Machine learning model to predict mental health crisis from electronic health records

Roger Garriga (roger.garrigacalleja@koahealth.com)
Koa Health
Aleksandar Matić
Koa Health
Javier Mas
Koa Health
Semhar Abraha
University of Warwick
Jon Nolan
Birmingham and Solihull mental Health Trust
Oliver Harrison
Koa Health
George Tadros
Birmingham and Solihull mental Health Trust

Article

Keywords: Operating Characteristic Curve, Precision-recall Curve, Sensitivity, Specificity, Continuous Prediction, Clinical Settings

Posted Date: March 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-275866/v1

Version of Record: A version of this preprint was published at Nature Medicine on May 16th, 2022. See the published version at https://doi.org/10.1038/s41591-022-01811-5.

Machine learning model to predict mental health crisis from electronic health records

DBPR

Abstract

Timely identification of patients who are at risk of mental health crises opens the door for improving the outcomes and for mitigating the burden and costs to the healthcare systems. Due to high prevalence of mental health problems, a manual review of complex patient records to make proactive care decisions is an unsustainable endeavour. We developed a machine learning model that uses Electronic Health Records to continuously identify patients at risk to experience a mental health crisis within the next 28 days. The model achieves an area under the receiver operating characteristic curve of 0.797 and an area under the precision-recall curve of 0.159, predicting crises with a sensitivity of 58% at a specificity of 85%. The usefulness of our model was tested in clinical practice in a 6-month prospective study, where the predictions were considered clinically useful in 64% of cases. This study is the first one to continuously predict the risk of a wide range of mental health crises and to evaluate the usefulness of such predictions in clinical settings.

Introduction

Nearly one billion people worldwide are living with a mental disorder – causing one lost life through suicide every 40 seconds[1]. With the global mental health emergency, further accelerated by the COVID-19 pandemic, healthcare systems are experiencing significantly increased demand paralleled with a shortage of skilled personnel[2–4]. The demand for mental health services is often triggered by mental health crises – defined as situations in which patients are no longer able to care for themselves and function effectively in the community and when they may hurt themselves or others[5]. It has long been established that timely treatment can prevent the exacerbation of symptoms that leads to a crisis onset and hospitalisation, i.e., by intervening before the crisis has occurred to prevent it or to mitigate its impact[6]. However, too often, patients access the urgent care pathways as their primary entry point to a hospital or psychiatric facility only when already in crisis. This is too late to apply preventative strategies and limits the ability of psychiatry services to plan their limited resources ahead of time. Therefore, identifying patients at risk of experiencing a crisis prior to its occurrence is of a paramount importance to improve patient outcomes and to mitigate the existing pressure on the healthcare systems[7].

In busy clinical settings, the manual review of large quantities of data across many patients to make proactive care decisions is impractical, unsustainable and error-prone[8]. Shifting such tasks to the automated analysis of electronic health records (EHRs) holds an important promise to revolutionise health services by enabling such continuous data review at scale. Research has demonstrated the feasibility of predicting critical events in a wide range of healthcare problems such as hypertension, diabetes, circulatory failure, hospital readmission and in-hospital death[9–13]. However, when it comes to mental health, the literature is still fragmented and it typically focuses on the prediction of very specific events including risk of suicide, self-harm or a propensity to experience the first psychosis crisis episode[14–24] rather than continuously predicting any kind of a mental health crisis that will require urgent care or hospitalisation at any point in time. In this regard, little is yet known about the feasibility of querying machine learning models continuously to estimate an imminent risk of mental health crisis, which would be a key enabler for improving triage processes, for optimising healthcare staff allocation and for preventing crisis onsets. Furthermore, even a highly accurate predictive model does not guarantee the improvement of mental health outcomes and saving long-term costs[25,26] – whether predictive technologies would provide a useful tool to the practitioners in the mental healthcare practice remain unanswered to date[27,28].

In this research, we explore the feasibility of predicting a mental health crisis regardless of its cause or an underlying mental disorder, and we investigate if such prediction provides an added value in clinical practice. The fundamental question is whether there are patterns in historical events that are predictive of future mental health crises and if such patterns are resembled in the real-world EHRs that suffer sparseness, noise, errors and systematic biases[29]. We developed the mental crisis risk model by applying a machine learning algorithm on EHRs collected over a 7 year period (from 2012 until 2018) for 17,122 patients. We evaluated the accuracy of the model in continuously predicting the risk of a mental health crisis within the next 28 days from any point in time, which is important for supporting dynamic care decisions. We analysed how the performance vary across a range of mental health disorders and with respect to the sparseness of data. Next, we validated the clinical usefulness of this model by implementing it in a prospective trial with 60 clinicians who were attending 1,011 patients over 6 months (from 26th

November 2018 until 12th May 2019). We juxtaposed the risk evaluation performed by clinicians and by the machine learning model, and we conducted a qualitative study to delve deeper into the usefulness of the prediction model in clinical settings.

Results

Prediction target

Our dataset included *crisis events* registered at an hourly granularity every time a patient had an urgent need of mental health services. These events frequently occur in bursts (one after the other) when the patient is undergoing a crisis. Hence, predicting each single *crisis event* registered in the EHR would be of little clinical relevance – notably, patients who experience one *crisis event* receive close clinical attention during successive days. Therefore, we defined the prediction target as the onset of a *crisis episode* which contains one or more *crisis events* preceded by at least one full stable week without any *crisis event* (Figure 1). We trained the machine learning model to predict the onset of a *crisis episode* – i.e. the first *crisis event* in an *episode* – within the upcoming 28 days given that the patient has been stable during the preceding week. This time window was selected according to clinical practice in our clinical setting with the goal to conduct a timely intervention and prevent the next *crisis episode*. Importantly, our definition of the crisis onset did not considerably reduce the generalisability, namely using different time periods reasonably selected for the prediction interval (other than 28 days) or for defining a stable period before a relapse (other than 7 days) did not have a major impact on the model performance (Supplementary Table 8).

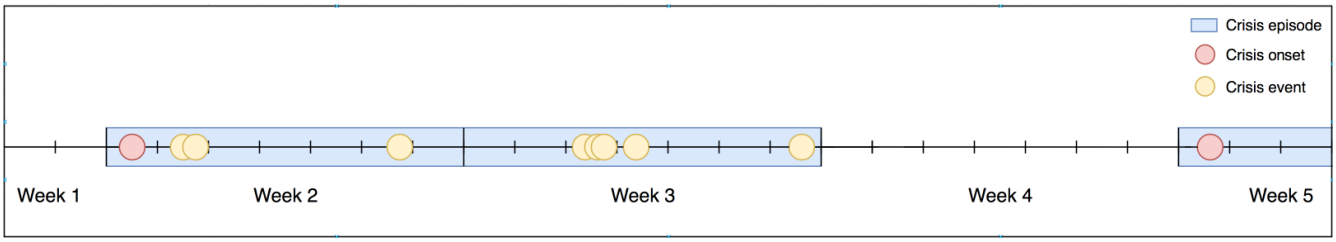

Figure 1. Timeline showing an example of crisis episode, where the crisis onset is the first crisis event of the crisis episode after a stable week without crisis events.

Dataset

Upon applying exclusion criteria (see Methods) the dataset contained 5,477,705 records collected between September 2012 and December 2018 from 17,122 unique patients within the age range between 16 and 102 years. The cohort included patients with a wide range of diagnosed disorders, including mood, psychotic, organic, substance abuse, neurotic and personality disorders. Demographic and other patient characteristics are summarized in Supplementary Table 1. The median number of records per patient was 115 (first and third quartile [40, 346]). In total, 1,448,542 records were crisis events and 942,017 of those events corresponded to hospitalisation. The data also contained timestamps of phone and in-person contacts with patients (2,239,632 of records), referrals (250,864 records), well-being and risks assessments (118,255 and 248,629 records respectively)– see Supplementary Table 2 for a detailed records break-down. Overall, 60,388 crisis episodes were included in the analysis, with a mean of 24 crisis events per episode. Our prediction target variable had a prevalence of 4.0% at average across the entire dataset, varying from 1.93% (Organic disorders) to 7.23% (Disorders of adult personality and behavior); Supplementary Table 5 presents the breakdown by diagnosis and train/test sets.

Development of a mental health crisis prediction model

The model was designed to be queried on a weekly basis and to infer the risk of experiencing a crisis onset for each patient for the following period of 4 weeks, i.e., 28 days. To build the model, we extracted three categories of features, namely: (i) static or semi-static patient information (such as age, gender, ICD-10 coded diagnoses), (ii) variables that encoded latest available assessments and interactions with the hospital (such as last risk assessments or well-being indicators, severity and number of crisis events in the last episode and similar) and (iii) variables that quantified the time elapsed since the registered events have occurred (such as crisis episode, contact, or a referral). In total, we extracted 198 features (see Supplementary Table 3 for a complete list of features) originating from 10 data tables with 74 routinely captured variables (see Supplementary Table 2). We queried the model on a week-by-week basis to predict the outcome in the upcoming 28 days (that is – the occurrence of a crisis onset or a stable period) and the performance metrics were averaged both in the validation and in test sets. When the system was implemented in practice, instead of a binary outcome the model generated a Predicted Risk Score (PRS) between 0 and 1 for each patient. Figure 2 shows the end-to-end process, from the data processing to delivering a dashboard in clinical settings during our prospective trial.

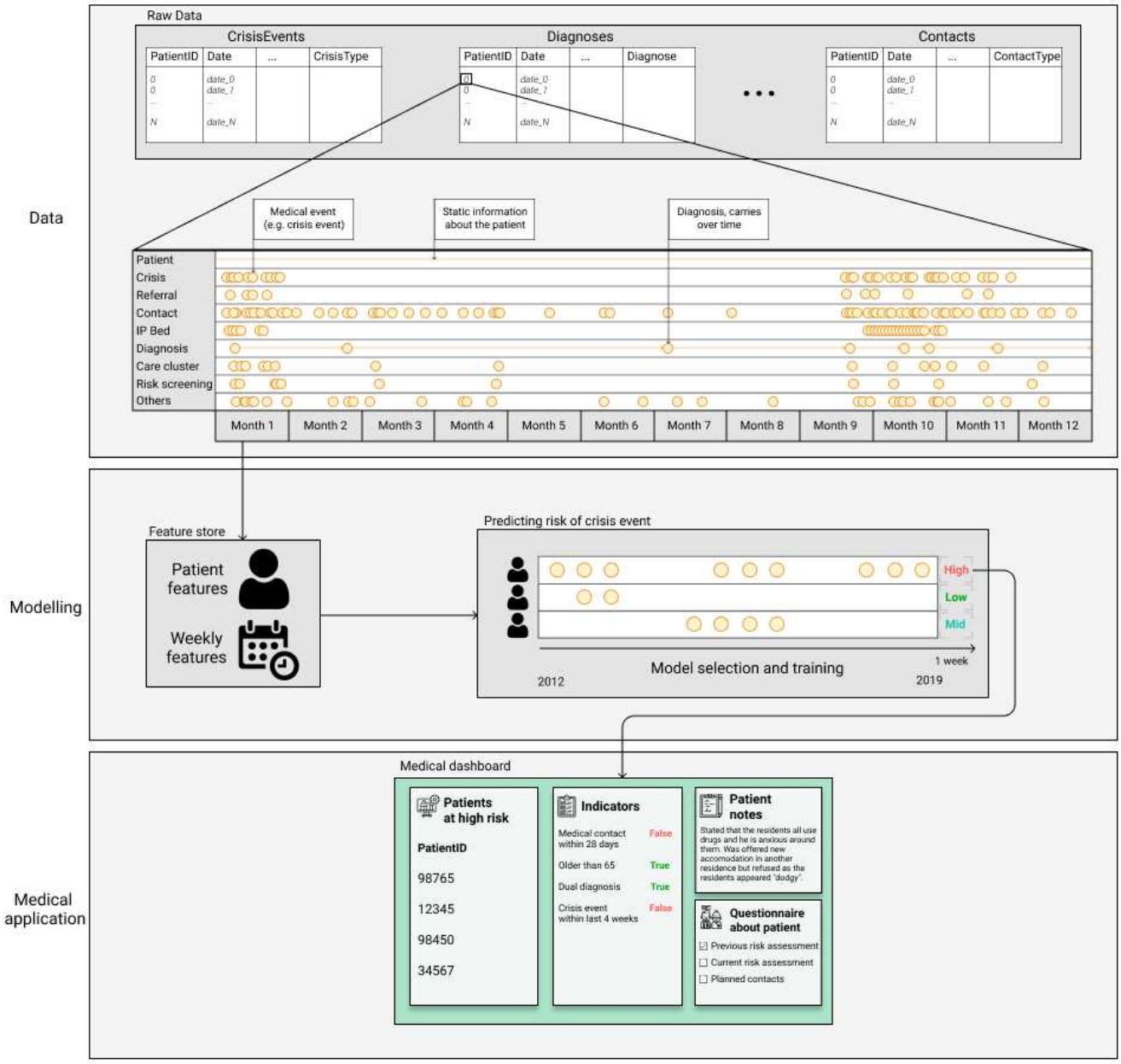

Figure 2. System diagram. Events over time are represented with their timestamp and characteristics in different SQL tables in the hospital's database. Those tables are processed into patient and weekly features for the modelling task. Models are trained, tuned and selected on data ranging from 2012 to 2019. The system then predicts every week - for every patient - the risk of crisis onset in the following four weeks. The patients with highest expected risk are displayed in the medical dashboard available to clinicians, along with key indicators, patient notes and a questionnaire to be filled by the clinician regarding the patient.

Consistent with the previous literature[9,12,30], XGBoost (eXtreme Gradient Boosting)[31] outpowered a range of machine learning techniques that we tested, including other decision trees and also geometrical, probabilistic, ensembles and deep learning based classifiers (Supplementary Table 6). The XGBoost model relied on an automatically selected subset of 104 features to predict mental health crises for all patients in our dataset regardless of a disorder (referred to as the general model). We benchmarked this model against two baseline classifiers, including: (i) clinical-practice based baseline model that was developed to impersonate doctor's decisions (specifically, a decision tree using a selection of indicators of the patient status that doctors in our clinical setting use to assess the risk of relapse), (ii) the diagnosis-based baseline model developed as a

logistical regression that was solely relying on diagnosis and time elapsed since the last crisis, resembling a threshold-based rule system (Supplementary Table 4 summarizes the list of features in each baseline). The Area Under the Receiver Operator Curves (AUROCs) of the general model, the clinical-practice-based baseline and the diagnosis-based baseline were 0.797 (95% CI 0.793-0.802), 0.736 (95% CI 0.733-0.740), 0.746 (95% CI 0.741-0.750) respectively (Figure 3). For predicting crisis episodes that occur infrequently compared to instances of stable periods (28 days without a crisis), the average precision (AP)[32] is a more informative metric and one of the most common metrics for unbalanced datasets[33]. The APs for the general model, the clinical-practice-based baseline and the diagnosis-based baseline were 0.159 (95% CI 0.154-0.165), 0.092 (95% CI 0.090-0.094), 0.092 (95% CI 0.089-0.094) respectively. Our general model significantly outperformed the two baseline models ($p<0.01$ for both AUROC and AP). We calibrated the predictions using isotonic regression[34] (Extended Figure 1), ensuring that the predicted risk reflects the actual expected risk of an individual to experience a relapse[35], and obtained a Brier score[36] of 0.028 (95% CI 0.028-0.029). The utility of the general model was also evaluated with the net benefit and the decision curve[37] was above the baseline models and default strategies (Extended Figure 1).

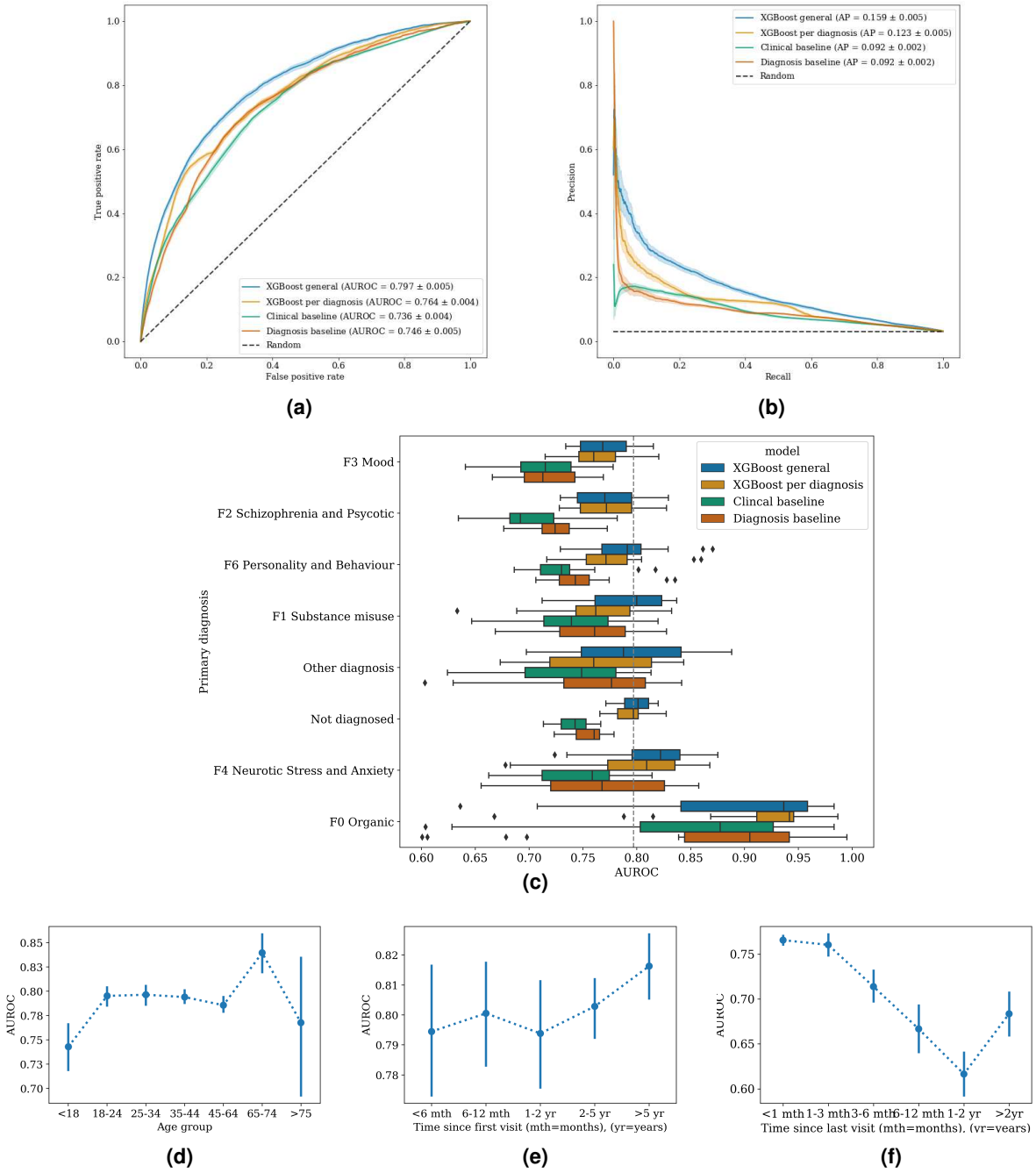

Figure 3. **a** Receiver operating characteristic curve for the crisis prediction task. Comparison between the proposed final model (XGBoost general), a proposed diagnosis specific model (XGBoost per diagnosis) and two baseline models. The solid lines and lighter-colored envelopes around each line were derived from the test evaluations ($n = 25$) as the mean and 95% confidence interval respectively. **b** Precision recall curve for the crisis prediction task with the same characteristics as **a**. **c** Box-plot of the area under the receiver operating curve evaluated per diagnosis. Comparison between the four models considered, same as in **a, b**. **d**, **e**, **f** Area under the receiver operator curve evaluated on different subsets of the cohort based on age group **(d)**, time since the patient had its first visit on the hospital **(e)** and time since last crisis episode **(f)**.

Prediction accuracy for different disorders, age groups and data availability

We evaluated the performance of the prediction model in subgroups of patients according to their latest diagnosed disorder, age group, time since the first visit and time since the last visit. To evaluate the model performance per each subgroup, we relied

only on AUROC as the AP is an inappropriate metric for comparing groups with different prevalence values[33].

To evaluate the model for each diagnosis, the patients were grouped by their latest diagnosis defined by the first level categorisation from the ICD-10[38]. The performance of the general model was significantly higher for Organic disorders (F0), with an AUC of 0.890 (95% CI 0.852-0.928) compared to the overall performance of 0.797 (95% CI 0.793-0.802) and the rest of the disorders. For the other diagnostic groups, the performance ranged between 0.770 (95% CI 0.760-0.779) to 0.814 (95% 0.796-0.831). The lowest performance was observed for mood affective disorders (ICD-10 F3.X), followed by schizophrenia, schizotypal and delusional disorders (ICD-10 F2.X). Separate models for each diagnosis subgroup were developed and compared with the general model that is diagnosis agnostic. The general model consistently outperformed the baseline models and no disorder-specific model was significantly better than the general model (see Figure 3c and Supplementary Figure 2).

The model exhibits a comparable accuracy for most of the age groups, with an AUROC between 0.782 (95% CI 0.771-0.793) and 0.796 (95% CI 0.786-0.806). The model performance dropped to 0.743 (95% CI 0.718-0.767) for patients below 18 years and increased to 0.840 (95% CI 0.820-0.859) for the group of patients who are between 65 and 74 year old (see Figure 3d and Supplementary Figure 3).

As expected, data availability was positively correlated with model accuracy – if there was no information about a patient for one year or more, the accuracy dropped to 0.617 (95% CI 0.592-0.641) whereas for patients who had at least one record within the previous month, the average accuracy was 0.765 (95% CI 0.761-0.771). Longer history of patient data in the EHR of the hospital improved the model accuracy, which ranged from 0.794 (95% CI 0.772-0.817) when a patient had his first visit on the last 6 months to 0.816 (95% CI 0.805-0.827)for patients whose first record dated to 5 or more years. (see Figure 3e, 3f and Supplementary Figure 3)

Analysis of the most predictive features

We used SHAP values[39] to assess the contribution of each feature on the general model output. Figure 4 shows the relative impact of the top 20 features on the model accuracy at each data point in the validation set according to the mean absolute SHAP value. The historical severity of symptoms (specifically, the total number of crisis events and episodes as well as the duration of the last episode), interactions with the hospital (including the unplanned contacts, missed appointments or a recent crisis), combined with patients' characteristics (such as their age and the individual risk indices), and total time since they have been registered in the hospital system carried most of the predictive power in the general model – see Figure 4a.

To further examine the impact of each variable, we represented SHAP values of the top 20 features separately (Figure 4c-4h and Supplementary Figure 4). The recency of records and in particular of the important events (such as crises, unplanned contacts, etc.) had the major impact on the risk prediction model. The presence of important events were positively contributing to the risk calculation until a certain point after which they start driving the risk score down. The impact of the different events exhibited different patterns of fading and some had a long lasting effect whereas the other events impacted the risk score only during the very first weeks after their occurrence. For instance, unplanned contacts with a patient had the highest short-term effect yet their impact almost disappeared after only two weeks. We observed a long effect for the events that encoded contacts with the carer as well as missed appointments – they had a sustained impact on the risk score for 10 and 16 weeks respectively. Referrals and crises also carried a considerable weight on the predicted risk, positive during approximately 6 months (precisely, referrals for 25 weeks and crises for 29 weeks), after which they drove the risk score down. The variables labelling severe symptoms typically had a long lasting impact on the risk score. The examples include referrals from acute services and risks of suicide, for which positive influence to the risk score persisted in the model for more than a year, whereas the substance misuse events impacted the modelled risk for two years after it has been identified.

In most cases, the presence of important events is associated with a deteriorated condition in the past, therefore the absence of certain types of events (denoted in the data with NaN values) for a particular patient suggests historically less severe symptoms and had a negative impact to the PRS. Similarly, if a patient has never been hospitalized has a negative influence on the PRS, whereas the influence is positive if the patient had at least one hospitalization in the past. The age of the patient carried a positive effect on the risk score if a patient was below 60 and a negative for patients above 60 years. The greatest positive influence of the age on the PRS was observed for the individuals younger than 21 and the greatest negative influence started from the age 73 and on-wards, with a very small influence for the patents aged between 21 and 60 years. The total number of crisis episodes experienced by a patient represents the fourth most important feature overall according to the SHAP values. Its influence on the risk scores increases steeply as the number becomes higher, negative for patients who had three crisis episodes and positive for the patents above that threshold. Although with a smaller effect, a similar pattern was observed for the number

of years since the first visit, which has a negative influence on the risk score for patients with less than seven years in the system and positive for nine or more years.

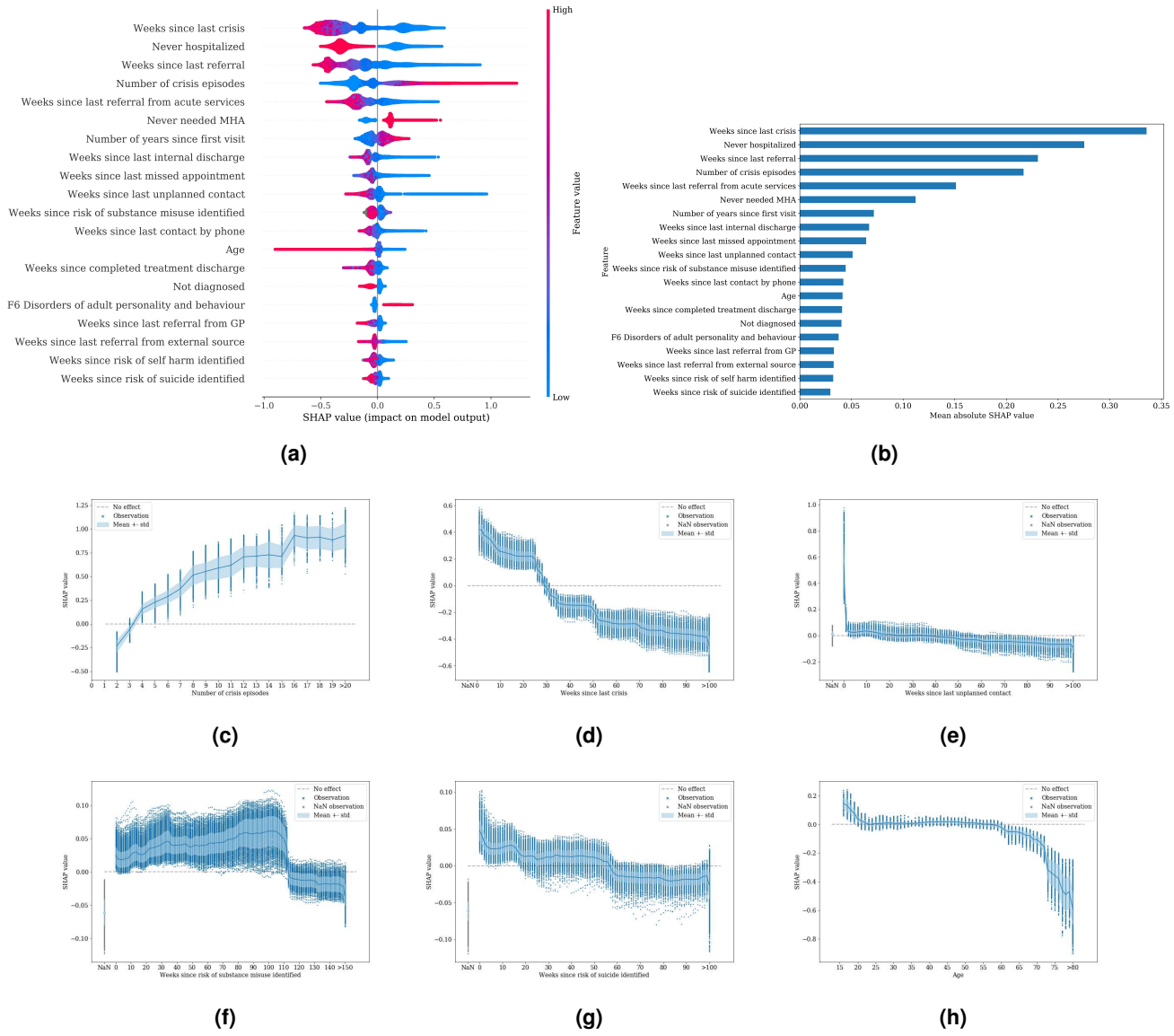

Figure 4. **a** Complete distribution of the SHAP values for the top 20 features based on the highest mean absolute SHAP value. Each sample of the test set is represented as a datapoint per feature and the *x* axis shows the positive or negative impact on the model's prediction of the feature. The color coding depicts the value of the feature and is scaled independently based on its range observed in the data. **b** Absolute feature contribution of the 20 features with the highest mean absolute SHAP value. **c-h** Six examples of dependence plots, showing the impact on the PRS with respect to the feature value. Each datapoint represents a sample in the test set, the solid lines and the lighter-colored envelops represent respectively the mean impact and its standard deviation per feature value. The variability at each feature value is related to the interaction with the rest of features. Missing values are colored in grey.

Clinical evaluation and usefulness

In the prospective trial, we queried our prediction model every two weeks to rank patients based on the PRS – from the highest to lowest predicted risk of experiencing a crisis episode within the upcoming 28 days. Four clinical teams received a dashboard with the top ranked patients – namely, 100 patients with the highest PRS (25 patients were presented to each team; see 1 for the team composition). When evaluated retrospectively, the average precision at 100 of our model was 0.457 (95% CI 0.422-0.493) whereas the precision at 100 was 0.338 (95% CI 0.322 – 0.354). Interestingly, in a prospective evaluation, clinicians disagreed

Table 1 Participants and completion rate of prospective study per team.

N^o (%)	Team 1	Team 2	Team 3	Team 4	Total
Clinicians	n=13	n=19	n=14	n=14	n=60
Nurses	12 (92)	15 (79)	11 (79)	13 (93)	51 (85)
Doctors	1 (8)	2 (11)	0 (0)	1 (7)	4 (7)
Occupational Therapists	0 (0)	1 (5)	1 (7)	0 (0)	2 (3)
Duty workers	0 (0)	1 (5)	1 (7)	0 (0)	2 (3)
Social workers	0 (0)	0 (0)	1 (7)	0 (0)	1 (2)
Form completion	n=292	n=279	n=196	n=244	n=1011
F1	292 (100)	246 (87)	177 (90)	220 (89)	935 (92)
F2	274 (94)	221 (78)	159 (81)	202 (80)	856 (84)

only with 7% ($n = 65$) of all the predictions provided over 6 months, ranging from 3% ($n = 6$) to 12% ($n = 27$) across the four mental health teams. In total, clinicians rated 38% ($n = 351$) of the cases as low risk, 44% ($n = 407$) as medium risk, 13% ($n = 119$) as high risk and less 0.1% ($n = 3$) as being at an imminent risk of experiencing a mental health crisis, whereas 6% ($n = 55$) of the reviewed cases were patients already experiencing a crisis. The risk assessment was part of the feedback form delivered upon an initial review of the presented cases (F1 in Table 2) with the completion rate of 92% ($n = 935$).

One week after the initial review, clinicians rated the usefulness of the risk prediction model with respect to either identifying patients who may be entering a crisis or in managing the caseload priority (the second feedback form, F2, in Table 2). Predictions were rated useful in 64% ($n = 602$) of the presented cases – more than 70% of cases in three out of the four teams. Only one team (Team 4 in Table 2) reported no utility in significantly higher percentage (71%, $n = 145$) than all the other teams that reported less than 30%. In particular, clinicians reported that the model was clinically useful to prevent a crisis in 19% ($n = 175$) of the cases and in identifying deterioration in patients' condition in 17% ($n = 159$) of the presented cases. The model output was used to manage caseload priorities in 28% ($n = 268$) of the cases (see Table 2 for a detailed summary and a break-down by team). The completion rate of F2 was 84% ($n = 846$) (see Table 1 for a detailed partition among the teams).

Table 2 Responses of the feedback forms F1 and F2 per team of clinicians involved in the prospective trial.

N^o (%)	Team 1	Team 2	Team 3	Team 4	Total
F1 Responses	n=292	n=246	n=177	n=220	n=935
Assessment of patient's risk of crisis					
Low risk	99 (34)	89 (36)	48 (27)	115 (52)	351 (38)
Medium risk	136 (47)	96 (39)	92 (52)	83 (38)	407 (44)
High risk	29 (10)	59 (24)	21 (12)	10 (5)	119 (13)
Imminent risk	2 (1)	0 (0)	1 (1)	0 (0)	3 (0)
Already in crisis	26 (9)	2 (1)	15 (8)	12 (5)	55 (6)
Have you taken /do you intend to take any actions as a result of this notification?					
Yes contact to be made (Telf)	9 (3)	15 (6)	11 (6)	8 (4)	43 (5)
Yes contact to be made (F2F)	12 (4)	38 (15)	11 (6)	10 (5)	71 (8)
No, contact made in last 7 days	46 (16)	28 (11)	41 (23)	29 (13)	144 (15)
No, risk already being managed	202 (69)	156 (63)	109 (61)	146 (66)	613 (65)
No, do not agree with assessment	23 (8)	9 (4)	6 (3)	27 (12)	65 (7)
F2 responses	n=274	n=221	n=159	n=202	n=856
What is your current assessment of this patient's condition?					
Low risk	110 (40)	102 (46)	47 (30)	110 (54)	369 (43)
Medium risk	124 (45)	72 (33)	83 (52)	73 (36)	352 (41)
High risk	25 (9)	42 (19)	16 (10)	7 (3)	90 (11)
Imminent risk	1 (0)	2 (1)	0 (0)	0 (0)	3 (0)
Already in crisis	14 (5)	3 (1)	13 (8)	12 (6)	42 (5)
Do you think that this additional information has helped you with..?					
Trying to prevent a crisis	36 (12)	75 (28)	45 (26)	19 (9)	175 (19)
Identifying patient's deterioration	57 (20)	62 (23)	32 (18)	8 (4)	159 (17)
Managing caseload priorities	125 (43)	62 (23)	48 (27)	33 (16)	268 (28)
Nothing, it was not useful	73 (25)	72 (27)	50 (29)	145 (71)	340 (36)

Discussion

We demonstrated the feasibility of predicting mental health crises by applying machine learning techniques on longitudinally collected EHR, reaching an AUROC of 0.797 for the general model. Despite the sparsity of EHR (related to the periods of having no record about a patient), querying the prediction model continuously (week by week) outperformed the baseline models, and the examination of the Brier score and the net benefit suggested the relevance of the decisions based on our predictive model. The lack of records for more than 3 months resulted in a drop of 7% in AUROC whereas not having any records about a patient for more than 6 months or 1 year caused a drop of 13% and 20% respectively. Like in previous predictive analytics experiments[9,11,12,30], gradient boosting (concretely the XGBoost algorithm, eXtreme Gradient Boosting)[31] was superior to the other machine learning models. We attempted to further increase the model performance by leveraging specificity of mental health disorders and by developing disorder specific models – however, training different models for each group of disorders did not prove superior to the general model despite the differences in the performance of the general model along different disorders (Figure 3c).

Beyond the technical feasibility assessed in a retrospective manner, we evaluated whether the mental health risk predictions were perceived as clinically useful when implemented prospectively. The clinicians disagreed with only 7% of the model predictions at average, whereas the model outputs were found clinically useful in 64% of individual cases. We did not succeed in identifying the reasons for witnessing considerably lower scores in one out of the four clinical teams. Importantly, the risk predictions were relevant for preventing crises in 19% of the cases and in identifying the deterioration in a patient's condition in 17% of the cases. A relatively high percentage of cases (36%) in which the predictions were not perceived useful was highly impacted by the number of serious cases that were already recognised and managed by the hospital staff. Nevertheless, the clinicians at our clinical site opted to receive the list of patients who are at the highest risk of experiencing a crisis even though it implies including patents that are already on their radar. In the initial trial design phase, they rejected the option of receiving predictions for the extended (or an entire) list of patients due to an increase in caseload. It is reasonable to expect that the requirements for the practical implementation would not be considerably different in the other clinical settings as hospital teams are typically stretched thin and able to review a limited number of cases at time. Periodically reviewing predictions queried for all the patients registered in the hospital system would perhaps introduce more work for clinicians and be of a little benefit.

The main limitation of our study is related to known and potentially unknown specificity of the single-centre test-bed. EHRs are characterised by high dimensionality and heterogeneity, hence most of the risk prediction may suffer from overfitting the model to the data and limit the generalizability of the results. However, we covered a wide spectrum of different mental health disorders and we also evaluated the usefulness of our model prospectively unlike in a plethora of the previous EHR-based predictive models. Whereas the collection of all the data fields in our dataset and the derived features may depend on the decision by a healthcare provider (such as referrals, change in teams, severity of each crisis, etc.), a great deal of data fields are expected to be routinely captured by typical mental health centers that only register crisis emergencies, visits and hospitalisations. Under this assumption, we selected 8 out of the top 20 features derived solely from the events related to crises, contacts and hospitalisation (see list in Supplementary Material) and evaluated the corresponding model. The resulting AUROC was 0.781 compared to 0.797 in the general model.

We were unable to quantitatively evaluate what percentage of patients labeled by the model with a high risk in the prospective trial ended up experiencing a crisis. Although the clinicians reported 19% of cases in which the prediction model helped in preventing the crisis, the actual outcome was not witnessed in practice as this would have implied the lack of clinicians' reaction to the predictions which would have been ethically and legally unacceptable. The semi-structured interviews further shed the light onto the clinical usefulness of the predictions to identify risky cases that would have been otherwise missed (Clinician Id12: " ... when I reviewed her, I didn't understand why she was on [the list of high-risk patients provided by the model] so did not act on this but on my next follow up with her, I had to call an ambulance").

Typically, healthcare systems have high-inertia and a strong resistance to innovation. Yet, the rising demand for mental healthcare is increasingly prompting hospitals to actively work on identifying novel methods for anticipating demand and better deploying their limited resources to improve patient outcomes and decrease long-term costs[7,40]. Evaluating the technical feasibility and the clinical usefulness are inevitable steps to be made before integrating prediction models into care[28]. Along this vision, our study paves the way towards a more optimised allocation of the mental healthcare staff as well as towards an enablement of a long awaited shift in the mental health paradigm – from reactive care (delivered in the Emergency Room) to preventative care (delivered in the community).

Methods

Study design

This study was conducted in two phases. The first phase included a retrospective study in which we developed and validated a mental health crisis prediction model that relies on EHRs. In the second phase, this model was used during a prospective study to evaluate the perceived usefulness of the predicted risk model derived from the first study in regular clinical settings. The retrospective study relied on the data collected from September 2012 until November 2018. The prospective study was conducted from the 26th of November 2018 until the 12th of May 2019.

Dataset

The dataset comprised anonymized clinical records extracted from a retrospective cohort of all patients who attended the Hospital, one of the largest mental health trust in the country operating over 40 sites and serving a culturally and socially diverse population of over a million. The data included patient's demographics, contacts with the hospital, referrals, diagnosis, hospitalizations, risk and wellbeing assessments and crisis events from all inpatients and outpatients. No exclusion criteria based on age or diagnosed disorder was applied, patients' age ranged from 16 to 102 years. Patients who had less than two crisis episodes or had been in the system for less than three months were excluded from the study. This left the total of 5,816,586 electronic records from 17,122 patients in the database used for this study. Supplementary Table 2 shows a breakdown of the amount of records per type.

Features and labels generation

To prepare the data for the modelling task, all records of each patient were consolidated at a weekly level. Following this process, we generated evenly spaced time series for each patient that spanned from the patient's first interaction with the hospital until the last week of the study.

Label generation

To construct the binary prediction target, each patient-week was assigned a positive label whenever there was a relapse during the time-window of the following four weeks, given that the patient had no crisis during the current week, and a negative label otherwise. Analogously, we built 47 additional labels to validate a certain level of generalisability of the modeling framework. These labels were defined in a similar way as the main target, varying three components:

- The number of stable weeks (without crisis) necessary to consider a crisis episode concluded: from 1 to 4 weeks.
- The prediction time-window length: from 1 to 4 weeks.
- The number of weeks since the prediction is made until the start of the prediction time-window: ranging from 0 to 2 weeks.

Features generation

We extracted a total of 198 features from the 10 data tables (see Supplementary Table 3). The feature extraction was done following six procedures:

- Static or semi-static features. Information from the patient's demographics was accounted as a constant value attributed to the patient for all the weeks. A special case regards to the age, which changes accordingly each year.
- Diagnosed disorder features. Each patient was assigned their latest valid diagnosed disorder, or a not diagnosed label. The diagnosed disorder was mapped to its corresponding first level category according to the ICD-10[38] code system. For instance, F200 Paranoid schizophrenia disorder was mapped to F2 Schizophrenia and Psychotic category.
- EHR weekly aggregations. EHR related to patient-hospital interactions were aggregated in a weekly basis for each patient. The resulting features were counts per type of interaction and one hot encoding features according to their categorization.
- Time elapse features. At each patient-week, for each type of interaction and category we constructed a feature that counted the number of weeks passed since the last time that type occurred. We used NaN values to indicate that up to that point in time the patient never had such type of event.
- Last crisis episode descriptors. For each crisis episode, a set of descriptors summarizing the length and severity of the crisis episode were built. At each patient-week, the values of those descriptors from the latest crisis episode experience up to that week were used as features. We used NaN values to indicate that up to that point in time the patient never had a crisis episode.

- Status features. For those EHR that have an start and end date, the patient was assigned the value or category of the record during all the weeks it lasted for.

Crisis prediction modeling and evaluation

Model evaluation

To emulate the clinical settings, we defined the crisis prediction task as a binary classification problem to be performed in a weekly basis. We employed a time based 80%/10%/10% train/validation/test split.

- Train data started the first week of September 2012 and ended the last week of December 2017.
- Validation data started the first week of January 2018 and ended the last week of June 2018.
- Test data started the first week of July 2018 and ended the third week of November 2018.

Performance evaluation was done on a weekly basis and the results of each week were used to build confidence intervals on the metrics evaluated. All reported results were computed using the test set if not otherwise indicated.

Machine Learning classifiers

For our final models, we used XGBoost[31], an implementation of Gradient Boosting Machines (GBM)[41] as it was the algorithm that performed the best. GBM models are tree based algorithms that build a sequence of trees such that every new tree is constructed to improve the performance from previous iterations. Given that XGBoost is able to handle missing data and is not sensible to scaling factors, we did not apply any imputation or scaling techniques for these models. For comparison, we also evaluated the performance of other commonly state of the art Machine learning classifiers, in particular: Logistic Regression, Naive Bayes, Random Forest and Neural Networks. For those classifiers that benefited from it, standard scaling and imputation of missing values was performed to make a fair comparison. As described in the following section, we did 100 trials of hyperparameter optimization for each of the classifiers to find the best hyperparameters, the search spaces are provided in Supplementary Materials (see Supplementary table 7).

Hyperparameter tuning and feature selection

To select the optimal hyperparameters for each model trained, we maximized AUROC on the validation set through a Bayesian optimization technique. For this purpose, we used Hyperopt[42], a Sequential Model-Based Optimization algorithm that performs Bayesian optimization through Tree Parzen Estimator[43]. This technique has a wide range of distributions available to accommodate to most search spaces. This flexibility makes the algorithm very powerful and appropriate for performing hyperparameter tuning on all the classifiers used. The same methodology was used for feature selection. To that end, we grouped the features on categories based on the information they brought and added a binary parameter to select the feature or not (see Supplementary table 3).

Model interpretation

To measure the contribution of each feature to the main model we used the SHAP values[39]. This technique is based on Shapely values from game theory, which quantifies the individual contributions of all the participants of a game to the outcome, and is the state of the art to interpret Machine Learning models. SHAP values were computed using the Python package shap v0.35.0 through the TreeExplainer algorithm, which is an additive feature attribution method that satisfies the properties of local accuracy, consistency, and allowance for missing data[44]. Feature attributions are computed for every particular prediction, assigning each feature an importance score that takes into account interactions with the rest of the features. The resulting SHAP values provide an overview of the feature's contribution based on its value and allow for both local and global interpretation. All SHAP values presented have been computed on the test set.

Statistical Methods

All reported metrics in text, tables and figures refer to the performance evaluation on the test set, if not otherwise indicated, corresponding to the temporal splits realized following the process described in Crisis prediction modeling and evaluation section in Methods. Confidence intervals of the reported performance metrics were computed using the $n = 25$ temporal splits. Statistical analysis for model comparison was done through analysis of AUROC using its equivalence to Mann-Whitney U-statistic and following the theory on generalized U-statistics to compare correlated ROC curves[45]. For figures showing curves (Figures 3a, 3b and 4c-4h, and Supplementary Figures 1c and 4), solid lines and shaded areas correspond respectively to the mean and standard deviation of the performance metrics across the temporal splits in the test set. For figures displaying point plots (Figures 3d and 3e), center points and vertical bars correspond respectively to the mean and 95% confidence interval across the temporal splits in the test set. For box-plot figures (Figure 3c), the solid line corresponds to the median value, the box limits to the first Q1 (left limit) and third (right limit) quartiles, the whiskers denote the rest of the distribution range from

Q1-1.5(Q3-Q1) (left whisker) to Q3+1.5(Q3-Q1) (right whisker) and the points displayed correspond to the outliers.

We evaluated the calibration of our proposed model and the model per diagnosis, i.e. compared the PRS of the model against the observed risk aggregating the observed labels. In order to calibrate the risk scores, we fitted an isotonic regression model[34] on the validation set's predictions and transformed the test set's predictions. As a result, a rank-preserving transformation that minimizes the deviation between the actual target variable and the PRS was applied. We used 25 evenly spaced bins on the PRS to generate the calibration curve in Supplementary Figure 1a-b .

Clinical evaluation

The prospective study included 4 community mental health teams (CMHTs) within the Hospital. In total, 60 clinicians from the the 4 CMHT participated in the study of which 4 were doctors, 2 occupational therapists, 2 duty workers, 1 social worker, and the majority (51) were nurses – including clinical leads and team managers. Each team had at least 2 coordinators, referred to as "digital champions". They served as the first contact point in their respective teams and they were in charge of assigning individual cases to the clinical participants in this study.
Prior to the study, all the clinicians involved in this study attended a set of training sessions. This sessions included a presentation of the project and its aims, a description of how the risk prediction tool worked, what was expected from them and a demonstration on how to use the tool. Additionally, there was a period of four weeks before the study in which the end-to-end process was tested to ensure that potential technical problems were timely addressed and to fine tune the process based on the early feedback from participants.

During the prospective study, the general model was applied on a biweekly basis to generate the PRS for all the patients. The model relied on the latest available data. Subsequently, the patients were ranked based on the PRS and the top 25 patients from each of the CMHT were presented to the clinicians. The tool used by the participants contained a list of patient names, patient identifier, risk score and relevant clinical and demographic information (Supplementary Table 9). Upon reviewing the list of patients, the participants completed F1 feedback form where they provided their estimation of the patient's risk level and specified their intended action as result of the prediction. A week after the initial review, participants were asked to record, on the F2 feedback forms, if they had changed their assessment of the patient's risk level upon further assessment and feedback on how useful the information provided to them was in either preventing a crisis or identifying a patient deteriorating to provide support or in managing their caseload priority. See the questions in Table 2.

The prospective pilot study included a total number of 1,011 patients. The initial objective was to include 1,200 patients, however 189 patients were discarded from the analysis due to an internal technical error in the crises coding. Importantly, this error did not impact the results of this study beyond reduction in the sample size.

Ethics approval

Ethics review and institutional review boards approved the study with waiver of informed consent or exemption at each institution.

References

1. World Health Organization. Depression and other common mental disorders: Global health estimates (2017).
2. Wainberg, M. *et al.* Challenges and opportunities in global mental health: a research-to-practice perspective. *Curr. Psychiatry Reports* **19**, DOI: 10.1007/s11920-017-0780-z (2017).
3. Fiorillo, A. & Gorwood, P. The consequences of the covid-19 pandemic on mental health and implications for clinical practice. *Eur. Psychiatry* **63**, e32, DOI: 10.1192/j.eurpsy.2020.35 (2020).
4. Duan, L. & Zhu, G. Psychological interventions for people affected by the covid-19 epidemic. *The Lancet Psychiatry* **7**, DOI: 10.1016/S2215-0366(20)30073-0 (2020).
5. National Alliance of Mental Illness. Navigating a mental health crises (2018).
6. Miller, V. & Robertson, S. A role for occupational therapy in crisis intervention and prevention. *Aust. Occup. Ther. J.* **38**, 143–146, DOI: https://doi.org/10.1111/j.1440-1630.1991.tb01710.x (1991). https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1440-1630.1991.tb01710.x.
7. Horwitz, L. I., Kuznetsova, M. & Jones, S. A. Creating a learning health system through rapid-cycle, randomized testing. *New Engl. J. Medicine* **381**, 1175–1179, DOI: 10.1056/NEJMsb1900856 (2019). PMID: 31532967, https://doi.org/10.1056/NEJMsb1900856.

8. Van Le, D., Montgomery, J., Kirkby, K. C. & Scanlan, J. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. *J. biomedical informatics* **86**, 49–58 (2018).

9. Ye, C. *et al.* Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning. *J Med Internet Res* **20**, e22, DOI: 10.2196/jmir.9268 (2018).

10. Arcadu, F. *et al.* Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *npj Digit. Medicine* **2**, DOI: 10.1038/s41746-019-0172-3 (2019).

11. Hyland, S. *et al.* Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Medicine* **26**, 1–10, DOI: 10.1038/s41591-020-0789-4 (2020).

12. Lin, H. E., Tan, I.-H., Lee, I., Wu, P. & Chong, H. Predicting readmission at early hospitalization using electronic health data: A customized model development. *Int. J. Integr. Care* **17(5)**, DOI: http://doi.org/10.5334/ijic.3826 (2017).

13. Rajkomar, A. *et al.* Scalable and accurate deep learning for electronic health records. *npj Digit. Medicine* **1**, DOI: 10.1038/s41746-018-0029-1 (2018).

14. Walsh, C. G., Ribeiro, J. & Franklin, J. Predicting risk of suicide attempts over time through machine learning. *Clin. Psychol. Sci.* **5**, 457 – 469 (2017).

15. Simon, G. *et al.* Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *The Am. journal psychiatry* **175**, appiajp201817101167, DOI: 10.1176/appi.ajp.2018.17101167 (2018).

16. Barak-Corren, Y. *et al.* Predicting suicidal behavior from longitudinal electronic health records. *Am. journal psychiatry* **174**, 154–162 (2017).

17. Chen, Q. *et al.* Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: A machine learning study using swedish national registry data. *PLoS medicine* **17**, DOI: 10.1371/journal.pmed.1003416 (2020).

18. Kessler, R. *et al.* Predicting suicides after psychiatric hospitalization in us army soldiers the army study to assess risk and resilience in servicemembers (army starrs). *JAMA psychiatry* **72**, 49–57, DOI: 10.1001/jamapsychiatry.2014.1754 (2015).

19. Poulin, C. *et al.* Predicting the risk of suicide by analyzing the text of clinical notes. *PLOS ONE* **9**, 1–7, DOI: 10.1371/journal.pone.0085733 (2014).

20. Su, C. *et al.* Machine learning for suicide risk prediction in children and adolescents with electronic health records. *Transl. Psychiatry* **10**, DOI: 10.1038/s41398-020-01100-0 (2020).

21. Fernandes, A. C. *et al.* Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Sci. reports* **8**, 1–10 (2018).

22. Olfson, M., Marcus, S. & Bridge, J. Emergency department recognition of mental disorders and short-term outcome of deliberate self-harm. *The Am. journal psychiatry* **170**, DOI: 10.1176/appi.ajp.2013.12121506 (2013).

23. Raket, L. L. *et al.* Dynamic electronic health record detection (detect) of individuals at risk of a first episode of psychosis: a case-control development and validation study. *The Lancet Digit. Heal.* **2**, e229 – e239, DOI: https://doi.org/10.1016/S2589-7500(20)30024-8 (2020).

24. Suchting, R., Green, C. E., Glazier, S. M. & Lane, S. D. A data science approach to predicting patient aggressive events in a psychiatric hospital. *Psychiatry Res.* **268**, 217–222, DOI: https://doi.org/10.1016/j.psychres.2018.07.004 (2018).

25. Mohr, D. C., Riper, H. & Schueller, S. M. A solution-focused research approach to achieve an implementable revolution in digital mental health. *JAMA psychiatry* **75**, 113–114 (2018).

26. Graham, A. *et al.* Lessons learned from service design of a trial of a digital mental health service: Informing implementation in primary care clinics. *Transl. Behav. Medicine* **10**, 598–605, DOI: 10.1093/tbm/ibz140 (2020).

27. Bardram, J. E. & Matic, A. A decade of ubiquitous computing research in mental health. *IEEE Pervasive Comput.* **19**, 62–72, DOI: 10.1109/MPRV.2019.2925338 (2020).

28. Salazar de Pablo, G. *et al.* Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophr. Bull.* DOI: 10.1093/schbul/sbaa120 (2020). Sbaa120, https://academic.oup.com/schizophreniabulletin/advance-article-pdf/doi/10.1093/schbul/sbaa120/35274544/sbaa120.pdf.

29. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. reports* **6**, 1–10 (2016).

30. Nielsen, D. *Tree boosting with xgboost-why does xgboost win" every" machine learning competition?* Master's thesis, NTNU (2016).

31. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794, DOI: 10.1145/2939672.2939785 (Association for Computing Machinery, New York, NY, USA, 2016).

32. Boyd, K., Eng, K. H. & Page, C. D. Area under the precision-recall curve: Point estimates and confidence intervals. In Blockeel, H., Kersting, K., Nijssen, S. & Železný, F. (eds.) *Machine Learning and Knowledge Discovery in Databases*, 451–466, DOI: 10.1007/978-3-642-40994-3_29 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).

33. Ozenne, B., Subtil, F. & Maucort-Boulch, D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol.* **68**, 855 – 859, DOI: https://doi.org/10.1016/j.jclinepi.2015.02.010 (2015).

34. Zadrozny, B. & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, 694–699, DOI: 10.1145/775047.775151 (Association for Computing Machinery, New York, NY, USA, 2002).

35. Steyerberg, E. *et al.* Assessing the performance of prediction models a framework for traditional and novel measures. *Epidemiol. (Cambridge, Mass.)* **21**, 128–38, DOI: 10.1097/EDE.0b013e3181c30fb2 (2010).

36. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weather. Rev.* **78**, 1–3, DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2 (1950).

37. Vickers, A. J. & Elkin, E. B. Decision curve analysis: A novel method for evaluating prediction models. *Med. Decis. Mak.* **26**, 565–574, DOI: 10.1177/0272989X06295361 (2006). PMID: 17099194, https://doi.org/10.1177/0272989X06295361.

38. World Health Organization. Icd-10 : international statistical classification of diseases and related health problems : tenth revision (2004).

39. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc., 2017).

40. Graham, A. K. *et al.* Implementation strategies for digital mental health interventions in health care settings. *Am. Psychol.* **75**, 1080 (2020).

41. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals Stat.* **29**, 1189–1232 (2001).

42. Bergstra, J., Yamins, D. & Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Dasgupta, S. & McAllester, D. (eds.) *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 of *Proceedings of Machine Learning Research*, 115–123 (PMLR, Atlanta, Georgia, USA, 2013).

43. Bergstra, J. S., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. & Weinberger, K. Q. (eds.) *Advances in Neural Information Processing Systems 24*, 2546–2554 (Curran Associates, Inc., 2011).

44. Lundberg, S. M. *et al.* Explainable ai for trees: From local explanations to global understanding (2019). 1905.04610.

45. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845 (1988).

Figures

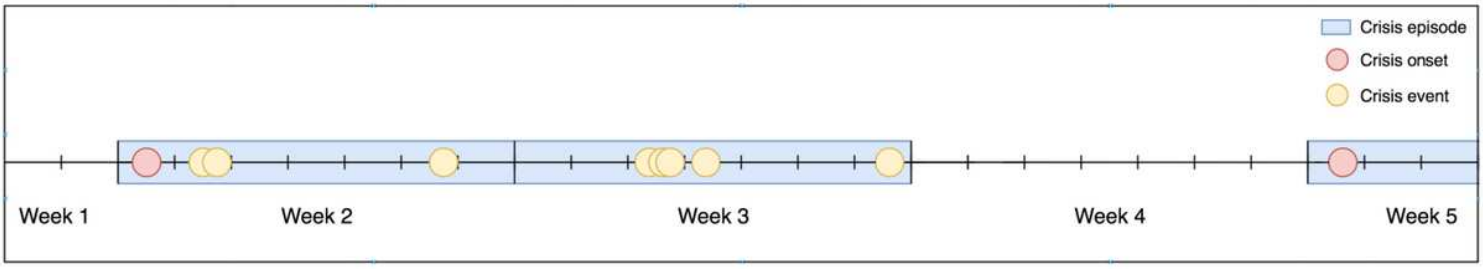

Figure 1

Timeline showing an example of crisis episode, where the crisis onset is the first crisis event of the crisis episode after a stable week without crisis events.

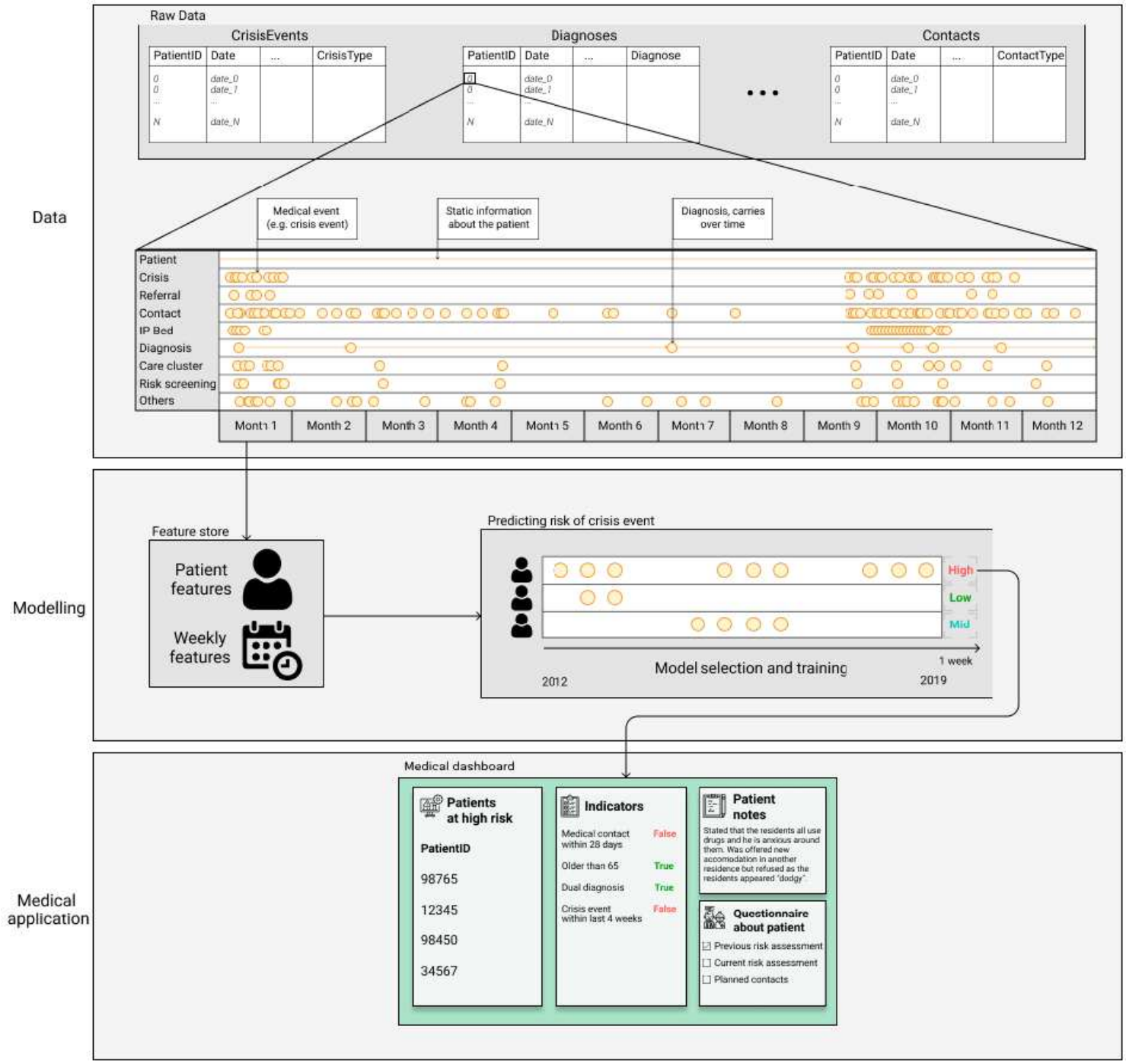

Figure 2

System diagram. Events over time are represented with their timestamp and characteristics in different SQL tables in the hospital's database. Those tables are processed into patient and weekly features for the modelling task. Models are trained, tuned and selected on data ranging from 2012 to 2019. The system then predicts every week - for every patient - the risk of crisis onset in the following four weeks. The patients with highest expected risk are displayed in the medical dashboard available to clinicians, along with key indicators, patient notes and a questionnaire to be filled by the clinician regarding the patient.

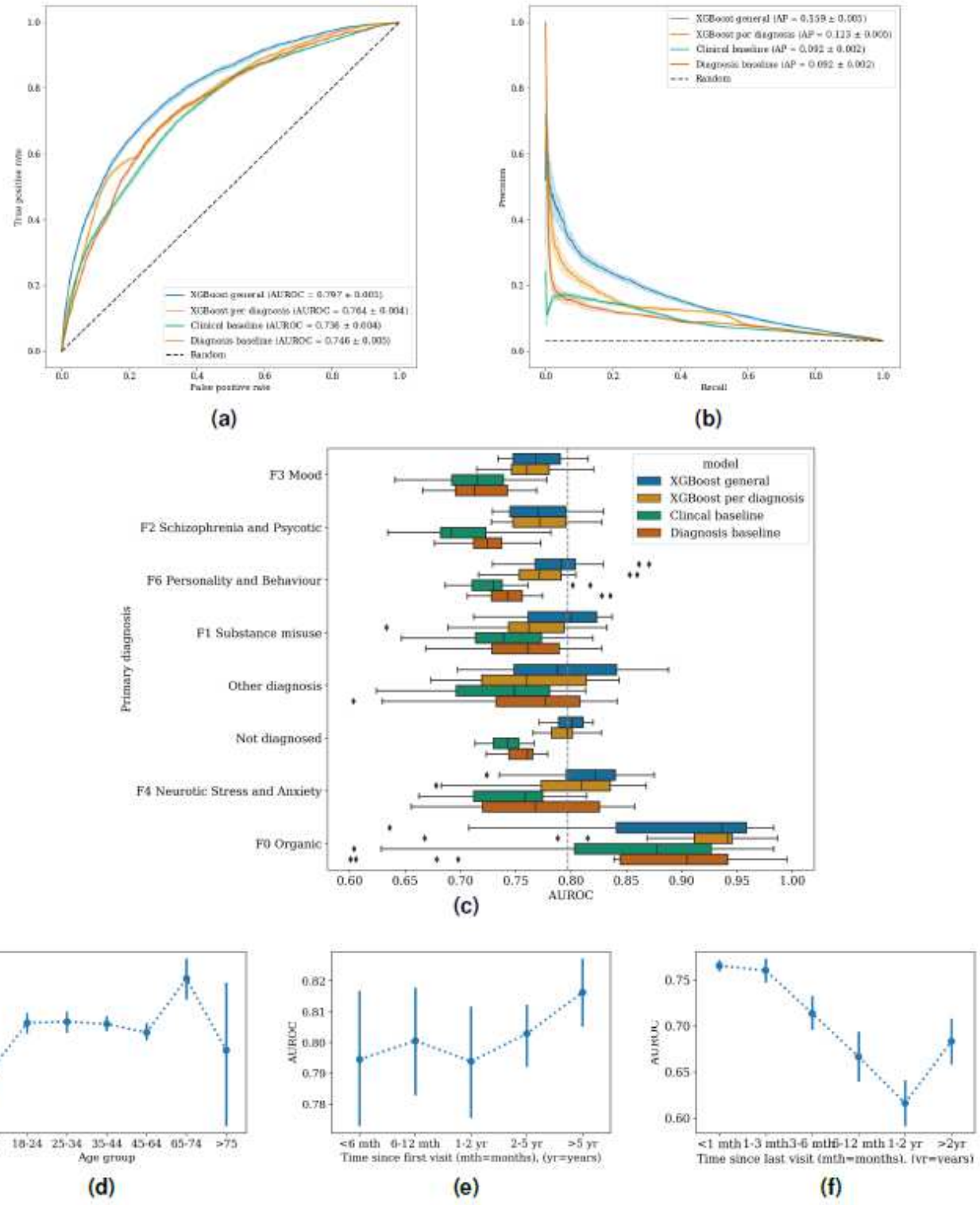

Figure 3

a Receiver operating characteristic curve for the crisis prediction task. Comparison between the proposed final model (XGBoost general), a proposed diagnosis specific model (XGBoost per diagnosis) and two baseline models. The solid lines and lighter-colored envelopes around each line were derived from the test evaluations (n = 25) as the mean and 95% confidence interval respectively. b Precision recall curve for the crisis prediction task with the same characteristics as a. c Box-plot of the area under the receiver

operating curve evaluated per diagnosis. Comparison between the four models considered, same as in a, b. d, e, f Area under the receiver operator curve evaluated on different subsets of the cohort based on age group (d), time since the patient had its first visit on the hospital (e) and time since last crisis episode (f).

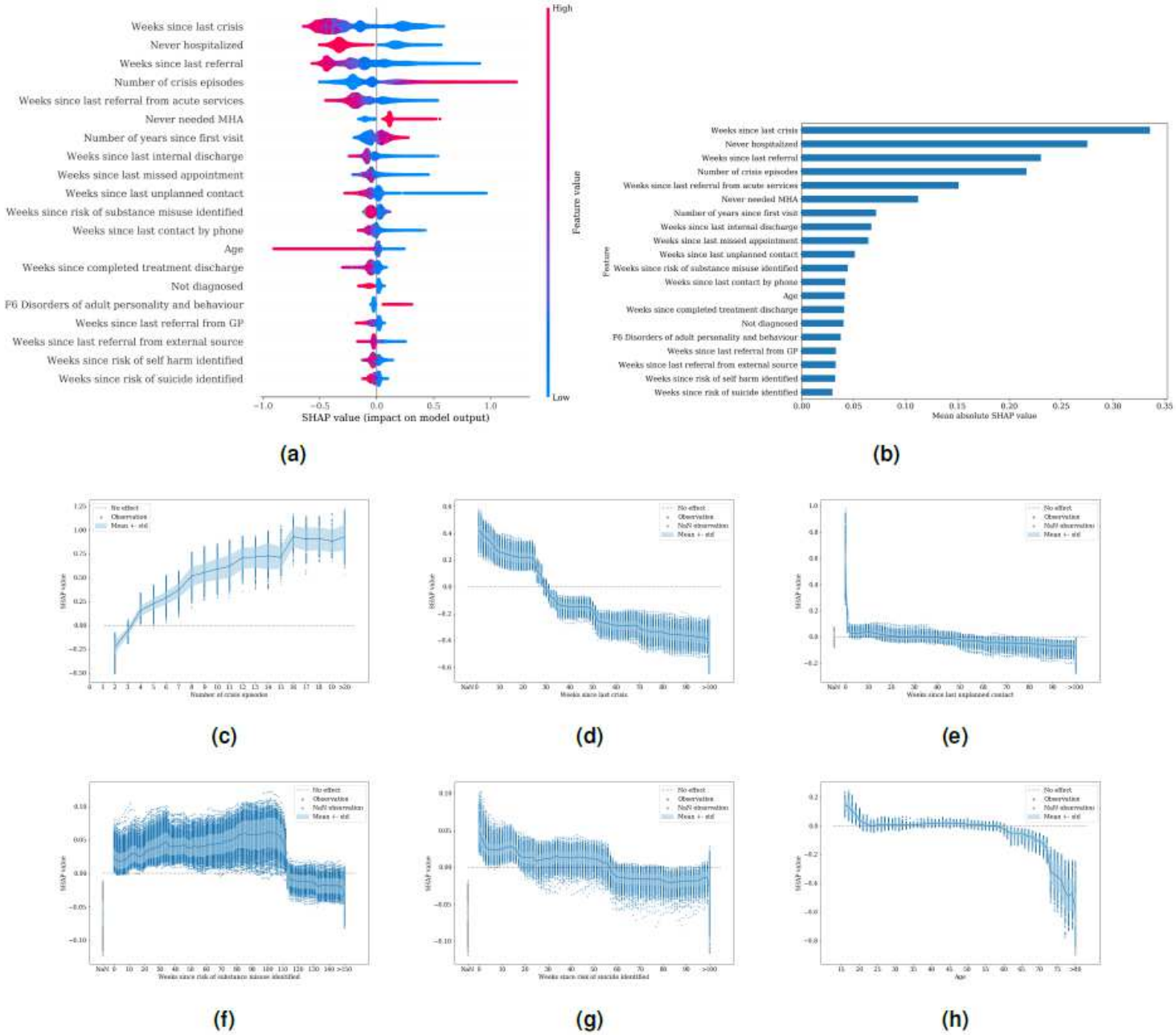

Figure 4

a Complete distribution of the SHAP values for the top 20 features based on the highest mean absolute SHAP value. Each sample of the test set is represented as a datapoint per feature and the x axis shows the positive or negative impact on the model's prediction of the feature. The color coding depicts the value of the feature and is scaled independently based on its range observed in the data. b Absolute feature contribution of the 20 features with the highest mean absolute SHAP value. c-h Six examples of dependence plots, showing the impact on the PRS with respect to the feature value. Each datapoint

represents a sample in the test set, the solid lines and the lighter-colored envelops represent respectively the mean impact and its standard deviation per feature value. The variability at each feature value is related to the interaction with the rest of features. Missing values are colored in grey.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MachinelearningmodeltopredictmentalhealthcrisisfromelectronichealthrecordsSupplementary.pdf
- Table3SuppMatFeaturesList.csv
- Table8SuppMatMultipleTargetsPerformance.csv