UNIVERSITÀ
DI PAVIA

SCUOLA DI ALTA FORMAZIONE DOTTORALE
MACRO-AREA SCIENZE DELLA VITA

PhD in Psychology, Neuroscience and Data Science
Department of Brain and Behavioural Sciences

Improving psychosis detection and management using precision medicine.

Academic year 2022-2025
Cycle 38th

Coordinatore
Prof.ssa Gabriella Bottini

Doctoral candidate
Kamil Krakowski

Tutor
Prof. Paolo Fusar-Poli

Co-tutor
Prof. Daniel Stahl

Table of Contents

ACKNOWLEDGMENTS

I owe tremendous thanks to my partner and love, Julia, who supported me throughout the entire process of this PhD. If it weren't for her support, motivation, and kindness, I might not have finished this project. I am equally grateful to my mum and dad for all their care and dedication in supporting my education up to this point.

I would like to express my deepest gratitude to my supervisors, Paolo Fusar-Poli and Daniel Stahl. Paolo, thank you for the fantastic opportunity to pursue research in this PhD project and for always pushing me to produce high-quality work. Daniel, thank you for introducing me to the field of clinical prediction models and for broadening my horizons. The advice and direction you have given me have been invaluable to my development.

My special thanks go to Dom Oliver. Thank you, Dom, for your great support throughout the entire PhD and for your commitment to helping and mentoring students. Your help in navigating the intricacies of health data and polishing the papers was invaluable.

To Maite, Yanakan, Andrea, Sun, Rashmi, and all the wonderful collaborators and co-authors I have met during my PhD—my deepest gratitude for your contributions. You have been my greatest source of inspiration, and I hope we will continue working together in the future.

ABSTRACT

Psychosis is a debilitating disorder that causes substantial societal cost and disrupts the lives of predominantly young people. Comprehensive management at all clinical stages including indicated prevention during prodromal period, first-episode interventions, and care for chronic patients is crucial to improve prognosis, symptoms and functioning of this vulnerable population. Over the past years advances in precision medicine have demonstrated potential for improving detection of at-risk individuals and optimising treatment selection. This thesis expands the knowledge on precision medicine methods and introduces scalable dynamic detection of at-risk individuals (Part A) and presents new methods for incorporating shared decision-making into the precision treatment rules for the first episode of psychosis (Part B).

Part A of this thesis aims to improve detection of at-risk individuals using dynamic survival modelling. Chapter 1 of this thesis introduces the concept of indicated prevention and Clinical High Risk of Psychosis. The motivation for effective detection strategies and main challenges are also presented. The overview of clinical prediction modelling in psychiatry with special emphasis on the dynamic survival modelling and its importance is also discussed. Chapter 2 of this thesis showcases a study that develops and internally-externally validates a dynamic risk calculator for psychosis using dynamic survival modelling. The study involves data of 158,139 patients extracted form electronic health records in secondary mental health care to construct a Cox Landmark model. By using the innovative method based on multi-level meta regression farmwork, the dynamic model was compared to the static approach. The dynamic approach compared to the static one improved discrimination performance, averaged over all time points by 0.035 (95% CI = 0.031–0.043, $p < 0.001$), measure in Harrell's C score. This corresponds to a 20% error reduction. The improvements to calibration and potential clinical utility were also achieved for predictions made at later time points. The presented approach advances the knowledge on dynamic, scalable and systematic approaches to detection of individuals at risk of psychosis. By refining the existing risk calculators, the findings have high translational potential.

Part B aims to develop precision medicine methods that incorporate shared decision making and improve treatment selection for first-episode patients. Chapter 3 of this thesis introduces the main concepts of psychopharmacology in the first episode of psychosis and presents a comparison of existing decision support systems for antipsychotic selection. The barriers to individualisation of treatment selection in psychosis is also outlined. Finally, the chapter presents the farmwork of pragmatic precision psychiatry and causal inference methods for observational data that can be used for treatment optimization. Chapter 4 of this thesis showcases a study that presents the first development and validation of precision treatment rule for the first episode of psychosis that incorporate patients' preferences. Electronic health records consisting of individual level data of 1709 patients from early interventions for psychosis services was used to construct innovative precision treatment rules combining causal forest, ranking of patient preferences alongside effectiveness and side effects outcomes. Across different preferences, precision treatment rules predominately recommended aripiprazole, as the optimal option for 80% to 98% of patients. If the treatment rules were implemented, the

rates of most side effects included in the study would significantly fall. However, extrapyramidal side effects would increase. No effects on the rate of hospitalisation and change of medication were found.

Part C of the thesis consists of summary of the findings of this thesis, discusses the impact of the results on the field of psychiatry and precision medicine, presents the limitations of the study, and outlines the direction for further research.

STATEMENT OF PERSONAL CONTRIBUTION

For the dynamic transdiagnostic risk calculator of psychosis study that is presented in chapter 2, I performed all the analysis and wrote the first draft of the manuscript. For the study presented in the study on precision treatment rules in first episode of psychosis presented in chapter 4, I have contracted the data set by applying for the extraction of the data from the electronic health records. I linked the data, performed all the analysis and produced the first draft of the manuscript. Finally, I have wrote this thesis in its entirety, with the one exception. The published paper, after being drafted by me, underwent peer review and were edited by co-authors prior to acceptance by the journal.

PART A: IMPROVING DETECTION OF INDIVIDUALS AT RISK OF PSYCHOSIS USING DYNAMIC SURVIVAL MODELLING

1. Introduction

1.1 Indicated prevention of psychosis and the clinical high risk of psychosis.

Since the end of the 1990s there has been a growing effort to establish the primary indicated prevention of psychosis. Indicated prevention targets individuals who are experience signs and symptoms that indicate increased risk of developing a disorder. The Clinical High-Risk of Psychosis state (CHR-P) has been proposed as a way to diagnose individuals who are help-seeking, experiencing positive psychotic symptoms (such as unusual thought content, suspiciousness, perceptual abnormalities like hallucinations, and disorganized communication) and who also exhibit functional impairment (1). Retrospectively, 78% of patients who developed psychosis recall experiencing prodromal symptoms prior to the onset of their First Episode of Psychosis (FEP) (2). Thus, this state is of particular importance for indicated prevention. Currently, the CHR-P construct is widely adopted worldwide and has been included in numerous clinical guidelines. According to a recent study (3), guidelines on CHR-P have been published in Italy (4), Catalonia (5), the United Kingdom (6,7), Australia (8,9), New Zealand (9), Canada (10), Germany (11) and by the European Psychiatric Association (12). CHR-P services that provide specialist care for this vulnerable group have been implemented around the world. In 2020 there were 47 CHR-P services providing care for over 22,000 individuals (13).

The primary aim of CHR-P services is prevention of psychosis onset, and the transition to a first episode of psychosis has been used as a main outcome in most studies on the CHR-P population (14). It has been proposed that 'disease-modifying' interventions could alert the course of the disorder, especially during the CHR-P stage of psychosis (15), before the symptoms of psychosis become too severe and persistent. Individuals meeting the CHR-P criteria are at substantially higher risk of developing psychosis, with around 25% of cases transitioning to full-blown psychosis within three years (16). An additional aim of CHR-P services is to reduce the severity of attenuated positive and negative symptoms, provide support to improve social and vocational functioning, and enhance the quality of life for CHR-P patients. Another important function of CHR-P services is rapid detection of psychotic episodes and referral to FEP services that provide antipsychotic treatment. The duration of untreated psychosis (DUP) - the period between psychosis onset and the initiation of antipsychotic treatment – is associated with an increased probability of relapse (17) and worse clinical and social outcomes for FEP patients (17). CHR-P services can reduce the duration of DUP and therefore improve prognosis and outcomes for their patients.

The improvement of care provided by CHR-P services depends on three interconnected factors: (i) efficient detection of individuals at risk of psychosis, (ii) accurate prognosis of clinical outcomes, and (iii) preventive interventions (18). In particular, improved detection can expand the benefits of the CHR-P construct to a larger number of people who may be at risk of this

debilitating disorder and facilitate prevention at a larger scale. In Part A of this thesis, I present a study designed to improve the detection of psychosis using dynamic survival modelling.

In this introduction to part A, I will outline the current approaches for detecting individuals at risk of psychosis and their limitations. I will present how the limitations of current methods can be addressed using precision medicine approaches, including a transdiagnostic risk calculator for psychosis. I will introduce dynamic survival modelling and its role in clinical prediction models in psychosis. In the Chapter 2, I will present how the transdiagnostic risk calculator can be refined using dynamic modelling to improve the detection of at-risk individuals.

1.2 Detection of individuals at risk of psychosis.

Effective detection of individuals at risk of psychosis is the first crucial step in increasing the number of patients that can benefit from the CHR-P construct (18). Even the most effective prognostic methods and preventive interventions would have very limited impact if the detection rate was low. Unfortunately, current detection approaches are suboptimal and face multiple challenges.

The first challenge is that not all patients who will later develop psychosis present with subthreshold psychotic symptoms before the onset. It is estimated that around one third of patients develop psychosis without having previously experienced the CHR-P state (19,20). Therefore, additional information beyond the CHR-P criteria is needed to detect all FEP patients at the critical stage before the onset of psychosis.

The second challenge is that current strategies are ineffective. The detection approaches are highly varied and differ significantly between clinical services. Some services rely solely on clinical referrers and others use outreach campaigns. Several interventions in the early interventions services have been implemented to improve detection of both FEP (21,22), and CHR-P (23,24) patients. These included screening assessments and recruitment, workshops with community partners, roadshows, student internships, advertisement in print media, and social media campaigns. Emerging evidence suggests that current detection strategies are ineffective. In South London and Maudsley (SLaM) National Health Service (NHS) Trust only 5% of patients who presented with FEP were identified by the local CHR-P services before the onset of psychosis (25). In Australia, only 12% of FEP patients were detected at CHR-P stage (26). This implies that the vast majority of patients who develop psychosis will not receive the care offered by the CHR-P paradigm and highlights the need for better detection methods.

The third challenge is associated with the influence of recruitment strategies on pre-test risk dilution and with the fact that the accuracy of CHR-P assessments is largely driven by their ability to rule out psychosis (i.e., high sensitivity). Most individuals undergoing CHR-P assessment are selected either by clinicians based on a suspected risk of psychosis (27) or because of help-seeking behaviour (28). However, depending on the recruitment strategy adopted, the pre-test risk can vary substantially. The pre-test risk is highest when patients in

secondary mental health services are targeted, intermediate when primary care patients are targeted, and lowest when individuals who are not help-seeking are targeted. The pre-test risk can vary from 0.43% to 15%, which represents a difference of over 35 times (18). Psychometric CHR-P assessments performed by specialised services demonstrate good accuracy, with the area under the curve at 34 months of 0.85 (95% CI: 0.81–0.88), according to the metanalytical evidence (29). CHR-P assessments have high sensitivity for ruling in psychosis (0.93, 95% CI: 0.87–0.96) but lower specificity for ruling out psychosis (0.58, 95% CI: 0.50–0.66) (29). When individuals have a pre-test risk of psychosis of 15%, those meeting the CHR-P criteria have a 26% risk of developing psychosis at 3 years (odds ratio 1.7), while those not meeting the CHR-P criteria have a 1.56% risk of developing psychosis (odds ratio 0.1). When the pre-test risk is diluted by assessing the general population, those meeting the CHR-P criteria have only a 5% risk of developing psychosis within 3 years (30,31). This indicates that more aggressive recruitment strategies are not a viable solution and can result in CHR-P services providing help to individuals at lower risk of psychosis. Given limited capacity of those services this can divert recourses from the individuals at greatest need.

Given these challenges, there is a need for an effective solution that can improve the number of detected individuals at risk of psychosis while being systematic and avoiding pre-test risk dilution.

1.3 Clinical prediction models in psychiatry

Improvement in detection of at-risk individuals while avoiding pre-test risk dilution can be achieved with precision medicine methods. The aim of precision medicine is to individualise healthcare based on each patient unique and evolving health status and characteristics (32). Individualised prediction of psychosis risk can be provided by clinical risk prediction models that relate multiple patient characteristics (predictors) to the outcome of interest, such as the risk of developing the disorder (33). The process of creating clinical prediction models can be conceptually divided into three main parts: (i) model development, (ii) internal and external validation, and (iii) implementation. There are detailed tutorials (34,35) and guidelines (36) on clinical prediction models for different fields of medicine, and more specifically for psychiatry (33). Here, I explain the main steps of the development of clinical prediction models in the context of psychiatry.

The development of the clinical prediction models begins with appropriate model design. The risk prediction model must be motivated by a specific clinical uncertainty (37), and the outcome of the model must be a well-defined clinical event such as the onset of a disease, hospitalisation or death. The type of the outcome defines the estimation method that relates the predictors with the risk of an event. According to the recent review (38) the most common estimation methods in psychiatry research are regression methods, machine learning algorithms or ensemble methods that combine multiple regression or machine learning predictions into one optimally weighed average. Among regression methods, the most common are logistic

regression for binary outcomes and Cox proportional hazard method for time-to-event outcomes (such as time to the onset of psychosis).

Following the model design, the predictors are selected. Mental disorders are multifaceted and require multiple predictors for effective risk prediction. The selection of predictors with a priori knowledge is recommended over automatic step-wise selection methods (33). To select predictors several evidence synthesis methods can be used. Meta-analysis and umbrella reviews are the most robust sources of association between different factors and the outcome of interest while minimizing biases (39), making them best suited for predictor selection. The automatic step-wise methods for predictors selection are not recommended (33). True predictors can be excluded by the step-wise method due to insufficient power, while random predictors can be included due to multiple testing (false positives) (40). If data driven predictor selection is desirable (due to e.g. limited a priori evidence) then a penalisation methods such as Least Absolute Shrinkage and Selection Operator (LASSO), elastic net, or different machine learning algorithms (41) involving penalisation (also called regularisation or shrinkage), can be used (34).

The next step is data preparation i.e. the appropriate coding of predictors and outcomes, ensuring that all predictors are available and the treatment of missing data. Continuous predictors should not be categorised, and merging and splitting groups must be avoided. These practices reduce the information available in the data and cause biases in prognosis research (35). Predictors or outcomes with missing data can be included in the model, however, appropriate multiple (42) or single imputation methods (43) must be used.

Next, the model is fitted in development or derivation data set. The model's error in the fitted data can be deconstructed into two parts. The error due to bias (model's inability to capture patterns and complexity in the fitted data) and error due to variance (model's inability to generalise to data outside the derivation set). The bias-variance trade-off is a key issue in the prediction modelling research (41) and the aim is to select the appropriate level of model's complexity that will minimize both types of errors. The performance of the fitted model can be evaluated by multiple metrics. The overall performance can be evaluated using measures of explained variability, such as R^2 or Brier score (44) that measure the distance between the predicted and actual outcome. The model's ability to correctly separate those who experience the outcome and those who do not can be assessed using the discrimination measures such as sensitivity, specificity, area under the curve or Hearrell's C statistic for the survival outcomes (44). The agreement between the predicted risk and the observed outcomes is measured with calibration metrics (regression slope of prognostic index, calibration plot) (44). Finally, the potential clinical utility of the prediction model is evaluated with the net benefit analysis (45).

After the development stage, validation is essential to ensure the models' generalisability, which consists of reproducibility and transportability (46). Reproducibility is the model's ability to accurately predict outcomes for the new patients from the same population. Transportability is the model's ability to provide accurate prediction in the new but related populations. To assess different components of generalisability several validation approaches are used. The

internal validation evaluates the model's reproducibility and is performed by using the data from the same patient population as the one that was used to develop the model. To perform the internal validation the data can be randomly split into two parts – development set and validation set. Alternatively, more efficient methods such as k-fold cross validation, leave-one-out cross validation or bootstrapping can be used. Regardless of the method used to split the data, the model must be fitted and evaluated on separate partitions to avoid optimism and to provide honest assessment of the model performance. Another validation approach is internal-external validation. It involves partitioning the data into clusters based on a feature of the data (e.g. by region, by hospital or by study source). One cluster is used as the test set and the remaining data serve as the training set in the iterative manner. The internal-external validation provides information on model's generalisability as clusters used in this method are heterogenous and thus evaluates model performance in different populations. Finally, external validation is essential in confirming model transportability to new but related populations. It is performed by evaluating the predictions of the model in a new patient population drawn from a different setting. Importantly, the model must use the original coefficients derived in the original development setting. According to a recent review (38) only 20% of clinical prediction models in psychiatry had some form of external validation and the problem is the most acute for low and middle-income countries.

The ultimate aim of clinical prediction models is to improve patient's care and healthcare cost-effectiveness. That can only be achieved through implementing the model in a healthcare system and informing decisions regarding patients' care and organisational practices (47). Prospective comparative and, ideally, randomised cluster studies are designed to evaluate the effect of clinical prediction model implementation (48). As identified by the review (38) in psychiatry literature only two implementation studies of prediction models have been published (49,50). Both of these studies evaluate the transdiagnostic risk calculator for psychosis, the refinement of which is the focus of this thesis.

1.4 Transdiagnostic risk calculator for psychosis

Clinical predictions models provide a crucial solution for systematic detection of at-risk individuals while mitigating the pre-test risk dilution problem. The transdiagnostic risk calculator for psychosis has been originally developed in SLaM NHS Trust in the UK by Paolo Fusar-Poli et al. (25). The risk calculator was developed in the secondary mental healthcare service where at the time of the risk calculator development 95% of individuals who developed psychosis went undetected. This highlighted a critical gap in psychosis detection at the site. Fusar-Poli et al. used the Cox proportional hazard model (51) to relate predictors at the index time (the time of admission to the secondary mental healthcare service) to the time of developing psychosis. The predictors have been selected a priori based on meta-analytical evidence and included index time diagnosis, age, sex, ethnicity and age by sex interaction. The model development and validation have been performed by splitting the data into two parts based on geographical location. The patients for the borough of Lambeth and Southwark (n= 33820, 37.08%) were used for the model's development and patients from the remining

boroughs (n = 54716, 60%) were used as validation sample. Patients with missing borough data (n = 2663, 2.92%) were excluded.

The transdiagnostic risk calculator core characteristics provide many desirable features for improving psychosis detection. The calculator is transdiagnostic meaning it can detect individuals at risk of psychosis outside the CHR-P samples. The tool can overcome the limitation of the CHR-P construct by allowing detection among all individuals receiving their first diagnosis in secondary mental healthcare service. The calculator is clinically focused as it uses routinely collated predictors that are available in the Electronic Health Records (EHRs), not requiring any additional effort for their collection. The tool is lifespan-inclusive as it can provide predictions for patients across different ages including the age range associated with the peak risk of psychosis (52). The detection process with the calculator can be automated within the EHRs. The automation ensures low cost of the implementation, allows for a large scale of operation (all patients receiving the first diagnosis in SLaM are automatically screened) and ensures standardisation. The aforementioned characteristics of the calculator enhance its implementation potential and help address the main barriers to improving care in psychiatry through clinical prediction models.

The transdiagnostic risk calculator demonstrated fair to good discrimination ability, achieving a Harrell's C index of 0.79 in the validation sample. The decision curve analysis of potential clinical usefulness showed that the model provides significant net benefit in both the derivation and the validation sample. The risk calculator has been further externally validated in different settings at the Camden and Islington NHS trust in the UK (53), where it retained good discrimination performance (Harrell's C 0.73). The risk calculator has been also externally validated outside the UK in the primary and secondary healthcare EHR's form the United States (54). In this international external validation, the risk calculator achieved prognostic accuracy above chance (Harrell's C 0.67). Together, these studies show the transportability potential of the risk calculator in both domestic and international settings, which is rarely assessed in psychiatry research (38). The transdiagnostic risk calculator was not only extensively externally validated but also the implementation feasibility studies have been performed. The risk calculator has been implemented in the EHR's at SLaM (50) and the automatic email alert system informing clinicians of the elevated psychosis risk of their patients have been put into operation (49). The results of prospective use of the transdiagnostic risk calculator in the first year were published (49). 3722 patients were screened, and 115 individuals were detected. The clinicians' adherence was 74% without outreach and 85% with outreach. These studies show that the implementation barriers for the risk calculator can be overcome, however further comparative studies with longer follow-up time that are now underway for full impact evaluation of the risk calculator.

The transdiagnostic risk calculator was refined with additional symptoms and substance use predictors (55). The structured EHR's data is an excellent source of sociodemographic predictors and diagnostic codes, however most of the clinical information is stored in the free text, event notes and uploaded attachments. The Natural Language Processing (NLP) methods applications have been implemented in the SLaM NHS Trust to access the information stored in

the free text notes and to convert it to a structured form that can be used in the prediction models research (56). 14 additional symptoms and substance use predictors extracted by the NLP methods have been added to the transdiagnostic model. This resulted in the NLP enhanced model (55) which significantly improved the prognostic ability achieving the Harrel's C of 0.85 on the external validation sample and demonstrating the value of the NLP methods in psychosis detection.

1.5 Moving from static to dynamic survival modelling in psychiatry

Current approaches to risk detection of psychosis and in psychiatry more broadly have been characterised by the pragmatic assumption that single cross section of the data (most often at the time of service entry) can be sufficient for effective risk prediction (57). However, psychosis and meatal disorders are dynamic entities that evolve over time and the risk of transitions changes as time progresses (58). The EHRs contain a history of patients' information from subsegment contacts with the healthcare service, that are by their nature temporal and represent much richer source of information than a single cross section of the data (59). In the context of medical research and psychosis risk prediction the survival models that estimate the time to event and allow for incorporation of censoring information are especially useful and popular (60). The Cox regression model (51) is the most used time-to-event method in the clinical prediction models in psychiatry (38) and have been used to develop the transdiagnostic risk calculator for psychosis (25,55). However, the Cox regression in its most used form is designed to use only a single cross-section of the data and different dynamic survival modelling methods are required to take advantage of the longitudinal predictors stored in the EHRs.

Over the past few decades there have been a growing interest in the longitudinal survival modelling. Many statistical (61) and machine learning methods have been proposed (62). Among statistical methods the Cox landmark regression (61,63), the Cox regression with time varying covariates models (64) and joint models (65) are most prominent. The other group can be broadly labelled as machine learning approaches. They consist of newly developed methods such as Random Forest for Survival and Longitudinal Data 'RF -SLAM' (66), boosted hazard estimator with dynamic covaries 'BoXHED' (67), recurrent neural network for dynamic survival analysis with competing risk 'DeepHit' (68) and combination of landmarking and machine learning assembly (69,70). The statistical methods such as Cox regression with landmarking or time-varying covariates and joint models provide better interpretability, lower computational cost and ease of use compared to the machine learning models. The landmarking approach is especially interpretable (71) compared to the Cox model with time-varying covariates and the landmarking approach can easily accommodate large sample size of the EHRs and large number of variables. This is challenging for the join models as they are more computationally expensive (72). The machine learning approaches offer better flexibility and allow for the modelling of more complex data structures and the incorporation of different data modalities (images, text), potentially offering better prediction performance. However, that comes at the cost of greater computational cost and the lack of inherited interpretability (73). Machine learning models must rely on the interpretability methods (74) that approximate the original model. In the

context of tabular EHRs it is not clear if the complex machine learning methods provide better prediction accuracy compared to the simpler regression methods (75).

In the following Chapter 2 I will be presenting the first development and validation of the dynamic survival model for psychosis detection in the secondary mental health. The Landmark Cox regression method was used as it is the most interpretable and computationally cost-effective method. This study refines the existing transdiagnostic risk calculator (25,55), which have been extensively externally validated and has been implemented in the SLaM NHS Trust. This study improves the previous model by the usage of dynamic survival modelling, imputation of missing data and more efficient internal-external validation.

1.6 References

1. Fusar-Poli P (2017): The Clinical High-Risk State for Psychosis (CHR-P), Version II. *Schizophr Bull* 43: 44–47.

2. Benrimoh D, Dlugunovych V, Wright AC, Phalen P, Funaro MC, Ferrara M, *et al.* (2024): On the proportion of patients who experience a prodrome prior to psychosis onset: A systematic review and meta-analysis. *Mol Psychiatry* 29: 1361–1381.

3. Poletti M, Pelizza L, Preti A, Raballo A (2024): Clinical High-Risk for Psychosis (CHR-P) circa 2024: Synoptic analysis and synthesis of contemporary treatment guidelines. *Asian J Psychiatry* 100: 104142.

4. Sistema Nazionale Linee Guida (2007): *Sistema Nazionale Linee Guida. Linea Guida Del Sistema Nazionale Linee Guida. Gli Interventi Precoci Nella Schizofrenia.* Italy.

5. Catalan Agency for Health Technology Assessment and Research. (2009): *Clinical Practice Guideline for Schizophrenia and Incipient Psychotic Disorder.* Barcelona.

6. National Institute for Health and Care Excellence (NICE) (2014): *NICEGuidelines, Psychosis and Schizophrenia in Adults: Prevention and Management Clinical Guideline.* London, UK.

7. Barnes TR, Drake R, Paton C, Cooper SJ, Deakin B, Ferrier IN, *et al.* (2020): Evidence-based guidelines for the pharmacological treatment of schizophrenia: Updated recommendations from the British Association for Psychopharmacology. *J Psychopharmacol Oxf Engl* 34: 3–78.

8. Early Psychosis Guidelines Writing Group and EPPIC National Support Program (2016): *Early Psychosis Guidelines Writing Group and EPPIC National Support Program, Australian Clinical Guidelines for Early Psychosis, 2nd Edition Update*. Melbourn, Australia.

9. Galletly C, Castle D, Dark F, Humberstone V, Jablensky A, Killackey E, *et al.* (2016): Royal Australian and New Zealand College of Psychiatrists clinical practice guidelines for the management of schizophrenia and related disorders. *Aust N Z J Psychiatry* 50: 410–472.

10. Addington J, Addington D, Abidi S, Raedler T, Remington G (2017): Canadian Treatment Guidelines for Individuals at Clinical High Risk of Psychosis. *Can J Psychiatry Rev Can Psychiatr* 62: 656–661.

11. German Association for Psychiatry, Psychotherapy and Psychosomatics (2019): *DGPPNS3 Guideline for Schizophrenia*. Germany.

12. Schmidt SJ, Schultze-Lutter F, Schimmelmann BG, Maric NP, Salokangas RKR, Riecher-Rössler A, *et al.* (2015): EPA guidance on the early intervention in clinical high risk states of psychoses. *Eur Psychiatry J Assoc Eur Psychiatr* 30: 388–404.

13. Kotlicka-Antczak M, Podgórski M, Oliver D, Maric NP, Valmaggia L, Fusar-Poli P (2020): Worldwide implementation of clinical services for the prevention of psychosis: The IEPA early intervention in mental health survey. *Early Interv Psychiatry* 14: 741–750.

14. Fusar-Poli P, Borgwardt S, Bechdolf A, Addington J, Riecher-Rössler A, Schultze-Lutter F, *et al.* (2013): The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry* 70: 107–120.

15. Millan MJ, Andrieux A, Bartzokis G, Cadenhead K, Dazzan P, Fusar-Poli P, *et al.* (2016): Altering the course of schizophrenia: progress and perspectives. *Nat Rev Drug Discov* 15: 485–515.

16. Salazar de Pablo G, Radua J, Pereira J, Bonoldi I, Arienti V, Besana F, *et al.* (2021): Probability of Transition to Psychosis in Individuals at Clinical High Risk: An Updated Meta-analysis. *JAMA Psychiatry* 78: 970–978.

17. Howes OD, Whitehurst T, Shatalina E, Townsend L, Onwordi EC, Mak TLA, *et al.* (2021): The clinical significance of duration of untreated psychosis: an umbrella review and random-effects meta-analysis. *World Psychiatry* 20: 75–95.

18. Fusar-Poli P, Sullivan SA, Shah JL, Uhlhaas PJ (2019): Improving the Detection of Individuals at Clinical Risk for Psychosis in the Community, Primary and Secondary Care: An Integrated Evidence-Based Approach. *Front Psychiatry* 10: 774.

19. Shah JL, Crawford A, Mustafa SS, Iyer SN, Joober R, Malla AK (2017): Is the Clinical High-Risk State a Valid Concept? Retrospective Examination in a First-Episode Psychosis Sample. *Psychiatr Serv Wash DC* 68: 1046–1052.

20. Schultze-Lutter F, Rahman J, Ruhrmann S, Michel C, Schimmelmann BG, Maier W, Klosterkötter J (2015): Duration of unspecific prodromal and clinical high risk states, and early help-seeking in first-admission psychosis patients. *Soc Psychiatry Psychiatr Epidemiol* 50: 1831–1841.

21. McGorry PD, Edwards J, Mihalopoulos C, Harrigan SM, Jackson HJ (1996): EPPIC: an evolving system of early detection and optimal management. *Schizophr Bull* 22: 305–326.

22. Marshall M, Husain N, Bork N, Chaudhry IB, Lester H, Everard L, *et al.* (2014): Impact of early intervention services on duration of untreated psychosis: Data from the National EDEN prospective cohort study. *Schizophr Res* 159: 1–6.

23. Valmaggia LR, Byrne M, Day F, Broome MR, Johns L, Howes O, *et al.* (2015): Duration of untreated psychosis and need for admission in patients who engage with mental health services in the prodromal phase. *Br J Psychiatry J Ment Sci* 207: 130–134.

24. Srihari VH, Tek C, Pollard J, Zimmet S, Keat J, Cahill JD, *et al.* (2014): Reducing the duration of untreated psychosis and its impact in the U.S.: the STEP-ED study. *BMC Psychiatry* 14: 335.

25. Fusar-Poli P, Rutigliano G, Stahl D, Davies C, Bonoldi I, Reilly T, McGuire P (2017): Development and Validation of a Clinically Based Risk Calculator for the Transdiagnostic Prediction of Psychosis. *JAMA Psychiatry* 74: 493–500.

26. McGorry PD, Hartmann JA, Spooner R, Nelson B (2018): Beyond the "at risk mental state" concept: transitioning to transdiagnostic psychiatry. *World Psychiatry* 17: 133–142.

27. Fusar-Poli P, Schultze-Lutter F, Cappucciati M, Rutigliano G, Bonoldi I, Stahl D, *et al.* (2016): The Dark Side of the Moon: Meta-analytical Impact of Recruitment Strategies on Risk Enrichment in the Clinical High Risk State for Psychosis. *Schizophr Bull* 42: 732–743.

28. Falkenberg I, Valmaggia L, Byrnes M, Frascarelli M, Jones C, Rocchetti M, *et al.* (2015): Why are help-seeking subjects at ultra-high risk for psychosis help-seeking? *Psychiatry Res* 228: 808–815.

29. Oliver D, Arribas M, Radua J, Salazar de Pablo G, De Micheli A, Spada G, *et al.* (2022): Prognostic accuracy and clinical utility of psychometric instruments for individuals at clinical high-risk of psychosis: a systematic review and meta-analysis. *Mol Psychiatry* 27: 3670–3678.

30. Fusar-Poli P, Schultze-Lutter F, Addington J (2016): Intensive community outreach for those at ultra high risk of psychosis: dilution, not solution. *Lancet Psychiatry* 3: 18.

31. Fusar-Poli P (2017): Why ultra high risk criteria for psychosis prediction do not work well outside clinical samples and what to do about it. *World Psychiatry Off J World Psychiatr Assoc WPA* 16: 212–213.

32. Kosorok MR, Laber EB (2019): Precision Medicine. *Annu Rev Stat Its Appl* 6: 263–286.

33. Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW (2018): The Science of Prognosis in Psychiatry: A Review. *JAMA Psychiatry* 75: 1289–1297.

34. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G (2024): Developing clinical prediction models: a step-by-step guide. *BMJ* 386: e078276.

35. Steyerberg EW (2019): *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-16399-0

36. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, *et al.* (2024): TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385: e078378.

37. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, *et al.* (2013): Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 10: e1001381.

38. Meehan AJ, Lewis SJ, Fazel S, Fusar-Poli P, Steyerberg EW, Stahl D, Danese A (2022): Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Mol Psychiatry* 27: 2700–2708.

39. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research - PubMed (n.d.): Retrieved July 10, 2025, from https://pubmed.ncbi.nlm.nih.gov/23393429/

40. Ew S, Mj E, Jd H (1999): Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol* 52. https://doi.org/10.1016/s0895-4356(99)00103-1

41. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition | SpringerLink (n.d.): Retrieved July 10, 2025, from https://link.springer.com/book/10.1007/978-0-387-84858-7

42. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, *et al.* (2009): Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *The BMJ* 338: b2393.

43. Sisk R, Sperrin M, Peek N, van Smeden M, Martin GP (2023): Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Stat Methods Med Res* 32: 1461–1477.

44. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, *et al.* (2010): Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiol Camb Mass* 21: 128–138.

45. Vickers AJ, Calster BV, Steyerberg EW (2016): Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 352: i6.

46. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M (2021): External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* 14: 49–58.

47. Reilly BM, Evans AT (2006): Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 144: 201–209.

48. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM (2018): Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagn Progn Res* 2: 11.

49. Oliver D, Spada G, Colling C, Broadbent M, Baldwin H, Patel R, *et al.* (2021): Real-world implementation of precision psychiatry: Transdiagnostic risk calculator for the automatic detection of individuals at-risk of psychosis. *Schizophr Res* 227: 52–60.

50. Wang T, Oliver D, Msosa Y, Colling C, Spada G, Roguski Ł, *et al.* (2020): Implementation of a real-time psychosis risk detection and alerting system based on Electronic Health Records using CogStack. *J Vis Exp JoVE* 10.3791/60794.

51. R. D. Cox (1972): Regression Models and Life-Tables. *J R Stat Soc Ser B Methodol* Vol. 34, No. 2: 187-220.

52. Solmi M, Radua J, Olivola M, Croce E, Soardo L, Salazar de Pablo G, *et al.* (2022): Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Mol Psychiatry* 27: 281–295.

53. Fusar-Poli P, Werbeloff N, Rutigliano G, Oliver D, Davies C, Stahl D, *et al.* (2019): Transdiagnostic Risk Calculator for the Automatic Detection of Individuals at Risk and the Prediction of Psychosis: Second Replication in an Independent National Health Service Trust. *Schizophr Bull* 45: 562–570.

54. Oliver D, Wong CMJ, Bøg M, Jönsson L, Kinon BJ, Wehnert A, *et al.* (2020): Transdiagnostic individualized clinically-based risk calculator for the automatic detection of individuals

at-risk and the prediction of psychosis: external replication in 2,430,333 US patients. *Transl Psychiatry* 10: 364.

55. Irving J, Patel R, Oliver D, Colling C, Pritchard M, Broadbent M, *et al.* (2020): Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Schizophr Bull* 47: 405–414.

56. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, *et al.* (2017): Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* 7: e012012.

57. Nelson B, McGorry PD, Wichers M, Wigman JTW, Hartmann JA (2017): Moving From Static to Dynamic Models of the Onset of Mental Disorder: A Review. *JAMA Psychiatry* 74: 528–534.

58. Scheffer M, Bockting CL, Borsboom D, Cools R, Delecroix C, Hartmann JA, *et al.* (2024): A Dynamical Systems View of Psychiatric Disorders—Practical Implications: A Review. *JAMA Psychiatry* 81: 624–630.

59. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR (2010): Temporal pattern discovery in longitudinal electronic patient records. *Data Min Knowl Discov* 20: 361–387.

60. Le-Rademacher J, Wang X (2021): Time-To-Event Data: An Overview and Analysis Considerations. *J Thorac Oncol* 16: 1067–1074.

61. Houwelingen H van, Putter H (2011): *Dynamic Prediction in Clinical Survival Analysis*. CRC Press.

62. Wiegrebe S, Kopper P, Sonabend R, Bischl B, Bender A (2024): Deep learning for survival analysis: a review. *Artif Intell Rev* 57: 65.

63. Putter H (2014): Handbook of Survival Analysis - chapter 21 Landmarking. *Handbook of Survival Analysis*. Chapman and Hall/CRC.

64. Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Oudshoorn CGM (2018): Time-varying covariates and coefficients in Cox regression models. *Ann Transl Med* 6: 121.

65. Henderson R, Diggle P, Dobson A (2000): Joint modelling of longitudinal measurements and event time data. *Biostat Oxf Engl* 1: 465–480.

66. Wongvibulsin S, Wu KC, Zeger SL (2019): Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med Res Methodol* 20: 1.

67. Pakbin A, Wang X, Mortazavi BJ, Lee DKK (2023, September 6): BoXHED2.0: Scalable boosting of dynamic survival analysis [no. arXiv:2103.12591]. arXiv. https://doi.org/10.48550/arXiv.2103.12591

68. Lee C, Yoon J, Schaar M van der (2020): Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data. *IEEE Trans Biomed Eng* 67: 122–133.

69. Tanner KT, Sharples LD, Daniel RM, Keogh RH (2021): Dynamic Survival Prediction Combining Landmarking with a Machine Learning Ensemble: Methodology and Empirical Comparison. *J R Stat Soc Ser A Stat Soc* 184: 3–30.

70. Devaux A, Genuer R, Peres K, Proust-Lima C (2022): Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history: a landmark approach. *BMC Med Res Methodol* 22: 188.

71. Putter H, van Houwelingen HC (2017): Understanding Landmarking and Its Relation with Time-Dependent Cox Regression. *Stat Biosci* 9: 489–503.

72. Li W, Li L, Astor BC (2023): A Comparison of Two Approaches to Dynamic Prediction: Joint Modeling and Landmark Modeling. *Stat Med* 42: 2101–2115.

73. Rudin C (2019): Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1: 206–215.

74. Linardatos P, Papastefanopoulos V, Kotsiantis S (2021): Explainable AI: A Review of Machine Learning Interpretability Methods [no. 1]. *Entropy* 23: 18.

75. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B (2019): A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 110: 12–22.

2. Dynamic and transdiagnostic risk calculator based on Natural Language Processing for the prediction of psychosis in secondary mental health care: development and internal-external validation cohort study.

Published in Biological Psychiatry 2024
doi: 10.1016/j.biopsych.2024.05.022

Short title: Dynamic risk calculator for psychosis in EHRs

Kamil Krakowski[1,2], Dominic Oliver[2,3,4,5], Maite Arribas[2], Daniel Stahl*[6], Paolo Fusar-Poli*[1,2,7,8]
[1] Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy;
[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] Department of Psychiatry, University of Oxford, Oxford, UK;
[4] NIHR Oxford Health Biomedical Research Centre, Oxford, UK;
[5] OPEN Early Detection Service, Oxford Health NHS Foundation Trust, Oxford, UK;
[6] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, London, UK;
[7] OASIS Service, South London and the Maudsley National Health Service Foundation Trust, London, UK;
[8] Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University Munich, Germany.
*Daniel Sthal and Paolo Fusar-Poli are shared last authors

Abstract: 249/250
Word count: 3,990/4,000

Corresponding author: Paolo Fusar-Poli (paolo.fusar-poli@kcl.ac.uk)
Keywords: precision psychiatry, electronic health records, psychosis, prediction modelling, dynamic modelling, implementation

2.1 Abstract

Background: Automatic transdiagnostic risk calculators can improve detection of individuals at risk of psychosis. However, they rely on a single point in time assessment and can be refined with dynamic modelling techniques that account for changes in risk over time.

Methods: We included n=158,139 patients (n=5,007 events) receiving a first index diagnosis of a non-organic and non-psychotic mental disorder within Electronic Health Records from the SLaM NHS Foundation Trust between 01/01/2008 and 10/08/2021. A dynamic Cox landmark model was developed to estimate the 2-year risk of developing psychosis according to TRIPOD statement. The dynamic model included 24 predictors extracted at nine landmark points (baseline, 0, 6, 12, 24, 30, 36, 42, and 48 months): three demographic, one clinical, and 20 Natural Language Processing (NLP) based symptom and substance use predictors. Performance was compared to a static Cox regression model with all predictors assessed at baseline only, indexed via discrimination (C-index), calibration (calibration plots), and potential clinical utility (decision curves) in internal-external validation.

Results: The dynamic model improves discrimination performance compared to the static model at baseline (dynamic: C-index=0.9; static: C-index=0.87) to the final landmark point (dynamic: C-index=0.79; static: C-index=0.76). The dynamic model was also significantly better calibrated (calibration slope=0.97-1.1) than the static model at later landmark points (≥24 months). Net benefit was higher in the dynamic compared to the static model at later landmark points (≥24 months).

Conclusion: These findings suggest that dynamic prediction models can improve detection of individuals at risk for psychosis in secondary mental health care.

2.2 Introduction

Psychosis is a substantial cause of disability, reduced life expectancy, and subjective burden worldwide (1–3). Despite the associated acute personal and societal costs, current treatments are not fully effective for all established cases, particularly if they are delayed (4,5). Primary indicated prevention, as implemented in the Clinical High Risk for Psychosis (CHR-P) state (6–10), has the potential to benefit patients' outcomes by ameliorating symptoms, preventing psychosis onset and shortening the duration of untreated psychosis (11–14).

For effective preventive care, the first critical step is the detection of individuals at risk of psychosis; however, current approaches are ineffective, with only a minority of individuals identified prior to psychosis onset (15,16). To address this, a transdiagnostic risk calculator was originally developed and validated to screen Electronic Health Records (EHRs) and detect those at risk in secondary mental health care by predicting psychosis risk within six years (17). The calculator used predictors routinely collected in clinical practice, achieving good discrimination performance (Harrell's C-index=0.79). The transdiagnostic risk calculator was further externally validated in different settings, maintaining good discrimination ability (C-index=0.68-0.79) (18–20). Also, the feasibility of the real-world implementation of the transdiagnostic risk calculator was demonstrated, a very rare instance in precision psychiatry research (21), highlighting its potential for improvement in CHR-P detection (22). More recently, the transdiagnostic risk calculator was further refined using more fine-grained symptoms and substance use predictors extracted by advanced and automatic data-mining apps based on Natural Language Processing (NLP) algorithms that extract information from the free text stored in the EHRs clinical notes and letters (23). The NLP-enhanced model achieved substantial improvement in discrimination ability (C-index=0.85).

The transdiagnostic risk calculator (17,23), as well as most clinical prediction models in early psychosis research (21,24,25), estimate the probability of psychosis using only the information collected at a single point in time (baseline). However, prior to psychosis onset, signs and symptoms evolve dynamically over time (26,27), and recent research (28) has recommended including changes in clinical information to dynamically update the prognosis as time progresses. The previously developed transdiagnostic risk calculators are based on a Cox regression model and do not accommodate time-varying data (17,23). In this study, we refine this approach by employing a landmark Cox regression model (29). This enables the incorporation of dynamic, time-dependent predictors, thereby providing a more detailed and temporally sensitive analysis of time-to-event data. Unlike joint models (30) that have been used in the CHR-P populations to predict psychosis (31–33), landmark models are less computationally demanding, allowing handling large datasets with numerous time-varying covariates. Studies have shown that the Cox landmark models often perform similarly or better compared to joint models (34). Also, in contrast to machine learning models (35), Cox landmark models offer a transparent and interpretable framework which facilitates acceptance among clinicians and patients *(36)* although there has been progress in recent years in demystifying ML/AI algorithms through the development of explainable AI methods (37).

In this study we aim to refine the existing risk calculator (23) by combining longitudinal EHRs, NLP data and a Cox landmark model to develop and internally-externally validate a dynamic transdiagnostic risk calculator of psychosis that updates risk estimates over time. We also statistically compared this model to a static version to directly quantify the potential performance benefits of a dynamic over a static approach throughout the course of follow-up.

2.3 Methods

The model development, internal-external validation, and reporting were performed following the TRIPOD guidelines (33) (eTable 2.8.1).

2.3.1 Setting and study population

The retrospective cohort study population was extracted from the EHRs at the South London and Maudsley NHS Foundation Trust (SLaM). SLaM is one of the largest secondary mental health care providers in Europe. The catchment area of the Trust consists of four socioeconomically diverse boroughs: Croydon, Lambeth, Lewisham, and Southwark, as well as tertiary referrals from other parts of London. Clinical Record Interactive Search (CRIS) was implemented in SLaM's EHR to facilitate research with full but anonymised clinical information (39) and has been extensively validated (40–42). CRIS received ethical approval as an anonymised dataset for secondary analyses from Oxfordshire REC C (Ref: 23/SC/0257). Incidence of psychosis in SLaM (from 58.3 to 71.9 cases per 100,000 person-years) (43,44) is one of the highest worldwide (45). The inclusion criteria for model development and validation were the assignment of the first (index) diagnosis of non-organic and non-psychotic mental disorder within SLaM services between 1st January 2018, and 10th August 2021. Patients were followed for a maximum time of 6 years and administratively censored at that time. The loss to follow-up time (censoring before 6 years) has been defined as the date of the last entry in the EHRs for a given patient.

2.3.2 Outcome

The main outcome was defined as the time until diagnosis of a non-organic psychotic disorder as defined in the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10) (eMethods 2.8.1).

2.3.3 Predictors

We extracted the following predictors: (i) demographic predictors: age at index date, sex, self-assigned ethnicity (eTable 2.8.5); (ii) clinical predictor: index ICD-10 diagnostic spectra (eTable 2.8.8); and (iii) NLP-based predictors: symptoms and substance use information extracted by text mining algorithms (eMethods 2.8.4, eTable 2.8.7). The static model included the predictors (i-ii-iii) extracted at baseline only. The dynamic model (Figure 1) included the predictors (i-ii-iii) extracted at baseline and different follow-up times, resulting in a maximum of nine landmark

points per patient (baseline, 6, 12, 18, 24, 30, 36, 42 and 48 months). NLP predictors were measured twofold: (a) as a sum of occurrences of each NLP predictor in the six months preceding the landmark point, representing the patient's current state (s, Figure 1) and (b) as the cumulative average of occurrences of each NLP predictor up to the landmark point (46,47), representing patients' histories (Figure 1). NLP predictors were included if they performed with a minimum of 80% precision threshold that had been used before (23) at the time of the data extraction. However, due to routine updates in the available NLP applications, we have used an expanded set of NLP predictors compared to our previous paper (23) (eTable 2.8.6). NLP predictors included agitation, appetite loss, blunted affect, cannabis use, cocaine use, delusional thinking, disturbed sleep, guilt, hallucinations (any), hallucinations (auditory), hallucinations (olfactory-gustatory-tactile), hallucinations (visual), hopelessness, insomnia, irritability, negative symptoms, paranoia, poverty of speech, tearfulness, weight loss. Notably, these predictors only index the presence or absence of a feature, regardless of its severity or frequency. This means that while positive psychotic symptoms are present, they may not be experienced at the severity or frequency required to meet criteria for an ICD-10 diagnosis (eMethods 2.8.2, eFigure 2.8.2-2.8.4).

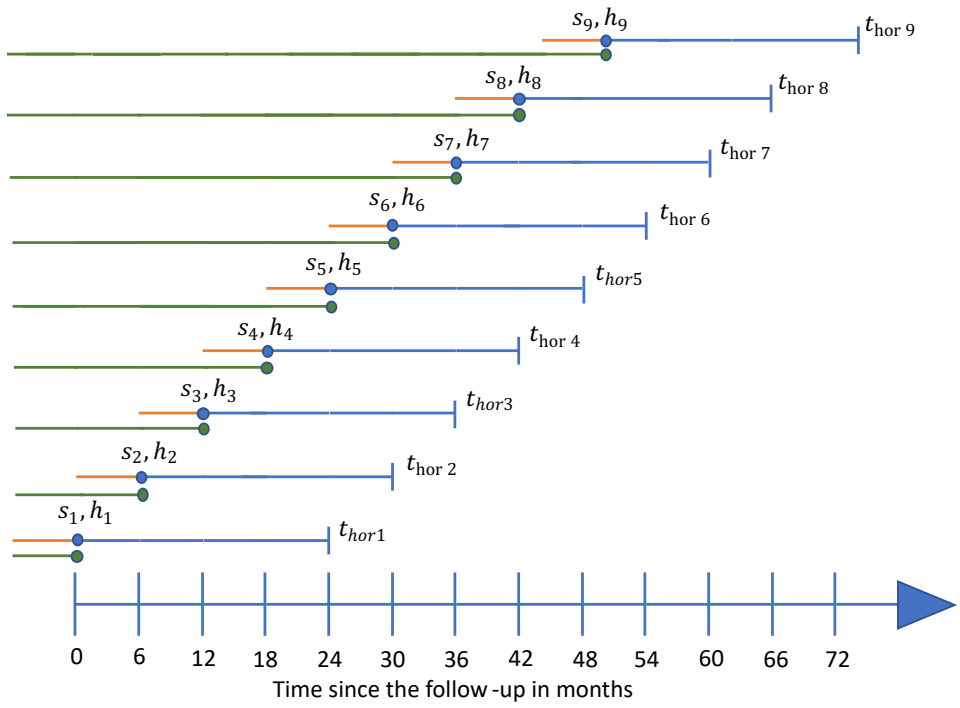


Figure 1 Distribution of the time-varying parameters in the dynamic model. The orange lines represent patients' current state(s) defined as the sum of the occurrences of each predictor in the six months preceding each landmark point. The green lines show the period at which the cumulative average of occurrences was obtained, representing patients' histories (h). The blue lines represent the prediction time horizons t_{hor} .

2.3.4 Missing data

For predictors with missing data (age, gender, ethnicity) a single imputation was performed using random forests within the Multivariate Imputation with Chain Equations framework (48) following the recommendations of Sisk et al. (49) (eMethod 8).

2.3.5 Statistical analysis

The static model was based on the Cox proportional hazards model (49), which examines the relationship between the time until an event of interest and predictor variables recorded at baseline. In this study, we extended the static model into a dynamic model by adopting a landmarking approach that is suitable for dynamically updating prognosis at predefined landmark time points throughout follow-up (29). The landmarking model analysis was performed in two stages. In the first stage, participant data from all landmark points were merged into a single 'stacked' dataset using a four-step approach (29):

1. Fixing the prediction window v;
2. Selecting a set of landmarking time points to make predictions at $\{s_{1,} \dots, s_n\}$;
3. Creating a data set for each landmarking time point s by truncation (keep only those still at risk at time s_n) and administrative censoring at time horizon $t_{hor\,n} = s_n + v$;
4. Stacking vertically all data sets created in point 3 into the stacked data set.

The landmark points $\{s_{1,} \dots, s_n\}$ were selected at 6-month intervals, starting with s_1 at baseline and ending with s_9 at 48 months since the start of follow-up (Figure 1). Next, we defined a time window for the predictions denoted as v, which was set at two years.
In the second stage, a single Cox regression model stratified by baseline hazards at each sliding window between s_1 and s_9 is fitted on the 'stacked' data set (29). According to Houwelingen (51) the Cox landmark model used can be called a 'supermodel' (51) and can be expressed as:

$$h_i(t|X_i, Y_i(s), s, v) = h_{0,s}(\mathrm{t|s})\exp\left(\gamma^\top X_i + \alpha Y_i(s)\right)$$

where $h_{0,s}(\mathrm{t|s})$ is a baseline hazard stratified on landmark s; v is the period between landmark point and time horizon; X_i is a vector of variables obtained at the baseline for an individual i; $Y_i(s)$ is a vector of dynamic variables depending on s for an individual i; γ is a parameter vector associated with the baseline variables X_i. This vector quantifies the impact of the baseline variables on the hazard (or risk) of the event α is a parameter vector associated with the dynamic variables $Y_i(s)$. α quantifies the impact of the dynamic variables on the hazard. The parameters γ and α are the same for all landmark points. Parameters are estimated through partial likelihood maximization (30).

2.3.6 Model development and internal-external validation.

The study population consists of patients from five different catchment areas – four London boroughs: Croydon, Lambeth, Lewisham, Southwark and tertiary referrals from the rest of London (Other). The internal-external validation (52) of both the static and dynamic models was performed by splitting the data by boroughs, generating 5 train-test splits (eFigure 2.8.1). One at a time, each borough was executed, developing the model on the four others, then evaluating it in the single excluded borough ('leave one borough out cross-validation). The final model was fitted on the pooled dataset (38), providing coefficients for future external validation and clinical applications.

2.3.7 Performance evaluation

Discrimination. The ability of a prediction model to distinguish between different classes is defined as discrimination (53,54). Harrell's C is the recommended index for assessing the discriminative ability of prediction models in the context of survival analysis (55). To evaluate the performance of the dynamic model over time, the Harrell's C index is calculated at every landmark point $\{s_1, \dots, s_9\}$ over a two-year horizon, considering only individuals still at risk at the specific landmark points.

Calibration. The agreement between the observed and predicted outcomes was evaluated and quantified by the coefficients of the logistic recalibration framework model (56). Slopes of the logistic recalibration framework model were evaluated at every landmark point for consistency with discrimination assessment. Additionally, the calibration curves using lowess regressions (56) were visually examined at baseline, 24 and 48 months landmark points.

Pooling the internal-external validation results at each time point. The internal-external validation based on the five geographical splits resulted in five C-indices and calibration coefficient values and their corresponding standard errors at each landmark point. Pooled average estimates with confidence intervals for each landmark point were obtained by random effects meta-analyses combining the five geographical splits (57). The Hartung-Knapp Sidik-Jonkma (58) method was used because the number of train-test splits (n = 5) was small.

Direct comparison between the dynamic and static models. To compare the performance of the static and dynamic models, we fitted a regular Cox proportional hazards model (50) using all predictors defined above and recorded exclusively at baseline, and outcome data recorded for up to 6 years. We evaluated the static model among patients still at risk at each landmark point using the 2-year prediction window v for consistency with the dynamic model. This was achieved in four steps:

1. Using the static model and values of predictors recorded at baseline generate prognosis at multiple time horizons $t_{hor\,n}$ (24, 30, 36, 42, 48, 54, 60, 66 and 72 months).
2. For each landmark point s_n keep only patients still at risk, and administratively censor at time horizon $t_{hor\,n} = s_n + v$;

3. For patients selected at each landmark point s_n keep only prognosis corresponding to the time horizon $t_{hor\,n} = s_n + v$;
4. Using data obtained in points 2 and 3 compute C-index and calibration coefficient at each landmark point.

The static model was evaluated using the same internal-external validation procedure and results pooling method as the dynamic model.

Meta-regression for quantifying the performance difference. Evaluation of both the static and dynamic models at each of the 9 landmark points and 5 train-test splits resulted in 90 C-index values and their corresponding standard errors. To quantify the difference in discrimination performance C-index between the two models, a multilevel meta-regression was conducted (48). Three dependant variables were included: borough, landmark time and model type to explain the C-index (eMethods 2.8.5). The coefficient of the model type variable was interpreted as the mean difference between the two models, summarizing the results across various landmark points and train-test splits.

Potential Clinical Utility for Psychosis Prevention. The potential clinical utility of the risk calculators was assessed using decision curve analysis, which quantifies the model's expected relative net benefit across different classification thresholds compared to the default strategies of treating all or no patients (59). Net benefit in decision curve analysis is the quantitative trade-off between the benefit of true positive predictions and the harm of false positive predictions at a particular threshold probability. Net benefit was calculated and presented visually for the range of risk thresholds between 0 and 0.6 using the same internal-external validation approach as for discrimination and calibration. For both the static and the dynamic model, results from the five test sets were combined, and decision curves (59) at three landmark points (baseline, 24 and 48 months) were presented.

Mean absolute error of survival time was calculated and presented in eMethods 2.8.7. To demonstrate that the model is useful if the ethnicity variable is not available, we reran the model without it (eMethods 2.8.6).

2.4 Results

2.4.1 Participants

We included 158,139 individuals receiving a first index diagnosis of a non-organic and non-psychotic mental disorder (eTable 2.8.4). Table 1 shows the characteristics of the overall study population and those of each borough separately. The final study cohort had an average age of 33.95 years (SD = 19.19), with females constituting about half the population (48%) and a predominance of self-identified white ethnicity (66%). The most common diagnoses were anxiety disorders and non-bipolar mood disorders (both 25%). The boroughs had similar gender and age distribution but differed in self-assigned ethnicity. Other boroughs had a higher

proportion of white ethnicity and a lower proportion of black ethnicity compared to the remaining four boroughs. Additionally, this group also had a lower proportion of non-bipolar mood disorders (14%) and a higher proportion of individuals with substance use disorders (28%). The number of events was 5,007, and the mean (SD) follow-up period was 1,510 (705) days. The cumulative risk of psychosis at 2 years was 0.031 (95% CI, 0.03-0.033). The Other boroughs had a significantly lower cumulative risk 0.015 (95% CI, 0.014-0.017) than the reaming four.

Table 1 Sample size, outcomes and predictors of the study population and stratified by borough.

Variables	No. (proportion [%]) All boroughs	Croydon	Lambeth	Lewisham	Other	Southwark	Differences between boroughs (test, p value)
Sample size							
Number of patients	153139	26362	27192	25154	53663	25768	
Outcome							
Psychosis	5007 (3.17)	1036 (4.01)	1043 (4)	966 (3.99)	933 (1.77)	1029 (4.16)	χ^2 = 540, p < .001
Predictors							
Sociodemographic							
Gender							χ^2 = 620, p < .001
Male	81689 (51.7)	14119 (53.6)	13669 (50.3)	13933 (55.4)	21119 (47.9)	14233 (55.2)	
Female	76355 (48.3)	12234 (46.4)	13515 (49.7)	11211 (44.6)	22938 (52)	11529 (44.7)	
Age, mean (SD)	33.95 (19.19)	33.31 (21.12)	35.26 (19.19)	33.12 (20.33)	33.79 (16.87)	34.33 (19.62)	F-test = 54, p<.001
Self-assigned ethnicity							χ^2>999, p < .001
White	88416 (55.9)	14758 (56)	13944 (51.3)	13722 (54.6)	27889 (63.2)	12850 (49.9)	
Black	23074 (14.6)	3669 (13.9)	5674 (20.9)	4929 (19.6)	2900 (6.6)	4492 (17.4)	
Asian	7768 (4.9)	1999 (7.6)	1151 (4.2)	1267 (5)	1782 (4)	1106 (4.3)	
Mixed	5881 (3.7)	1202 (4.6)	1295 (4.8)	1202 (4.8)	1007 (2.3)	908 (3.5)	
Other	8585 (5.4)	1011 (3.8)	2127 (7.8)	955 (3.8)	1344 (3)	2722 (10.6)	
Clinical							
ICD-10 diagnostic spectra							χ^2 > 999, p < .001
Acute and transient psychotic disorders	2796 (1.8)	458 (1.7)	640 (2.4)	533 (2.1)	498 (1.1)	439 (1.7)	
Anxiety disorders	39459 (25)	6771 (25.7)	5823 (21.4)	7255 (28.8)	10749 (24.4)	6487 (25.2)	
At Risk Mental State (ARMS)	446 (0.3)	44 (0.2)	153 (0.6)	70 (0.3)	23 (0.1)	124 (0.5)	
Bipolar mood disorders	5232 (3.3)	947 (3.6)	1009 (3.7)	839 (3.3)	1133 (2.6)	873 (3.4)	
Childhood/adolescence onset disorders	20729 (13.1)	3963 (15)	3315 (12.2)	3793 (15.1)	6070 (13.8)	2815 (10.9)	
Developmental disorders	8440 (5.3)	1828 (6.9)	908 (3.3)	1146 (4.6)	3222 (7.3)	1028 (4)	
Mental retardation	2548 (1.6)	517 (2)	509 (1.9)	628 (2.5)	262 (0.6)	416 (1.6)	
Non bipolar mood disorders	39646 (25.1)	7739 (29.4)	6613 (24.3)	7617 (30.3)	6112 (13.9)	8709 (33.8)	
Personality disorders	5937 (3.8)	962 (3.6)	928 (3.4)	969 (3.9)	1632 (3.7)	932 (3.6)	

Physiological syndromes	9335 (5.9)	925 (3.5)	1720 (6.3)	1125 (4.5)	4035 (9.1)	1316 (5.1)	
Substance use disorders	23571 (14.9)	2208 (8.4)	5574 (20.5)	1179 (4.7)	10368 (23.5)	2629 (10.2)	
Natural Language Processing (NLP)							
Agitation	26269 (18)	5069 (19.23)	4206 (15.47)	4069 (16.18)	8462 (19.19)	4463 (17.32)	χ^2 = 172, p < .001
Appetite loss	9523 (6)	1996 (7.57)	1652 (6.08)	1862 (7.4)	1968 (4.46)	2045 (7.94)	χ^2 = 429, p < .001
Blunted affect	5625 (4)	985 (3.74)	1138 (4.19)	928 (3.69)	1341 (3.04)	1233 (4.79)	χ^2 = 139, p < .001
Cannabis use	26185 (18)	4035 (15.31)	4776 (17.56)	4037 (16.05)	9104 (20.65)	4233 (16.43)	χ^2 = 307, p < .001
Cocaine use	19223 (13)	2280 (8.65)	4019 (14.78)	2195 (8.73)	7791 (17.67)	2938 (11.4)	χ^2 > 999, p < .001
Delusional thinking	8987 (6)	1840 (6.98)	1623 (5.97)	1657 (6.59)	2285 (5.18)	1582 (6.14)	χ^2 = 99, p < .001
Disturbed sleep	54068 (36)	10558 (40.05)	9850 (36.23)	9999 (39.75)	13028 (29.54)	10633 (41.27)	χ^2 = 683, p < .001
Guilt	21078 (14)	3496 (13.26)	3666 (13.48)	3540 (14.07)	6124 (13.89)	4252 (16.5)	χ^2 = 109, p < .001
Hopelessness	22473 (15)	3855 (14.62)	3811 (14.02)	4199 (16.69)	6082 (13.79)	4526 (17.57)	χ^2 = 190, p < .001
Hallucinations (all)	15147 (1)	2911 (11.04)	2603 (9.57)	2617 (10.4)	4090 (9.28)	2926 (11.36)	χ^2 = 91, p < .001
Hallucinations (auditory)	7956 (5)	1586 (6.02)	1394 (5.13)	1387 (5.51)	2134 (4.84)	1455 (5.65)	χ^2 = 49, p < .001
Hallucinations (olfactory, gustatory & tactile)	920 (1)	167 (0.63)	178 (0.65)	160 (0.64)	231 (0.52)	184 (0.71)	χ^2 = 11, p = .028
Hallucinations (visual)	4968 (3)	985 (3.74)	866 (3.18)	856 (3.4)	1375 (3.12)	88 (3.44)	χ^2 = 21, p < .001
Insomnia	11154 (8)	2070 (7.85)	1931 (7.1)	1768 (7.03)	3287 (7.45)	2098 (8.14)	χ^2 = 30, p < .001
Irritability	19091 (13)	3518 (13.35)	3124 (11.49)	3459 (13.75)	5472 (12.41)	3518 (13.65)	χ^2 = 71, p < .001
Negative symptoms	36594 (25)	7065 (26.8)	6570 (24.16)	6555 (26.06)	9186 (20.83)	7218 (28.02)	χ^2 = 366, p < .001
Paranoia	23727 (16)	4384 (16.63)	4280 (15.74)	4007 (15.93)	6631 (15.04)	4425 (17.18)	χ^2 = 48, p < .001
Poverty of speech	754 (1)	213 (0.81)	160 (0.59)	127 (0.5)	126 (0.29)	128 (0.5)	χ^2 =93, p < .001
Weight loss	16998 (11)	3044 (11.55)	2983 (10.97)	2896 (11.51)	4954 (11.23)	3121 (12.11)	χ^2 = 16, p < .004
Tearfulness	38958 (26)	7564 (28.7)	6726 (24.74)	7391 (29.39)	9220 (20.91)	8057 (31.27)	χ^2 = 724, p < .001

2.4.2 Model Specification

Table 2 presents the hazard ratios (eTable 2.8.3 presents coefficients) of the Cox landmark model fitted on the entire dataset. For features' importance the results of ANOVA Wald test (proportion of total χ^2 explained by each variable) is also presented in Table 2. The regression coefficients remain constant for the entire duration of the analyses, but the impact of the predictors on the hazard rate may vary as time progresses. The regression coefficients in the model can be interpreted as weighted average of the time-varying covariate effects (30). The values of landmark stratified baseline hazards are presented in the supplementary (eTable 2.8.2).

Diagnosis variable differentiates the estimated risk the most and corresponds to the greatest improvement in fit (χ^2 proportion). The individuals with acute and transient psychotic disorders have over 50 times higher risk than patients with developmental disorders. Ethnicity also played a significant role, with Black, Asian, mixed, and other ethnic groups showing a higher risk when compared to the white ethnicity. Female gender is a protective factor (HR = 0.91). Among the NLP-derived predictors, the presence of poverty of speech (HR = 1.4), hallucinations (olfactory, gustatory & tactile) (HR = 1.25), hallucinations (auditory) (HR = 1.17), and paranoia (HR = 1.13) produced the highest risk increase. Among the cumulative averages of the NLP variables, paranoia corresponds to the highest hazard (HR = 1.25), also paranoia in both currant state NLPs and cumulative averages resulted in the highest improvement in fit.

Table 2 The hazard ratios corresponding to the coefficients of the final model fitted on the entire dataset. The coefficient can be interpreted as some weighted average of the time-varying covariate effects. The proportion of explained total χ^2 is a measure of relative variable importance for the model fit.

Predictor	**Hazard Ratio (95% CI)**	***P* Value**	***Proportion of total*** χ^2[%]
Sociodemographic predictors			
Gender			0.17
Male	1 [Reference]	NA	
Female	0.912 (0.876 - 0.949)	<.001	
Age	1.002 (1.001 - 1.003)	<.001	0.1
Self-assigned ethnicity			12.71
White	1 [Reference]	NA	
Black	2.33 (2.23 - 2.44)	<.001	
Asian	1.36 (1.22 - 1.51)	<.001	
Mixed	1.84 (1.72 - 1.98)	<.001	
Other	1.17 (1.08 - 1.27)	<.001	
Clinical predictors			
Diagnosis			85.3
Acute and transient psychotic disorders	1 [Reference]	NA	
Anxiety disorders	0.07 (0.06 - 0.07)	<.001	
At Risk Mental State (ARMS)	0.32 (0.27 - 0.37)	<.001	

Bipolar mood disorders	0.25 (0.24 - 0.27)	<.001	
Childhood/adolescence onset disorders	0.02 (0.02 - 0.03)	<.001	
Developmental disorders	0.02 (0.02 - 0.03)	<.001	
Mental retardation	0.05 (0.04 - 0.06)	<.001	
Non bipolar mood disorders	0.09 (0.09 - 0.10)	<.001	
Personality disorders	0.10 (0.09 - 0.11)	<.001	
Physiological syndromes	0.03 (0.03 - 0.04)	<.001	
Substance use disorders	0.08 (0.07 - 0.08)	<.001	
Natural Language Processing (NLP) predictors			
Appetite loss	0.93 (0.84 - 1.03)	0,17	0.02
Agitation	1.06 (1.00 - 1.12)	0,042	0.03
Blunted affect	1.02 (0.91 - 1.14)	0,741	<0.01
Cannabis uses	1.01 (0.96 - 1.07)	0,63	<0.01
Cocaine use	0.92 (0.86 - 0.99)	0,024	0.04
Delusional thinking	1.10 (1.03 - 1.17)	0,004	0.07
Disturbed sleep	1.00 (0.96 - 1.05)	0.99	<0.01
Guilt	0.90 (0.83 - 0.97)	0.005	0.07
Hopelessness	0.95 (0.88 - 1.02)	0,164	0.02
Hallucinations (all)	1.07 (0.98 - 1.16)	0.122	0.02
Hallucinations (auditory)	1.17 (1.07 - 1.27)	0.001	0.1
Hallucinations (olfactory, gustatory & tactile)	1.25 (1.04 - 1.50)	0.015	0.05
Hallucinations (visual)	1.04 (0.94 - 1.15)	0.497	<0.01
Insomnia	0.94 (0.86 - 1.04)	0.234	0.01
Irritability	1.05 (0.99 - 1.11)	0.102	0.02
Negative symptoms	1.02 (0.96 - 1.08)	0.495	<0.01
Paranoia	1.13 (1.07 - 1.18)	<.001	0.2
Poverty of speech	1.40 (1.13 - 1.74)	0.002	0.08
Weight loss	0.97 (0.90 - 1.05)	0.404	0.01
Tearfulness	0.94 (0.89 - 0.99)	0.02	0.04
Cumulative agitation	1.02 (0.96 - 1.09)	0.466	<0.01
Cumulative appetite loss	1.18 (1.05 - 1.33)	0.007	0.06
Cumulative blunted affect	0.95 (0.83 - 1.08)	0,4	0.01
Cumulative cannabis use	1.03 (0.96 - 1.09)	0,412	0.01
Cumulative cocaine use	0.98 (0.91 - 1.06)	0.627	<0.01
Cumulative delusional thinking	1.08 (1.01 - 1.17)	0,035	0.04
Cumulative disturbed sleep	1.00 (0.95 - 1.06)	0.939	<0.01
Cumulative guilt	1.08 (0.99 - 1.18)	0.086	0.02
Cumulative hopelessness	1.03 (0.95 - 1.13)	0,47	<0.01
Cumulative hallucinations (any)	1.12 (1.02 - 1.23)	0.02	0.01
Cumulative hallucinations (auditory)	0.96 (0.86 - 1.06)	0.386	<0.01
Cumulative hallucinations (olfactory, gustatory & tactile)	0.89 (0.72 - 1.10)	0.296	<0.01
Cumulative hallucinations (visual)	1.04 (0.92 - 1.16)	0.539	<0.01
Cumulative insomnia	1.14 (1.03 - 1.27)	0.016	0.01
Cumulative irritability	0.99 (0.93 - 1.06)	0.788	<0.01

Cumulative negative symptoms	1.00 (0.94 - 1.07)	0.95	<0.01
Cumulative paranoia	1.25 (1.18 - 1.33)	<.001	0.49
Cumulative poverty of speech	0.65 (0.51 - 0.83)	0.001	0.1
Cumulative weight loss	1.17 (1.07 - 1.28)	0.001	0.1
Cumulative tearfulness	1.04 (0.97 - 1.10)	0.269	<0.01

2.4.3 Model Performance

In Figure 2, the discrimination indices (C-indices) for each hold-out borough and the pooled index at each landmark are presented for both the dynamic Cox landmark model and the static Cox model. The pooled and train-test split specific C-indices at each landmark for the static and dynamic model are presented in Figure 2. The static model obtained pooled C-indices ranging from 0.87 (95% CI = 0.85 – 0.88) at baseline to 0.76 (95% CI = 0.74 – 0.78) at 48 months. The dynamic model obtained the pooled C indices ranging from 0.89 (95% CI = 0.88 – 0.91) at baseline and gradually decreasing to 0.79 (95% CI = 0.76 - 0.82) at the last landmark point (48 months). The multilevel meta-regression revealed an average improvement in C-index of 0.035(95% CI = 0.031 – 0.043, $p < 0.001$) for the dynamic model compared to the static model over the follow-up period.

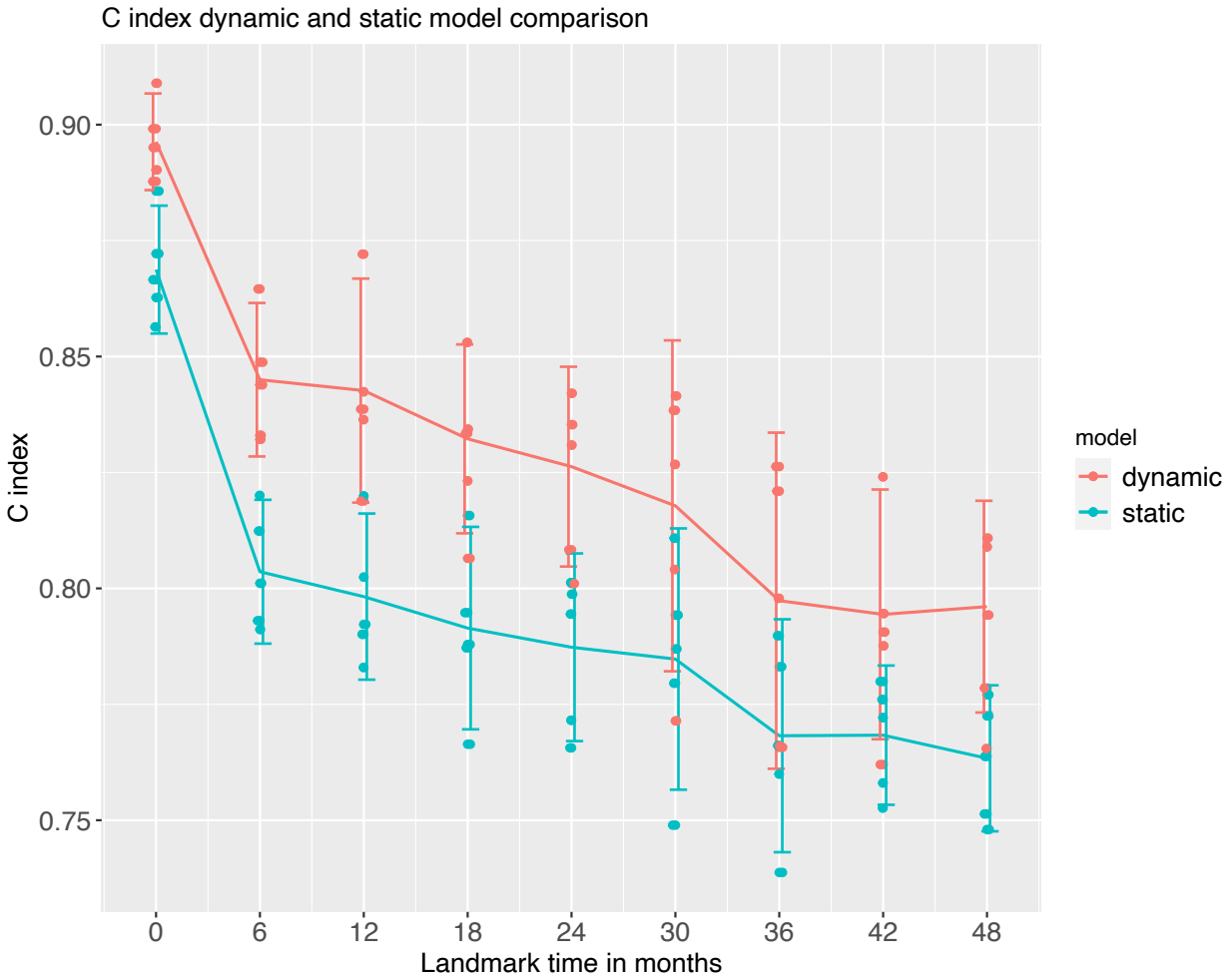


Figure 2 Discrimination (C-index) results for all train-test splits and evaluated at each landmark point in the internal-external cross-validation. The landmark points represent the time at which predictions are made. The error bars represent the 95% CI for the C-indices pooled at each landmark point. Dots indicate borough-specific C-indices. The dynamic model has on average larger C-index, but the variance of the dynamic model increases with time.

Calibration slope results are presented in Figure 3. A slope greater than 1 suggests an underestimation of high risk and an overestimation of low risk. Conversely, a slope less than 1 implies an underestimation of low risk and an overestimation of high risk. The dynamic model maintains a consistent and adequate calibration slope over the entire period, ranging from 1.1 (95% CI = 0.97 – 1.24) at baseline to 0.97 (95% CI 0.82 – 1.11) at the last landmark point. The static model shows a consistently decreasing calibration slope from 1.11 (95% CI 1.00 – 1.23) at baseline to 0.64 (95% CI 0.49 – 0.78) at the final landmark point, indicating significant miscalibration at later follow-ups. As indicated by the calibration plots (supplementary, eFigure 2.8.5 – 2.8.10), both models are adequately calibrated at the baseline; however, at later landmark points, the static model significantly overestimates the predicted risks (the calibration recalibration framework intercept below -1). At the 24- and 48-month landmark time, the

dynamic model is well calibrated for observed risks below 0.2 and for higher risks it overestimates the predictions.

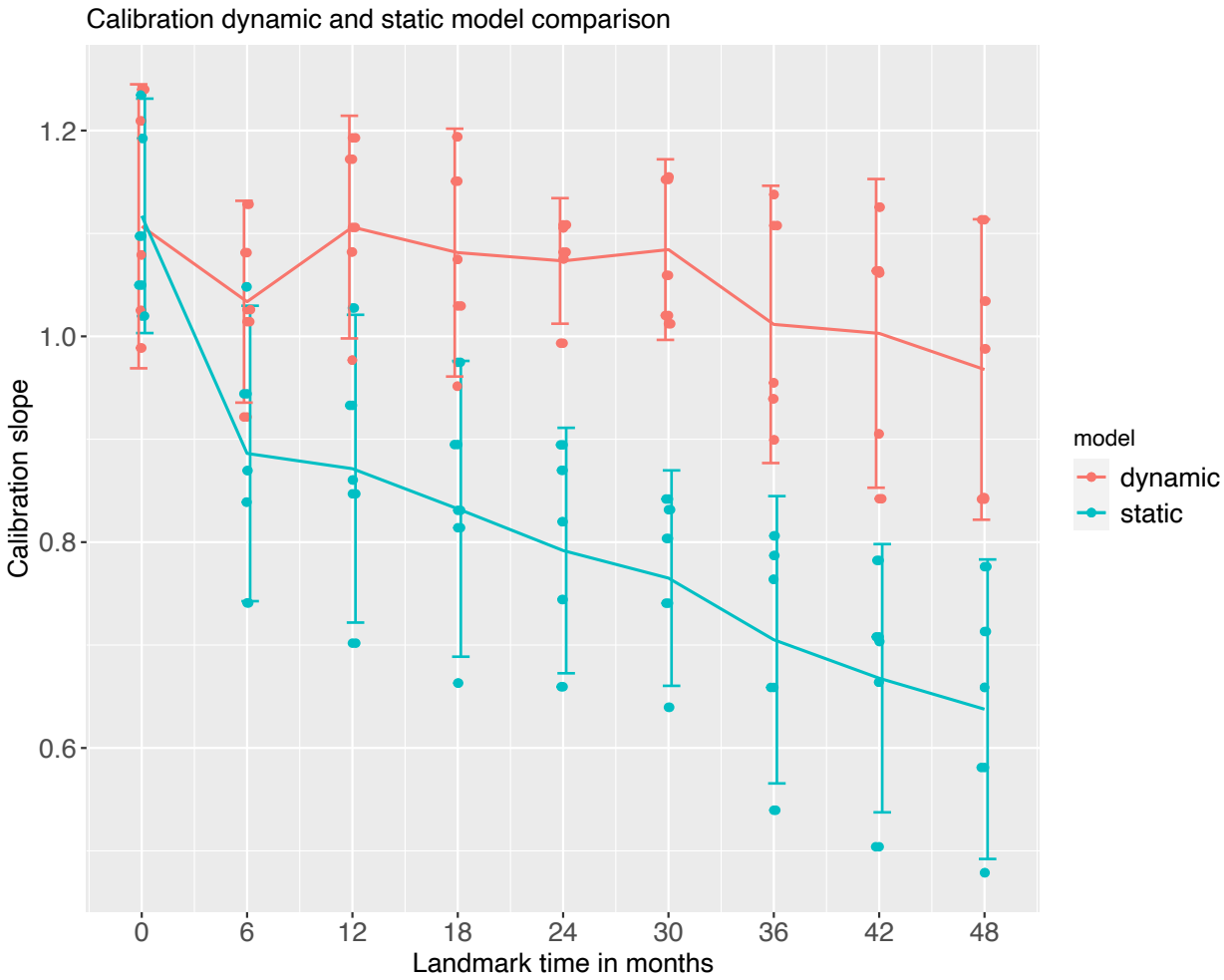


Figure 3 Calibration slopes computed for each train-test split and landmark point for the dynamic and static models in the internal-external cross validation. Lines represent pooled average of the calibration slopes with 95% confidence intervals at each landmark point. Dots indicate borough-specific calibration slopes.

The decision curve analysis (eFigure 2.8.11 – 2.8.13) was conducted at three landmark points (baseline, 24 and 48 months). At baseline, both models achieved very similar net benefits and are associated with larger net benefits compared to treating all or treating none at risk thresholds ranging from 1% to 69% (dynamic) and 71% (static). At later landmark points, the models' performance diverges. At 24 months, the dynamic model has positive net benefits between 1% and 44% threshold, and the static between 1% and 31%. At 48 months, the dynamic model is beneficial between 1% and 26% threshold, whereas the static model is beneficial only between 1% and 3%.

2.5 Discussion

This is the first study to show that a refined dynamic transdiagnostic model using the landmarking approach and NLP data can deliver a robust improvement in detecting individuals at risk of psychosis in secondary mental health care, compared to a static approach. Dynamic updating of risk significantly improves discrimination at all landmark points. The calibration and clinical utility improved past 24-month landmark point. It is significant as vast majority of at-risk individuals who later develop psychosis currently remains undetected (17).

The landmark model presented in this study is well suited for the automatic screening of individuals at risk for psychosis in EHRs. This represents a substantial advancement compared to joint modelling-based approaches developed among CHR-P populations (31–33), which used significantly smaller sample sizes (maximum n=488), employed predictors that are not routinely collected in clinical practice, and only one study used multiple longitudinal predictors (33). Together, this reduces the likelihood of generalisability of the model and increases the logistical and financial burden of data collection, which in turn impacts feasibility of routine clinical use (60). One other study used a case healthy-control data set to develop a dynamic prediction model for the first episode of psychosis in primary and secondary health care in the United States (35) using machine learning methods (gated recurrent neural networks) (61). The landmarking model achieved performance improvement over static model with a less complex and simpler to use approach than (35). Additionally, this model employs automated NLP predictors, and it is a refinement of one of the few prediction models ever tested for implementation and externally validated (eMethods 2.8.3) in this patient group (21,22).

This dynamic model shows good discrimination performance for 30 months post-baseline, contrasting with 6 months for the static model. The dynamic model also demonstrates superior calibration and clinical utility at later landmark points. At the 48-month landmark point, the dynamic model shows a positive net benefit for the Number Needed to Treat (NNT) between 100 and 4, whereas the static model is beneficial only for NNT between 100 and 33. Together, these aspects increase the potential for this model to be clinically valuable and extend the benefits of automated detection strategies in secondary mental health care (22).

Furthermore, this study implements several methodological innovations. Firstly, this is the first use of landmark modelling in psychosis research and the first dynamic model using NLP data in EHRs. Secondly, the presented model improves the performance of previous transdiagnostic risk calculators (17,23) by optimizing data usage through internal-external validation. This provides an estimate of model stability by reporting pooled results from all train-test splits rather than relying on a single partition. Lastly, this is the first use of meta-regression in psychosis research to quantify the benefit of dynamic modelling compared to static approaches.

This study had several limitations. Firstly, EHR data has high ecological validity, but the diagnoses, symptoms and substance use recorded in clinical notes are not psychometrically validated. However, meta-analyses suggest that EHR data are generally predictive of true validated diagnoses (62). Secondly, NLP tools are not able to extract data from free text with 100% precision. There will be inter-individual differences in how symptoms and substance use

are recorded in the notes, meaning that standardisation will not be high (63). We mitigated this issue by pre-selecting NLP algorithms for an adequate level of precision (≥80%). Thirdly, the discretisation of data to six-month intervals during model development may limit the accuracy of predictions made outside of those landmark points. Selecting the appropriate baseline hazard closest to the prediction date ensures the most accurate prediction possible using this model. Fourthly, currently, no guidelines exist for the use of dynamic prediction models in clinical practice. The clinical and ethical questions regarding the frequencies of updates and risk thresholds for triggering automated alerts require further research. Due to the limitations of the data (eLimitations), we were unable to ascertain severity of any symptoms experienced. In future clinical practice, this would need to be assessed with an in-depth clinical interview following any individual being detected as at-risk, as in previous work (22). Finally, even though the internal-external validation evaluates the performance in different London boroughs an independent external validation in different health care setting, i.e. with lower incidence rates, is a necessary further step to test generalisability.

An extension to the current methodology could involve integrating linear mixed models (47), employing machine learning techniques specific to landmark analysis (34,64) or different machine learning based longitudinal methods (65). Investigating multiple outcomes such as transition to schizophrenia or affective psychotic disorders and different severe mental disorders may also be clinically useful extensions of the model. Further work in assessing model performance in sub-populations (e.g. different ethnic groups) should be performed to investigate and mitigate potential algorithmic biases (66,67).

2.6 Conclusions

This is the first study demonstrating improved performance through dynamic modelling employing NLP in EHRs compared to static approaches. This contributes novel evidence supporting the refining of detection strategies for individuals at risk for psychosis in secondary mental health care using dynamic prediction models. Future research should explore complementary techniques for modelling longitudinal data, addressing missing data, and conducting implementation research.

Acknowledgments
MA is supported by the UK Medical Research Council (MR/N013700/1) and King's College London member of the MRC Doctoral Training Partnership in Biomedical Sciences. DS and PFP were part-funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Disclosures
MA has been employed by F. Hoffmann-La Roche AG outside of the current study. PFP has received research fees from Lundbeck and received honoraria from Lundbeck, Angelini,

Menarini and Boehringer Ingelheim outside of the current study. All other authors report no potential conflicts of interest.

2.7 References

1. Charlson FJ, Ferrari AJ, Santomauro DF, Diminic S, Stockings E, Scott JG, *et al.* (2018): Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophr Bull* 44: 1195–1203.

2. Fusar-Poli P, Estradé A, Stanghellini G, Venables J, Onwumere J, Messas G, *et al.* (2022): The lived experience of psychosis: a bottom-up review co-written by experts by experience and academics. *World Psychiatry Off J World Psychiatr Assoc WPA* 21: 168–188.

3. Estradé A, Onwumere J, Venables J, Gilardi L, Cabrera A, Rico J, *et al.* (2023): The Lived Experiences of Family Members and Carers of People with Psychosis: A Bottom-Up Review Co-Written by Experts by Experience and Academics. *Psychopathology* 1–12.

4. Millan MJ, Andrieux A, Bartzokis G, Cadenhead K, Dazzan P, Fusar-Poli P, *et al.* (2016): Altering the course of schizophrenia: progress and perspectives [no. 7]. *Nat Rev Drug Discov* 15: 485–515.

5. Fusar-Poli P, McGorry PD, Kane JM (2017): Improving outcomes of first-episode psychosis: an overview. *World Psychiatry Off J World Psychiatr Assoc WPA* 16: 251–265.

6. Salazar de Pablo G, Estradé A, Cutroni M, Andlauer O, Fusar-Poli P (2021): Establishing a clinical service to prevent psychosis: What, how and when? Systematic review. *Transl Psychiatry* 11: 43.

7. Estradé A, Salazar de Pablo G, Zanotti A, Wood S, Fisher HL, Fusar-Poli P (2022): Public health primary prevention implemented by clinical high-risk services for psychosis. *Transl Psychiatry* 12: 43.

8. Estradé A, Spencer TJ, De Micheli A, Murguia-Asensio S, Provenzani U, McGuire P, Fusar-Poli P (2022): Mapping the implementation and challenges of clinical services for psychosis prevention in England. *Front Psychiatry* 13: 945505.

9. Kotlicka-Antczak M, Podgórski M, Oliver D, Maric NP, Valmaggia L, Fusar-Poli P (2020): Worldwide implementation of clinical services for the prevention of psychosis: The IEPA early intervention in mental health survey. *Early Interv Psychiatry* 14: 741–750.

10. Fusar-Poli P, Spencer T, De Micheli A, Curzi V, Nandha S, McGuire P (2020): Outreach and support in South-London (OASIS) 2001-2020: Twenty years of early detection, prognosis and preventive care for young people at risk of psychosis. *Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol* 39: 111–122.

11. Oliver D, Davies C, Crossland G, Lim S, Gifford G, McGuire P, Fusar-Poli P (2018): Can We Reduce the Duration of Untreated Psychosis? A Systematic Review and Meta-Analysis of Controlled Interventional Studies. *Schizophr Bull* 44: 1362–1372.

12. Catalan A, Salazar de Pablo G, Vaquerizo Serrano J, Mosillo P, Baldwin H, Fernández-Rivas A, *et al.* (2021): Annual Research Review: Prevention of psychosis in adolescents - systematic review and meta-analysis of advances in detection, prognosis and intervention. *J Child Psychol Psychiatry* 62: 657–673.

13. Fusar-Poli P, Salazar de Pablo G, Correll CU, Meyer-Lindenberg A, Millan MJ, Borgwardt S, *et al.* (2020): Prevention of Psychosis: Advances in Detection, Prognosis, and Intervention. *JAMA Psychiatry* 77: 755–765.

14. Fusar-Poli P, Correll CU, Arango C, Berk M, Patel V, Ioannidis JPA (2021): Preventive psychiatry: a blueprint for improving the mental health of young people. *World Psychiatry Off J World Psychiatr Assoc WPA* 20: 200–221.

15. Fusar-Poli P (2017): Extending the Benefits of Indicated Prevention to Improve Outcomes of First-Episode Psychosis. *JAMA Psychiatry* 74: 667–668.

16. McGorry PD, Hartmann JA, Spooner R, Nelson B (2018): Beyond the "at risk mental state" concept: transitioning to transdiagnostic psychiatry. *World Psychiatry Off J World Psychiatr Assoc WPA* 17: 133–142.

17. Fusar-Poli P, Rutigliano G, Stahl D, Davies C, Bonoldi I, Reilly T, McGuire P (2017): Development and Validation of a Clinically Based Risk Calculator for the Transdiagnostic Prediction of Psychosis. *JAMA Psychiatry* 74: 493–500.

18. Fusar-Poli P, Werbeloff N, Rutigliano G, Oliver D, Davies C, Stahl D, *et al.* (2019): Transdiagnostic Risk Calculator for the Automatic Detection of Individuals at Risk and the Prediction of Psychosis: Second Replication in an Independent National Health Service Trust. *Schizophr Bull* 45: 562–570.

19. Oliver D, Wong CMJ, Bøg M, Jönsson L, Kinon BJ, Wehnert A, *et al.* (2020): Transdiagnostic individualized clinically-based risk calculator for the automatic detection of individuals at-risk and the prediction of psychosis: external replication in 2,430,333 US patients [no. 1]. *Transl Psychiatry* 10: 1–10.

20. Puntis S, Oliver D, Fusar-Poli P (2021): Third external replication of an individualised transdiagnostic prediction model for the automatic detection of individuals at risk of psychosis using electronic health records. *Schizophr Res* 228: 403–409.

21. Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, Irving J, Catalan A, Oliver D, *et al.* (2021): Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophr Bull* 47: 284–297.

22. Oliver D, Spada G, Colling C, Broadbent M, Baldwin H, Patel R, *et al.* (2021): Real-world implementation of precision psychiatry: Transdiagnostic risk calculator for the automatic detection of individuals at-risk of psychosis. *Schizophr Res* 227: 52–60.

23. Irving J, Patel R, Oliver D, Colling C, Pritchard M, Broadbent M, *et al.* (2021): Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Schizophr Bull* 47: 405–414.

24. Worthington MA, Cannon TD (2021): Prediction and Prevention in the Clinical High-Risk for Psychosis Paradigm: A Review of the Current Status and Recommendations for Future Directions of Inquiry. *Front Psychiatry* 12: 770774.

25. Sanfelici R, Dwyer DB, Antonucci LA, Koutsouleris N (2020): Individualized Diagnostic and Prognostic Models for Patients With Psychosis Risk Syndromes: A Meta-analytic View on the State of the Art. *Biol Psychiatry* 88: 349–360.

26. Paquin V, Cupo L, Malla AK, Iyer SN, Joober R, Shah JL (2023): Dynamic association of the first identifiable symptom with rapidity of progression to first-episode psychosis. *Psychol Med* 53: 2008–2016.

27. Paquin V, Malla AK, Iyer SN, Lepage M, Joober R, Shah JL (2023): Transsyndromic trajectories from pre-onset self-harm and subthreshold psychosis to the first episode of psychosis: A longitudinal study. *J Psychopathol Clin Sci* 132: 198–208.

28. Nelson B, McGorry PD, Wichers M, Wigman JTW, Hartmann JA (2017): Moving From Static to Dynamic Models of the Onset of Mental Disorder: A Review. *JAMA Psychiatry* 74: 528–534.

29. Putter H van H Hein (2012): *Dynamic Prediction in Clinical Survival Analysis*. Boca Raton: CRC Press. https://doi.org/10.1201/b11311

30. Ibrahim JG, Chu H, Chen LM (2010): Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *J Clin Oncol* 28: 2796–2801.

31. Yuen HP, Mackinnon A, Hartmann J, Amminger GP, Markulev C, Lavoie S, *et al.* (2018): Dynamic prediction of transition to psychosis using joint modelling. *Schizophr Res* 202: 333–340.

32. Studerus E, Beck K, Fusar-Poli P, Riecher-Rössler A (2020): Development and Validation of a Dynamic Risk Prediction Model to Forecast Psychosis Onset in Patients at Clinical High Risk. *Schizophr Bull* 46: 252–260.

33. Zhang T, Tang X, Zhang Y, Xu L, Wei Y, Hu Y, *et al.* (2023): Multivariate joint models for the dynamic prediction of psychosis in individuals with clinical high risk. *Asian J Psychiatry* 81: 103468.

34. Tanner KT, Sharples LD, Daniel RM, Keogh RH (2021): Dynamic Survival Prediction Combining Landmarking with a Machine Learning Ensemble: Methodology and Empirical Comparison. *J R Stat Soc Ser A Stat Soc* 184: 3–30.

35. Raket LL, Jaskolowski J, Kinon BJ, Brasen JC, Jönsson L, Wehnert A, Fusar-Poli P (2020): Dynamic ElecTronic hEalth reCord deTection (DETECT) of individuals at risk of a first

episode of psychosis: a case-control development and validation study. *Lancet Digit Health* 2: e229–e239.

36. Rudin C (2019): Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead [no. 5]. *Nat Mach Intell* 1: 206–215.

37. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR (2022): Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Comput Methods Programs Biomed* 226: 107161.

38. Collins GS, Reitsma JB, Altman DG, Moons KG (2015): Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 13: 1.

39. Lähteenmäki S, Aalto-Setälä T, Suokas JT, Saarni SE, Perälä J, Saarni SI, *et al.* (2009): Validation of the Finnish version of the SCOFF questionnaire among young adults aged 20 to 35 years. *BMC Psychiatry* 9: 5.

40. Roberts E, Wessely S, Chalder T, Chang C-K, Hotopf M (2016): Mortality of people with chronic fatigue syndrome: a retrospective cohort study in England and Wales from the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Clinical Record Interactive Search (CRIS) Register. *The Lancet* 387: 1638–1643.

41. Oram S, Khondoker M, Abas M, Broadbent M, Howard LM (2015): Characteristics of trafficked adults and children with severe mental illness: a historical cohort study. *Lancet Psychiatry* 2: 1084–1091.

42. Fusar-Poli P, Rutigliano G, Stahl D, Schmidt A, Ramella-Cravaro V, Hitesh S, McGuire P (2016): Deconstructing Pretest Risk Enrichment to Optimize Prediction of Psychosis in Individuals at Clinical High Risk. *JAMA Psychiatry* 73: 1260–1267.

43. Fusar-Poli P, Lai S, Di Forti M, Iacoponi E, Thornicroft G, McGuire P, Jauhar S (2020): Early Intervention Services for First Episode of Psychosis in South London and the Maudsley (SLaM): 20 Years of Care and Research for Young People. *Front Psychiatry* 11: 577110.

44. Fusar-Poli P, Estradé A, Spencer TJ, Gupta S, Murguia-Asensio S, Eranti S, *et al.* (2019): Pan-London Network for Psychosis-Prevention (PNP). *Front Psychiatry* 10: 707.

45. Jongsma HE, Turner C, Kirkbride JB, Jones PB (2019): International incidence of psychotic disorders, 2002-17: a systematic review and meta-analysis. *Lancet Public Health* 4: e229–e244.

46. Reinikainen J, Laatikainen T, Karvanen J, Tolonen H (2015): Lifetime cumulative risk factors predict cardiovascular disease mortality in a 50-year follow-up study in Finland. *Int J Epidemiol* 44: 108–116.

47. Paige E, Barrett J, Stevens D, Keogh RH, Sweeting MJ, Nazareth I, *et al.* (2018): Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. *Am J Epidemiol* 187: 1530–1538.

48. Buuren S van, Groothuis-Oudshoorn K (2011): mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw* 45: 1–67.

49. Sisk R, Sperrin M, Peek N, van Smeden M, Martin GP (2023): Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Stat Methods Med Res* 32: 1461–1477.

50. Cox DR (1972): Regression Models and Life-Tables. *J R Stat Soc Ser B Methodol* 34: 187–202.

51. Van Houwelingen HC (2007): Dynamic Prediction by Landmarking in Event History Analysis. *Scand J Stat* 34: 70–85.

52. Steyerberg EW, Harrell FE (2016): Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 69: 245–247.

53. Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW (2018): The Science of Prognosis in Psychiatry: A Review. *JAMA Psychiatry* 75: 1289–1297.

54. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, *et al.* (2010): Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiol Camb Mass* 21: 128–138.

55. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011): On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 30: 1105–1117.

56. Calster BV, Nieboer D, Vergouwe Y, Cock BD, Pencina MJ, Steyerberg EW (2016): A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 74: 167–176.

57. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS (2016): External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 353: i3140.

58. IntHout J, Ioannidis JPA, Borm GF (2014): The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 14: 25.

59. Vickers AJ, Van Calster B, Steyerberg EW (2016): Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *The BMJ* 352: i6.

60. Baldwin H, Loebel-Davidsohn L, Oliver D, Salazar de Pablo G, Stahl D, Riper H, Fusar-Poli P (2022): Real-World Implementation of Precision Psychiatry: A Systematic Review of Barriers and Facilitators [no. 7]. *Brain Sci* 12: 934.

61. Goodfellow I, Bengio Y, Courville A (2016): *Deep Learning*. MIT Press.

62. Davis KAS, Sudlow CLM, Hotopf M (2016): Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. *BMC Psychiatry* 16: 263.

63. Webb JR, Addington J, Perkins DO, Bearden CE, Cadenhead KS, Cannon TD, *et al.* (2015): Specificity of Incident Diagnostic Outcomes in Patients at Clinical High Risk for Psychosis. *Schizophr Bull* 41: 1066–1075.

64. Devaux A, Genuer R, Peres K, Proust-Lima C (2022): Individual dynamic prediction of clinical endpoint from large dimensional longitudinal biomarker history: a landmark approach. *BMC Med Res Methodol* 22: 188.

65. Pande A, Li L, Rajeswaran J, Ehrlinger J, Kogalur UB, Blackstone EH, Ishwaran H (2017): Boosted Multivariate Trees for Longitudinal Data. *Mach Learn* 106: 277–305.

66. Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, *et al.* (2023): Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. *JAMA Netw Open* 6: e2345050.

67. Khor S, Haupt EC, Hahn EE, Lyons LJL, Shankaran V, Bansal A (2023): Racial and Ethnic Bias in Risk Prediction Models for Colorectal Cancer Recurrence When Race and Ethnicity Are Omitted as Predictors. *JAMA Netw Open* 6: e2318495.

2.8 Supplementary content

eTable 2.8.1 TRIPOD Checklist: Prediction Model Development and validation

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

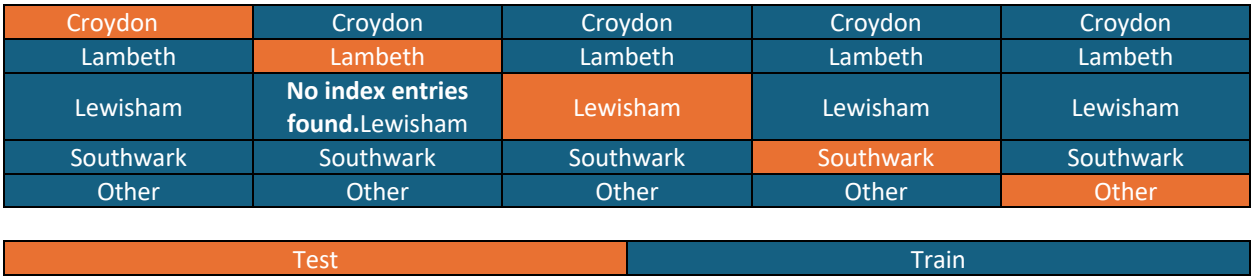
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	Method/ Model development and 'Leave one borough out' Internal-external cross validation
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Results/ Participants Table 1
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	Results/ Participants Table 1 Figure 2
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Results/ Participants Table 1
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	Results/ Participants Table 1 Figure 2
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Results/ Model specification Table 2 Supplementary
	15b	D	Explain how to the use the prediction model.	Results/ Model specification Supplementary
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	Results/Model Performance
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	Discussion
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	Discussion
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	Discussion
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	Discussion
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	NA
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	Acknowledgement

eMethods 2.8.1 Outcome

Outcome The primary outcome of interest was the hazard ratio (HR) of developing any psychotic disorder. This was defined as the emergence of the first ICD-10 primary diagnosis of non-organic psychotic disorder, occurring at least three months after the index diagnosis as recorded in the electronic medical records: schizophrenia spectrum psychoses (schizophrenia [F20.x, except F20.4/F20.5], schizoaffective disorder [F25.x], delusional disorders [F22.x, F24], acute and transient psychotic disorders [F23.x]), unspecified nonorganic psychosis (F28/F29), psychotic disorders due to psychoactive substance use ([F10-F19].5), and affective psychoses (mania with psychotic symptoms [F30.2], bipolar affective disorder with psychotic symptoms [F31.2, F31.5], and depression with psychotic symptoms [F32.3/F33.3]).

eFigure 2.8.1 Internal-external validation

The specific train-test splits based on the geographical (boroughs) divisions are visualised below.



eFigure 2.8.1 Internal-external geographical cross-validation. The blue shading represents borough in the test set and the orange represent boroughs allocated to the test set. It can be noted than contrary to the previous studies in the internal-external cross validation the train-test spit is performed five times as opposed to just once.

eTable 2.8.2 Landmark stratified baseline hazards

Landmark point [months]	Baseline hazard value at two years past landmark point
0	0.32
6	0.237
12	0.24
18	0.238
24	0.24
30	0.259
36	0.259
42	0.272
48	0.281

eTable 2.8.2 Baseline hazards stratified at the landmark points. The presented hazards values were taken at 2 years after each landmark point.

eTable 2.8.3 Cox landmark model coefficients

Predictor	Cox landmar coefficient
Sociodemographic predictors	
Female Gender	-0,09237
Age	0,0022
Self-assigned ethnicity	
White	0
Black	0,84695
Asian	0,61081
Mixed	0,30445
Other	0,15939
Clinical predictors	
Diagnosis	
Acute and transient psychotic disorders	0
At Risk Mental State (ARMS)	-1,14943
Anxiety disorders	-2,70443
Bipolar mood disorders	-1,37605
Childhood/adolescence onset disorders	-3,76846
Developmental disorders	-3,75154
Mental retardation	-3,03301
Non bipolar mood disorders	-2,37496
Personality disorders	-2,31221
Physiological syndromes	-3,37269
Substance use disorders	-2,54772
Natural Language Processing (NLP) predictors	
Appetite loss	-0,07369
Agitation	0,05562
Blunted affect	0,01948
Cannabis uses	0,01395
Cocaine use	-0,07806
Delusional thinking	0,09414
Disturbed sleep	-0,00004
Guilt	-0,10786
Hopelessness	-0,05376
Hallucinations (all)	0,0647
Hallucinations (auditory)	0,15272
Hallucinations (olfactory, gustatory & tactile)	0,22478
Hallucinations (visual)	0,03554
Insomnia	-0,0567
Irritability	0,04738
Negative symptoms	0,01933
Paranoia	0,11908
Poverty of speech	0,33658

Weight loss	-0,03276
Tearfulness	-0,06486
Cumulative agitation	0,02364
Cumulative appetite loss	0,16367
Cumulative blunted affect	-0,05587
Cumulative cannabis use	0,02665
Cumulative cocaine use	-0,01958
Cumulative delusional thinking	0,08023
Cumulative disturbed sleep	0,00207
Cumulative guilt	0,07619
Cumulative hopelessness	0,03189
Cumulative hallucinations (any)	0,11484
Cumulative hallucinations (auditory)	-0,04509
Cumulative hallucinations (olfactory, gustatory & tactile)	-0,11208
Cumulative hallucinations (visual)	0,03639
Cumulative insomnia	0,1324
Cumulative irritability	-0,00928
Cumulative negative symptoms	0,00228
Cumulative paranoia	0,22473
Cumulative poverty of speech	-0,4348
Cumulative weight loss	0,15396
Cumulative tearfulness	0,03525

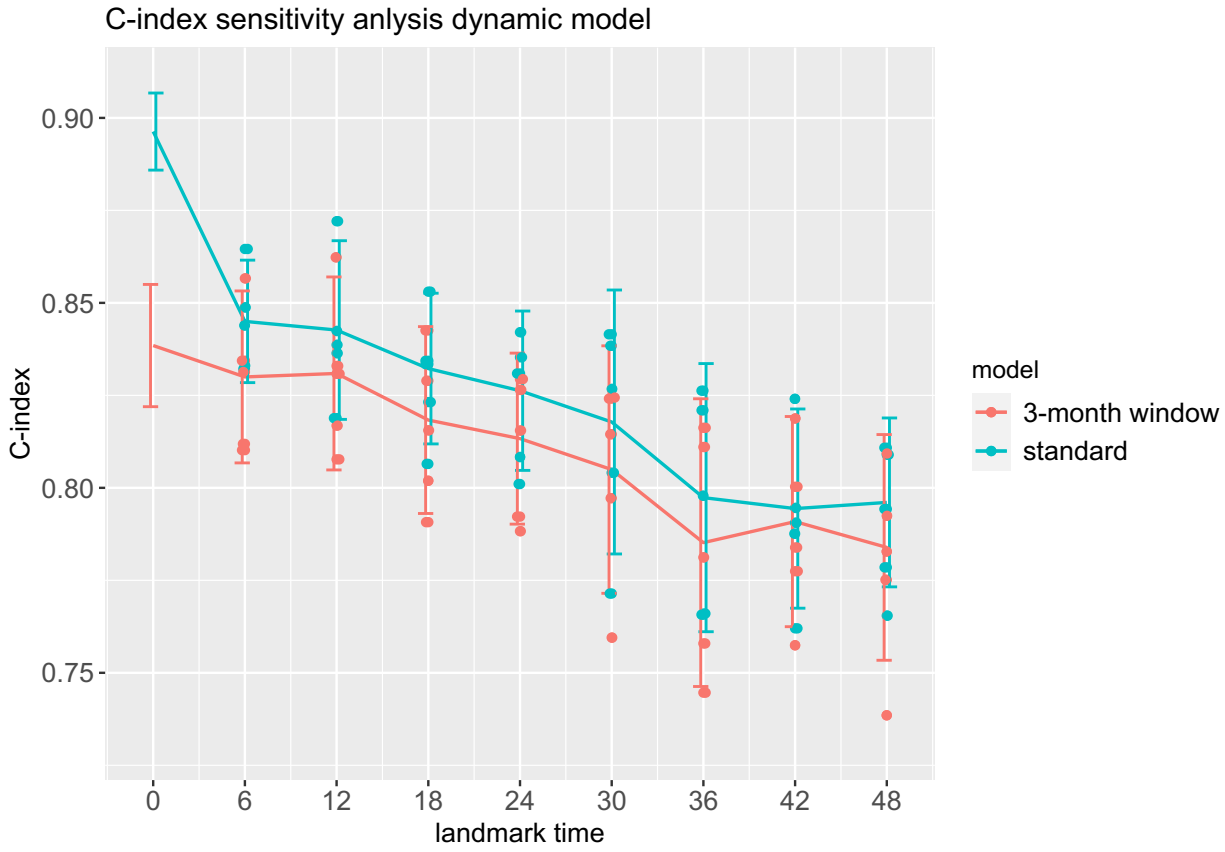
eMethods 2.8.2 Sensitivity analysis - ***excluding patients that developed psychosis within 3 months following a landmark point across all test sets to further distinguish predictors and outcomes.***

Method: For the static model in the test set we excluded patients who developed psychosis within 3 months following baseline. For the dynamic model in the test set we excluded patients who developed psychosis within 3 months following any landmark point. The training set was kept unchanged for both models to ensure optimal model performance. The mean difference in the C-index between the standard test set and the 3-month window exclusion test set was quantified by the meta regression.

Results: A meta-regression revealed that for the dynamic model the standard test set had a mean C-index difference of 0.022 (95% CI 0.014 - 0.03) higher compared to the 3-month window exclusion test set, as summarized in eFigure 2.8.2. A difference of 0.026 (95% CI 0.019 - 0.032) was observed between dynamic and static models in their respective 3-month exclusion window sets, detailed in eFigure 2.8.3. The calibration results show that the dynamic model is well calibrated at the entire follow-up period when evaluated using the modified test set.

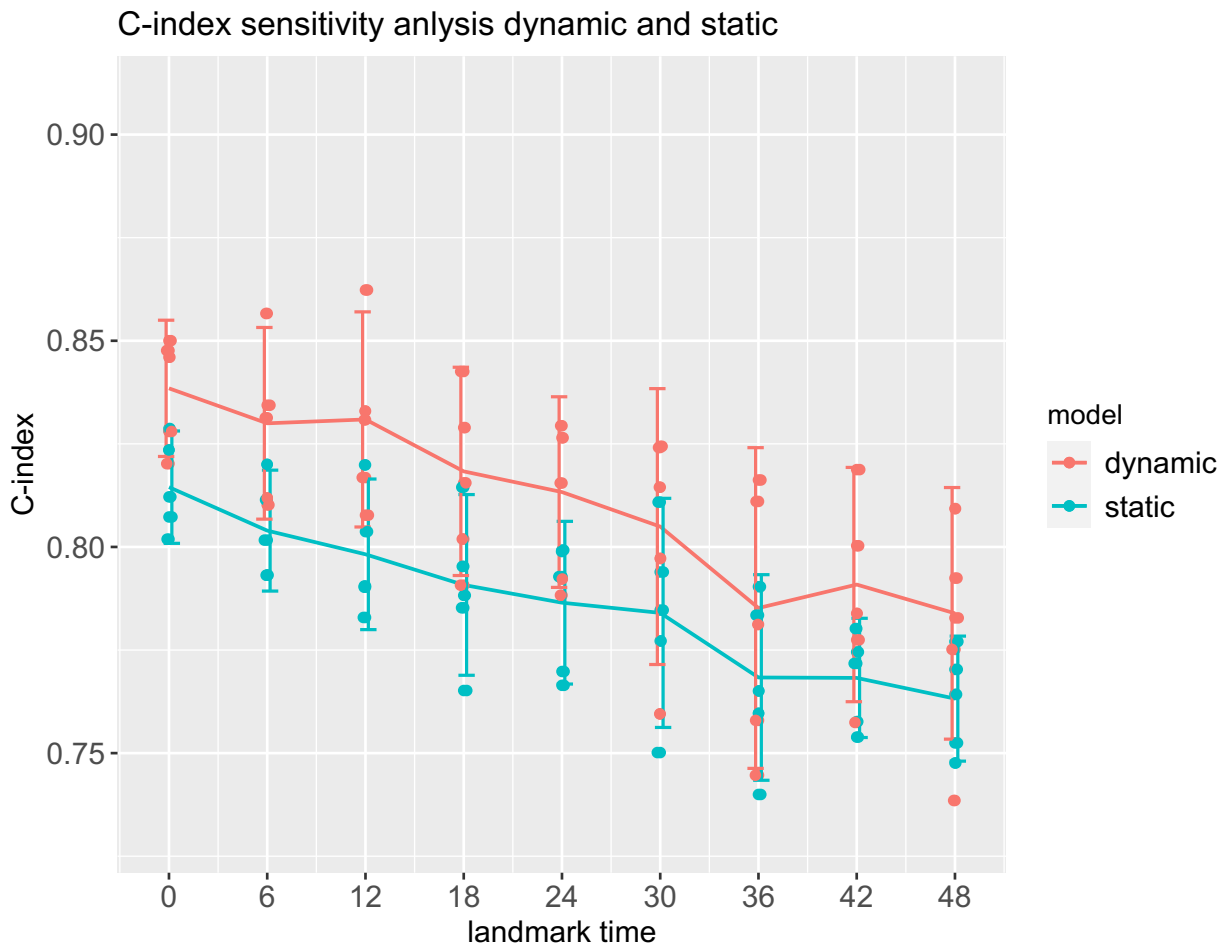
Interpretation: The dynamic model shows a small but significant performance drop without the 3-month window exclusion in the test set. However, it still outperforms the static model on modified test sets, with an average C-index 0.026 (95% CI 0.019 - 0.032) higher. The dynamic model is also significant better calibrated eFigure 2.8.4. The results indicate that the static model's time-varying NLP predictors hold significant predictive value beyond mere confirmation, though their predictive ability is slightly reduced when short transitions are excluded. The results show that the time-varying NLP predictors of the static model do hold significant predictive value and are not merely confirmatory.

eFigure 2.8.2 Sensitivity analysis for dynamic model. C-index on the standard test set and the test set with short term psychosis transitions removed – 3-month window.



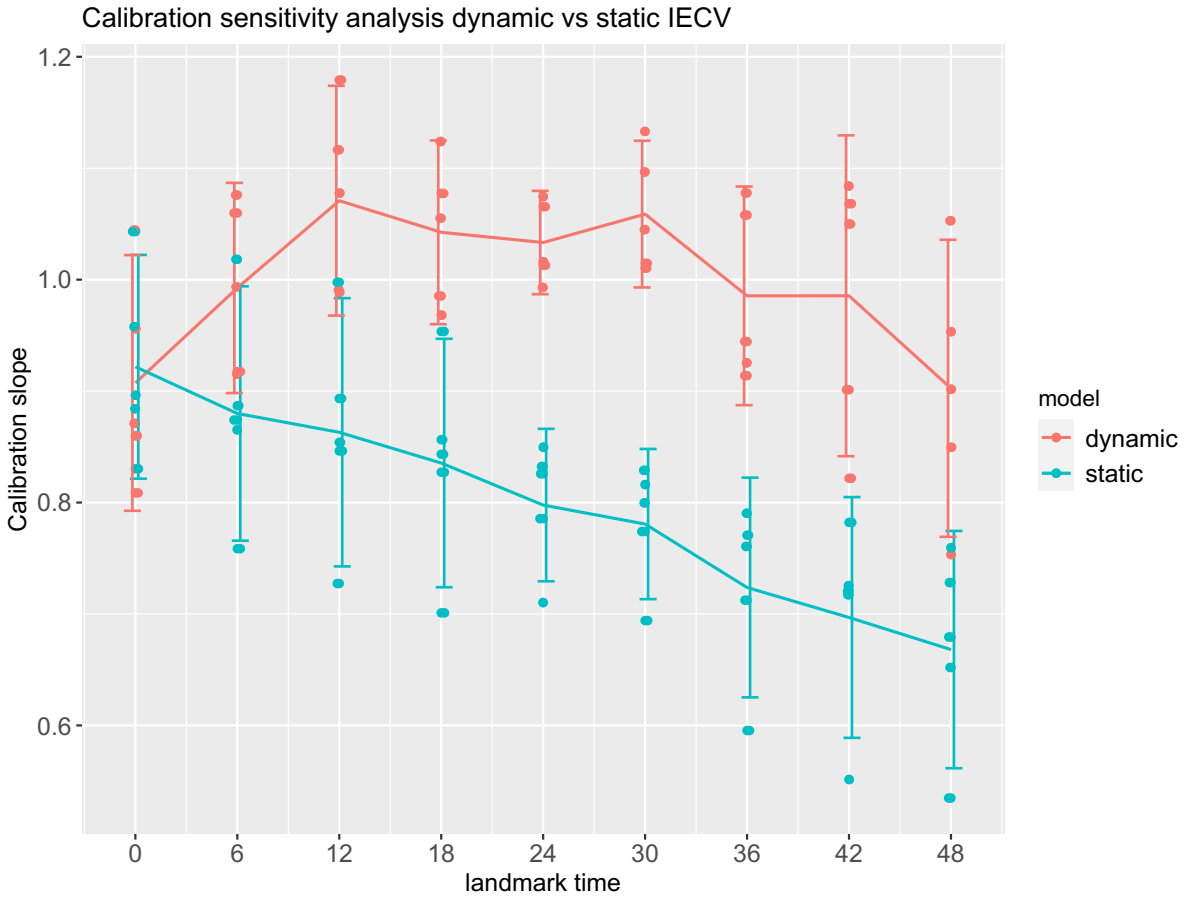
eFigure 2.8.2. C-index results for the dynamic model in internal -external cross validation. The 'standard' represents the unmodified test set. The '3-month window' represents the C-index of the dynamic model evaluated on the test set with the transitions to psychosis under 3-month excluded. On average as indicated by the meta regression coefficient the standard model has 0.022 (95% CI 0.014 - 0.03) higher C-index. For landmark times 6-48 month the mean difference is 0.013 (95% CI 0.005 -0.02).

eFigure 2.8.3 Sensitivity analysis results - excluding patients that developed psychosis within 3 months following a landmark point across all test sets in both static and dynamic model.



eFigure 2.8.3 C -index results for the static and dynamic model in internal-external cross validation evaluated on the test sets that exclude patients that developed psychosis within 3 months following a landmark point. On average as indicated by the meta regression coefficient the dynamic model has 0.026 (95% CI 0.019 - 0.032) higher C-index.

eFigure 2.8.4 Sensitivity analysis calibration results - excluding patients that developed psychosis within 3 months following a landmark point across all test sets in both static and dynamic model.



eFigure 2.8.4. Calibration slope for the static and dynamic model in internal-external cross validation evaluated on the test sets that exclude patients that developed psychosis within 3 months following any landmark point.

eMethods 2.8.3 External validations of the prediction's models developed using SLaM EHRs.

The transdiagnostic risk calculator developed In SLaM NHS Trust was externally validated in three different locations. The model has performed well in external validation despite other NHS Trusts having different service configurations (1-2). The model also replicated in the US (3) with a combination of primary/secondary care so again, what provides evidence for the robustness to these differences. The table presents results of the external validations:

Site	Sample size	Number of events	C-index
Camden and Islington NHS Trust UK (1)	13,702	490	0.73
Oxford Health NHS Foundation Trust UK (2)	33,710	868	0.79
IBM® MarketScan® Commercial Database US (3)	2,430,333	24,941	0.68

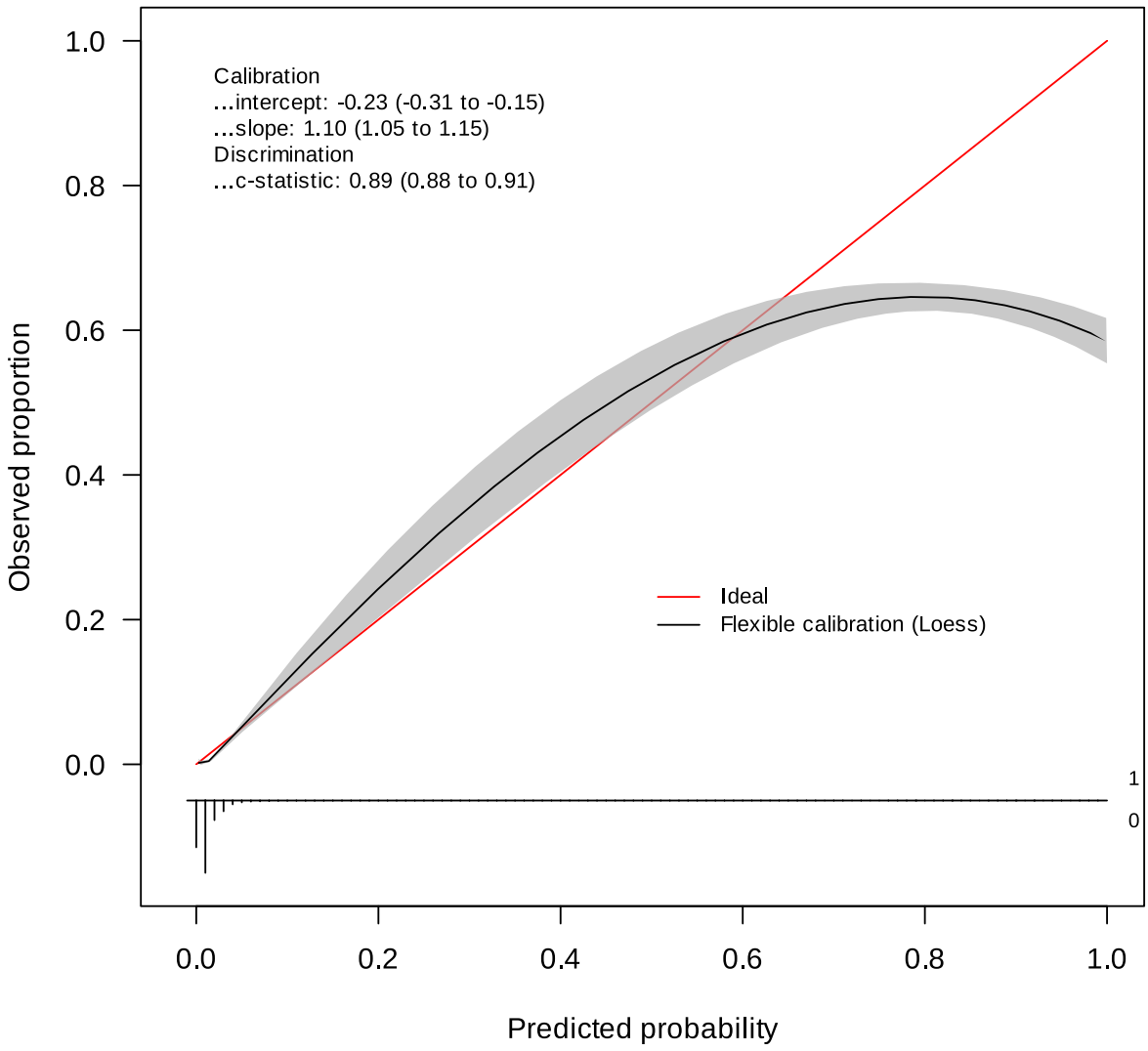
Ref:

(1) Fusar-Poli, Paolo, et al. "Transdiagnostic risk calculator for the automatic detection of individuals at risk and the prediction of psychosis: second replication in an independent national health service trust." Schizophrenia Bulletin 45.3 (2019): 562-570.

(2) Puntis, Stephen, Dominic Oliver, and Paolo Fusar-Poli. "Third external replication of an individualised transdiagnostic prediction model for the automatic detection of individuals at risk of psychosis using electronic health records." Schizophrenia Research 228 (2021): 403-409.

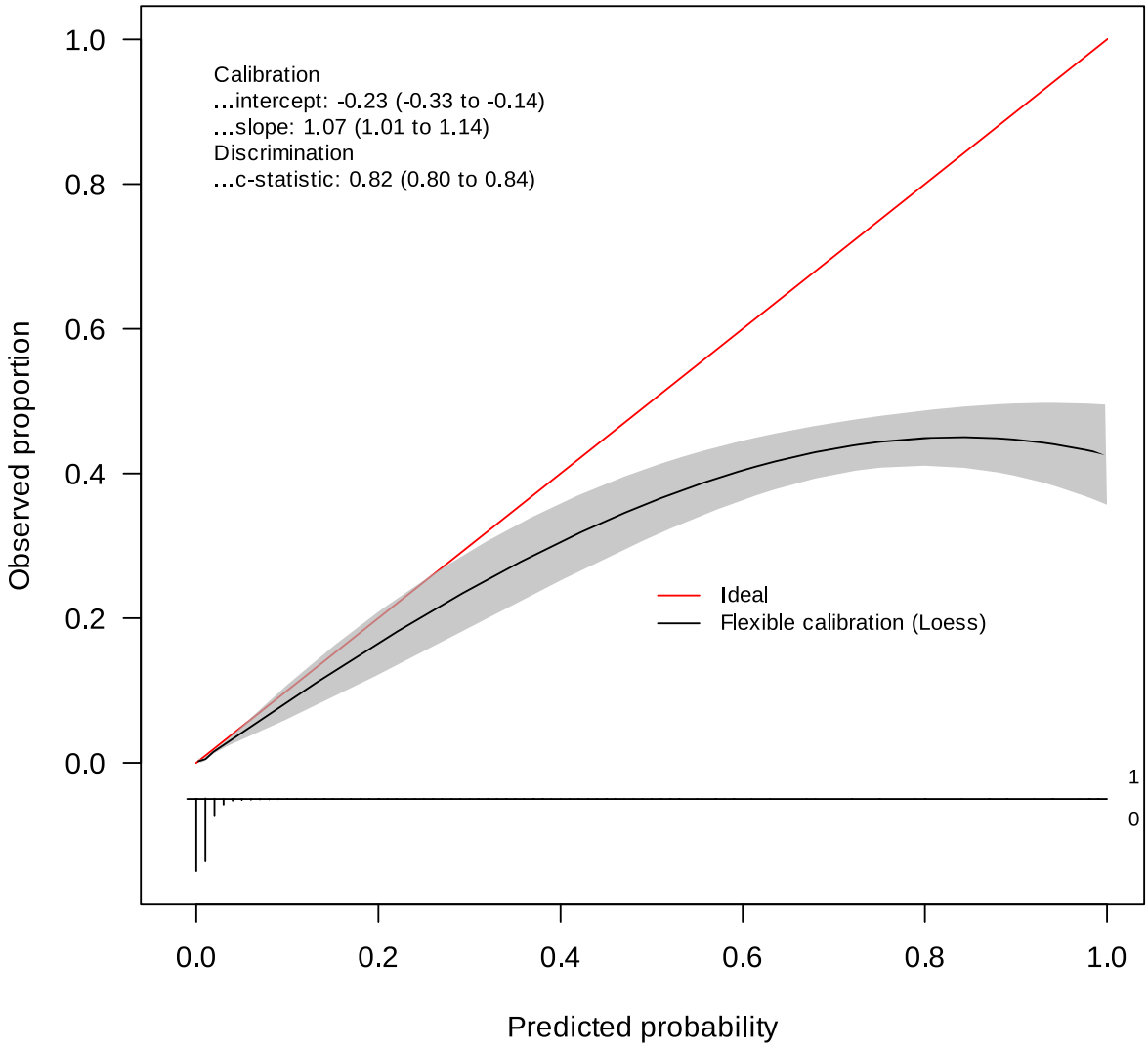
(3) Oliver, Dominic, et al. "Transdiagnostic individualized clinically-based risk calculator for the automatic detection of individuals at-risk and the prediction of psychosis: external replication in 2,430,333 US patients." Translational psychiatry 10.1 (2020): 364.

eFigure 2.8.5 Dynamic model calibration plot at baseline



eFigure 2.8.5 Black line shows non-parametric estimate of the calibration relationship between actual and predicted probability using lowess regression. The red line shows the ideal relationship (intercept of zero and slope of one). Lines at the bottom of the calibration plot provide a visual representation of the distribution of the predicted probabilities for our data points.

eFigure 2.8.6 Dynamic model calibration plot at 24-month landmark point



eFigure 2.8.6 Black line shows non-parametric estimate of the calibration relationship between actual and predicted probability using lowess regression. The red line shows the ideal relationship (intercept of zero and slope of one). Lines at the bottom of the calibration plot provide a visual representation of the distribution of the predicted probabilities for our data points.

eFigure 2.8.7 Dynamic model calibration plot at 48-month landmark point

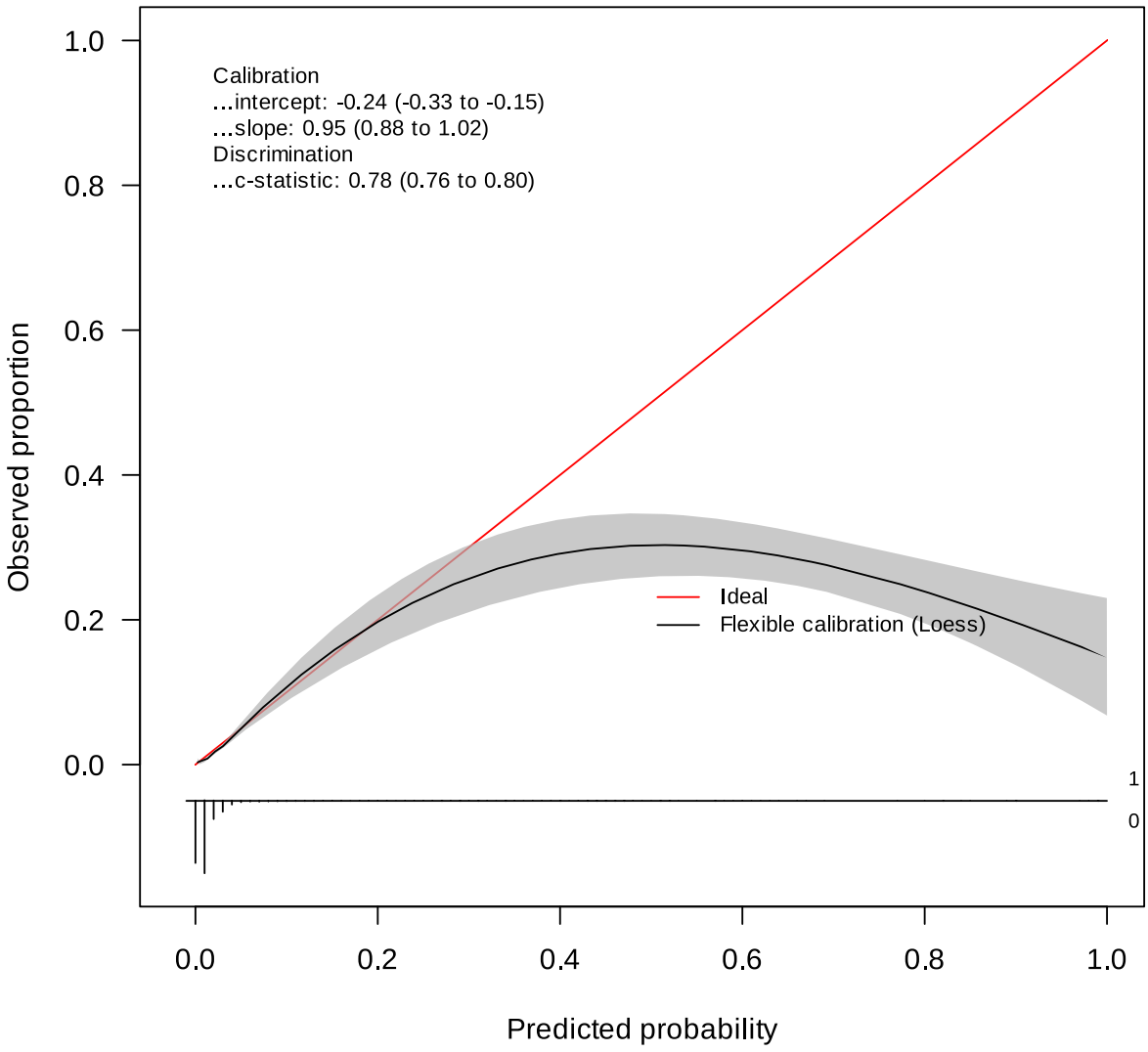
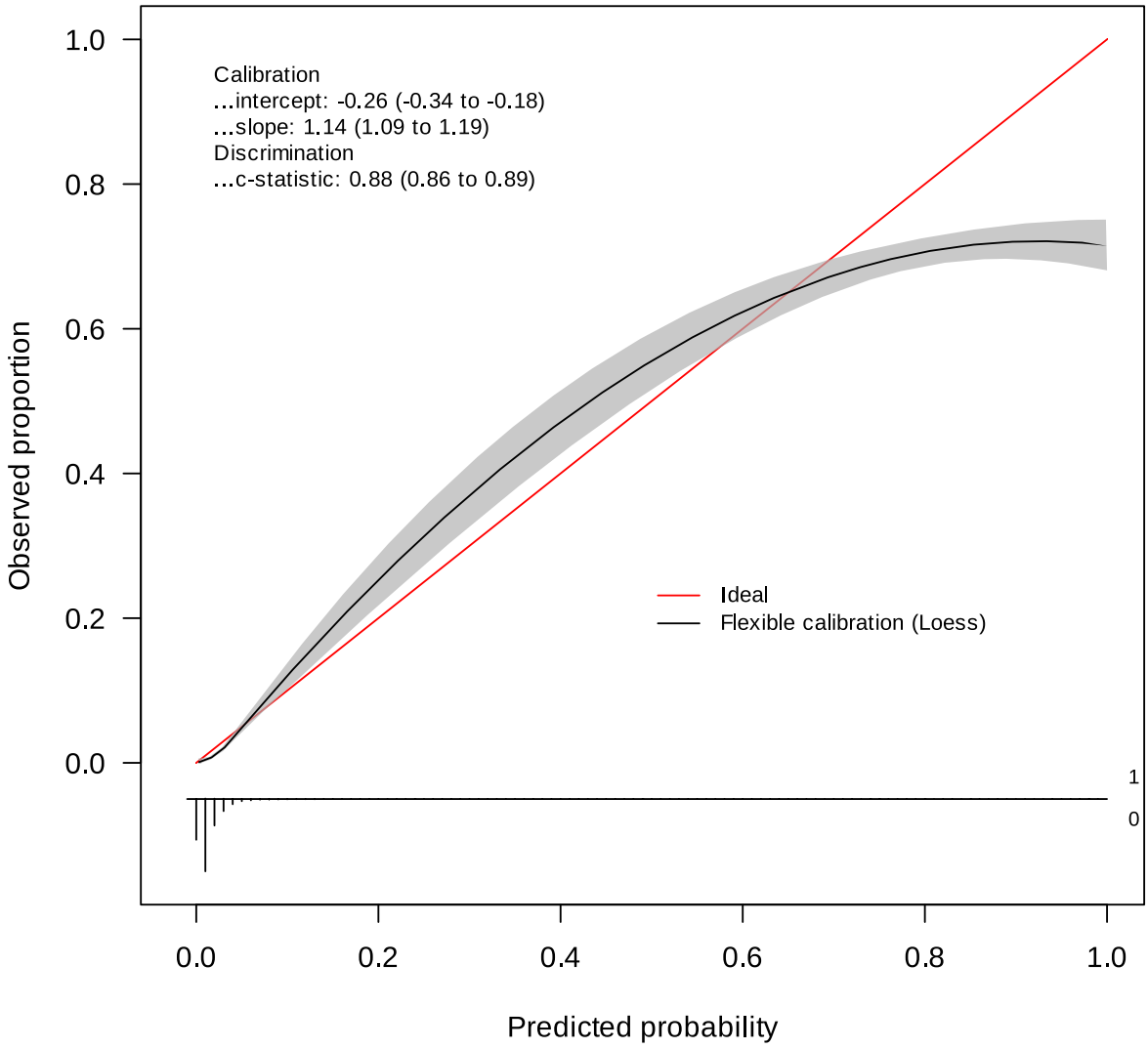


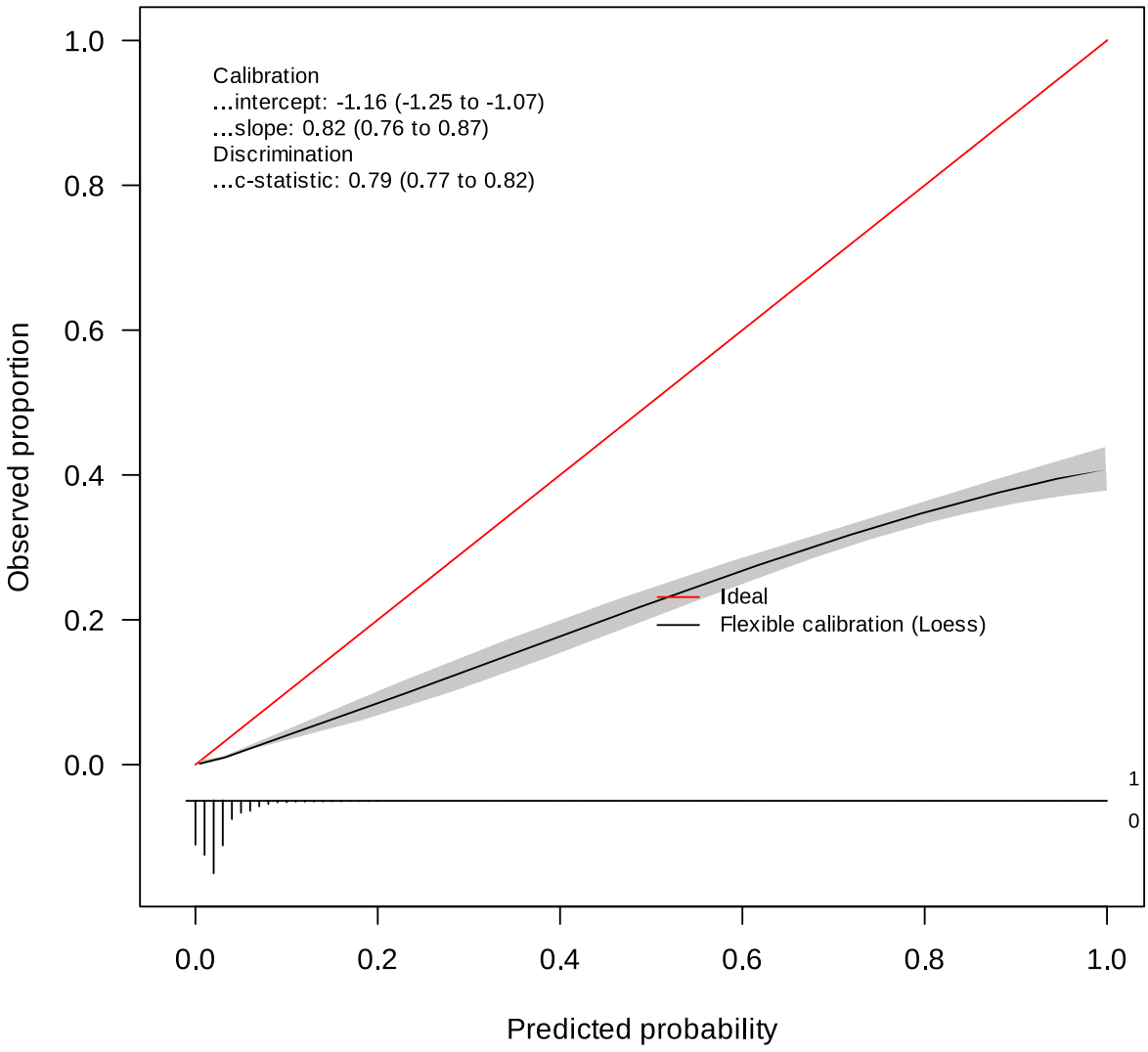
Figure 2.8.7 Black line shows non-parametric estimate of the calibration relationship between actual and predicted probability using lowess regression. The red line shows the ideal relationship (intercept of zero and slope of one). Lines at the bottom of the calibration plot provide a visual representation of the distribution of the predicted probabilities for our data points.

eFigure 2.8.8 Static model calibration plot at baseline



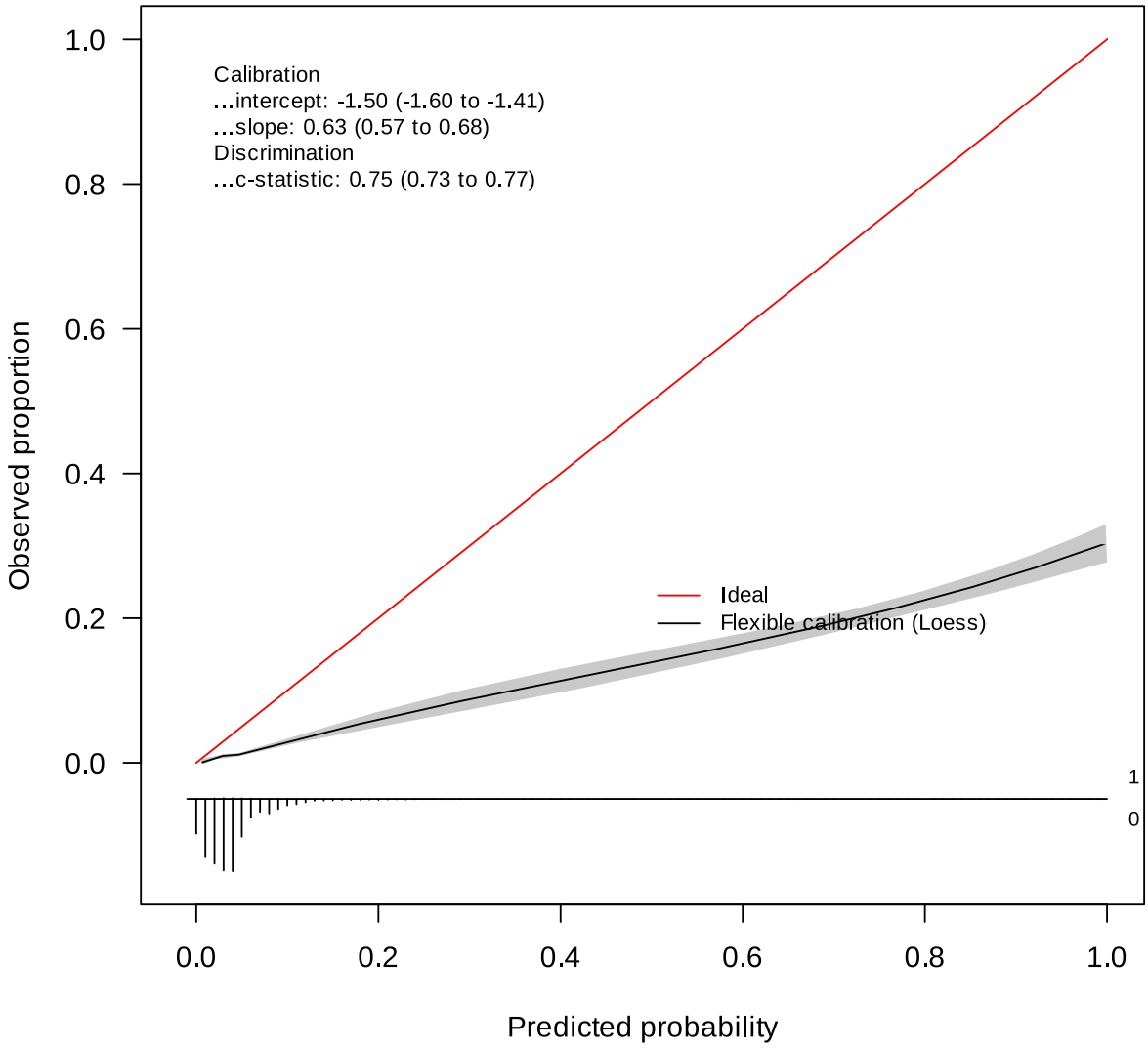
eFigure 2.8.8 Black line shows non-parametric estimate of the calibration relationship between actual and predicted probability using lowess regression. The red line shows the ideal relationship (intercept of zero and slope of one). Lines at the bottom of the calibration plot provide a visual representation of the distribution of the predicted probabilities for our data points.

eFigure 2.8.9 Static model calibration plot at 24-month landmark point



eFigure 2.8.9 Black line shows non-parametric estimate of the calibration relationship between actual and predicted probability using lowess regression. The red line shows the ideal relationship (intercept of zero and slope of one). Lines at the bottom of the calibration plot provide a visual representation of the distribution of the predicted probabilities for our data points.

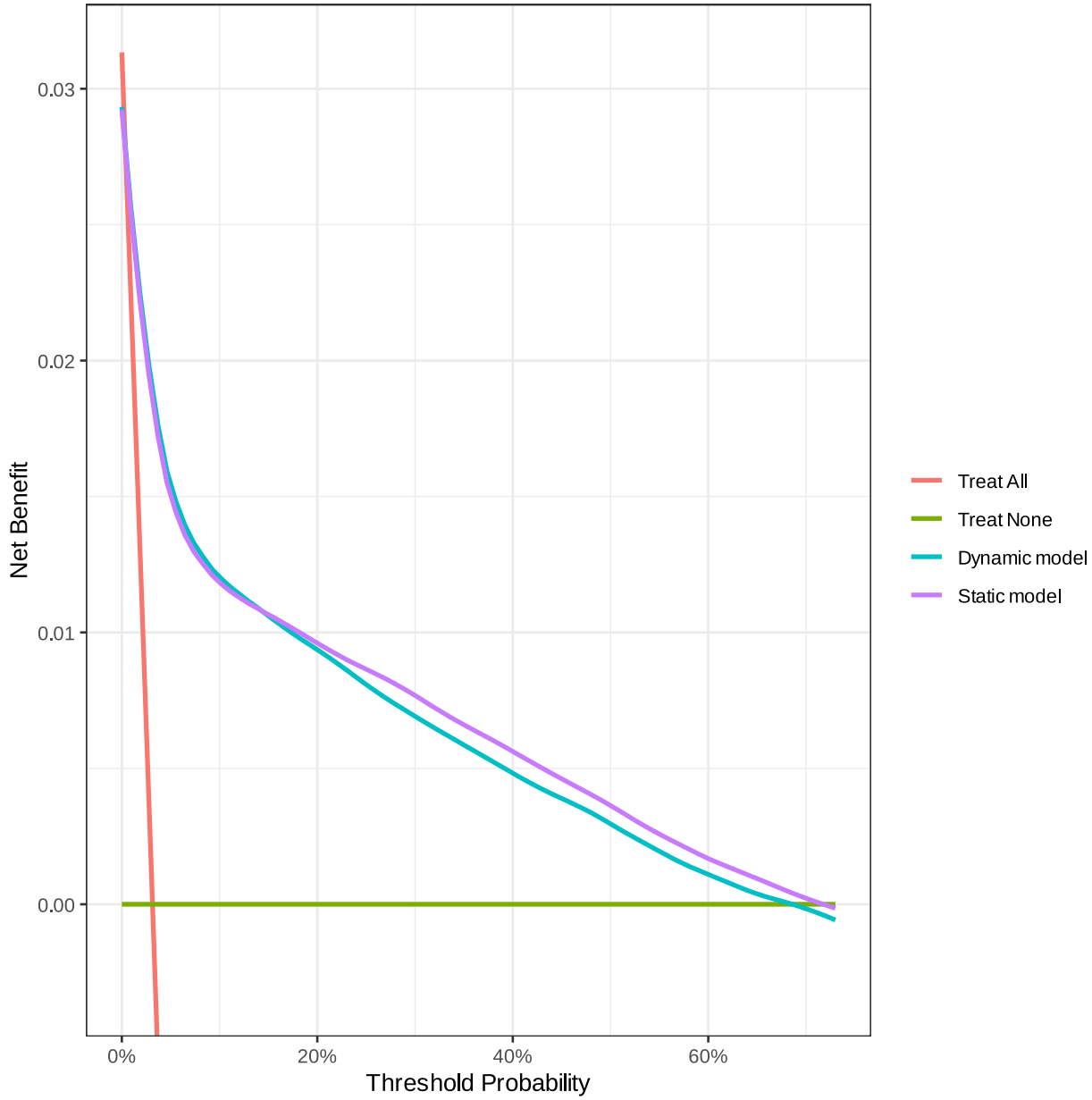
eFigure 2.8.10 Static model calibration plot at 48-month landmark point



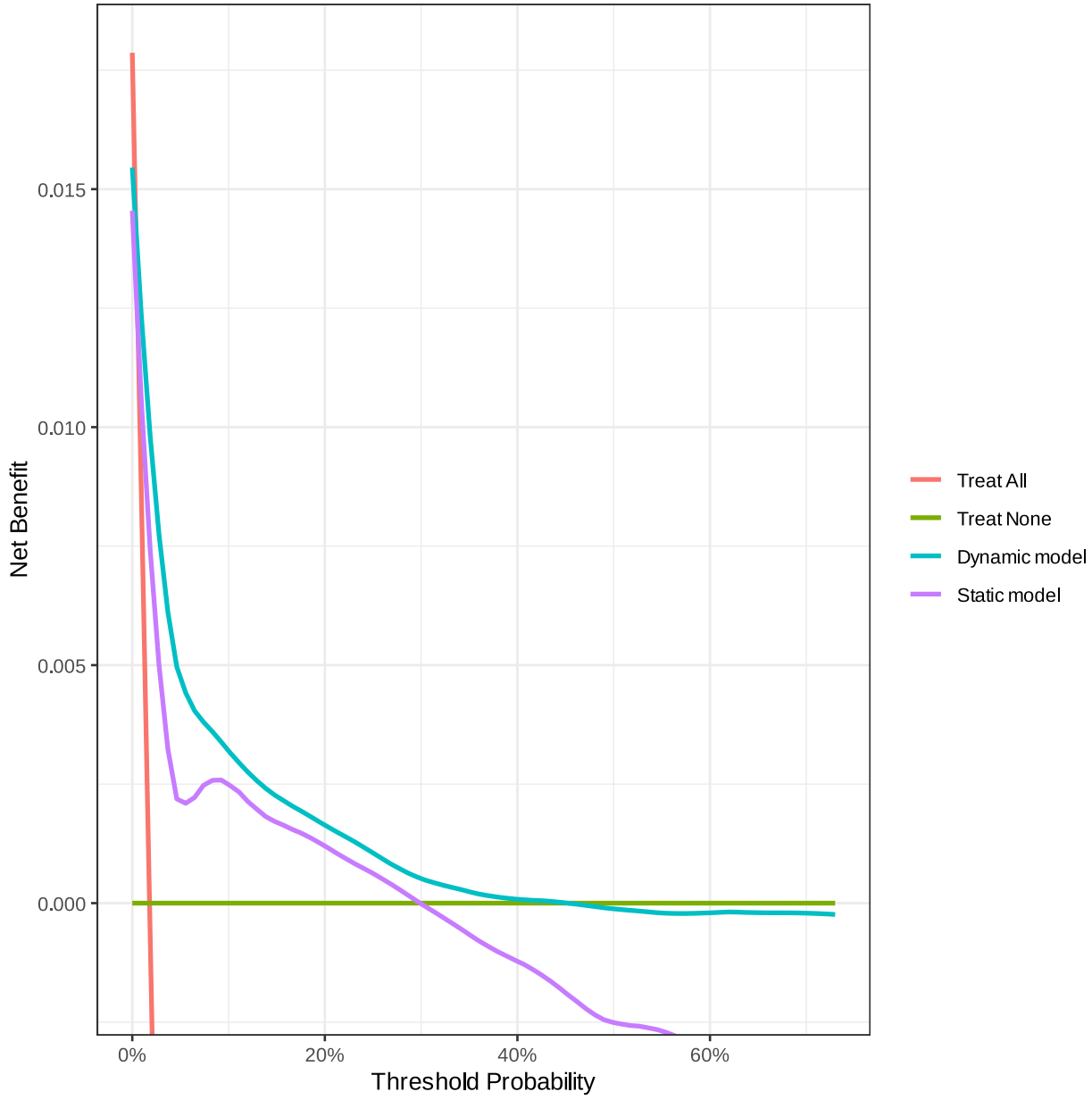
eFigure 2.8.10 Black line shows non-parametric estimate of the calibration relationship between actual and predicted probability using lowess regression. The red line shows the ideal relationship (intercept of zero and slope of one). Lines at the bottom of the calibration plot provide a visual representation of the distribution of the predicted probabilities for our data points.

eFigure 2.8.11 to 2.8.13: The decision curve shows the net benefit of using the dynamic and the static model across different threshold probabilities, as well as the two clinical alternatives: treating all or none. The y-axis shows the net benefit, which takes into account the true positives and weighs the false positives by the odds of the threshold probability. A higher net benefit implies that the model provides better decision-making support. The x-axis represents the threshold probability.

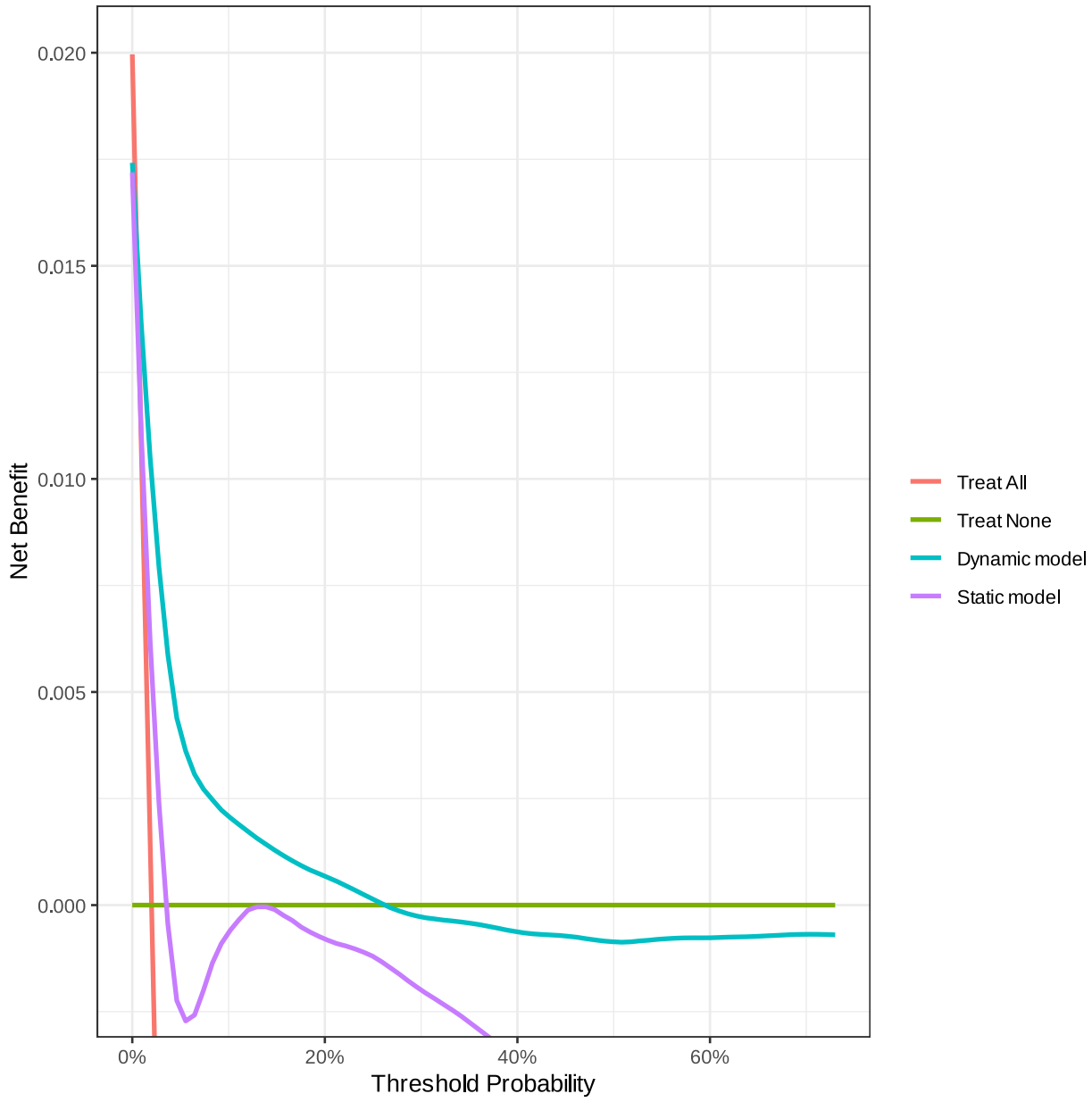
eFigure 2.8.11 Potential Clinical Utility Decision Curve at baseline



eFigure 2.8.12 Potential Clinical Utility Decision Curve at 24-month landmark point



eFigure 2.8.13 Potential Clinical Utility Decision Curve at 48-month landmark point



eTable 2.8.4 Missing data distribution:

Landmark point [month]	Total cases [n]	Total missing [n]	Missing ethnicity [n]	Missing gender[n]	Missing age[n]
0	158139	24441	22415	95	31
6	116898	16831	16741	66	6
12	98802	13898	13825	54	1
18	85807	11997	11938	41	1
24	75652	10436	10383	36	1
30	67265	9065	9022	27	1
36	60668	8002	7960	26	1
42	54490	7116	7083	19	1
48	48709	6366	6337	15	1

eTable 2.8.5 Definition of self-reported ethnicity according to UK Office of National Statistics

Ethnic group	Self-reported ethnicity as recorded in EHR
Black	Black or Black British - African Black or Black British - Caribbean Black or Black British - Any other Black background
White	White - British White - Irish White - Any other White background
Asian	Asian or Asian British - Bangladeshi Asian or Asian British - Indian Asian or Asian British - Pakistani Asian or Asian British - Any other Asian background Other Ethnic Groups - Chinese
Mixed	Mixed - White and Asian Mixed - White and Black African Mixed - White and Black Caribbean Mixed - Any other mixed background
Other	Other Ethnic Groups - Any other ethnic group
Missing	Not Known Not Recorded

eMethods 2.8.4 NLP algorithm development and validation

The CRIS symptom algorithms (e.g. 'guilt') have been developed using machine learning approaches against gold standard training sets manually annotated for positive, negative and unknown (irrelevant) mentions. As such, they are able to exclude language features such as negation (e.g. 'patient denies guilt', 'patient has no guilt') and irrelevant mentions (e.g. 'his mother felt guilty'). Patterns of failure driving false positives are identified through manual testing of algorithm output (e.g. 'ZZZZZ was found guilty of stealing'); the machine learning classifier is then trained on these false positives to ignore these and similar statements in an iterative process of testing and redeveloping until acceptable precision is achieved. Patterns of failure identified through testing can be found in the CRIS service's comprehensive online NLP algorithm library provided at https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/cris-natural-language-processing/.

The performance of each NLP algorithm was measured with precision (proportion of true positive instances of total NLP-labelled positive instances) and recall (proportion of true positive instances of all positive instances in the text). As EHRs provide multiple opportunities for term detection, we favour precision over recall, using only NLP algorithms with at least 80% precision (see eTable 4 for a final list of NLP algorithms employed).

These algorithms were manually validated by an independent researcher at the SLaM Biomedical Research Centre Nucleus prior to the current research project. The programme for algorithms validation was responsive to the specific needs of scheduled CRIS research activities and therefore the approach was not standardised. For example, depression symptom algorithms have been validated against records for SLaM individuals who had ever had a depression diagnosis; other algorithms have been validated against records for all individuals on the SLaM register.

eMethods 2.8.5 Meta-regression for quantifying the performance difference.

The meta-regression was performed to provide a single measure that quantifies the improvement in C-index between the static and dynamic model. Evaluation of both the static and dynamic models at each of the 9 landmark points and 5 train-test splits resulted in 90 C-index values and their corresponding standard errors. To quantify the difference in discrimination performance C-index between the two models, a multilevel meta-regression was conducted (1). We have included three dependant variables: borough, landmark time and model type to explain the C-index. There were 5 boroughs categories (Croydon, Lambeth, Lewisham, Other, and Southwark). The landmark time variable corresponded to one of the 9 nine landmark points (0,1,2, … ,7, 8, 9). The model variable was a binary either static or dynamic model. 'Borough' was included as a random effect to account for the dependency of the estimates of C-index between the two modelling approaches within the same validation datasets over time. Model type (dynamic versus static) and landmark point were included as categorical fixed-effect variables to assess differences in the C-index values. The coefficient of the model type variable was interpreted as the mean difference between the two models, summarizing the results across various landmark points and train-test splits. The full results of the meta regression model:

variable	estimate (95% CI)	p value
Intercept	0.8992 (0.889 - 0.9106)	<0.001
Model type (C-index diffrance dynamic vs static)	0.0351 (0.029 - 0.04)	<0.001
Landmark 6-month	-0.0586 (-0.0673 -0.0499)	<0.001
Landmark 12-month	-0.0626 (-0.0720 -0.0532)	<0.001
Landmark 18-month	-0.0711 (-0.0812 -0.0610)	<0.001
Landmark 24-month	-0.0766 (-0.0873 -0.0659)	<0.001
Landmark 30-month	-0.0809 (-0.0923 -0.0695)	<0.001
Landmark 36-month	-0.0989 (-0.1116 -0.0863)	<0.001
Landmark 42-month	-0.1014 (-0.1148 -0.0879)	<0.001
Landmark 48-month	-0.1033 (-0.1148 -0.0879)	<0.001

eTable 2.8.6 Comparison between NLP predictors used in the presented model and the previous static risk calculator by Irving et al.

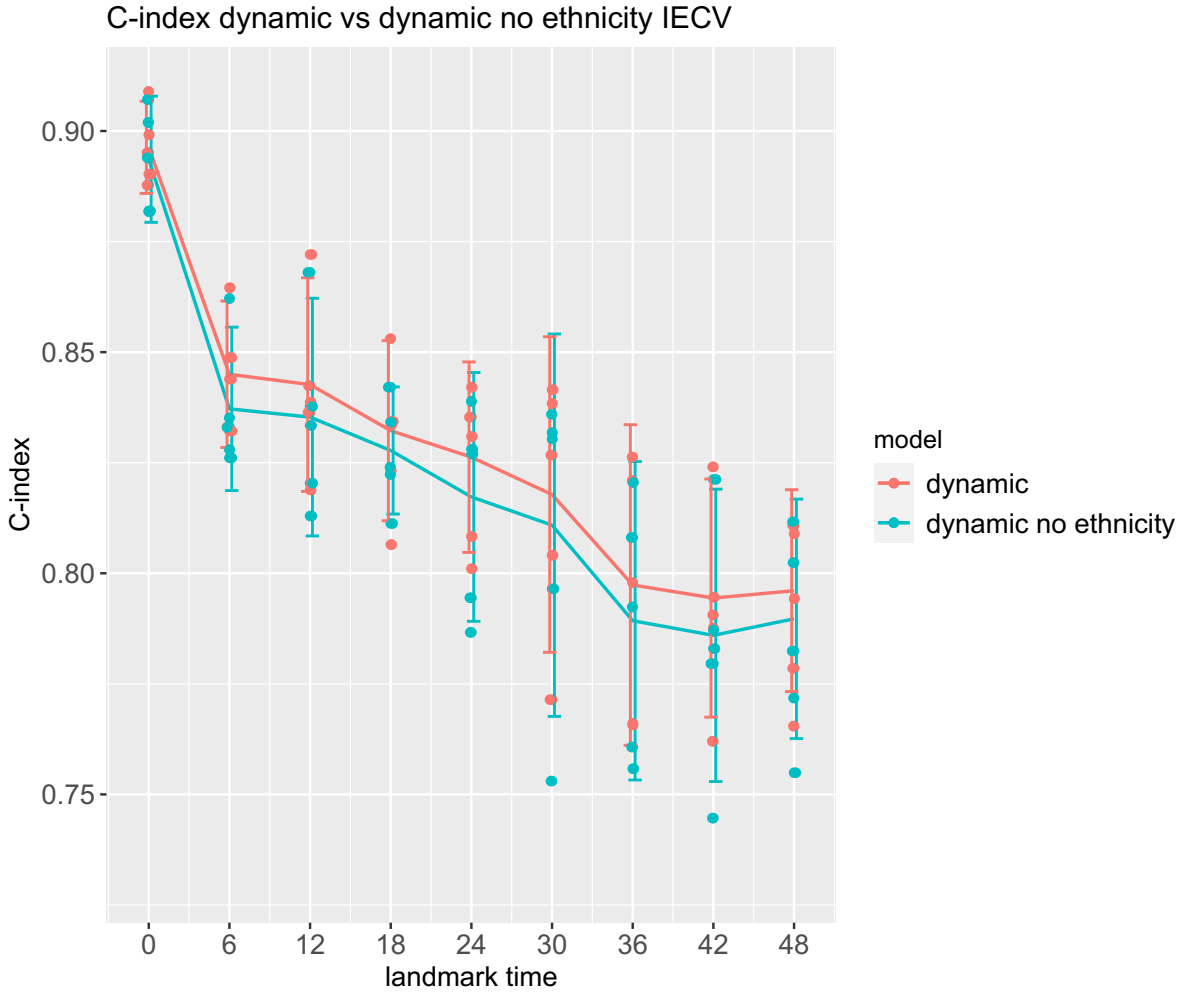
NLP predictors were included if they performed with a minimum of 80% precision threshold (eTable 4) that had been used before (Irving et al. 2021) at the time of the data extraction. However, due to routine updates in the available NLP applications, we have used an expanded set of NLP predictors compared to our previous paper as presented below:

Presented model	Irving et al. 2021*
Agitation	Agitation
Appetite loss	Appetite loss
Blunted affect	
Cannabis use	Cannabis use
Cocaine use	Cocaine use
Delusional thinking	Delusional thinking
Disturbed sleep	Disturbed sleep
Guilt	Guilt
Hallucinations (all)	
Hallucinations (auditory)	
Hallucinations (olfactory, gustatory & tactile)	
Hallucinations (visual)	
Hopelessness	Hopelessness
Insomnia	Insomnia
Irritability	Irritability
Negative symptoms	
Paranoia	Paranoia
	Poor insight
Poverty of speech	
Tearfulness	Tearfulness
Weight loss	Weight loss

* Irving J, Patel R, Oliver D, Colling C, Pritchard M, Broadbent M, *et al.* (2021): Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Schizophr Bull* 47: 405–414.

eMethods 2.8.14 Sensitivity analysis ethnicity variable removal

To evaluate the impact of the ethnicity variable on the overall model performance we have conducted a sensitivity analysis where we have removed the ethnicity variable form the model in the development and internal-external validation. The C-index results of the standard dynamic model and model without the ethnicity variable are presented below in eFigure 2.8.12. The mean difference between the standard model and the one with ethnicity removed was 0.0055 (95% CI -0.0006 - 0.0117) as quantified by the meta-regression.

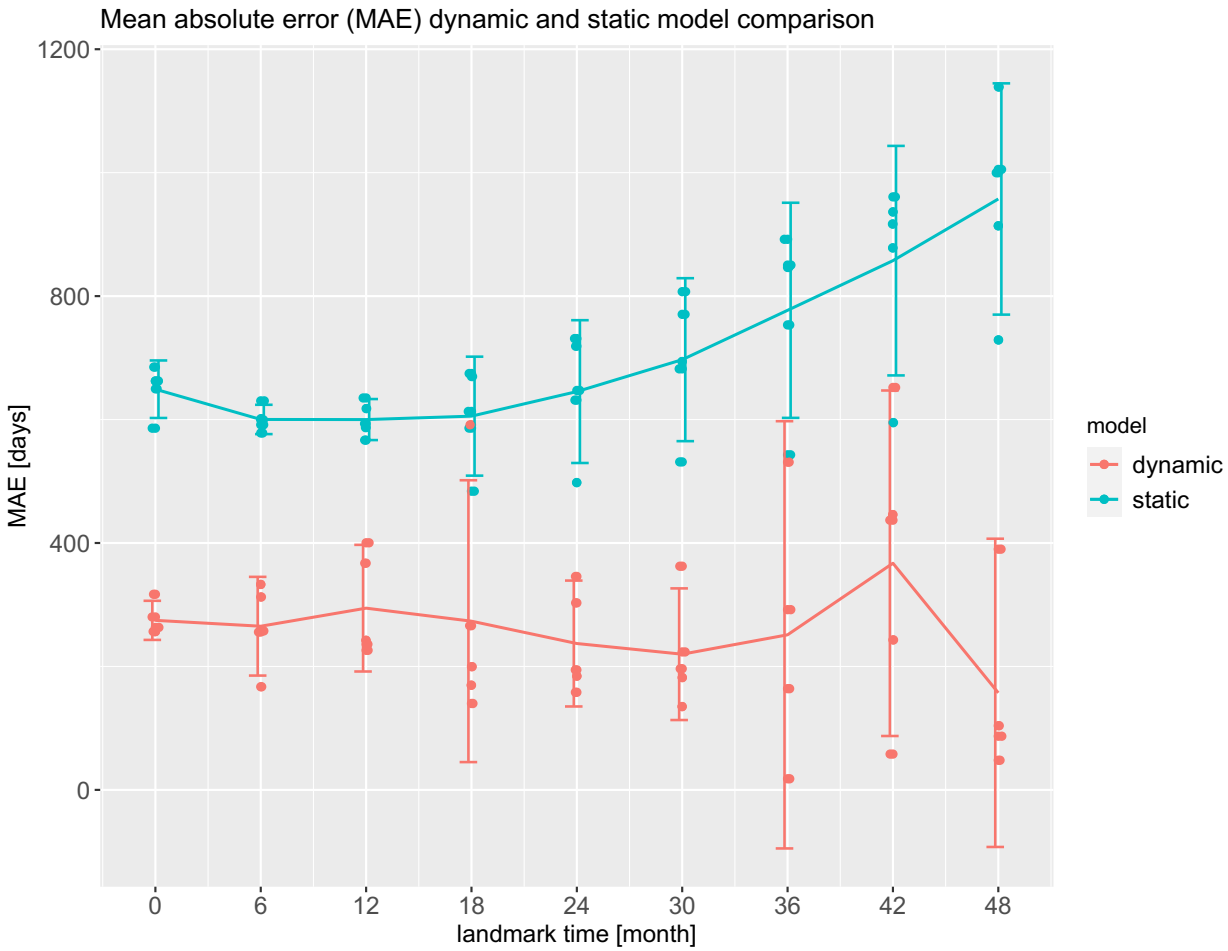


eFigure 2.8.14 C-index results for the dynamic model in internal -external cross validation. The 'dynamic' represents the unchanged model. The 'dynamic no ethnicity' represents the C-index of the dynamic model developed without the ethnicity variable. On average as indicated by the meta regression coefficient the standard dynamic model 0.0055 (95% CI -0.0006 - 0.0117) higher C-index.

eMethods 2.8.7 Mean absolute error (MAE)

To evaluate the mean absolute error (MAE) of predicted time to psychosis we computed the MAE in the internal-external validation in the same way as for the main dynamic and static model in the manuscript. We present the results for both static and dynamic model at each landmark obtained in internal-external validation. As MAE can be computed only for uncensored cases, the presented results included less than 3% of study population. Due to the low number of cases available for the MAE calculation it is difficult to reliably interpret the results.

eFigure 2.8.15



eFigure 2.8.15. Mean absolute error (MAE) results for the static and dynamic model in internal-external cross validation. Lower is better. The dynamic model has lower MAE, however due to small proportion of uncensored data it is difficult to interpret these results.

eMethods 2.8.8 Missing data imputation

Three predictors (age, gender, ethnicity) out of all used in the model had missing observations. We performed a single imputation using random forest within the Multivariate Imputation with Chain Equations framework (1) following the recommendations of Sisk et al. (2). Following (2), we excluded the outcome from imputation and developed an imputation model using training data in internal-external validation. This ensured reliable performance estimates for both train and test sets, mirroring real-world scenarios where imputation models are essential for handling missing data. The clinical predictor (diagnosis) was not missing as it was part of the inclusion criteria. The NLP predictors were recorded as present or absent in each month over the follow-up time, therefore there were no missing data associated with these variables.
Ref:

(1) Buuren S van, Groothuis-Oudshoorn K (2011): mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw* 45: 1–67.

(2) Sisk R, Sperrin M, Peek N, van Smeden M, Martin GP (2023): Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Stat Methods Med Res* 32: 1461–1477.

eTable 2.8.7 Type and precision for NLP algorithms selected in the current study

Precision values are taken from the CRIS Natural Language Processing Library (2021), and were obtained by randomly selecting *n* positive annotations from each algorithm for a specified cohort, limited to one annotation per patient ID. Precision was then calculated as the ratio of the number of relevant (true positive) instances retrieved out of the total NLP-labelled positive instances (including irrelevant [false positive] and relevant [true positive] instances) for each NLP algorithm.

NLP algorithms	Cohort	Annotations validated (n)	Precision (%)
Agitation	Random sample	100	85
Appetite loss	Random sample	100	89
Blunted affect	Random sample	100	98
Cannabis use	All patients	100	88
Cocaine use	Random sample	30	97
Delusional thinking	Random sample	100	90
Disturbed sleep	Random sample	100	89
Guilt	Random sample	100	84
Hopelessness	Random sample	100	88
Hallucinations (all)	Random sample	100	90
Hallucinations (auditory)	Random sample	100	92
Hallucinations (OTG: olfactory, tactile, gustatory)	Random sample	100	86
Hallucinations (visual)	Random sample	100	83
Insomnia	Random sample	100	97
Irritability	Random sample	100	99
Negative symptoms	Random sample	100	87
Paranoia	Random sample	100	89
Poverty of speech	Random sample	100	88
Weight loss	Random sample	100	80
Tearfulness	Random sample	100	94

eLimitations

Applying automated text extraction to Electronic Health Records (EHR) presents opportunities for enhancing prognostic accuracy. However, Natural Language Processing (NLP) tools introduce some level of noise due to the inherent impossibility of achieving 100% precision in extracting data from free text. Moreover, these tools fail to capture variations in symptom severity. The subjective nature of clinician input, influenced by structural or unconscious biases, can affect how symptoms are documented for individual patients, thus reducing output standardization. Nevertheless, our selection of NLP applications ensured an acceptable level of precision, and the symptom and substance use data derived from NLP contributed to improved prognostic accuracy, demonstrating generalizability across different contexts and timeframes. External validation is now essential to determine whether NLP tools maintain similar prognostic capabilities beyond the SLaM Trust, necessitating the creation of equivalent tools in mental health care settings worldwide with accessible EHR systems. Unfortunately, validation beyond the UK is currently unfeasible due to the unique nature of the algorithms used with CRIS data and the absence of data-sharing policies enabling their transfer to non-CRIS sites. However, our utilization of transdiagnostic symptom data could mitigate disparities in diagnostic practices across countries, thereby enhancing the model's applicability across diverse settings. Additional constraints largely stem from the original model and have been addressed in prior publications (1-2). For instance, forthcoming adequate size trials assessing effectiveness will be necessary to evaluate whether this transdiagnostic risk estimator enhances outcomes for individuals experiencing their first psychotic episode. Furthermore, there is no reference standard for psychosis diagnoses recorded in EHRs. This creates potential for the reification of systemic biases in patient factors and service factors, such as overrepresentation of people of minority ethnic or socially deprived backgrounds, and issues relating to reimbursement. It is essential that future efforts to incorporate risk algorithms into psychiatric clinical practice are accompanied by user instructions detailing the potential for bias and further research into evaluating potential algorithmic bias.

Ref:

(1) Irving, Jessica, et al. "Using natural language processing on electronic health records to enhance detection and prediction of psychosis risk." *Schizophrenia bulletin* 47.2 (2021): 405-414.

(2) Fusar-Poli, P., Rutigliano, G., Stahl, D., Davies, C., Bonoldi, I., Reilly, T. and McGuire, P., 2017. Development and validation of a clinically based risk calculator for the transdiagnostic prediction of psychosis. *JAMA psychiatry*, *74*(5), pp.493-500.

eTable 2.8.8 Primary index diagnoses of non-organic and non-psychotic mental disorder

Diagnostic group	ICD-10 code	ICD-10 diagnosis name
ARMS	APS BLIPS GRD	At Risk Mental State, Attenuated Psychosis Syndrome Subgroup At Risk Mental State, Brief and Limited Intermittent Psychotic Symptoms At Risk Mental State, Genetic Risk and Deterioration syndrome
Acute and transient psychotic disorders	F23.x	Acute and transient psychotic disorders
Substance use disorders	F10 (excluding *.5, *.4, *.7) F11 (excluding *.5, *.4, *.7) F12 (excluding *.5, *.4, *.7) F13 (excluding *.5, *.4, *.7) F14 (excluding *.5, *.4, *.7) F15 (excluding *.5, *.4, *.7) F16 (excluding *.5, *.4, *.7) F17 (excluding *.5, *.4, *.7) F18 (excluding *.5, *.4, *.7) F19 (excluding *.5, *.4, *.7)	Non psychotic mental and behavioural disorders due to use of alcohol Non psychotic mental and behavioural disorders due to use of opioids Non psychotic mental and behavioural disorders due to use of cannabinoids Non psychotic mental and behavioural disorders due to use of sedatives or hypnotics Non psychotic mental and behavioural disorders due to use of cocaine Non psychotic mental and behavioural disorders due to use of other stimulants, including caffeine Non psychotic mental and behavioural disorders due to use of hallucinogens Non psychotic mental and behavioural disorders due to use of tobacco Non psychotic mental and behavioural disorders due to use of volatile solvents Non psychotic mental and behavioural disorders due to multiple drug use and use of other psychoactive substances
Bipolar mood disorders	F31.x (excluding F31.2 and F31.5) F34.0 F30.x (excluding *.2)	Non psychotic bipolar disorder Cyclothymia Non psychotic mania or hypomania
Non bipolar mood disorders	[F32-F33].x (excluding F32.3 and F33.3) F34.1 F34.8, F34.9, F38.x, F39	Non psychotic depressive disorder Dysthymia Unspecified mood disorders
Anxiety disorders	F40.x F41.0 F41.1 F41.2-F41.9 F42.x F43.x F44.x F45.x	Phobic anxiety disorders Panic disorder Generalized anxiety disorder Other anxiety disorders F42.x Obsessive compulsive disorders Reaction to severe stress, and adjustment disorders Dissociative [conversion] disorders Somatoform disorders

Personality disorders	F60.0 F60.1 F60.2 F60.3 F60.4 F60.5 F60.6 F60.7 F60.8-F60.9, F61, F62.x, F68.x F69 F21 F63.x F64.x, F65.x, F66.x	Paranoid personality disorder Schizoid personality disorder Dissocial personality disorder Emotionally unstable personality disorder Histrionic personality disorder Anankastic personality disorder Anxious [avoidant] personality disorder Dependent personality disorder Other personality disorders Schizotypal Disorder Habit and impulse disorders Sexual disorders
Developmental disorders	F80.x F81.x, F82, F83 F84.x F88, F89	Specific developmental disorders of speech and language Other specific developmental disorders Pervasive developmental disorders Other and unspecified disorders of psychological development
Physiological syndromes	F50.x F51.x F52.x F53.x (excluding F53.1) F54.x, F55, F59	Eating disorders Nonorganic sleep disorders Sexual dysfunction, not caused by organic disorder or disease Non psychotic Mental and behavioural disorders associated with the puerperium, not elsewhere classified Other physiological syndromes
Childhood/adolescence onset disorders	F90.x F91.x F92.x, F93.x, F94.x, F98.x F95.x	Hyperkinetic disorders Conduct disorders Other emotional and behavioural disorders with childhood or adolescence onset Tic disorders
Mental retardation	F70.x F71.x F72.x F73.x F78.x F79.x	Mild mental retardation Moderate mental retardation Severe mental retardation Profound mental retardation Other mental retardation Unspecified mental retardation

PART B: INDIVIDUALISING FIRST-LINE ANTIPSYCHOTICS PRESCRIPTIONS WITH PRECISION TREATMENT RULES INCORPORATING PATIENTS' PREFERENCES

3. Introduction

3.1 Pharmacology in first episode of psychosis

Psychotic disorders most commonly develop in adolescence and early adulthood with the highest risk of onset occurring around the age of 20.5 years (1). Emergent psychosis often disrupts maturation of young people, impairs their social and family relationships, and negatively impacts their education and employment outcomes, contributing to substantial personal and societal burden (2). Psychotherapy, psychosocial interventions and antipsychotic medications are recommended by the clinical guidelines for the management of first-episode cases (3,4). Pharmacological treatments with first- and second-generation antipsychotics have been demonstrated to be effective and there are numerous prescribing guidelines related to psychotic disorders recommending their use for the first-episode patients (5,6). The evidence for the superiority of a specific antipsychotic drug for first-episode patients is inconclusive (7). Antipsychotics are well known to differ significantly by their side effects profiles, however their efficacy is believed to be similar (8). Guidelines recommend the use of second- and first-generation antipsychotics (other than clozapine) as a class for the first episode patients and do not specify which specific medication should be prescribed as the first-line treatment (5). Clozapine is reserved for patients with treatment-resistant schizophrenia, after two unsuccessful trials with other antipsychotics. The general principles of the prescribing guidelines emphasize shared decision-making (9) and the importance of involving the patient and their careers in the decision-making process. The patient should be clearly informed about the available treatment options and the risks and benefits associated with each treatment plan. The patient should be able to provide feedback on the ongoing effects of the treatment, and the decisions should be adjusted accordingly. The second general principle is that the pharmacological treatments should be individualised to each patient (10). The specific treatment should be selected based on the current state of the patient's symptoms and their preferences regarding the side effects and efficacy profiles of each treatment option. Various patient-specific factors such as age, gender, ethnicity, comorbidities and substance use information should be included in the decision.

The selection of the first-line antipsychotic for the patients with no prior exposure to antipsychotic treatment (antipsychotics naïve) is a very challenging task as there is no prior history to guide decision-making. Most patients discontinue their first medication within a year (11) and the process of finding the suitable treatment option often relies on an iterative trial and error process. What complicates the treatment decision even further is the need to integrate shared decision-making and individual patient features and preferences into the process. Many patients do not feel sufficiently involved in the decision-making (12) and clinicians often do not implement shared decision-making due to lack of time and heavy clinical workload (13). Clinicians often doubt the patient's capacity to make decisions (14), and the fact that first-episode patients can present with cognitive impairments and lack of insight at different stages of the disease poses additional challenges. The prescribing guidelines do not provide a clear procedural algorithm for handling this complex decision-making process, and patients and clinicians may make suboptimal decisions.

3.2 Decision support systems for psychosis

Clinical decision support systems CDSS have been introduced to provide situation-specific advice to healthcare professionals and patients facing complex treatment decisions (15). The CDSS can be

categorised into systems aiding medication dosing, providing ordering of facilitators, systems for point-of-care alerts and reminders, dashboards displaying relevant information, expert systems, and healthcare workflow support (16). Despite potential to improve care and reduce costs in psychiatry and mental health the use of CDSS systems remains rare (17). In psychosis psychopharmacology, six unique computer or paper-based systems to aid antipsychotic selection process have been identified (18). The main aim of these tools (19–24) is to present the evidence on the efficacy and side effects of different antipsychotic options for patients and clinicians and to enhance the shared decision-making. In two of the systems (23,24), carers are also included as part of the target audience. A detailed comparison of these tools is presented in Table 3.1. The COMPASS (20) and "The Personal Antipsychotic Choice Index" (22) support tools were designed specifically for first-episode patients, whereas the remaining tools were intended for long-term psychosis or general psychotic disorders. The evidence base of the available support systems includes meta and network analyses, randomised trials, guidelines, expert opinion and pre-clinical studies. These sources, especially the meta-analyses and randomised trials, can provide high-quality evidence; however, this is limited to aggregate summaries and average treatment effects at the population level. Although individualised prescribing is one of the main principles of clinical guidelines, all analysed support systems lack the capability to generate personalised treatment recommendations. The evidence they rely on does not allow for identification of heterogeneous treatment effects (HTEs), stratified by individual patient characteristics. Moreover, the tools do not use appropriate algorithms or statistical methods to generate such individualised recommendations based on patient-level data.

System name	**Clinical target population**	**Intended users**	**Evidence base**	**Average treatment effects**	**Heterogenous Treatment Effects**	**First author, year, reference**
WEGWEIS	First-episode Psychosis, Long-term Psychosis	Patients	Guidelines, Expert Opinion, Service User Experiences	Yes	No	van der Krieke, 2012, (19)
COMPASS	First-episode Psychosis	Clinicians, Patients	Literature Review	Yes	No	Mueser, 2015, (20)
TREAT	General Psychotic Disorders	Clinicians, Patients	Guidelines, Expert Opinion	Yes	No	Tasma, 2018, (21)
The Personal Antipsychotic Choice Index	First-episode Psychosis	Clinicians, Patients	Cochrane Systematic Review, Network Meta-Analysis, Expert Opinion	Yes	No	van Dijk, 2018, (22)
Encounter Decision Aid	First-episode Psychosis, Long-term Psychosis	Clinicians, Patients, Care Givers	Systematic Review, Meta-analysis, Guidelines, Randomised Trials	Yes	No	Zisman-Ilani, 2018, (23)
In Control of Effects	Not given	Clinicians, Patients, Care Givers	Network Meta-Anlysis	Yes	No	Henshall, 2019, (24)

Table 3.1. Decision support systems in psychosis. Comparison of the main characteristics and evidence base. All the analysed systems do not rely on evidence that supports identification of heterogenous treatment effects.

3.3 Barriers to the individualisation of treatment selection

There are numerous barriers to the individualisation of treatment selection in psychosis. First, the available data from the randomised controlled trials (RCTs) are limited in size, and especially so for the first-episode group. The stable identification of HTEs requires significantly larger sample sizes than those needed to estimate average treatment effects. It has been estimated that for the stable estimation of between-patient differences in optimal treatment for major depressive disorder, a sample size of 300 patients per treatment arm is required (25). Even large RCTs investigating antipsychotics in first-episode patients that are performed in multiple centres and in many countries include at most 100 patients per treatment arm (26). In contrast to other fields of medicine, such as cardiology, where very large trials have been conducted (27), mental health research lacks appropriate funding and support to carry out such trials.

The second barrier to individualisation is the lack of consistency in measurements between trials. One way of overcoming the small sample size of individual trials is to combine them and perform an analysis on the pooled data. This approach has been tried in depression and post-traumatic stress disorder research (28,29). However, it was mostly unsuccessful, due to insufficient and inconsistent inclusion of key predictors of HTEs across mental disorder trials. To the best of my knowledge, there are no individual-level meta-analyses of RCTs involving patients experiencing a first episode of psychosis. Therefore, different solutions are needed.

The emerging solution to overcome these barriers is to use patient data routinely collected in clinical practice stored in Electronic Health Records (EHRs). EHRs offer access to very large sample sizes and include a wide array of potential HTE predictors. However, this approach presents a different challenge. Unlike RCTs, EHRs are based on observational data, rather than experimental design. In clinical practice, clinicians prescribe medications based on individual patient characteristics, such as age, gender, ethnicity and location, which can result in unequal distribution of those characteristics across different treatment groups. These characteristics can also be prognostic of clinical outcomes, and lead to confounding bias (30). To address this challenge, multiple statistical techniques have been developed to adjust for measured differences between treatment groups. These techniques range from multivariable regression and matching to inverse probability weighting and doubly robust methods (31). However, in any observational study design, some confounders may not be observed; therefore, any causal inference requires the assumption that all important variables influencing both treatment selection and outcomes have been observed, referred to as the no unmeasured confounding assumption (31).

Another challenge of using observational data for causal inference is the risk of immortal time bias (32) and selection bias due to differential loss to follow-up (33,34). Immortal time bias can arise from misalignment between treatment initiation, eligibility determination, and follow-up periods. In an RCT, baseline characteristics assessment and treatment initiation occur exactly at the time of randomisation; this is also the starting point of the follow-up period. In contrast, in observational studies aiming to estimate the treatment effect of exposure versus non-exposure, there may be a delay between the time at which eligibility criteria are met and the time of treatment initiation. This may create a follow-up period during which participants in the exposure group could not have experienced the outcome (e.g., disease onset or death). This period is referred to as ‘immortal time’, and inappropriate design of the observational study can lead to ‘immortal time bias’. Another selection bias may arise when different exposure groups differ systematically in their

rates of loss to follow-up. Loss to follow-up bias also must be addressed in the design and analysis of observational studies.

3.4 Pragmatic precision psychiatry

To address these challenges, including the limited sample size and measurement inconsistencies between trials, as well as the methodological challenges of observational EHR data, a pragmatic precision psychiatry approach has been proposed (35). The advocated approach is to develop precision treatment rules (PTRs) using a trial emulation design and observational data from EHRs, and then prospectively validate the PTRs in a pragmatic trial. The development and subsequent validation in a pragmatic trial are intended as an iterative process cycling between the observational and experimental designs to both test and refine the PTRs.

Trial emulation was introduced by Hernán and Robins in 2016 (36), and is a crucial component of this approach, as it presents a systematic way to design an observational study that can address selection bias. The trial emulation framework can be summarised in three main steps. The first step is a clear definition of the causal question regarding treatment options. The second step involves specifying eligibility criteria, defining the treatment strategies to compare, identifying the treatment assignment procedure, and establishing the start of follow-up, which mirrors the design of an RCT. In the final step, the observational data meeting these criteria are analysed using appropriate causal inference methods.

The second crucial component of pragmatic precision psychiatry is the use of PTRs. The main aim of PTRs is to map patient characteristics to the optimal treatment. To achieve this, the first step in creating the PTRs is to estimate treatment effects for a given patient under each of the competing treatment options and then select the treatment that is expected to yield the most benefit or minimise harm. This problem is closely related to the estimation of HTEs, and numerous methodologies exist for estimating HTEs using observational data. The three most frequently used approaches are treatment specific modelling, risk modelling, and effect modelling. The treatment specific models are fitted exclusively on participants that received a given treatment, and therefore they can not to be used to compare the treatment group with the no-exposure group. As such these models are not recommended for the construction of PTRs (35). The risk modelling approach consists of fitting the models on participants prior to treatment exposure. Then the estimated risk of an outcome can be used to analyse the effects of treatment across different risk strata. The rationale of this approach is that individuals with lower risk of an outcome are likely to experience lower benefit from the treatment, and conversely individuals with higher risk of an outcome may experience higher benefits. The risk modelling approach relies on estimation of a single interaction between the risk score and the treatment to evaluate the HTE, and as such can be especially useful in small sample sizes. A different approach is needed to capture multiple interactions between prescriptive predictors and treatments. The effect modelling approach estimates a separate interaction for each prescriptive predictor and then combines them to create the PTRs. In effect modelling the risk score from the risk model can be included as one of the predictors. The PTRs are constructed using the counterfactual logic, wherein the effects estimated by the multi-coefficient model are compared between each treatment option for each patient given their individual-level data.

Effect modelling of PTRs can most fully realise the aim of precision medicine, that is, individualisation of treatment based on patient-specific characteristics. It is also the most

recommended approach among the three. However, estimating multiple interactions can be prone to instability and overfitting. Therefore, sample-splitting methods such as cross-fitting are need when estimating the PTRs with effects modelling.

3.5 Estimation of HTEs in observational data with causal machine learning

The counterfactual framework is the backbone of causal inference methods that are used in observational studies and in developing PTRs. It is defined by the potential outcomes framework (37) that postulates the existence of potential outcomes $Y_i(0)$ and $Y_i(1)$ which record the hypothetical outcome Y_i for a given patient i in two different treatment states. The treatment indicator is denoted as W and in the simplest case it can take two values $W_i = 0$, referring to the control unit, and $W_i = 1$, referring to the treated unit. Under the stable unit treatment value assumption, the average treatment effect (ATE) is defined as the difference between the two potential outcomes:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)],$$

where $\mathbb{E}$ is the average over the population. Inferring causation in observational data requires an additional assumption that all important covariates influencing both treatment assignment and outcome (confounders) have been measured and adjusted for. This is often referred to as the no unmeasured confounders assumption, also known as the unconfoundedness assumption (38). This assumption is untestable; however, sensitivity analyses exist to assess the sensitivity of the results to the potential impact of unmeasured confounding (39). Numerous techniques have been developed to estimate treatment effects in observational studies. These methods are based on the central idea that, under the unconfoundedness assumption, observed differences between treatment groups can be used to recover causal effects after adjusting for observed covariates. Techniques such as matching (40) aim to construct comparable treatment and control groups by pairing units with similar covariate values, thereby reducing bias due to observable differences. Matching is intuitive and transparent, but its performance can deteriorate in high-dimensional settings, and it may discard a substantial portion of the data. Multiple regression adjusts for confounding by modelling the outcome as a function of treatment and covariates, offering efficiency and ease of implementation; however, it relies heavily on correct model specification and may extrapolate beyond regions of limited common support (41). Inverse probability weighting addresses confounding by reweighting observations using the estimated propensity score to create a pseudo-population in which treatment assignment is independent of covariates (42). While inverse probability weighting can consistently estimate average treatment effects under correct propensity score specification, it is sensitive to extreme weights and can suffer from high variance when overlap between treatment groups is weak (43).

The Robinson transformation, a key method discussed in this presentation, simplifies estimation by first removing the influence of covariates from both the outcome and the treatment variables (44). Assuming a plausible set of patients' characteristics to account for confounding has been collected, we can then run a linear regression relating the potential outcome to the treatment effect and the patients' covariates:

$$Y_i \sim \tau W_i + \beta X_i,$$

and interpret the τ as the ATE, given the following assumptions:

1. W_i is unconfounded given X_i(a sufficient number of important confounders has been captured).
2. The effect of X_i on the potential outcome Y_i is linear.
3. The treatment effect τ is constant.

To relax the second assumption on the liner effect of X_i on Y_i, a partially linear model can be posited:

$$Y_i = \tau W_i + f(X_i) + \epsilon_i, \qquad \mathbb{E}[\epsilon_i \mid X_i, W_i] = 0.$$

Where it is assumed that the Y_i can be expressed as a potentially complex function f that relates the covariates to the baseline outcome, and that Y_i is shifted by a constant treatment effect, τ. To estimate τ from this model, a crucial step is to use the Robinson transform (44). For that, one needs to define the propensity score, which is the probability of treatment assignment given a patient's characteristics:

$$e(x) = \mathbb{E}[W_i \mid X_i = x].$$

Secondly, the probability of an outcome given a patient's characteristics is defined as:

$$m(x) = \mathbb{E}[Y_i \mid X_i = x] = f(x) + \tau e(x).$$

The two functions, $e(x)$ and $m(x)$, are called nuisance parameters and are essential for ATE estimation in observational study design. One can now rewrite the above equation to obtain:

$$Y_i - m(x) = \tau\big(W_i - e(x)\big) + \epsilon_i.$$

This equation allows for estimation of τ using only the nuisance parameters. The terms $Y_i - m(x)$ and $W_i - e(x)$ represent the residual outcomes and treatments, respectively:

$$\tilde{Y}_i = Y_i - m(x),$$
$$\tilde{W}_i = W_i - e(x).$$

The ATE τ can be estimated using residual-on-residuals regression:

$$\hat{\tau} = \arg\min_{\tau} \sum_{i=1}^{n} (\tilde{Y}_i - \tau \tilde{W}_i)^2$$

The residual-on-residual regression, also known as orthogonal construction, removes or partials out the effect of baseline variables on the outcome, and when combined with an appropriate data splitting procedure, can provide high-quality inference on the treatment effect. It was shown (44) that this approach allows for efficient estimation of the parameter of interest, τ, even if the nuisance parameters $m(x)$ and $e(x)$ converge at a lower rate, meaning that precise estimation of τ can be obtained even when the nuisance parameters are 'noisy'.

To estimate τ in this semiparametric approach, all that is needed are good predictions of residual outcome and residual treatment assignment. This can be achieved by machine learning methods (such as random forests, boosting, or neural networks) instead of traditional regression models to improve prediction accuracy and allow for flexibility in capturing nonlinear effects. It has been shown (45) that when the modern machine learning methods are used to predict the nuisance parameters the estimated τ can be biased due to the regularisation. However, when data splitting techniques (e.g., cross-fitting) are used such that models to predict the outcome and the treatment assignment are fitted without using data from unit i for which τ is estimated, unbiased estimation can be obtained (45,46).

The limitation of the ATE is that it cannot explain whether patients respond differently to the same treatment. This heterogeneity of treatment effects can be described by identifying subgroups of patients who have different responses based on their baseline characteristics. This is defined as the Conditional Average Treatment Effect (CATE):

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x],$$

Which represents the excepted difference in outcomes between treated and untreated conditions for individuals with covariates $X_i = x$.

The partial linear model, relaxing the assumption of constant τ, is as follows:

$$Y_i = \tau(X_i)W_i + f(X_i) + \epsilon_i, \;\; \mathbb{E}[\epsilon_i \mid X_i, W_i] = 0,$$

where $f(X_i)$ captures the baseline effects of covariates, and $\tau(X_i)$ allows for the treatment effect to vary with covariates.

If one had access to some neighbourhoods, denoted by $\ell(x)$, where τ is constant, it would be possible to estimate $\tau(x)$ using the residual-on-residual method described above; however, this would involve restricting the estimation to samples belonging to the neighbourhood $\ell(x)$:

$$\hat{\tau}(\mathrm{x}) = \arg\min_{\tau} \sum_{i \in \ell(x)} \left(\tilde{Y}_i^{(-i)} - \tau \widetilde{W}_i^{(-i)} \right)^2,$$

with

$$\tilde{Y}_i^{(-i)} = Y_i - \hat{m}^{(-i)}(X_i),$$
$$\widetilde{W}_i^{(-i)} = W_i - \hat{e}^{(-i)}(X_i),$$

where supersxript $(-i)$ indicates that the nuisance parameters were fitted on a sample that excludes the unit i for which the treatment effect is estimated.

A method that uses this approach to CATE estimation and has been gaining popularity in observational studies in mental health research is causal forest (47–49). The causal forest modifies the original random forest (50) in several ways. The standard random forest builds an ensemble of threes to predict Y form X. Each tree is grown by recursively splitting the data at nodes in a way

that maximizes the homogeneity of outcomes within each child node, and the final prediction is obtained by averaging the predictions across all trees in the forest.

In casual forest node splits maximize the partition heterogeneity of treatment effect across children, rather than outcome homogeneity. Secondly, the CATE for a given X_i is estimated by the residual-on-residual regression weighted by the frequency with which the i-th training sample falls into the same leaf.

$$\hat{\tau}(\mathrm{x}) = \arg\min_{\tau} \sum_{i \in \ell(x)} \alpha_i(x)(\tilde{Y}_i^{(-i)} - \tau \tilde{W}_i^{(-i)})^2$$

where $\alpha_i(x)$ represents the weight of observation i that corresponds to how frequently it falls within the same leaf as covariate x across the forest.

Another crucial feature of the causal forest is referred to as 'honesty' (49). This involves separating the construction of tree partitions from the effect estimation within the leaves by using different samples for these two tasks. This approach allows for data-driven identification of unspecified heterogeneous subgroups while providing unbiased estimation of the CATE. This property is especially valuable in research on optimal treatment allocation, notably in settings with limited prior knowledge of prescriptive predictors.

3.6 PTRs using causal forest and shared decision-making

The causal forest offers a well-developed and conceptually rigorous way to identify heterogeneous subgroups and estimate CATE from observational data, which can be used to improve treatment selection in psychiatry. It is also well suited to provide individualised predictions of treatment effects that can support PTRs in accordance with the principle of treatment individualisation outlined in the clinical guidelines for psychosis management. Another advantage of the causal forest method is its applicability to large EHRs, which helps mitigate the problem of restricted sample sizes in RCTs. However, to meet the need for shared decision-making and patient involvement, the PTR based on the causal forest requires specific modifications. In the next chapter of this thesis, I outline how patient preferences can be incorporated into the PTR to generate treatment recommendations that satisfy both individualisation and shared decision-making criteria. Additionally, I demonstrate how multiple outcomes, reflecting the multidimensionality of antipsychotic selection, can be integrated into the causal forest framework to construct PTRs.

3.7 References

1. Solmi M, Radua J, Olivola M, Croce E, Soardo L, Salazar de Pablo G, *et al.* (2022): Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Mol Psychiatry* 27: 281–295.

2. Crespo-Facorro B, Such P, Nylander A-G, Madera J, Resemann HK, Worthington E, *et al.* (2021): The burden of disease in early schizophrenia – a systematic literature review. *Curr Med Res Opin* 37: 109–121.

3. National Institute for Health and Care Excellence (NICE) (2014): *NICEGuidelines, Psychosis and Schizophrenia in Adults: Prevention and Management Clinical Guideline.* London, UK.

4. Keepers GA, Fochtmann LJ, Anzia JM, Benjamin S, Lyness JM, Mojtabai R, *et al.* (2020): The American Psychiatric Association Practice Guideline for the Treatment of Patients With Schizophrenia. *Am J Psychiatry* 177: 868–872.

5. Taylor, David M, Thomas RE Barnes,, Allan H. Young (2021): *The Maudsley Prescribing Guidelines in Psychiatry 14th Edition*. Chichester UK: John Wiley & Sons.

6. Keating D, McWilliams S, Schneider I, Hynes C, Cousins G, Strawbridge J, Clarke M (2017): Pharmacological guidelines for schizophrenia: a systematic review and comparison of recommendations for the first episode. *BMJ Open* 7: e013881.

7. Y Z, C L, M H, P R, M K, I B, *et al.* (2017): How well do patients with a first episode of schizophrenia respond to antipsychotics: A systematic review and meta-analysis. *Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol* 27. https://doi.org/10.1016/j.euroneuro.2017.06.011

8. Leucht S, Schneider-Thoma J, Burschinski A, Peter N, Wang D, Dong S, *et al.* (2023): Long-term efficacy of antipsychotic drugs in initially acutely ill adults with schizophrenia: systematic review and network meta-analysis. *World Psychiatry* 22: 315–324.

9. Hamann J, Leucht S, Kissling W (2003): Shared decision making in psychiatry. *Acta Psychiatr Scand* 107: 403–409.

10. Guinart D, Fagiolini A, Fusar-Poli P, Giordano GM, Leucht S, Moreno C, Correll CU (2024): On the Road to Individualizing Pharmacotherapy for Adolescents and Adults with Schizophrenia – Results from an Expert Consensus Following the Delphi Method. *Neuropsychiatr Dis Treat*.

11. Szmulewicz AG, Martínez-Alés G, Logan R, Ferrara M, Kelly C, Fredrikson D, *et al.* (2024): Antipsychotic drugs in first-episode psychosis: a target trial emulation in the FEP-CAUSAL Collaboration. *Am J Epidemiol* 193: 1081–1087.

12. Do Patients With Schizophrenia Wish to Be Involved in Decisions About Their Medical Treatment? | American Journal of Psychiatry (n.d.): Retrieved July 31, 2025, from https://psychiatryonline.org/doi/full/10.1176/appi.ajp.162.12.2382

13. Kal EF (2009): Paternalism or lack of time? *Psychiatr Serv Wash DC* 60: 1403.

14. Hamann J, Mendel R, Cohen R, Heres S, Ziegler M, Bühner M, Kissling W (2009): Psychiatrists' use of shared decision making in the treatment of schizophrenia: patient characteristics and decision topics. *Psychiatr Serv Wash DC* 60: 1107–1112.

15. Berner ES (Ed.) (2016): *Clinical Decision Support Systems*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-31913-1

16. Beeler PE, Bates DW, Hug BL (2014): Clinical decision support systems. *Swiss Med Wkly* 144: w14073.

17. Tong F, Lederman R, D'Alfonso S (2025): Clinical decision support systems in mental health: A scoping review of health professionals' experiences. *Int J Med Inf* 199: 105881.

18. K M, F S, A R, S S, S L, J H (2023): How should patient decision aids for schizophrenia treatment be designed? - A scoping review. *Schizophr Res* 255. https://doi.org/10.1016/j.schres.2023.03.025

19. van der Krieke L, Emerencia AC, Boonstra N, Wunderink L, de Jonge P, Sytema S (2013): A web-based tool to support shared decision making for people with a psychotic disorder: randomized controlled trial and process evaluation. *J Med Internet Res* 15: e216.

20. Mueser KT, Penn DL, Addington J, Brunette MF, Gingerich S, Glynn SM, *et al.* (2015): The NAVIGATE Program for First-Episode Psychosis: Rationale, Overview, and Description of Psychosocial Components. *Psychiatr Serv Wash DC* 66: 680–690.

21. Tasma M, Roebroek LO, Liemburg EJ, Knegtering H, Delespaul PA, Boonstra A, *et al.* (2018): The development and evaluation of a computerized decision aid for the treatment of psychotic disorders. *BMC Psychiatry* 18: 163.

22. van Dijk F, de Wit I, Blankers M, Sommer I, de Haan L (2018): The Personal Antipsychotic Choice Index. *Pharmacopsychiatry* 51: 89–99.

23. Zisman-Ilani Y, Shern D, Deegan P, Kreyenbuhl J, Dixon L, Drake R, *et al.* (2018): Continue, adjust, or stop antipsychotic medication: developing and user testing an encounter decision aid for people with first-episode and long-term psychosis. *BMC Psychiatry* 18: 142.

24. Henshall C, Cipriani A, Ruvolo D, Macdonald O, Wolters L, Koychev I (2019): Implementing a digital clinical decision support tool for side effects of antipsychotics: a focus group study. *Evid Based Ment Health* 22: 56–60.

25. Luedtke A, Sadikova E, Kessler RC (2019): Sample Size Requirements for Multivariate Models to Predict Between-Patient Differences in Best Treatments of Major Depressive Disorder. *Clin Psychol Sci* 7: 445–461.

26. Fleischhacker WW, Keet IPM, Kahn RS, EUFEST Steering Committee (2005): The European First Episode Schizophrenia Trial (EUFEST): rationale and design of the trial. *Schizophr Res* 78: 147–156.

27. null null (2015): A Randomized Trial of Intensive versus Standard Blood-Pressure Control. *N Engl J Med* 373: 2103–2116.

28. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, *et al.* (2021): Internet-Based Cognitive Behavioral Therapy for Depression: A Systematic Review and Individual Patient Data Network Meta-analysis. *JAMA Psychiatry* 78: 361–371.

29. Qi W, Ratanatharathorn A, Gevonden M, Bryant R, Delahanty D, Matsuoka Y, *et al.* (2018): Application of data pooling to longitudinal studies of early post-traumatic stress disorder (PTSD): the International Consortium to Predict PTSD (ICPP) project. *Eur J Psychotraumatology* 9: 1476442.

30. Meuli L, Dick F (2018): Understanding Confounding in Observational Studies. *Eur J Vasc Endovasc Surg* 55: 737.

31. Hernan MA, Robins JM (n.d.): Causal Inference: What If.

32. Lévesque LE, Hanley JA, Kezouh A, Suissa S (2010): Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ* 340: b5087.

33. Ma H, S H-D, Jm R (2004): A structural approach to selection bias. *Epidemiol Camb Mass* 15. https://doi.org/10.1097/01.ede.0000135174.63482.43

34. Scola G, Chis Ster A, Bean D, Pareek N, Emsley R, Landau S (2023): Implementation of the trial emulation approach in medical research: a scoping review. *BMC Med Res Methodol* 23: 186.

35. Kessler RC, Luedtke A (2021): Pragmatic Precision Psychiatry-A New Direction for Optimizing Treatment Selection. *JAMA Psychiatry* 78: 1384–1390.

36. Hernán MA, Robins JM (2016): Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol* 183: 758–764.

37. Gudio W. Imbens, Donald B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*. Harvard University, Massachusetts.

38. Rubin DB, Imbens GW (Eds.) (2015): Assessing Unconfoundedness. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, pp 479–495.

39. D'Agostino McGowan L (2022): Sensitivity Analyses for Unmeasured Confounders. *Curr Epidemiol Rep* 9: 361–375.

40. Abadie A, Imbens G (2006): Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica* 74: 235–267.

41. Shi AX, Zivich PN, Chu H (2024): A Comprehensive Review and Tutorial on Confounding Adjustment Methods for Estimating Treatment Effects Using Observational Data. *Appl Sci* 14. https://doi.org/10.3390/app14093662

42. Hirano K, Imbens GW, Ridder G (2003): Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71: 1161–1189.

43. Mao H, Li L, Greene T (2019): Propensity score weighting analysis and treatment effect discovery. *Stat Methods Med Res* 28: 2439–2454.

44. Robinson PM (1988): Root-N-Consistent Semiparametric Regression. *Econometrica* 56: 931–954.

45. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018): Double/debiased machine learning for treatment and structural parameters. *Econom J* 21: C1–C68.

46. Zheng W, van der Laan MJ (2011): Cross-Validated Targeted Minimum-Loss-Based Estimation. In: van der Laan MJ, Rose S, editors. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer, pp 459–474.

47. Athey S, Tibshirani J, Wager S (2019): Generalized random forests. *Ann Stat* 47: 1148–1178.

48. Tibshirani J, Athey S, Sverdrup E, Wager S (2024): grf: Generalized Random Forests. p 2.4.0.

49. Athey S, Imbens G (2016): Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci* 113: 7353–7360.

50. Breiman L (2001): Random Forests. *Mach Learn* 45: 5–32.

4. Development and validation of a precision treatment rules for first-line antipsychotic recommendations in first episode of psychosis jointly incorporating effectiveness, side effects and patient preferences.

Short title: Precision treatment rules in first-episode psychosis

Accepted for publication in Translational Psychiatry

Kamil Krakowski[1,2], Dominic Oliver[2,3,4,5], Maite Arribas[2], Yanakan Logeswaran[2,7], Andrea de Micheli[2], Rashmi Patel[3,8], Daniel Stahl*[7], Paolo Fusar-Poli*[1,2,9,10]

[1] Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy;
[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] Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK;
[4] Department of Psychiatry, University of Oxford, Oxford, UK;
[5] NIHR Oxford Health Biomedical Research Centre, Oxford, UK;
[6] OPEN Early Detection Service, Oxford Health NHS Foundation Trust, Oxford, UK;
[7] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK;
[8] Department of Psychiatry, University of Cambridge, Cambridge, UK;
[9] OASIS Service, South London and the Maudsley National Health Service Foundation Trust, London, UK;
[10] Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University Munich, Germany.
*Daniel Stahl and Paolo Fusar-Poli are shared last authors

Corresponding author: Kamil Krakowski (kamil.krakowski01@universitadipavia.it)

4.1 Abstract

Selecting first-line antipsychotic medication for first episode of psychosis patients is a very challenging task requiring the clinicians to empirically weight multiple criteria. Precision treatment rules developed using health records offer a pragmatic approach to support clinicians' treatment selection, however, they don't incorporate side effects and patient preferences. We used Electronic Health Records from Early Intervention for Psychosis services in South London and the Maudsley NHS Trust and followed the RECORD and TRIPOD+AI guidelines. Precision treatment rules were developed using causal machine learning methods and estimated effectiveness (change of medication, hospitalisation) and side effects (extrapyramidal side effects, hyperprolactinemia, sedation, sexual side effects, and weight gain) using clinical, demographic, symptom and substance use predictors. Patient preferences regarding side effects were incorporated by ranking method. 1709 patients (mean age 26.7 years and 64% male) were included. If treatment rules were implemented, hyperprolactinemia would be reduced by 4.7 percentage points (pp), sedation by 15.8 pp, sexual side effects by 4.3 pp and weight gain by 15.2 pp with no change in hospitalisation and change of medications outcomes. However, extrapyramidal side effects were estimated to increase by 5.5 pp. Aripiprazole was recommended to between 80% and 98% of patients depending on selected patients' preferences. This study presents the first precision treatment rules for early psychosis that integrate effectiveness, side effects and patient preferences. If replicated, an implementation could substantially reduce side effects, support personalized medicine and shared decision-making.

4.2 Introduction

Psychosis is a highly burdensome disorder causing reduced life expectancy, and substantial suffering worldwide. The management of first episode of psychosis by second generation antipsychotics is the mainstream treatment recommended by clinical guidelines (1). Choosing the specific first-line medication for antipsychotic naïve patients is a very challenging task, and the process often relies on trial and error, with the majority of patients discontinuing the first medication within a year (2). It is recommended that the specific antipsychotic should be selected in a shared decision making framework (3) and consider effectives (4), side effects profiles (5), individual patient's characteristics and their preferences (1,6). Integrating all these domains on a case-by-case basis in routine care is very challenging and clinical decision support systems (CDSS) offer a potential solution to optimize treatment selection and improve patients' outcomes. Recent scoping review (7) lists several CDSS for schizophrenia or first episode of psychosis management developed using evidence from randomised experiments. However, these approaches do not incorporate individual patient's characteristics and don't allow for individualised treatment selection. This can be achieved with the precision medicine framework (8) as operationalized by the precision treatment rules (PTRs) (9) which map patient's characteristics to treatments with the greatest estimated benefit. Due to the required sample size and representativeness, the development of PTRs in Electronic Health Records (EHRs) with the machine learning and causal inference methods is recommended as a pragmatic approach for optimising treatment selection in psychiatry (10). To date only one PTR using this framework was developed for psychosis or schizophrenia management (11). The PTR was estimated to reduce hospitalisations and changes of medication (11), however it did not incorporate side effects nor patient preferences to enable share-decision making.
In this study, we aimed to develop and validate PTRs for first-line antipsychotic treatment selection in first episode psychosis that integrate effectiveness, side effects, individual patient characteristics and their preferences using causal machine learning approaches. We also provide estimates of the average treatment effects across pairs of the three most used antipsychotics on the risk of changing medication, hospitalisation, and five common antipsychotics' side effects.

4.3 Methods

4.3.1 Data source

This is a cohort study that used data from the EHRs at South London and Maudsley (SLaM) National Health Service (NHS) Trust in the UK (eMethods 1). SLaM is a specialist secondary care mental health provider to an area of 1.3 million inhabitants in South London. Patients' records at SLaM are fully digitalised and anonymised by the Clinical Record Interactive Search (CRIS) to facilitate research (12). CRIS received ethical approval as an anonymised dataset for secondary analyses from Oxfordshire REC C (Ref: 23/SC/0257).

Early Intervention Services (EIS) for patients with first episode of psychosis run at SLaM to improve detection and offer quick psychological and pharmacological care (13). Data inclusion criteria consisted of acceptance to the SLaM EIS between 1st April 2008 and 31st March 2019 and prescription of one of the three second generation antipsychotics (Aripiprazole, Olanzapine, and Risperidone) at or above the minimum effective dose for psychosis (1). We restricted the analysis

to three most prescribed antipsychotics to ensure the recommended by Luedtke et. al sample size requirement of 300 patients per medication arm (14). The start of the follow-up was defined as the date of accepted referral to the SLaM EIS and the end date was 31st March 2021. Patients with missing data on gender and ethnicity were excluded from the cohort. The flow chart of patients included in the study is presented in eMethods 2. We linked the included patient's EHRs with the side effects data that were manually extracted from unstructured EHRs' free text for a previous study (15). The manual extraction was performed by two trained mental health researchers and discrepancies were resolved by a trained psychiatrist (RP). We followed the TRIPOD+AI (16) and RECORD (17) guidelines.

4.3.2 Outcome

This is a multi-outcome study with 7 outcomes in total. Two measures of effectiveness were selected: (i) changing a medication within 2 years; (ii) hospitalisation within 2 years. As with previous studies (2,11) all cause hospitalisation was included as psychosis is associated with increased risk of both physical and mental health problems (18). Changing medication was included as it indicates lack of effectiveness or major side effects.
Additionally, manually extracted (15) clinician-recorded side effects associated with antipsychotics use as documented in the Maudsley Prescribing Guidelines (1) were included as outcomes. These were (iii) extrapyramidal side effects (EPSE), (iv) hyperprolactinemia, (v) sedation, (vi) sexual side effects, and (vii) weight gain. The extracted side effects were in binary (yes/no) format and indicate presence or absence during the administration period of the first antipsychotic prescribed at EIS.

4.3.3 Predictors

We extracted the following predictors form the structured EHRs to form an adjustment set intended to account for confounding and allow for interpretation of causal effects: (i) demographic predictors: age at index date, sex, self-assigned ethnicity (eTable 1); (ii) clinical predictor: index ICD-10 diagnostic spectra (eTable 1). We also extracted (iii) symptoms and substance use information by text mining algorithms - Natural Language Processing (NLP) form the unstructured clinical notes in the EHRs (eMethods 3, eTable 2). NLP based predictors were included if they performed with a minimum of 80% precision (19) at the time of the data extraction as used in previous studies (19,20). We included 66 NLP based predictors (see eTable 2 for the full list). By including this set of covariates, we aim to block all back-door paths between treatment and outcome, satisfying the assumptions required for valid causal inference.

4.4 Analysis Methods

4.4.1 Average treatment effects of antipsychotics

We estimated the average treatment effect (ATE) for each of the seven individual outcomes across every pair of antipsychotics in the cohort. To achieve this, we fitted a separate model for each outcome independently, allowing the relationships between predictors and each outcome to vary freely across models. ATE reflects the estimated average difference in outcome if all patients received one treatment versus another and it is typically estimated in randomized controlled trials. In the absence of RCTs, observational data can be used to estimate ATE given the assumption that

non-random treatment assignment can be adjusted by controlling for the baseline covariates (21). We used Causal Forest (CF) the special case of the Generalised Random Forest (GRF) (22) to estimate the effects of antipsychotic treatments as it allows for adjustment of baseline differences between treatment groups. CF is a causal machine learning approach that modifies the Random Forest (RF) (23) algorithm to estimate conditional average treatment effects (CATE). The method uses predictions of probabilities of receiving treatments (propensity scores) (24) and outcomes for baseline covariates adjustments. We used RF models including all collected predictors to obtain both probabilities of treatment and outcome. The CF estimates treatment effects using the orthogonal construction (25) that separates the treatment effects form the baseline covariates. Then the CF constructs trees that partition the data to minimize variance in treatment effects within each node, rather than maximizing prediction accuracy as in traditional random forests, thereby identifying subgroups with heterogeneous treatment effects.

We obtained the ATEs by pooling the CATEs estimated by CF with the Augmented Inverse Probability Weighted (AIPW) (26), which is considered doubly robust (27), providing unbiased effect estimates if either the propensity score model or the outcome model is correctly specified. However, AIPW relies on the assumption that all confounding variables influencing both treatment assignment and outcomes are observed and appropriately adjusted for (ignorability assumption). The analysis was conducted with the grf R package version 2.4.0.

4.4.2 Estimating precision treatment rules

The PTRs developed in this study introduces a novel approach to recommending antipsychotics by incorporating individual patient preferences and CATEs of seven distinct outcomes as described above. To address this complex decision-making process, we used a scalarization approach that balances multiple objectives by combining them into a single score, referred to as 'utility function'. This function is calculated by assigning weights to each outcome and summing them together. The weighted sum method is a common approach to approximate multi-objective optimization (28).
We categorize the outcomes into effectiveness (hospitalisation, change of medication) and side effects (EPSE, hyperprolactinemia, sedation, sexual side effects, weight gain). We assigned both categories equal relative importance to prevent recommendations from being dominated on a single side effect with little or no regard for effectiveness. Our method allows patients to select up to three side effects that they want to avoid the most, ranking from the most to the least important. We also allow selecting only one or two most important side effects or where all side effects are ranked equally. This procedure results in 86 possible side effect preferences combinations to be selected by patients. We then convert the rank ordering of the side effects to scalar weights using the reciprocal weight method (29). These weights are then used to construct utility functions. Each preferences ranking produces a separate utility function outcome which is then used to develop a precision treatment rule. All rankings and their corresponding weights are presented in eMethods 4.
We then constructed Precision Treatment Rules (PTRs) to recommend the optimal treatment based on patients' baseline characteristics and preferences. PTRs were developed by estimating Conditional Average Treatment Effects (CATE), the expected treatment effect for a patient conditional on the given patient's characteristics using CF (22) with the utility function score used as an outcome. In the last step, for a given patient we select the treatment with the lowest predicted CATEs. This corresponds to the largest decrease in the utility function and to the treatment associated with the lowest weighted risk across the seven negative outcomes.

4.4.3 Estimating performance of the PTRs

We evaluated the performance of the PTRs by comparing estimated outcomes under PTRs-recommended treatments to the outcomes under observed treatment decisions. To do this, we generated 500 bootstrap samples (resampling with replacement) that were used as training sets. For each bootstrap sample, the cases not included in the resample (out of bag observations) were used as the corresponding test set.

In each training sample, we fitted a CF to generate treatment recommendations for each individual in the corresponding test set. While the PTRs were based on CATEs, we evaluated their performance by estimating the ATE of following the recommended treatments versus observed treatment decisions, averaged across all patients in the test set. This was done separately for each of the seven outcomes using outcome-specific CF models fitted on the test data. The 95% confidence intervals of these ATEs were obtained by percentile method based on the bootstrapped estimates. The bootstrapping and evaluation procedure was performed 86 times, once for each utility function corresponding to each specific preference ranking.

4.4.4 Stability of the PTR

We also assessed the stability of the predicted treatment recommendations. For each preference ranking a PTR that used utility function as an outcome was developed using the full data set and compared to the recommendations predicted by the bootstrapped sample PTRs. A *classification instability index* (30) to quantify variability in treatment recommendations across resampled models. R code including all the analysis steps can be shared on request.

4.5 Results

4.5.1 Sample characteristics

1709 patients were included in the sample (eMethods 4.2). Mean age was 26.7 years and 64% were male. Detailed sample characteristics are presented in the eTable 4.1.

The observed proportions of outcomes in the cohort and estimated average treatment effects are presented in table 2. In terms of effectiveness, 63.7% of patients changed their medication within 2 years and 39.9% were hospitalised within 2 years. The most prevalent side effects were sedation (32.9%) and weight gain (19.9%), followed by extrapyramidal side effects (8.2%), hyperprolactinemia (4.5%) and sexual side effects (4.2%).

4.5.2 Estimated average treatment effects of antipsychotics

Estimated ATEs indicated no significant differences between antipsychotics in the risk of hospitalization or medication change (Table 4.1). However, Aripiprazole significantly reduced the prevalence of four side effects hyperprolactinemia, sedation, sexual, and weight gain compared to Risperidone or Olanzapine. Olanzapine reduced the prevalence of EPSE by 12.5 percentage points (pp) (95% CI 9 pp – 16 pp) compared to Aripiprazole and by 14.1 pp (95% CI 10.8 pp - 17.3 pp) compared to Risperidone.

Table 4.1 Estimated Average Treatment Effects in the full cohort

Outcome	Prevalence in the cohort %	Average Treatment Effect Contrast, percentage points (95% CI)		
		Risperidone - Aripiprazole	Olanzapine - Aripiprazole	Olanzapine – Risperidone
Changing a medication within 2 years	63.7	1.4 (-1.7 - 4.5)	-0.2 (-3.1 - 2.7)	-1.6 (-4.3 - 1.1)
Hospitalisation within 2 years	39.9	-4.4 (-10.6 - 1.9)	0.5 (-5.2 - 6.2)	5.1 (-0.2 - 10.3)
Extrapyramidal side effects	8.2	1.5 (-3.1 - 6.2)	-12.5 (-16 - -9)	-14.1 (-17.3 - -10.8)
Hyperprolactinemia	4.5	13.7 (10.6 - 16.8)	1.8 (0.9 - 2.7)	-11.8 (-14.9 - -8.6)
Sedation	32.9	11.3 (5.8 - 16.7)	26.5 (21.5 - 31.5)	15.1 (9.9 - 20.3)
Sexual side effects	4.3	10.9 (8 - 13.8)	1.8 (0.6 - 3)	-9 (-12 - -6)
Weight gain	19.9	6.7 (3.4 - 10.1)	28.7 (25 - 32.4)	22 (17.9 - 26.2)

Table 4.1. The presented estimated average treatment effects contrasts represent a scenario when the first antipsychotic is used as intervention and the second as reference. Positive values indicate that the intervention increases the outcome proportion compared to reference. Negative values indicate that the intervention decreases the outcome proportion compared to reference.

4.5.3 Estimated precision treatment rules

In our cohort, Olanzapine was the most frequently prescribed antipsychotic (49.2%), followed by Risperidone (28.3%) and Aripiprazole (22.5 %) (Table 2). In contrast, recommendations from the PTRs were substantially different from the observed treatment decisions. Averaged across all 86 PTRs preference rankings, Aripiprazole was recommended as a first-line treatment for 93.9% of patients, Olanzapine for 2.4% and Risperidone for 3.5% (Table 4.2).

Table 4.2 Observed prevalence of prescription and proportion of antipsychotic medication recommended by the PTRs averaged over all preferences.

Medication	Observed prevalence the cohort, %	PTRs recommended proportions averaged over all preferences, %	Binominal test adjusted pi value
Aripiprazole	22.5	94.7	$p < 0.001$
Olanzapine	49.2	3.2	$p < 0.001$
Risperidone	28.3	1.9	$p < 0.001$

Table 4.2 Observed prevalence refers to antipsychotic medications prescriptions as observed in the cohort and PTR recommended proportions reflect the distribution of antipsychotics medication recommended by the PTRs over all preferences rankings. Column "PTRs recommended proportions averaged over all preferences" represents aggregate distribution of 86 PTRs.

To explore how treatment recommendations varied by specific preferences, we divided the PTRs into five mutually exclusive groups of 17 rules, each prioritising one side effect preference (eMethods 4.5). For the EPSE-priority group, Aripiprazole was recommended for 79.83% of patients; Olanzapine for 15.98%, and Risperidone for 4.19% (Table 4.3). The remaining four priority groups (hyperprolactinemia, sedation, sexual side effects, and weight gain) recommended Aripiprazole for nearly all patients (96.77% to 98.03%), with Olanzapine recommended for 0% to 1.13% and Risperidone 0.93% to 3.23% (Table 4.3).

Table 4.3 Proportion of antipsychotic medication recommended by the groups of PTRs with the highest priority for a given side effect

Medication	**Estimated antipsychotics distribution if patients received medications recommended by the PTRs prioritising given side effect, %**				
	EPSE	Hyperprolactinemia	Sedation	Sexual side effects	Weight gain
Aripiprazole	81.28	98.54	98.32	98.04	97.49
Olanzapine	14.96	0.85	0	0.82	0
Risperidone	3.76	0.62	1.68	1.14	2.51

Table 4.3 Antipsychotic medications distribution spited by PTRs groups. Columns "EPSE" to "Weight gain" represent a single group of 17 PTRs that rank given side effect the highest.

4.5.4 Estimated performance of the PTRs

Table 4.4 presents the estimated ATEs effects of following PTRs recommendations compared to observed treatment decisions. Averaged across all 86 PTRs, the estimated mean of ATEs was 0.6 pp (95% CI -4 pp - 2.3 pp, $P = 0.71$) for changing the medication and 0.2 pp (95% CI -5 pp. - 5.2 pp, $P = 0.93$) for hospitalisation, indicating no significant benefit or harm under PTRs for the two effectiveness outcomes.

In contrast, following PTR recommendations yielded significant risk reduction for four side effects: hyperprolactinemia -4.7 pp (95% CI -6.2 pp - -2.8 pp, $P < 0.001$), sedation, -15.8 pp (95% CI -21.1 pp - -9.4 pp, $P < 0.001$), sexual side effects -3.9 pp (-5.5 pp - -2.2 pp, $P < 0.001$), and weight gain -15.2 pp (95% CI -19.2 pp - -10.1 pp, $P < 0.001$). However, these improvements were accompanied by a significant estimated risk increase for EPSE 5.5 pp (95% CI 1.4 pp - 97 pp, $P = 0.009$).

Table 4.4 Estimated effect of implementing PTRs as compared to the observed treatment decisions in the cohort average over all preferences.

Outcome	Estimated effect between implementing PRTs recommendations and observed treatment decisions	
	Percentage points (95% CI)	***P* value[a]**
Changing a medication within 2 years	-0.6 (-4 - 2.3)	0.71
Hospitalisation within 2 years	0.2 (-5 - 5.2)	0.93
Extrapyramidal side effects	5.5 (1.4 - 9.7)	0.009
Hyperprolactinemia	-4.7 (-6.2 - -2.8)	<0.001
Sedation	-15.8 (-21.1 - -9.4)	<0.001
Sexual side effects	-3.9 (-5.5 - -2.2)	<0.001
Weight gain	-15.2 (-19.2 - -10.1)	<0.001

Table 4.4. Values represent the nominal change in outcome rates between implementing the PTRs and observed treatment decisions. Positive values represent increase in estimated risk and negative values represent risk reduction. [a] Two-sided nonparametric bootstrap test.

The comparison of estimated PTRs effects across the five priority groups—each avoiding a specific side effect—showed similar results for hospitalization, medication change, hyperprolactinemia and sexual side effects across all groups (Table 4.5). All priority groups showed an increase in the risk of ESPE under PTRs. The EPSE-priority group had the smallest increase at 4.4 pp (95 % CI –1.5 to 9.5 pp) compared to 5.7 pp to 6 pp in the other four groups. However, the estimated risk reduction of sedation and weight gain was lower for the PTRs in the EPSE priority group with -11.8 pp (95% CI -19.7 pp – -0.1 pp) and -11.5 pp (95% CI -17.9 pp - 0.1 pp) respectively compared with larger reductions in the other four groups.

Table 4.5 Estimated effect of implementing PTRs as compared to the observed treatment decisions in the cohort, stratified by prioritized side effect preference group

Outcome	Estimated effect between implementing PRTs recommendations and observed treatment decisions, stratified by the PTRs prioritizing given side effect, percentage points (95% CI)				
	EPSE	Hyperprolactinemia	Sedation	Sexual side effects	Weight gain
Changing a medication within 2 years	-0.6 (-3.6 - 2.3)	-0.7 (-4 - 2.2)	-0.7 (-4 - 2.2)	-0.7 (-4 - 2.2)	-0.7 (-3.9 - 2.2)
Hospitalisation within 2 years	1.1 (-4 - 6)	0 (-5.1 - 5.2)	0 (-5.5 - 5.3)	0.1 (-5.1 - 5.2)	0.1 (-5.2 - 5.3)
Extrapyramidal side effects	4.4 (-1.5 - 9.5)	5.7 (2.1 - 9.7)	5.8 (2.2 - 9.7)	5.7 (2.1 - 9.7)	6 (2.4 - 9.9)
Hyperprolactinemia	-4.3 (-6 - -1.7)	-4.9 (-6.3 - -3.3)	-4.7 (-6.3 - -2.9)	-4.8 (-6.3 - -3.2)	-4.6 (-6.2 - -2.7)
Sedation	-11.8 (-19.7 - -0.1)	-16.7 (-21.5 - -11.4)	-16.7 (-21.4 - -11.8)	-16.6 (-21.4 - -11.3)	-16.7 (-21.4 - -11.7)
Sexual side effects	-3.5 (-5.3 - -1.3)	-4 (-5.6 - -2.6)	-3.9 (-5.5 - -2.3)	-4 (-5.5 - -2.5)	-3.9 (-5.5 - -2.2)
Weight gain	-11.5 (-17.9 - 0.1)	-15.9 (-19.4 - -12.2)	-16.1 (-19.5 - -13)	-15.9 (-19.4 - -12.3)	-16.1 (-19.5 - -13)

Table 4.5. The nominal change in outcome rates of implementing PTRs compered to observed treatment decisions spited by the PTRs prioritized side effect preference group. Columns "EPSE" to "Weight gain" represent a single group of 17 PTRs that rank given side effect the highest. Values are mean percentages with 95% confidence intervals based on 500 bootstrap iterations

4.5.5 Stability of the PTR

The obtained classification instability index indicated that treatment recommendations changed in 5.9% of cases across bootstrap sample (eTable 4.3). The PTRs in the EPSE priority group had the highest classification instability index 19.4%. The other four priority groups had the index between 1.9% and 2.7% (eTable 4.3).

4.6 Discussion

This is the first study that develops PTRs for first episode of psychosis to optimise first-line antipsychotic prescriptions that jointly incorporates effectiveness; side effects; and wide spectrum of patient preferences into treatment recommendations. The incorporation of subjective preferences representing the lived perspective of persons receiving mental health care is currently a mainstream recommendation in clinical psychiatry (31). This is also the first development of PTRs for psychosis that uses fine-grained symptom and substance use predictors extracted by text mining algorithms from real-world clinical notes representing secondary mental health care. Our results show that PTRs averaged across all preferences can significancy reduce hyperprolactinemia by 4.7 pp, sedation by 15.8 pp, sexual side effects by 4.3 pp, and weight gain by 15.2 pp compared to the observed treatment decisions. That would result in the potential elimination of hyperprolactinemia and sexual side effects, a two-fold decrease in the rate of sedation and a four-fold decrease in weight gain. However, under the PTRs estimated risk of EPSE was estimated to increase by 5.5 pp compared to the observed treatment decisions. In the cohort Olanzapine was the most prescribed antipsychotic by clinicians at (49.2%), however under the PTRs Aripiprazole was most recommended at (94.7%). The substantial difference between the recommended and observed treatments can explain the increased risk of EPSE under PTRs and benefit to the other side effects. Olanzapine carries one of the lowest EPSE risk among all antipsychotics (32) and Aripiprazole is best tolerated for all other measured side effects. The PTRs recommendations averaged over all preferences achieved high stability with classification instability index (5.9%).

The results of PTRs stratified by patient's preferences show that when EPSE was selected as top priority to avoid, we predicted larger proportion of Olanzapine recommendations (16%) compared to the average over all preferences (3.2%). Despite that, Aripiprazole was most recommended at (79.8%) for the EPSE priority group. This indicates that for most patients even if high weight is assigned to EPSEs the benefits of Aripiprazole for the remaining four side effects outweigh the increased EPSE risk.

Despite the positive impact of our PTRs in reducing four out of five side effects outcomes, we found no effect for PTRs on the prevalence of hospitalisation or changing medication compared to the observed treatment decisions. This is in line with the meta-analytical evidence that antipsychotics, both of first, second and third generation, differ predominantly by side effects profiles rather than by their intrinsic effectiveness (33).

The presented study presents many methodological innovations. Contrary to the previous work (11), the presented PTRs solve the complex task of multiple criteria decision making by

integrating, effectiveness and side effects outcomes, individual patient characteristics and patient preferences all into a single PTR recommendation. The PTRs allow for easy selection of preferences by rank ordering the top 3 side effects of most concern, minimising cognitive load and facilitating patient's involvement in the shared decision making with clinicians. The single recommendation integrating all domains can be presented along the estimated individualised effects for each of the seven included outcomes under all antipsychotics options to avoid 'computer paternalism' (34) and provide full information for the final decision by patients.

Limitations of the study include that the conditional treatment effects underpinning the PTRs are estimated using the nonexperimental data. However, the real-world data offers better representatives of clinical routine practice than randomised control trials (35). We also used best practice causal machine learning methods to adjust for non-random assignment. There may be an effect of confounding by indication as the choice of first-line antipsychotic may vary according to perceived clinical severity. However, we mitigated against this by adjusting for a range of confounders including demographics, clinical variables as well as symptoms and substance use. In every observational study the effects of unobserved confounders cannot be ruled out. As recommended (10), the development and independent validation the of PTRs with the large EHRs should be followed by a pragmatic trial to experimentally evaluate the performance of PTRs recommendations. Secondly, the side effects data relies on what clinicians decided to document in the notes during routine care. Information is limited by accurate sharing and recording in the clinical setting. Thirdly, due to sample size requirements only data on three antipsychotic medications were available for the PTRs development. These three medications represent over 74% of prescriptions of first-line antipsychotics during the study period, however further research with more treatment options when a sufficient sample size will be available is suggested.

Further research can involve reproduction of the PTR in different clinical setting for more rigorous assessment of preliminary results before the targeted trial evaluation. The model can be extended to provide recommendations for multiple rounds of treatments. The recommendations for the second-line antipsychotics can be provided by refining the PTRs with dynamic treatment rules methods (36). The presented approach of incorporating multiple outcomes and patient preferences into PTRs can be used to optimize treatment selection in different fields of medicine.

4.7 Conclusions

We present the first development of PTRs for first episode of psychosis incorporating effectiveness, side effects, and patient preferences. The presented treatment rules can significantly reduce hyperprolactinemia, sedation, sexual side effects, and weight gain, without any reduction in effectiveness, but increase extrapyramidal side effects. If implemented the PTRs can support personalized medicine and shared decision making.

ACKNOWLEDGMENTS AND CONFLICT OF INTEREST

MA (Grant No. MR/ N013700/1) and YL Council (MR/W006820/1) are supported by the UK Medical Research Council and King's College London member of the Medical Research Council Doctoral Training Partnership in Biomedical Sciences. DS and PF-P were partly funded by the National Institute for Health and Care Research (NIHCR) Maudsley Biomedical Research Centre and King's College London. The views expressed are those of the author(s) and not necessarily those of the (NIHCR) or the Department of Health and Social Care. MA has been employed by F. Hoffmann-La Roche AG outside of the current study. PF-P has received research fees from Lundbeck and honoraria from Lundbeck, Angelini, Menarini, and Boehringer Ingelheim outside the current study. All other authors report no biomedical financial interests or potential conflicts of interest.

4.8 References

1. Taylor, David M, Thomas RE Barnes,, Allan H. Young (2021): *The Maudsley Prescribing Guidelines in Psychiatry 14th Edition*. Chichester UK: John Wiley & Sons.

2. Szmulewicz AG, Martínez-Alés G, Logan R, Ferrara M, Kelly C, Fredrikson D, *et al.* (2024): Antipsychotic drugs in first-episode psychosis: a target trial emulation in the FEP-CAUSAL Collaboration. *Am J Epidemiol* 193: 1081–1087.

3. Stovell D, Morrison AP, Panayiotou M, Hutton P (2016): Shared treatment decision-making and empowerment-related outcomes in psychosis: systematic review and meta-analysis. *Br J Psychiatry J Ment Sci* 209: 23–28.

4. Leucht S, Schneider-Thoma J, Burschinski A, Peter N, Wang D, Dong S, *et al.* (2023): Long-term efficacy of antipsychotic drugs in initially acutely ill adults with schizophrenia: systematic review and network meta-analysis. *World Psychiatry* 22: 315–324.

5. Leucht S, Cipriani A, Spineli L, Mavridis D, Örey D, Richter F, *et al.* (2013): Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *The Lancet* 382: 951–962.

6. Guinart D, Fagiolini A, Fusar-Poli P, Giordano GM, Leucht S, Moreno C, Correll CU (2024): On the Road to Individualizing Pharmacotherapy for Adolescents and Adults with Schizophrenia – Results from an Expert Consensus Following the Delphi Method. *Neuropsychiatr Dis Treat*.

7. Müller K, Schuster F, Rodolico A, Siafis S, Leucht S, Hamann J (2023): How should patient decision aids for schizophrenia treatment be designed? - A scoping review. *Schizophr Res* 255: 261–273.

8. Kosorok MR, Laber EB (2019): Precision Medicine. *Annu Rev Stat Its Appl* 6: 263–286.

9. Tsiatis AA, Davidian M, Holloway ST (2021): *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC Press Boca Raton USA.

10. Kessler RC, Luedtke A (2021): Pragmatic Precision Psychiatry-A New Direction for Optimizing Treatment Selection. *JAMA Psychiatry* 78: 1384–1390.

11. Wu C-S, Luedtke AR, Sadikova E, Tsai H-J, Liao S-C, Liu C-C, *et al.* (2020): Development and Validation of a Machine Learning Individualized Treatment Rule in First-Episode Schizophrenia. *JAMA Netw Open* 3: e1921660.

12. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, *et al.* (2009): The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* 9: 51.

13. Fusar-Poli P, Lai S, Di Forti M, Iacoponi E, Thornicroft G, McGuire P, Jauhar S (2020): Early Intervention Services for First Episode of Psychosis in South London and the Maudsley (SLaM): 20 Years of Care and Research for Young People. *Front Psychiatry* 11: 577110.

14. Luedtke A, Sadikova E, Kessler RC (2019): Sample Size Requirements for Multivariate Models to Predict Between-Patient Differences in Best Treatments of Major Depressive Disorder. *Clin Psychol Sci* 7: 445–461.

15. Patel R, Brinn A, Irving J, Chaturvedi J, Gudiseva S, Correll CU, *et al.* (2023): Oral and long-acting injectable antipsychotic discontinuation and relationship to side effects in people with first episode psychosis: a longitudinal analysis of electronic health record data. *Ther Adv Psychopharmacol* 13: 20451253231211575.

16. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, *et al.* (2024): TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385: e078378.

17. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, *et al.* (2015): The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med* 12: e1001885.

18. Firth J, Siddiqi N, Koyanagi A, Siskind D, Rosenbaum S, Galletly C, *et al.* (2019): The Lancet Psychiatry Commission: a blueprint for protecting physical health in people with mental illness. *Lancet Psychiatry* 6: 675–712.

19. Irving J, Patel R, Oliver D, Colling C, Pritchard M, Broadbent M, *et al.* (2021): Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Schizophr Bull* 47: 405–414.

20. Krakowski K, Oliver D, Arribas M, Stahl D, Fusar-Poli P (2024): Dynamic and Transdiagnostic Risk Calculator Based on Natural Language Processing for the Prediction of Psychosis in Secondary Mental Health Care: Development and Internal-External Validation Cohort Study. *Biol Psychiatry* 96: 604–614.

21. Gudio W. Imbens, Donald B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*. Harvard University, Massachusetts.

22. Athey S, Tibshirani J, Wager S (2019): Generalized random forests. *Ann Stat* 47: 1148–1178.

23. Breiman L (2001): Random Forests. *Mach Learn* 45: 5–32.

24. Desai RJ, Franklin JM (2019): Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ* 367: l5657.

25. Robinson PM (1988): Root-N-Consistent Semiparametric Regression. *Econometrica* 56: 931–954.

26. Glynn AN, Quinn KM (2010): An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Polit Anal* 18: 36–56.

27. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M (2011): Doubly Robust Estimation of Causal Effects. *Am J Epidemiol* 173: 761–767.

28. Marler RT, Arora JS (2010): The weighted sum method for multi-objective optimization: new insights. *Struct Multidiscip Optim* 41: 853–862.

29. Roszkowska E (2013): Rank Ordering Criteria Weighting Methods – a Comparative Overview. *Optim Stud Ekon* 14–33.

30. Riley RD, Collins GS (2023): Stability of clinical prediction models developed using statistical or machine learning methods. *Biom J* 65: 2200302.

31. Fusar-Poli P, Estradé A, Stanghellini G, Venables J, Onwumere J, Messas G, *et al.* (2022): The lived experience of psychosis: a bottom-up review co-written by experts by experience and academics. *World Psychiatry Off J World Psychiatr Assoc WPA* 21: 168–188.

32. Siafis S, Wu H, Wang D, Burschinski A, Nomura N, Takeuchi H, *et al.* (2023): Antipsychotic dose, dopamine D2 receptor occupancy and extrapyramidal side-effects: a systematic review and dose-response meta-analysis. *Mol Psychiatry* 28: 3267–3277.

33. Zhu Y, Krause M, Huhn M, Rothe P, Schneider-Thoma J, Chaimani A, *et al.* (2017): Antipsychotic drugs for the acute treatment of patients with a first episode of schizophrenia: a systematic review with pairwise and network meta-analyses. *Lancet Psychiatry* 4: 694–705.

34. Leucht S, Siafis S, Rodolico A, Peter NL, Müller K, Waibel J, *et al.* (2023): Shared Decision Making Assistant (SDMA) and other digital tools for choosing antipsychotics in schizophrenia treatment. *Eur Arch Psychiatry Clin Neurosci* 273: 1629–1631.

35. Taipale H, Schneider-Thoma J, Pinzón-Espinosa J, Radua J, Efthimiou O, Vinkers CH, *et al.* (2022): Representation and Outcomes of Individuals With Schizophrenia Seen in Everyday Practice Who Are Ineligible for Randomized Clinical Trials. *JAMA Psychiatry* 79: 1–9.

36. Mahar RK, McGuinness MB, Chakraborty B, Carlin JB, IJzerman MJ, Simpson JA (2021): A scoping review of studies using observational data to optimise dynamic treatment regimens. *BMC Med Res Methodol* 21: 39.

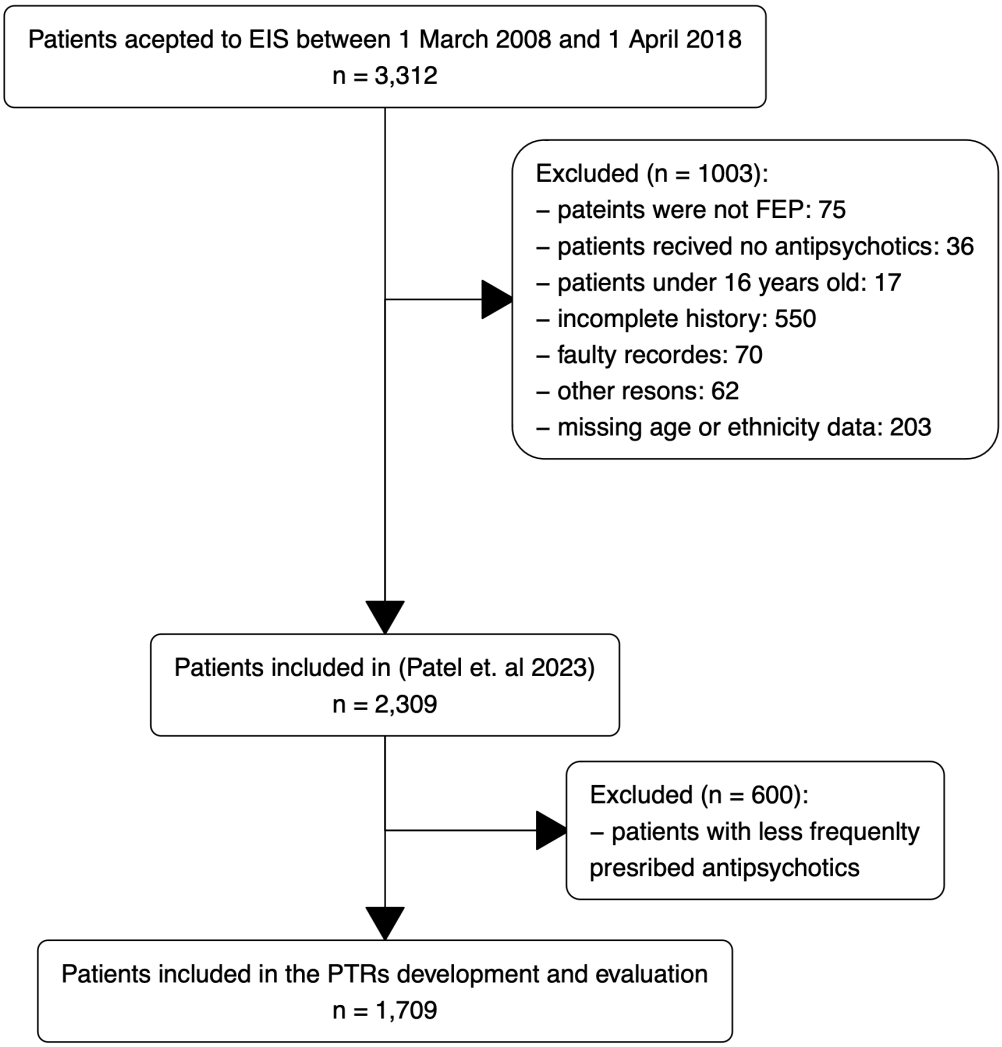
4.9 SUPPLEMENTARY MATERIAL

eMethods 4.1 SLaM Service Characteristics

With respect to service characteristics, SLaM early intervention teams serve an overall catchment area of 443,050 people aged 16–35 years (2017)[31], and are amongst the largest of their kind in the UK and worldwide. The incidence of psychosis in SLaM (from 58.3 to 71.9 cases per 100,000 person-years[31]) is greater than the national average, and it is one of the highest worldwide[32]. Moreover, the incidence of BMD in South London boroughs has also been shown to be over twice that in other areas of the UK such as Bristol and Nottingham[33]. This may be accounted for by the accumulation of several risk factors for PSY and BMD such as immigration, higher numbers of people from ethnic minorities associated with psychosis (e.g., African/Afro-Caribbean, prevalence around 50%), low socioeconomic status, high co-morbidity rates with other mental disorders, high urbanicity and illicit substance use. [28,34–37]

The trust is digitized and paper-free[38], with each patient having a personal EHR since 2007[39,40]. SLaM healthcare professionals are legally required to update these records[38]. The SLaM register contains all these clinical records which are constantly updated throughout the patient's care, irrespective of any discharges from and/or referrals to other services.

eMethods 4.2 Flow chart of patients included in the study



This study uses patients' sample extracted by (Patel et. al 2023). Due to the sample size requirement for the Precision Treatment Rules development, we restricted the sample to patients prescribed Aripiprazole, Olanzapine or Risperidone as their first antipsychotic.

(Patel et. al 2023) Patel R, Brinn A, Irving J, Chaturvedi J, Gudiseva S, Correll CU, *et al.* (2023): Oral and long-acting injectable antipsychotic discontinuation and relationship to side effects in people with first episode psychosis: a longitudinal analysis of electronic health record data. *Ther Adv Psychopharmacol* 13:

eTable 4.1 Sample characteristics stratified by first antipsychotic prescription

	No. (proportion [%])		
Antipsychotic	**Aripiprazole**	**Olanzapine**	**Risperidone**
Sample size			
Number of patients	348	841	484
Predictors ***Sociodemographic***			
Gender			
Female	126 (32.8)	308 (36.6)	144 (29.8)
Male	258 (67.2)	533 (63.4)	340 (70.2)
Age, mean (SD)	28.5 (9.94)	26.88 (7.87)	26.22 (8.4)
Self-assigned ethnicity			
Asian	29 (7.6)	66 (7.8)	53 (11)
Black	200 (52.1)	420 (49.9)	223 (46.1)
Mixed	19 (4.9)	35 (4.2)	22 (4.5)
White	27 (7)	57 (6.8)	45 (9.3)
Other	101 (26.3)	254 (30.2)	140 (28.9)
missing	8 (2.1)	9 (1.1)	1 (0.2)
Clinical			
ICD-10 psychosis type			
Acute and transient psychotic disorders	84 (21.9)	209 (24.9)	120 (24.8)
Delusional disorder	13 (3.4)	22 (2.6)	12 (2.5)
Mania with psychotic symptoms	6 (1.6)	34 (4)	5 (1)
Bipolar affective disorders with psychotic symptoms	4 (1)	24 (2.9)	4 (0.8)
Depressive disorders with psychotic symptoms	14 (3.6)	39 (4.6)	27 (5.6)
Puerperal psychosis	0 (0)	4 (0.5)	1 (0.2)
Schizoaffective disorder	2 (0.5)	18 (2.1)	14 (2.9)
Schizophrenia	48 (12.5)	101 (12)	67 (13.8)
Substance use with psychosis	7 (1.8)	57 (6.8)	27 (5.6)
Unspecified nonorganic psychosis	206 (53.6)	333 (39.6)	207 (42.8)
Natural Language Processing (NLP)			
Mean (SD)			
Aggression	1.04 (1.11)	1.05 (1.08)	1.02 (1.22)
Agitation	1.09 (1.13)	1.21 (1.1)	1.07 (1.11)
Anergia	0.03 (0.22)	0.01 (0.08)	0.02 (0.23)
Anhedonia	0.14 (0.44)	0.11 (0.36)	0.15 (0.47)
Anxiety	2.14 (1.91)	1.94 (1.32)	2.07 (1.43)
Apathy	0.1 (0.35)	0.07 (0.31)	0.09 (0.35)
Appetite	0.23 (0.48)	0.24 (0.52)	0.21 (0.48)
Arousal	0.33 (0.6)	0.45 (0.73)	0.3 (0.69)

Auditory hallucination	1.07 (1.03)	0.98 (0.98)	1.06 (1.09)
Bad dreams	0.04 (0.24)	0.04 (0.26)	0.04 (0.22)
Blunted flat affect	0.29 (0.56)	0.3 (0.66)	0.38 (0.75)
Circumstantial speech	0.19 (0.49)	0.16 (0.45)	0.16 (0.53)
Cognitive impairment	2.11 (1.42)	1.94 (1.33)	2.14 (1.5)
Concrete thinking	0.02 (0.14)	0.04 (0.24)	0.02 (0.16)
Delusion	1.2 (1.11)	1.14 (1.06)	1.13 (1.13)
Derailment of speech	0.09 (0.35)	0.1 (0.35)	0.08 (0.35)
Disturbed sleep	1.29 (1.04)	1.37 (1.02)	1.36 (1.08)
Diurnal	0.01 (0.07)	0.01 (0.1)	0.01 (0.1)
Echolalia	0.02 (0.14)	0.03 (0.24)	0.03 (0.17)
Elation	0.28 (0.63)	0.45 (0.79)	0.33 (0.73)
Emotional withdrawn	0.55 (0.89)	0.5 (0.77)	0.58 (0.84)
Flight of ideas	0.17 (0.47)	0.24 (0.58)	0.18 (0.56)
Formal thought disorder	0.13 (0.4)	0.11 (0.4)	0.09 (0.31)
Grandiosity	0.34 (0.68)	0.42 (0.8)	0.35 (0.76)
Guilt	0.25 (0.6)	0.33 (0.61)	0.29 (0.65)
Hallucination	1.3 (1.1)	1.21 (1.05)	1.3 (1.17)
Helpless	0.06 (0.27)	0.07 (0.27)	0.04 (0.19)
Hopeless	0.37 (0.74)	0.3 (0.63)	0.28 (0.66)
Hostility	0.4 (0.69)	0.53 (0.83)	0.42 (0.82)
Insight	1.07 (0.94)	1.13 (1.01)	1 (0.98)
Insomnia	0.24 (0.52)	0.19 (0.51)	0.19 (0.45)
Irritability	0.74 (0.95)	0.89 (1.03)	0.79 (1.02)
Loneliness	0.14 (0.43)	0.13 (0.39)	0.15 (0.4)
Loss of coherence	0.31 (0.68)	0.31 (0.65)	0.26 (0.61)
Low energy	0.15 (0.43)	0.12 (0.38)	0.14 (0.42)
Mood instability	0.56 (0.78)	0.68 (0.87)	0.62 (0.95)
Mutism	0.22 (0.62)	0.26 (0.69)	0.22 (0.6)
Negative symptoms	0.21 (0.64)	0.14 (0.45)	0.16 (0.5)
Nightmares	0.12 (0.39)	0.13 (0.43)	0.13 (0.41)
Paranoia	1.69 (1.21)	1.64 (1.24)	1.68 (1.3)
Passivity	0.21 (0.49)	0.16 (0.48)	0.15 (0.43)
Persecutory	0.82 (0.92)	0.74 (0.91)	0.78 (1.04)
Poor concentration	0.66 (0.79)	0.69 (0.81)	0.73 (0.88)
Poor motivation	0.2 (0.47)	0.2 (0.5)	0.25 (0.53)
Poverty of speech	0.12 (0.4)	0.11 (0.39)	0.11 (0.37)
Poverty of thought	0.05 (0.24)	0.05 (0.29)	0.06 (0.3)
Social withdrawal	0.23 (0.56)	0.19 (0.53)	0.2 (0.55)
Stupor	0.01 (0.09)	0.01 (0.11)	0.01 (0.09)
Suicidal thoughts	0.35 (0.63)	0.38 (0.68)	0.43 (0.82)
Tangential speech	0.46 (0.77)	0.47 (0.76)	0.39 (0.79)
Tearful	0.61 (0.87)	0.73 (0.95)	0.58 (0.9)
Thought block	0.26 (0.53)	0.3 (0.61)	0.24 (0.53)
Thought broadcast	0.24 (0.52)	0.19 (0.54)	0.17 (0.51)
Thought insert	0.22 (0.52)	0.19 (0.52)	0.15 (0.43)

Thought withdrawal	0.11 (0.44)	0.07 (0.3)	0.06 (0.27)
Visual hallucination	0.4 (0.69)	0.35 (0.66)	0.38 (0.69)
Waxy flexibility	0.01 (0.09)	0.01 (0.13)	0.01 (0.08)
Weightless	0.27 (0.57)	0.38 (0.73)	0.33 (0.66)
Worthless	0.06 (0.31)	0.07 (0.3)	0.05 (0.22)
Cannabis	1.09 (1.23)	1.13 (1.28)	1.18 (1.33)
Cocaine	0.29 (0.71)	0.36 (0.75)	0.32 (0.73)
Mdma	0.07 (0.31)	0.08 (0.38)	0.05 (0.34)
Lives alone	0.19 (0.52)	0.17 (0.49)	0.16 (0.5)
Smoking	0.78 (0.95)	0.84 (1.08)	0.85 (1.12)

eTable 4.2 Type and precision for 66 NLP algorithms

Precision values are taken from the CRIS Natural Language Processing Library (2021) [90], and were obtained by randomly selecting *n* positive annotations from each algorithm for a specified cohort, limited to one annotation per patient ID. Precision was then calculated as the ratio of the number of relevant (true positive) instances retrieved out of the total NLP-labelled positive instances (including irrelevant [false positive] and relevant [true positive] instances) for each
NLP algorithm.

NLP algorithms	Cohort	Annotations validated (n)	Precision (%)
Aggression	Random sample	100	90
Agitation	Random sample	100	85
Anergia	Random sample	100	84
Anhedonia	Random sample	100	94
Anxiety	Random sample	100	94
Apathy	Random sample	100	94
Arousal	Random sample	100	89
Bad dreams	CAMHS events	100	92
Blunted affect	Random sample	100	98
Cannabis use	All patients	100	88
Circumstantiality	Random sample	100	97
Cocaine use	Random sample	30	97
Cognitive impairment	Patients with F20	100	84
Concrete thinking	Random sample	146	91
Delusional thinking	Random sample	100	90
Derailment	Random sample	100	87
Disturbed sleep	Random sample	100	89
Diurnal mood	Random sample	100	86
Early morning wakening	Random sample	100	96
Echolalia	Random sample	100	96
Elation	Random sample	100	95
Emotional withdrawal	Random sample	100	87
Feeling helpless	Random sample	100	92
Feeling hopeless	Random sample	100	88
Feeling lonely	Random sample	100	87
Feeling worthless	Random sample	100	91
Flight of ideas	Random sample	100	89
Formal thought disorder	Random sample	100	85
Grandiosity	Random sample	100	89
Guilt	Random sample	100	84
Hallucinations (all)	Random sample	100	90
Hallucinations (auditory)	Random sample	100	92
Hallucinations (OTG: olfactory, tactile, gustatory)	Random sample	100	86
Hallucinations (visual)	Random sample	100	83
Hostility	Random sample	100	86

Insomnia	Random sample	100	97
Irritability	Random sample	100	99
Lives Alone	Random sample	50	100
Loss of coherence	Random sample	158	85
Low energy	CAMHS events	100	89
MDMA use	Random sample	100	94
Mood instability	Random sample	100	91
Mutism	Random sample	100	95
Negative symptoms	Random sample	100	87
Nightmares	Random sample	100	89
Paranoia	Random sample	100	89
Passivity	Random sample	100	88
Persecutory ideation	Random sample	100	80
Poor appetite	Random sample	100	89
Poor concentration	Random sample	100	88
Poor insight	Random sample	100	85
Poor motivation	Random sample	100	95
Poverty of speech	Random sample	100	88
Poverty of thought	Random sample	100	98
Social withdrawal	Random sample	100	98
Stupor	Random sample	100	88
Suicidality	CAMHS events	100	87
Tangential speech	Random sample	100	90
Tearfulness	Random sample	100	94
Thought block	Random sample	100	92
Thought broadcast	Random sample	100	84
Thought insertion	Random sample	100	84
Thought withdrawal	Random sample	100	84
Tobacco use	Random sample	118	90
Waxy flexibility	Random sample	100	81
Weight loss	Random sample	100	80

eMethods 4.3 NLP algorithm development and validation

The CRIS symptom algorithms (e.g. 'guilt') have been developed using machine learning approaches against gold standard training sets manually annotated for positive, negative and unknown (irrelevant) mentions. As such, they are able to exclude language features such as negation (e.g. 'patient denies guilt', 'patient has no guilt') and irrelevant mentions (e.g. 'his mother felt guilty'). Patterns of failure driving false positives are identified through manual testing of algorithm output (e.g. 'ZZZZZ was found guilty of stealing'); the machine learning classifier is then trained on these false positives to ignore these and similar statements in an iterative process of testing and redeveloping until acceptable precision is achieved. Patterns of failure identified through testing can be found in the CRIS service's comprehensive online NLP algorithm library provided at https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/cris-natural-language-processing/.

The performance of each NLP algorithm was measured with precision (proportion of true positive instances of total NLP-labelled positive instances) and recall (proportion of true positive instances of all positive instances in the text). As EHRs provide multiple opportunities for term detection, we favour precision over recall, using only NLP algorithms with at least 80% precision (see eTable xx for a final list of NLP algorithms employed).

These algorithms were manually validated by an independent researcher at the SLaM Biomedical Research Centre Nucleus prior to the current research project. The programme for algorithms validation was responsive to the specific needs of scheduled CRIS research activities and therefore the approach was not standardised. For example, depression symptom algorithms have been validated against records for SLaM individuals who had ever had a depression diagnosis; other algorithms have been validated against records for all individuals on the SLaM register.

Antipsychotics Minimum Effective Doses

Extracted from the Maudsley Prescribing Guidelines Version 14.

Antipsychotics

Drug	Minimum effective dose (mg/day)
Aripiprazole	10
Olanzapine	5
Risperidone	2

eMethods 4.4 Preference elicitation rankings and weights Antipsychotics

To facilitate preference elicitation, we allow users to select a ranking of side effects according to the users' concern. We allow them to select top three side-effects that they want to avoid the most in order form the most important to the least. We also allow elicitation when only single most important side-effect is selected, two most important ones or all side effects are ranked equally. That creates 86 individual rankings that are then converted to scalar weights by the reciprocal weight method. All individual rankings and the corresponding weights are presented in the table below. The details on weights derivation are described below the table.

All preference rankings and corresponding weights

Effectiveness outcomes (weights used in the weighted sum function)		Ranking positions of side-effects preferences (weights used in the weighted sum function)				
Change of medication	Hospitalisation	EPSE	Hyperprolactinemia	Sedation	Sexual side effects	Weight gain
(0.25)	(0.25)	1st (0.1)	1st (0.1)	1st (0.1)	1st (0.1)	1st (0.1)
(0.25)	(0.25)	1st (0.22)	2nd (0.07)	2nd (0.07)	2nd (0.07)	2nd (0.07)
(0.25)	(0.25)	2nd (0.07)	1st (0.22)	2nd (0.07)	2nd (0.07)	2nd (0.07)
(0.25)	(0.25)	2nd (0.07)	2nd (0.07)	1st (0.22)	2nd (0.07)	2nd (0.07)
(0.25)	(0.25)	2nd (0.07)	2nd (0.07)	2nd (0.07)	1st (0.22)	2nd (0.07)
(0.25)	(0.25)	2nd (0.07)	2nd (0.07)	2nd (0.07)	2nd (0.07)	1st (0.22)
(0.25)	(0.25)	1st (0.22)	2nd (0.11)	3rd (0.056)	3rd (0.056)	3rd (0.056)
(0.25)	(0.25)	1st (0.22)	3rd (0.056)	2nd (0.11)	3rd (0.056)	3rd (0.056)
(0.25)	(0.25)	1st (0.22)	3rd (0.056)	3rd (0.056)	2nd (0.11)	3rd (0.056)
(0.25)	(0.25)	1st (0.22)	3rd (0.056)	3rd (0.056)	3rd (0.056)	2nd (0.11)
(0.25)	(0.25)	2nd (0.11)	1st (0.22)	3rd (0.056)	3rd (0.056)	3rd (0.056)
(0.25)	(0.25)	2nd (0.11)	3rd (0.056)	1st (0.22)	3rd (0.056)	3rd (0.056)
(0.25)	(0.25)	2nd (0.11)	3rd (0.056)	3rd (0.056)	1st (0.22)	3rd (0.056)
(0.25)	(0.25)	2nd (0.11)	3rd (0.056)	3rd (0.056)	3rd (0.056)	1st (0.22)
(0.25)	(0.25)	3rd (0.056)	1st (0.22)	2nd (0.11)	3rd (0.056)	3rd (0.056)
(0.25)	(0.25)	3rd (0.056)	1st (0.22)	3rd (0.056)	2nd (0.11)	3rd (0.056)
(0.25)	(0.25)	3rd (0.056)	1st (0.22)	3rd (0.056)	3rd (0.056)	2nd (0.11)
(0.25)	(0.25)	3rd (0.056)	2nd (0.11)	1st (0.22)	3rd (0.056)	3rd (0.056)
(0.25)	(0.25)	3rd (0.056)	2nd (0.11)	3rd (0.056)	1st (0.22)	3rd (0.056)
(0.25)	(0.25)	3rd (0.056)	2nd (0.11)	3rd (0.056)	3rd (0.056)	1st (0.22)
(0.25)	(0.25)	3rd (0.056)	3rd (0.056)	1st (0.22)	2nd (0.11)	3rd (0.056)
(0.25)	(0.25)	3rd (0.056)	3rd (0.056)	1st (0.22)	3rd (0.056)	2nd (0.11)
(0.25)	(0.25)	3rd (0.056)	3rd (0.056)	2nd (0.11)	1st (0.22)	3rd (0.056)
(0.25)	(0.25)	3rd (0.056)	3rd (0.056)	2nd (0.11)	3rd (0.056)	1st (0.22)
(0.25)	(0.25)	3rd (0.056)	3rd (0.056)	3rd (0.056)	1st (0.22)	2nd (0.11)
(0.25)	(0.25)	3rd (0.056)	3rd (0.056)	3rd (0.056)	2nd (0.11)	1st (0.22)
(0.25)	(0.25)	1st (0.22)	2nd (0.11)	3rd (0.07)	4th (0.05)	4th (0.05)
(0.25)	(0.25)	1st (0.22)	2nd (0.11)	4th (0.05)	3rd (0.07)	4th (0.05)
(0.25)	(0.25)	1st (0.22)	2nd (0.11)	4th (0.05)	4th (0.05)	3rd (0.07)
(0.25)	(0.25)	1st (0.22)	3rd (0.07)	2nd (0.11)	4th (0.05)	4th (0.05)
(0.25)	(0.25)	1st (0.22)	3rd (0.07)	4th (0.05)	2nd (0.11)	4th (0.05)
(0.25)	(0.25)	1st (0.22)	3rd (0.07)	4th (0.05)	4th (0.05)	2nd (0.11)
(0.25)	(0.25)	1st (0.22)	4th (0.05)	2nd (0.11)	3rd (0.07)	4th (0.05)
(0.25)	(0.25)	1st (0.22)	4th (0.05)	2nd (0.11)	4th (0.05)	3rd (0.07)
(0.25)	(0.25)	1st (0.22)	4th (0.05)	3rd (0.07)	2nd (0.11)	4th (0.05)
(0.25)	(0.25)	1st (0.22)	4th (0.05)	3rd (0.07)	4th (0.05)	2nd (0.11)

(0.25)	(0.25)	1st (0.22)	4th (0.05)	4th (0.05)	2nd (0.11)	3rd (0.07)
(0.25)	(0.25)	1st (0.22)	4th (0.05)	4th (0.05)	3rd (0.07)	2nd (0.11)
(0.25)	(0.25)	2nd (0.11)	1st (0.22)	3rd (0.07)	4th (0.05)	4th (0.05)
(0.25)	(0.25)	2nd (0.11)	1st (0.22)	4th (0.05)	3rd (0.07)	4th (0.05)
(0.25)	(0.25)	2nd (0.11)	1st (0.22)	4th (0.05)	4th (0.05)	3rd (0.07)
(0.25)	(0.25)	2nd (0.11)	3rd (0.07)	1st (0.22)	4th (0.05)	4th (0.05)
(0.25)	(0.25)	2nd (0.11)	3rd (0.07)	4th (0.05)	1st (0.22)	4th (0.05)
(0.25)	(0.25)	2nd (0.11)	3rd (0.07)	4th (0.05)	4th (0.05)	1st (0.22)
(0.25)	(0.25)	2nd (0.11)	4th (0.05)	1st (0.22)	3rd (0.07)	4th (0.05)
(0.25)	(0.25)	2nd (0.11)	4th (0.05)	1st (0.22)	4th (0.05)	3rd (0.07)
(0.25)	(0.25)	2nd (0.11)	4th (0.05)	3rd (0.07)	1st (0.22)	4th (0.05)
(0.25)	(0.25)	2nd (0.11)	4th (0.05)	3rd (0.07)	4th (0.05)	1st (0.22)
(0.25)	(0.25)	2nd (0.11)	4th (0.05)	4th (0.05)	1st (0.22)	3rd (0.07)
(0.25)	(0.25)	2nd (0.11)	4th (0.05)	4th (0.05)	3rd (0.07)	1st (0.22)
(0.25)	(0.25)	3rd (0.07)	1st (0.22)	2nd (0.11)	4th (0.05)	4th (0.05)
(0.25)	(0.25)	3rd (0.07)	1st (0.22)	4th (0.05)	2nd (0.11)	4th (0.05)
(0.25)	(0.25)	3rd (0.07)	1st (0.22)	4th (0.05)	4th (0.05)	2nd (0.11)
(0.25)	(0.25)	3rd (0.07)	2nd (0.11)	1st (0.22)	4th (0.05)	4th (0.05)
(0.25)	(0.25)	3rd (0.07)	2nd (0.11)	4th (0.05)	1st (0.22)	4th (0.05)
(0.25)	(0.25)	3rd (0.07)	2nd (0.11)	4th (0.05)	4th (0.05)	1st (0.22)
(0.25)	(0.25)	3rd (0.07)	4th (0.05)	1st (0.22)	2nd (0.11)	4th (0.05)
(0.25)	(0.25)	3rd (0.07)	4th (0.05)	1st (0.22)	4th (0.05)	2nd (0.11)
(0.25)	(0.25)	3rd (0.07)	4th (0.05)	2nd (0.11)	1st (0.22)	4th (0.05)
(0.25)	(0.25)	3rd (0.07)	4th (0.05)	2nd (0.11)	4th (0.05)	1st (0.22)
(0.25)	(0.25)	3rd (0.07)	4th (0.05)	4th (0.05)	1st (0.22)	2nd (0.11)
(0.25)	(0.25)	3rd (0.07)	4th (0.05)	4th (0.05)	2nd (0.11)	1st (0.22)
(0.25)	(0.25)	4th (0.05)	1st (0.22)	2nd (0.11)	3rd (0.07)	4th (0.05)
(0.25)	(0.25)	4th (0.05)	1st (0.22)	2nd (0.11)	4th (0.05)	3rd (0.07)
(0.25)	(0.25)	4th (0.05)	1st (0.22)	3rd (0.07)	2nd (0.11)	4th (0.05)
(0.25)	(0.25)	4th (0.05)	1st (0.22)	3rd (0.07)	4th (0.05)	2nd (0.11)
(0.25)	(0.25)	4th (0.05)	1st (0.22)	4th (0.05)	2nd (0.11)	3rd (0.07)
(0.25)	(0.25)	4th (0.05)	1st (0.22)	4th (0.05)	3rd (0.07)	2nd (0.11)
(0.25)	(0.25)	4th (0.05)	2nd (0.11)	1st (0.22)	3rd (0.07)	4th (0.05)
(0.25)	(0.25)	4th (0.05)	2nd (0.11)	1st (0.22)	4th (0.05)	3rd (0.07)
(0.25)	(0.25)	4th (0.05)	2nd (0.11)	3rd (0.07)	1st (0.22)	4th (0.05)
(0.25)	(0.25)	4th (0.05)	2nd (0.11)	3rd (0.07)	4th (0.05)	1st (0.22)
(0.25)	(0.25)	4th (0.05)	2nd (0.11)	4th (0.05)	1st (0.22)	3rd (0.07)
(0.25)	(0.25)	4th (0.05)	2nd (0.11)	4th (0.05)	3rd (0.07)	1st (0.22)
(0.25)	(0.25)	4th (0.05)	3rd (0.07)	1st (0.22)	2nd (0.11)	4th (0.05)
(0.25)	(0.25)	4th (0.05)	3rd (0.07)	1st (0.22)	4th (0.05)	2nd (0.11)
(0.25)	(0.25)	4th (0.05)	3rd (0.07)	2nd (0.11)	1st (0.22)	4th (0.05)
(0.25)	(0.25)	4th (0.05)	3rd (0.07)	2nd (0.11)	4th (0.05)	1st (0.22)
(0.25)	(0.25)	4th (0.05)	3rd (0.07)	4th (0.05)	1st (0.22)	2nd (0.11)
(0.25)	(0.25)	4th (0.05)	3rd (0.07)	4th (0.05)	2nd (0.11)	1st (0.22)
(0.25)	(0.25)	4th (0.05)	4th (0.05)	1st (0.22)	2nd (0.11)	3rd (0.07)
(0.25)	(0.25)	4th (0.05)	4th (0.05)	1st (0.22)	3rd (0.07)	2nd (0.11)
(0.25)	(0.25)	4th (0.05)	4th (0.05)	2nd (0.11)	1st (0.22)	3rd (0.07)
(0.25)	(0.25)	4th (0.05)	4th (0.05)	2nd (0.11)	3rd (0.07)	1st (0.22)
(0.25)	(0.25)	4th (0.05)	4th (0.05)	3rd (0.07)	1st (0.22)	2nd (0.11)
(0.25)	(0.25)	4th (0.05)	4th (0.05)	3rd (0.07)	2nd (0.11)	1st (0.22)

We allocated equal weight to the two groups of outcomes. Sum of weights for the effectiveness outcomes (change of medication and hospitalisation) equals 0.5. And sum of the 5 side effects outcomes' weights (EPSE, hyperprolactinemia, sedation, sexual side effects, weight gain) also equals 0.5.
The weights of effectiveness outcomes are kept constant throughout all allowed preference elections.
The reciprocal weights method (RR) uses the reciprocal of the ranks which are normalized by dividing each term by the sum of the reciprocal according to the formula:

$$w_j(\mathrm{RR}) = \frac{1/r_j}{\sum_{k=1}^{n}(1/r_k)}$$

where r_j is the rank of the j – th selected side effects, $j = 1,2, \dots, n$.

Values for the reciprocal weights for number of criteria n = 5:

Rank ordering	1st	2nd	3rd	4th	5th	Sum
Weight value	0.44	0.22	0.14	0.11	0.09	1

Values for the reciprocal weights for number of criteria n = 5 and sum of weights 0.5:

Rank ordering	1st	2nd	3rd	4th	5th	Sum
Weight value	0.22	0.11	0.07	0.55	0.045	0.5

The allowed preference elicitation can be divided into four scenarios:

(i) All side effects are ranked equally
Equal weights for number of criteria n = 5 and sum of weights 0.5:

Rank ordering	1st	1st	1st	1st	1st	Sum
Weight value	0.1	0.1	0.1	0.1	0.1	0.5

(ii) Top one side effect is selected and the remaining four are ranked equally

Rank ordering	1st	2nd	2nd	2nd	2nd	Sum
Weight value	0.22	0.07	0.07	0.07	0.07	0.5

(iii) Top one and top two side effect of most concerned are selected. Remaining three are ranked equally.

Rank ordering	1st	2nd	3rd	3rd	3rd	Sum
Weight value	0.22	0.11	0.056	0.056	0.056	0.5

(iv) Top one, top two and top three side effect of most concerned are selected. Remaining two are ranked equally.

Rank ordering	1st	2nd	3rd	4th	4th	Sum
Weight value	0.22	0.11	0.07	0.05	0.05	0.5

eMethods 4.5 Stratifying PTRs into the high priority groups

To evaluate the performance of the Precision Treatment Rules (PTRs) across different preferences selected by patients we stratified the PTRs into five mutually exclusive groups. Each group includes 17 PTRs that assigns the highest preferences (weight) to a given side effect. E. g the Extra Pyramidal Side Effects (EPSE) high priority group consists of 17 PTRs where EPSE is selected as the first ranking positions of side-effects preferences. Similarly, the high priority groups of PTRs were created for hyperprolactinemia, sedation, sexual side effects, and weight gain. The single PTR that ranks all side effects equally was not included in neither of the high priority groups.

eTable 4.3 Classification instability index averaged over all preferences and for sub-groups of preferences prioritizing given side effect.

Scenario	Classification instability index, %
All preferences average	5.9
EPSE high priority preferences	19.4
Hyperprolactinemia high priority preferences	1.9
Sedation high priority preferences	2.4
Sexual side effects high priority preferences	2.7
Weight gain high priority preferences	3.2

Table 4.3 Classification instability index. Average for all preferences combination and for groups of preferences with the highest priority for a given side effect.

TRIPOD + AI Guidelines

Section/Topic	Item	Development / evaluation	Checklist item	Reported on page
Title				
Title	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	1
Abstract				
Abstract	2	D;E	See TRIPOD+AI for Abstracts checklist	2
Introduction				
Background	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	**3**
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	**3**
	3c	D;E	Describe any known health inequalities between sociodemographic groups	**N/A**
Objectives	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	**3**
Methods				
Data	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	**3**
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	**3**
Participants	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	**3**
	6b	D;E	Describe the eligibility criteria for study participants	**3**
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	**3**
Data preparation	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	**3**
Outcome	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	**4**
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	**N/A**
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	**N/A**
Predictors	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	**4**
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	**4**
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	**N/A**
Sample size	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	**3/Data source**
Missing data	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	**3/Data source**
Analytical methods	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	**6/Estimating performance of the PTRs**

	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	**5**
	12c	D	Specify the type of model, rationale2 , all model-building steps, including any hyperparameter tuning, and method for internal validation	**5**
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations3	**N/A**
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	**6**
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	**N/A**
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	**6**
Class imbalance	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	**N/A**
Fairness	14	D;E	Describe any approaches that were used to address model fairness and their rationale	**N/A**
Model output	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	**5**
Training versus evaluation	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	**6**
Ethical approval	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	**4**
OPEN SCIENCE				
Funding	18a	D;E	Give the source of funding and the role of the funders for the present study	**9**
Conflicts of interest	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	**9**
Protocol	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	**Protocol not prepared**
Registration	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	**No registration**
Data sharing	18e	D;E	Provide details of the availability of the study data	**N/A**
Code sharing	18f	D;E	Provide details of the availability of the analytical code	**6**
PATIENT & PUBLIC INVOLVEMENT				
Patient & Public Involvement	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	**No involvment**
RESULTS				
Participants	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	**6/eMethods 2**
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups	**6**
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	**N/A**
Model development	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	**6**

Model specification	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary)	**6**
Model performance	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	**7**
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details.	**N/A**
Model updating	24	E	Report the results from any model updating, including the updated model and subsequent performance.	**N/A**
DISCUSSION				
Interpretation	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	**8**
Limitations	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	**8/9**
Usability of the model in the context of current care	27a	D;E	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	**9**
	27b	D;E	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	**8**
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	**9**

The RECORD statement – checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data.

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Title and abstract					
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found	Title/Abstract	RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.	Title/Abstract
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	Introduction		
Objectives	3	State specific objectives, including any prespecified hypotheses	Introduction		
Methods					
Study Design	4	Present key elements of study design early in the paper	Methods/Data source		
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Methods/Data source		
Participants	6	*(a) Cohort study* - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	Methods/Data source	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If	Methods/Data source and eMethods 2

		Case-control study - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study* - Give the eligibility criteria, and the sources and methods of selection of participants *(b) Cohort study* - For matched studies, give matching criteria and number of exposed and unexposed *Case-control study* - For matched studies, give matching criteria and the number of controls per case		this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.	eMethods 2
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	Methods/Predictors	RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	Methods/Outocmes and Methods/Predictors
Data sources/ measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Methods/Data source and eTable 1		
Bias	9	Describe any efforts to address potential sources of bias	Analysis Methods		
Study size	10	Explain how the study size was arrived at	Analysis Methods/Estimating Precision Treatment Rules		
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	Methods/Data source		

Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) *Cohort study* - If applicable, explain how loss to follow-up was addressed *Case-control study* - If applicable, explain how matching of cases and controls was addressed *Cross-sectional study* - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses	Analysis Methods		
Data access and cleaning methods		.. Methods/Data source and eMethods 2		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	Methods/Data source and eMthods 2
Linkage		..		RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	Methods/Data source and eMthods 2
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the study (*e.g.*, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed)	eMethods 2	RECORD 13.1: Describe in detail the selection of the persons included in the study (*i.e.,* study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	Results/sample charachteristics eMthods 2

		(b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram			
Descriptive data	14	(a) Give characteristics of study participants (*e.g.*, demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest (c) *Cohort study* - summarise follow-up time (*e.g.*, average and total amount)	Results/sample characteristics and eTable 1		
Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report numbers in each exposure category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures	Results/Sample characteristics		
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	Results/Estimatated Average Treatment Effects; Results/Estimated Performance of the PTRs		
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	Results/Stability of the PTRs		
Discussion					
Key results	18	Summarise key results with reference to study objectives	Discussion		

Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Discussion	RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	Discussion
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Discussion		
Generalisability	21	Discuss the generalisability (external validity) of the study results	Discussion		
Other Information					
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Declaration of Interest		
Accessibility of protocol, raw data, and programming code		..	Methods	RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	Methods

*Reference: Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM, the RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Medicine* 2015; in press.

PART C: GENERAL DISCUSSION

5. General Discussion

In this section, I will present a summary of the findings of this thesis, discuss the impact of the results on the field of psychiatry and precision medicine, and outline the limitations along with recommendations for further research.

5.1 Summary of findings

The work presented in this thesis is intended to show how precision medicine can be used to improve detection (Part A) and management of psychosis (Part B). Part A demonstrates that dynamic modelling, leveraging longitudinal data, and Natural Language Processing (NLP) predictors improve the performance of the psychosis risk calculator. The comparison between the static and dynamic models reveals that the Cox landmark model achieves higher discrimination, better calibration, and greater potential clinical utility than the standard Cox model. Improvements in discriminatory performance (measured by Harrell's C score) between the static and dynamic models, averaged over all time points, were 0.035 (95% CI = 0.031–0.043, $p < 0.001$), representing a 20% error reduction. Significant gains in calibration performance and potential net benefit were achieved when predictions were made at later time points, further from the baseline. The dynamic model maintained an adequate calibration slope throughout all time points; however, the static model provided miscalibrated predictions beyond the baseline assessment. Similarly, the dynamic model demonstrated superior potential clinical utility beyond 24 months from the baseline assessment. Part A also demonstrates the generalisability of the risk calculator's accuracy across London's boroughs, as the model achieved adequate performance in all geographical regions tested through internal-external cross-validation.

Part B demonstrates how principles of shared decision-making and treatment individualisation can be incorporated into Precision Treatment Rules (PTRs) for first-line antipsychotics. The causal forest method, which estimates treatment effects conditional on patient characteristics, was extended to include patient preferences and multiple outcomes. It is the first demonstration of how patient characteristics and preferences, together with effectiveness and side effects outcomes, can be jointly used to optimise treatment selection for individuals with first-episode psychosis. Additionally, this thesis introduces the first PTRs in psychiatry to use predictors extracted by text mining algorithms from free-text clinical notes stored in Electronic Health Records (EHRs). The results demonstrate that, if the treatments had been allocated according to the PTRs' recommendations, patients would experience significantly lower rates of side effects compared to the outcomes under observed treatment decisions. Averaged across all preference profiles, hyperprolactinemia would be reduced by 4.7 percentage points (pp), sedation by 15.8 pp, sexual side effects by 4.3 pp, and weight gain by 15.2 pp. However, extrapyramidal side effects (EPSE) are projected to increase by 5.5 pp. No effect on the rates of hospitalisation or change of medication was found. The PTRs predominantly recommend aripiprazole, which is estimated to be the optimal treatment for 80% to 98% of patients, depending on their selected preferences.

5.2 Impact of findings

5.2.1 Dynamic modelling of psychosis risk

Precision medicine has great potential to improve patient care in psychosis and other mental disorders. In recent years, a relatively large number of risk models for psychosis prediction have been published (1). The vast majorly of prediction models rely on only a single cross-section of data and are unable to generate predictions continuously throughout the course of patient care. The few dynamic models published to date have been predominantly developed in Clinical High Risk for Psychosis (CHR-P) populations, using small sample sizes (2–4) . In most cases, these models relied on just a single predictor (2,3) or on predictors not routinely collected in clinical practice. The presented findings advance knowledge of psychosis detection and dynamic risk modelling in several important ways. The presented Cox landmarking model offers a pragmatic approach to processing complex longitudinal data including multiple sociodemographic, clinical and NLP predictors to generate dynamic predictions. The approach is computationally effective, which have not been achieved with previous Joint Modelling methods. It was developed and internally-externally validated in the biggest secondary mental health care facility in Europe (South London and Maudsley NHS Trust) and is especially well suited for large scale and automated detection, which is not the case for models developed in CHR-P samples. By enhancing existing risk calculator (5,6) that have been extensively externally validated (7) and implemented in routine clinical care (8,9) the model is ready for implementation impact study and advances translational knowledge that can improve patient care and outcomes.

Moreover, by using the internal-external validation approach (10), this thesis further advances knowledge of the generalisability of the transdiagnostic risk calculator across multiple catchment areas in London, demonstrating adequate performance in all regions. The improved validation method, compared to the previously used single data partition, allows for obtaining final coefficients of the model using the entire data set, enabling more efficient use of the available data. The thesis also advances knowledge by presenting an innovative method to evaluate performance differences between static and dynamic models. By using the meta-regression framework (11,12) and generating predictions at multiple time points for both models, the difference in performance has been quantified, representing, to the best of my knowledge, the first such result in precision psychiatry.

5.2.2 PTRs for first episode of psychosis

Research in precision psychiatry has been predominantly focused on predicting patient outcomes using clinical prediction models. However, moving form risk prognosis of an outcome to specific clinical decisions is not always clear. Very little work has been done to develop systems that directly model and estimate the individualised effects of competing treatment options and recommend optimal treatment for a given individual, representing substantial research gap (13). Chapter 3 outlines numerous barriers to treatment individualisation in psychiatry such as insufficient sample sizes in Randomised Clinical Trials (RCTs), inconsistent measurements across trials, inadequate methodology for statistical analysis and the design of observational studies. The recently published precision psychiatry framework offers a solution to the aforementioned problems (14). To date there have been only one study (15) that developed a PTR for first-line antipsychotics recommendation that used large EHRs data and sophisticated statistical methods (16) capable of estimating conditional treatments effects under unconfoundedness assumption. My thesis expands knowledge by demonstrating how causal machine learning methods can be extended to incorporate patient preferences to develop PTRs that facilitate shared decision-

making. I presented the first study that used multiple side effects outcomes extracted form EHRs, together with effectiveness outcomes that were used before. It is the first demonstration of how individual patient characteristics, preferences, side effects and effectiveness outcomes can all be incorporated into one PTRs recommendation of an optimal first-line antipsychotic for first episode population. The knowledge on incorporation of shared decision-making into the PTRs have been advanced by presenting the ranking method that is not cognitively demanding for patients yet lets the patients select outcomes' weights that are subsequently used for joint optimalization of multiple included outcomes. The patient is asked to select three side effects of most concern, and the PTRs do the heavy lifting by generating optimal antipsychotic recommendation and detailed estimates of each outcome under all treatment options.

The method of developing multiple PTRs for each preference combination was presented with the causal forest method (17), however, it can be used with other causal inference approaches e.g. doable machine learning (18) or targeted learning (19). Moreover, these are the first PTRs in psychiatry that use fine grain symptoms and substance use data extracted by NLP.

Additionally, the study further strengthens the emerging evidence that Aripiorazle (20) is the optimal first-line antipsychotic as it was recommended to between 80% and 98% of patients depending on their preferences. The results are the first estimates of antipsychotics effects in real world data, using both effectiveness and side effects outcomes, accounting for heterogeneity of treatment effects, and wide spectrum of patient preferences. However, this result is limited to the comparison of the three most used antipsychotics in the UK. The other contribution to the field is the quantification of the effects when following the treatment recommendations, compared to the observed treatment assignment and the estimated benefits to side effects rates reduction.

5.3 Limitations

5.3.1 Precision of routinely collected data in EHRs

This thesis relies on routinely collected data stored in EHRs to both improve detection of psychosis and treatment management of first-episode patients. However, the diagnosis, symptoms and substance use information recorded in EHRs are collected for administrative reasons and are not psychometrically validated. Additionally, the information stored in EHRs is dependent on the extend and accuracy of what patients and clinicals decide to share in their notes. This is partially mitigated by the meta-analytical evidence suggesting that diagnostic codes recorded in EHRs are predictive of true validated diagnosis (21). Secondly, the NLP tools that were used for predictors extraction are not perfectly accurate, and identify signs and symptoms present in the clinical notes with various precision. We mitigated this by including only predictors that were extracted by NLP tools with above 80% precision. Thirdly, the information on symptoms and substance use do not reflect severity and intensity of recorded features and was used as a binary variable indicating presence or absence in a given month.

5.3.2 Confounding in observational data

A fundamental problem of causal inference with observational data is that treatment groups may systematically differ and there may be confounding factors that affect both treatment assignment and outcomes. Methods for causal effect estimation rely on the unconfoundedness assumption,

which states that all important confounders have been measured and appropriately controlled (22). This assumption can be reasonable if a rich set of variables affecting both treatment and outcome is included in the model (23). However, the assumption is untestable, and there may be sources of unmeasured confounding unknown to the analyst. I have mitigated this by using a wide range of pre-treatment covariates, and by using cutting edge causal machine learning methods (24). As outlined by the pragmatic precision psychiatry framework observational data are necessary for PTRs discovery, as RCTs have too small sample sizes and do not have consistent measurements across trials (14). The approach of cycling between PTRs discovery in observational data and evaluation in RCTs is advised as the most efficient method for using available data resources (14).

5.3.3 Restricting analysis to three most used antipsychotics

The PTRs in this study were developed by restricting the data set to the three most frequently used antipsychotics, which limits their clinical utility. This was done to ensure a sufficient number of patients for each treatment option, enabling stable estimation of conditional average treatment effects and, subsequently, the construction of PTRs. The Maudsley Prescribing Guideline lists 19 different antipsychotics approved for the treatment of psychosis (25); therefore, further research is needed to explore PTRs across a wider range of treatment options. However, the three antipsychotics included in the study account for over 74% of all prescriptions within the study population. Another limitation is that data on new treatments are not available in the EHRs. For PTRs to include a new treatment option, a sufficient sample of patients who have received a new medication must first be gathered. The presented PTRs also do not incorporate information on dosage or route of administration, such as oral or long-acting injectable forms.

5.3.4 Limited guidelines and implementations studies for novel models

Currently, there are no available guidelines for the use of either dynamic prediction models for psychosis or PTRs for first-episode psychosis. Numerous clinical and ethical questions regarding the implementation remain unanswered. E.g. the determination of optimal frequency and risk thresholds of alerts generated by a dynamic model is essential. It is not known if more frequent dynamic updates would lead to alert fatigue among clinicians. The optimal visual presentation of the PTR recommendations and estimates has yet to be established. The most suitable method for election of preferences remains unknown. Various approaches to the proposed ranking method should be compared and evaluated in studies involving patients.

5.4 Further Research

5.4.1 Integrating new types of information stored in EHRs.

The presented model for dynamic risk prediction of psychosis consists of a wide spectrum of predictors, ranging from diagnosis, demographics to substance use and symptoms extracted by NLP. However, EHRs are a rich source of information and additional variables exit that, with appropriate modelling techniques could enhance prediction accuracy. Future research could aim to include data on hospitalizations or suicide attempts, which would be particularly well-suited to the presented dynamic model, as such events occur throughout the course of secondary mental

health care. The other substantial source of information are data on pharmacological and psychological interventions that are administered to patients throughout their contact with secondary mental health care. New methods for clinical prediction models that involve causal inference techniques have been recently presented and can allow for the incorporation of treatment data that were administered after the baseline assignment (26). Additionally, patient histories on their visits with General Practitioners and their somatic health data can be linked with the psychiatric information to provide holistic clinical picture.

5.4.2 Sequential testing for psychosis risk across clinical stages

Detection of individuals at high risk of psychosis in secondary mental health care is an important component of wider prevention and subsequent management of psychosis. However, this detection approach is targeted to patients at a single clinical stage. The psychosis detection and subsequent prognosis and management could be in the future a continues process across age groups and clinical stages. The detection can start earlier at primary care (27,28). In such scenarios when patients enter the secondary mental health care the presented dynamic model can be used for multiple screenings. As presented in the implementation study of the transdiagnostic risk calculator (8), once the individual is detected by the risk model, the alter is send to the clinician and the patient can be referred for a psychometric CHR-P assessment. As explained earlier theses assessments have high sensitivity, however their specificity is low. The prognostication of psychosis can be further refined with the use Psychosis Poly Risk Scours (29), and additional psychometric and biological testing. The genetic (30), and blood tests (31), together with the neuroimaging (32) have been showed to discriminate between transiting and non-transitioning individuals with various accuracy. However, any additional testing is associated with additional cost and burden for the patients and health care system. Further research can explore sequential testing approach where each subsequent test is performed only for individuals meeting certain risk threshold and is conditional on all previously gathered information (33).

5.4.3 PTRs for multiple rounds of antipsychotics

The presented thesis describes how PTRs for first-line antipsychotics recommendations incorporating patient preferences can be developed. However, psychosis disorders for many patients are long term and chronic disease. After 5.5 years around 32% of first-episode patients recover, and most patients experience subsequent episodes (34). The discontinuation rates of all antipsychotics for first-episode patients are high (35) and, therefore, further research should explore PTRs for second-line antipsychotics and possibly further rounds including clozapine. In recent years methods on Dynamic Treatment Regimes that can support multiple rounds of treatment have been published (36,37). The Dynamic Treatment Regimes can incorporate patient histories consisting of time varying confounders to optimise a sequence of optimal treatments. The information on the previously prescribed antipsychotics and other treatment together with the reason for specific antipsychotic discontinuation can be incorporated using these methods. Current evidence and recommendations in the guidelines are inconclusive on how to guide selection of second-line antipsychotic (25). PTRs can close this knowledge gap and contribute to better patient care.

5.4.4 Transdiagnostic approaches

Prodromal periods of severe mental disorders share similar clinical presentation across different disorders. In South London and Maudsley NHS Trust in the UK there are only neglectable or small differences between patients experiencing the prodromal of non-psychotic unipolar mood disorders, non-psychotic bipolar mood disorders and psychotic disorders (38). Further research on psychosis detection should be extended to include risk assessment for all severe mental disorders (39). Risk prediction models can either predict severe mental disorders as a group or provide individual risk assessments for each disorder. As patients share similar clinical presentation, similar preventive intervention can be offered and benefit patients across severe mental disorders, stressing the need for transition to transdiagnostic detection strategies.

Similarly, in the future PTRs can be extended across mental disorders. Comorbidity among first episode of psychosis cohort is high and patients are often prescribed other psychopharmacological treatments alongside antipsychotics (40). Similarly to the detection of psychosis, the PTRs can be implemented at different clinical stages and include multiple treatment options for multiple different diagnosis. To accomplish such an ambitious project much larger and richer data sets than the one used in this study are need. This might become possible with nationwide integration of EHRs for research purposes which is has already been achieved for the covid- 19 study (41).

5.5 Conclusions

This thesis presents work that advances knowledge on psychosis detection and management of first episode cases using innovative precision medicine methods. Part A presents a first quantitative comparison between static and dynamic approaches to large scale detection of individuals at risk of psychosis in secondary mental health care. The presented Cox Landmark model improves discrimination, calibration, and potential clinical utility compared to the previously used static risk calculator. The model is well suited for implementation study and for improving the large-scale detection of at-risk individuals. The part B demonstrates an innovative method for treatment optimization in the first episode of psychosis. The presented PTRs are the first to jointly integrate effectiveness, side effects, and individual patients' characteristics and preferences. By allowing patients to express their preferences with the ranking method the presented PTRs advance knowledge on the incorporation of shared decision making into precision medicine. The results further strengthen the evidence that among the three compared antipsychotics aripiprazole is the optimal option and the analysis is the first to jointly consider effectiveness, side effects, patients' preferences and heterogeneity of treatment effects. This part also estimates that under the PTRs recommendations significant benefits to most side effects can be expected. Together, these findings advance knowledge on innovative precision medicine methods improving the detection and management of psychosis. Consequently, this work can be used to improve outcomes and quality of life for young people at risk of psychosis or going through antipsychotics treatments.

5.6 References

1. Meehan AJ, Lewis SJ, Fazel S, Fusar-Poli P, Steyerberg EW, Stahl D, Danese A (2022): Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Mol Psychiatry* 27: 2700–2708.

2. Yuen HP, Mackinnon A, Hartmann J, Amminger GP, Markulev C, Lavoie S, *et al.* (2018): Dynamic prediction of transition to psychosis using joint modelling. *Schizophr Res* 202: 333–340.

3. Studerus E, Beck K, Fusar-Poli P, Riecher-Rössler A (2020): Development and Validation of a Dynamic Risk Prediction Model to Forecast Psychosis Onset in Patients at Clinical High Risk. *Schizophr Bull* 46: 252–260.

4. Zhang T, Tang X, Zhang Y, Xu L, Wei Y, Hu Y, *et al.* (2023): Multivariate joint models for the dynamic prediction of psychosis in individuals with clinical high risk. *Asian J Psychiatry* 81: 103468.

5. Fusar-Poli P, Rutigliano G, Stahl D, Davies C, Bonoldi I, Reilly T, McGuire P (2017): Development and Validation of a Clinically Based Risk Calculator for the Transdiagnostic Prediction of Psychosis. *JAMA Psychiatry* 74: 493–500.

6. Irving J, Patel R, Oliver D, Colling C, Pritchard M, Broadbent M, *et al.* (2020): Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Schizophr Bull* 47: 405–414.

7. Oliver D, Wong CMJ, Bøg M, Jönsson L, Kinon BJ, Wehnert A, *et al.* (2020): Transdiagnostic individualized clinically-based risk calculator for the automatic detection of individuals at-risk and the prediction of psychosis: external replication in 2,430,333 US patients. *Transl Psychiatry* 10: 364.

8. Oliver D, Spada G, Colling C, Broadbent M, Baldwin H, Patel R, *et al.* (2021): Real-world implementation of precision psychiatry: Transdiagnostic risk calculator for the automatic detection of individuals at-risk of psychosis. *Schizophr Res* 227: 52–60.

9. Wang T, Oliver D, Msosa Y, Colling C, Spada G, Roguski Ł, *et al.* (2020): Implementation of a real-time psychosis risk detection and alerting system based on Electronic Health Records using CogStack. *J Vis Exp JoVE* 10.3791/60794.

10. Steyerberg EW, Harrell FE (2016): Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 69: 245–247.

11. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS (2016): External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *The BMJ* 353: i3140.

12. IntHout J, Ioannidis JP, Borm GF (2014): The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 14: 25.

13. Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, *et al.* (2024): Causal machine learning for predicting treatment outcomes. *Nat Med* 30: 958–968.

14. Kessler RC, Luedtke A (2021): Pragmatic Precision Psychiatry-A New Direction for Optimizing Treatment Selection. *JAMA Psychiatry* 78: 1384–1390.

15. Wu C-S, Luedtke AR, Sadikova E, Tsai H-J, Liao S-C, Liu C-C, *et al.* (2020): Development and Validation of a Machine Learning Individualized Treatment Rule in First-Episode Schizophrenia. *JAMA Netw Open* 3: e1921660.

16. Zheng W, van der Laan MJ (2011): Cross-Validated Targeted Minimum-Loss-Based Estimation. In: van der Laan MJ, Rose S, editors. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer, pp 459–474.

17. Tibshirani J, Athey S, Sverdrup E, Wager S (2024): grf: Generalized Random Forests. p 2.4.0.

18. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018): Double/debiased machine learning for treatment and structural parameters. *Econom J* 21: C1–C68.

19. Van Der Laan MJ, Rose S (2018): *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65304-4

20. McCutcheon RA, Pillinger T, Varvari I, Halstead S, Ayinde OO, Crossley NA, *et al.* (2025): INTEGRATE: international guidelines for the algorithmic treatment of schizophrenia. *Lancet Psychiatry* 12: 384–394.

21. Davis KAS, Sudlow CLM, Hotopf M (2016): Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. *BMC Psychiatry* 16: 263.

22. Rubin DB, Imbens GW (Eds.) (2015): Assessing Unconfoundedness. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, pp 479–495.

23. Ding P (2024): Observational Studies, Selection Bias, and Nonparametric Identification of Causal Effects. *A First Course in Causal Inference*. Chapman and Hall/CRC.

24. Athey S, Tibshirani J, Wager S (2019): Generalized random forests. *Ann Stat* 47: 1148–1178.

25. Taylor, David M, Thomas RE Barnes,, Allan H. Young (2021): *The Maudsley Prescribing Guidelines in Psychiatry 14th Edition*. Chichester UK: John Wiley & Sons.

26. van Geloven N, Swanson SA, Ramspek CL, Luijken K, van Diepen M, Morris TP, *et al.* (2020): Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol* 35: 619–630.

27. Fusar-Poli P, Sullivan SA, Shah JL, Uhlhaas PJ (2019): Improving the Detection of Individuals at Clinical Risk for Psychosis in the Community, Primary and Secondary Care: An Integrated Evidence-Based Approach. *Front Psychiatry* 10: 774.

28. Woodberry KA, Johnson KA, Shrier LA (2022): Screening for Early Emerging Mental Experiences (SEE ME): A Model to Improve Early Detection of Psychosis in Integrated Primary Care. *Front Pediatr* 10. https://doi.org/10.3389/fped.2022.899653

29. Oliver D, Radua J, Reichenberg A, Uher R, Fusar-Poli P (2019): Psychosis Polyrisk Score (PPS) for the Detection of Individuals At-Risk and the Prediction of Their Outcomes. *Front Psychiatry* 10: 174.

30. Legge SE, Jones HJ, Kendall KM, Pardiñas AF, Menzies G, Bracher-Smith M, *et al.* (2019): Association of Genetic Liability to Psychotic Experiences With Neuropsychotic Disorders and Traits. *JAMA Psychiatry* 76: 1256–1265.

31. Fuentes-Claramonte P, Estradé A, Solanes A, Ramella-Cravaro V, Garcia-Leon MA, de Diego-Adeliño J, *et al.* (2024): Biomarkers for Psychosis: Are We There Yet? Umbrella Review of 1478 Biomarkers. *Schizophr Bull Open* 5: sgae018.

32. Zhu Y, Maikusa N, Radua J, Sämann PG, Fusar-Poli P, Agartz I, *et al.* (2024): Using brain structural neuroimaging measures to predict psychosis onset for individuals at clinical high-risk. *Mol Psychiatry* 29: 1465–1477.

33. Schmidt A, Cappucciati M, Radua J, Rutigliano G, Rocchetti M, Dell'Osso L, *et al.* (2017): Improving Prognostic Accuracy in Subjects at Clinical High Risk for Psychosis: Systematic Review of Predictive Models and Meta-analytical Sequential Testing Simulation. *Schizophr Bull* 43: 375–388.

34. Catalan A, Richter A, Pablo GS de, Vaquerizo-Serrano J, Mancebo G, Pedruzo B, *et al.* (2021): Proportion and predictors of remission and recovery in first-episode psychosis: Systematic review and meta-analysis. *Eur Psychiatry* 64: e69.

35. Szmulewicz AG, Martínez-Alés G, Logan R, Ferrara M, Kelly C, Fredrikson D, *et al.* (2024): Antipsychotic drugs in first-episode psychosis: a target trial emulation in the FEP-CAUSAL Collaboration. *Am J Epidemiol* 193: 1081–1087.

36. Tsiatis AA, Davidian M, Holloway ST (2021): *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC Press Boca Raton USA.

37. Liang D, Paul AK, Weir DL, Deneer VHM, Greiner R, Siebes A, Gardarsdottir H (2025): Methods in dynamic treatment regimens using observational healthcare data: A systematic review. *Comput Methods Programs Biomed* 263: 108658.

38. Arribas M, Barnby JM, Patel R, McCutcheon RA, Kornblum D, Shetty H, *et al.* (2025): Longitudinal evolution of the transdiagnostic prodrome to severe mental disorders: a dynamic temporal network analysis informed by natural language processing and electronic health records. *Mol Psychiatry* 30: 2931–2942.

39. Oliver D (2024): The future of preventive psychiatry is precise and transdiagnostic. *Neurosci Biobehav Rev* 160: 105626.

40. Strålin P, Hetta J (2021): First episode psychosis: register-based study of comorbid psychiatric disorders and medications before and after. *Eur Arch Psychiatry Clin Neurosci* 271: 303–313.

41. Wood A, Denholm R, Hollings S, Cooper J, Ip S, Walker V, *et al.* (2021): Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 373: n826.

APPENDIX: WORK PUBLISHED DURING PHD

Arribas, M., Oliver, D., Patel, R., Kornblum, D., Shetty, H., Damiani, S., **Krakowski, K**., Provenzani, U., Stahl, D., Koutsouleris, N. and McGuire, P., 2024. A transdiagnostic prodrome for severe mental disorders: an electronic health record study. *Molecular Psychiatry*, *29*(11), pp.3305-3315.

Arribas, M., Barnby, J.M., Patel, R., McCutcheon, R.A., Kornblum, D., Shetty, H., **Krakowski, K**., Stahl, D., Koutsouleris, N., McGuire, P. and Fusar-Poli, P., 2025. Longitudinal evolution of the transdiagnostic prodrome to severe mental disorders: a dynamic temporal network analysis informed by natural language processing and electronic health records. *Molecular Psychiatry*, pp.1-12.

Krakowski, K., Oliver, D., Arribas, M., Stahl, D. and Fusar-Poli, P., 2024. Dynamic and transdiagnostic risk calculator based on natural language processing for the prediction of psychosis in secondary mental health care: development and internal-external validation cohort study. *Biological psychiatry*, *96*(7), pp.604-614.

Arribas, M., de Micheli, A., **Krakowski, K.**, Stahl, D., Correll, C. U., Young, A. H., ... & Fusar-Poli, P. (2026). Joint detection of risk for psychotic disorders or bipolar disorders in clinical practice in the UK: development and validation of a clinical prediction model. *The Lancet Psychiatry, 13(1)*, 14-23.z

Oliver, D., Arribas, M., Perry, B.I., Whiting, D., Blackman, G., **Krakowski, K**., Seyedsalehi, A., Osimo, E.F., Griffiths, S.L., Stahl, D. and Cipriani, A., 2024. Using electronic health records to facilitate precision psychiatry. *Biological psychiatry*, *96*(7), pp.532-542.

De Micheli, A., Provenzani, U., **Krakowski, K**., Oliver, D., Damiani, S., Brondino, N., McGuire, P. and Fusar-Poli, P., 2024. Physical Health and Transition to Psychosis in People at Clinical High Risk. *Biomedicines*, *12*(3), p.523.

Publications accepted for publication, pending release

Krakowski, K., Oliver, D., Aribas, M., Logeswaran, Y.L., de Micheli, A., Patel, R., Stahl, D. and Fusar-Poli, P., 2025. Development and Validation of a Precision Treatment Rules for First-Line Antipsychotic Recommendations in First Episode of Psychosis Jointly Incorporating Effectiveness, Side Effects and Patient Preferences. Preprint available at http://dx.doi.org/10.2139/ssrn.5286089, Translational Psychiatry, 2025.