International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

Machine Learning for Chronic Disease Case Identification in Primary Care: Protocol for a Scoping Review

INPLASY2025120002
doi: 10.37766/inplasy2025.12.0002
Received: 1 December 2025
Published: 1 December 2025

Pham, ANQ; Lindeman, C; Cummings, M; Aponte-Hao, S; Kjelland, K.

Corresponding author:
Anh Pham

nqp@sfu.ca

Author Affiliation:
Simon Fraser University.

ADMINISTRATIVE INFORMATION

Support - None.

Review Stage at time of this submission - Completed but not published.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY2025120002

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 1 December 2025 and was last updated on 1 December 2025.

INTRODUCTION

Review question / Objective This scoping review will examine how machine learning (ML) methods are used to develop case definitions for chronic diseases using primary care electronic medical record (EMR) or electronic health record (EHR) data.

The specific objectives are to:

Primary objectives

- To describe the machine learning methods used to develop chronic disease case definitions in primary care.
- To describe the metrics found in these studies (e.g., case definition accuracy).
- To describe the transparency and interpretability of resulting case definitions (e.g., rules-based vs black-box approaches).

Secondary objective

- To describe knowledge translation and dissemination of these results to knowledge users (e.g., health care practitioners, policy makers, sponsors, etc.).

Background Primary care EMRs are longitudinal clinical records documenting patient encounters, diagnoses, procedures, and prescribed medications for individuals receiving care from family physicians and other primary care providers. Their increasing adoption has made primary care EMR/EHR data a key resource for health services research, chronic disease surveillance, and quality improvement in many jurisdictions (Birkhead et al., 2015).

Accurate identification of chronic disease cases in EMR data (EMR phenotyping) is methodologically challenging. Variability in coding practices, evolving diagnostic criteria, multimorbidity, and the mix of structured and unstructured data all complicate case definition development (Weiskopf & Weng, 2013). Traditional rule-based phenotyping relies heavily on expert-defined combinations of diagnosis codes, medications, and sometimes laboratory results. While transparent, these approaches can be time-consuming to develop, may be static, and can be difficult to adapt across settings.

ML methods offer an alternative, data-driven route to EMR phenotyping by learning case definitions from labelled reference datasets and using high-dimensional feature representations (e.g., codes, medications, demographics, text-derived features) (Pendergrass & Crawford, 2019). However, their use in primary care EMR data for chronic disease case identification has not previously been mapped comprehensively in a scoping review. Existing studies vary widely in ML techniques, validation practices, and reporting, which limits comparability and clinical uptake.

Rationale Chronic diseases (e.g., cardiometabolic conditions, chronic respiratory disease, mental disorders, frailty) account for a substantial proportion of global morbidity, health care utilization, and mortality. Primary care is often the first point of contact and the locus of long-term management for these conditions, making accurate EMR-based case definitions particularly important for surveillance, risk stratification, and policy planning.

Despite growing enthusiasm for ML in health care, there is limited consolidated evidence about how ML has actually been applied to develop chronic disease case definitions in primary care EMR datasets, how these models are validated, and whether their outputs are interpretable and actionable for end-users. The reviewed literature suggests inconsistent reporting of ML workflows, variable use of validation metrics, and limited attention to knowledge translation or clinical implementation.

A scoping review methodology was therefore chosen to:

- map the breadth and characteristics of existing ML-based phenotyping studies in primary care EMR/EHR data;
- identify common methodological patterns and gaps; and
- highlight implications for transparency, validation, and clinical relevance, rather than estimating pooled effect sizes. (Peters et al., 2020)

This protocol formalizes the methods used in the completed review and supports transparent post-study registration.

METHODS

Strategy of data synthesis The scoping review is conducted according to the methodological framework of Arksey and O'Malley, with refinements by Levac et al., and reported in line with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidance (Arksey & O'Malley, 2005; Levac et al., 2010).

Strategy of data synthesis

After completing searches in all specified databases, records were imported into Covidence systematic review software, and duplicates were removed. Screening and data extraction proceeded in the following steps:

1. Title and abstract screening

o Two reviewers independently screened titles and abstracts against pre-specified inclusion and exclusion criteria.

o A short pilot phase was conducted to calibrate understanding of the criteria.

o Disagreements were resolved through discussion, and, if needed, with input from a third reviewer.

2. Full-text screening

o Potentially eligible articles were retrieved in full.

o Each full text was independently assessed by two reviewers.

o Discrepancies in eligibility judgments were resolved by discussion within the review team.

3. Data extraction (charting)

o A structured data extraction form (codebook) was developed and piloted on a subset of included articles (n ≈ 5) by multiple team members, then iteratively refined.

o The final data extraction form captured:

bibliographic and study characteristics (authors, year, country, data source);

database and data type details (primary care EMR/EHR name, linkage to other datasets);

population characteristics;

chronic disease(s) or condition(s) targeted;

feature types and feature reduction methods;

ML model types and training approaches;

reference standard (gold standard) used;

handling of class imbalance (e.g., case–control sampling, resampling);

internal and external validation strategies;

performance metrics (e.g., sensitivity, specificity, PPV, NPV, AUC, F1);

interpretability of the resulting case definition (human-readable rules vs black-box);

reporting of KT or dissemination activities to clinical or policy audiences.

4. Synthesis

o Extracted data were summarized primarily using descriptive statistics (frequency counts, proportions) for categorical variables.

o Results were presented in summary tables describing databases, ML methods, validation practices, and case definition properties.

o A narrative synthesis explored patterns across studies, methodological gaps (e.g., limited external validation, inconsistent metrics), and implications for transparency, interpretability, and KT.

o No formal meta-analysis or quantitative pooling was performed due to heterogeneity in methods, outcomes, and reporting.

Eligibility criteria Eligibility criteria
Study type and focus
Inclusion criteria
Studies were included if they:
1. Used primary care EMR or EHR data as a core component of the data source; linked datasets (e.g., hospital discharge, laboratory, administrative data) were eligible if the primary care component remained central.
2. Applied machine learning methods (e.g., tree-based models, neural networks, support vector machines, Naïve Bayes, regularized regression, ensemble methods) to develop or validate case definitions for chronic diseases or long-term conditions.
3. Focused on case identification / case definition (phenotyping), i.e., identifying prevalent or existing cases or refining diagnostic criteria, rather than predicting incident future disease.
4. Were original, data-based research articles published in peer-reviewed journals.
5. Were published from 2000 onwards and in English.

Exclusion criteria
Studies were excluded if they:
• Primarily used data from non-primary-care settings (e.g., specialty outpatient clinics, biobanks) without a central primary care EMR/EHR component.
• Focused on predictive modelling of incident disease (risk prediction) rather than defining or validating case definitions for existing chronic disease cases.
• Were not primary research (e.g., reviews, editorials, commentaries, letters, conference abstracts, theses, grey literature).
• Were published prior to 2000 or were not available in English.
This focus was chosen to keep the scope specifically on recent phenotyping/case definition in primary care EMR data and to avoid conflating this with prediction-focused ML work.

Sample
Included studies could involve any adult or paediatric patient population represented in primary care EMR/EHR data, provided that the study developed or validated a chronic disease case definition using ML methods. No restrictions were placed on sex/gender, ethnicity, or geographic location of patients.
Chronic diseases/conditions included, but were not limited to:
• cardiometabolic diseases (e.g., diabetes, heart failure, hypercholesterolemia),
• chronic respiratory diseases (e.g., COPD, asthma),
• autoimmune conditions (e.g., rheumatoid arthritis, primary Sjögren's syndrome),
• frailty,
• mental health conditions (e.g., dementia, PTSD),
• other long-term conditions explicitly treated as chronic in the source articles.

Source of evidence screening and selection
Source of evidence screening and selection
A health sciences librarian supported the design and refinement of the search strategy. Searches were conducted in the following databases:
• MEDLINE
• EMBASE
• CINAHL
• Scopus
• Web of Science
Searches combined four main concept groups:
1. machine learning / ML methods,
2. EMR/EHR,
3. case definition / case selection / phenotyping,
4. primary care / general practice / family medicine.
A global 'chronic disease' search filter was not used because chronic disease phenotyping studies are typically indexed using disease-specific terms rather than the generic term 'chronic,' and applying a filter would have reduced sensitivity.

The initial search was performed on 15 December 2022, with an updated search on 5 July 2024 to capture more recent studies. The full search strategy is reported in an appendix to the main manuscript. Grey literature and conference abstracts were not included, in order to focus on peer-reviewed research and ensure a minimum threshold of methodological reporting.
No filters were applied for study design or chronic disease type beyond the eligibility criteria outlined above; chronic disease phenotypes were captured via disease-specific and phenotyping-related terms rather than a global "chronic disease" filter.
A PRISMA-style flow diagram will summarize the number of records identified, screened, excluded (with reasons at full-text stage), and included.

Data management
All records and screening decisions were managed using Covidence. Extracted data were stored on secure institutional servers with access restricted to the review team.
Formal risk-of-bias or quality assessment tools were not applied, in line with accepted scoping review methodology that emphasizes mapping the

evidence base rather than appraising study quality. This lack of formal quality assessment is acknowledged as a limitation in the review.

Language restriction Only studies available in English were included. Non-English full texts and non-professionally translated materials were excluded.

Country(ies) involved The review team is based in Canada. Included studies may originate from any country, provided they meet the stated eligibility criteria.

Keywords Primary care; electronic medical records; machine learning; chronic disease; case definition; phenotyping.

Contributions of each author
Author 1 - Anh N.Q. Pham.
Author 2 - Cliff Lindeman.
Author 3 - Micheal Cummings.
Author 4 - Sylvia Aponte-Hao.
Author 5 - Katie Kjelland.