manuscript\_v1

Nicole White, Rex Parsons, David Borg, Adrian Barnett

08/07/2022

# Introduction

Medical decision-making requires accurate information about patient risk to guide appropriate management and care. Patient risk estimates support decision-making by providing information on the chances of developing a disease or health condition [ref] or experiencing a future event based on an individual’s current health [ref]. For example, early knowledge of an individual’s increased risk of heart disease may prompt preventative measures to delay or reduce the likelihood of onset [ref]. Conversely, identifying patients at low risk of cancer recurrence may help prioritize high-cost treatments and monitoring for vulnerable patients [ref]. Routinely collected data sources, including demographics, medical history, and clinical test results, are often used to infer patient risk. As advances in health technology add to the volume and variety of data available, advances in quantitative methods are also needed to transform these data into reliable and interpretable outputs. Clinical prediction models offer a general framework for risk estimation based on multiple data sources. When successfully validated and implemented, clinical prediction models have the potential to progress the shift from evidence-based to personalized medicine [ref].

Despite their perceived value for personalized medicine, clinical prediction models are a growing source of research waste. Systematic reviews of diagnostic and prognostic prediction models have consistently found that most models are poorly reported and unsuitable for use in practice [refs]. The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement was introduced in 2015 to help researchers better report the development, validation and updating of clinical prediction models [ref]. Related guidelines to assess risk of bias have also been developed [ref]. Whilst the TRIPOD statement has led to modest improvements in conduct and reporting [refs], poor quality models are still being published [ref]. At the start of the COVID-19 pandemic, a statistician-led systematic review found 20 diagnostic and 10 prognostic models for predicting COVID-19 outcomes [refs]. All models were assessed as being at high risk of bias due to low sample sizes, inappropriate choice of predictors, and inadequate model assessment. By July 2022, the number of published prognostic models had increased to 606, comprising 381 newly developed models and 225 validation studies on existing models. Updated results showed that 90% of models were unsuitable for clinical use. These findings imply that best practices for clinical prediction modelling are being ignored, and errors made during the design stage are major contributors to poor research quality [ref].

Current reviews of clinical prediction models are likely affected by publication bias, as studies that report favourable results are more likely to be submitted and published [ref]. Publication bias severely hinders progress in health and medical research, as the failure to publish “negative” findings inevitably leads to duplicated efforts by other researchers to address the same research question(s). If publication bias exists in the clinical prediction modelling literature, then many more models will have been planned but never completed or published. Approaches to identify planned studies are an opportunity to examine non-publication and current areas of research focus, including studies aiming to predict the same outcome(s). Sources of bias in proposed models may also be identified early and remedied, to support improvements in research quality.

Study registries are a useful way to evaluate current research before it is published [ref]. Clinicaltrials.gov is an online study registry launched in 2000 to improve access to information on planned, ongoing and completed clinical studies. Since then, over 420,000 studies have been registered from 221 countries (<https://www.clinicaltrials.gov/>; last viewed: 8 July 2022). Previous research has analysed data collected by clinicaltrials.gov to assess reporting completeness and outcomes, including non-publication [refs]. In this paper, we analysed registered studies that included the development and/or validation of a clinical prediction model as part of their planned research. Our analysis addressed three objectives. First, we summarised trends in diagnostic and prognostic prediction models registered for all diseases and health conditions. Second, we reviewed reporting of planned versus actual sample sizes, as a common source of bias in published clinical prediction models. Third, we estimated the time to publication for all registered studies, by linking clinicaltrials.gov study identifiers to publication records in PubMed [ref].

# Data and Methods

Clinicaltrials.gov is an online database designed to improve access to information about planned, ongoing and completed clinical studies. The database was launched in 2000 by the National Institute of Health’s National Library of Medicine, in response to United States legislation mandating the registration of funded clinical trials. Since then, over 420,000 studies have been registered, comprising both interventional and observational studies from 221 countries (<https://www.clinicaltrials.gov/>; last viewed: 8 July 2022).

Study registration is completed by a nominated study investigator using a standardised template. Compulsory fields cover the disease or condition of interest, planned commencement and end dates, study type and study design details including interventions (if any), outcomes measures and planned sample size. Additional information on participant recruitment, funding and regulatory oversight is also required. A statistical analysis plan is not required at the time of registration, but can be included as part of an optional detailed summary. Study records can be updated at any time up until project completion, including providing details of resulting publications.

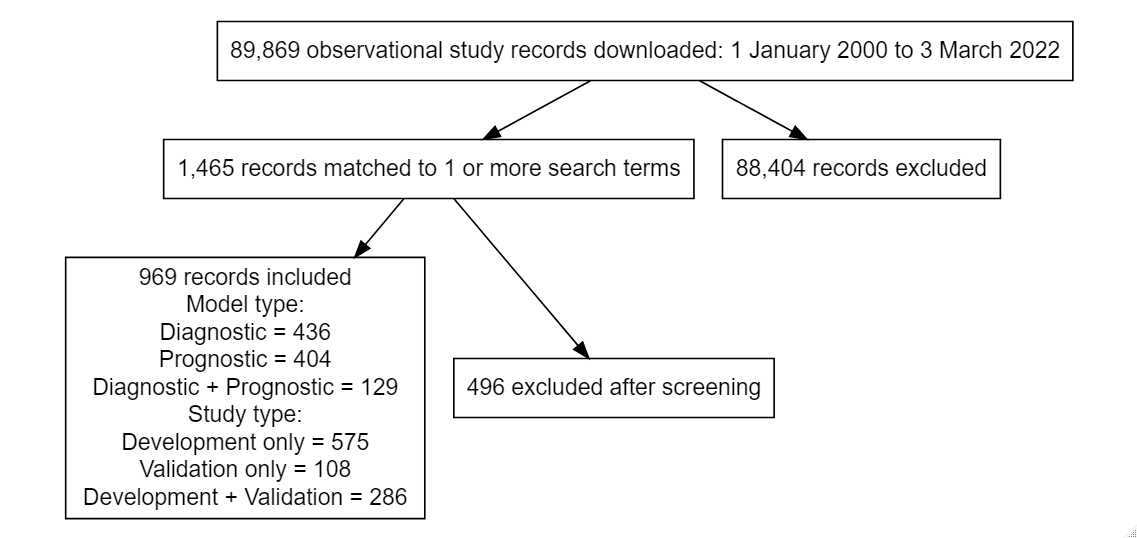
For this analysis, we identified and analysed outcomes of registered studies that planned to develop or validate a clinical prediction model (data downloaded on 3 March 2022). A clinical prediction model was defined as any multivariable model designed to estimate the individual-level risk of being diagnosed with a single disease or health condition (diagnostic model) or experiencing a future health-related outcome following diagnosis (prognostic model) (Hemingway et al. (2013)). Future outcomes considered by prognostic models included both pre-defined clinical endpoints, such as treatment response and mortality, and measures of an individual’s health state following diagnosis, such as health-related quality of life.

All records classified as observational study designs were included. All other study types were excluded [need to provide rationale here for excluding RCTS]. Dependent variables could be defined as a binary, continuous or time-to-event outcome. No restrictions were placed on model data sources or choice of independent variables, provided variables were defined at the individual level and at least two variables were planned for inclusion in the proposed model. This meant that independent variables could take the form of patient-level characteristics (e.g., demographics, clinical parameters) or processed features (e.g, extracted from images).

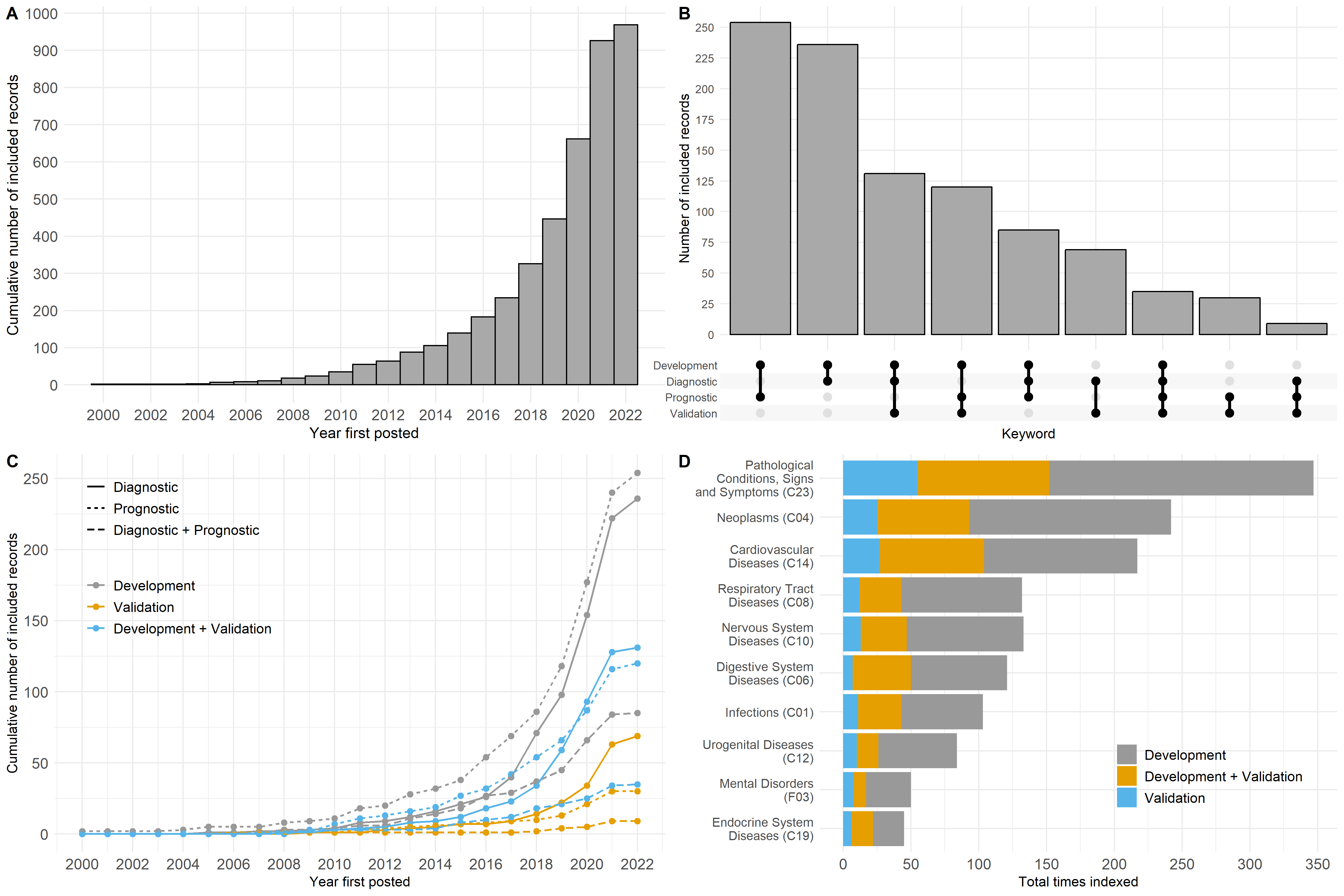
Eligible studies proposed a clinical prediction model as the primary study objective, or as an objective within a larger study. Studies that planned to identify or validate independent risk factors associated with outcomes without mention of developing a subsequent prediction model were excluded.

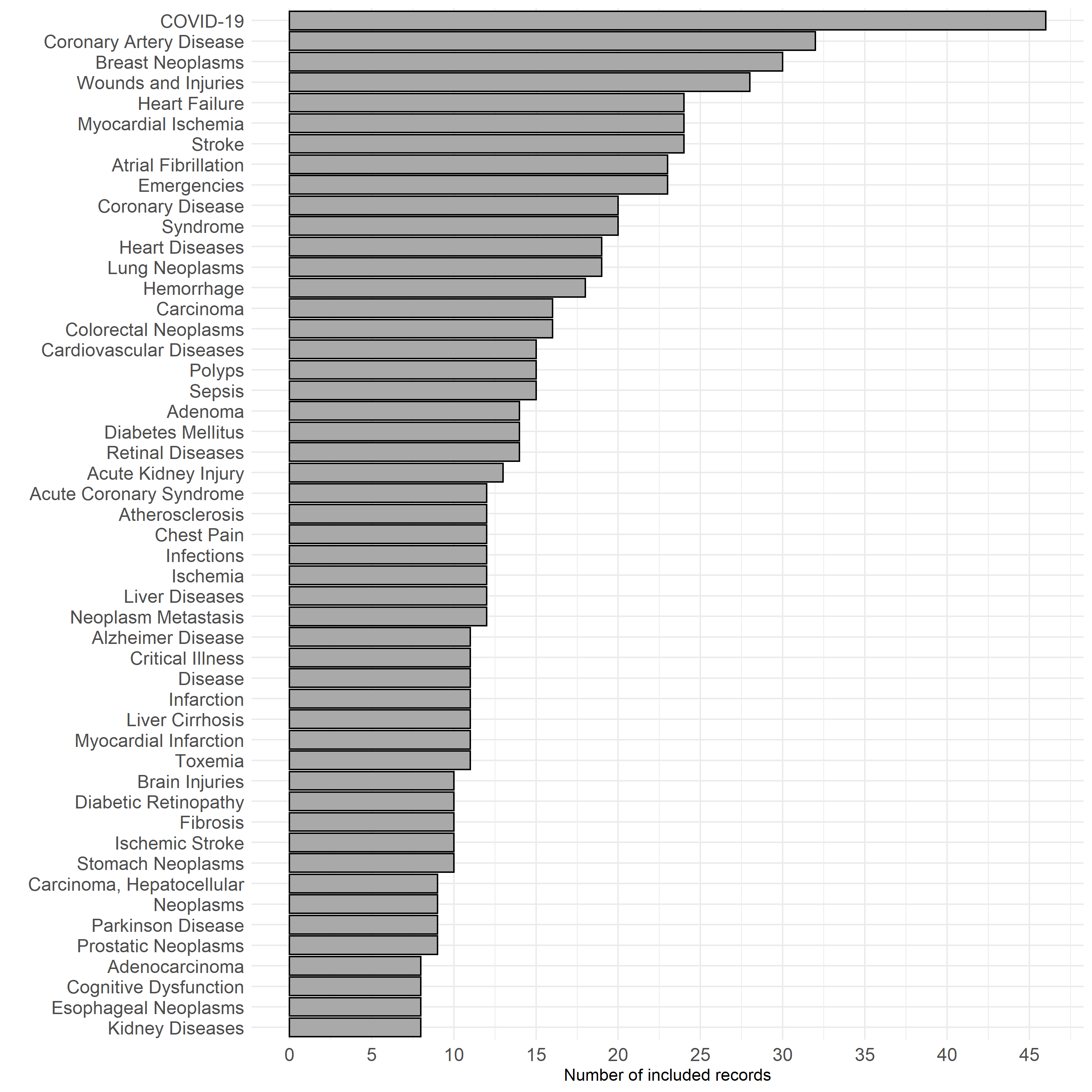
All study records posted on clinicaltrials.gov until 3 March 2022 were downloaded in XML for analysis (n = xxx). Relevant studies were identified over two stages. In the first stage, we scanned study record fields for keywords reflecting approaches to clinical prediction modelling and their application to diagnosis and prognosis (Table here - search terms + fields searched). Matching studies were then manually screened for eligibility using the web application rayyan. All studies were independently reviewed by at least two study authors.

* As table: Search terms were:“machine learning”,“artificial intelligence”,“deep learning”, “prediction model”,“predictive model”,“prediction score”,“predictive score”, “warning score”,“risk score”,“risk prediction”, “prognostic model”,“diagnostic model”

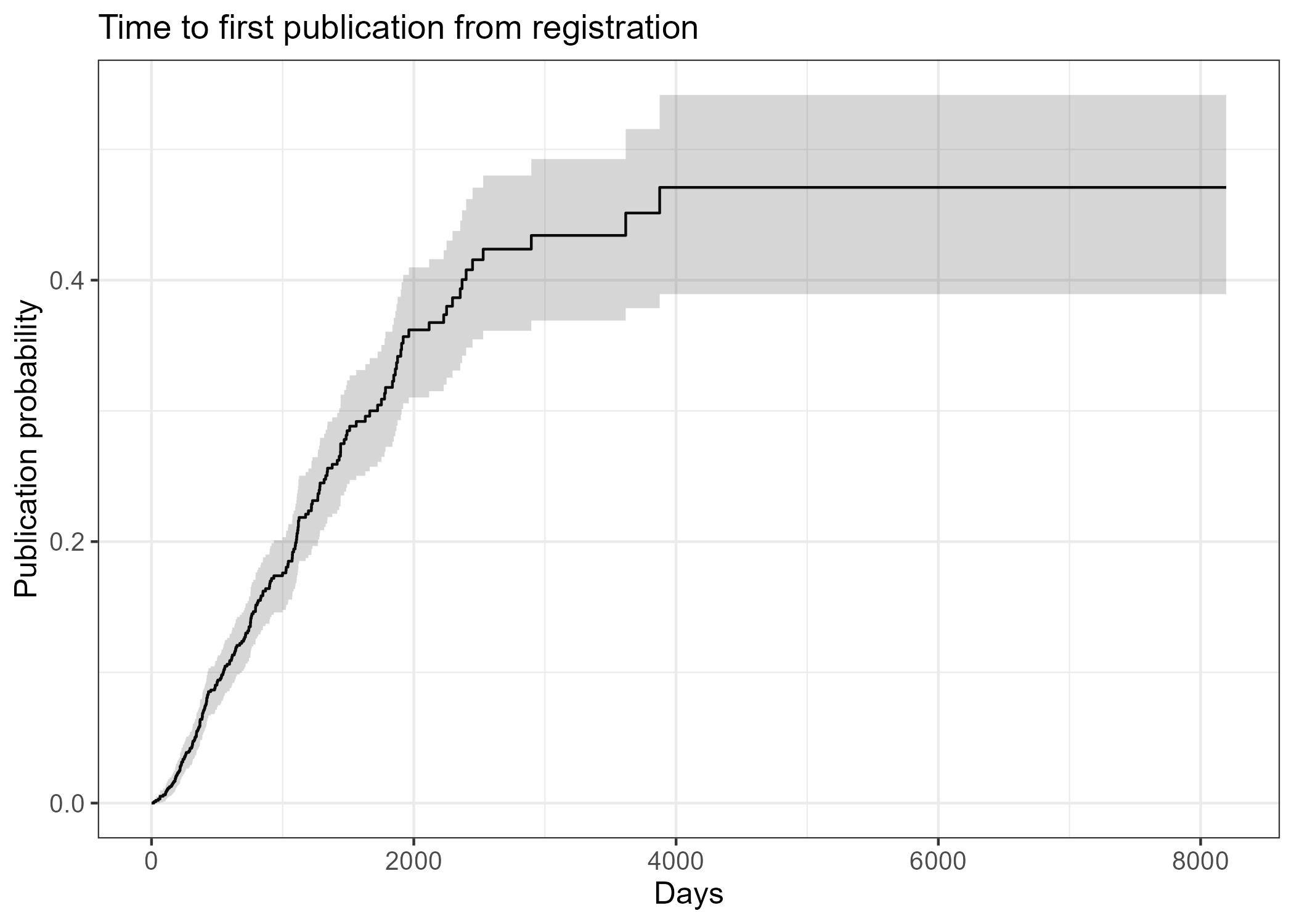


Flowchart note pretend numbers and fix typos!

 summary here



Top 50 MeSH terms



CIF: time to first publication (check caption)

## Analysis

* Study status; all identified studies, possibly compared with random sample of observational studies that don’t return match to search terms?
* Sample size history for final sample
* Frequency of targeted keywords (e.g, (.\*)validation, TRIPOD)
* Links to publications using python tool (Rex to lead)

# Discussion

* need to scope existing published models for a given outcome before developing another one! [refs - riley]
* Before embarking on a prediction modelling study, its important to ensure a sufficient sample size is available and that proposed predictors/outcome data collection is of sufficient quality. Just as per any planned statistical analysis. Poor choice will likley lead to poor performance down this line (garbage in = garbage out). ref: <https://www.bmj.com/content/338/bmj.b604> (royston2009)

# References

Hemingway, Harry, Peter Croft, Pablo Perel, Jill A Hayden, Keith Abrams, Adam Timmis, Andrew Briggs, et al. 2013. “Prognosis Research Strategy (PROGRESS) 1: A Framework for Researching Clinical Outcomes.” *Bmj* 346.