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

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

James Liley[1,2,3,⊥,*], Gergo Bohner[1,4,⊥], Samuel R. Emerson[3], Bilal A. Mateen[1,5], Katie Borland[6], David Carr[6], Scott Heald[6], Samuel D. Oduro[6b], Jill Ireland[6], Keith Moffat[6,7], Rachel Porteous[6], Stephen Riddell[6b], Simon Rogers[8], Ioanna Thoma[1,2], Nathan Cunningham[1,9], Chris Holmes[1,10], Katrina Payne[1], Sebastian J. Vollmer[1,4], Catalina A. Vallejos[1,2,*,†], and Louis J. M. Aslett[1,3,*,†]

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

JL: *james.liley@durham.ac.uk*
CAV: *catalina.vallejos@ed.ac.uk*
LJMA: *louis.aslett@durham.ac.uk*

September 2, 2024

Abstract

Emergency admissions (EA), where a patient requires urgent in-hospital care, are a major challenge for healthcare systems. The development of risk prediction models can partly alleviate this problem by supporting primary care interventions and public health planning. Here, we introduce SPARRA$v4$, a predictive score for EA risk that will be deployed nationwide in Scotland. SPARRA$v4$ was derived using supervised and unsupervised machine-learning methods applied to routinely collected electronic health records from approximately 4.8M Scottish residents (2013-18). We demonstrate improvements in discrimination and calibration with respect to previous scores deployed in Scotland, as well as stability over a 3-year timeframe. Our analysis also provides insights about the epidemiology of EA risk in Scotland, by studying predictive performance across different population sub-groups and reasons for admission, as well as by quantifying the effect of individual input features. Finally, we discuss broader challenges including reproducibility and how to safely update risk prediction models that are already deployed at population level.

Introduction

Emergency admissions (EA), where a patient requires urgent in-hospital care, represent deteriorations in individual health and are a major challenge for healthcare systems. For example, approximately 395,000 Scottish residents (≈1 in 14) had at least one EA between 1 April 2021 and 31 March 2022 [1]. In total, around 600,000 EAs were recorded for these individuals, nearly 54% of all hospital admissions in that period, and they resulted in longer hospital stays (6.8 days average) compared to planned elective admissions (3.6 days average). Modern health and social care policies aim to implement proactive strategies [2], often by appropriate primary care intervention [3, 4, 5]. Machine learning (ML) can support such interventions by identifying individuals at risk of EA who may benefit from anticipatory care. If successful, such interventions can be expected to improve patient outcomes and reduced pressures on secondary care (Figure 1a).

A range of risk prediction models have been developed in this context [6, 7, 8, 9, 10, 11]. However, transferability across temporal and geographical settings is limited due to differing demographics and data availability [8]. Development of models in the setting in which they will be used is thus preferable to reapplication of models trained in other settings. In Scotland, the Information Services Division of the National Services Scotland (now incorporated into Public Health Scotland; PHS) developed SPARRA (Scottish Patients At Risk of Re-admission and Admission) — an algorithm to predict the risk of EA in the next 12 months. SPARRA was derived using national electronic health records (EHR) databases and has been in use since 2006. The current version of the algorithm (SPARRA$v3$) [12] was deployed in 2012/13 and is calculated monthly by PHS for almost the entire Scottish population. Individual-level SPARRA scores can be accessed by general practitioners (GPs), helping them to plan mitigation strategies for individuals with complex care needs. Collectively, SPARRA scores may be used to estimate future demand, supporting planning and resource allocation. SPARRA has also been used extensively in public health research [13, 14, 15, 16, 17, 18].

In this paper we update the SPARRA algorithm to version 4 (SPARRA$v4$) using contemporary supervised and unsupervised ML methods. In particular, we use an ensemble of machine learning methods [19], and use a topic model [20] to derive further information from prescriptions and diagnostic data. This represents a large scale ML risk score, fitted and deployed at national level, and widely available in clinical settings. We develop SPARRA$v4$ using EHRs collected for around 4.8 million (after exclusions) Scottish residents between 2013 and 2018. Among other variables, this includes data about past hospital admissions, long term conditions (e.g. asthma) and prescriptions. We use cross-validation to evaluate the validity of SPARRA$v4$ and its stability over time. This shows an improvement of performance with respect to SPARRA$v3$ in terms of discrimination and calibration, including a stratified analysis across different subpopulations. We also perform extensive analyses to determine what reasons for emergency admission are predictable, and use Shapley values [21] to quantify the effect of individual input factors. Finally, we discuss some of the practical challenges that arise when developing and deploying models of this kind, including issues associated to updating risk scores that are already deployed at population level.

Reproducibility is critical to ensure reliable application of ML in clinical settings [22]. To provide a transparent description of our pipeline, this manuscript conforms to the TRIPOD guidelines [23] (Supplementary Table 1). Moreover, all code is publicly available at `github.com/jamesliley/SPARRAv4`. This includes non-disclosive outputs used to generate all the figures and tables presented in this article.

Results

Data overview

The input data prior to any exclusions combines multiple national EHR databases held by Public Health Scotland for 5.8 million Scottish residents between 1 May 2013 and 30 April 2018 (Supplementary Table 2), some of whom died during the observation period. These comprised 468 million records, comprising interactions with the Scottish healthcare system and deaths. The number of total available records varies across sex, age, and SIMD (Figure 1b), and when records are grouped by database (Supplementary Figure 1a). In particular, marginally more records are available for individuals in the most deprived areas (as measured by deciles of the 2016 Scottish Index of Multiple Deprivation (SIMD); [24]), particularly within accidents and emergency and mental health hospital records. Two additional tables (see Supplementary Table 2) containing historic data about long term conditions (LTC, back to 1981) and mortality records were also used as input.

We selected three time cutoffs for model fitting (1 May and 1 December 2016, and 1 May 2017) leading to 17.4 million individual-time pairs, hereafter referred to as samples (Figure 1c). This choice was informed by the extent of data required to define the input features used by the score (3 years prior the time cutoff) and the prediction target (1 year after the time cutoff). We used the earliest (1 May 2016) and latest (1 May 2017) possible time cutoffs, and a third time cutoff halfway between these. Although we could have used more than one time cutoff between the earliest and latest, we deemed that this would add little because, for most patients, we expect to have negiblible variation in their input features and EA status from month to month. After exclusions (which were predominantly due to samples without SPARRA v3 scores; see Methods), the data comprise 12.8 million samples corresponding to 4.8 million individuals. Overall, the study cohort is slightly older, has more females, and is moderately more deprived than the general population (Table 1). The prediction target was defined as a recorded EA to a Scottish hospital or death in the year following the time cutoff (see Methods). In total, 1,142,169 EA or death events (9%) were observed across all samples. This includes 57,183 samples for which a death was recorded (without a prior EA within that year) and 1,084,986 samples for which an EA was recorded (amongst those, 107,827 deaths were observed after the EA). As expected, the proportion of deaths amongst the observed events increases with age (Supplementary Figure 1b). Moreover, patients with an EA or death event (in at least one time cutoff) are, on average, older and more deprived than those without an event (Table 1).

Overall predictive performance

In held out test data, SPARRA$v4$ was effective at predicting EA, and outperformed SPARRA$v3$ on the basis of area-under-receiver-operator-characteristic (AUROC) and area-under-precision-recall-curve (AUPRC) (Figure 2a-b). SPARRA$v4$ was also better calibrated, particularly for samples with observed risk ≈ 0.5 (Figure 2c). Whilst SPARRA$v3$ and SPARRA$v4$ scores were highly correlated, large discrepancies were observed for some samples (Supplementary Figure 2). In samples for whom $v3$ and $v4$ disagreed (defined as $|v3 - v4| > 0.1$), we found that $v4$ was better-calibrated than $v3$ (Figure 2d).

We also assessed the potential population-wide benefit of SPARRA$v4$ over SPARRA$v3$ directly. Amongst the 50,000 individuals judged to be at highest risk by SPARRA$v3$, around 4,000 fewer individuals were eventually admitted that were amongst the 50,000 individuals judged to be at highest risk by SPARRA$v4$ (Figure 2e). For another perspective, if we

		Cohort			
Variable	Scottish population	Input data	After exclusions	EA or death	No EA or death
Sex (%)					
Male	48.5	48.2	45.4	46.2	45.3
Female	51.5	51.8	54.6	53.7	54.7
Age at time cutoff (%)					
0-19	16.9	21.1	19.6	11.8	20.4
20-70	71.2	64.2	64.9	50.1	66.4
71+	11.9	14.7	15.4	38.1	13.2
SIMD decile (%)					
1-5	50.0	50.8	52.0	59.5	51.2
6-10	50.0	49.2	48.0	40.5	48.8
Any LTC (%)	Unknown	29.4	32.1	58.8	29.5

Table 1: **Demographic summary for the different cohorts**: the whole Scottish population (approximately 5.8 million), those present in the input databases at least one (17,488,596 samples comprising 5,829,532 unique individuals), our study cohort after exclusions (12,866,084 samples comprising 4,835,428 unique individuals) and our study cohort after stratifying by event status (EA or death: 1,142,169 samples comprising 667,566 unique individuals; no EA or death: 11,723,915 samples comprising 4,670,756 unique individuals). Summary statistics were calculated using sample-level data. The EA or death cohort includes individual-time pairs for which the individual at least one EA or died during the year after the time. LTC denotes long-term conditions (e.g. epilepsy). Data for the Scottish population is from the 2011 Census [25].

simply assume that 20% of admissions are avoidable [value taken from 26], that avoidable admissions are as predictable as non-avoidable admissions, and that we wish to pre-empt 3,000 avoidable admissions by targeted intervention on the highest risk patients (the second assumption is conservative, since avoidable admissions are often predictable due to other medical problems). Then, by using SPARRA$v4$, we would need to intervene on approximately 1,500 fewer patients than if we were to use SPARRA$v3$ in the same way, in order to achieve the target of avoiding 3,000 admissions (Figure 2f).

SPARRA$v4$ comprises an ensemble of models (see Methods), so we also explored a breakdown of AUROC/AUPRC (Table 2) and calibration (Supplementary Figure 3) across constituent models. The ensemble had slightly better performance ($>$ 1 standard error) than the best constituent models (XGB and RF) and substantially better performance than simple statistical models (GLM and NB), which can be considered as benchmarks. Note that some constituent models (ANN, GLM, NB) had ensemble coefficients which were regularised to be vanishingly small, so in practice scores for those models need not be computed when calculating SPARRA$v4$. We investigated whether performance could be improved by using separate sets of coefficients for each SPARRA$v3$ cohort, but found that the improvement was so small that we judged this to be unnecessary (Supplementary Note 3).

Stratified performance of SPARRA$v3$ and SPARRA$v4$

To examine differences in performance more closely, we explored the performance of SPARRA$v3$ and SPARRA$v4$ across different patient subcohorts defined by age, SIMD deciles and the four subcohorts defined as part of SPARRA$v3$ development. Generally, we observed that SPARRA$v4$ had better discrimination performance across all subcohorts (Figure 3a).

Conditional performance of SPARRA$v4$ by admission type and imminence

Figure 3b displays the distribution of SPARRA$v4$ scores stratified according to event status and, for those with an EA, according to the diagnosis that was assigned to the patient during admission (Supplementary Table 5). When comparing samples with and without an event (defined by the composite EA or death outcome), we observed the former had generally lower SPARRA$v4$ scores. Amongst those with an event, all-cause mortality was associated with high SPARRA$v4$ scores. If the event was an EA, we found that samples with certain medical classes of admission tended to have particularly high SPARRA scores, suggesting that such admissions can be predicted disproportionately well (Figure 3b): in particular, those with mental/behavioural, respiratory and endocrine/metabolic related admissions. As one would expect, we were less able to predict external causes of admissions (e.g., S21: open wound of thorax [27]). Obstetric and puerperium-related admissions were particularly challenging to predict by SPARRA$v4$. When further analysing SPARRA$v4$ scores, we also found that among individuals who had an EA during the 1 year outcome period, those with higher risk scores were likelier to have the first EA near the start of the period (Figure 3c). We did not use an absolute threshold to determine who is at high risk. Instead, we ranked individuals according to their scores and looked at those in the top part of the ranking (i.e. with the highest risk scores).

Deployment scenario stability and performance attenuation

We next addressed two crucial aspects pertaining to practical usage of SPARRA$v4$. Firstly, we assess the durability of performance for a model trained once (at the time cutoff 1 May 2014, using a one-year lookback) and employed to generate scores at future times (1 May and 1 December 2015, 1 May and 1 December 2016, 1 May 2017), confirming it does not deteriorate. This is the way in which SPARRA$v4$ will be deployed by PHS, generating new scores each month but without repeated model updating, akin to SPARRA$v3$'s monthly use without update from 2013–2023. Secondly, we demonstrate that it is none-the-less necessary to update scores despite the absence of model updates, since evolving patient covariates lead to the performance attenuation of any point-in-time score.

We firstly used a *static model* M_0 (Methods) to predict risk at future time-points (i.e. new scores are generated as the features are updated). M_0 performed essentially equally well over time (Figure 4a-c), with no statistically significant decrease in performance (adjusted p-values > 0.05), or improved performance with time for all comparisons of AUROCs. With stability under the deployment scenario confirmed, we also explored the distribution of scores over time. In line with expectations, the quantiles of scores generated by the static model increased as the cohort grew older (Figure 4d). The mean risk scores of individuals in the highest centiles of risk at t_0 decreased over time (Figure 4e), suggesting that very high risk scores tend to be transient. The bivariate densities of time-specific scores (Figure 4f) also show lower scores to be more stable than higher scores, and that subjects 'jump' to higher scores (upper left in Figure 4f) more than they drop to lower scores (bottom right).

Finally, we examined the behaviour of *static scores* (computed at t_0 using M_0) to predict future event risk (note that the model is also static in this setting, though we will call it *static scores* for brevity). We observed that the static scores performed reasonably well even 2-3 years after t_0, although discrimination and calibration were gradually lost (Supplementary Figure 4a-c). More generally, we observe that scores fitted and calculated at a fixed time cutoff had successively lower AUROCs for predicting EA over future periods (Supplementary Figure 4d). Although the absolute differences in AUROC over time with static scores are small, they are visibly larger than those seen between SPARRA$v3$ and SPARRA$v4$ (Figure 2a), indicating that comparisons analogous to Figure 2e,f would similarly show much larger differences. This affirms the need for updated scores in deployment, despite the static model.

Feature importance

The features with the largest mean absolute Shapley value (excluding SPARRA$v3$ and the features derived from the topic model) were age, the number of days since the last EA, the number of previous A&E attendances, and the number of antibacterial prescriptions (Table 3). Most features had non-linear effects (see e.g. Supplementary Figure 5a-b). For example, the risk contribution from age was high in infancy, dropping rapidly from infancy through childhood, then remaining stable until around age 65, and rising rapidly thereafter (Figure 5a). We also found a non-linear importance of SIMD (Figure 5b) and number of previous emergency hospital admissions (Supplementary Figure 5c).

We further investigated the contribution of SIMD by comparing Shapley values between features. We computed the mean difference in contribution of SIMD to risk score between individuals in the most deprived and least deprived SIMD decile areas, and the additional years of age which would contribute an equivalent amount. This was generally around 10-40 additional years (Figure 5d). In terms of raw admission rates, disparity was further

apparent: individuals aged 20 in lowest SIMD decile areas had similar admission rates to individuals aged 70 in the 3 highest SIMD decile areas (Figure 5e).

When exploring the added value (in terms of AUROC) of including the features derived using the topic model (Supplementary Table 4), we observed slightly better performance than the model without such features (p-value $= 3 \times 10^{-29}$; Supplementary Figure 5e-F). In some cases, topic features led to substantial changes in overall score: for example, a topic relating to skin disease contributed more than 2% to the SPARRA$v4$ score (roughly equivalent to the mean contribution to the score from age for individuals aged 75; see Figure 5a) for around 0.43% of individuals with the resultant SPARRA$v4$ scores better-calibrated than the SPARRA$v3$ scores, which did not use a topic model (Supplementary Note 1). Analogously to Figure 2e, we also computed the additional number of samples correctly identified as having an event amongst the top scores by the two models. Although the absolute difference in AUROC was small, we found that the use of topic features increased the number of EAs detected in the top 500,000 scores by around 200.

Deployment

SPARRA$v4$ was developed in a remote data safe haven (DSH) environment [28] without access to internet or modern collaboration tools (e.g. git version control). Whilst our analysis code and a summary of model outputs (e.g. AUROC values) could be securely extracted from the DSH, this was not possible for the actual trained model due to potential leaks of sensitive patient information [29]. This introduced reproducibility challenges, since the model had to be retrained in a different secure environment before it was deployed by PHS. In particular, this re-development outside the DSH had two distinct phases. Firstly, the raw data transformations (to convert the original databases into a format that is suitable for ML algorithms) were reproduced from scratch from the same source data. Once the output of the transformations matched perfectly between the DSH and the external environment for all features, the topic and predictive models were re-trained. The training process could not be exactly matched due to differing compute environments, package versions and training/validation split. However, after training, the external models were validated by comparing the performance (via AUROC) and the calibration with the results obtained within the DSH.

Another practical issue that arises when developing and deploying a new version of SPARRA is due to potential *performative prediction* effects [30]. Since SPARRA$v3$ is already visible to GPs (who may intervene to reduce the risk of high-risk patients), v3 can alter observed risk in training data used for v4, with v3 becoming a *'victim of its own success'* [31, 32]. This is potentially hazardous: if some risk factor R confers high v3 scores prompting GP intervention (e.g., enhanced follow-up), then in the training data for v4, R may no longer apparently confer increased risk. Should v4 replace v3, some individuals would therefore have their EA risk underestimated, potentially diverting important anticipatory care away from them. This highlights a critical problem in the theory of model updating [33], which we expand on in Methods and illustrate in Figure 6a-d. As a practical solution, during deployment, GPs could receive the maximum between v3 and v4 scores. This would avoid the potential hazard of risk underestimation, at the cost of mild loss of AUROC (Figure 6e) and score calibration (Figure 6f).

Discussion

We used routinely collected EHRs from around 5.8 million Scottish residents to develop and evaluate SPARRA$v4$, a risk score that quantifies 1-year EA risk based on age, deprivation (using SIMD as a geographic-based proxy) and a wide range of features derived from a patient's past medical history. SPARRA$v4$ constitutes a real-world use of ML, derived from population-level data and embedded in clinical settings across Scotland (Figure 1).

While the increases in AUROC and AUPRC over the previous version of SPARRA may be small (Figure 2a,b), the improvement provided by SPARRA$v4$ in terms of absolute benefit to population is substantial (Figure 2e,f). This arises from the use of more flexible ML methods (e.g. to capture non-linear patterns between features and EA risk) and the incorporation of features derived by a topic model which extracts more granular information (with respect to the manually curated features used by SPARRA$v3$) from past diagnoses and prescriptions data. The latter can be thought of as a proxy for multi-morbidity patterns, in that topic models identify patterns of diagnoses and prescriptions which commonly occur together [34], which can be seen to occur in our data (Supplementary Table 4). The use of an ensemble of models also allows stronger models and methods to dominate the final predictor, and weaker models to be discarded.

Our analysis also provides insights into the epidemiology of EA risk, highlighting predictable patterns in terms of EA type (as defined by the recorded primary diagnosis; Figure 3b) and the imminence of EA (Figure 3c), in that those at high risk of an admission are likely to have an imminent admission rather than equally likely to have an admission over the year-long prediction period. Moreover, we studied the contribution of each feature, revealing a complex relationship between age, deprivation and EA risk (Figure 5). Note, however, that we cannot assign a causal interpretation for any reported associations. In particular, the link between SIMD and EA risk is complex; SIMD includes a 'health' constituent [24], and individuals in more-deprived SIMD decile areas (1: most deprived; 10: least deprived) miss more primary care appointments [35].

One important strength of SPARRA$v4$ is its nationwide coverage, using existing healthcare databases without the need for additional bespoke data collection. This, however, prevents the use of primary care data (beyond community prescribing) as it is not currently centrally collected in Scotland. Due to privacy considerations, we were also unable to access geographic location data, precluding the study of potential differences between e.g. rural and urban areas and the use of a geographically separated test set [8]. Limited data availability also limits a straightforward comparison of predictive performance (e.g. in terms of AUROC) with respect to similar models developed in England [10, 6] (this is also complicated because of different model choices, e.g. [6] modelled time-to-event data but we used a binary 1-year EA indicator). For example, we do not have information about marital and smoking status, blood test results and family histories; all of which were found to be predictive of EA risk by [6]. Our training dataset is non-representative of our raw dataset (which in turn is non-representative of the Scottish population, as per Table 1, as is typical of studies based on electronic health records [36, 37]), but it does generally include individuals at higher EA risk.

Beyond model development and evaluation, our work also highlights broader challenges that arise in this type of translational project using EHR. In particular, as SPARRA$v4$ has the potential to influence patient care, we have placed high emphasis on transparency and reproducibility while ensuring compliance with data governance constraints. Providing our code in a publicly available repository will also allow us to transparently document

future changes to the model (e.g. if any unwanted behaviour is identified during the early stages of deployment). SPARRA$v4$ also constitutes a real-world example in which potential performative effects need to be taken into account when updating an already deployed risk prediction model (Figure 6).

It is critical to emphasise that SPARRA$v4$ will not replace clinical judgement, nor does it direct changes to patient management made solely based on the score. Indeed, any potential interventions must be decided jointly by medical professionals and patients, balancing the underlying risks and benefits. Moreover, lowering EA risk does not necessarily entail overall patient benefit as e.g. long-term oral corticosteroid use in mild asthmatics would reduce EA risk, but the corticosteroids themselves cause an unacceptable cost of long-term morbidity [38].

Optimal translation into clinical action is a vital research area and is essential for quantifying the benefit of such scores in clinical practice. Indeed, any benefit is dependent on widespread uptake and the existence of timely integrated health and social care interventions, and identification of EA risk is only the first step in this pathway. As such, the evaluation of real-world effectiveness for SPARRA$v4$ and similar risk scores is complex, and requires a multi-disciplinary approach that considers a variety of factors (e.g. the local health economy and the capacity to deliver pre-emtive interventions in primary care) Therefore, we will continue to collaborate to achieve successful deployment of SPARRA$v4$ and will carefully consider the feedback from GPs to improve the model and the communication of its results further (e.g. via informative dashboards). As the COVID-19 pandemic resolves, it will also be important to assess potential effects of dataset shift [39] due to disproportionate mortality burden in older individuals and long-term consequences of COVID-19 infections. In an era where healthcare systems are under high stress, we hope that the availability of robust and reproducible risk scores such as SPARRA$v4$ (and its future developments) will contribute to the design of proactive interventions that reduce pressures on healthcare systems and improve healthy life expectancy.

Methods

Ethics and data governance

The project was covered under National Safe Haven Generic Ethical Approval (favourable ethical opinion from the East of Scotland NHS Research Ethics Service). This study was conducted in accordance with UK data governance regulations and the use of patient-level EHR was approved by the Public Benefit and Privacy Panel (PBPP) for Health and Social Care (study number 1718-0370; approval evidenced in application outcome minutes for 2018/19 at `https://www.informationgovernance.scot.nhs.uk/pbpphsc/application-outcomes/`). Data access was also approved by the PHS National Safe Haven, through the electronic Data Research and Innovation Service (eDRIS).

All studies have been conducted in accordance with information governance standards; data had no patient identifiers available to the researchers. Due to the confidential nature of the data, all analysis took place on a remote "data safe haven", 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 on all data exports, including the outputs presented in this manuscript.

SPARRA$v3$

SPARRA$v3$ [12], deployed in 2012, uses separate logistic regressions on four subcohorts of individuals: frail elderly conditions (FEC; individuals aged > 75); long-term conditions (LTC; individuals aged 16-75 with prior healthcare system contact), young emergency department (YED; individuals aged 16-55 who have had at least one A&E attendance in the previous year) and under-16 (U16; individuals aged < 16). If an individual belongs to more than one of these groups, the maximum of the associated scores is reported. SPARRA$v3$ was fitted once (at its inception in 2012) with regression coefficients remaining fixed thereafter. Most input features were manually dichotomised into two or more ranges for fitting and prediction. The prediction target for SPARRA$v3$ is EA within 12 months. People who died in the pre-prediction period, and who therefore do not have an outcome for use in the analysis, are excluded. PHS calculated SPARRA$v3$ scores and provided them as input for the analysis described herein. Any GP in Scotland can access SPARRA scores after attaining information governance approval.

Exclusion criteria

The exclusion criteria were applied per sample (defined as individual-time pairs; Figure 1c). Samples were excluded if: (i) they were excluded from SPARRA$v3$ (these are individuals for which PHS did not calculate a SPARRA$v3$ score and largely correspond to individuals with no healthcare interactions or that were not covered by the four SPARRA$v3$ subcohorts; [12]), (ii) when the individual died prior to the prediction time cutoff, (iii) when the SIMD for the individual was unknown, or (iv) those associated to individuals whose Community Health Index [CHI; 40] changed during the study period ('Unmatched' in Figure 1). The CHI number is a unique identifier which is used in Scotland for health care purposes. Rates of EA and death in the follow-up period were generally lower in excluded samples than in included samples (3.40% versus 8.88%, only considering exclusions which were not due to the individual having died prior to the time cutoff; Supplementary Table 6). Exclusion criteria (i) and (ii) were applied at the sample level, while exclusion criteria (iii) and (iv) were applied at the individual level.

Feature engineering

A typical entry in the source EHR tables (Supplementary Table 2) recorded a single interaction between a patient and NHS Scotland (e.g. hospitalisation), comprising a unique individual identifier (an anonymised version of the CHI number), the date on which the interaction began (admission), the date it ended (discharge), and further details (diagnoses made, procedures performed). For each sample, entries from up to three years before the time cutoff were considered when building input features, except long-term condition (LTC) records, which considered all data since recording began in January 1981. A full feature list is described in Supplementary Table 3. This includes SPARRA$v3$ [12] features, e.g. age, sex, SIMD deciles and counts of previous admissions (e.g. A&E admissions, drug-and-alcohol-related admissions). Additional features encoding time-since-last-event (e.g. days since last outpatient attendance) were included following findings in [6]. From community prescribing data, we derived predictors encoding the number of prescriptions of various categories (e.g. respiratory), extending the set of predictors beyond a similar set used in SPARRA$v3$. Similarly to SPARRA$v3$, we also derived the total number of different prescription categories, the total number of filled prescription items, and the number of British National

Formulary (BNF) sections from which a prescription was filled [41]. From LTC records, we extracted the number of years since diagnosis of each LTC (e.g. asthma), the total number of LTCs recorded, and the number of LTCs resulting in hospital admissions.

Data from prescription records and recorded diagnoses tend to be sparse, in that most medications and diagnoses will only be recorded for a small proportion of the population. We used our topic model [20] to assimilate this data, by jointly modelling prescriptions and diagnoses using 30 topics (effectively clusters of prescriptions and diagnoses), considering samples as 'documents' and diagnoses/prescriptions as 'words'. This enabled a substantial reduction in feature dimensionality, given the number of diagnoses/prescription factor levels. Using the map from documents to topic probabilities, we used derived topic probabilities as additional features in SPARRA$v4$, which corresponded to sample-wise membership of each topic.

Choice of prediction target for SPARRA$v4$

The primary target for SPARRA is to predict whether an individual will experience an EA within 12 months from the prediction cutoff. A problem arises due to the deaths during the follow-up year for which the target may be unknown (e.g. if someone died within 6 months, without a prior EA). Broadly, there are four options for how to treat such individuals during model training and testing:

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

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.

Option 1 would exclude the most critically ill individuals from the dataset and hence was discarded. 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 3 would effectively classify death as a 'desirable' outcome, so we avoided it. The consequences from coding community deaths as non-EA would be 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. Instead, option 4 allows the general description of the target as 'a catastrophic breakdown in health'. In this case, our model would not be able to distinguish community deaths from emergency admissions: we may assign high 'EA' 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 [42] 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. As the philosophy of

the SPARRA score is to avert breakdowns in health, of which death can be considered an example, we decided to use a composite prediction target (EA or death within 12 months) which is consistent with option 4.

Machine learning prediction methods

For SPARRAv4, we had no prior belief that any ML model class would be best, so considered a range of binary prediction approaches (hereafter referred to as constituent models). The following models were fitted using the *h2o* [43] R package (version 3.24.0.2): an artificial neural network (ANN), two random forests (RF) (depth 20 and 40), an elastic net generalised linear model (GLM) and a naive Bayes (NB) classifier. The *xgboost* [44] R package (version 1.6.0.1) was used to train three gradient-boosted trees (XGB) models (maximum tree depth 3, 4, and 8). Hyper-parameter choices are described in Methods. SPARRAv3 was used as an extra constituent model.

Rather than selecting a single constituent model, we used an ensemble approach. Similar to [19], we calculated an optimal linear combination (L_1-penalised regression, using the R package `glmnet`, version 4.1.4) of the scores generated by each constituent model. Ensemble weights were chosen to optimise the AUROC. Finally, we monotonically transformed the derived predictor to improve calibration by inverting the empirical calibration function (Supplementary Note 2).

Data imputation

As all non-primary care interactions with NHS Scotland are recorded in the input databases, there was no missingness for most features. For 'time-since-interaction' type features, samples for which there was no recorded interaction were coded as twice the maximum lookback time. There was minor non-random missingness in topic features ($\sim$0.8%) due to individuals in the dataset with no diagnoses or filled prescriptions, for whom topic probabilities could not be calculated. 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 [43]) in RF, used sample-wise imputation in XGB (as per [44]) and dropped during fitting (default in [43]) in NB (omitted missing values for prediction). All imputation rules were determined using training sets only.

Particular care was required for features encoding total lengths of hospital stays. In some cases, a discharge date was not recorded, which could lead to an erroneous assumption of a very long hospital stay (from admission until the time cutoff). To address this, we truncated apparently spuriously long stays at data-informed values (Supplementary Note 4).

Hyperparameter choice for ML prediction methods

We used a range of constituent models. The h2o [43] R package (version 3.24.0.2) was used to train ANN, RF, GLM and NB models. The xgboost [44] R package (version 1.6.0.1) was used to train the XGB models. Unless otherwise specified, hyperparameters were set as the software defaults. When tuned, hyperparameter values were chosen to optimise the default objective functions implemented for each method: log-loss or the ANN, RFs and GLM, likelihood for the NB model; and a logistic objective for the XGB trees. 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. Details for each method are provided

below. Only limited hyperparameter tuning was possible due to the restricted computational environment in the data safe haven (see Results)

SPARRA$v3$

SPARRA$v3$ scores were calculated by PHS using their existing algorithm [12].

Artificial neural network (ANN)

We used a training dropout rate of 20% to reduce generalisation error. We optimised over the number of layers (1 or 2) and the number of nodes in each layer (128 or 256).

Random forest (RF)

We fitted two RF: 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).

Gradient-boosted trees (XGB)

We fitted three boosted tree models with three maximum depths: 3, 4, and 8. For the deeper-tree model, we set a low step size shrinkage $\eta = 0.075$ and a positive minimum loss reduction $\gamma = 5$ in order to avoid overfitting. In the other two models, we used default values of $\eta = 0.3$, $\gamma = 0$.

Naive Bayes (NB)

The only hyperparameter we tuned was a Laplace smoothing parameter, varying between 0 and 4.

Penalised Generalised linear model (GLM)

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 $\{0, 0.5, 1\}$.

Cross-validation

We fitted and evaluated SPARRA$v4$ using three-fold cross-validation (CV). We considered three-fold cross validation acceptable in our case given the size of our dataset [45]. This was designed such that all elements of the model evaluated on a test set were agnostic to samples in that test set. Individuals were randomly partitioned into three data folds (F1, F2 and F3). At each CV iteration, F1 and F2 were combined and used as a training dataset, F3 was used as a test dataset. The training dataset (F1+F2) was used to fit the topic model and to train all constituent models (except SPARRA$v3$, whose training anyhow pre-dates the data used here). The ensemble weights and re-calibration transformation were learned using F1 + F2, i.e. without using the test set from the test set (Supplementary Note 2).

Predictive performance

Our primary endpoint for model performance was AUROC. We also considered area-under-precision-recall curves (PRC) and calibration curves. We plotted calibration curves using a kernelised calibration estimator (Supplementary Note 5).

For simplicity, figures show ROC/PRC that were calculated by combining all samples from the three *test* CV folds (that is, all scores and observed outcomes were merged to draw a single curve). Quoted AUROC/AUPRC values were calculated as an average across the three *test* CV folds to avert problems from between-fold differences in models [46]. For ease of comparison, we also used mean-over-folds to compute quoted AUROCs and AUPRCs for SPARRA$v3$, although the latter was not fitted to our data.

Deployment scenario stability and performance attenuation

Using the same analysis pipeline as for the development of SPARRA$v4$, we trained a static model M_0 to an early time cutoff (t_0=1 May 2014), and using one year of data prior to t_0 to derive predictors (the restricted lookback is the only deviation from the actual model pipeline, due to limited temporal span of the training data).

We studied the performance of M_0 as a *static model* to repeatedly predict risk at future time cutoffs, which mirrors the way in which PHS will deploy the model. To do this, we assembled test features from data 1 year prior to 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, applying M_0 to predict EA risk in the year following each time-point. In this analysis, the comparison of the distribution of scores over time only considered the cohort of patients who were alive and had valid scores at $t_1, \ldots, t_5$.

To ensure a fair comparison when evaluating the performance of *static scores* (computed at t_0 using M_0) to predict future event risk (at $t_1, \ldots, t_5$), we only considered a subsample of 1 million individuals with full data across all time-points, selected such that global admission rates matched those at t_0.

Assessment of feature importance

We examined the contribution of feature to risk scores at an individual level by estimating Shapley values [21] for each feature. For simplicity, this calculation was done using 20,000 randomly-chosen samples in the first cross-validation fold (F1). We treated SPARRA$v3$ scores as fixed predictors rather than as functions of other predictors.

We also 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 the model to F2+F3 with topic-derived features excluded from the predictor matrix. We compared the performance of these models using F1 as test data. We compared the performance of predictive models with and without the features derived from the topic model by comparing AUROC values using DeLong's test [47].

Model	Fold 1		
	AUROC	AUPRC	Coef.
ANN	0.7613	0.346	0
Penalised GLM	0.7879	0.3657	0
Naive Bayes	0.7471	0.2233	0
RF, depth: 20	0.7927	0.3787	0.3624
RF, depth: 40	0.7845	0.3666	0
SPARRA$v3$	0.7812	0.3568	0
XGB depth: 4	0.7981	0.3839	0.6626
XGB depth: 8	0.7984	0.3873	2.004
XGB depth 3	0.7984	0.3864	1.363
Ensemble	0.7989	0.3888	

Model	Fold 2		
	AUROC	AUPRC	Coef.
ANN	0.7698	0.3479	0
Penalised GLM	0.7874	0.367	0
Naive Bayes	0.7468	0.2238	0
RF, depth: 20	0.7928	0.3799	0.3749
RF, depth: 40	0.7844	0.3678	0
SPARRA$v3$	0.7809	0.3584	0
XGB depth: 4	0.7975	0.3839	0.6579
XGB depth: 8	0.798	0.3881	1.162
XGB depth 3	0.7981	0.387	1.727
Ensemble	0.7987	0.3895	

Model	Fold 3			Mean over folds	
	AUROC	AUPRC	Coef.	AUROC	AUPRC
ANN	0.7693	0.3525	0	0.7668	0.3488
Penalised GLM	0.7878	0.3661	0	0.7877	0.3663
Naive Bayes	0.7468	0.2246	0	0.7469	0.2239
RF, depth: 20	0.7926	0.3791	0.5013	0.7927	0.3792
RF, depth: 40	0.784	0.3674	0	0.7843	0.3672
SPARRA$v3$	0.7809	0.3572	0	0.7810	0.3574
XGB depth: 4	0.7973	0.3837	0.9105	0.7976	0.3838
XGB depth: 8	0.7978	0.3877	1.116	0.7981	0.3877
XGB depth 3	0.798	0.3867	1.418	0.7982	0.3867
Ensemble	0.7985	0.3891		0.7987	0.3891

Table 2: **Overall discrimination performance for SPARRA$v4$ and its constituent models.** Areas under ROC curves and PR curves by fold for each constituent predictor and ensemble. Columns 'Coef.' indicate estimated coefficients (weights) in the final ensemble (see Methods section for details). All standard errors for AUROCs are $< 5 \times 10^{-4}$ and for AUPRCs are $< 8 \times 10^{-4}$

Variable	Importance
Age at time cutoff	1.530
Days since last emergency admission	0.752
Number of previous A&E attendances	0.509
Number of antibacterial prescriptions	0.376
Number of central nervous system related prescriptions	0.375
Male sex	0.373
Days since last A&E attendance	0.321
SIMD decile	0.310
Number of emergency bed days	0.299
Days since last acute admission of any type	0.285
Days since last outpatient attendance	0.257
Number of diuretic prescriptions	0.213
Number of lipid lowering drug prescriptions	0.194
Number of previous first outpatient appointments	0.190
Number of recorded long term conditions	0.173
Number of emergency admissions	0.161
Total number of filled prescriptions	0.160
Number of antianaemic prescriptions	0.159
Number of bronchodilator prescriptions	0.152
Number of BNF sections from which a prescription was filled	0.141

Table 3: **Top 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.

Model updating in the presence of performative effects

We aim to produce the SPARRA score to accurately estimate EA risk over a year under normal medical care. In other words, the score should represent the EA risk if GPs do not already have access to such a risk score. Because GPs see a SPARRA score (SPARRA$v3$) and may act on it, the observed risk may be lower than predicted - the score may become a 'victim of its own success' [31, 32] due to performative effects [48]. Unfortunately, since the SPARRA$v3$ score is widely available to Scottish GPs, and may be freely acted on, we cannot assess the behaviour of the medical system in its absence. This is potentially hazardous [33].

Formally, at a given fixed time, for each individual, the value of 'EA in the next 12 months' is a Bernoulli random variable. The probability of the event for individual i is 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 SPARRA$v4$ scores as functions of covariates, and assume a causal structure shown in Figure 6 (for simplicity, we assume there are no unobserved confounders but the same argument applies in their presence). With no SPARRA-like predictive score in place, there is only one causal pathway $X_i \rightarrow EA$. It is to this system (coloured red) that $v3$ was fitted. Here, $v3(X_i)$ estimates the 'native' risk $Pr(EA|X_i)$ (ignoring previous versions of the SPARRA score, which covered $< 30\%$ of the population). Although $v3(X_i)$ is determined entirely by X_i, the act of distributing values of $v3(X_i)$ to GPs opens a second causal pathway from X_i to EA (Figure 6) 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 SPARRA$v4$ naively replaced SPARRA$v3$ (Figure 6), we would be using $v4(X_i)$ to predict behaviour of a system different to that on which it was trained (Figure 6). To amend this problem, we propose to use SPARRA$v4$ in *conjunction* with SPARRA$v3$ rather than to completely replace it (Figure 6). 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 6. This is impractical, so instead, we propose to distribute a single value (given by the maximum between $v3(X_i)$ and $v4(X_i)$), avoiding the potential hazard of risk underestimation, at the cost of mild loss of score calibration (Figure 6).

Data availability

Raw data for this project are patient-level EHR, which have been anonymised for confidentiality ahead of any analysis being undertaken. Enquiries about access to this data may be directed to `phs.edris@phs.scot`. However, the summary data required to draw figures included in our manuscript is publically available from our GitHub repository. All publicly available data summaries were reviewed by an independent team to avoid the risk of accidental disclosure of sensitive information.

Code availability

All analysis code and co-ordinates required to reproduce our Figures are available in `github.com/jamesliley/SPARRA`$v4$

This manuscript conforms to the TRIPOD guidelines [23] (Supplementary Table 1).

Acknowledgements

The authors note that this project's success was entirely contingent on close co-operation between the Alan Turing Institute and PHS. We thank individuals involved in primary care in Scotland for the continued support of the SPARRA project and the Public Benefit and Privacy Panel for Health and Social Care (study number 1718-0370) for Information Governance approval on behalf of the Health Boards in NHS Scotland.

Computing for this project was performed in the Scottish National Safe Haven (NSH), which is commissioned by eDRIS, Public Health Scotland from EPCC, based at The University of Edinburgh. The authors would like to acknowledge the support of the eDRIS Team for their involvement in obtaining approvals, provisioning and linking data and the use of the secure analytical platform within the National Safe Haven. This work uses data provided by patients and collected by the NHS as part of their care and support

We thank the Alan Turing Institute, PHS, 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, IT, 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, IT, 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.

Author contributions

All author contributions were significant and essential to the completion of this work. Author contributions were as follows:

Manuscript preparation	:	JL, SRE, BAM, SJV, CAV, LJMA, IT;
Project initiation	:	SJV, CAV, LJMA, CH;
Model design	:	JL, GB, SJV, CAV, LJMA;
Code and scripts	:	JL, GB, LJMA, NC, IT, SDR;
Code review and checking	:	SRE, IT, SDR;
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.

Authors JL and GB contributed equally to this work. Authors CAV and LJMA had an equal role in supervising this work.

Competing interests

The authors declare no competing interests.

References

[1] Public Health Scotland. Acute hospital activity and NHS beds information for Scotland, 2022.

[2] Rural Access Action Team. The national framework for service change in NHS Scotland. *Scottish Executive, Edinburgh*, 2005.

[3] Marian S McDonagh, David H Smith, and Maria Goddard. Measuring appropriate use of acute beds: a systematic review of methods and results. *Health policy*, 53(3):157–184, 2000.

[4] Colin Sanderson and Jennifer Dixon. Conditions for which onset or hospital admission is potentially preventable by timely and effective ambulatory care. *Journal of health services research & policy*, 5(4):222–230, 2000.

[5] Joanna Coast, Abby Inglis, and Stephen Frankel. Alternatives to hospital care: what are they and who should decide? *BMJ*, 312(7024):162–166, 1996.

[6] Fatemeh Rahimian, Gholamreza Salimi-Khorshidi, Amir H Payberah, Jenny Tran, Roberto Ayala Solares, Francesca Raimondi, Milad Nazarzadeh, Dexter Canoy, and Kazem Rahimi. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS medicine*, 15(11):e1002695, 2018.

[7] David Lyon, Gillian A Lancaster, Steve Taylor, Chris Dowrick, and Hannah Chellaswamy. Predicting the likelihood of emergency admission to hospital of older people: development and validation of the emergency admission risk likelihood index (EARLI). *Family practice*, 24(2):158–167, 2007.

[8] Emma Wallace, Ellen Stuart, Niall Vaughan, Kathleen Bennett, Tom Fahey, and Susan M Smith. Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. *Medical care*, 52(8):751, 2014.

[9] Alex Bottle, Paul Aylin, and Azeem Majeed. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *Journal of the Royal Society of Medicine*, 99(8):406–414, 2006.

[10] John Billings, Jennifer Dixon, Tod Mijanovich, and David Wennberg. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ*, 333(7563):327, 2006.

[11] Julia Hippisley-Cox and Carol Coupland. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ open*, 3(8):e003482, 2013.

[12] 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), 2011. `https://www.isdscotland.org/Health-Topics/Health-and-Social-Community-Care/SPARRA/2012-02-09-SPARRA-Version-3.pdf`, Accessed: 6-3-2020.

[13] Attakrit Leckcivilize, Paul McNamee, Christopher Cooper, and Robby Steel. Impact of an anticipatory care planning intervention on unscheduled acute hospital care using difference-in-difference analysis. *BMJ health & care informatics*, 28(1), 2021.

[14] Gill Highet, Debbie Crawford, Scott A Murray, and Kirsty Boyd. Development and evaluation of the supportive and palliative care indicators tool (SPICT): a mixed-methods study. *BMJ supportive & palliative care*, 4(3):285–290, 2014.

[15] N Bajaj, S Jauhar, and J Taylor. Scottish patients at risk of readmission and admission-mental health (SPARRA MH) case study of users and non-users of a national information source. *Health Syst Policy Res*, 3:3, 2016.

[16] Anne Canny, Frances Robertson, Peter Knight, Adam Redpath, and Miles D Witham. An evaluation of the psychometric properties of the indicator of relative need (IoRN) instrument. *BMC geriatrics*, 16(1):1–10, 2016.

[17] S Manoukian, S Stewart, N Graves, H Mason, C Robertson, S Kennedy, J Pan, L Haahr, SJ Dancer, B Cook, et al. Evaluating the post-discharge cost of healthcare-associated infection in NHS Scotland. *Journal of Hospital Infection*, 114:51–58, 2021.

[18] Emma Wallace, Susan M Smith, Tom Fahey, and Martin Roland. Reducing emergency admissions through community based interventions. *BMJ*, 352, 2016.

[19] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

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

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

[22] Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586):eabb1655, 2021.

[23] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Journal of British Surgery*, 102(3):148–158, 2015.

[24] Scottish Government. Scottish index of multiple deprivation, 2016.

[25] Office for National Statistics, National Records of Scotland, and Northern Ireland Statistics and Research Agency. 2011 census aggregate data. UK data service (edition: June 2011), 2011.

[26] Ian Blunt. Focus on preventable admissions. *London: Nuffield Trust*, 2013.

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

[28] Public Health Scotland. eDRIS Products and Services, Public Health Scotland, 2020.

[29] Emily Jefferson, James Liley, Maeve Malone, Smarti Reel, Alba Crespi-Boixader, Xaroula Kerasidou, Francesco Tava, Andrew McCarthy, Richard Preen, Alberto Blanco-Justicia, et al. GRAIMATTER green paper: Recommendations for disclosure control of trained machine learning (ML) models from trusted research environments (TREs). *arXiv preprint arXiv:2211.01656*, 2022.

[30] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[31] Matthew C Lenert, Michael E Matheny, and Colin G Walsh. Prognostic models will be victims of their own success, unless.... *Journal of the American Medical Informatics Association*, 26(12):1645–1650, 2019.

[32] Matthew Sperrin, David Jenkins, Glen P Martin, and Niels Peek. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *Journal of the American Medical Informatics Association*, 26(12):1675–1676, 2019.

[33] James Liley, Samuel R Emerson, Bilal A Mateen, Catalina A Vallejos, Louis JM Aslett, and Sebastian J Vollmer. Model updating after interventions paradoxically introduces bias. *AISTATS proceedings*, 2021.

[34] Ron Kremer, Syed Mohib Raza, Fabiola Eto, John Casement, Christian Atallah, Sarah Finer, Dennis Lendrem, Michael Barnes, Nick J Reynolds, and Paolo Missier. Tracking trajectories of multiple long-term conditions using dynamic patient-cluster associations. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4390–4399. IEEE, 2022.

[35] David A Ellis, Ross McQueenie, Alex McConnachie, Philip Wilson, and Andrea E Williamson. Demographic and practice factors predicting repeated non-attendance in primary care: a national retrospective cohort analysis. *The Lancet Public Health*, 2(12):e551–e559, 2017.

[36] Robert A Verheij, Vasa Curcin, Brendan C Delaney, and Mark M McGilchrist. Possible sources of bias in primary care electronic health record data use and reuse. *Journal of medical Internet research*, 20(5):e185, 2018.

[37] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

[38] NICE guidelines. Asthma: diagnosis, monitoring and chronic asthma management. *National Institute of Health and Care Excellence*, November 2017.

[39] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

[40] ISD Scotland Data Dictionary. CHI - Community Health Index, 2023. `https://www.ndc.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?ID=128`, Accessed: 17-3-2023.

[41] Anne B Prasad. British National Formulary. *Psychiatric Bulletin*, 18(5):304–304, 1994.

[42] Rafael D Romo, Theresa A Allison, Alexander K Smith, and Margaret I Wallhagen. Sense of control in end-of-life decision-making. *Journal of the American Geriatrics Society*, 65(3):e70–e75, 2017.

[43] Erin LeDell, Navdeep Gill, Spencer Aiello, Anqi Fu, Arno Candel, Cliff Click, Tom Kraljevic, Tomas Nykodym, Patrick Aboyoun, Michal Kurka, and Michal Malohlava. *h2o: R Interface for 'H2O'*, 2019. R package version 3.26.0.2.

[44] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. *xgboost: Extreme Gradient Boosting*, 2019. R package version 0.90.0.2.

[45] Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, pages 1–12, 2023.

[46] George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *Acm Sigkdd Explorations Newsletter*, 12(1):49–57, 2010.

[47] 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. *Biometrics*, pages 837–845, 1988.

[48] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

Figure Legends

Figure 1: **Data and model fitting overview.** (A) Illustration of how SPARRA can support primary care intervention with the goal of improving patient outcomes. (B) Distribution of the number of input EHR entries (prior to exclusions) according to age, sex and SIMD deciles (1: most deprived; 10: least deprived). (C) Flow chart summarising data and model fitting pipelines.

Figure 2: **Comparison of overall predictive performance between SPARRA$v3$ and SPARRA$v4$.** (A) ROC. (B) PRC. Lower sub-panels show differences in sensitivity and precision, respectively. Confidence intervals are negligible. (C) Calibration curves. (D) Calibration curves for samples in which $|v4 - v3| > 0.1$. Lower sub-panels show the difference between curves and the $y = x$ line (perfect calibration). Confidence envelopes are pointwise (that is, for each x-value, not the whole curve). Predicted/true value pairs are combined across cross-validation folds in all panels for simplicity. (E) Difference in the number of individuals who had an event amongst individuals designated highest-risk by v3 and v4. The repeating pattern is a rounding effect of v3. (F) Difference in the number of highest-risk individuals to target to avoid a given number of admissions.

Figure 3: **Stratified performance of SPARRA$v3$ and SPARRA$v4$.** (A) Performance of SPARRA$v3$ and SPARRA$v4$ in subcohorts defined by age, SIMD and the original subcohorts defined during SPARRA$v3$ development (Methods). Top: AUROC (blue: SPARRA$v3$; red: SPARRA$v4$). Vertical bars denote plus/minus 3 standard deviations. Middle: AUROC increase for SPARRA$v4$ with respect to SPARRA$v3$. For context, bottom sub-panels show the proportion of samples with an event within each group. (B) Distribution of SPARRA$v4$ scores (in log-scale) based on the type of diagnosis recorded during the admission (see Supplementary Table 5 for definitions). Black points indicate the associated medians. Groups were defined according to whether an event was observed (grey violin plots) or, for those with an EA, based on the diagnosis recorded during the admission (black violin plots). (C) Density of time-to-first-EA (that is, days between time cutoff and first EA date) in subsets of individuals who had an EA in the year following the time cutoff and had a SPARRA$v4$ score above a given cutoff. For instance, the lightest line shows density of time-to-first-EA in samples who had an EA and had SPARRA$v4$ >0.8

Figure 4: **Performance of a static model with changing scores over time.** (A-C) Performance of scores calculated at $t_1 - t_5$ from static model M_0. (A) ROC curves. Lower panel shows differences in sensitivity with respect to t_1. (B) PRC curves. Lower panel shows differences in precision with respect to t_1. (C) Calibration curves. Lower panel shows the difference between observed and expected EA frequency. (D) Centiles (grey) and deciles (black) of risk scores (calculated using M_0) over time, across all individuals with data available at all time cutoffs. (E) Average score over time for groups of individuals defined by risk centiles (grey) and deciles (black) at time t_0 (2 May 2015). (F) Density (low to high: white-grey-red-yellow) of scores generated using the static model M_0 to predict EA risk 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.

Figure 5: **Analysis of Shapley values.** Distribution of Shapley values by (A) age and (B) SIMD deciles (1: most deprived; 10: least deprived). (C) Number of additional years of age needed to match the difference in Shapley values between SIMD deciles 1 and 10. (D) 'Effective ages' calculated to match EA rates: for an (age, SIMD decile) pair, the age at mean SIMD with the equivalent EA rate.

Figure 6: **Model updating in the presence of performative effects.** (A-D) Causal structure for the training and deployment of SPARRA$v3$ and SPARRA$v4$. X_i represents covariates for a patient-time pair; $v3(fit)/v4(fit)$ and $v3(X_i)/v4(X_i)$ represent the fitting and deployment of v3 and v4 respectively. (A) Training setting for SPARRA$v3$. (B) Training setting for SPARRA$v4$. (C) Deployment setting if SPARRA$v4$ were to naively replace SPARRA$v3$. (D) Deployment setting in which SPARRA$v4$ is used as an adjuvant to SPARRA$v3$. (E) Comparison of discrimination (ROC) between SPARRA$v4$ and the maximum of both scores. (F) Comparison of calibration between SPARRA$v4$ and the maximum of both scores.