in XGB (as per [8]) and dropped during fitting (default in [17]) in NB (omitted missing values for prediction).

Predictive performance

We optimised various metrics for constituent models, which were generally the default in the relevant functions. For the ANN, RFs and GLM, we minimised log-loss; for the NB model, likelihood; and for the XGB trees, a logistic objective. For fitting model weights in the ensemble ensemble, we optimised area under an ROC curve (AUROC) as a metric for model fit.

Although our primary endpoint for model performance was AUROC, we also considered area-under-precision-recall curve and calibration curves for visualisation of results. We plotted calibration curves using a kernelised calibration estimator; see Supplementary Note 5. We compared the performance of predictive models with and without topic features by evaluating AUROCs for prediction on a given year, and compared AUROCs using DeLong's test [11].

For simplicity, figures show ROC/PR curves for combined cross-validation folds (that is, given fold-wise predictions $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$ for targets Y_1, Y_2, Y_3, for ROC curves figures show $ROC\left(Y_1 Y_2 Y_3, \hat{Y}_1 \hat{Y}_2 \hat{Y}_3\right)$ and similarly for PR curves). Quoted AUROC/AUPRC values are means over folds (that is, $\frac{1}{3}\left(ROC(Y_1,\hat{Y}_1)+ROC(Y_2,\hat{Y}_2)+ROC(Y_e,\hat{Y}_3)\right)$ or similar), which averts problems from between-fold differences in models [14]. For fairness of comparison, we also used mean-over-folds to compute quoted AUROCs and AUPRCs for SPARRA v3, although SPARRA v3 was not fitted to our data.

Cross-validation

We fitted and evaluated the model using three-fold cross-validation (see supplementary Figure 1). Each third of the data was used as a training, validation and test set in turn. The full dependency structure is shown in supplementary figure 1. For each fold, the prediction function used was independent of data in that fold. A full description of the cross-validation procedure is given in Supplementary note 3.

Model updating

Our aim in producing the SPARRA score is to produce an accurate estimate of EA risk in the coming year under normal medical care. In other words, the score should represent the true EA risk if the score was not visible to the GPs. Because GPs see the score and may act on it, the observed risk may be lower than predicted - the score may become a 'victim of its own success' [18, 27]. Unfortunately, since the SPARRA v3 score is universally available to GPs, and may be acted on, we cannot assess the behaviour of the medical system in its absence. As illustrated in [19], this is potentially hazardous.

Formally, at a given fixed time, for each individual, EA in the next 12 months is a Bernoulli random variable. The probability of the event for individual i is modelled conditional on a set of covariates X_i derived from their EHR. We denote $v3(X_i)$, $v4(X_i)$ the derived SPARRA v3 and v4 scores as functions of covariates, and assume a causal structures shown in Figure 6. With no SPARRA-like predictive score in place, there is only one causal pathway $X_i \rightarrow EA$ (Figure 6A). It is to this system (coloured red) that $v3$ was fitted. In this setting, $v3(X_i)$ is an estimator of the 'native' risk $Pr(EA|X_i)$ (this ignores the effect of previous versions of the SPARRA score, which covered less than 30% of the Scottish population). Although $v3(X_i)$ is determined entirely by X_i, the act of calculating and distributing values of $v3(X_i)$ to GPs opens a second causal pathway from X_i to EA (Figure 6B). This is driven by GP interventions made in response to $v3(X_i)$ scores. It is to this system (coloured red) that SPARRA v4 is fitted. Hence, $v4(X_i)$ is an estimator of $Pr(EA|X_i, v3(X_i))$, a 'conditional' risk after interventions driven by $v3(X_i)$ have been implemented. If $v4$ naively replaced $v3$ (Figure 6C), we would be using v4 to predict behaviour of a system different to that on which it was trained (Figure 6B). To amend this problem, we propose to use v4 in *conjunction* with v3 rather than to replace it (Figure 6D). Ideally, GPs would be given $v3(X_i)$ and $v4(X_i)$ simultaneously and asked to *firstly* observe and act on $v3(X_i)$, *then* observe and act on $v4(X_i)$, thereby only using $v4(X_i)$ as per Figure 6D. This is impractical, so instead, we disseminate to GPs the single value $\max(v3, v4)$, avoiding the potential hazard of risk underestimation, at the cost of loss of score calibration.

Code and data availability and reproducibility.

Raw data for this project are patient-level NHS Scotland health records, and are confidential. Due to the confidential nature of the data used, all analysis took place on remote 'safe havens' [2, 4], without access to internet, software updates or unpublished software. Information Governance training was required for all researchers accessing the analysis environment. Moreover, to avoid the risk of accidental disclosure of sensitive information, an independent team carried out statistical disclosure control checks to all data exports, including the outputs presented in this manuscript. All analysis code and co-ordinates required to reproduce our Figures are available in `github.com/jamesliley/SPARRAv4`.

ETHICS STATEMENT

This work was conducted in accordance with UK data governance regulations under PBPP application number eDRIS 1718-0370

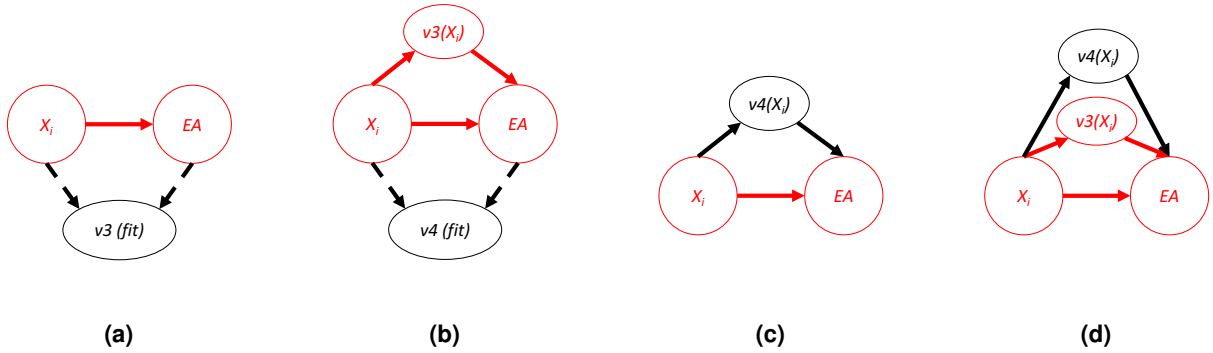


Figure 6. Causal structure for fitting of SPARRA v3 and v4. X_i represents covariates for a patient-time pair; Y represents the EA event; $v3(fit)/v4(fit)$ and $v3(X_i)/v3(X_i)$ represent the fitting and deployment of v3 and v4 respectively. Note v3 and v4 are fitted to different systems, so v4 cannot be used to directly replace v3.

ACKNOWLEDGEMENTS

The authors note that this project's success was entirely contingent on close co-operation between the Alan Turing Institute (JL, GB, NC, SV, LA, CV) and Public Health Scotland (RP, JI, DC, KB, SH, KP, SO, SR). We thank all individuals involved in primary care in Scotland for the continued support of the SPARRA project, as well as the general Scottish population.

Authorial contributions were as follows: Manuscript preparation: JL, SRE, BAM, SJV, CAV, LJMA; Project initiation: SJV, CAV, LJMA; Model design: JL, GB, SJV, CAV, LJMA; Code and scripts: JL, GB, LJMA; NC Code review and checking: SRE; Setup of computational system: GB, LJMA; Data access management: DC, RP; EHR access: KB, DC, JI, RP, SO, SR; Public health input: KB, DC, SO, JI, RP, SR; Medical input: JL, BAM, KM; Core planning group: JL, GB, SRE, BAM, KB, DC, JI, KM, RP, SJV, CAV, LJMA; Logistical and legal oversight of project: SH, KP.

All authorial contributions were significant and essential to the completion of this work.

Computing for this project was performed on the Scottish National Safe Haven (NSH), supported by the electronic Data Research and Innovation Service (eDRIS), itself a subsidiary of Public Health Scotland, and the Edinburgh Parallel Computing Centre (EPCC), based at the University of Edinburgh. The authors would like to acknowledge the support of the eDRIS Team (Public Health Scotland) for their involvement in obtaining approvals, provisioning and linking data and the use of the secure analytical platform within the National Safe Haven.

We thank the Alan Turing Institute, Public Health Scotland, the MRC Human Genetics Unit at the University of Edinburgh, Durham University, University of Warwick, Wellcome Trust, Health Data Research UK, and King's College Hospital, London for their continuous support of the authors.

JL, CAV and LJMA were partially supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Health" theme within that grant and The Alan Turing Institute; JL, BAM, CAV, LJMA and SJV were partially supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England), the devolved administrations, and leading medical research charities; SJV, NC and GB were partially supported by the University of Warwick Impact Fund. SRE is funded by the EPSRC doctoral training partnership (DTP) at Durham University, grant reference EP/R513039/1; LJMA was partially supported by a Health Programme Fellowship at The Alan Turing Institute; CAV was supported by a Chancellor's Fellowship provided by the University of Edinburgh.

REFERENCES

[1] ”Health and Social Care Information Programme”. *A Report on the Development of SPARRA Version 3 (Developing Risk Prediction to Support Preventative and Anticipatory Care in Scotland)*. https://www.isdscotland.org/Health-Topics/Health-and-Social-Community-Care/SPARRA/2012-02-09-SPARRA-Version-3.pdf, Accessed: 6-3-2020. 2011.

[2] ”NHS Scotland Information Services Division”. *National Safe Haven*. https://www.isdscotland.org/Products-and-Services/EDRIS/Use-of-the-National-Safe-Haven/, Accessed: 6-3-2020. 2020.

[3] ”Rural Access Action Team”. “The National framework for service change in NHS Scotland”. In: *Scottish Executive, Edinburgh* (2005).

[4] Diego Arenas et al. “Design choices for productive, secure, data-intensive research at scale in the cloud”. In: *arXiv e-prints*, arXiv:1908.08737 (Aug. 2019), arXiv:1908.08737. arXiv: 1908.08737 [cs.CR].

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent Dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[6] Ian Blunt. “Focus on preventable admissions”. In: *London: Nuffield Trust* (2013).

[7] Alex Bottle, Paul Aylin, and Azeem Majeed. “Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis”. In: *Journal of the Royal Society of Medicine* 99.8 (2006), pp. 406–414.

[8] Tianqi Chen et al. *xgboost: Extreme Gradient Boosting*. R package version 0.90.0.2. 2019. URL: https://CRAN.R-project.org/package=xgboost.

[9] Joanna Coast, Abby Inglis, and Stephen Frankel. “Alternatives to hospital care: what are they and who should decide?” In: *Bmj* 312.7024 (1996), pp. 162–166.

[10] Gary S Collins et al. “Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement”. In: *Circulation* 131.2 (2015), pp. 211–219.

[11] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach”. In: *Biometrics* (1988), pp. 837–845.

[12] David A Ellis et al. “Demographic and practice factors predicting repeated non-attendance in primary care: a national retrospective cohort analysis”. In: *The Lancet Public Health* 2.12 (2017), e551–e559.

[13] Scottish Executive. “Scottish index of multiple deprivation 2006 technical report”. In: *Office of the Chief Statistician, Scottish Executive* (2006), p. 10.

[14] George Forman and Martin Scholz. “Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement”. In: *Acm Sigkdd Explorations Newsletter* 12.1 (2010), pp. 49–57.

[15] Scottish Government. *Scottish index of multiple deprivation*. 2016.

[16] Devan Kansagara et al. “Risk prediction models for hospital readmission: a systematic review”. In: *Jama* 306.15 (2011), pp. 1688–1698.

[17] Erin LeDell et al. *h2o: R Interface for ’H2O’*. R package version 3.26.0.2. 2019. URL: https://CRAN.R-project.org/package=h2o.

[18] Matthew C Lenert, Michael E Matheny, and Colin G Walsh. “Prognostic models will be victims of their own success, unless...” In: *Journal of the American Medical Informatics Association* 26.12 (2019), pp. 1645–1650.

[19] James Liley et al. “Model updating after interventions paradoxically introduces bias”. In: *AISTATS proceedings* (2021).

[20] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems*. 2017, pp. 4765–4774.

[21] David Lyon et al. “Predicting the likelihood of emergency admission to hospital of older people: development and validation of the Emergency Admission Risk Likelihood Index (EARLI)”. In: *Family practice* 24.2 (2007), pp. 158–167.

[22] Marian S McDonagh, David H Smith, and Maria Goddard. “Measuring appropriate use of acute beds: a systematic review of methods and results”. In: *Health policy* 53.3 (2000), pp. 157–184.

[23] Office for National Statistics, National Records of Scotland, and Northern Ireland Statistics and Research Agency. *2011 Census aggregate data. UK Data Service (Edition: June 2016)*. 2016. DOI: http://dx.doi.org/10.5257/census/aggregate-2011-1.

[24] World Health Organization. *International statistical classification of diseases and related health problems*. Vol. 1. World Health Organization, 2004.

[25] Fatemeh Rahimian et al. "Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records". In: *PLoS medicine* 15.11 (2018), e1002695.

[26] Colin Sanderson and Jennifer Dixon. "Conditions for which onset or hospital admission is potentially preventable by timely and effective ambulatory care". In: *Journal of health services research & policy* 5.4 (2000), pp. 222–230.

[27] Matthew Sperrin et al. "Explicit causal reasoning is needed to prevent prognostic models being victims of their own success". In: *Journal of the American Medical Informatics Association* 26.12 (2019), pp. 1675–1676.

[28] Adarsh Subbaswamy and Suchi Saria. "From development to deployment: dataset shift, causality, and shift-stable models in health AI". In: *Biostatistics* 21.2 (2020), pp. 345–352.

[29] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1 (2007).

[30] Emma Wallace et al. "Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review". In: *Medical care* 52.8 (2014), p. 751.

Supplementary material

Our supplementary material comprises

1. Table 1: Data sources,
2. Table 2: All predictors and definitions
3. Table 3: Areas under ROC/PRC curves for each constituent model by cross-validation fold
4. Table 4: Exploration of contributors to each topic.
5. Figure 1: Diagram of cross-validation setup,
6. Figure 2: ROC curves for constituents of SPARRA model
7. Figure 3: PR curves for constituents of SPARRA model
8. Figure 4: Calibration curves for constituents of SPARRA model
9. Figure 5: Density plot showing SPARRA v3 against SPARRA v4.
10. Figure 6: Non-attenuation of model performance with time: ROC, PRC and calibration
11. Figure 8: Attenuation of out-of-date risk scores at predicting EA in subsequent years.
12. Figure 7: Bivariate densities of SPARRA scores computed for the same individuals at different time points
13. Figure 9: Performance of SPARRA v3 and v4 restricting to first and final time points
14. Figure 10: Distribution of individuals at a given SPARRA score cutoff across admission types
15. Figure 11: PPVs for predicting each cause of admission given SPARRA score cutoff
16. Figure 12: ROC, PRC and calibration curves comparing models fitted with and without topic features
17. Figure 13: Shapley values vs value for number of emergency and elective admissions.
18. Figure 14: Shapley value vs value for predictors 1-9
19. Figure 15: Shapley value vs value for predictors 10-18
20. Figure 16: Shapley value vs value for predictors 19-27
21. Figure 17: Shapley value vs value for predictors 28-36
22. Figure 18: Shapley value vs value for predictors 37-45
23. Figure 19: Shapley value vs value for predictors 46-54
24. Figure 20: Shapley value vs value for predictors 55-63
25. Note 1: Details of models
26. Note 3: Details of cross-validation procedure
27. Note 4: Details of logistics of work
28. Note 5: Details of assessment of calibration.
29. TRIPOD: Checklist for TRIPOD guidelines [11], and associated pages.

SUPPLEMENTARY TABLES

Name	Reference	Description	Notes
SMR00	[4]	Outpatient attendance	
SMR01	[5]	Acute hospital admissions	Non-psychiatric, non-obstetric
SMR01E	[6]	Geriatric long stay	
SMR04	[5]	Psychiatric hospital admissions	Acute only
AE2	[1]	Accident and emergency records	
PIS	[3]	Prescribing information	
SystemWatch	[7]	Urgent care monitoring	
LTC	[2]	Long-term conditions	

Table 1. Data sources used to derive SPARRA v4 predictors. SMR - Scottish Morbidity Records. Also see Figure 1b

Class	Variable name	Name	Type
Demographics	age	Age at time cutoff	Int
	SIMD_DECILE_2016_SCT	Scottish Index of Multiple Deprivation (SIMD) decile in 2016	Int
Previous admissions	emergency_bed_days	Number of emergency bed days	Int
	num_emergency_admissions	Number of emergency admissions	Int
	elective_bed_days	Number of elective bed days	Int
	num_elective_admissions	Number of elective admissions	Int
	emergency_drugAndalcohol_admin	Number of emergency drug and alcohol related admissions	Int
	num_psych_admissions	Number of previous psychiatric admissions	Int
	num_ae2_attendances	Number of previous A&E attendances	Int
	num_outpatient_appointment_general	Number of previous outpatient appointments	Int
	num_outpatient_appointment_mental	Number of previous outpatient mental health appointments	Int
	days_since_last_AE2	Days since last A&E attendance	Int
	days_since_last_SMR00	Days since last outpatient attendance	Int
	days_since_last_SMR01	Days since last acute/inpatient or day case attendance	Int
	days_since_last_SMR01E	Days since last geriatric long stay attendance	Int
	days_since_last_SMR04	Days since last mental health or mental health day case attendance	Int
	days_since_last_SMR01_emergency_only	Days since last emergency inpatient or day case attendance	Int
	days_since_last_SMR01E_emergency_only	Days since last emergency geriatric long stay attendance	Int
Prescriptions	pis_PAID_GIC_INCL_BB	Total amount paid in prescription costs	Int
	pis_NUMBER_OF_PAID_ITEMS	Total number of paid prescription items	Int
	pis_countBNFsections	Number of BNF sections from which a prescription was filled	Int
	pis_Respiratory	Number of respiratory related prescriptions	Int

pis_CetralNervousSystem	Number of central nervous system related prescriptions	Int
pis_Infections	Number of infection related prescriptions	Int
pis_EndocrineSystem	Number of endocrine related prescriptions	Int
pis_IncontinenceDevices	Number of incontinence device prescriptions	Int
pis_CombinedStomaDevices	Number of stoma device prescriptions	Int
pis_Anticoagulants_And_Protamine	Number of anticoagulant and protamine prescriptions	Int
pis_Antiepileptic_Drugs	Number of antiepileptic prescriptions	Int
pis_Antifibrinolytic_Drugs._Haemostatics	Number of antifibrinolytic and haemostatic prescriptions	Int
pis_Antisecretory_Drugs_Mucosal_Protectants	Number of antisecretory and mucosal protectant prescriptions	Int
pis_Antispasmod_Other_Drugs_Alt_Gut_Motility	Number of antispasmodic and gut motility altering prescriptions	Int
pis_Arm_Sling_Bandages	Number of arm sling bandage prescriptions	Int
pis_Catheters	Number of catheter prescriptions	Int
pis_Corticosteroids_Respiratory	Number of respiratory corticosteroid prescriptions	Int
pis_Dementia	Number of dementia related prescriptions	Int
pis_Drugs_Affecting_Intestinal_Secretions	Number of prescriptions for drugs affecting intestinal secretions	Int
pis_Drugs_Used_In_Diabetes	Number of prescriptions for drugs used in diabetes mellitus	Int
pis_Drugs_Used_In_Neuromuscular_Disorders	Number of prescriptions for drugs used in neuromuscular disorders	Int
pis_Drugs_Used_In_Park.ism_Related_Disorders	Number of prescriptions for drugs used in Parkinsonism and related disorders	Int

	pis_Drugs_Used_In_Substance_Dependence	Number of prescriptions for drugs used in substance dependence	Int
	pis_Fluids_And_Electrolytes	Number of fluid and electrolyte prescriptions	Int
	pis_Minerals	Number of mineral prescriptions	Int
	pis_Mucolytics	Number of mucolytic prescriptions	Int
	pis_Oral_Nutrition	Number of oral nutrition prescriptions	Int
	pis_Vitamins	Number of vitamin prescriptions	Int
Long-term conditions	parkinsons_indicated	Parkinsons disease	Bin
	MS_indicated	Multiple Sclerosis	Bin
	epilepsy_indicated	Epilepsy	Bin
	dementia_indicated	Dementia	Bin
	ltc_FIRST_ARTHRITIS_EPISODE_yearssincediag	Years since first arthritis diagnosis	Int
	ltc_FIRST_ASTHMA_EPISODE_yearssincediag	Years since first asthma diagnosis	Int
	ltc_FIRST_ATRIAL_FIBRILLATION_EPISODE_yearssincediag	Years since first atrial fibrillation diagnosis	Int
	ltc_FIRST_CANCER_EPISODE_yearssincediag	Years since first cancer diagnosis	Int
	ltc_FIRST_CHRONIC_LIVER_DISEASE_EPISODE_yearssincediag	Years since first chronic leiver disease diagnosis	Int
	ltc_FIRST_COPD_EPISODE_yearssincediag	Years since first chronic obstructive pulmonary disease diagnosis	Int
	ltc_FIRST_DEMENTIA_EPISODE_yearssincediag	Years since first dementia diagnosis	Int
	ltc_FIRST_DIABETES_EPISODE_yearssincediag	Years since first diabetes mellitus diagnosis	Int
	ltc_FIRST_EPILEPSY_EPISODE_yearssincediag	Years since first epilepsy diagnosis	Int
	ltc_FIRST_HEART_DISEASE_EPISODE_yearssincediag	Years since first heart disease diagnosis	Int
	ltc_FIRST_HEART_FAILURE_EPISODE_yearssincediag	Years since first heart failure diagnosis	Int
	ltc_FIRST_MULTIPLE_SCLEROSIS_EPISODE_yearssincediag	Years since first multiple sclerosis diagnosis	Int
	ltc_FIRST_PARKINSON_DISEASE_EPISODE_yearssincediag	Years since first Parkinsons disease diagnosis	Int

	ltc_FIRST_RENAL_FAILURE_EPISODE_yearssincediag	Years since first renal failure diagnosis	Int
	ltc_FIRST_CEREBROVASCULAR_DISEASE_EPISODE_yearssincediag	Years since first cerebrovascular disease diagnosi	Int
	ltc_rawdata_NUMBEROFLTC	Number of recorded long term conditions	Int
	numLTCs_resulting_in_admin	Number of long-term conditions resulting in admissions	Int
Outcome	target	Emergency admission in year following cutoff date	Bin

Table 2. Predictors and target used in SPARRA v4 model. Variable name is the name used in any code. Int: integer; Bin: binary

	Fold 1			Fold 2			Fold 3			Mean	
	AUROC	AUPRC	Coef.	AUROC	AUPRC	Coef.	AUROC	AUPRC	Coef.	AUROC	AUPRC
ANN	0.7653	0.3593	0	0.7788	0.3644	0	0.7692	0.3516	0	0.7711	0.3584
GLM	0.7879	0.3717	0	0.7879	0.3727	0	0.7877	0.3726	0	0.7879	0.3723
Naive Bayes	0.752	0.2388	0	0.7519	0.2391	0	0.7523	0.2401	0	0.7521	0.2393
RF, max. 20	0.7963	0.3921	0.3353	0.7966	0.3935	0.2502	0.7966	0.3927	0.984	0.7965	0.3928
RF, max. 40	0.7867	0.3808	0.2828	0.7864	0.3816	0	0.7859	0.3809	0.1452	0.7863	0.3811
SPARRA v3	0.7811	0.3585	0	0.7807	0.3598	0	0.7808	0.3591	0	0.7809	0.3591
XG-boost, max. 4	0.8002	0.3976	1.37	0.8001	0.3989	0.7382	0.8	0.3983	0.9133	0.8001	0.3983
XG-boost, max. 8	0.801	0.3998	1.523	0.8006	0.4002	1.5	0.8005	0.3997	1.914	0.8007	0.3999
XG-boost, max. 3	0.7999	0.3976	1.288	0.8004	0.3994	1.132	0.7996	0.3973	0.9066	0.7999	0.3981
Ensemble	0.8015	0.402	-	0.8016	0.4033	-	0.8012	0.4021	-	0.8014	0.4025

Table 3. Areas under ROC curves and PR curves by fold for each constituent predictor and ensemble. Columns 'Coef.' indicate coefficients in the ensemble. For models, 'max' indicates maximum depth. All standard errors for AUROCs are $< 5 \times 10^{-4}$ and for PRCs are $< 8 \times 10^{-4}$

Words	Label
Vitamins Drugs Affecting Bone Metabolism Corticosteroids (Endocrine)	Osteoporosis-related
Anaemias + Other Blood Disorders Drugs Affecting Intestinal Secretions	*Malabsorption*
Antibacterial Drugs Wound Management & Other Dressings *Essential (primary) hypertension* *Urinary tract infection, site not specified* *Type 2 diabetes mellitus without complications* *Acute renal failure, unspecified* *Unspecified acute lower respiratory infection* *Chronic ischaemic heart disease, unspecified* *Personal history of diseases of the circulatory system* *Old myocardial infarction* *Chronic obstructive pulmonary disease, unspecified* *Mental and behavioural disorders due to harmful use of tobacco*	Complications of Diabetes Mellitus
Diuretics	
Antidepressant Drugs	
Bronchodilators Corticosteroids (Respiratory) Antibacterial Drugs Corticosteroids (Endocrine) Cromoglycate,Rel,Leukotriene Antagonists Other Appliances Mucolytics	Asthma
Drugs Used In Psychoses & Rel.Disorders Minerals *Poisoning by 4-aminophenol derivatives*	Overdose
Antihist, Hyposensit & Allergic Emergen Drugs Acting On The Nose Corti'roids & Other Anti-Inflamm.Preps.	Atopy
Antibacterial Drugs Contraceptives Treatment Of Vaginal & Vulval Conditions Acne and Rosacea Antifungal Drugs Antiviral Drugs Anti-Infective Skin Preparations Antifibrinolytic Drugs & Haemostatics Hypothalamic&Pituitary Hormones&Antioest	Sexually transmitted infections
Drugs Used In Nausea And Vertigo Drugs Used In Neuromuscular Disorders Anti-Arrhythmic Drugs	

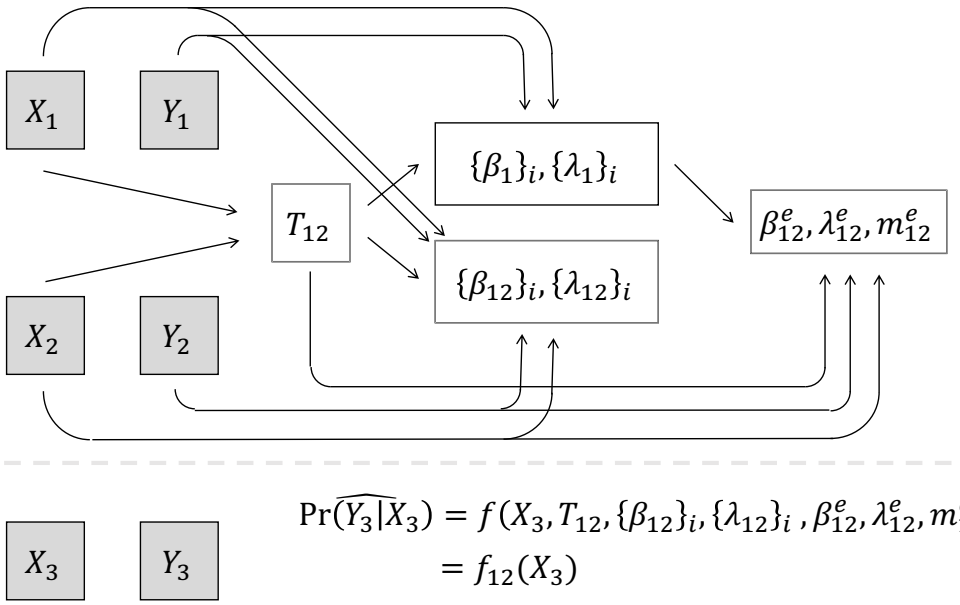
Hypnotics And Anxiolytics Antidepressant Drugs	Anxiety/depression
Sex Hormones & Antag In Malig Disease Skin Fillers And Protectives Local Anaesthesia Night Drainage Bags Catheters Swabs Ileostomy Bags Antibacterial Drugs Leg Bags Wound Management & Other Dressings Adhesive Removers (Sprays/Liquids/Wipes) Colostomy Bags Vaginal Moisturisers Two Piece Ostomy Systems *Chemotherapy session for neoplasm* *Other chemotherapy* *Malignant neoplasm, breast, unspecified*	Malignancy and treatment complications
(BNF) Unknown Antibacterial Drugs	
Laxatives Oral Nutrition Drugs Affecting The Immune Response Local Prepn for Anal & Rectal Disorders Antibacterial Drugs	
Hypertension and Heart Failure Lipid-Regulating Drugs	Cardiovascular disease
Thyroid And Antithyroid Drugs Miscellaneous Ophthalmic Preparations Antibacterial Drugs	*Graves' disease*
Antisecretory Drugs+Mucosal Protectants Antiprotozoal Drugs	Diarrheal infection
Nit,Calc Block & Other Antianginal Drugs Hypertension and Heart Failure	Cariovascular disease
Drugs Used In Diabetes Other Appliances Dementia *Type 2 diabetes mellitus without complications*	
Beta-Adrenoceptor Blocking Drugs Anticoagulants And Protamine Positive Inotropic Drugs	Arrythmias
Treatment Of Glaucoma Sunscreens And Camouflagers Vaccines And Antisera	Eye disease

Corti'roids & Other Anti-Inflamm.Preps. Mydriatics And Cycloplegics *Cataract, unspecified*	
Emollient & Barrier Preparations Topical Corticosteroids Emollients Antibacterial Drugs Anti-Infective Skin Preparations	Skin disease
Antibacterial Drugs Soft-Tissue Disorders & Topical Pain Rel Anti-Infective Skin Preparations Analgesics Drugs Acting On The Oropharynx Cough Preparations Anti-Infective Eye Preparations Drugs Acting On The Ear Topical Corticosteroids Fluids And Electrolytes Preparations For Warts And Calluses Antifungal Drugs Top Local Anaesthetics & Antipruritics Anthelmintics	Skin disease
Drugs Acting On The Nose Sex Hormones Drugs Used In Substance Dependence Antibacterial Drugs Nasal Products	
Lipid-Regulating Drugs Antiplatelet Drugs	Cardiovascular disease
Analgesics Antibacterial Drugs Soft-Tissue Disorders & Topical Pain Rel	
Antiepileptic Drugs Drugs Used In Park'ism/Related Disorders CNS Stimulants and drugs used for ADHD	Central nervous system disease
Antispasmod.&Other Drgs Alt.Gut Motility Eye Products Acute Diarrhoea Miscellaneous Ophthalmic Preparations Dry Mouth Products Antibacterial Drugs	*Sympathomimetics*
Dyspep&Gastro-Oesophageal Reflux Disease Shampoo&Other Preps For Scalp&Hair Cond Preparations For Eczema And Psoriasis Selective Preparations Topical Corticosteroids	Eczema/Psoriasis

Drugs For Genito-Urinary Disorders	
Chronic Bowel Disorders	
Antibacterial Drugs	
Cytotoxic Drugs	

Table 4. Details of derived topics for topic model used for prediction in fold 1 (fitted to folds 2 and 3). A topic model assumes that each 'document' (individual) in a 'corpus' (population) is associated with various 'topics' (roughly, illness categories) where each topic corresponds to a distribution over 'words' (ICD10 codes and medication types). We would expect that the 30 topics fitted to each fold roughly represent the major clusters of disease types which occur amongst those individuals. This tables shows the 'words' with the highest probability of membership in each topic ($> 1\%$, where probabilities over all words sum to 100%). In each topic, words are ordered by decreasing probability of topic membership. ICD10 codes are italicised; medication types are not. Topics are ordered by decreasing importance (mean absolute Shapley value). We assign labels to some topics which appear to code for clear disease types, italicised for tenuous links.

SUPPLEMENTARY FIGURES



$$\widehat{\Pr(Y_3|X_3)} = f(X_3, T_{12}, \{\beta_{12}\}_i, \{\lambda_{12}\}_i, \beta_{12}^e, \lambda_{12}^e, m_{12}^e) = f_{12}(X_3)$$

$$\widehat{\Pr(Y_3|X_3)} \perp\!\!\!\perp Y_3|X_3 \qquad f_{12}(\cdot) \perp\!\!\!\perp X_3, Y_3$$

Figure 1. Variable dependencies. Boxes denote random variables, arrows denote causal dependence. In general, $X_\cdot$ are predictors, $Y_\cdot$ are targets, $T_\cdot$ is a topic model, $\beta_\cdot$ are parameters, and $\lambda_\cdot$ are hyperparameters. See online methods for details.

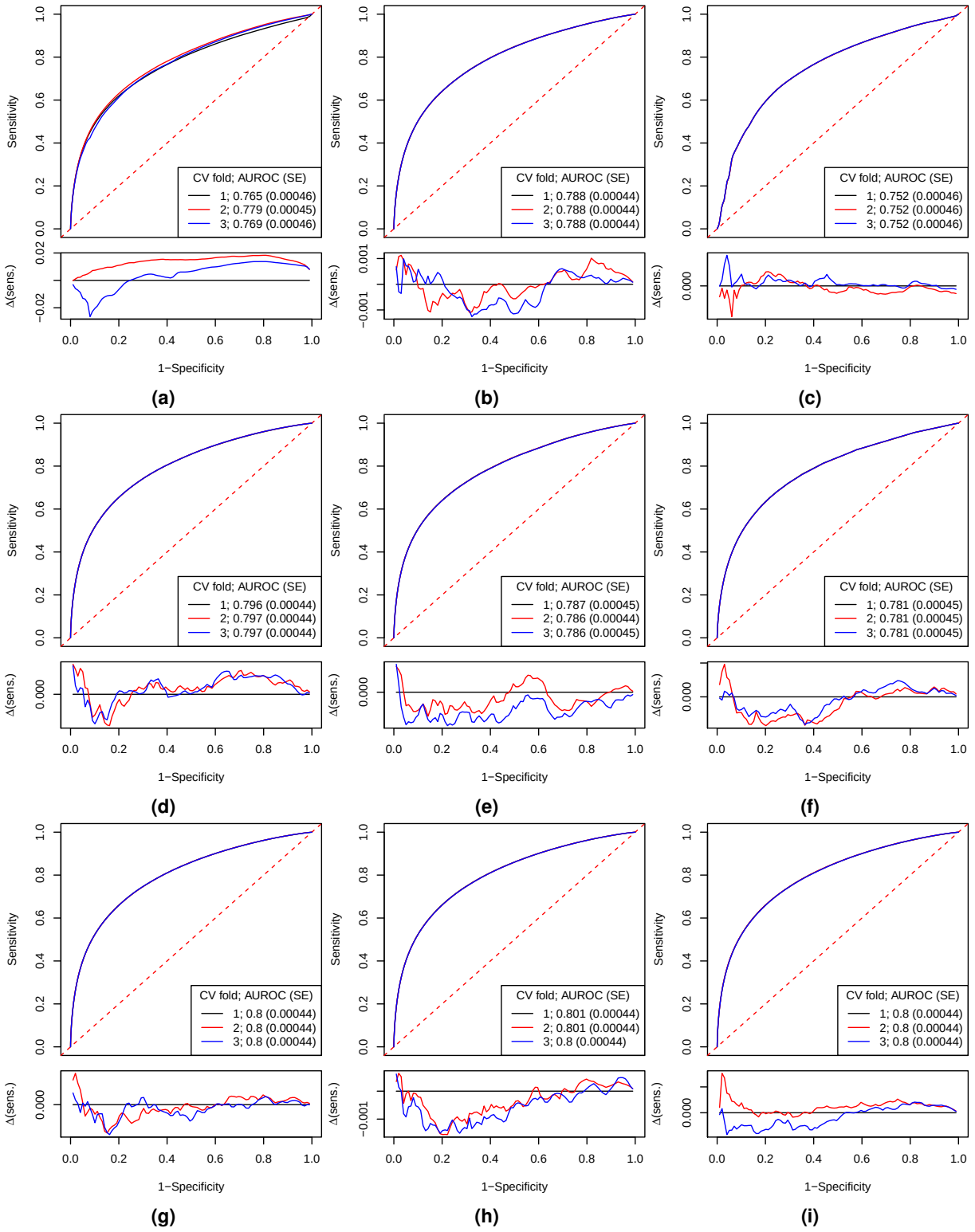


Figure 2. Receiver-operator characteristic curves showing discriminative ability of constituent models of SPARRA v4.

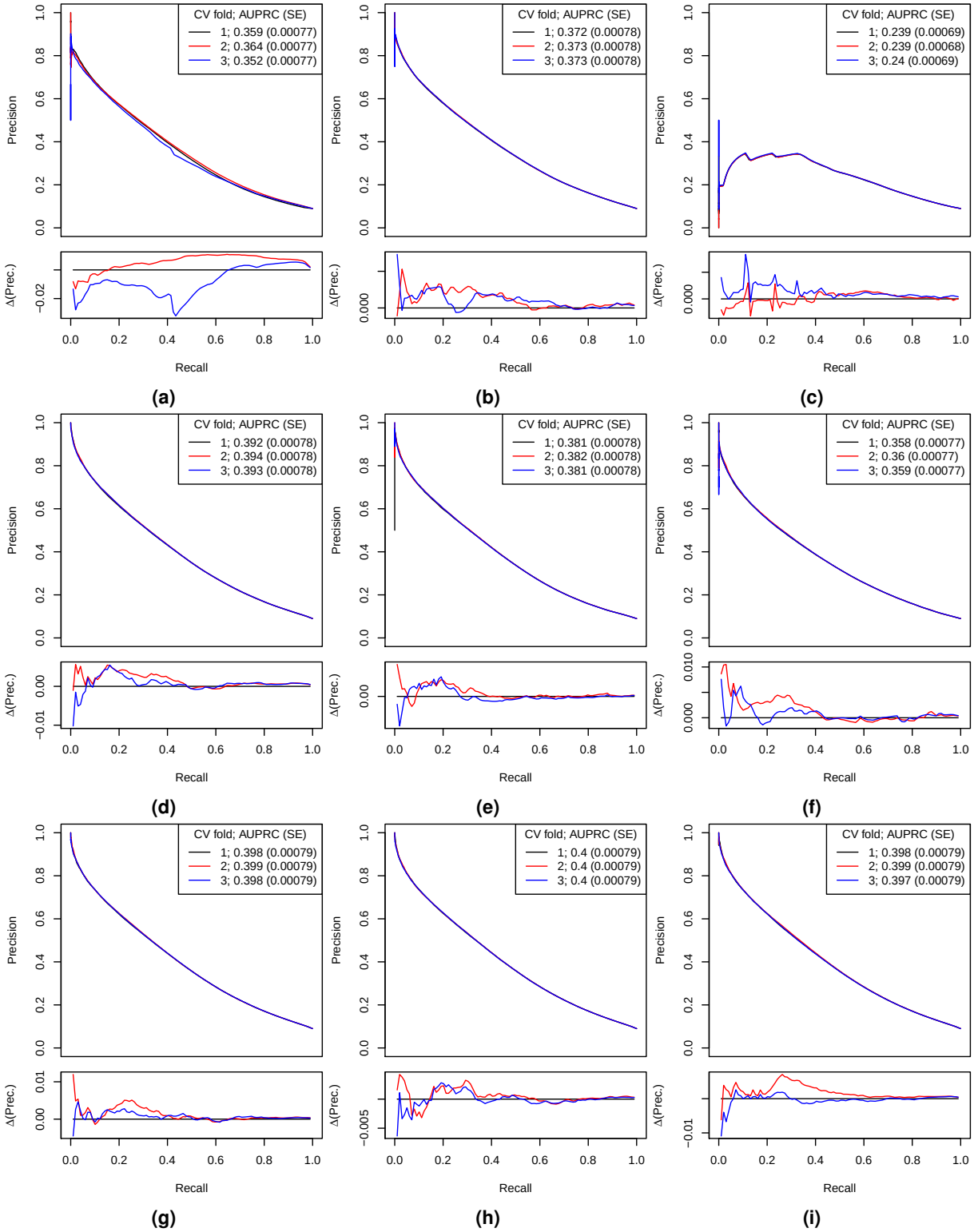


Figure 3. Precision recall curves showing discriminative ability of constituent models of SPARRA v4.

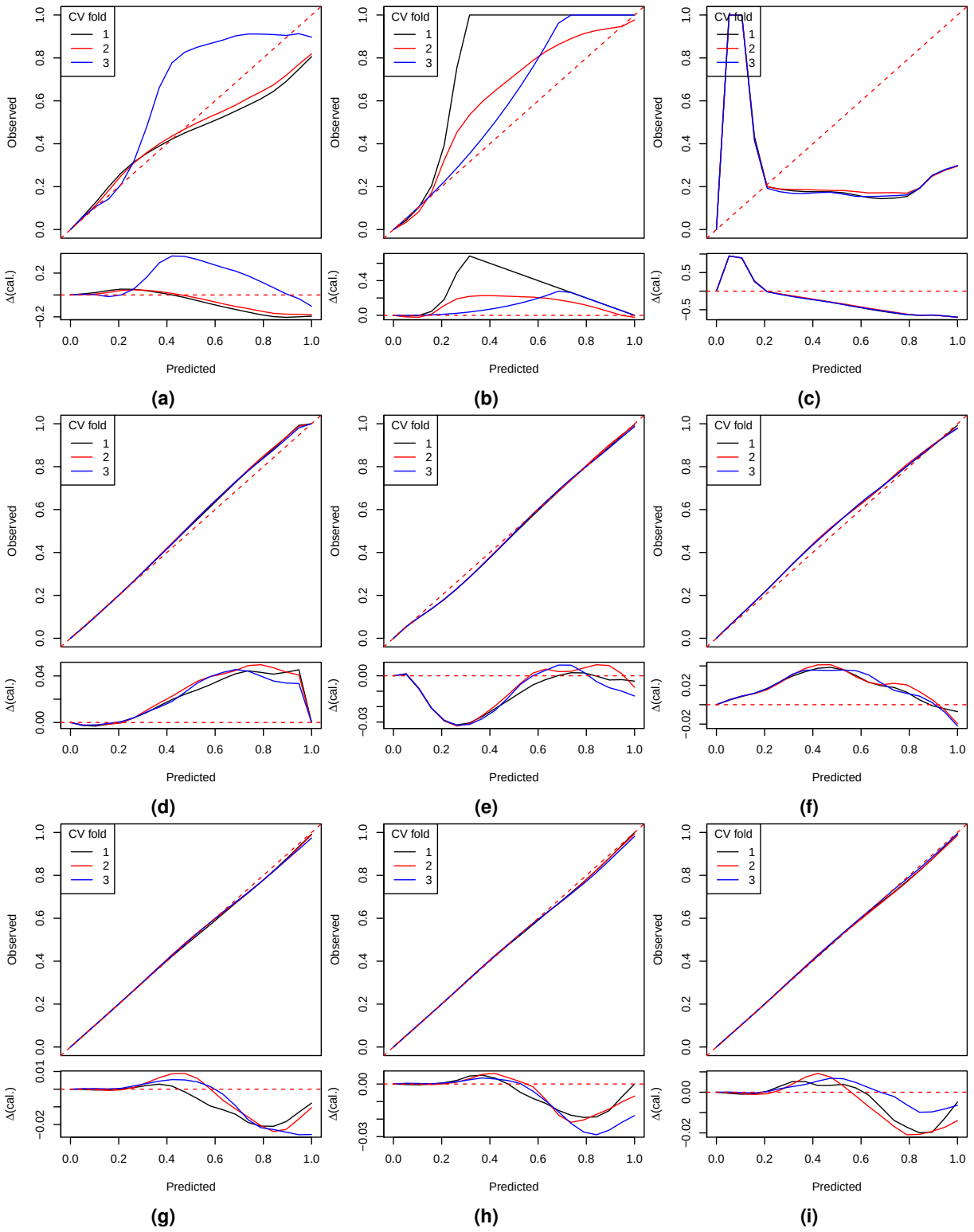


Figure 4. Calibration curves showing discriminative ability of constituent models of SPARRA v4. Confidence envelopes are pointwise.

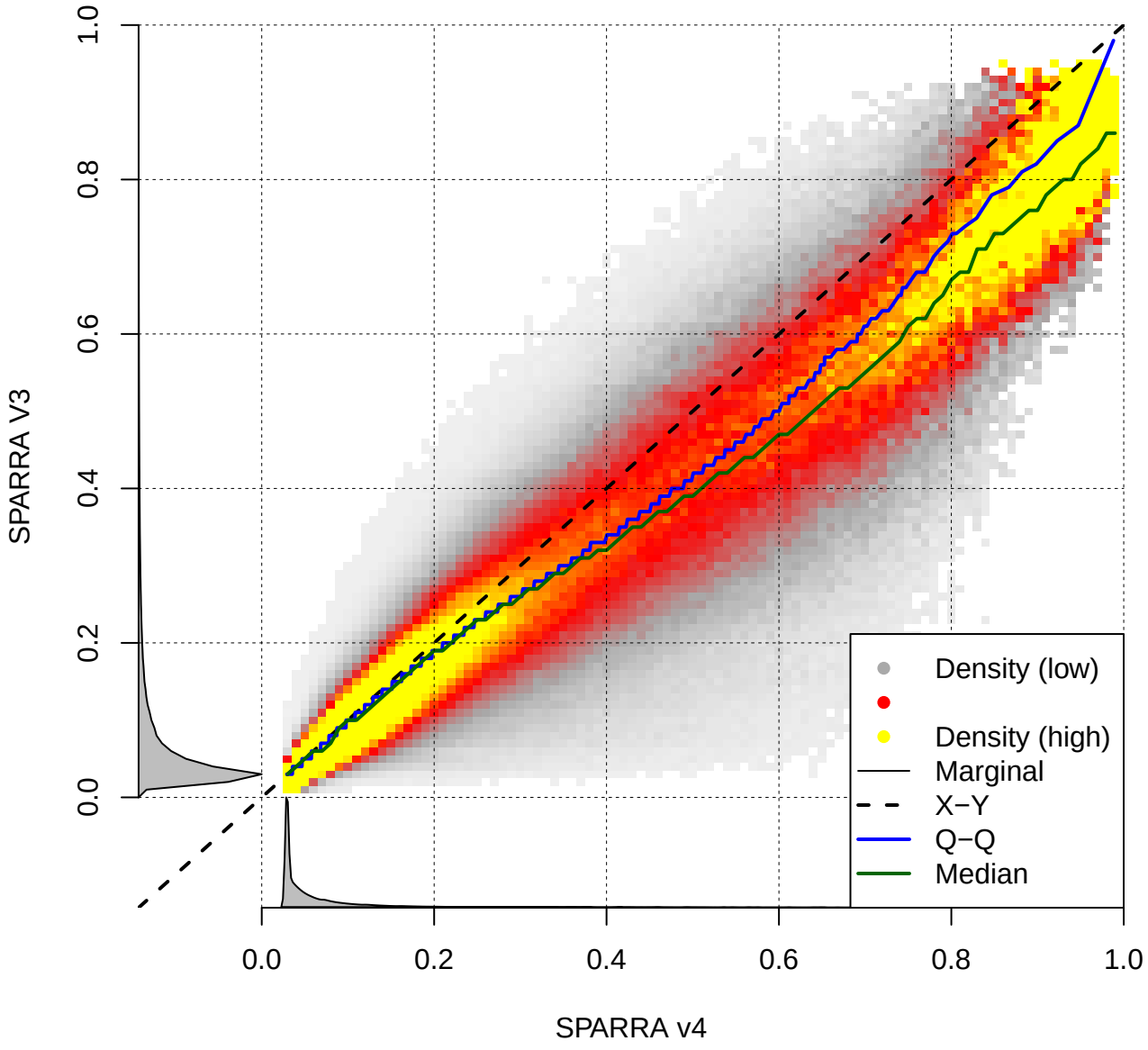


Figure 5. Joint density (low to high: white-grey-red-yellow) of individual SPARRA v3 and v4 scores. The density is normalised to uniform marginal on the Y axis, then the X axis; true marginal distributions of risk scores are shown alongside in grey.

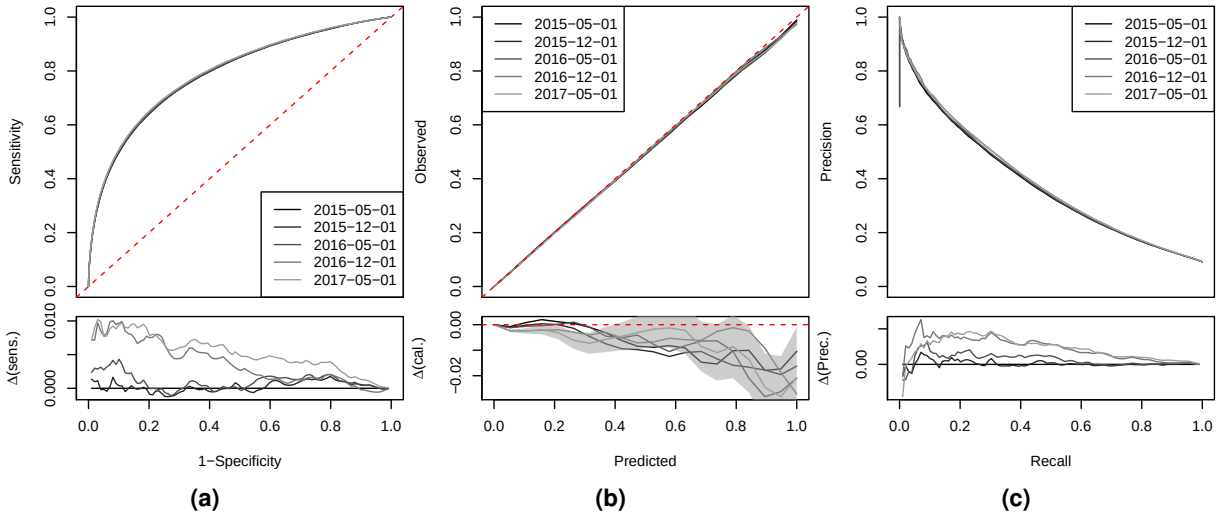
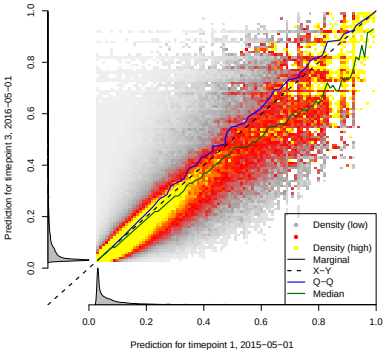
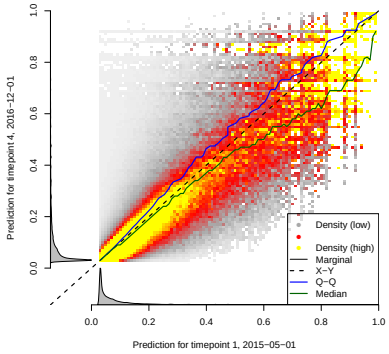


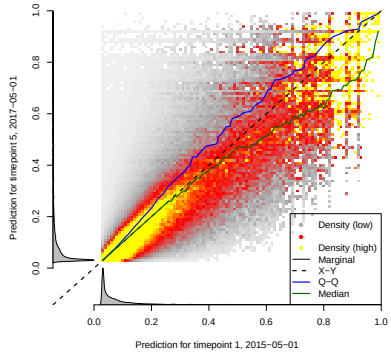
Figure 6. ROC curves, PR curves and calibration curves for a model M_0 fitted to an early time point (2 May 2014) and evaluated at later time points.



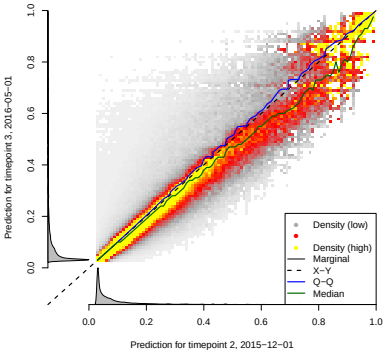
(a) t_1 vs t_3



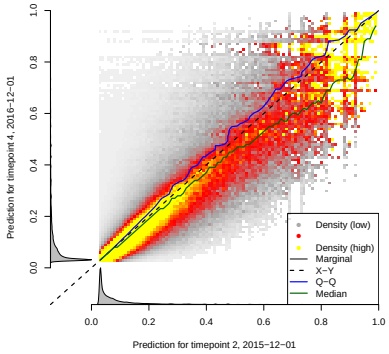
(b) t_1 vs t_4



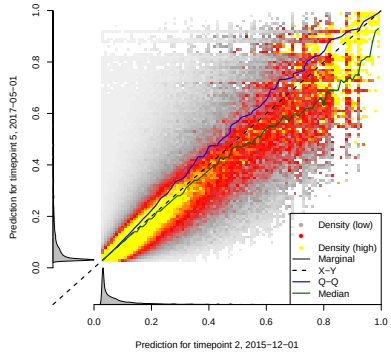
(c) t_1 vs t_5



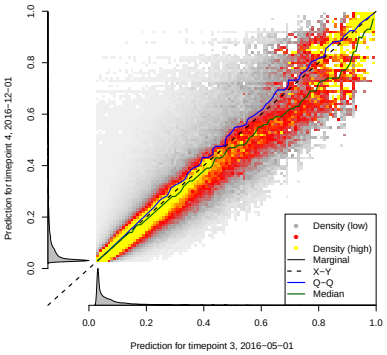
(d) t_2 vs t_3



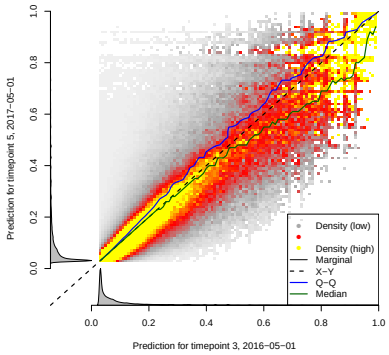
(e) t_2 vs t_4



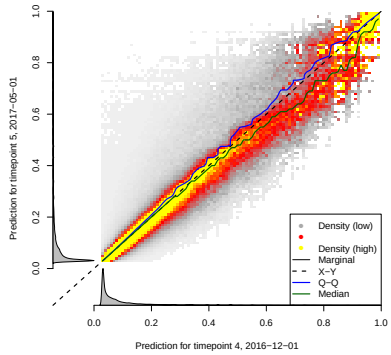
(f) t_2 vs t_5



(g) t_3 vs t_4



(h) t_3 vs t_5



(i) t_4 vs t_5

Figure 7. Joint density (low to high: white-grey-red-yellow) of individual risk scores at t_i and t_j for $(i, j) \in \{1,2,3,4,5\}$, $i < j$. The density is normalised to uniform marginals; true marginal distributions of risk scores are shown alongside in grey.

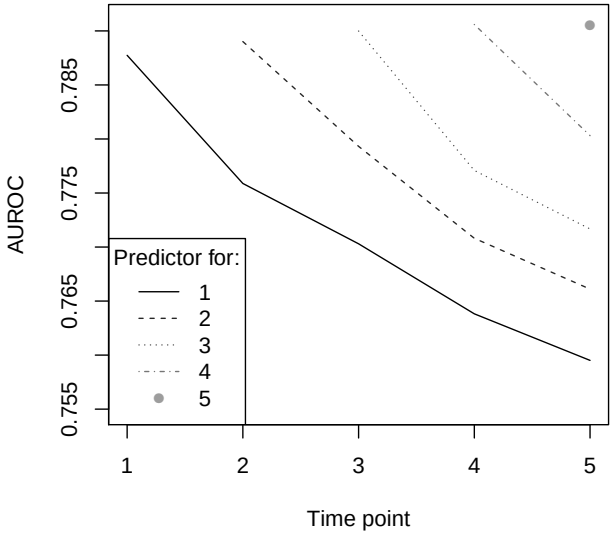
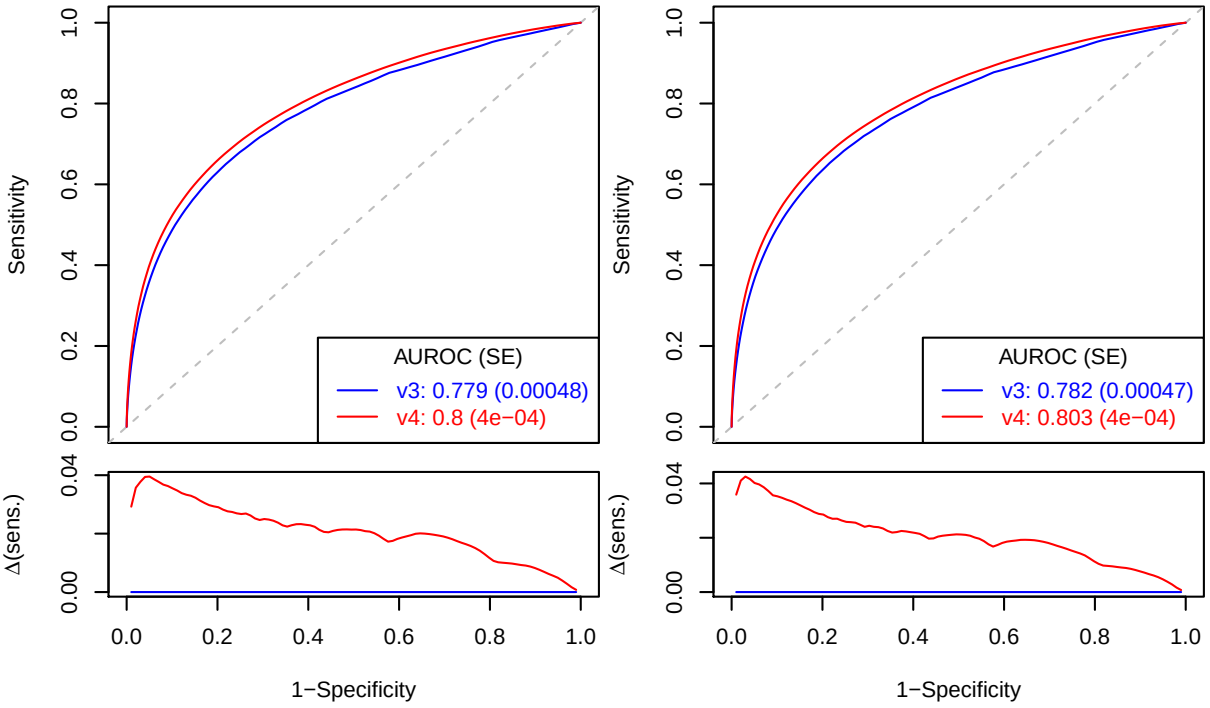


Figure 8. AUROCs for predictions calculated at each time point for prediction EA in years following subsequent time points.



(a) Final time point **(b)** First time point

Figure 9. Performance of SPARRA v4 (new model) and SPARRA v3 (existing model) in subcohorts: final time point only and first time point only. Plots show ROC curves.

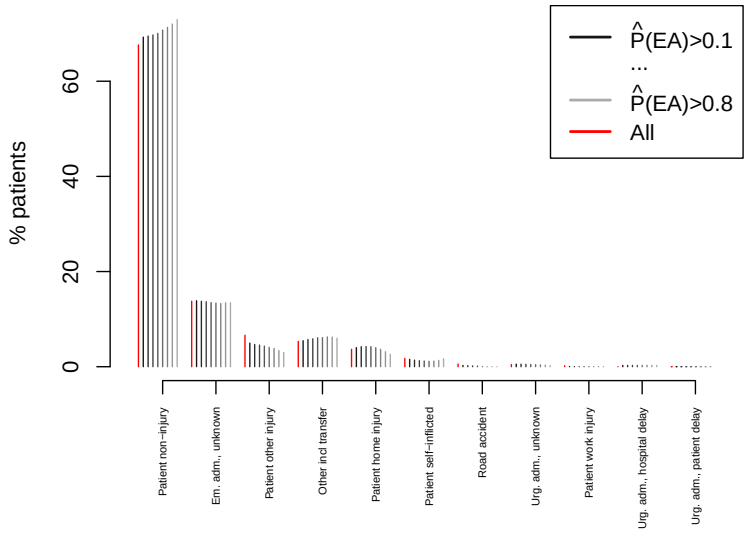


Figure 10. SPARRA v4 performance by category of admission type. Bar height for a given particular x-axis category and a given colour indicates the number of individuals with that x-axis category as a proportion of the cohort of individuals reaching the SPARRA v4 threshold indicated by the colour; note that percentages sum to 100 across bars of each *colour*, not each *category*. For a given x-axis category, a rising profile for greyer colours indicates relatively better prediction of this category and a dropping profile indicates worse prediction.

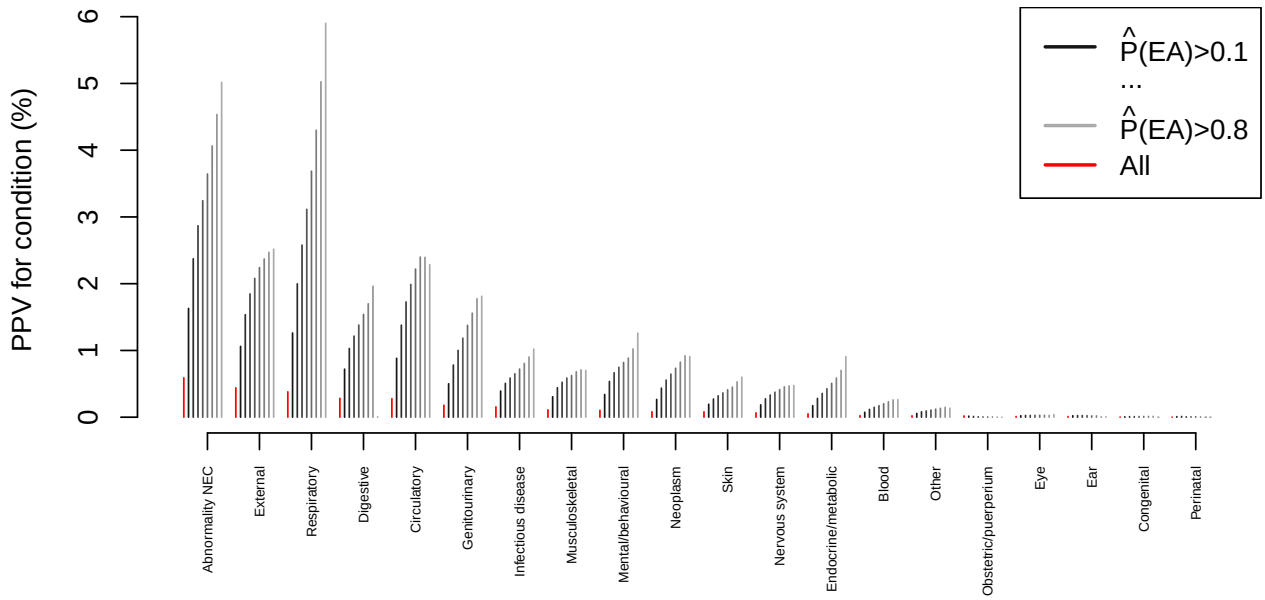
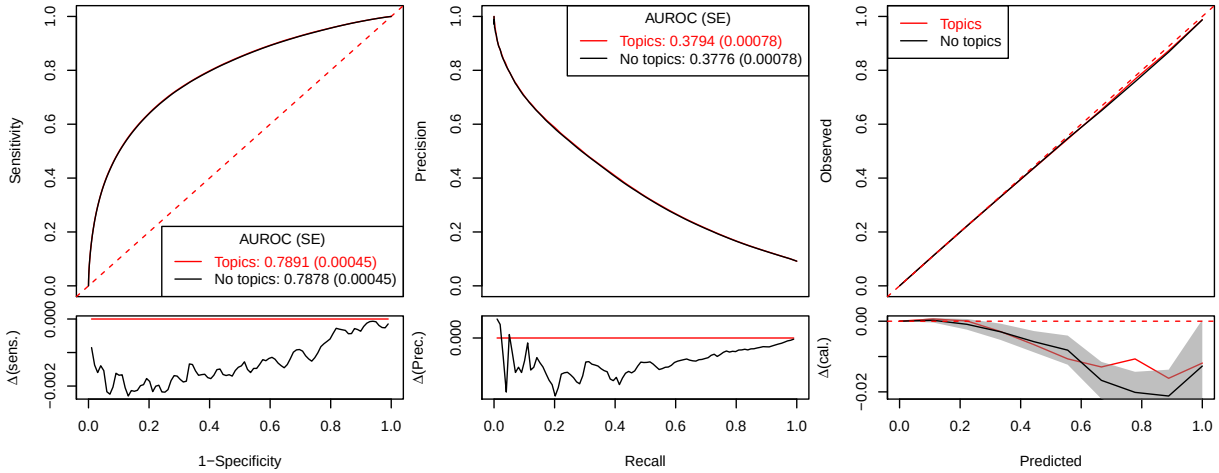


Figure 11. SPARRA v4 performance by type of admission; data as per figure 3d. Bar height for a given particular x-axis category and a given colour indicates the positive predictive value for predicting an admission of that type given a score $\geq$ the associated value. Note that, particularly for low scores, the positive predictive value for being admitted at all is low.



(a) ROC curve **(b)** PR curve **(c)** PR curve

Figure 12. Comparison of model performance with and without topic-model derived features. Models with topic features perform slightly better by AUROC and AUPRC, and are equivocally well-calibrated.

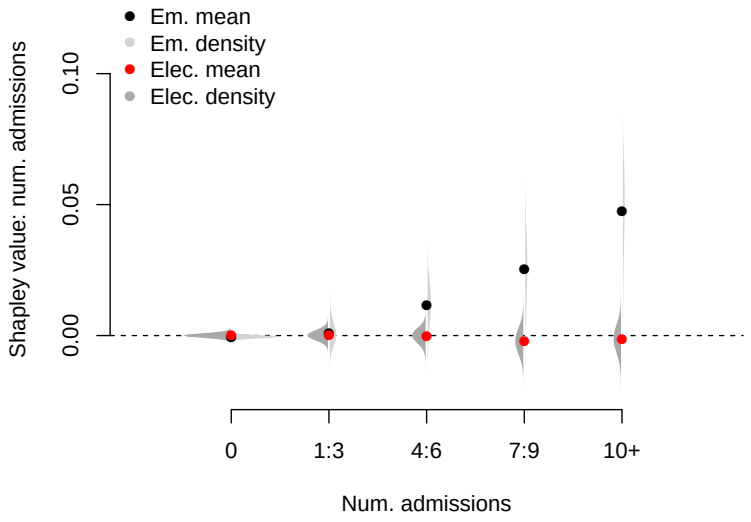


Figure 13. Influence of number of previous elective (Elec.) and emergency (Em.) admissions on risk scores. Also see supplementary table 2 and Supplementary Figures 14-20.

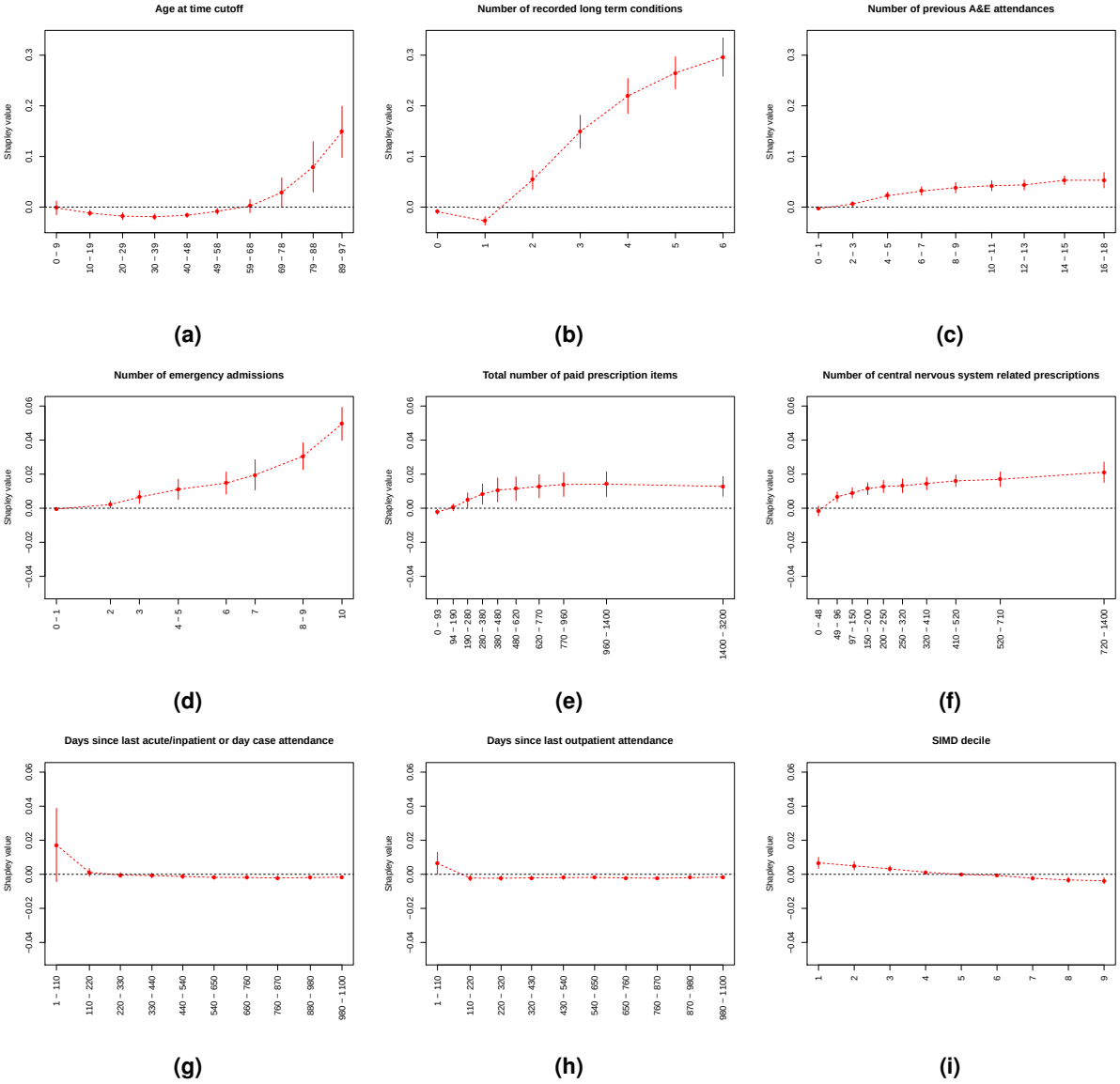


Figure 14. Influence profile on risk scores by variable and value, in decreasing order of mean absolute Shapley value. Vertical lines show standard deviations. Y axes are identical in all plots. Data are anonymised to have > 10 individuals for each x axis mark, so some X axis spacings are irregular. Plots are not shown if data are not sufficiently anonymous

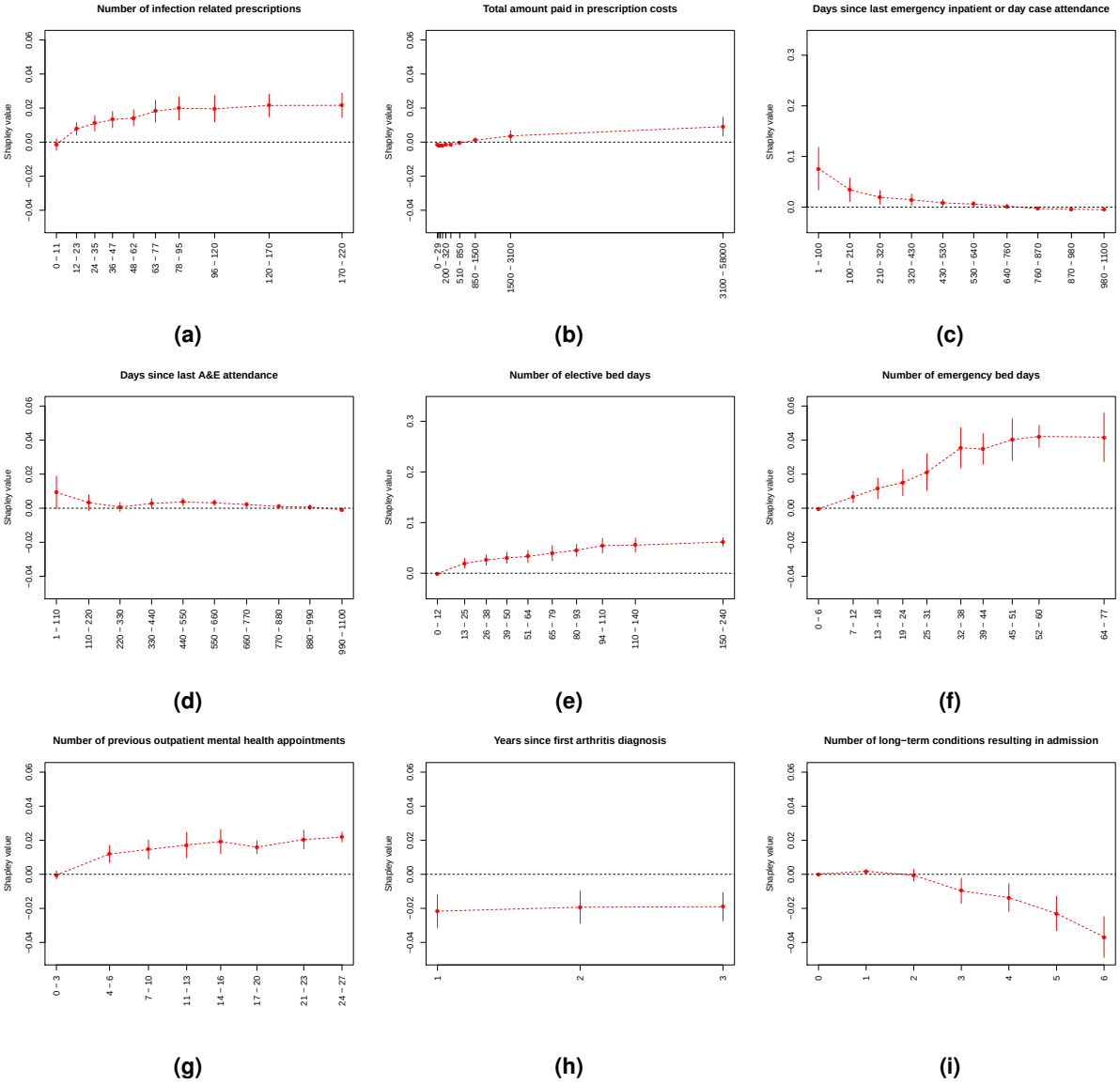


Figure 15. Influence profile on risk scores by variable and value, in decreasing order of mean absolute Shapley value. Vertical lines show standard deviations. Y axes are identical in all plots. Data are anonymised to have > 10 individuals for each x axis mark, so some X axis spacings are irregular. Plots are not shown if data are not sufficiently anonymous

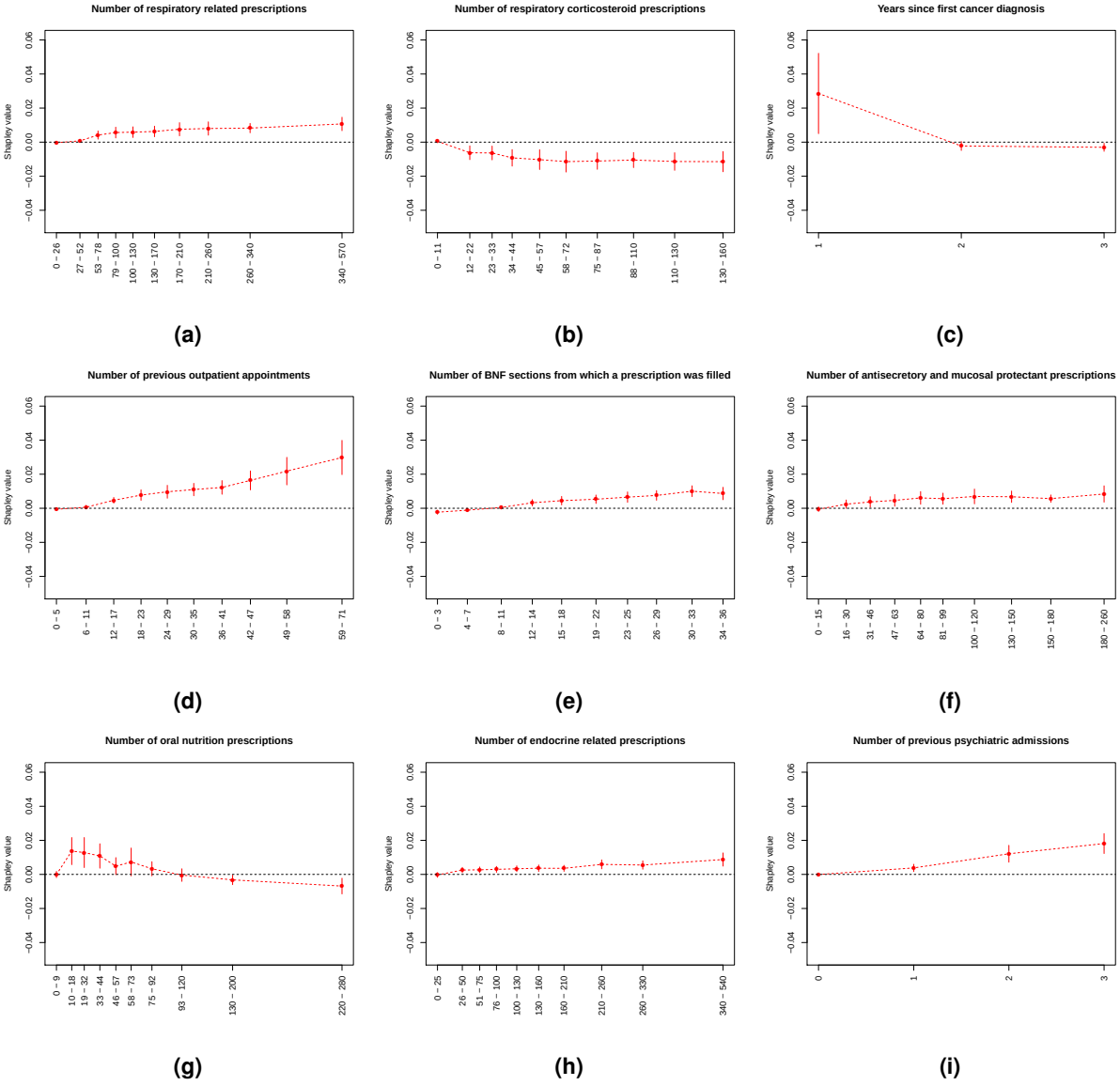


Figure 16. Influence profile on risk scores by variable and value, in decreasing order of mean absolute Shapley value. Vertical lines show standard deviations. Y axes are identical in all plots. Data are anonymised to have > 10 individuals for each x axis mark, so some X axis spacings are irregular. Plots are not shown if data are not sufficiently anonymous

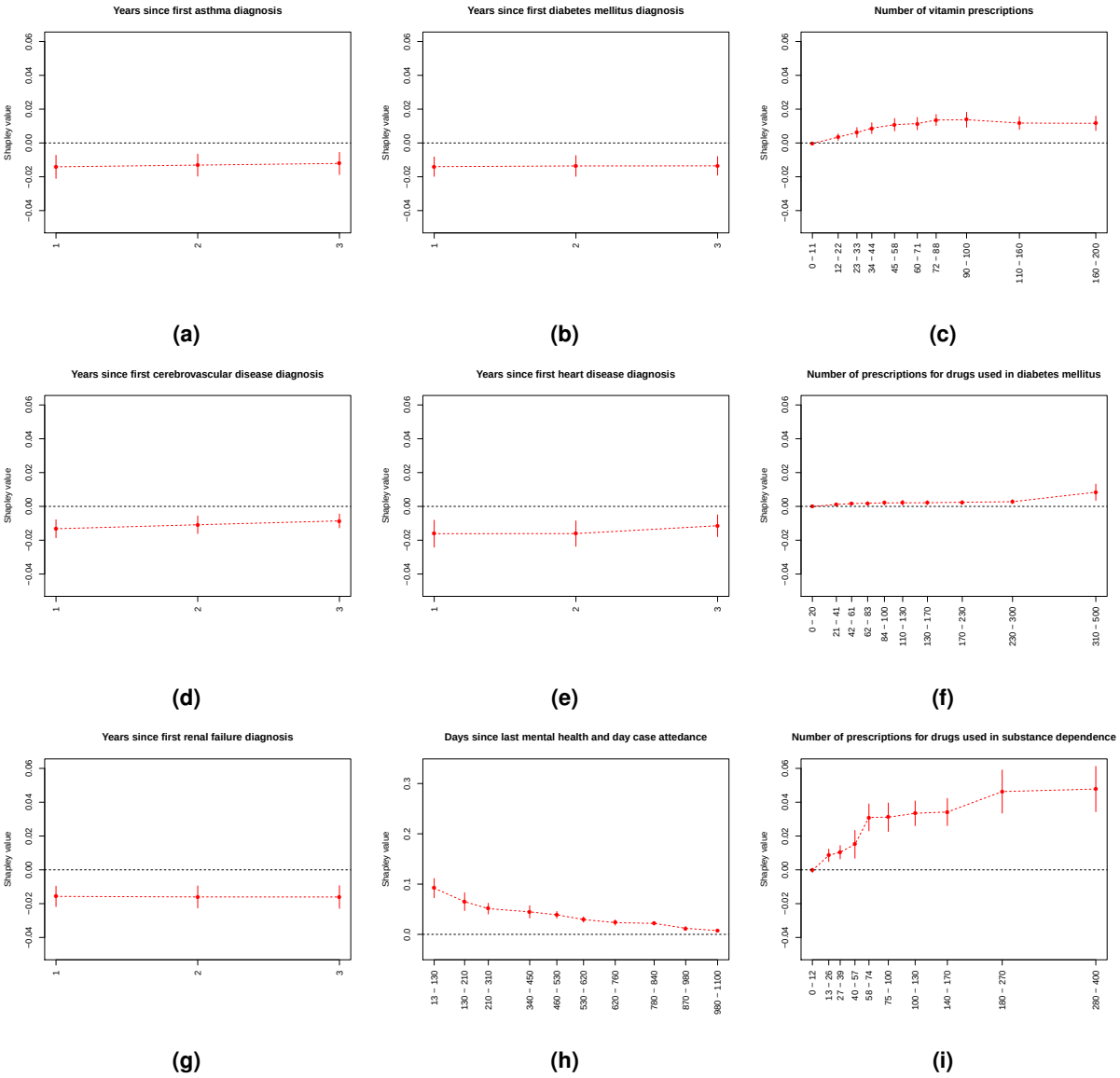


Figure 17. Influence profile on risk scores by variable and value, in decreasing order of mean absolute Shapley value. Vertical lines show standard deviations. Y axes are identical in all plots. Data are anonymised to have > 10 individuals for each x axis mark, so some X axis spacings are irregular. Plots are not shown if data are not sufficiently anonymous

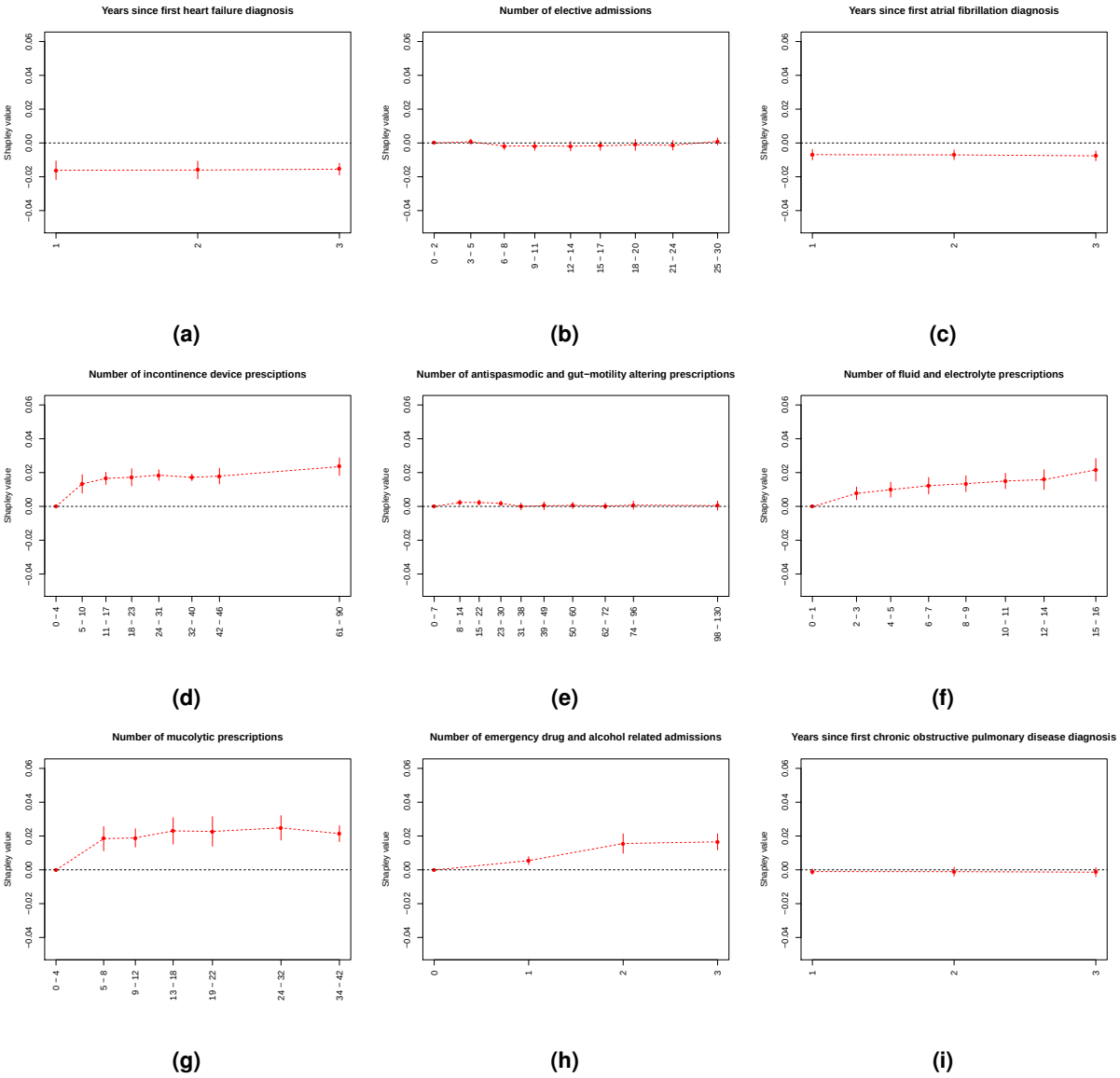


Figure 18. Influence profile on risk scores by variable and value, in decreasing order of mean absolute Shapley value. Vertical lines show standard deviations. Y axes are identical in all plots. Data are anonymised to have > 10 individuals for each x axis mark, so some X axis spacings are irregular. Plots are not shown if data are not sufficiently anonymous

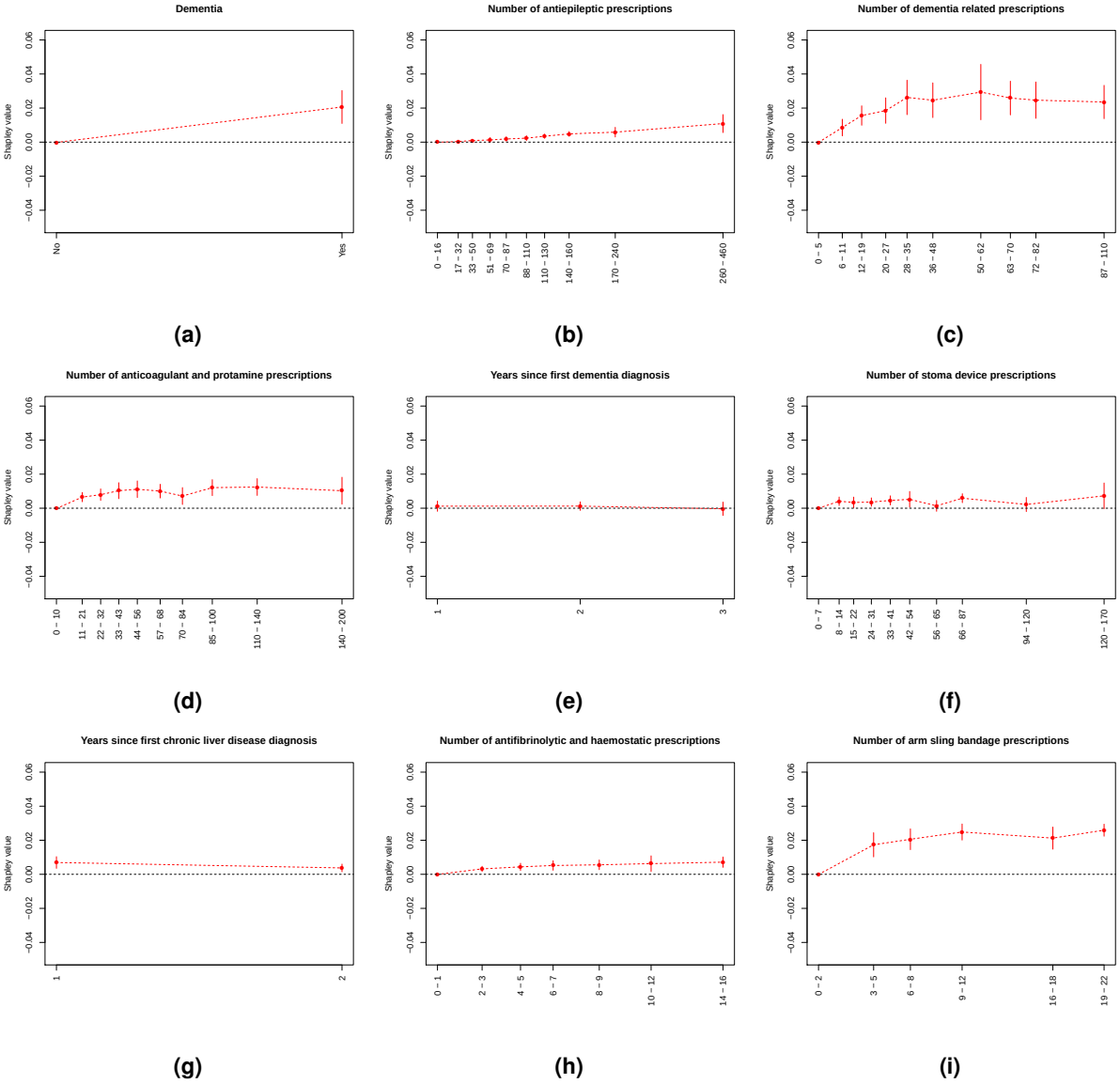


Figure 19. Influence profile on risk scores by variable and value, in decreasing order of mean absolute Shapley value. Vertical lines show standard deviations. Y axes are identical in all plots. Data are anonymised to have > 10 individuals for each x axis mark, so some X axis spacings are irregular. Plots are not shown if data are not sufficiently anonymous

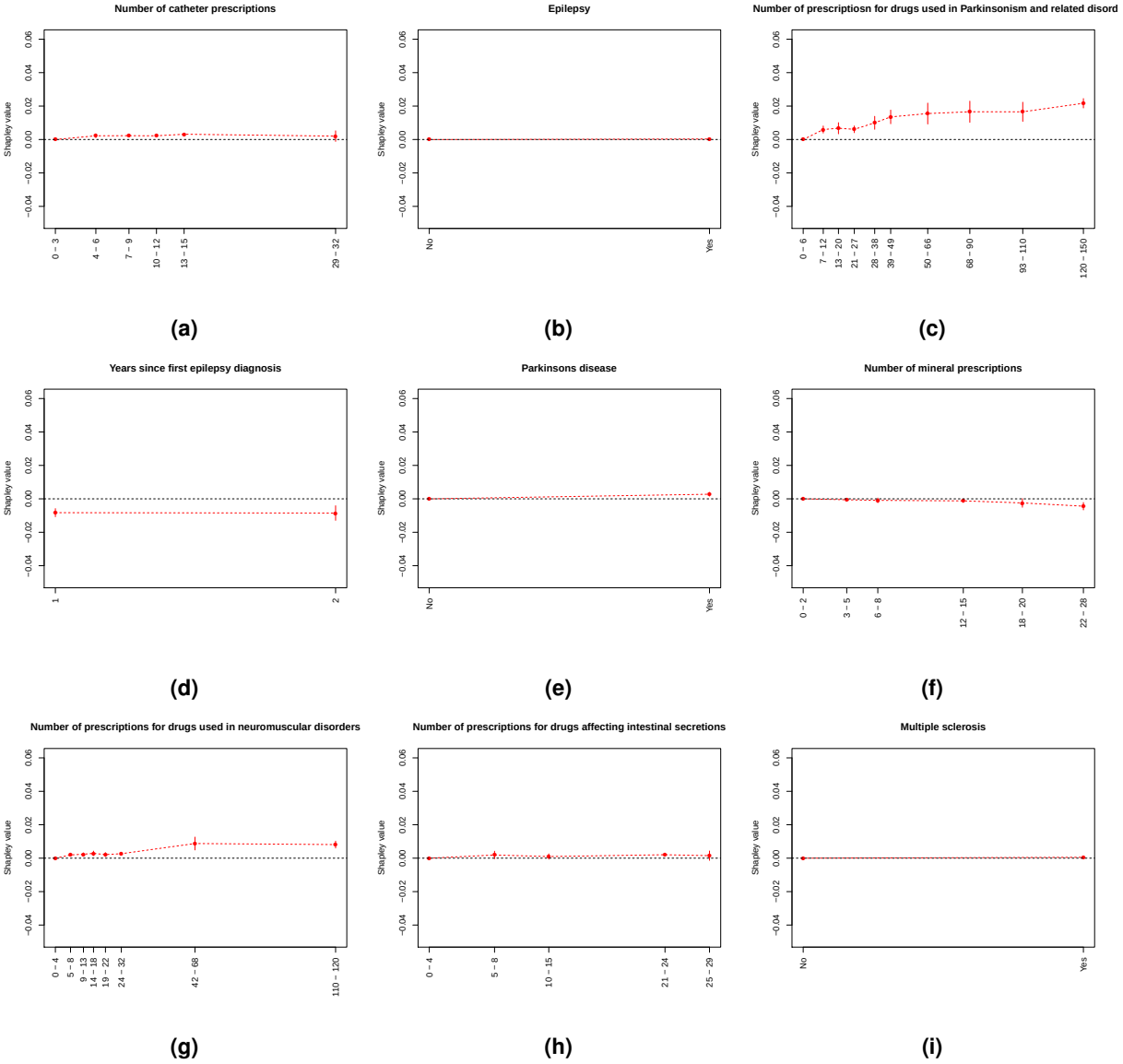


Figure 20. Influence profile on risk scores by variable and value, in decreasing order of mean absolute Shapley value. Vertical lines show standard deviations. Y axes are identical in all plots. Data are anonymised to have > 10 individuals for each x axis mark, so some X axis spacings are irregular. Plots are not shown if data are not sufficiently anonymous

SUPPLEMENTARY NOTE

1 MODEL DETAILS

We used a range of constituent models, detailed below. In all cases, hyperparameters were determined by randomly splitting the relevant dataset into a training and test set of 80% and 20% of the data respectively.

1.1 Existing model

We used SPARRA v3 as one constituent model.

1.2 Artificial neural network

We used an artificial neural network as implemented in the h2o platform (via the h2o R package [13]). We optimised over the number of layers (1 or 2) and the number of nodes in each layer (128 or 256). We used a training dropout rate of 20% to reduce generalisation error.

1.3 Random forest

We fitted two random forests using the h2o R package [13]. One had maximum depth 20 and 500 trees, and the other had maximum depth 40 and 50 trees (both taking a similar time to fit). Settings were otherwise left as h2o defaults (see [13]).

1.4 Boosted trees

We used the R package xgboost [10] to fit three boosted tree models, considering the number of trees as a hyperparameter. We considered trees at three maximum depths: 3, 4, and 8. For the deeper-tree model, we set a low step size shrinkage (η) of 0.075 and a positive minimum loss reduction (γ) of 5 in order to avoid overfitting. In the other two models, we used default values of η=0.3, $\gamma = 0$.

1.5 Naive Bayes

We used the h2o package [13] to fit naive Bayes models. The only hyperparameter we considered was a Laplace smoothing parameter, varying between 0 and 4.

1.6 Generalised linear model

We used the h2o package [13] to fit generalised linear models. We optimised L_1 and L_2 penalties (an elastic net), considering total penalty ($L_1 + L_2$) in $10^{-\{1,2,3,4,5\}}$, and a ratio L_1/L_2 in $\{0, 0.5, 1\}$.

1.7 Super-learner

We aggregated models by finding the linear combination which optimised AUROC in a separate training sample. We included an L_1 penalty as a hyperparameter in this step [12].

2 DEATH IN FOLLOW-UP PERIOD

A problem in choice or target arises due to death of individuals during the follow-up year. Broadly, there are four options for how to treat such individuals:

1. Exclude them from the dataset
2. Classify them as according to whether they had an emergency admission before they died
3. Classify them as no admission, or
4. Classify them as an admission

The philosophy of the SPARRA score is to avert breakdowns in health, of which death can be considered an example. Option 3 would effectively classify death as a 'desirable' outcome, so we avoided it. Option 2 would effectively mean such individuals have a follow-up time less than a year, and would classify individuals who died without a hospital admission as having had a 'desirable' outcome. Option 1 would exclude the most critically ill individuals from the dataset. Our choice of option 4 allows the general description of the target as 'a catastrophic breakdown in health'.

It would also be possible to code death in follow-up differentially; for instance, coding in-hospital death as EA and in-community death as exclusions or non-EA. Our choice not to code all deaths identically is in the interests of non-maleficence. If an individual is at risk of imminent death in the community they will typically be admitted to hospital if it is possible to react in time, with a possible exemption if this is not in their best interests.

The error in SPARRA scores from coding all deaths as EA is thus in assigning high scores to the very old and terminally ill, when in fact these individuals may be treated in the community rather than admitted. The potential harm from this option is small: it could mean that such individuals are excessively treated rather than palliated, but since palliation over treatment is an active decision [14] and such individuals are generally known to be high-risk it is unlikely that the SPARRA score will adversely affect any decisions in this case.

However, the error from coding community deaths as non-EA would be more severe, as it could mean that healthier individuals at risk of sudden death are either coded as non-EA or excluded from the dataset, potentially leading to inappropriately low scores being assigned to these individuals. This could draw treatment away from individuals in high need.

3 CROSS-VALIDATION PROCEDURE

Denote by

X_f, Y_f the data (predictors, outcome) in fold f,

$\mathbb{X}$ the domain of X,

$T.$ a topic model interpreted as a function $T. : \mathbb{X} \to [0,1]^{30}$,

$F^i(\cdot;\beta,\lambda)$ the ith constituent model ($i \in \{1\ldots9\}$ with parameters β_i and hyperparameters λ_i; $F^i : (\mathbb{X} \times T.(\mathbb{X})) \to [0,1]$),

$\mathbb{B}^i$ and $\mathbb{L}^i$ the spaces of parameters and hyperparameters for F^i,

$L^i : ([0,1] \times \{0,1\}) \to \mathbb{R}$ the loss function for F^i

For prediction on fold 3, we began by fitting a Latent Dirichlet Allocation-based topic model T_{12} to predictor data (X_1, X_2), retaining 30 topic features to be used as predictors. See [8] for details.

For each constituent model F^i we partitioned X_1, Y_1 into train/test sets in size ratio 4 : 1 and determined fold-specific parameters $\{\beta_1\}_i$ and hyperparameters $\{\lambda_1\}_i$ in the usual way according to

$$\begin{aligned}
\beta_{\max}(\lambda) &\triangleq \arg\min_{\beta\in\mathbb{B}^i} L^i\left\{F^i\left(X_1^{train}, T_{12}(X_1^{train});\beta,\lambda\right), Y_1^{train}\right\}\\
\{\lambda_1\}_i &= \arg\min_{\lambda\in\mathbb{L}^i} L^i\left\{F^i\left(X_1^{test}, T_{12}\left(X_1^{test}\right);\beta_{\max}(\lambda),\lambda\right), Y_1^{test}\right\}\\
\{\beta_1\}_i &= \arg\min_{\beta\in\mathbb{B}^i} L^i\left\{F^i\left(X_1, T_{12}(X_1);\beta,\lambda\right), Y_1\right\}
\end{aligned}$$

We then evaluated each model on X_2 and established an optimal L-1 penalised linear sum with parameters β_{12}^e and hyperparameter λ_{12}^e (e for 'ensemble'), after splitting into cross-validation folds:

$$\begin{aligned}
\hat{Y}_2^i &= F^i\left(X_2, T_{12}(X_2);\{\beta_1\}_i,\{\lambda_1\}_i\right)\\
\hat{\mathbf{Y}}_{\mathbf{2}} &= (\hat{Y}_2^1 \;\; \hat{Y}_2^2 \;\; \ldots \;\; \hat{Y}_2^9)\\
\beta^e(\lambda) &\triangleq \arg\max_{\beta\in\mathbb{R}^9}\left(\overline{\mathrm{AUROC}\left\{\hat{\mathbf{Y}}_{\mathbf{2}}^{train}\cdot\beta, Y_2^{train}\right\} - \lambda||\beta||_1}\right)\\
\lambda^e &= \arg\max_{\lambda\in\mathbb{R}}\left(\overline{\mathrm{AUROC}\left\{\hat{\mathbf{Y}}_{\mathbf{2}}^{test}\cdot\beta^e(\lambda), Y_2^{test}\right\} - \lambda||\beta^e(\lambda)||_1}\right)\\
\beta^e &\triangleq \arg\max_{\beta\in\mathbb{R}^9}\left(\mathrm{AUROC}\left\{\hat{\mathbf{Y}}_{\mathbf{2}}\cdot\beta, Y_2\right\} - \lambda^e||\beta||_1\right)
\end{aligned} \tag{1}$$

where bars indicate mean over each cross-validation permuation of train/test sets.

Given the output of the ensemble $\tilde{Y}_2(X_2) = \beta^e \cdot \hat{\mathbf{Y}}_{\mathbf{2}}$ (taken as a function of X_2), we then computed a transform $m_{12}^e(x)$ to optimise calibration, essentially using isotonic regression. We defined an empirical calibration function for an estimator $Y'(X)$ of $Y|X$:

$$\begin{aligned}
CAL_{Y'}(x) &= \mathrm{mean}\left(Y_2\middle|\left(|Y'(X_2)-x| < \frac{1}{100}\right)\right)\\
&\approx \mathbb{E}_{Y|X}\left(Y|Y'(X)=x\right)
\end{aligned} \tag{2}$$

We first found a, b such that the mean and mode of $(a\tilde{Y}_2 + b)$ were approximately correctly calibrated; that is, $CAL_{a\tilde{Y}_2+b}(x) = x$ for $x \in \{\text{mean}(a\tilde{Y}_2 + b), \text{mode}(a\tilde{Y}_2 + b)\}$, and scaled a, b such that $0 \leq a\tilde{Y}_2(X_2) + b \leq 1$. Across an evenly spaced grid G of 100 x-values we computed the function:

$$c(x) = (1 - 10^{-5}) \max_{x' \in G; x' \leq x} CAL_{a\tilde{Y}_2+b}(x') + 10^{-5}x \tag{3}$$

using the cumulative maximum of $CAL(x)$ to ensure c is non-decreasing, and adding a linear term to ensure c is increasing. We extended the domain of c to $[0, 1]$ using piecewise linear interpolation, and defined our calibrating transform m^e_{12} as the inverse of c:

$$m^e_{12}(x) = c^{-1}(ax + b) \tag{4}$$

Finally, we re-determined parameters and hyperparameters on folds 1 and 2 (denoting $X_{12} = (X_1, X_2)$, $Y_{12} = (Y_1, Y_2)$

$$\begin{aligned}
\beta_{\max}(\lambda) &\triangleq \arg\min_{\beta \in \mathbb{B}^i} L^i \left\{ F^i \left(X^{train}_{12}, T_{12}(X^{train}_{12}); \beta, \lambda \right), Y^{train}_{12} \right\} \\
\{\lambda_{12}\}_i &= \arg\min_{\lambda \in \mathbb{L}^i} L^i \left\{ F^i \left(X^{test}_{12}, T_{12}\left(X^{test}_{12} \right); \beta_{\max}(\lambda), \lambda \right), Y^{test}_{12} \right\} \\
\{\beta_{12}\}_i &= \arg\min_{\beta \in \mathbb{B}^i} L^i \left\{ F^i \left(X_{12}, T_{12}(X_{12}); \beta, \lambda \right), Y_1 \right\}
\end{aligned}$$

and defined the final predictor function $f_{12}(X)$ to be used on fold 3 as

$$\begin{aligned}
\hat{Y}^i(X) &= F^i \left(X, T_{12}(X); \{\beta_{12}\}_i, \{\lambda_{12}\}_i \right) \\
\hat{\mathbf{Y}}(X) &= (\hat{Y}^1(X) \;\; \hat{Y}^2(X) \;\; \ldots \;\; \hat{Y}^9(X)) \\
f_{12}(X) &= \widehat{\mathbb{E}(Y|X)} \\
&= m^e_{12} \left(\hat{\mathbf{Y}}(X) \cdot \beta^e \right)
\end{aligned} \tag{5}$$

Considered as a function of X, this function depends only on parameters determined using X_{12}, Y_{12}, and hence is independent of X_3 and Y_3 and conditionally independent of Y_3 given X_3.

4 PRACTICALITIES

Data study groups (DSGs) are week long collaborative efforts held in the Alan Turing Institute, where teams of students, early career researchers in data science combine to tackle the real-world problems of non-academic partners. In 2017, one such event was organised with the task of exploring the scope of improving SPARRA performance using modern approaches in machine learning. Despite the time restrictions of this project, results were promising, suggesting advanced machine learning techniques such as random forests offered improved performance over traditional approaches such as logistic regression, as used in the SPARRA v3 model.

This project arose from a collaboration between Public Health Scotland, the Scottish National Health Service, and the Alan Turing Institute, a UK government-founded artificial intelligence research organisation. Collaboration between groups was largely electronic, but face-to-face knowledge transfer meetings were essential in facilitating knowledge transfer between groups.

5 ASSESSMENT OF CALIBRATION

We assume in general that, for IID predictor/outcome pairs $(X_i, Y_i) \sim (X, Y)$, $i \in 1..n$, and an optimal predictor function p_{opt}, we have

$$Y|X \sim \text{Bernoulli}[p_{opt}(X)] \tag{6}$$

noting that this implies

$$p_{opt}(X) = E[Y|p_{opt}(X)] \tag{7}$$

We want to estimate $p_{opt}(X)$.

Since we only observe $Y = 1$ or $Y = 0$, we must estimate $E[Y|p(X) = z]$ as some kind of average of Y about observed values $p(X)$ close to z. A routine way to do this is to use 'reliability diagrams' [9] in which we bin values of $p(X)$ and estimate $E(Y|p(X))$ in each bin.

Since for small bin sizes there may be few or no values of $p(X)$ in some bins, we use a kernel estimate $\hat{c_p}(z)$ of $c_p(z) = E[Y|p(X) = z]$:

$$\hat{c_p}(z) = z\frac{\sum_i Y_i K_\delta[p(X_i), z)]}{\sum_i p(X_i) K_\delta[p(X_i), z]} \tag{8}$$

where $K_\delta : (0,1)^2 \to \mathbb{R}^+$ is some distance-measuring kernel with width δ. We avoid the simpler estimate of the K_δ-weighted mean of Y_is for reasons shown below. We note the following:

Proposition 1. *If $p(X)$ has Lesbegue-integrable positive density on $(0,1)$, $K(z,x)$ and $c_p(x)$ are Lesbegue-integrable functions of x for fixed $z > 0$, and the kernel 'narrows with δ' so*

$$E_X\{p(X)K_\delta[p(X), z]\} \xrightarrow{\delta\to 0} z$$
$$E_X\{c_p[p(X)]K_\delta[p(X), z]\} \xrightarrow{\delta\to 0} c_p(z)$$

then $\hat{c}(z)$ becomes a consistent estimator of $c(z)$ as $\delta \to 0$

Proof. From Slutsky's lemma, the law of total expectation and the strong law of large numbers

$$\hat{c_p}(z) = z\frac{\sum_i Y_i K_\delta(p(X_i), z)}{\sum_i f(X_i) K_\delta(f(X_i), z)} \xrightarrow[\text{prob}]{n\to\infty} z\frac{E_X\left\{c_p[p(X)]K_\delta[f(X), z]\right\}}{E_X\left\{p(X)K_\delta[f(X), z]\right\}} \xrightarrow{\delta\to 0} z\frac{c_p(z)}{z} = c_p(z) \tag{9}$$

□

We note that $\hat{c_p}(Z)$ is not generally consistent if $\delta > 0$. However, the inconsistency is not severe: we note

Proposition 2. *If, in addition to the above, $K_\delta(x,z) = K_\delta(x-z)$ is a symmetric density with second moment δ and negligible moments of higher order, and the densities of $p(X)$ and $c_p(X)$ are twice differentiable at z, then $\hat{c_p}(z) \to c_p(z) + O(\delta^2)$*

Proof. We have

$$\begin{aligned}
E_X\{c_p[p(X)]K_\delta[p(X), z]\} &= E_{x\sim p(X)}[c_p(x)K_\delta(x-z)] \\
&= \int_0^1 f_{p(X)}(x)c_p(x)K_\delta(x-z)dx \\
&= \int_0^1 (f_{p(X)}(z) + f'_{p(X)}(z)(x-z))(c_p(z) + c'_p(x-z))K_\delta(x-z)dx \\
&\quad + \int_0^1 O\left((x-z)^2\right)K_\delta(x-z)dx \\
&= f_{p(X)}(z)c_p(z) + \int_0^1 O\left((x-z)^2\right)K_\delta(x-z)dx \\
&\quad + (f_{p(X)}(z)c'_p(z) + f'_{p(X)}(z)c_p(z))\int_0^1 (x-z)K_\delta(x-z)dz \\
&= f_{p(X)}(z)c_p(z) + O\left(\delta^2\right)
\end{aligned} \tag{10}$$

noting the symmetry of K_δ. If we replace $c_p[p(X)]$ with $p(X)$, the expectation is $zf_{p(X)}(z) + O(\delta^2)$, and the result follows from the first part of 9. □

Remark 1. *In the ideal case where $c_p(z) = z$ (that is, our model is perfectly calibrated) estimator 8 is consistent even when $\delta > 0$, whereas the apparently simpler asymptotically consistent (as $\delta \to 0$) estimator of a weighted sum of Y_i's:*

$$c_p(z) \approx \frac{\sum_i Y_i K_\delta[p(X_i), z)]}{\sum_i K_\delta[p(X_i), z]} \tag{11}$$

is not.

Finally, we note the following:

Proposition 3. *Under the assumptions above, with fixed X_i, the bias of $\hat{c_p}(z)$ is*

$$\frac{\sum_i B(X_i,z)K_\delta[p(X_i),z])}{\sum_i p(X_i)K_\delta[p(X_i),z]} \tag{12}$$

where $B(X_i,z) = p(X_i)c_p(z) - zc_p(p(X_i))$.

Proof. With fixed X_i

$$\begin{aligned} E_Y[c_p(z) - \hat{c_p}(z)] &= E_Y\left[c_p(z) - z\frac{\sum_i Y_i K_\delta[p(X_i),z)]}{\sum_i p(X_i)K_\delta[p(X_i),z]}\right] \\ &= c_p(z) - z\frac{\sum_i c_p(p(X_i))K_\delta[p(X_i),z)]}{\sum_i p(X_i)K_\delta[p(X_i),z]} \\ &= \frac{\sum_i [p(X_i)c_p(z) - zc_p(p(X_i))]K_\delta[p(X_i),z])}{\sum_i p(X_i)K_\delta[p(X_i),z]} \end{aligned} \tag{13}$$

□

Remark 2. *This enables straightforward evaluation of bounds on bias given bounds on the form of c_p. The estimator $\hat{c_p}$ is unbiased if $c_p(x) = kx$ for some k, since $B(X_i,z) \equiv 0$.*

Remark 3. *An alternative way to draw a kernelised calibration curve is to simply plot a parametric curve*

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} \sum p(X_i)K_\delta[p(X_i),t] \\ \sum Y_i K_\delta[p(X_i),t] \end{pmatrix} \tag{14}$$

which, for each t, is an only-slightly biased estimate of some point $z, c_p(z)$. If a rectangular kernel is used, this is equivalent to binning values of $p(X_i)$ [9]. However, this method does not generally give a curve across the entire range of $p(X_i)$.

It is straightforward to estimate

$$\begin{aligned} \mathrm{var}(c_p(z) \mid \{X_1, X_2 ... X_n\}) &= \mathrm{var}\left(z\frac{\sum_i Y_i K_\delta[p(X_i),z]}{(\sum_i p(X_i)K_\delta[p(X_i),z])} \;\middle|\; X_1, X_2..X_n\right) \\ &= \mathrm{var}(\sum_i w_i Y_i \mid \{X_1, X_2..X_n\}) \\ &= \sum_i w_i^2 \mathrm{var}(Y_i \mid X_1, X_2, ..X_n) \\ &\approx \sum_i w_i^2 p(X_i)(1 - p(X_i)) \end{aligned}$$

where the approximation is exact if $c_p(z) = z$. Together with an estimate of maximum absolute bias b_z at z, this enables estimates of conservative confidence intervals on $\hat{c_p}(z)$ at level $1-\alpha$:

$$\hat{c_p}(z) \pm \left(b_z + \Phi^{-1}\left(\frac{\alpha}{2}\right)\mathrm{SE}(c_p(z)|X_i)\right) \tag{15}$$

In all plots in this paper, we bounded bias under the assumption that there existed k such that $|c_p(z) - kz| < z^2/10$.

The calibration estimator derived here is demonstrated in an R script `sparra_calibration.R` available with the attached R code for this manuscript.

REFERENCES

[1] "Public health information Scotland". *AE2 - accident and emergency records*. `hhttps://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=3`, Accessed: 6-3-2020. 2020.

[2] "Public health information Scotland". *LTC: Long-term conditions*. `https://www.isdscotland.org/Health-Topics/Hospital-Care/Diagnoses/`, Accessed: 6-3-2020. 2020.

[3] ”Public health information Scotland”. *PIS - prescribing information systems*. https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=9, Accessed: 6-3-2020. 2020.

[4] ”Public health information Scotland”. *SMR00 - Outpatient attendance*. https://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets/SMR00-Outpatient-Attendance, Accessed: 6-3-2020. 2020.

[5] ”Public health information Scotland”. *SMR01 - Acute Hospital Admissions*. https://www.scotpho.org.uk/publications/overview-of-key-data-sources/scottish-national-data-schemes/hospital-discharges, Accessed: 6-3-2020. 2020.

[6] ”Public health information Scotland”. *SMR01 - Acute Hospital Admissions; SMR04 - Psychiatric Hospital Admissions*. https://www.scotpho.org.uk/publications/overview-of-key-data-sources/scottish-national-data-schemes/hospital-discharges, Accessed: 6-3-2020. 2020.

[7] ”Public health information Scotland”. *System Watch: urgent care usage*. https://www.isdscotland.org/Products-and-Services/System-Watch/, Accessed: 6-3-2020. 2020.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent Dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[9] Jochen Bröcker and Leonard A Smith. “Increasing the reliability of reliability diagrams”. In: *Weather and forecasting* 22.3 (2007), pp. 651–661.

[10] Tianqi Chen et al. *xgboost: Extreme Gradient Boosting*. R package version 0.90.0.2. 2019. URL: https://CRAN.R-project.org/package=xgboost.

[11] Gary S Collins et al. “Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement”. In: *Circulation* 131.2 (2015), pp. 211–219.

[12] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *Journal of statistical software* 33.1 (2010), p. 1.

[13] Erin LeDell et al. *h2o: R Interface for 'H2O'*. R package version 3.26.0.2. 2019. URL: https://CRAN.R-project.org/package=h2o.

[14] Rafael D Romo et al. “Sense of control in end-of-life decision-making”. In: *Journal of the American Geriatrics Society* 65.3 (2017), e70–e75.

heading=References

TRIPOD GUIDELINES

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	1
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	1
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	2
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	2
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	2
	5b	D;V	Describe eligibility criteria for participants.	2
	5c	D;V	Give details of treatments received, if relevant.	-
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	2
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	-
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	2, table 2
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	-
Sample size	8	D;V	Explain how the study size was arrived at.	1
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	2
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	2
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	2,11
	10c	V	For validation, describe how the predictions were calculated.	12
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	12
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	12
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	-
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	4

Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	2, figure 1
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	2, figure 1
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Not applicable; see page 4
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	2
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	-
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Not possible: see page 12
	15b	D	Explain how to the use the prediction model.	4
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	2
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	12, 2
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	4
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	4
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	4
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	4
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	S1
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	13

TRIPOD guidelines and pages where discussed