*DOI:*
10.3791/60794

*Document Version*
Peer reviewed version

**TITLE:**

Implementation of a Real-Time Psychosis Risk Detection and Alerting System Based on Electronic Health Records using CogStack

**AUTHORS AND AFFILIATIONS:**

Tao Wang[1], Dominic Oliver[2], Yamiko Msosa[1], Craig Colling[3], Giulia Spada[2], Łukasz Roguski[4], Amos Folarin[1], Robert Stewart[3,5], Angus Roberts[1,3], Richard J.B. Dobson[1,3,4,6], Paolo Fusar-Poli[2,3,7,8]

[1]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[2]Early Psychosis: Interventions and Clinical-detection (EPIC) lab, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK
[3]National Institute for Health Research, Maudsley Biomedical Research Centre, South London and Maudsley National Health Service (NHS) Foundation Trust, London, UK
[4]Institute of Health Informatics, University College London, London, UK
[5]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[6]Health Data Research UK London, University College London, London, UK
[7]OASIS service, South London and Maudsley NHS Foundation Trust, London, UK
[8]Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy

tao.wang@kcl.ac.uk
dominic.a.oliver@kcl.ac.uk
yamiko.msosa@kcl.ac.uk
Craig.Colling@slam.nhs.uk
giulia.spada@kcl.ac.uk
l.roguski@ucl.ac.uk
amos.folarin@kcl.ac.uk
robert.stewart@kcl.ac.uk
angus.roberts@kcl.ac.uk
richard.j.dobson@kcl.ac.uk
paolo.fusar-poli@kcl.ac.uk

**KEYWORDS:**

Psychosis, psychosis risk calculation, electronic health records, real-time alerting, health informatics, clinical decision support, prevention

**SUMMARY:**

We demonstrate how to deploy a real-time psychosis risk calculation and alerting system based on CogStack, an information retrieval and extraction platform for electronic health records.

**ABSTRACT:**

Recent studies have shown that an automated, lifespan-inclusive, transdiagnostic, and clinically based, individualized risk calculator provides a powerful system for supporting the early

detection of individuals at-risk of psychosis at a large scale, by leveraging electronic health records (EHRs). This risk calculator has been externally validated twice and is undergoing feasibility testing for clinical implementation. Integration of this risk calculator in clinical routine should be facilitated by prospective feasibility studies, which are required to address pragmatic challenges, such as missing data, and the usability of this risk calculator in a real-world and routine clinical setting. Here, we present an approach for a prospective implementation of a real-time psychosis risk detection and alerting service in a real-world EHR system. This method leverages the CogStack platform, which is an open-source, lightweight, and distributed information retrieval and text extraction system. The CogStack platform incorporates a set of services that allow for full-text search of clinical data, lifespan-inclusive, real-time calculation of psychosis risk, early risk-alerting to clinicians, and the visual monitoring of patients over time. Our method includes: 1) ingestion and synchronization of data from multiple sources into the CogStack platform, 2) implementation of a risk calculator, whose algorithm was previously developed and validated, for timely computation of a patient's risk of psychosis, 3) creation of interactive visualizations and dashboards to monitor patients' health status over time, and 4) building automated alerting systems to ensure that clinicians are notified of patients at-risk, so that appropriate actions can be pursued. This is the first ever study that has developed and implemented a similar detection and alerting system in clinical routine for early detection of psychosis.

**INTRODUCTION:**

Psychotic disorders are serious mental health illnesses that lead to difficulties in distinguishing between the internal experience of the mind and the external reality of the environment[1], as well as a higher than average risk of self-harm and suicide[2]. Under standard care, these disorders result in major public health impact with a significant health and economic burden on individuals, families and societies worldwide[3]. Early interventions in psychosis can improve outcomes of this mental disorder[4]. In particular, detection, prognostic assessment and preventive treatment of individuals who are at clinical high risk for developing psychosis (CHR-P)[5] provides a unique potential to alter the course of the disorder, thereby improving the quality of life for many people and their families[3,6,3]. CHR-P individuals are help-seeking young people presenting with attenuated symptoms and functional impairment[7]: their risk of developing psychosis is 20% at 2-years[8] but it is higher in some specific subgroups[9,10]. Despite some substantial advancements, the impact of preventive approaches in routine clinical practice is limited by the ability to detect most individuals who are at-risk[11]. Current detection methods are based on help-seeking behaviors and referrals on suspicion of psychosis risk; these methods are highly inefficient in handling a large number of samples[11]. Thus, the scalability of current detection methods to the vast majority of the at-risk population is quite limited[12]. In fact, only 5% (standalone specialized early detection services) to 12% (youth mental health services) of individuals at-risk of developing a first psychotic disorder can be detected at the time of their at-risk stage by the current detection strategies[6].

To extend the clinical benefits of the preventive approaches in a larger number of at-risk individuals, we developed an automated, lifespan-inclusive (i.e., across all ages), transdiagnostic (i.e., across different diagnoses)[13], clinically-based individualized risk calculator, which can detect

individuals at-risk of psychosis in secondary mental health care at scale, beyond those meeting CHR-P criteria[14]. This risk calculator used a Cox proportional hazard model to predict the risk of developing a psychotic disorder over six years from five routinely collected clinical variables selected a priori, in line with methodological guidelines[15]: age, gender, ethnicity, age-by-gender and primary index diagnosis. These clinical variables were selected based on a priori knowledge obtained from meta analyses[16,17], as recommended by the state-of-the-art methodological guidelines[15]. The number of predictors is limited to preserve the Event Per Variable ratio and minimize overfitting biases; including too many variables without a priori filter leads to overfitting problems and poor prognostic accuracy[18]. The method used to develop this model provides similar prognostic accuracy to automatic machine learning methods[18]. Parameters of the Cox model were estimated based on a retrospective de-identified cohort from the South London and Maudsley National Health Service Foundation Trust (SLaM)[19]. SLaM is a National Health Service (NHS) mental health trust that provides secondary mental health care to a population of 1.36 million individuals in South London (Lambeth, Southwark, Lewisham and Croydon boroughs), and has one of the highest recorded rates of psychosis in the world[20]. All data used in the model development were extracted from the Clinical Record Interactive Search (CRIS) platform, a digital case register system, which provides researchers with retrospective access and analysis of anonymized clinical records[19]. The clinical information in CRIS is extracted from a bespoke Electronic Health Record (EHR) system, at SLaM, called electronic Patient Journey System (ePJS). SLaM is paper-free and ePJS represents the standard data collection platform for clinical routine. Thus, the transdiagnostic risk calculator leverages EHRs and has the potential to automatically screen large EHRs of patients accessing secondary mental healthcare, to detect those who may be at-risk of psychosis. The algorithm of this transdiagnostic risk calculator has been published previously[6,14,21]. The transdiagnostic risk calculator has been externally validated in two NHS Foundation Trusts[14,21] and optimized[22], demonstrating its adequate prognostic performance and generalizability across different populations.

According to methodological guidelines on the development of a risk prediction model[15,23], the next step after model development and validation is to implement the prediction model in routine clinical practice. Implementation studies are usually preceded by pilot or feasibility studies that address potential pragmatic limitations associated with the use of risk algorithms in clinical practice. For example, required data for running a calculator, such as age, gender and ethnicity, may not be available at the date of diagnosis or updated later. Effective methods for handling missing data and synchronizing frequent updates in real-time data streams should be considered to obtain the most reliable prediction results in an implementation. Furthermore, since the initial development of the risk calculator was based on retrospective cohort data, it is not known whether it can be used in a real-time data stream that is typical of a real-world clinical setting. Another challenge is ensuring that relevant clinicians receive the recommendations generated by the risk calculator within an appropriate time frame and within a shared and accepted communication pathway.

To overcome these limitations, we have completed a feasibility implementation study employing the individualized transdiagnostic risk calculator. The study included two phases: an in vitro phase that was conducted using data from the local EHR, without contacting clinicians or patients, and

an in vivo phase, which involved direct contact with clinicians. The in vitro phase had two manifold aims: (i) to address implementation barriers according to the Consolidated Framework for Implementation Research (CFIR)[27] and (ii) to integrate the transdiagnostic risk calculator into the local EHR. Implementation barriers included the communication of risk outcomes to clinicians. In SLaM, all patients are invited to register for Consent for Contact (C4C), which indicates their willingness to be contacted for research, without affecting the quality of care. This reduces the ethical issues relating to contacting patients. Further to this, working groups with clinicians aided tailoring of how this information was communicated. During the in vivo phase (May 14th 2018 to April 29th 2019), all individuals (i) older than 14 years (ii) who were accessing any SLaM service (boroughs of Lambeth, Southwark, Lewisham, Croydon), (iii) receiving a first ICD-10 index primary diagnosis of any non-organic, non-psychotic mental disorder (with the exception of Acute and Transient Psychotic Disorders; ATPD), or a CHR-P designation and (iv) with existing contact details were deemed eligible. During the in vivo phase, new patients accessing SLaM each week were automatically screened for their psychosis risk, and those with having a risk greater than a certain threshold were detected. The research team then contacted the patients' responsible clinicians to discuss further recommendations and eventually suggest a further face to face assessment[6]. If those assessed were considered to meet CHR-P criteria, they were referred to specialist CHR-P services, such as Outreach and Support in South London (OASIS)[28]. This would result in improved detection of individuals prior to the onset of a psychotic disorder and provide a significant opportunity for altering the course of the disorder. Crucially, this feasibility study involved the full integration of the calculator into the local EHR system, which is the topic of the current article. The full protocol of this feasibility study, including an overview of the plan for evaluating the proposed research, details on managing data security and ethical issues, has been presented in our previous work[6]. The current article, as a part of the feasibility study[6], selectively focuses on presenting the technical implementation of a real-time psychosis risk detection and alerting system based on the local EHR data. More specifically, the aim of this study is to investigate the technical feasibility of this risk calculator in timely detecting at-risk patients as soon as they access a secondary mental healthcare service. The full results of the feasibility study, in terms of clinicians' adherence to the recommendations made by the risk calculator, will be presented separately. A comprehensive evaluation of the effectiveness of the proposed research, which requires randomized designs, is outside the scope of the current research program. To our best knowledge, this is the first method describing the implementation of a risk calculator based on live EHR data for early detection of psychosis.

Our approach to psychosis risk detection and alerting takes advantage of the CogStack platform. The CogStack platform is a lightweight, distributed, and fault-tolerant information retrieval and text-extraction platform[24]. This platform consists of three key components: 1) the CogStack Pipeline that uses the Java Spring Batch framework to ingest and synchronize data from a pre-defined data source (both structured and unstructured EHR data in multiple formats such as Word, PDF files and images) to a predefined data sink in real time; 2) Elasticsearch, a search engine allowing for storage and querying of the full text of EHR data, as well as providing various application programming interfaces (APIs) to embed advanced analytics into the engine; and 3) Kibana, an interactive, web-based user interface that allows users to query data in Elasticsearch, build visualization dashboards and set alerting on anomalies or other patterns of interest from

data. Moreover, CogStack incorporates the ability to alert clinicians to potential problems by Email and SMS (text), allowing clinicians to receive timely notifications about at-risk patients reported by the risk calculator.

We present a model of psychosis risk detection and alerting based on ePJS at SLaM, leveraging the CogStack platform. Compared with the CRIS platform that provides a mechanism for retrospective access to de-identified health records from ePJS on a weekly basis[19], the CogStack platform at SLaM enables access to an identifiable EHR in real time, bringing the alerting closer to the point-of-care and the risk prediction in a prospective design, although both the CRIS and CogStack platforms use data sourced from ePJS in SLaM. In the section that follows, we provide details of the key steps in our approach, including preparing source data from the EHR, ingesting the source data into the CogStack platform to enable full-text search via Elasticsearch, running the psychosis risk calculator using a Python daemon thread, and setting interactive visualizations and real-time risk alerting via the Kibana user interface. Any researcher who aims to build a real-time risk detection and alerting system based on EHR data can follow the approach and its reference implementation. As we shall elaborate below, the proposed method exploits open-source, lightweight techniques with high flexibility and portability. This enables the risk calculator to be run in various locations and shows a high applicability to other risk estimation algorithms. Moreover, the method works as a straightforward approach to enhance the risk detection and alerting functionalities of an EHR embedded in a general healthcare system.

**PROTOCOL:**

This study was approved by East of England - Cambridgeshire and Hertfordshire Research Ethics Committee (Reference number: 18/EE/0066).

NOTE: We have developed this protocol based on the CogStack platform and the Python programming language. This system requires Docker (more specifically, Docker Compose https://docs.docker.com/compose/), Anaconda Python (https://www.anaconda.com/distribution/) and Git (https://git-scm.com/downloads) pre-installed on a device. The commands provided in this protocol are based on the Linux environment. In the following, we provide the details of preparing source data from an EHR database, ingesting the data to CogStack platform, and setting up a real-time risk calculation and alerting system for psychosis based on the CogStack platform. Moreover, an online version of the risk calculator was developed to facilitate numeric calculation of the probability of an individual developing psychosis in secondary mental health care on http://www.psychosis-risk.net.

**1. Source data preparation**

NOTE: In most use cases, CogStack ingests source data from a specified database view that can combine data from one or more source database tables, where a view is a searchable object in a database that contains the result set of a stored query on the data. The setup of the ingesting view is tailored by the specific use cases and deployment settings of a health record database system. This protocol is developed based on a psychosis risk calculator developed and externally

validated twice by Fusar-Poli et al.[14,21] and as part of a pilot implementation feasibility study[6]. The protocol is based on an EHR database deployed with Microsoft SQL Server 2014.

1.1 Create a view object (called "*vwPsychosisBase*" in this protocol) in an existing EHR database system to join necessary information of patients for psychosis risk calculation and alerting. Make sure that this view includes all patients receiving a first primary diagnosis of non-organic and non-psychotic mental disorder (recorded by the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision [ICD-10]), as defined in the original model[14,21].

1.2 Ensure that each record in the view involves three types of patient information: 1) the first primary diagnosis of a patient in the EHR system, including ICD-10 diagnosis index (diagnoses were clustered together into the following ten clusters: acute and transient psychotic disorders, anxiety disorders, bipolar mood disorders, childhood and adolescent onset disorders, developmental disorders, nonbipolar mood disorders, mental retardation, personality disorders, physiological syndromes, substance use disorders) and diagnosis date; 2) a patient's demographic data, including gender, ethnicity and date of birth; and 3) the most recent contact information of care team for a patient, such as details of general practice (GP), consultants and care coordinators. The first two types of information are vital for the psychosis risk calculator[14,21], and the third type of information is to enable timely risk alerting.

1.3 Make sure that each record in the view has a unique identifier (e.g., "*patient_id*" used in this protocol).

1.4 Select the last update timestamps of all source information related to a record in the view (e.g., the last update times of a patient's demographic information and the patient's first primary diagnosis information), and choose the latest timestamp as the last update date and time for the record in the view (denoted as "*etl_updated_dttm*" in this protocol). The last update date and time of a record allows CogStack to synchronize updates in the database, such as new and updated records.

## 2. Data ingestion

2.1 Download or clone the code repository from Github (https://github.com/cogstack-slam/psychosis) or by typing "git clone https://github.com/cogstack-slam/psychosis.git" in a terminal window. The downloaded folder contains the code for psychosis risk calculation and configuration files for deploying a CogStack instance.

2.2 Go to the "*cogstack_deploy/cogstack/*" directory and modify "*psychosis.properties*" to configure CogStack Pipeline for data ingestion. Modify the settings of section "*SOURCE: DB CONFIGURATIONS*" based on the EHR database setup, including specifying the IP address of the database server, database name, database username and password. Modify the view name (i.e. "*vwPsychosisBase*") and field names (e.g., "*patient_id*" and "*etl_updated_dttm*") if necessary. In case of error in configuring this file, follow the instructions at https://cogstack.atlassian.net/wiki/spaces/COGDOC/pages/38043684/Quickstart.

2.3 Go to the "*cogstack_deploy/common/elasticsearch/config/*" directory and modify the section "*xpack.notification.email.account*" in the "*elasticsearch.yml*" file to configure an Email address for sending alerts. A detailed instruction for Email configuration can be found on https://www.elastic.co/guide/en/kibana/6.4/watcher-create-threshold-alert.html.

2.4 Go to the "*cogstack_deploy/*" directory and type "*docker-compose up*" to run the CogStack platform. Execute this command with root access. If the process is completed successfully, there will be printed status logs of the currently running services, including CogStack Pipeline, Elasticsearch and Kibana, in the terminal. As a result, all data and updates in the source database view will be timely ingested to an Elasticsearch index called "*psychosis_base*" in the CogStack platform.

2.5 Open a web browser and access Kibana user interface by typing "http://localhost:5601/" (or replacing "*localhost*" with a specific IP address of the server running the CogStack platform). For the first time accessing Kibana, click the **Management** tab and **Index Patterns** tab to specify an Elasticsearch index that one wants to access with Kinaba. Type "*psychosis_base*" in the "*Index pattern*" field and click **Next step**. Select "*etl_updated_dttm*" for the "*Time Filter*" field name and click **Create index pattern** to add the "*psychosis_base*" index pattern for Kinana.

2.6 Once Kibana is connected to the Elasticsearch index (i.e., "*psychosis_base*"), search and browse the source data interactively through the "*Discover*" page. Kibana allows non-technical users to search for both structured metadata and free text. Detailed instructions of using "*Discover*" are available on https://www.elastic.co/guide/en/kibana/6.4/discover.html.

**3. Risk calculation**

3.1 Open a new terminal window and go to the "*psychosis/*" directory. Install all required Python packages (including "*elasticsearch*", "*elasticsearch_dsl*", "*pandas*" and "*numpy*") used in the risk calculator by typing "*conda install package-name*" or "*pip install package-name*" in the terminal.

3.2 Type "*python risk_calculator.py*" to run the psychosis risk calculator. If the process is completed successfully, logs of the risk calculation will be printed in the terminal and the risk results will be stored in a new Elasticsearch index called "*psychosis_risk*" within the CogStack platform.

3.3 Check the risk results by using the Kibana interface. Similar to Steps 2.5 and 2.6, add a new index pattern "*psychosis_risk*" to connect Kinbana with the "*psychosis_risk*" index, and explore the risk results through the "*Discover*" page. To facilitate identifying new patients at-risk, use "first_primary_diagnosis_date" as the "Time Filter" field in building the "*psychosis_risk*" index. When exploring data in the "*Discover*" page, make sure that the index pattern "*psychosis_risk*" is selected.

**4. Data visualization**

4.1 In addition to searching and accessing individual-level information via the "*Discover*" page in Kibana, one can build visualizations and dashboards to obtain an overview of characteristics for the whole population of at-risk patients. To do this, click on **Visualize** in the side navigation of Kibana. Then, click the **Create new visualization** button and choose a visualization type (e.g., pie and line charts). Select "*psychosis_risk*" as the index that one wants to visualize through Kibana. By default, visualizations will include all records/patients in the "*psychosis_risk*" index. Detailed instructions of building Kibana visualizations are available on "https://www.elastic.co/guide/en/kibana/6.4/visualize.html".

4.2 To select a specific subset of data for visualization, add a "*filter*". For example, selecting a filter filed as "*h_2_year*", choosing an operator as "*is not between*" and setting values from "*0.0*" to "*0.05*" will only include patients whose risk of psychosis in 2 years are higher than 0.05.

4.3 Once individual visualizations are built, click on **Dashboard** in the side navigation of Kibana to create a dashboard that displays a set of related visualizations together. Click **Create new dashboard** and the **Add** button to create a new dashboard panel. Click visualizations that one wants to show within the new dashboard panel. Click **Save** and type a title to save the panel. Instructions on building Kibana dashboards are available at https://www.elastic.co/guide/en/kibana/6.4/dashboard.html.

**5. Risk Alerting**

5.1 Click on **Management** in the side navigation of Kibana and then click **Watcher** under **Elasticsearch** to create alerting for clinicians when patients were at-risk of psychosis. If the **Watcher** button is not visible, click **License Management** and click **Start trail** or **Update license**.

5.2. Click **Create advanced watch** to set up a new Watcher. Type an "ID" and "Name". Delete the content of "*Watch JSON*" section and copy the content in the "*watcher.json*" file in the "psychosis" directory to the "*Watch JSON*" section. This watcher will send alerting Email to "clinician@nhs.uk" (which can be replaced with the Email address where one wants to send alerts) from "username@nhs.uk" (which was set in Step 2.3) if there are one or more patients whose risk of psychosis in 2 years are higher than 0.05 (a tentative threshold for feasibility testing) in every 24 hours.

5.3 Before saving the Watcher, click **Simulate** to test the Watcher execution. If the Watcher is set successfully, one will see the simulation output printed. In case of error in the settings, follow the instructions on https://www.elastic.co/guide/en/elastic-stack-overview/6.4/watcher-getting-started.html.

5.4 To stop a Watcher, permanently delete it or temporarily deactivate it from the "*Status*" page of the Watcher.

**REPRESENTATIVE RESULTS:**

In this section, we present implementation results focusing in practicality in handling live clinical data streams elaborated through the risk calculator and facilitating timely delivery of prognostic results to clinicians. Evaluations of the clinical utility of the system, such as the adherence of clinicians to the recommendations made by the risk calculator, will be presented in a separate report when complete.

**Ingestion of source data**

We deployed the psychosis risk calculation and alerting system based on a replica database of ePJS in SLaM. This replica database synchronizes the live data from ePJS every 10 minutes. A database view combining patients' information for psychosis risk calculation was built in this replica database, where each record contains information for a patient. All records in this view were ingested into the CogStack platform in real time (approximately 0.6 microsecond per record in a virtual machine with 8-core CPU and 16GB RAM). Until 13 July 2019 when this manuscript was prepared, all the records of 202,289 patients who received a first index diagnosis of non-organic and non-psychotic mental disorder in SLaM were ingested into CogStack for psychosis risk calculation, stored in the "*psychosis_base*" Elasticsearch index. **Figure 1** shows the number of records ingested into CogStack over time, in chronological order based on the last update date of a record. By comparing the numbers and content of records in the database and the Elasticsearch index, no missing and discrepant data were found, which confirms the reliability of CogStack Pipeline in data ingestion and synchronization.

**Validation of risk results**

To validate the implementation of the psychosis risk detector in this protocol, we compared at-risk patients detected by CogStack (called “CogStack version”) with those detected by the original risk calculator based on CRIS (called “CRIS version”). Since there were no thresholds developed to screen an at-risk patient[6,14,21], we here used a tentative threshold of 5% for the risk of psychosis in two years. Note that this tentative threshold is merely to test whether the system can pragmatically work in the NHS and is susceptible to change with future research. The actual threshold for an optimal detection of at-risk individuals will need to be identified in future large-scale studies. Specifically, we first retrieved all patients who had a risk for psychosis above the threshold in the CRIS version (the number of patients N=169). All these patients received a first index diagnosis of non-organic and non-psychotic mental disorder in SLaM from May 14th 2018 to April 29th 2019. By filtering patients who were diagnosed in the same time period, we then retrieved N=170 patients whose risk for psychosis in 2 years were higher than 0.05 in the CogStack version. Finally, we compared the difference between the two sets of patients, where the total number of unique patients in the two sets are N=173. We found that 161 patients (accounting for 93% of 173 patients) had the same scores in both versions. The high degree of agreement confirms the validity of this CogStack-based protocol in generating risk scores.

There were 12 patients having different risk scores in the two versions. By inspecting patients' EHRs, we found that this difference was because data for these patients were updated after the risk scores were calculated in the CRIS version. Specifically, although predictors used in the risk calculator, such as date of birth, gender and self-assigned ethnicity, were static variables, some patients' health records had a missing or default value for a variable (e.g., an unknown ethnicity)

at an earlier stage and these variables were entered or updated at a later stage. This can lead to different risk scores at two different stages. Similarly, the first primary index diagnoses of some patients were invalidated after an initial risk score was calculated based on these diagnoses. In this case, the risk calculator will look for the next valid primary diagnosis for such patient and re-calculate a risk score. The updated risk score can also differ from the initial one. As the original risk calculator was developed based on retrospective data in CRIS for research use, the original calculator pipelines did not synchronize these updates in EHR data and refresh the risk scores in a timely manner. In contrast, a patient's risk score will be re-calculated in the CogStack version if any source data of the patient is updated, which allows this CogStack-based calculator to provide the most reliable and up-to-date risk scores for patients. These results strongly highlight the reliability of risk scores in this protocol.

**Result visualization and risk alerting**

To demonstrate the capabilities of CogStack in data visualization, we built a dashboard for information about patients at-risk of psychosis. As used before for feasibility testing, we selected those who have a risk of psychosis in two years higher than 5% as at-risk patients. **Figure 2** shows the visualizations of characteristics for patients at-risk of psychosis, including patients' ethnicities, genders, ages and categories of diagnoses. Apart from visualizing risk results via Web interfaces (e.g., Kibana), this protocol allows risk alerts to be sent to users or clinicians through other notification channels such as Email. **Figure 3** shows the interface for setting a risk alerting service by using the Watch component in Kibana. Once this service is configured successfully, users can receive an Email notification if there were one or more patients whose risk of psychosis in two years are higher than 5%. **Figure 4** shows an example of these Email notifications, which reports the numbers of patients at-risk and these patients' boroughs. Since more work is needed to tailor how the predicted psychosis risk scores are communicated, we have not sent risk notifications directly to clinicians. For testing the technological feasibility, all notifications in this study were sent from a technical researcher (T.W.) to a clinical researcher (D.O.) via the SLaM's email system within a secure network. Only an aggregated statistic of patient information was included in a notification; no any personally identifiable information was included.

**FIGURE AND TABLE LEGENDS:**

**Figure 1: Source data ingested into CogStack.** There are 202,289 records in total ingested into the "*psychosis_base*" Elasticsearch index until 13 July 2019, and the histogram shows the numbers of records ingested over time, ordered by the last update data time of a record. One can also query both structured and unstructured information, and obtain search hits that match the query in this page.

**Figure 2: Dashboard of characteristics of patients at-risk of psychosis (i.e., the risk of psychosis in 2 years higher than 0.05).** (**a**) Distribution of ethnicities for patients at-risk, where outer pies are the subcategories of an ethnicity category in inner pies. (**b**) distribution of patients' gender, (**c**) distribution of patients' ages at diagnosis and (**d**) number of patients per diagnosis group.

**Figure 3: Setting and simulating Watch for risk alerting.**

**Figure 4: An example of risk alerting Email.** The numbers of patients at-risk of psychosis in each Clinical Commissioning Groups (CCG) are reported in parentheses.

**DISCUSSION:**

We have demonstrated the first EHR implementation of a real-time psychosis risk detection and alerting system based on CogStack, an open source information retrieval and extraction platform. Following this approach, one can transform and ingest a large set of clinical data in various formats, including structured and unstructured information, into a CogStack instance, so as to enable full-text search, interactive analyses and visualization of data, as well as real-time alerting to clinicians of patients that are at-risk of psychosis. Although the original psychosis risk calculator has been validated in pilot studies across several NHS Trusts, albeit using retrospective patient records[6,14,21], this experimental design provides the first evidence base that this risk calculator can be replicated and deployed for use in real time. This approach allows the automatic delivery of prognostic results to clinicians through existing clinical notification channels, such as Email, in real time. This clearly demonstrates the technical feasibility for conducting a large-scale effectiveness trial to evaluate the ultimate clinical utility of this risk calculator in the real world.

This protocol is empirically innovative, as there does not exist a similar risk detection and alerting system for psychosis. Moreover, this protocol has high generalizability in clinical use, particularly because of the unique strengths of our approach. From a theoretical perspective, we used a risk prediction model that was developed based on a large retrospective de-identified cohort from the SLaM NHS Trust. SLaM provides secondary mental health care to a population of 1.36 million individuals in South London and has one of the highest recorded rates of psychosis in the world. This large cohort, which has high diversity in sociodemographic and diagnostic characteristics, allows us to develop a risk prediction model that is unlikely to be biased towards a population with specific characteristics. This is supported by evidence that the prognostic accuracy of this risk calculator has already been replicated twice in two different databases[14,21], including one outside of SLaM. Another theoretical strength of this risk model is that basic demographic and clinical diagnosis information were used as predictors. Such information is ubiquitous in electronic clinical data and in fact missing data for these predictors have been shown to be relatively rare in our previous studies[14,21]. The high availability of information for building predictors makes it possible to run the risk calculator over a large number of patient samples across different secondary mental health care sectors. In addition, the risk calculator is a general algorithm which is suitable for all individuals at-risk of developing psychosis in secondary mental health care, regardless of individuals' ages. That is, this calculator is not only suitable for the 15-35 age range of peak psychosis risk[16], but also for those outside of this range, showing a high degree of generalizability.

From a practical perspective, both the risk calculator and the CogStack platform are light-weight and open-source services that do not involve resource-heavy techniques or costly infrastructure. Such a low-cost and easy-to-deploy platform can reduce the barriers to its adoption in real-world clinical settings. Also, our solution overcomes the main implementation barrier: risk estimation systems provide little value unless they are used by clinicians in day-to-day practice[25]. Specifically,

our approach accesses data from the EHR, performs analyses independent of an electronic medical record system and can send analysis results back to clinicians through existing notification channels. This method does not require that the business logic in pre-existing systems be modified and can work as a standalone service to support and extend existing clinical decision support systems. Thus, the protocol has high compatibility with pre-existing clinical systems and can be easily integrated into routine clinical practice. Moreover, the protocol provides user-friendly interfaces for searching, analyzing and visualizing of clinical data, which make it easy for clinicians to interpret and explore the risk results.

This protocol also has its limitations. First, the effectiveness of this protocol has not been evaluated in routine clinical practice. This study focused on technical feasibility tests of implementing a real-time psychosis risk detection and alerting system in a local EHR. To further evaluate the effectiveness of this system in routine clinical practice, future large scale randomized controlled trials are needed[6]. A second limitation is that the predictions of risk scores in this protocol were made based on the first primary diagnoses, which are static data collected at a single snapshot. However, the CHR-P symptoms are intrinsically evolving over time. A dynamic version of psychosis risk calculator, in which prediction models can be dynamically updated to reflect the changes, has been developed recently[26]. Future work will focus on integrating this dynamic calculator in the current protocol.

The most critical step in this approach was identifying EHR data that were used for extract predictors in the risk calculator. This may also involve creating data element mappings, when an EHR system used a data model different from that used in this protocol, such as distinct coding systems for patients' ethnic groups. We have open-sourced all the code and mapping definitions online (https://github.com/cogstack-slam/psychosis). Based on these materials, one would be able to replicate the deployment or adjust the calculator depending on one's own circumstance. Another critical step was creating a database view for data ingestion in CogStack. Since relational join operations (i.e., combining columns from one or more database tables) in Elasticsearch can lead to high computational cost, we conducted these join operations in the EHR database by creating a database view. This view combined all information that was needed to extract predictors in the risk calculator, and two vital fields that were used by CogStack pipelines for data partitioning in data ingestion. The first field is a unique primary key for each record in the view ("*patient_id*" used this protocol) and the second is a timestamp when a record was modified most recently. If these two fields were not set properly, CogStack might not synchronize data updates in an EHR database timely. Detailed instructions for troubleshooting issues on CogStack data ingestion are available on https://cogstack.atlassian.net/wiki/spaces/COGDOC/overview and https://github.com/CogStack/CogStack-Pipeline.

This protocol is highly transportable and can be easily deployed in NHS Trusts that have a CRIS or CogStack platform. So far, the CRIS platform—including the consenting procedures—has been fully described elsewhere and is under expansion across 12 NHS Trusts in the UK, harnessing over 2 million deidentified patient records (https://crisnetwork.co/). Similarly, the CogStack platform has been deployed not only in SLaM, but also other NHS Trusts across the UK such as University College London Hospitals (UCLH), King's College Hospital (KCH), Guy's and St Thomas' (GSTT), and

Mersey Care NHS Trusts. Those Trusts without such as platform can use an online version of risk calculator (http://psychosis-risk.net), or build this protocol from scratch based on this manuscript and our online documents. Although this protocol is developed for psychosis risk detection, the architectural design of this protocol is not tied-in to this specific use case. The protocol is flexible enough to allow for reconfiguration and repurposing of the real-time monitoring and alerting components for other risk measurement areas, such as adverse drug reactions, thereby allowing clinicians to timely take action to improve patient care, safety and experience.

**ACKNOWLEDGMENTS:**

This study is funded by and is a direct output of the King's College London Confidence in Concept award from the Medical Research Council (MRC) (MC_PC_16048) to PFP. RD and AR were supported by: (a) the Maudsley Charity; (b) the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London; (c) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust; (d) The BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No. 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA; it is chaired, by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC; and (e) The National Institute for Health Research University College London Hospitals Biomedical Research Centre. These funding bodies had no role in the design of the study, collection and analyses. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

**DISCLOSURES:**

The authors have nothing to disclose.

**REFERENCES:**

1. Lieberman, J.A., First, M.B. Psychotic disorders. *New England Journal of Medicine*. **379** (3), 270-280 (2018).
2. Oh, H., Koyanagi, A., Kelleher, I., DeVylder, J. Psychotic experiences and disability: findings from the collaborative psychiatric epidemiology surveys. *Schizophrenia Research*. **193**, 343-347 (2018).
3. Tsiachristas, A., Thomas, T., Leal, J., Lennox, B.R. Economic impact of early intervention in psychosis services: results from a longitudinal retrospective controlled study in England. *BMJ Open*. **6** (10), e012611 (2016).
4. Fusar-Poli, P., McGorry, P.D., Kane, J.M. Improving outcomes of first-episode psychosis: an overview. *World Psychiatry*. **16** (3), 251-265 (2017).
5. Fusar-Poli, P. The clinical high-risk state for psychosis (CHR-P), version II (2016).
6. Fusar-Poli, P. et al. Real-world Implementation of a Transdiagnostic Risk Calculator for the Automatic Detection of Individuals at-risk of Psychosis in Clinical Routine: Study Protocol.

572 *Frontiers in Psychiatry*. **10**, 109 (2019).
573 7. Fusar-Poli, P. et al. Disorder, not just state of risk: meta-analysis of functioning and quality
574 of life in people at high risk of psychosis. *British Journal of Psychiatry.* **207** (3), 198-206 (2015).
575 8. Fusar-Poli, P. et al. Heterogeneity of psychosis risk within individuals at clinical high risk:
576 a meta-analytical stratification. *JAMA Psychiatry*. **73** (2), 113-120 (2016).
577 9. Fusar-Poli, P. et al. Diagnostic and prognostic significance of brief limited intermittent
578 psychotic symptoms (BLIPS) in individuals at ultra high risk. *Schizophrenia Bulletin*. **43** (1), 48-56
579 (2016).
580 10. Fusar-Poli, P. et al. Prognosis of brief psychotic episodes: a meta-analysis. *JAMA*
581 *Psychiatry*. **73** (3), 211-220 (2016).
582 11. Fusar-Poli, P., Sullivan, S., Shah, J., Uhlhaas, P. Improving the detection of individuals at
583 clinical risk for psychosis in the community, primary and secondary care: an integrated evidence-
584 based approach. *Frontiers in Psychiatry*. **10**, 774 (2019).
585 12. Fusar-Poli, P. Extending the benefits of indicated prevention to improve outcomes of first-
586 episode psychosis. *JAMA Psychiatry*. **74** (7), 667-668 (2017).
587 13. Fusar-Poli, P. et al. Transdiagnostic psychiatry: a systematic review. *World Psychiatry*. **18**
588 (2), 192-207 (2019).
589 14. Fusar-Poli, P. et al. Development and validation of a clinically based risk calculator for the
590 transdiagnostic prediction of psychosis. *JAMA Psychiatry*. (2017).
591 15. Fusar-Poli, P., Hijazi, Z., Stahl, D., Steyerberg, E.W. The science of prognosis in psychiatry:
592 a review. *JAMA Psychiatry*. **75** (12), 1289-1297 (2018).
593 16. Radua, J. et al. What causes psychosis? An umbrella review of risk and protective factors.
594 *World Psychiatry*. (2018).
595 17. Fusar-Poli, P. et al. Deconstructing vulnerability for psychosis: Meta-analysis of
596 environmental risk factors for psychosis in subjects at ultra high-risk. *European Psychiatry*. **40**,
597 65-75 (2017).
598 18. Fusar-Poli, P. et al. Clinical-learning versus machine-learning for transdiagnostic
599 prediction of psychosis onset in individuals at-risk. *Translational Psychiatry*. **9** (1), 1-11 (2019).
600 19. Stewart, R. et al. The South London and Maudsley NHS foundation trust biomedical
601 research centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry*. **9**
602 (1), 51 (2009).
603 20. Jongsma, H.E. et al. Treated incidence of psychotic disorders in the multinational EU-GEI
604 study. *JAMA Psychiatry*. **75** (1), 36-46 (2018).
605 21. Fusar-Poli, P. et al. Transdiagnostic risk calculator for the automatic detection of
606 individuals at-risk and the prediction of psychosis: Second replication in an independent national
607 health service trust. *Schizophrenia Bulletin*. (2019).
608 22. Fusar-Poli, P. et al. Transdiagnostic individualized clinically based risk calculator for the
609 detection of individuals at-risk and the prediction of psychosis: Model refinement including
610 nonlinear effects of age. *Frontiers in Psychiatry*. **10**, 313 (2019).
611 23. Colditz, G.A., Wei, E.K. Risk prediction models: applications in cancer prevention. *Current*
612 *Epidemiology Reports*. **2** (4), 245-250 (2015).
613 24. Jackson, R. et al. CogStack - Experiences Of Deploying Integrated Information Retrieval
614 And Extraction Services In A Large National Health Service Foundation Trust Hospital. *BMC*
615 *Medical Informatics and Decision Making*. **18** (47) (2017).

25. McGorrian, C., Leong, T., D'Agostino, R., Graham, I. Risk estimation systems in clinical use: SCORE, Heart Score and the Framingham system. *Ther Strateg Cardiovasc Risk 1st ed Oxford, UK Clin Publ*. 159-172 (2008).
26. Studerus, E., Beck, K., Fusar-Poli, P., Riecher-Rössler, A. Development and Validation of a Dynamic Risk Prediction Model to Forecast Psychosis Onset in Patients at Clinical High Risk. *Schizophrenia Bulletin*. (2019).
27. Damschroder, L.J. et al. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implementation Science.* **4**, 50 (2009).
28. Fusar-Poli, P. et al. Outreach and support in south London (OASIS), 2001–2011: ten years of early diagnosis and treatment for young individuals at high clinical risk for psychosis. *European Psychiatry*. **28** (5), 315-326 (2013).