Trustworthy Data and AI Environments for Clinical Prediction: Application to Crisis-Risk in People with Depression

Yamiko Joseph Msosa, Arturas Grauslys, Yifan Zhou, Tao Wang, Iain Buchan, Paul Langan, Steven Foster, Michael Walker, Michael Pearson, Amos Folarin, Angus Roberts, Simon Maskell, Richard Dobson, Cecil Kullu and Dennis Kehoe [*†‡§¶‖]

September 6, 2023

Abstract

Depression is a common mental health condition that often occurs in association with other chronic illnesses, and varies considerably in severity. Electronic Health Records (EHRs) contain rich information about a patient's medical history and can be used to train, test and maintain predictive models to support and improve patient care. This work evaluated the feasibility of implementing an environment for predicting mental health crisis among people living with depression based on both structured and unstructured EHRs. A large EHR from a mental health provider, Mersey Care, was pseudonymised and ingested into the Natural Language Processing (NLP) platform CogStack, allowing text content in binary clinical notes to be extracted. All unstructured clinical notes and summaries were semantically annotated by MedCAT and BioYODIE NLP services. Cases of crisis in patients with depression were then identified. Random forest models, gradient boosting trees, and Long Short-Term Memory (LSTM) networks, with varying feature arrangement, were trained to predict the occurrence of crisis. The results showed that all the prediction models can use a combination of structured and unstructured EHR information to predict crisis in patients with depression with good

[*] This work was supported by Innovate UK under Grant number 104542; National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre (BRC); NIHR University College London Hospitals BRC; and Maudsley Charity.

[†] Y. J. Msosa, T. Wang, A. Folarin and R. Dobson are with the Department of Biostatistics & Health Informatics, King's College London, UK (e-mails: yamiko.msosa, tao.wang, amos.folarin, angus.roberts, richard.j.dobson@kcl.ac.uk).

[‡] A. Grauslys, Y. Zhou and S. Maskell are with the Department of Electrical Engineering and Electronics, University of Liverpool, UK (e-mails: a.grauslys, yifanz, s.maskell@liverpool.ac.uk).

[§] M. Pearson and I. Buchan are with the Institute of Population Health, University of Liverpool, UK (e-mails: mpearson, buchan@liverpool.ac.uk).

[¶] S. Foster and C. Kullu are with the Mersey Care NHS Foundation Trust, UK (e-mails: steven.foster, cecil.kullu@merseycare.nhs.uk).

[‖] P. Langan, M. Walker, and D. Kehoe are with AIMES, UK (e-mails: paul.langan, michael.walker, dennis.kehoe@aimes.net).

and useful accuracy. The LSTM network that was trained on a modified dataset with only 1000 most-important features from the random forest model with temporality showed the best performance with a mean AUC of 0.901 and a standard deviation of 0.006 using a training dataset and a mean AUC of 0.810 and 0.01 using a hold-out test dataset. Comparing the results from the technical evaluation with the views of psychiatrists shows that there are now opportunities to refine and integrate such prediction models into pragmatic point-of-care clinical decision support tools for supporting mental healthcare delivery.

Keywords: Depression, Digital Health, Artificial Intelligence, Natural Language Processing, Mental Health Crisis

1 Introduction

Depression can cause extensive periods of disability [1]. This mental illness continues to be one of the most common conditions amongst mental health patients in England [2,3]. Depression occurs in association with other chronic conditions, lifetime adversity, and in various grades of severity [2, 4]. Such cases cause a strain in mental health services that cost England alone an estimated average of 12 billion pounds a year to treat patients [5]. There are few diagnostic tests for mental health conditions, in contrast to physical health illnesses, and therefore clinicians increasingly rely on clinical notes during diagnosis and treatment of mental health patients [6, 7]. Frontline clinical mental health practitioners come from a range of professional backgrounds including psychiatrists, psychologists, mental health nurses, social workers, allied health professionals, and physician associates. Good mental health care depends on staff from such a diverse range of disciplines recording a careful patient history in clinical notes, which is used collectively with other clinical records and expert knowledge to formulate an appropriate care plan for the patient [8]. Each professional discipline can record and store their prose records in separate places using disparate systems, and sometimes with subtly different wording for similar concepts. When a clinician conducts an assessment they have to assimilate previously recorded clinical records before adding their own and creating a care plan [9].

Digital health promises to improve the way healthcare is delivered by providers, through the use of information and communications technologies to monitor and improve the health and wellbeing of patients [10]. Recent advances, including those in Artificial Intelligence (AI) and Natural Language Processing (NLP), make it possible to bring together clinical notes from different professionals and avail meaningful insights from the Electronic Health Record (EHR) [11, 12]. This work assessed the feasibility of applying NLP and AI in the management of depression to recognise patterns of patient behaviour that require intervention, thus providing a foundation for aiding the healthcare professional with indicators of important features from past care.

2 Background

This section discusses related work focusing on application of clinical NLP to EHRs to support predictive analytics for mental health.

2.1 Electronic Health Records for Mental Health

EHRs continue to gain traction across the globe as more evidence becomes available showing that they can improve healthcare whilst reducing costs [13,14]. This is also true for the UK, as evidenced in the National Health Service (NHS) Long Term Plan [15], where health institutions are being encouraged to incorporate more health and medical technologies including EHRs to support *"digital first"* care [16]. Specifically for mental health, the UK Government's strategy recommends increased use of digital health to improve care and access to services [17]. Although digital health systems in the UK are more diverse and fragmented, most mental health NHS trusts have achieved near-full digitisation of clinical information and have a greater capacity to share information [18].

2.2 Clinical Natural Language Processing

Natural Language processing can be used to surface structured patient data from routinely-collected clinical notes, from EHRs, which are a rich source of health information [19, 20]. Applying NLP techniques to identify and extract relevant information, from clinical text, can be challenging due to the inherent nature of written text that can contain typographical errors and linguistic nuances [21]. However, NLP applications can apply a wide range of text analytics methods to support the identification of contextualised mentions of biomedical concepts in clinical text [22, 23]. Such NLP tools and methods can support mental health research and care delivery in clinical practice [24].

Named entity recognition + linking (NER+L) for biomedical text allows the detection of concepts in clinical text without providing manual annotations [25, 26]. There are a growing number of biomedical NER+L tools. Metamap provides tooling for mapping biomedical text to the Unified Medical Language System (UMLS) [27] although it has limited support for handling ambiguous concepts and spelling mistakes [26, 28]. BioYODIE provides improved disambiguation capabilities when compared to Metamap, but requires supervised training with a labelled corpus [29]. SemEHR additional capabilities to those of BioYODIE by applying manual rules for improved performance though its manual-rule application can be quite labour-intensive and time-consuming [23]. cTAKES is built on top of existing open-source technologies: the Unstructured Information Management Architecture framework and the NLP toolkit, OpenNLP [30, 31]. However, cTakes does not handle mispellings and biomedical concept disambiguation without separate plugins [26,30]. SciSpacy provides a supervised model for NER with limited linking capabilities for scientific and biomedical text [32]. Bioportal annotator provides a fixed algorithm that is accessible via a web-hosted API with limited capabilities for customisation and support of non-english corpus [33]. MedCAT addresses many of the shortcomings in biomedical NER+L tools by supporting unsupervised training and clinician-driven concept contextualization [26].

2.3 Artificial Intelligence in Health Informatics

EHRs are a rich but underused source of health information that provide an opportunity for predictive modelling, thereby providing valuable insights for better patient management and care [19]. EHR data has been used to assess the feasibility of applying machine learning techniques to manage caseload priorities and to intervene so that the risk of crises can be mitigated [34]. But building a predictive model using clinical records is not a trivial task. Long-term time dependencies between clinical events such as diagnosis and treatment, in addition to inadequate interpretation of results, complicates the machine learning process and hinders model integration in the clinical setting [35].

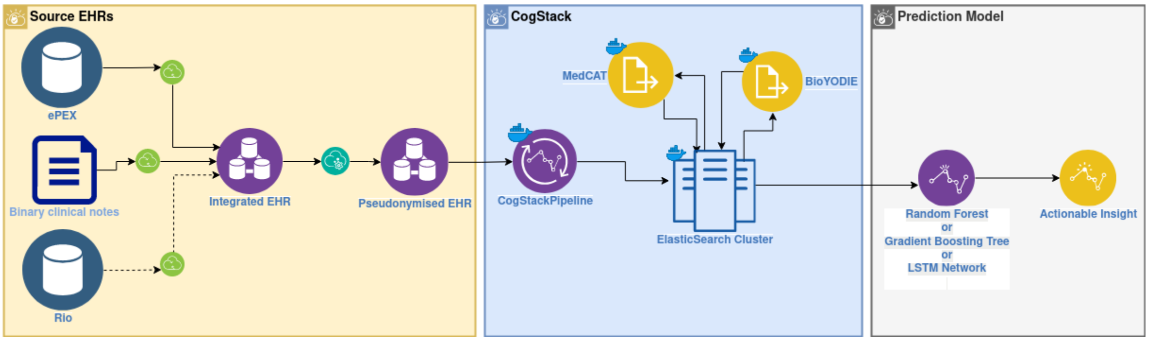


Figure 1: The architecture of a reference implementation of the AI prediction environment

Machine learning has been applied widely to support human health and wellbeing in areas such as: bioinformatics, medical imaging, pervasive sensing, medical informatics and public health [35].

3 Methods

This section details the materials and methods that were used in this work. Fig. 1 shows the overall architectural framework of a reference implementation for this work.

3.1 Scope and Context

This work was a collaboration of the Mersey Care NHS Foundation Trust, King's College London (KCL), University of Liverpool (UoL), and AIMES. Mersey Care is the largest mental health NHS Foundation Trust in the UK that provides specialist inpatient and community mental health services to more than 11 million people across 85 sites. AIMES is a spin-out company from the UoL that focuses on health data solutions. KCL and UoL are research institutions that carry out informatics research. The broader aim of this collaborative is to apply NLP and AI techniques to predict the onset of crisis amongst mental health patients and evaluate the impact of related healthcare interventions for mental health service users. This particular study focused on evaluating the technical feasibility of predicting crisis to support the management of patients suffering from depression. The performance of algorithms used in this work are explicitly compared with those advocated in recent related work [34].

3.2 Electronic Health Record Curation

EHR data covering a period of about 11 years, from 2007 up to 2018, were obtained from Mersey Care. The EHRs at Mersey Care are stored in two separate EHR systems. The first system, ePEX, is a legacy EHR system that contains ten years worth of historical data. Whereas the second system, Rio, is the successor EHR system to ePEX and is currently in use at Mersey Care. The care records covered both structured and unstructured data, where unstructured clinical notes were stored in various binary formats such as portable document format (PDF) and Microsoft Word document (MS Word).

Source system adapters were created to extract, pseudonymise and transform the health data into a standard format from the source EHR systems. Pseudonimisation allows personal details to be removed and substituted, thereby providing a mechanism that can maintain privacy, at the same time allowing linkage of the health data to other key information [36]. This was key for the AI environment so that it met information governance requirements that were set by the Mersey Care NHS Trust. The source system adapters allowed the standardised EHRs to be imported into a common data model, CareXML, that uses an NHS-compliant data dictionary to facilitate storage of integrated care records.

3.3 Natural Language Processing with CogStack

Both structured and unstructured data from the CareXML integrated care records were ingested into an indexed and searchable repository using the CogStack platform. CogStack provides a patient-centric view of the EHR that is availed through a number of document processing services, such as binary-to-text conversion, and NLP services that are based on open standards. The services within CogStack are orchestrated by a fault-tolerant batch processing framework to process and combine the structured and unstructured data from relational services into a configurable, searchable and patient-centric view of the EHR that is accessible through ElasticSearch [37]. Specifically, three EHR data pipelines were used in this process.

The first pipeline ingested structured data into its own searchable index. The second pipeline carried out binary-to-text conversion of the unstructured EHR data, and ingested the extracted data into a separate searchable index.

The third pipeline generated semantic annotations from the previously extracted unstructured text. This process was achieved by applying NER+L services with the UMLS over the unstructured EHR data. CogStack NLP services allow mentions that correspond to concepts in a given set of a controlled vocabulary, such as the UMLS, to be found [21].

3.4 Building and Evaluating the Prediction Models

A methodical approach was taken to define crisis and select suitable training data.

3.4.1 Definition of Crisis

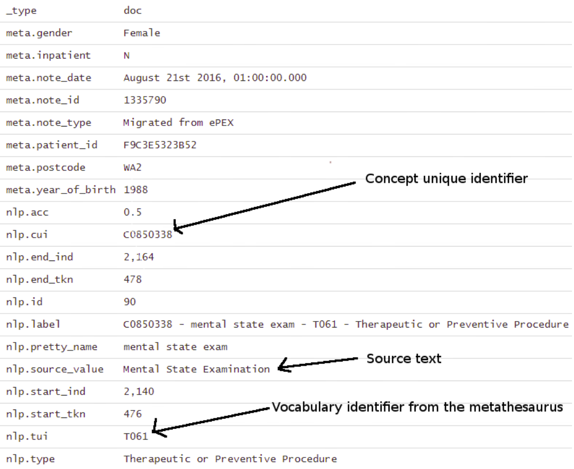
Crisis in the context of mental health is subjective as it is not a formal term. As a result and with guidance from clinicians from Mersey Care, an objective definition of crisis was agreed upon for the purposes of this study. A mental health *'crisis'* was defined as an event that was recorded in the EHR that led to one of the following:

1. A visit to the accident and emergency (A&E) department at the hospital;

2. Contact with the crisis resolution and intensive home treatment (CRHT) team;

3. Hospitalisation.

Any occurrence of at least one of these events was used as a proxy for mental health crisis. This definition of crisis is in line with the pragmatic service-oriented definition of crisis by the Joint Commissioning Panel for Mental Health [38].



_type	doc
meta.gender	Female
meta.inpatient	N
meta.note_date	August 21st 2016, 01:00:00.000
meta.note_id	1335790
meta.note_type	Migrated from ePEX
meta.patient_id	F9C3E5323B52
meta.postcode	WA2
meta.year_of_birth	1988
nlp.acc	0.5
nlp.cui	C0850338
nlp.end_ind	2,164
nlp.end_tkn	478
nlp.id	90
nlp.label	C0850338 - mental state exam - T061 - Therapeutic or Preventive Procedure
nlp.pretty_name	mental state exam
nlp.source_value	Mental State Examination
nlp.start_ind	2,140
nlp.start_tkn	476
nlp.tui	T061
nlp.type	Therapeutic or Preventive Procedure

Concept unique identifier
Source text
Vocabulary identifier from the metathesaurus

Figure 2: A sample MedCAT annotation entry

3.4.2 Cohorting

In order to identify a representative cohort of patients with depression, a search of depression-related diagnoses was carried out on unstructured clinical notes. This approach was taken because diagnoses were often not explicitly recorded in the structured EHR. A list of semantic annotations for depressive disorders were identified and their related patient notes and documents from the EHR were selected. All the annotations that had a negation flag, and their related EHR data, were removed from the dataset. Only those patients from the dataset that had affirmed annotations of depressive disorders regardless of comorbidity, were included in the cohort.

3.4.3 Training and Hold-out Test Data

Patient profiles were characterised with the following types of features:

1. Medically-relevant semantic annotations that were generated from patient notes by the MedCAT [26] NLP service (Fig. 2 shows a sample MedCAT annotation entry),

2. Event type of each patient note,

3. Demographic information for each patient.

All patients in the cohort were assigned to one of two groups depending on whether their clinical history included at least one 'crisis' event or not.

For patients in the *'crisis cohort'*, a six-month (24 weeks) period was selected preceding a crisis event. A two-week-buffer period was created before each crisis event to prevent possible inclusion of clinical notes related to oncoming crisis, such as hospitalisation notes in advance of a hospitalisation event. As for patients in the *'no crisis cohort'*, an equivalent period to the *'crisis cohort'* was chosen at random, but not less than six months before a patient was registered into the EHR system.

The semantic annotations for the patient notes and clinical documents in the six-month period were then counted and collected as features. The features were extended by adding counts of *'patient note types'* and *'contact types'* from the selected period. A *'patient note type'* refers to a recorded kind of contact with the mental health service, such as a telephone call or a community nurse visit. The dataset that was extended with *'patient note type'* in its feature set was tagged as a *'standard dataset'*. A *'contact type'* refers to a kind of intervention with the mental health service that can include instances where crisis is predicted by a clinician and measures are being taken. For example, forensic psychological services contacts for offenders seeking mental health services or inpatient contacts for supporting routine engagement with inpatients. The dataset that was extended with *'contact type'* in its feature set was tagged as an *'extended dataset'*. Furthermore, structured EHR data was added as features to both the standard and extended datasets by including age, gender, ethnic background and presence of disability. Individual records in the dataset were labelled with a binary classification that indicated the presence or absence of a crisis event after the selected time period.

The dataset was divided into two, a *training* dataset and a *hold-out test* dataset, using timestamps from the EHRs. The cross-validation training set included data that was recorded in the EHR between Jan 2007 and June 2017 (85% of the overall dataset). The hold-out test set included the data recorded in the EHR after June 2017 (15% of the overall dataset) with the aim of assessing the performance that would have resulted from deploying the predictive models (after they were trained).

3.4.4 Model Training and Evaluation

An iterative approach was taken to progressively construct and evaluate prediction models. Through that process, random forests (RFs), gradient boosting (GB) trees (which were also considered in [34]), and Long Short-Term Memory (LSTM) networks were trained to predict crisis from the clinical text and structured EHR features. We considered six scenarios with varying kinds of feature arrangement and prediction algorithms.

Random forests and gradient boosting trees classify data using a number of decision trees as their predictive analytics method and they have been widely used with good performance [39–42]. Specifically, a random classifier adds an additional layer of randomness to bagging by splitting each node using the best among a subset of predictors that have been randomly chosen at that node [40]. In the first scenario, with a random forest, a model $M1$ was trained by the standard dataset which consisted of around 41641 features which was depending on different fold in the 10-fold cross-validation. This random forest utilised 1000 estimators with a tree depth of up to eight. Thereafter, feature importance was assessed using the mean reduction in the Gini index [43]. All predictors that were assigned zero importance were filtered out from the dataset. In the second scenario, with a separate random forest, a model $M2$ was trained to investigate the influence of using temporal information in the training data. In this scenario, temporal-indexed features were added to the training dataset. As random forests are not designed to process temporal data, the input data was adapted to include week-long time windows. Each 6-month time period preceding a crisis event was separated into $4 \times 6 = 24$ subsets. This resulted in a training dataset that comprised a stacked vector of all the features in all the time windows. Specifically, a $24 \times D$-dimensional vector where D is the number of features. The second random forest model was simplified to ensure it did not demand prohibitive and expensive compute which was achieved by using the $D = 1000$ most-important features that were identified by the random forest model $M1$. Such that the resultant number of features was $24 \times 1000 = 24000$.

A gradient boosting tree is an ensemble model of weak predictors which work in an additive stage-wise manner to improve on the deficiencies of previous predictors [42,44]. The third scenario, with a gradient boosting tree, entailed a model $M3$ considering the 1000 most-important features from model $M1$. The number of estimators was set to 100 due to a smaller set of features. In the fourth scenario, with a separate gradient boosting tree, a model $M4$ again considered the temporal information which extended the features to 24000, with an increased number of estimators to 1000.

LSTM networks are a type of recurrent neural network (RNN) that are capable of learning long-term dependencies [45]. An RNN is a powerful sequence model that allows previous outputs to be used as inputs at each timestep where its hidden state is updated in order to make a prediction [46]. Rather than considering 1-dimensional data vector for each training and testing example as is considered by the tree-based models, the LSTM network considers temporal information that links data in multiple dimensional vectors. LSTM networks have been applied successfully in various domains with notable impacts in areas that include language modelling, speech-to-text transcription and machine translation [47]. For the LSTM networks' training datasets, the data setup is the same with the tree-based models, except that the input is multiple dimensional vectors depending on the time windows. This resulted in an input to the LSTM networks that took a form of a sequence of 24 D-dimensional data points where D is the number of features. Recall that a *'contact type'* can be regarded as an event that modulates a patient's state as it is a type of intervention where crisis is predicted and measures are already being taken. One can use this information to approximate knowledge that contacts are, at the time of prediction, pre-planned to occur during the two-week-buffer period that precedes the time when a crisis is to be predicted. Note that we do not use the notes associated with these contacts since, at the time of prediction, the notes would not exist. Experiments that included and excluded all contact type information during the one-week-buffer period were therefore carried out in order to evaluate the extent to which knowledge of the pre-planned interventions alters the prediction of crisis. Of the two LSTM networks that were trained, one model, $M5$, used the standard training dataset (which was used for all the tree-based models) whilst the other model, $M6$, considered interventions for comparison. The two LSTM networks were also simplified by using the $D = 1000$ most important features identified by $M1$ as inferred from performance on the first fold of the training set, which is to make them less computationally expensive in order to reduce training time as long as not influencing the performance obviously. The architecture of the LSTM networks are shown in figure 3. To train both models, we used identical parameters: the batch size is 36 and the number of epochs is 10. Both numbers were empirically estimated using a validation set whose size is 5% of the training data. The details of all the six models are summarised in table 1.

The random forests, gradient boosting trees, and LSTM networks that were trained were evaluated by calculating their accuracy, sensitivity, specificity, precision, recall and the area under the curve (AUC) for both receiver operating characteristic (ROC) and precision-recall curves. A 10-fold cross-validation was performed on each model with the training dataset to evaluate its robustness. Each model was further tested using the hold-out test dataset (with no other use of the hold-out dataset) as part of the performance evaluation: the aim was to provide a final performance estimate of the likely performance of the model after its training and validation had been completed. We emphasise that, to avoid a 'double-dip' risk (eg stemming from the validation set for all other folds intersecting with the training set for the first fold and this first fold being used to select the features used by *M1*), all hyper-parameter optimisation only used data in the training set and did not use the hold-out test set.

The scikit-learn and Tensorflow 2.0 software libraries were used for model training and evaluation. Both scikit-learn and Tensorflow python libraries integrate a wide range of state-of-the-art

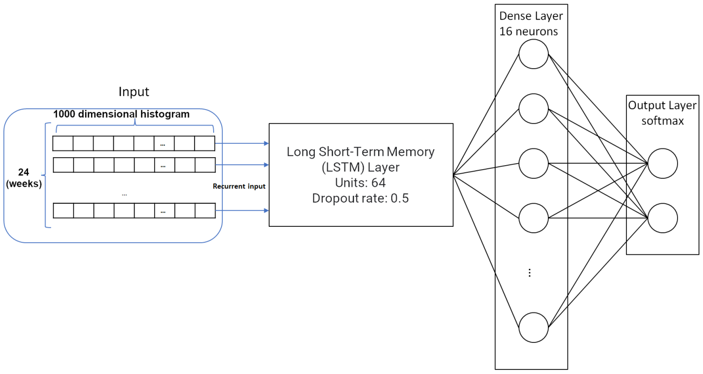


Figure 3: The architectural building blocks of the LSTM networks

Table 1: The setup of the scenarios for the models

Model	**Description**	**Algorithm**	**Number of features**	**Number of estimators/trees**
$M1$	Random forest with standard dataset	Random forest	around 41641	1000
$M2$	Random forest with time windows	Random forest	24000	1000
$M3$	Gradient boosting tree with standard dataset	Gradient boosting tree	1000	100
$M4$	Gradient boosting tree with time windows	Gradient boosting tree	24000	1000
$M5$	LSTM network with time windows	LSTM network	24000	N/A
$M6$	LSTM network with time windows and interventions	LSTM network	24000	N/A

machine learning algorithms [48, 49]. *'Prediction explanations'* were constructed and added as value importance measures, using SHapley Additive exPlanation (SHAP) [50] values and Gini importance [43], to aid the performance evaluation of the different algorithms. SHAP value estimation methods provide an intuitive and unified measure of feature importance [50,51]. In each prediction model, Shapley coefficients were calculated for every data point. Additionally, Gini coefficients were extracted from the random forests. This allowed feature importance to be analysed. Features that corresponded to coefficients with higher absolute values were considered more important. Furthermore, those features were considered to have had greater contribution to their corresponding classification results.

A qualitative audit was carried out by two actively-practising psychiatrists to establish clinical plausibility of the prediction models. A positive prediction was defined as $P(A) > 0.5$, *where* $A = crisis$ and crisis did or did not happen within two weeks. Forty cases with a prediction result from the LSTM network model were selected for the audit using stratified random sampling: 10 each of true-positives, true-negatives, false-positives, and false-negatives. The psychiatrists were asked asked to complete a semi-structured questionnaire for each case. First, they were presented

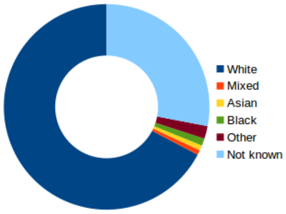


Figure 4: Distribution of recorded ethnicity in the EHR

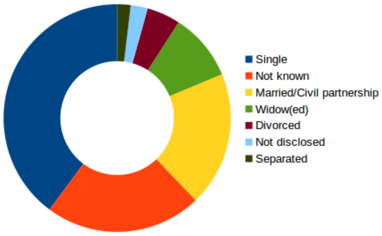


Figure 5: Distribution of marital status categories

with data prior to the trigger date and asked to formulate their clinical view as if they were in clinic, focusing on diagnosis and risk of imminent admission, without the knowledge of the AI prediction output. Thereafter, with knowledge of the LSTM model predictions, and the clinical data for the two weeks following, the psychiatrists compared their recorded views to the model prediction output and what actually transpired.

3.5 Legal and Ethical Considerations

This work subscribes to the *'code of conduct for data-driven health and care technology'* in order to account for ethical challenges associated with the use of data-driven technologies in the NHS and the wider health and care system [52]. In addition, the *'standards for commissioning or developing personal health records'*, that include confidential information and security standards, were applied throughout this work [53]. The use of data and the application of privacy and confidentiality measures were approved by the Mersey Care Trust ethics committee and its patient support group under protocol number 254711.

4 Results

This section presents the key results that were obtained from the study.

4.1 Characteristics of the Ingested EHRs

94605 patients were ingested and processed. Of the ingested patients, 50.8% were female whilst 49.2% were male. There were varying sizes of ethnic and marital status categories amongst patients, including those whose patients chose to not identify themselves or were not known in the EHR as shown in Fig. 4–5. The number of days that the patients were tracked in the EHR was variable with a mean of 1120, a standard deviation of 1190, and a median of 662. Table 2 details the distribution of the duration of the period that individual patients were managed in the EHR system at Mersey Care.

The ingested EHR data included a large amount of unstructured text, where 28.7 million were note entries and 4.4 million were binary documents containing clinical notes. The unstructured text entries that were directly captured in the source EHR system generated 464.6 million and 1 billion worth of MedCAT and BioYODIE annotations respectively. Whereas, the uploaded binary attachments generated 902.7 million and 2.7 billion MedCAT and BioYODIE annotations respectively.

There were 41419 patients that were identified as having depressive disorders. Those patients' medical concept annotations were from 41967 unique medical concepts and 338 unique concept categories of the UMLS. The three most-frequent medical concept annotations from the health record were: *'Contacts'* that occurred in 29650 (71.58%) individuals' notes and belonging to the category *'Health care activity'*, *'Time'* which occurred in 29647 (71.57%) individuals' notes and belonged to the category *'Temporal concept'*, and *'Reviewed'* which occurred in 28333 (68.41%) individuals' notes and belonged to *'Qualitative concept'* category of the UMLS. There were 11579 annotations that occurred in only one individual's notes which was 27.59% of all the annotations. A few random examples of such annotations included: *'Santonin'*, *'Munsee race'* and *'Embezzler'*.

There were 326 kinds of interventions or contact types, in addition to the aforementioned medical concept annotations. The three most-frequent interventions from the health record were: *'Telephone contacts'* which occurred in 23235 (56.09%) individuals' notes, *'Clinic'* which occurred in 13959 (33.70%) individuals' notes, and *'Home visit'* which occurs in 13287 (32.07%) individuals' notes. There were 85 interventions that occurred in only one individual's notes.

A total of 1447 and 39972 records were included in the *'crisis cohort'* and *'no crisis cohort'* respectively. Due to the highly imbalanced sample between the *'crisis cohort'* and *'no crisis cohort'*, a higher weight for the *'crisis cohort'* was configured based on the ratio of the sizes.

Table 2: Length of period individual patients were tracked in the EHR

Duration	**Days**	**Years**
Mean	1120	3.1
Standard deviation	1190	3.3
Minimum	0	0
25%	75	0.2
50%	662	1.8
75%	1903	5.2
Maximum	4107	11.2

4.2 Performance of the Prediction Models

The average ROC curves and precision-recall curves for 10-fold cross validations of the six prediction models with varying kinds of feature arrangement using the *training* dataset are shown in Fig. 6

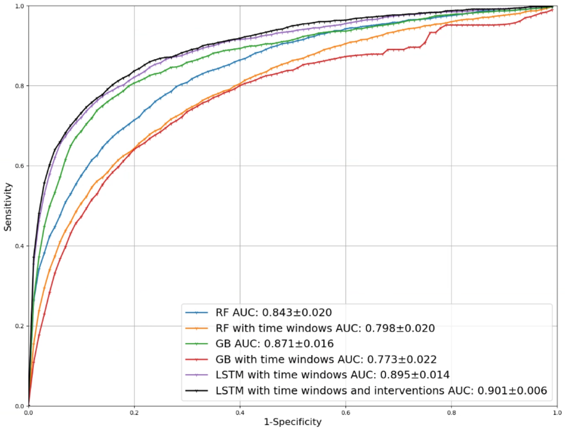


Figure 6: The average ROC curves for 10-fold cross-validations of the prediction models with varying kinds of feature arrangement using the **training** dataset. The Mean and standard deviation of AUC are shown in the legend.

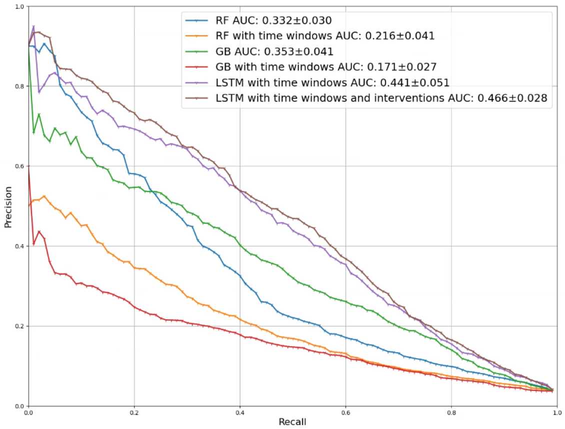


Figure 7: The average precision and recall curves for 10-fold cross-validations of the prediction models with varying kinds of feature arrangement using the **training** dataset. The Mean and standard deviation of AUC are shown in the legend.

and Fig. 7, respectively. Similarly, the average ROC curves and precision-recall curves using the *hold-out test* dataset are shown in Fig. 8 and Fig. 9, respectively. Recall the models summarised in table 1. The crisis prediction by all proposed models and their related features was more effective than chance: the ROC curves were all above the diagonal line as shown in Fig. 6 and Fig. 8. Using the *training* dataset, the random forest model $M1$ performed with an average AUC of 0.843 and a standard deviation 0.02 across the 10 validation folds. Model $M2$, a random forest considering a time window, performed less well with an average AUC of 0.798 and a standard deviation 0.02

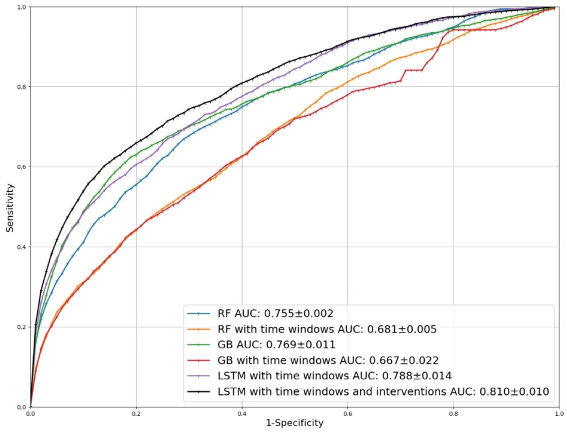


Figure 8: The average ROC curves of the prediction models with varying kinds of feature arrangement using the **hold-out test** dataset. The Mean and standard deviation of AUC are shown in the legend.

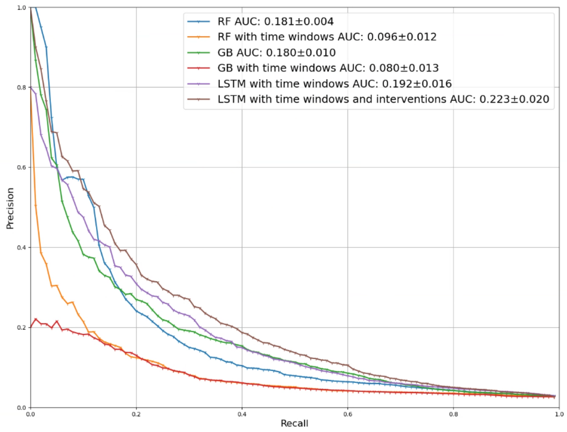


Figure 9: The average precision and recall curves of the prediction models with varying kinds of feature arrangement using the **hold-out test** dataset. The Mean and standard deviation of AUC are shown in the legend.

across the folds. The gradient boosting tree, Model $M3$, showed a large increase in performance, when compared with models $M1$ and $M2$, with an average AUC of 0.871 and a standard deviation of 0.016. Model $M4$, another gradient boosting tree, showed a downgrade in performance with an average AUC of 0.773 and a standard deviation of 0.022. The two LSTM-based models performed the best. When standard features with time windows were considered in model $M5$, the average AUC was 0.895 with standard deviation of 0.014. When the intervention features were included in model $M6$, the average AUC increased to 0.901 with a standard deviation of 0.006.

Table 3: Selected classification metrics when the thresholds were configured to maximise the F1-score from the cross-validation set. For the hold-out test set, the thresholds were those derived from cross-validation set.

Model	**Sensitivity/Recall**		**Specificity**		**Precision**		**F1 score**	
	A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
$M1$	0.316	0.140	0.991	0.993	0.605	0.345	0.415	0.199
$M2$	0.316	0.188	0.975	0.966	0.338	0.129	0.326	0.153
$M3$	0.367	0.441	0.987	0.921	0.549	0.130	0.441	0.200
$M4$	0.382	0.273	0.953	0.940	0.247	0.107	0.301	0.155
$M5$	0.492	0.266	0.981	0.984	0.508	0.304	0.500	0.284
$M6$	0.553	0.293	0.973	0.975	0.444	0.234	0.493	0.260

Dataset descriptions: ***A*** *− Training* dataset; ***B*** *− Hold-out test* dataset;

Table 3 presents some classification metrics that include precision, recall, specificity, and F1 score. When the *training* dataset was used, both LSTM networks had the top-two high F1 scores, 0.5 and 0.49 for models $M5$ and $M6$, respectively. The gradient boosting trees had F1 scores 0.441 and 0.301 for models $M3$ and $M4$, respectively. The F1 scores of the random forests were 0.415 and 0.326, for models $M1$ and $M2$, respectively. When the *hold-out test* dataset was used, the performance difference between the random forests and gradient boosting trees was marginal with F1 scores of 0.199 and 0.2 for models $M1$ and $M3$, respectively. However, the LSTM networks' performance improved by approximately 0.06 when compared to the random forests and gradient boosting trees. In addition, the results showed that there was no advantage in considering interventions when models $M5$ and $M6$ were compared.

With regards to model performance, similar conclusions can be drawn from the precision-recall curves shown in Fig. 7 and Fig. 9. The tree-based algorithms, in models $M1$ to $M4$, showed a downgrade in performance when dimensionality is increased from inclusion of time windows. The LSTM networks performed the best, followed by the gradient boosting trees which were better than the random forests.

When the *hold-out test* dataset was used, it can be observed that the the average AUCs for each of the models was downgraded by approximately 0.1 when compared to using the *training* dataset. We note that this is the case for all models and attribute this to the temporally disjoint nature of the hold-out test set. We also note that the standard deviations of the performance metrics also increased on the gradient boosting trees and LSTM networks relative to their values on the training set. However, the random forests did not show significant evidence of such differences. Nevertheless, the standard deviations were still very small such that they did not change the ranking of the models' performance.

4.3 Feature Importance

To illustrate the role that individual features played in the prediction models as described in Section 3.4.4, a subset of the features restricted to the union of eight most-important features (denoted by letters of the alphabet) for each of the models $M1$ to $M6$ are presented. The extracted features in relation to their importance in predicting a crisis outcome in all the six models are summarised in Table 4. It should be noted that the exact order of the extracted features is arguably less important given that many of the features make a significant contribution to the final prediction outcome. The important features that are shown in Table 4 include those semantic annotations related to engagement with the healthcare system (e.g. 'Hospital Admission'), health conditions (e.g. 'Dementia'), and pharmaceutical products (e.g. 'Bromides').

The ranks of the 19 features shown in Table 4 varied widely across the six models. Nevertheless, most of the features used the prediction models are among the top-100 features across all the models. The feature importance ranking in the LSTM networks $M5$ and $M6$ were similar. However, the ranking in the gradient boosting trees $M3$ and $M4$ were very different. Whereas the ranking in the random forests $M1$ and $M2$, had mild differences as most of their features were amongst the top-100 important features across the two models.

Although models $M1$ and $M3$ are both tree-based, the feature-importance ranking of the two models were different. Of the 19 features listed in Table 4 for model $M1$, all of those features were among the top-100 features across the six prediction models. Whereas, model $M3$ only had 13 of its listed features amongst the top-100 features across the models. Surprisingly, the feature-importance ranking between model $M1$ and model $M5$ were quite similar despite their different model architectures. Such that all of the listed features in model $M5$ were among the top-100 features which is the same to the listed features for model $M1$.

It is also worth noting that feature *G: - Gamma-glutamyl transferase - Enzyme*, an important marker of liver damage, was misclassified as *Guatemala - Geographic Area* in the clinical notes by the NLP service. This was due to a commonly used abbreviation in the clinical notes not having a corresponding synonym entry in the UMLS metathesaurus. This should not be a major problem as the NLP services can be trained iteratively to annotate task-specific language patterns correctly.

4.4 Qualitative Audit

The qualitative clinical assessment showed that depression mostly occurs with other conditions such as schizophrenia, alcohol and substance abuse, and dementia ($n > 50\%$). In the 10 cases who were true positives, the prediction scores had very high values (mean 0.98) and conversely very low values (mean 0.13) for the 10 true negatives. Of the 10 true positives, the clinicians agreed completely with the prediction model output in four, felt the primary diagnosis was not depression for two, and in four agreed with the depression diagnosis but did not think a crisis was imminent (i.e. the prediction model was suggesting they might have missed something). Of the 20 false positive and negatives the prediction model output were less clear (means of 0.73 and 0.25) and there was a wider range of diagnoses. Only two of these had a strong prediction of crisis ($P(A) > 0.9$, where $A = crisis$): one was a person with severe learning disability (different care options) and the other a man in prison being attended by prison psychiatrists. There were no cases where an alert would have been deleterious to care.

5 Discussion

The prediction models use annotated EHR data of a particular patient from any six-month period. The patient record data was used to predict the probability of a patient experiencing crisis in the immediate weeks following the selected six-month period. The ROC curves for the 10-fold cross-validation of all the prediction models show that the models can predict, with good performance,

Table 4: The union of 6 most-important features that were summarised by models 1 - 6 and their ranks in the corrsponding model

Model	**Textual Features**																		
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*
$M1$	n/a	9	18	5	14	3	1	83	15	2	16	8	12	4	6	10	54	13	27
$M2$	n/a	2	12	7	14	24	1	62	5	4	6	3	16	34	15	48	112	9	82
$M3$	n/a	45	185	208	10	1	3	125	91	12	66	321	25	2	575	4	6	56	5
$M4$	n/a	13	30	4	37	3	1	23	5	10	8	11	6	2	95	38	25	18	12
$M5$	n/a	25	4	22	9	2	7	5	17	3	15	6	14	24	18	29	27	1	73
$M6$	1	2	3	4	5	6	9	11	12	13	16	19	22	28	29	39	41	57	123

Textual feature descriptions:
A: Telephone Contact (intervention); **B:** Bromides - Inorganic Chemical; **C:** Mental health - Mental Process;
D: Health - Idea or Concept; **E:** Agree - Finding; **F:** Hospital Admission - Health Care Activity;
G: Gamma-glutamyl transferase - Enzyme; **H:** International Unit - Quantitative Convept;
I: Contacts - Health Care Activity; **J:** Structure of left thigh - Body Part, Organ, or Organ Component;
K: Psyche structure - Mental Process; **L:** liter - Quantitative Concept; **M:** Alchohols - Pharmacologic Substance;
N: Detox - Immunologic Factor; **O:** Greater Than - Quantitative Concept; **P:** Coordinator - Professional or Occupational Group;
Q: Memory - Mental Process; **R:** Attending (action) - Functional Concept; **S:** Dementia - Mental or Behavioral Dysfunction;
Notes: Rank entry *"n/a"* in the table means that the feature is not considered in the model.
Top 6 features are highlighted in red.

the likelihood of a patient experiencing crisis related to depression. Specifically, the mean AUCs, which all have a relatively small standard deviation, show good performance and robustness across the mainstream predictive models.

Although the current reduced feature set of an average of 41641 across the folds for the first random forest $M1$ performs as well as it does, it is in no way exhaustive or sufficient. Furthermore, it is in no way an optimal set as this feature set was simply comprised of features that had a decrease in Gini index by having a zero score in importance. Therefore, further work is required to work closely with a variety of clinical workers to derive an optimal set of features. This would involve the explaining, refining and reduction of features for a pragmatic and usable prediction model that can be easily integrated into a point-of-care clinical decision support tool.

But it can be observed that the LSTM networks were particularly effective at improving prediction performance when the specificity was low. Whereas when the specificity was high, the performance of model $M3$, a gradient boosting tree, improved and was marginally better than the LSTM networks at very high sensitivity. This could be attributed to the difference in feature arrangement of the two algorithms. However, the overall performance of the LSTM networks was better than that of the gradient boosting trees as evidence on both ROC curves and precision-recall curves.

It should be noted that using the temporal-indexed features with the random forest $M2$ did not impact the AUC positively. This can be observed in the average AUC decrease of 0.045 in model $M2$ when compared to $M1$. The reason that the performance was inferior can be attributed to the following factors: 1) the time-indexed features do not typically offer an edge in performance to tree-based algorithms, 2) random forests can hardly deal with very high dimensional features, and

3) only 1000-most important features were considered in model $M2$. Similarly, models $M3$ and $M4$ showed a similar kind of performance as they were tree-based models that considered 1000-most important features.

The results further indicate that the LSTM network's improvement in performance comes from the LSTM algorithm and not only benefit from the temporal-indexed features. Considering the pre-planned interventions that occurred in the time window resulted in a small further increase in the performance of the LSTM with an average AUC increase by approximately 0.006. This modest improvement implies that the LSTM is able to anticipate the impact of the pre-planned interventions. This could be caused by how pre-planned interventions and their related motivating observations are stored in the EHR. It is noted that the pre-planned interventions, as features, are mostly among the top 1000 important features and they are effective on reducing the standard deviation associated with the AUC.

Testing the models using the *hold-out test* dataset does not affect their relative performance significantly. The LSTM networks still yield the best results. Although there is a noticeable improvement in performance when gradient boosting trees are used in place of random forests, the performance is still under par. The overall performance of the models when tested using the *hold-out test* dataset is downgraded when compared to the tests carried out using the *training* dataset. This result is not surprising as implicit patterns in data set can change over time, giving room for new and emerging features.

The small qualitative assessment, comparing clinical professionals to the model prediction output, showed that when the prediction score was high ($P(A) \geq 0.9$, $where\ A = crisis$) and there was a depressive diagnosis, an admission did follow. Importantly in half of this small sample, the clinician using the same data did not anticipate this outcome. Clinicians observed that if the alert had been offered to the clinician as they saw the patient, it may have caused clinicians to reconsider what actions might be appropriate to mitigate that risk. There were two false-negative cases with high prediction scores – and both were unusual and required care on different clinical pathways that were already in place. Noting that the electronic health record was filtered to the depression pathways, future iterations of the digital health solution will have more clinical data to assist in the management of other mental health conditions. Nevertheless, if the present solution was used to alert clinicians when there is prediction score was high ($P(A) \geq 0.9$, $where\ A = crisis$), it could be a useful aide for the busy short-of-time clinician and there no cases where that alert could have been deleterious. This is not to criticise the clinicians but rather to recognise how difficult it can be to assimilate the full detail of a person's record in the few minutes at the start of a consultation. If such alerts led the clinician to rethink the management plan for even one case per week, that could have a significant contribution to more efficient care. And it should be noted that there would be no downside to the care pathway since the clinician can always overrule any suggestions if they think it appropriate.

The CogStack component of the AI environment provides a flexible framework for adding Extract-Transform-Load (ETL) services to manipulate EHR data. This allows a variety of text analytics tools and services, such as NLP applications, to be easily added or removed in order to support scaling, extension, and refinement of functionality. Although the current NLP pipeline in the AI environment can successfully generate semantic annotations from both MedCAT and BioYODIE NLP applications, the current prediction model is based on a feature set from structured EHR data and semantic annotations from the MedCAT application only. The MedCAT and BioYODIE applications can perform differently or provide diferent contextual information, such as negation and temporality, given an identical source text data. Hence, additional work is required to create a prediction model that uses BioYODIE semantic annotations for its feature set. This will allow for a thorough and comprehensive comparison, between MedCAT- and BioYODIE-based prediction models, to be performed with the aim of using the best performing of the two methods for clinical decision support.

The current prediction models use random forest algorithms, gradient boosting trees, or LSTM networks, which are supervised machine learning approaches that have been widely used with good performance. Noting the increased number of features and the complexity of the EHR data, alternative machine learning approaches that are based on supervised and/or unsupervised machine learning methods can be applied and compared with each other in order to select the best performing prediction model. Other approaches such as deep learning can provide state-of-the-art and sophisticated training methods, where in some applications, an unsupervised approach can be used to extract the most relevant features, and then use them for classification by exploiting a supervised learning step with dramatic results [35, 54]. The extreme complexity and sparse observation of mental, and common comorbidities, means that EHR data will contain only limited amounts of predictive information, therefore external statistical structure from prior knowledge needs careful incorporation, where causal, counterfactual machine learning advances need to be considered [55]. Therefore, further work should be carried out to compare prediction performance from the various learning approaches.

6 Conclusion

This work presents a state-of-the-art framework and a reference implementation of a digital health environment that can enable text and predictive analytics. A data pipeline with connectors for SSAs was implemented, providing an extensible mechanism for adding multiple EHR sources to the environment. The pipeline was further extended with NLP services that included text extraction and semantic annotation applications. Mental health records covering a period of 11 years were ingested into the AI environment and annotated using the UMLS corpus. Thereafter, two random forest models, two gradient boosting trees, and two LSTM networks were trained on the EHR data and its related semantic annotations. The results demonstrate that it is feasible to ingest a large mental health record, apply text analytics operations, and build a predictive model in a secure environment with good performance. This work will be extended to support optimisation the current prediction models, and their evaluation against other models that are based on other approaches such as deep learning methods. Furthermore, the best performing model, with the help of a variety of clinical workers, will be further refined and integrated into a usable point-of-care clinical decision support tool for deployment and validation in a real clinical setting with a view to provide a 'blind test' of the algorithms.

References

[1] D. M. Howard, M. J. Adams, M. Shirali, T.-K. Clarke, R. E. Marioni, G. Davies, J. R. Coleman, C. Alloza, X. Shen, M. C. Barbu *et al.*, "Genome-wide association study of depression phenotypes in uk biobank identifies variants in excitatory synaptic pathways," *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[2] A. Brailean, J. Curtis, K. Davis, A. Dregan, and M. Hotopf, “Characteristics, comorbidities, and correlates of atypical depression: evidence from the uk biobank mental health survey,” *Psychological medicine*, vol. 50, no. 7, pp. 1129–1138, 2020.

[3] D. M. Howard, L. Folkersen, J. R. Coleman, M. J. Adams, K. Glanville, T. Werge, S. P. Hagenaars, B. Han, D. Porteous, A. Campbell *et al.*, “Genetic stratification of depression in uk biobank,” *Translational Psychiatry*, vol. 10, no. 1, pp. 1–8, 2020.

[4] H. Castelijns, V. Eijsbroek, A. T. Cees, H. W. van Marwijk, C. M. van der Feltz-Cornelis *et al.*, “Illness burden and physical outcomes associated with collaborative care in patients with comorbid depressive disorder in chronic medical conditions: A systematic review and meta-analysis,” *General Hospital Psychiatry*, vol. 50, pp. 1–14, 2018.

[5] Full Fact, “Mental health spending in the english nhs,” 2019. [Online]. Available: https://fullfact.org/health/mental-health-spending-england/

[6] C. M. Gillan and N. D. Daw, “Taking psychiatry research online,” *Neuron*, vol. 91, no. 1, pp. 19–23, 2016.

[7] R. H. McAllister-Williams, D. Cousins, and B. Lunn, “Clinical assessment and investigation in psychiatry,” *Medicine*, vol. 44, no. 11, pp. 630–637, 2016.

[8] A. Takian, A. Sheikh, and N. Barber, “We are bitter, but we are better off: case study of the implementation of an electronic health record system into a mental health hospital in england,” *BMC health services research*, vol. 12, no. 1, pp. 1–13, 2012.

[9] F. Röhricht, G. K. Waddon, P. Binfield, R. England, R. Fradgley, L. Hertel, P. James, J. Littlejohns, D. Maher, and M. Oppong, “Implementation of a novel primary care pathway for patients with severe and enduring mental illness,” *BJPsych bulletin*, vol. 41, no. 6, pp. 314–319, 2017.

[10] G. E. Iyawa, M. Herselman, and A. Botha, “Digital health innovation ecosystems: From systematic literature review to conceptual framework,” *Procedia computer science*, vol. 100, no. C, pp. 244–252, 2016.

[11] R. Patel, T. Lloyd, R. Jackson, M. Ball, H. Shetty, M. Broadbent, J. Geddes, R. Stewart, P. Mcguire, and M. Taylor, “Mood instability and clinical outcomes in mental health disorders: A natural language processing (nlp) study,” *European psychiatry*, vol. 33, no. sS, pp. S224–S224, 2016.

[12] L. Vogel, “AI opens new frontier for suicide prevention.(news: Mental health)(artificial intelligence)(report),” *CMAJ: Canadian Medical Association Journal*, vol. 190, no. 4, p. E119, 2018.

[13] P. C. Webster, “The rise of open-source electronic health records,” *The lancet*, vol. 377, no. 9778, pp. 1641–1642, 2011.

[14] J. G. Shull, “Digital health and the state of interoperable electronic health records,” *JMIR medical informatics*, vol. 7, no. 4, p. e12712, 2019.

[15] H. Alderwick and J. Dixon, “The nhs long term plan,” 2019.

[16] S. Asthana, R. Jones, and R. Sheaff, “Why does the nhs struggle to adopt ehealth innovations? a review of macro, meso and micro factors,” *BMC Health Services Research*, vol. 19, no. 1, pp. 1–7, 2019.

[17] C. Hollis, R. Morriss, J. Martin, S. Amani, R. Cotton, M. Denis, and S. Lewis, “Technological innovations in mental healthcare: harnessing the digital revolution,” *The British Journal of Psychiatry*, vol. 206, no. 4, pp. 263–265, 2015.

[18] M. Honeyman, P. Dunn, and H. McKenna, “A digital nhs,” *An introduction to the digital agenda and plans for implementation. London: Kings Fund*, 2016.

[19] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G.-Z. Yang, “Big data for health,” *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, pp. 1193–1208, 2015.

[20] B. Fonferko-Shadrach, A. S. Lacey, A. Roberts, A. Akbari, S. Thompson, D. V. Ford, R. A. Lyons, M. I. Rees, and W. O. Pickrell, “Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the exect (extraction of epilepsy clinical text) system,” *BMJ open*, vol. 9, no. 4, p. e023232, 2019.

[21] H. C. Tissot, A. D. Shah, D. Brealey, S. Harris, R. Agbakoba, A. Folarin, L. Romao, L. Roguski, R. Dobson, and F. W. Asselbergs, “Natural language processing for mimicking clinical trial recruitment in critical care: a semi-automated simulation based on the leopards trial,” *IEEE Journal of Biomedical and Health Informatics*, 2020.

[22] E. Scheurwegs, K. Luyckx, L. Luyten, W. Daelemans, and T. Van den Bulcke, “Data integration of structured and unstructured sources for assigning clinical codes to patient stays,” *Journal of the American Medical Informatics Association*, vol. 23, no. e1, pp. e11–e19, 2016.

[23] H. Wu, G. Toti, K. I. Morley, Z. M. Ibrahim, A. Folarin, R. Jackson, I. Kartoglu, A. Agrawal, C. Stringer, D. Gale *et al.*, “Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research,” *Journal of the American Medical Informatics Association*, vol. 25, no. 5, pp. 530–537, 2018.

[24] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet *et al.*, “Machine learning and natural language processing in mental health: systematic review,” *Journal of Medical Internet Research*, vol. 23, no. 5, p. e15708, 2021.

[25] L. Weber, J. Münchmeyer, T. Rocktäschel, M. Habibi, and U. Leser, “Huner: improving biomedical ner with pretraining,” *Bioinformatics*, vol. 36, no. 1, pp. 295–302, 2020.

[26] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts *et al.*, “Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit,” *Artificial intelligence in medicine*, vol. 117, p. 102083, 2021.

[27] O. Bodenreider, “The unified medical language system (umls): integrating biomedical terminology,” *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[28] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[29] G. Gorrell, X. Song, and A. Roberts, "Bio-yodie: A named entity linking system for biomedical text," *arXiv preprint arXiv:1811.04860*, 2018.

[30] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[31] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon, "A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS).* IEEE, 2019, pp. 338–343.

[32] M. Neumann, D. King, I. Beltagy, and W. Ammar, "Scispacy: fast and robust models for biomedical natural language processing," *arXiv preprint arXiv:1902.07669*, 2019.

[33] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W541–W545, 2011.

[34] R. Garriga, J. Mas, S. Abraha, J. Nolan, O. Harrison, G. Tadros, and A. Matic, "Machine learning model to predict mental health crises from electronic health records," *Nature medicine*, vol. 28, no. 6, pp. 1240–1248, 2022.

[35] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning.* MIT press Cambridge, 2016, vol. 1.

[36] S. de Lusignan, "Effective pseudonymisation and explicit statements of public interest to ensure the benefits of sharing health data for research, quality improvement and health service management outweigh the risks." *Inform Prim Care*, vol. 21, no. 2, pp. 61–63, 2014.

[37] R. Jackson, I. Kartoglu, C. Stringer, G. Gorrell, A. Roberts, X. Song, H. Wu, A. Agrawal, K. Lui, T. Groza *et al.*, "Cogstack-experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital," *BMC medical informatics and decision making*, vol. 18, no. 1, p. 47, 2018.

[38] F. Paton, K. Wright, N. Ayre, C. Dare, S. Johnson, B. Lloyd-Evans, A. Simpson, M. Webber, and N. Meader, "Improving outcomes for people in mental health crisis: a rapid synthesis of the evidence for available models of care," *Health Technologyl Assessment*, vol. 20, no. 3, 2016.

[39] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[40] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[41] D. Paul, R. Su, M. Romain, V. Sébastien, V. Pierre, and G. Isabelle, “Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier,” *Computerized Medical Imaging and Graphics*, vol. 60, pp. 42–49, 2017.

[42] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[43] S. Nembrini, I. R. König, and M. N. Wright, “The revival of the gini importance?” *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018.

[44] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[45] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[46] I. Sutskever, J. Martens, and G. E. Hinton, “Generating text with recurrent neural networks,” in *ICML*, 2011.

[47] A. Sherstinsky, “Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[49] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: A system for large-scale machine learning,” in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[50] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[51] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable ai for trees,” *Nature machine intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[52] Department of Health and Social Care, “Code of conduct for data-driven health and care technology,” 2018. [Online]. Available: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology

[53] NHS Digital, “Standards for commissioning or developing personal health records,” 2020. [Online]. Available: https://digital.nhs.uk/services/personal-health-records-adoption-service/personal-health-records-adoption-toolkit/developing-a-personal-health-record/standards-for-developing-personal-health-records

[54] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, “Deep learning for health informatics,” *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2016.

[55] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, “Causal inference and counterfactual prediction in machine learning for actionable healthcare,” *Nature Machine Intelligence*, vol. 2, no. 7, pp. 369–375, 2020.