Modifiable risk factors in administrative health data: validity of smoking identification algorithms and patterns of social determinants of health documentation

By

Md Ashiqul Haque

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Department of Community Health Sciences

Rady Faculty of Health Sciences

University of Manitoba

Winnipeg

Copyright © 2025 by Md Ashiqul Haque

ABSTRACT

Introduction: Modifiable risk factors such as smoking, alcohol abuse, and psychosocial circumstances often increase the risk of having chronic health conditions, such as cardiovascular diseases. Administrative health data (AHD), including hospital discharge abstracts, physician billing claims, and drug dispensation records, offer opportunities for evaluating the impact of these factors on disease burden and identifying trends for surveillance. Given that AHD primarily supports healthcare management, validation of modifiable risk factor identification algorithms is important to prevent erroneous findings when repurposing these data for research or surveillance. The purpose of this research was to assess the validity of using AHD to ascertain modifiable risk factors. The objectives were to: 1) synthesize evidence about the validity of electronic health data for ascertaining smoking status, 2) develop and evaluate model-based algorithms (MBAs) to identify smoking status for AHD, and 3) identify the patterns of social determinants of health (SDOH) documentation in AHD to assess their face validity.

Methods: Four related studies were conducted to achieve the research objectives. The first study evaluated the validity of using electronic health records and AHD to capture smoking status in observational studies by undertaking a systematic review and meta-analysis of articles published from 1990 to 2022. The remaining studies followed the retrospective cohort study design using AHD and clinical registry data from the Population Research Data Repository housed at the Manitoba Centre for Health Policy at the University of Manitoba. The second study tested the validity of Least Absolute Shrinkage and Selection Operator and Random Forest machine learning (ML) MBAs to identify current smoking status using AHD. The third study explored the use of synthetic data to improve the validity of diagnostic algorithms for AHD in model-based smoking status identification. This study applied three synthetic data generation techniques-Synthetic Minority Oversampling Technique (SMOTE), Conditional Tabular Generative Adversarial Networks, and Classification and

Regression Trees to augment algorithm training data. The last study examined the patterns of documentation of SDOH, such as homelessness and unemployment related to economic conditions in AHD, to assess their face validity.

Results: The first study identified 57 articles that assessed the validity of 116 algorithms to ascertain smoking status. ML MBAs, particularly those using electronic medical records, performed better than traditional rule-based (i.e., deterministic) algorithms; however, sensitivity was often low. The findings of the second study indicate that ML models can improve the accuracy of the identification of smoking status in healthcare databases. In this study, a comparison of the performance of ML MBAs with their rule-based counterparts found the use of ML models in algorithm construction significantly improved sensitivity, but positive predictive value (PPV) was lower for MBAs. By balancing training datasets, the third study results demonstrate the use of synthetic data to enhance the fairness and accuracy of diagnostic algorithms in healthcare research. Notably, SMOTE improved specificity and PPV, particularly for underrepresented groups in the validation cohort. The fourth study identified variations in SDOH documentation across sociodemographic groups, with temporal shifts following coding policy change and adverse events such as the onset of the COVID-19 pandemic.

Significance: This research evaluates smoking identification algorithms, demonstrates the use of ML models and synthetic data generation techniques to enhance the validation of diagnostic algorithms for AHD, and highlights the variation in SDOH diagnosis. The results support more informed public health monitoring and data-driven policy development. Specifically, the findings are valuable for researchers constructing disease risk prediction models and for agencies such as the Public Health Agency of Canada, which can utilize them to develop surveillance systems.

ACKNOWLEDGEMENTS

The completion of this doctoral thesis would not have been possible without the invaluable support, guidance, and encouragement of numerous individuals to whom I owe my deepest gratitude.

First and foremost, I extend my sincerest appreciation to my advisor and mentor, Dr. Lisa Lix, for her unwavering support, exceptional guidance, and steadfast commitment to academic excellence. Her mentorship has been instrumental in shaping this research and my growth as a scholar.

I am profoundly grateful to my thesis committee members, Dr. Nathan C. Nickel and Dr. Maxime Turgeon, for their insightful feedback, constructive critiques, and invaluable mentorship throughout this journey. Their expertise and encouragement greatly enriched the quality of this work.

I extend my sincere gratitude to the faculty, staff, and fellow students at the George & Fay Yee Centre for Healthcare Innovation for their support and enriching academic environment. Special thanks to: Dr. Jiang for his encouragement; Dr. Romanescu for his mentorship during my VADA internship; Josie for her administrative support; Azizur for his help with my candidacy exam; and Yixiu, Muditha, Joy, Hassan, Rozi, & Nasiba for their camaraderie. I also thank the Manitoba Centre for Health Policy, particularly Charles Burchill, Heather Prior, and Gilles Detillieux, for their support.

This research was supported by the VADA Program and Dr. Lix's generous funding, for which I am deeply grateful.

Finally, I thank my family and friends for their unwavering love and encouragement. To all who contributed—thank you.

DEDICATION

To my brother, Mahfuj Hoque, who encouraged me to enter graduate school.

To my wife, Shaima Afroz, who left behind her dreams in Bangladesh so that I could pursue mine in Canada.

To those who lost their lives in Bangladesh's 2024 monsoon uprising, their sacrifice is a testament to the struggle for a just and equitable society.

TABLE OF CONTENTS

ABSTRACT i

ACKNOWLEDGEMENTS iii

DEDICATION iv

TABLE OF CONTENTS v

LIST OF TABLES x

LIST OF FIGURES xiv

CHAPTER 1: INTRODUCTION 16

1.1 Overview 16

1.2 Purpose and Objectives 19

1.3 Organization of Thesis 20

1.4 References 24

CHAPTER 2: THE VALIDITY OF ELECTRONIC HEALTH DATA FOR MEASURING SMOKING STATUS: A SYSTEMATIC REVIEW AND META-ANALYSIS 16

2.1 Overview 30

2.2 Abstract 31

2.3 Background 33

2.4 Methods 34

2.4.1 Data Sources and Search Strategy 34

2.4.2 Study Selection 35

2.4.3 Data Extraction 36

2.4.5 Statistical Analysis 37

2.4.6 Risk of Bias Assessment 38

2.5 Results 39
2.5.1 Search Results 39
2.5.2 Article Characteristics 40
2.5.3 Characteristics of the Identified Algorithms 42
2.5.4 Validity Measures 43
2.6 Discussion 50
2.6.1 Strengths and Limitations 53
2.6.2 Conclusions 54
2.7 References 56
2.8 SUPPLEMENTARY MATERIAL 72
CHAPTER 3: MODEL-BASED ALGORITHMS TO ASCERTAIN SMOKING IN ADMINISTRATIVE HEALTH DATA: A REGISTRY-BASED VALIDATION STUDY 86
3.1 Overview 86
3.2 Abstract 87
3.3 Background 89
3.4 Methods 90
3.4.1 Data Sources 90
3.4.2 Study Cohort 92
3.4.3 Target Measure 92
3.4.4 Indicators of Smoking Status in Administrative Health Data 92
3.4.5 Rule-based Algorithm (RBA) 93
3.4.6 Statistical Analysis 94
3.4.7 Model-based Algorithm (MBA) 94
3.4.8 Performance Evaluation 96

3.5 Results 96
3.5.1 Cohort Characteristics 97
3.5.2 Overall Performance of Algorithms 99
3.5.3 Performance of Algorithms by Demographic Characteristics 103
3.6 Discussion 104
3.6.1 Performance of RBAs 105
3.6.2 Performance of MBAs 105
3.6.3 Comparison Between RBAs and MBAs 106
3.6.4 Impact of Number of Years of Administrative Data and Sociodemographic Factors 106
3.6.5 Study Implications 106
3.6.6 Strengths and Limitations 107
3.6.7 Future Directions 107
3.7 Conclusions 108
3.8 References 109
3.9 SUPPLEMENTARY MATERIAL 117
CHAPTER 4. LEVERAGING SYNTHETIC DATA TO MITIGATE COHORT IMBALANCE IN DIAGNOSTIC ALGORITHM VALIDATION STUDIES FOR ADMINISTRATIVE HEALTH DATA 130
4.1 Overview 130
4.2 Abstract 131
4.3 Background 133
4.4 Methods 135
4.4.1 Data Sources 135
4.4.2 Population and Validation Cohort 136

4.4.3 Response Variable 136

4.4.4 Indicators of Smoking in AHD 137

4.4.5 Synthetic Data Generation 137

4.4.6 Statistical Analysis 141

4.5 Results 143

4.5.1 Estimates of Model Hyperparameters 144

4.5.2 Estimates of Validity Metrics 144

4.5.3 Comparison of Performance 149

4.6 Discussion 151

4.6.1 Strengths and Limitations 153

4.6.2 Future Directions 154

4.6.3 Study Implications 154

4.6.4 Conclusions 154

4.7 References 156

4.8 SUPPLEMENTARY MATERIAL 167

CHAPTER 5: SOCIAL DETERMINANTS OF HEALTH DOCUMENTATION IN ADMINISTRATIVE HEALTH DATA: A POPULATION-BASED STUDY IN MANITOBA, CANADA 183

5.1 Overview 183

5.2 Abstract 184

5.3 Background 185

5.4 Methods 187

5.4.1 Data Sources 187

5.4.2 Population and Study Cohort 188

5.4.3 Study Outcome 188
5.4.4 Covariates 189
5.4.5 Statistical Analysis 190
5.5 Results 191
5.5.1 Study Cohort Characteristics 191
5.5.2 Association Between Cohort Characteristics and SDOH Documentation 196
5.5.3 Trends in SDOH Documentation 199
5.6 Discussion 202
5.6.1 Face Validity Assessment 203
5.6.2 Strengths and Limitations 205
5.6.3 Future Directions 205
5.6.4 Conclusions 206
5.7 References 207
5.8 SUPPLEMENTARY MATERIAL 213
CHAPTER 6: SUMMARY 183
6.1 Overview 226
6.1 Summary and Interpretation of Research Findings 226
6.1 Thematic Integration and Research Implications 226
6.2 Strengths and Limitations 233
6.3 Conclusions and Future Directions 234
6.4 References 237
Appendix A: Contributions of Authors 241
Appendix B: Ethics Approval and Data Access 242

LIST OF TABLES

Table 2.1 Summary of article characteristics (n=57) 41

Table 2.2 Meta-regression model parameter estimates (SE) for validity measures of PPV, sensitivity, and specificity 48

Table 2.3 Risk of bias (ROB) assessment results 49

Supplemental Table 2.1 Article Characteristics 74

Supplemental Table 2.2 Meta-regression model parameter estimates (SE) for validity measures of PPV, sensitivity, and specificity 82

Supplemental Table 2.3 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist 83

Table 3.1 Characteristics of the study cohort stratified by self-reported current smoking status (N = 24,718) 98

Table 3.2 Random forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression, and rule-based algorithm (RBA) performance for test data (n=7414) by number of years of administrative health data used for identifying independent variables 101

Supplemental Table 3.1 The Reporting of studies Conducted using Observational Routinely-collected Data (RECORD) statement – checklist of items, extended from the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, that should be reported in observational studies using routinely collected health data 117

Supplemental Table 3.2 List of indicators used to develop rule-based and model-based algorithms for current smoking status by data source 123

Supplemental Table 3.3 Area under the curve (AUC) of the receiver operating characteristic

curve estimates (95% confidence intervals [CI]) for model-based algorithms by the number of years of administrative health data used for independent variable identification (%)..... 125
Supplemental Table 3.4 Sex-stratified validity measure estimates (95% confidence intervals) for algorithms by number of years of administrative health data used for independent variable identification 126
Supplemental Table 3.5 Place of residence-stratified validity measure estimates (95% confidence intervals) for algorithms by number of years of administrative health data used for independent variable identification 128
Table 4.1 Validity metric estimates (95% confidence intervals; CI) by demographic characteristics of sex and age group in test dataset (N=7414; Female=6567, Male=847; 18-45 years=293, >45 years=7121) for Least Absolute Shrinkage and Selection Operator logistic regression model based on non-augmented training data 145
Table 4.2 Sex-stratified validity metric estimates (95% confidence intervals; CI) in test dataset (N=7414; Female=6567, Male=847) for Least Absolute Shrinkage and Selection Operator logistic regression model trained with augmented data in 500 synthetic data generation experiments 146
Table 4.3 Age group-stratified validity estimates (95% confidence intervals) in test dataset (N=7414; 18-45 years=293, > 45 years=7121) for Least Absolute Shrinkage and Selection Operator logistic regression models trained with augmented data in 500 synthetic data generation experiments 148
Supplemental Table 4.1 The RECORD statement – checklist of items, extended from the STROBE statement, to report in observational studies using routinely collected health data.... 167
Supplemental Table 4.2 Indicators used to develop rule-based and model-based algorithms for

current smoking status by data source 173
Supplemental Table 4.3 Characteristics of the study cohort stratified by self-reported current smoking status (N = 24,718) 176
Supplemental Table 4.4 Sex-stratified validity measure estimates (95% confidence intervals) in test dataset (N=7414; Female=6567, Male=847) for Least Absolute Shrinkage and Selection Operator logistic regression models trained with augmented data in 500 synthetic data generation experiments 178
Supplemental Table 4.5 Age group-stratified validity measure estimates (95% confidence intervals) in test dataset (N=7414; 18-45 years=293, >45 years=7121) for Least Absolute Shrinkage and Selection Operator logistic regression trained with augmented data in 500 synthetic data generation experiments 180
Table 5.1 Characteristics of the study cohort stratified by social determinants of health (SDOH) documentation and data sources 194
Table 5.2 Odds ratios (ORs) and 95% confidence intervals (CIs) for the association of cohort characteristics with social determinants of health (SDOH) documentation, stratified by data sources 196
Supplemental Table 5.1 The Reporting of studies Conducted using Observational Routinely-collected Data (RECORD) statement – checklist of items, extended from the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, that should be reported in observational studies using routinely collected health data 213
Supplemental Table 5.2 Social determinants of health (SDOH) code documentation in hospital discharge abstracts, 2012/13 to 2022/23 (n=5988) 220
Supplemental Table 5.3 Social determinants of health (SDOH) code documentation in hospital

discharge abstracts by SDOH category (n = 5,988)..... 221

Supplemental Table 5.4 Social determinants of health (SDOH) code documentation in physician billing claims, 2015/16 to 2022/23 (n=12,064) 222

Supplemental Table 5.5 Social determinants of health (SDOH) code in physician billing claims, by SDOH category (n=12,064)..... 224

Table 6.1 Summary of novel contributions to literature..... 226

LIST OF FIGURES

Figure 2.1 Three-level meta-regression model 38

Figure 2.2 Flowchart of the study selection process 40

Figure 2.3 Percent (%) of smoking status algorithms characterized by validity measures (n=116) 43

Figure 2.4 Distribution of selected algorithm validity measures, stratified by algorithm data source 44

Figure 2.5 Distribution of selected algorithm validity measures, stratified by algorithm data structure 45

Figure 2.6 Distribution of selected algorithm validity measures, stratified by use of predictive model in algorithm construction 47

Figure 3.1 Cohort flow diagram 97

Figure 3.2 Comparison of model-based algorithms (independent variables identified based on five years of administrative health data) for area under the receiver operating characteristic curve in test data 99

Figure 3.3 Top ten most important variables for Random Forest (RF) and Least Absolute Shrinkage and Selection Operator (LASSO) models using independent variables identified based on five years of administrative health data 100

Figure 4.1 Marginal and joint distribution (%) for demographic characteristics of sex and age group in the Manitoba population data (N=435,120) and non-augmented training data (N=17,304) 143

Figure 4.2 Boxplot of the absolute distance of fairness ratio from perfection (1.0) by sex and age group 149

Supplemental Figure 4.1 Cohort flow diagram..... 175

Figure 5.1 Study cohort based on administrative health data from Manitoba, Canada (2012-2023) 191

Figure 5.2 Temporal trends in social determinants of health (SDOH) documentation in hospital discharge abstracts 199

Figure 5.3 Temporal trends in social determinants of health (SDOH) documentation in physician billing claims..... 200

Figure 5.4 Segmented regression models for trends in types of social determinants of health (SDOH) documentation in hospital discharge abstracts (any code corresponding to the "Economic" category of codes) 201

Figure 5.5 Segmented regression models for trends in types of social determinants of health (SDOH) documentation in hospital discharge abstracts (any code corresponding to the "Economic" category of codes) 202

CHAPTER 1: INTRODUCTION

Overview

Administrative health data (AHD) are systematically recorded healthcare information generated through routine operations related to the delivery of healthcare services (1). These data are primarily collected for purposes such as reimbursement processes and monitoring regulatory compliance (2). Common forms of AHD include hospital discharge abstracts, physician billing claims, pharmaceutical dispensations, and emergency department records (3). While these data were not originally designed for research purposes, their inherent characteristics, including the availability of information for a long period of time, population-level coverage, and standardized collection protocols, mean that they are an important resource for population health and health services research (4).

The research utility of AHD is particularly pronounced in regions with universal healthcare systems, where these data sources encompass virtually all healthcare encounters across the entire population and over time (5). This comprehensive coverage facilitates both methodologically robust and highly generalizable analysis results while minimizing selection biases that often arise in traditional research designs (6). Most often, the standardized nature of AHD collection enables researchers to construct retrospective and prospective cohorts, examine healthcare utilization patterns, and evaluate health outcomes across diverse demographic groups with remarkable cost-efficiency compared to primary data collection methods (7). These advantages make AHD valuable for studying chronic noncommunicable diseases (NCDs) (8).

NCDs, particularly cardiovascular diseases, diabetes mellitus, various cancer types, and chronic respiratory conditions, collectively represent the leading causes of premature mortality and account for the majority of global healthcare expenditures (9). While certain risk factors such as

advanced age, biological sex, and genetic predisposition are immutable, a substantial proportion of the disease burden comes from modifiable behavioral and environmental factors (10). These include tobacco use, suboptimal dietary patterns, physical inactivity, and excessive alcohol consumption, as well as broader social determinants such as socioeconomic deprivation, environmental pollution, and housing instability. These modifiable risk factors are also associated with mental disorders (11) and can contribute to the spread of infectious diseases (12).

Accurately measuring and monitoring modifiable risk factors at population levels presents considerable methodological challenges. Traditional surveillance methods, including population health surveys and clinical screenings, are frequently constrained by substantial financial costs, participant burden, and potential recall biases (13). Given these constraints, AHD could be a valuable alternative, albeit limited to individuals with documented healthcare interactions only. However, to effectively use AHD for research and to monitor modifiable risk factors requires the development and validation of methods to ascertain health behaviors and risk exposures from these routinely collected records (14). This methodological imperative aligns closely with the Behavioral Epidemiology Framework (BEF), a systematic model proposed by Sallis and colleagues in 2000 for studying health-related behaviors (15). The BEF consists of five phases: (a) establishing links between behaviors and health; (b) developing measures of the behavior; (c) identifying influences on the behavior; (d) evaluating interventions to change the behavior; and (e) translating research into practice. Of particular relevance is phase (b) of this framework, which is about robust measurement methodologies and is related to research about valid methods to identify behavioral risk factors in AHD (16).

Despite significant methodological advancements, existing methods for modifiable risk factor identification in AHD continue to face several limitations. Methods for identifying modifiable

risk factors in AHD can be broadly categorized into two distinct approaches (deterministic and model-based). Deterministic or rule-based algorithms (RBAs) use explicit, predefined classification criteria typically derived from clinical guidelines or expert consensus (17). These algorithms utilize specific combinations of diagnostic codes, procedure records, and medication dispensations occurring within defined temporal windows to classify individuals. For instance, validated algorithms for diabetes identification commonly require either one hospitalization record or two outpatient physician claims containing diabetes-specific diagnostic codes within a 24-month period (18). While these approaches benefit from transparency and straightforward implementation, they may demonstrate limited sensitivity in identifying certain behaviors. To illustrate, when applied to US Medicare claims data, RBAs achieved a sensitivity below 10% for tobacco use detection, leading to substantial underestimations of smoking prevalence and introducing misclassification bias that could affect the results of downstream analyses (19).

In contrast, model-based approaches incorporate statistical and machine learning (ML) techniques to predict the outcome of interest (modifiable risk factors) using predictor variables (20). These methods may integrate demographic characteristics, patterns of healthcare utilization, medication histories, and various proxy indicators to generate probabilistic classifications. Advanced techniques, including random forests, neural networks, and ensemble methods, have shown particular promise in handling the complexity and heterogeneity inherent in AHD (21–23). However, these data-driven approaches often require extensive training datasets, may produce less interpretable results, and have limited transportability compared to their rule-based counterparts.

Algorithm validation methods also present challenges. The assessment of the validity of AHD for ascertaining modifiable risk factors represents a critical methodological consideration, given that these data sources were not originally intended for research (24,25). The gold standard,

criterion validation, involves comparing algorithm-based classifications against an accurate external reference standard, such as clinician-confirmed information, to calculate performance metrics such as sensitivity, specificity, and predictive values (26). However, these reference standards often suffer from representation biases, disproportionately excluding or overrepresenting specific demographic groups (27). As a result, validation study findings may lack generalizability, particularly for vulnerable populations most affected by data-driven health policies (28). The use of synthetic data for underrepresented groups in a validation cohort may be a potential strategy to tackle this challenge. In the absence of a reference standard, complementary methods such as face validation assess whether documentation patterns in AHD plausibly reflect the target constructs such as social determinants (26). A common concern is the limited attention to sociodemographic context in studies examining the documentation of social determinants of health (SDOH) in AHD (29,30).

AHD sources represent a robust and cost-efficient resource for population health research and surveillance of modifiable risk factors. However, the utilization of their full potential requires ongoing methodological innovation in identification algorithm development and validation. By adhering to structured frameworks, such as the BEF, and employing data-driven analytical techniques, the utility of AHD in informing evidence-based public health interventions and healthcare policy decisions can be enhanced.

Purpose and Objectives

The purpose of the research was to evaluate the validity of AHD for identifying modifiable risk factors. This research pursued three objectives:

(1) Conduct a synthesis of published evidence regarding the validity of electronic health data, including AHD and electronic medical records (EMRs), to capture smoking status;

(2) Develop and validate model-based smoking status identification algorithms for AHD using a registry-based cohort; and

(3) Examine the distribution of SDOH documentation in AHD to assess their face validity.

Organization of Thesis

This thesis includes four related manuscripts that address the research objectives. Manuscript 1, "The validity of electronic health data for measuring smoking status: a systematic review and meta-analysis," evaluated algorithms designed to ascertain smoking status from electronic health records and administrative databases. Following the guidelines in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement (31), the study synthesized evidence from multiple validation studies, comparing different methodological approaches for identifying smoking behavior in routine healthcare data. The analysis found that algorithms incorporating ML techniques, particularly those based on EMRs, demonstrated improved performance over traditional rule-based methods. Key findings highlighted substantial variation in algorithm accuracy across studies, with model-based approaches showing relatively high estimates of algorithm validity metrics, such as sensitivity and specificity, compared to their rule-based counterparts. The study identified significant heterogeneity in validity metric estimates, partially explained by data source (EMRs vs. administrative), algorithm type (model-based vs. rule-based), and clinical population focus. Additionally, the review identified geographic bias (studies used data mainly from the US), inconsistent adherence to the Standards for Reporting of Diagnostic Accuracy Studies reporting

standards (32), and reliance on single-jurisdiction validation. Despite these challenges, the results underscored the potential of integrating multi-source data and advanced modeling techniques to improve the ascertainment of smoking status.

Manuscript 2, "Model-based algorithms to ascertain smoking in administrative health data: a registry-based validation study," developed and validated ML MBAs to identify smoking status in AHD, comparing their performance to RBAs. Using a retrospective cohort design, the research integrated hospital records, physician claims, and prescription drug data, validated against a clinical registry with self-reported smoking status. The MBAs were constructed using random forest (33) and least absolute shrinkage and selection operator (LASSO) logistic regression (34) models, incorporating sociodemographic variables and smoking-related indicators, while RBAs used diagnosis and medication codes alone. Findings demonstrated that MBAs significantly improved sensitivity in detecting smokers compared to RBAs, though RBAs achieved higher specificity and positive predictive value (PPV) estimates. Performance varied across demographic groups, with MBAs showing stronger accuracy for males and rural residents. Overall, the study identified key trade-offs: MBAs better captured smokers (higher sensitivity), while RBAs more accurately confirmed non-smokers (higher specificity). Algorithm performance was mostly consistent regardless of the duration of AHD (in years) considered. Methodologically, the study highlighted the value of ML approaches for smoking ascertainment in AHD, particularly when integrating diverse data sources and sociodemographic factors.

Chapter 4 contains Manuscript 3, “Leveraging synthetic data to mitigate cohort imbalance in diagnostic algorithm validation studies for administrative health data”, which pertained to Objective 2. The study aimed to address cohort imbalances in diagnostic algorithm validation studies for AHD by using synthetic data for MBAs training data augmentation. The methodology involved linking

AHD with clinical registry data to validate a smoking status identification algorithm (same as for Manuscript 2), focusing on underrepresented groups (males and younger adults). Three synthetic data generation models, namely, the synthetic minority oversampling technique (SMOTE) (35), conditional tabular generative adversarial networks (CTGAN) (36), and classification and regression trees (CART) (37), were used to augment the training dataset, ensuring the population's representativeness of sex and age distributions. The algorithm was developed using LASSO logistic regression and evaluated on a test dataset (30% of the validation cohort) for sensitivity, specificity, PPV, and NPV. The findings show that synthetic data augmentation improved algorithm performance, with SMOTE enhancing specificity and PPV for underrepresented groups, while CTGAN improved sensitivity. CART showed balanced improvements across metrics. SMOTE also demonstrated greater fairness, with validity estimates more consistent across subgroups. The study highlighted trade-offs between sensitivity and specificity, emphasizing the need for context-specific selection of synthetic data generation techniques. The study results underscored the potential of synthetic data to mitigate biases and enhance the generalizability of AHD validation studies.

Chapter 5, which addresses Objective 3, contains Manuscript 4, "Social determinants of health diagnoses in administrative health data: A population-based study in Manitoba, Canada" which assessed the face validity of AHD records containing SDOH documentation. It examines by examining patterns of SDOH records across geographic regions, data sources, and time. A retrospective cohort design was used. AHD used in the study included hospital discharge abstracts, physician billing claims, and health insurance registrations. Records of SDOH were identified using the International Statistical Classification Of Diseases And Related Health Problems, tenth revision, Canada (ICD-10-CA) codes from the ICD chapter on "Factors influencing health status and contact with health services (38)" and ICD ninth revision, clinical modification (ICD-9-CM) codes from the ICD chapter on "supplementary classification of factors influencing health status and contact with

health services" (39). Logistic regression models tested the association between SDOH code presence/absence and sociodemographic factors. Segmented regression (40,41) was used to analyze temporal trends in SDOH codes captured in AHD. The study revealed low overall SDOH prevalence, which is consistent with previous research. There were variations in SDOH prevalence by income quintile and geographic location of residence. Urban residents had higher odds of SDOH documentation in AHD records, while individuals from lower-income neighborhoods showed an increased likelihood of SDOH documentation, particularly in hospital data. Geographic location of residence was also associated with SDOH documentation. Temporal trends indicated a sharp increase in economic SDOH documentation in hospital data, likely due to updated coding guidelines by the Canadian Institute for Health Information. Documentation of problems related to primary support groups, such as family-related circumstances, declined immediately after the onset of the COVID-19 pandemic. In contrast, the use of psychosocial-related ICD-9-CM codes in physician billing claims, particularly those pertaining to psychological or physical stress, increased during the same period. Although the overall documentation prevalence was low, sociodemographic gradients were observed, suggesting the face validity of SDOH records in AHD and the potential to use these data for future SDOH algorithm development.

Chapter 6 summarizes the key research findings. It also describes the strengths and limitations of the research. This chapter also contains a section about the main conclusions of the research and directions for further research.

References

1. Cadarette SM, Wong L. An introduction to healthcare administrative data. Canadian Journal of Hospital Pharmacy. 2015;68(3):232–7.

2. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Medical Research Methodology. 2022 Nov 5;22(1):287.

3. Al-Azazi S, Singer A, Rabbani R, Lix LM. Combining population-based administrative health records and electronic medical records for disease surveillance. BMC Medical Informatics and Decision Making. 2019 Jul 2;19:120.

4. Li L, Novillo-Ortiz D, Azzopardi-Muscat N, Kostkova P. Digital data sources and their impact on people's health: A systematic review of systematic reviews. Frontiers in Public Health. 2021 May 5;9:645260.

5. Emerson SD, McLinden T, Sereda P, Yonkman AM, Trigg J, Peterson S, et al. Secondary use of routinely collected administrative health data for epidemiologic research: Answering research questions using data collected for a different purpose. International Journal of Population Data Science. 2024;9(1):2407.

6. Lim WM. What Is Quantitative research? An overview and guidelines. Australasian Marketing Journal. 2024 Aug 1;14413582241264622.

7. Feldman K, Johnson RA, Chawla NV. The state of data in healthcare: Path towards standardization. Journal of Healthcare Informatics Research. 2018 May 22;2(3):248–71.

8. Bai J, Cui J, Shi F, Yu C. Global epidemiological patterns in the burden of main non-communicable diseases, 1990–2019: Relationships with socio-demographic index. International Journal of Public Health. 2023 Jan 16;68:1605502.

9. World Health Organization. Noncommunicable diseases [Internet]. Available from: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

10. Ng R, Sutradhar R, Yao Z, Wodchis WP, Rosella LC. Smoking, drinking, diet and physical activity—modifiable lifestyle risk factors and their associations with age to first chronic disease. International Journal of Epidemiology. 2020 Feb;49(1):113–30.

11. Kioumourtzoglou MA. Identifying modifiable risk factors of mental health disorders: The importance of urban environmental exposures. JAMA Psychiatry. 2019 Jun 1;76(6):569–70.

12. Weiss RA, McMichael AJ. Social and environmental risk factors in the emergence of infectious diseases. Nature Medicine. 2004 Dec;10(12 Suppl):S70-76.

13. Dal Grande E, Chittleborough CR, Campostrini S, Taylor AW. Bias of health estimates obtained from chronic disease and risk factor surveillance systems using telephone population surveys in Australia: Results from a representative face-to-face survey in Australia from 2010 to 2013. BMC Medical Research Methodology. 2016 Apr 18;16(1):44.

14. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: A review of methods and applications. Annual Reviews in Public Health. 2016 Mar 18;37(Volume 37, 2016):61–81.

15. Sallis JF, Owen N, Fotheringham MJ. Behavioral epidemiology: A systematic framework to classify phases of research on health promotion and disease prevention. Annals of Behavioral Medicine. 2000 Dec 1;22(4):294–8.

16. Bedson J, Skrip LA, Pedi D, Abramowitz S, Carter S, Jalloh MF, et al. A review and agenda for integrated disease models including social and behavioural factors. Nature Human Behaviour. 2021 Jul;5(7):834–46.

17. Lix LM, Renoux C, Moriello C, Choi KL, Dormuth CR, Fisher A, et al. Validity of diagnoses of SARS-CoV-2 infection in Canadian administrative health data: a multiprovince, population-based cohort study. Canadian Journal of Medical Association. 2023 Sep 1;11(5):E790–8.

18. Lix L, Yogendran M, Mann J. Defining and validating chronic diseases: An administrative data approach [Internet]. Available from: http://mchp-appserv.cpe.umanitoba.ca/viewReference.php?referencePaperID=76208

19. Desai RJ, Solomon DH, Shadick N, Iannaccone C, Kim SC. Identification of smoking using Medicare data-a validation study of claims-based algorithms. Pharmacoepidemiology and Drug Safety. 2016 Apr;25(4):472–5.

20. Hamm NC, Marrie RA, Jiang D, Irani P, Lix L. Building and validating trend-based multiple sclerosis case definitions: A population-based cohort study for Manitoba, Canada. BMJ Open. 2024 Jul 19;14(7):e083141.

21. Badhoutiya A, Verma RP, Shrivastava A, Laxminarayanamma K, Rao ALN, Khan AK. Random forest classification in healthcare decision support for disease diagnosis. In: 2023 International

Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI) [Internet]. p. 1–7. Available from: https://ieeexplore.ieee.org/abstract/document/10489244

22. Paul SG, Saha A, Hasan MdZ, Noori SRH, Moustafa A. A systematic review of graph neural network in healthcare-based applications: Recent advances, trends, and future directions. IEEE Access. 2024;12:15145–70.

23. Rahim MJ, Afroz A, Akinola O. Predictive analytics in healthcare: Big data, better decisions. International Journal of Scientific and Technological Research. 2025 Jan 17;4(1):1–21.

24. Shiff NJ, Jama S, Boden C, Lix LM. Validation of administrative health data for the pediatric population: a scoping review. BMC Health Services Research. 2014 May 22;14:236.

25. Richard L, Hwang SW, Forchuk C, Nisenbaum R, Clemens K, Wiens K, et al. Validation study of health administrative data algorithms to identify individuals experiencing homelessness and estimate population prevalence of homelessness in Ontario, Canada. BMJ Open. 2019 Oct 1; e030221

26. Lim WM. A typology of validity: content, face, convergent, discriminant, nomological and predictive validity. J Trade Sci. 2024 Jul 17;12(3):155–79.

27. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns. 2024 Apr 12;5(4):100946.

28. Kang S. Model validation failure in class imbalance problems. Expert Syst Appl. 2020 May 15;146:113190.

29. Prophetis ED, Beck K, Ridgeway D, Chuang J, Richard L, Durbin A, et al. Trends in hospital coding for people experiencing homelessness in Canada, 2015–2020: a descriptive study. Canadian Medical Association Open Access Journal. 2023 Nov 1;11(6):E1188–96.

30. Agarwal AR, Prichett L, Jain A, Srikumaran U. Assessment of use of ICD-9 and ICD-10 codes for social determinants of health in the US, 2011-2021. JAMA Netw Open. 2023 May 9;6(5):e2312538.

31. PRISMA Statement Organization. PRISMA Endorsers [Internet]. Available from: http://www.prisma-statement.org/Endorsement/PRISMAEndorsers?AspxAutoDetectCookieSupport=1

32. Korevaar DA, Cohen JF, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. Updating standards for reporting diagnostic accuracy: The development of STARD 2015. Research Integrity and Peer Review. 2016 Jun 7;1(1):7.

33. Hastie T, Tibshirani R, Friedman J. Random Forests. In: Hastie T, Tibshirani R, Friedman J, editors. The elements of statistical learning: Data mining, inference, and prediction [Internet]. p. 587–604. (Springer Series in Statistics). Available from: https://doi.org/10.1007/978-0-387-84858-7_15

34. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of Royal Statistical Society Series B Methodology. 1996;58(1):267–88.

35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002 Jun 1;16:321–57.

36. Jia J, Wu P, Zhang K, Zhong J. Imbalanced disk failure data processing method based on CTGAN. In: Huang DS, Jo KH, Jing J, Premaratne P, Bevilacqua V, Hussain A, editors. Intelligent Computing Theories and Application. Cham: Springer International Publishing; 2022. p. 638–49.

37. Sallier K. Toward more user-centric data access solutions: Producing synthetic data of high analytical value by data synthesis. Stat J IAOS. 2020 Jan 1;36(4):1059–66.

38. Version 2022 ICD-10-CA/CCI classifications, Canadian coding standards and related products | CIHI [Internet]. Available from: https://www.cihi.ca/en/version-2022-icd-10-cacci-classifications-canadian-coding-standards-and-related-products

39. ICD-9-CM - Medical Codes [Internet]. Available from: https://www.findacode.com/code-set.php?set=ICD9&i=2432

40. Feder PI. On asymptotic distribution theory in segmented regression problems: Identified case. Annals of Statistics. 1975;3(1):49–83.

41. Tzeng IS, Chien KL, Tu YK, Chen JY, Ng CY, Chien CY, et al. Segmented regression analysis of emergency departments patient visits from Septicemia in Taiwan. Health Policy Technology. 2018 Jun 1;7(2):149–55.

CHAPTER 2: THE VALIDITY OF ELECTRONIC HEALTH DATA FOR MEASURING SMOKING STATUS: A SYSTEMATIC REVIEW AND META-ANALYSIS

2.1 Overview

This chapter presents a systematic review and meta-analysis of 57 studies that evaluate 116 algorithms used to determine smoking status in electronic health data. It examines the characteristics and performance of model-based and rule-based methods across electronic medical records (EMRs) and administrative data. The analysis found considerable variation in sensitivity, specificity, and positive predictive value, with model-based algorithms, especially those using EMR data, achieving higher sensitivity and more consistent performance than rule-based methods. Differences in algorithm design, coding practices, and validation methods reflect the heterogeneity of data sources and populations. Using multilevel meta-regression, the study quantifies how data type, algorithm development methods, and study design influence validity estimates. A notable gap in Canadian and multi-jurisdictional research is observed despite the availability of data sources based on universal healthcare systems across provinces. The findings underscore the importance of rigorous validation, consistent reporting standards, and methodological refinement to support accurate measurement of behavioral risk factors. This chapter provides an evidence base to inform the development of valid and generalizable smoking status ascertainment algorithms for use in population-based health research utilizing routinely collected data.

Publication Details

Haque MA, Gedara MLB, Nickel N, Turgeon M, Lix LM. The validity of electronic health data for measuring smoking status: a systematic review and meta-analysis. BMC Medical Informatics and Decision Making. 2024 Feb 2;24(1):33.

2.2 Abstract

Background: Smoking is a risk factor for many chronic diseases. Multiple smoking status ascertainment algorithms have been developed for population-based electronic health databases such as administrative databases and electronic medical records (EMRs). Evidence syntheses of algorithm validation studies have often focused on chronic diseases rather than risk factors. We conducted a systematic review and meta-analysis of smoking status ascertainment algorithms to describe the characteristics and validity of these algorithms.

Methods: The Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines were followed. We searched articles published from 1990 to 2022 in EMBASE, MEDLINE, Scopus, and Web of Science with key terms such as validity, administrative data, electronic health records, smoking, and tobacco use. The extracted information, including article characteristics, algorithm characteristics, and validity measures, was descriptively analyzed. Sources of heterogeneity in validity measures were estimated using a meta-regression model. Risk of bias (ROB) in the reviewed articles was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool.

Results: The initial search yielded 2086 articles; 57 were selected for review and 116 algorithms were identified. Almost three-quarters (71.6%) of algorithms were based on EMR data. The algorithms were primarily constructed using diagnosis codes for smoking-related conditions, although prescription medication codes for smoking treatments were also adopted. About half of the algorithms were developed using machine-learning models. The pooled estimates of positive predictive value, sensitivity, and specificity were 0.843, 0.672, and 0.918 respectively. Algorithm sensitivity and specificity were highly variable and ranged from 3% to 100% and 36% to 100%, respectively. Model-based algorithms had significantly greater sensitivity (p=0.006) than rule-based algorithms. Algorithms for EMR data had higher sensitivity than algorithms for administrative data (p=0.001). The ROB was low in most of the articles (76.3%) that underwent the assessment.

Conclusions: Multiple algorithms using different data sources and methods have been proposed to ascertain smoking status in electronic health data. Many algorithms had low sensitivity and positive predictive value, but the data source influenced their validity. Algorithms based on machine-learning models for multiple linked data sources have improved validity.

2.3 Background

Electronic health databases, including electronic medical records (EMRs) and administrative data, contain routinely-collected information that is widely used for health research (1,2) even though they were not originally intended for this purpose. EMRs typically include a diverse amount of information about the patient, including medical history, family history, immunization status, laboratory test results, and radiology images (3). Administrative data also include large amounts of information, including insurance enrollment dates, inpatient and outpatient contacts, and vital statistics (4). Data quality is an important consideration when using electronic health databases for research, given that the data are used for secondary purposes.

Smoking is responsible for more than 8 million deaths worldwide each year (5) and is the leading cause of preventable diseases and premature deaths (6). Smoking is a significant risk factor for cancers, cardiovascular diseases, and respiratory diseases. Valid measurement of smoking status contributes to accurate estimates from risk prediction models and other outcome studies about these diseases (7). Valid smoking status measures can also aid in accurately estimating disease trends at the population level.

Population-based surveys are typically used to capture information about smoking status. However, they are expensive to conduct and are not always conducted on a routine basis (8). Routinely collected administrative data often contain indirect information about smoking status, such as diagnosis codes for related diseases (e.g., chronic bronchitis) and substance use disorders. EMRs capture information on smoking status through free-text information about one's health history as well as diagnosis codes. Therefore, many studies have investigated the validity of electronic health databases, including administrative databases and EMRs, for capturing information on smoking status. For example, a study based on Medicare claims data reported smoking status ascertainment algorithms with limited sensitivity but very high specificity (9). In a different study, sensitivity estimates of algorithms

for EMRs within an integrated healthcare system varied widely by years of data used (10). However, the positive predictive value (PPV) consistently remained high.

To date, there have been few, if any studies that have systematically examined validation studies about smoking status in electronic health databases. Summary information about smoking status ascertainment algorithms might be used to develop recommendations about the optimal algorithm(s) to use and opportunities for further research. The latter is particularly timely, given increasing interest in novel approaches to mine new information from electronic health databases using machine-learning methods (11–13).

Given this background, the purpose of our study was to synthesize information about smoking status algorithms developed for electronic health databases. The objectives were to describe smoking status algorithm characteristics, methods to construct the algorithms, and estimates of their validity. A systematic review methodology was used to provide a comprehensive summary of the algorithms to ascertain smoking status (14). Meta-analysis (15) of algorithm validity measure estimates was conducted to assess the potential sources of heterogeneity in them.

2.4 Methods

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline for this review (16) (see Additional/ Supplementary table file 4). This guideline is widely recognised for ensuring rigorous and consistent reporting in systematic reviews and meta-analyses (17).

2.4.1 Data Sources and Search Strategy

EMBASE, MEDLINE, Scopus, and Web of Science were searched from 1990 to November 22, 2022. The target sources were English-language peer-reviewed journal articles. We excluded review articles. The search strategy was developed by the research team in consultation with an experienced university librarian. The search strategy was based on three concepts, each built with specific sets of keywords. The concepts are validity measures (valid*, quality, accuracy, sensitivity, specificity),

electronic health data (electronic medical record*, electronic health record*, administrative health data, administrative data, health care data, administrative billing record*, administrative claims data, claims data, hospital data, hospital discharge data, medicare data, medicaid data), and smoking status (smok*, tobacco). The * at the end of the words valid, record, and smok indicates the use of truncation to capture variant endings. The keywords within each set were connected with OR and the concepts were connected with AND. Article titles, abstracts, and keywords were reviewed to identify potentially relevant articles. The detailed search strategy implemented for each database is available in Additional file 1.

2.4.2 Study Selection

The titles, abstracts, and keywords of the selected articles were uploaded to Rayyan (18) for deduplication and screening for inclusion. An article was included if it reported the results of a validation study for one or more smoking status ascertainment algorithms developed for EMRs or administrative health data (e.g., hospital records, physician claims, prescription drug records). There were no restrictions on geography of the data or population characteristics (e.g., age, sex). To ensure an acceptable level (>80%) (19) of agreement between the two reviewers who conducted the screening, two rounds of abstract and title screening training were undertaken; each training session was conducted on a random sample of 10% of the identified articles. Both reviewers independently screened all the articles. Inter-reviewer agreement was assessed with Cohen's kappa (20). Disagreements on study selection for full-text review were resolved by consensus. Articles were retained when there was uncertainty regarding the eligibility to be included based on title, abstract, and keywords alone. The final decision about the inclusion of an article in this study was made with full text review of the articles identified based on titles, abstracts, and keywords screening. The reference lists of the selected articles were searched for additional articles.

2.4.3 Data Extraction

Two reviewers independently extracted data from two randomly selected articles in a training session to maintain high inter-reviewer reliability (21). Disagreements on data extraction were resolved by consensus and discussion between the reviewers. The remainder of the articles were equally distributed to the two reviewers for data extraction. We extracted information from the selected studies about article characteristics, algorithm characteristics, and algorithm validation estimates.

Article characteristics included year of publication, geographical data source, whether data from multiple jurisdictions were used, and journal discipline. The latter was determined based on subject terms from the United States National Library of Medicine catalog in PubMed.

Algorithm characteristics included International Statistical Classification of Diseases and Related Health Problems (ICD) codes, procedure or intervention codes, data source, data structure, and the use of a predictive model to develop the algorithm. The algorithm data source was categorized as EMR or administrative data. Data structure was classified as structured (e.g., diagnosis codes), unstructured text (e.g., clinical notes), or both structured and unstructured. Algorithms were classified as model-based or deterministic/rule-based on the basis of the method of construction (22). Model-based algorithms implemented predictive statistical and/or machine-learning models (e.g., support vector machine). Rule-based approaches relied on measuring the type and frequency of diagnosis/billing codes in the records of an individual.

Information about the validation data source and validity measures (e.g., sensitivity, specificity) was also extracted from the articles. For model-based algorithms, the validity measure estimates for test data were extracted. The validation source was classified as self-reported data (e.g., survey), chart review data (e.g., patient charts reviewed to extract smoking status based on

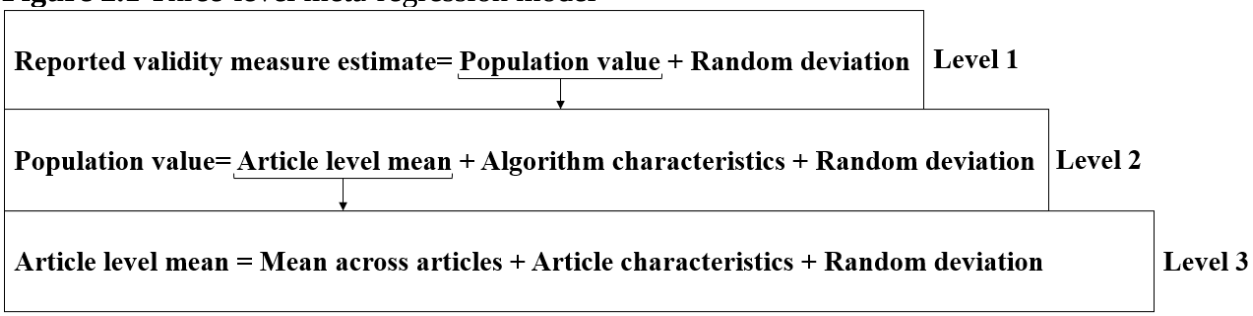
assessments by clinical or domain-knowledge experts), and clinical data (e.g., blood test results). The reported validity measures and their respective estimates and 95% confidence intervals (CI) were recorded. If estimates were reported for more than one sub-group or category (e.g., by demographic characteristics, by years of data), only the overall value of the validity measure was extracted. Finally, we assessed whether the Standards for Reporting Diagnostic Accuracy (STARD) criterion (23) about the number of measures recommended for reporting were fulfilled.

2.4.5 Statistical Analysis

We analyzed the extracted data at both the study level and algorithm level. At the study level, data about the articles were descriptively analyzed using frequencies and percentages. At the algorithm level, we conducted descriptive analyses overall, and then stratified by algorithm characteristics. The distributions of algorithm validity measures were visually summarized using boxplots; the median and interquartile range (IQR) were used to describe the data.

Sources of heterogeneity in algorithm validity estimates, including algorithm characteristics and article characteristics, were examined using a three-level meta-regression model (24). The first level accounts for sampling error, the second level examines algorithm characteristics, and the third level considers article characteristics. The structure of this model is depicted in Figure 2.1. The random deviations at each level were assumed to follow a normal distribution with zero mean and constant variance.

Figure 2.1 Three-level meta-regression model



Standard errors for the reported estimates were calculated from CIs or number of positive/negative cases in the validation data (25). Likelihood ratio (LR) tests were conducted to compare three-level models against their two-level counterparts (26). Null models were fitted to calculate pooled estimate and variance in the reported validity measures. To further investigate the sources of heterogeneity in these estimates, predictor variables were included in the model. At level 2, data source (1=EMRs, 0=Administrative) and the use of a predictive model were included in the model (1=Yes, 0=No) to account for algorithm characteristics. The variables included at level 3 were article characteristics: reference standard (1=Chart review/clinical data, 0=Self-report), clinical population (1= Yes (e.g., HIV patients), 0=No), study population age (1= Restricted to only a specific age-group (e.g., 15-45 years (27)), 0=all ages), and country of data origin (1=US,0=non-US). Data structure was not included in the model since it has a strong association with data source; administrative data are generally structured while EMRs can be either structured or unstructured. Test of residual heterogeneity was conducted to find if the heterogeneity not explained by the models is significant or not (28). To assess model fit, reduction in variance estimates for the models with predictors relative to the initial random-effects pooling models (28) were calculated. An R package metafor (29) was used to conduct the meta-analysis.

2.4.6 Risk of Bias Assessment

The articles included in the meta-analysis underwent a risk of bias (ROB) assessment, utilizing the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool (30). This tool comprises four domains that evaluate patient selection, index test, reference standard, and flow of patients through the study and timing of the index test(s). Each domain includes specific signaling questions to aid in determining the ROB. For each article, the ROB was evaluated as high, low, or unclear for each domain. A low ROB was assigned if all signaling questions within a domain were answered affirmatively (i.e., best practices were followed). Conversely, a high or unclear ROB was assigned if any signaling question received a negative or unclear response. To ensure accuracy, two reviewers independently performed the ROB assessment on a 5% random training sample of eligible articles. Discrepancies between the reviewers were addressed by a third reviewer, to reach consensus. The remaining articles were then evenly distributed between the two reviewers for ROB assessment.

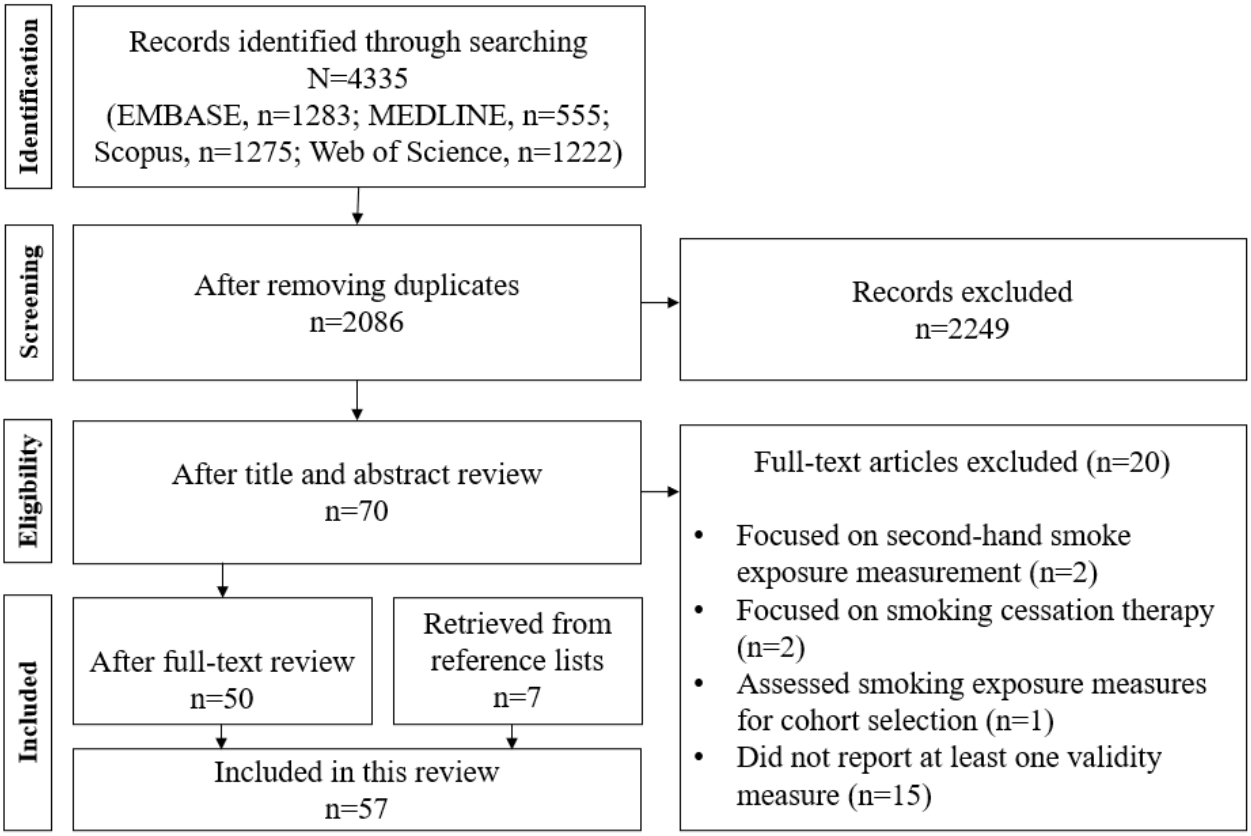
A sensitivity analysis of the meta-regression models was conducted to assess robustness of the synthesized results. The sensitivity analysis excluded the articles with the presence of high/unclear ROB in any of the four domains of QUADAS-2. A publication bias test was conducted by regressing the null model residuals on their corresponding variance estimates (31).

2.5 Results

2.5.1 Search Results

As shown in Figure 2.2, a total of 4335 articles were retrieved from the literature search. After removing duplicates, the titles and abstracts of 2086 studies were screened for study inclusion. The screening process left 70 articles for full-text review. Cohen's kappa for study inclusion/exclusion was 0.97 (95% [CI]: 0.93, 1.00). After full-text review, 20 articles were removed. An additional seven articles were included after the review of reference lists of the remaining 50 articles. Therefore, a total of 57 articles were included in our systematic review (see Additional file 2).

Figure 2.2 Flowchart of the study selection process



2.5.2 Article Characteristics

Only one (1.8%) of the included articles was published before 2000. The majority of the articles (35, 61.4%) were published after 2014 (Table 2.1). In most of the articles (44, 77.2%), algorithms were constructed using US data. A large number (37, 64.9%) of articles used clinical population (e.g., lung cancer patients) data. Very few articles (8, 14.0%) reported the use of data from more than one jurisdiction. Overall, the largest number of articles (24, 42.1%) were published in medical informatics/electronic data journals. The majority of the studies (27, 47.4%) validated algorithms using self-report data, followed by (26, 45.6%) chart review data, and (4, 7.0%) clinical data (e.g., serum cotinine in the blood). Only three (5.3%) articles reported that the study population included exclusively a single biological sex group.

Table 2.1 Summary of article characteristics (n=57)

Characteristic	%	Article Reference
Year of publication		
≥2015	61.4	9,12,27,32–63
<2015	38.6	10,64–84
Geographical location of data source		
US	77.2	9,10,27,32-34,36-40,42-48,50-55,57,59-61,66-73,75-83
Australia	5.3	64,65,74
UK	3.5	49,81
Canada	3.5	41,56
Other	10.5	12,35,46,58,62,84
Clinical population		
Yes	64.9	27,32,34,36,38,39,41-48,52,53,55-58,60–63,66-68,70,72-74,76-78,82,83
No	35.1	10,12,33,35,37,40,49–51,54,59,64,65,69,71,75,79–81,84
Data from multiple jurisdictions		
Yes	14.0	43,44,48,50,68,70,72,73
No	86.0	9,10,12,27,32,33,34-42,45-47,49,51-67,71,74-84
Journal discipline		
Medical informatics/Electronic data	42.1	27,33,41,45,47-55,57,60,66,67,73,74,76,78-80,83
Medicine/Clinical	29.8	34,36,38–40,43,44,46,56,58,59,64,65,70,71,75,77

Public health and epidemiology	10.5	9,12,35,37,42,81
Health services	8.8	62,63,68,72,82
Substance-related disorders	5.3	10,61,69
Biomedical	3.5	32,84

2.5.3 Characteristics of the Identified Algorithms

The 57 articles reported on validity estimates for 116 algorithms. Overall, 50 (43.1%) algorithms used ICD codes; of this number only ten used the 10th revision (e.g., tobacco dependence syndrome, personal history of tobacco use disorder) of this classification system and the remainder used the 9th revision (e.g., tobacco use complicating pregnancy, toxic effect of tobacco). Only 11 (9.5%) algorithms used procedure or intervention codes such as advisement to quit and screening for tobacco use followed by an intervention (i.e., smoking cessation program).

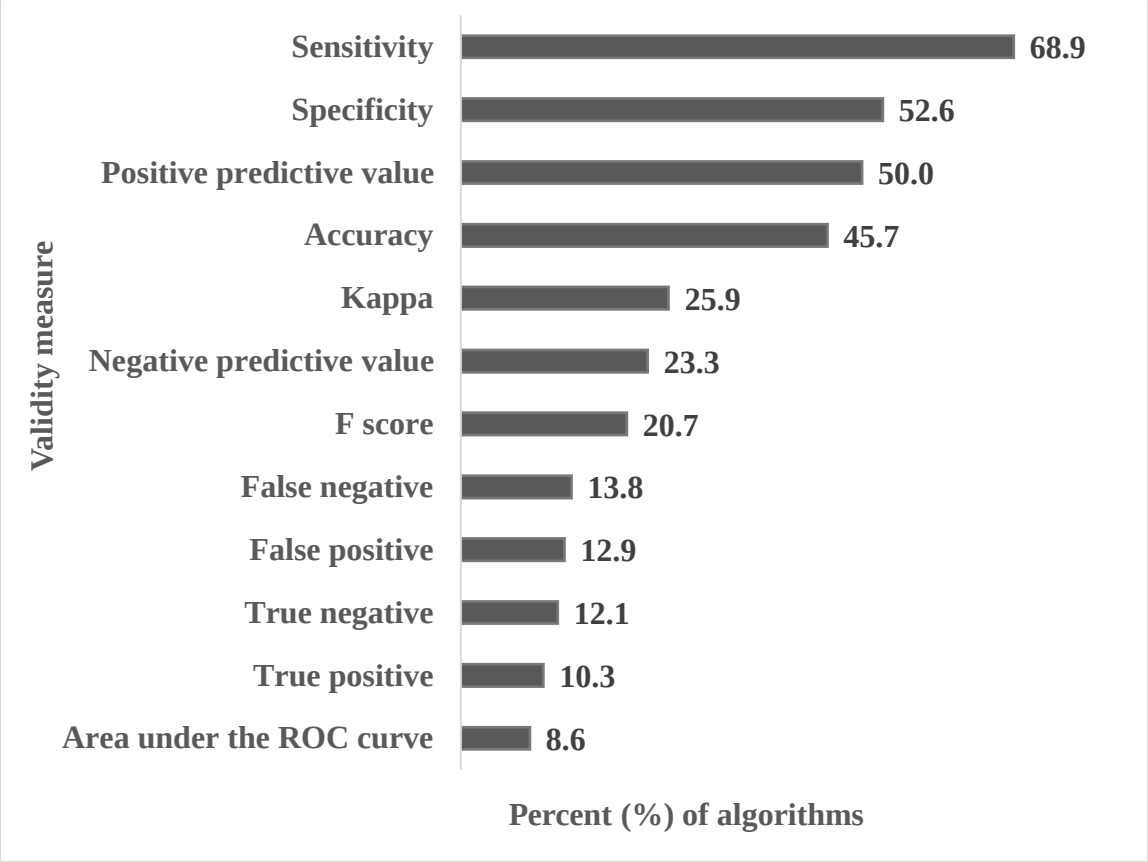
Almost three-quarters of the algorithms (83, 71.6%) were constructed using EMR data; the remaining 33 (28.4%) algorithms were constructed using administrative data. Nearly half of the algorithms (54, 46.5%) used structured data such as diagnosis codes, while unstructured EMR data were used to construct 41 (35.3%) algorithms. Only 21 (18.1%) algorithms were based on both structured and unstructured data.

More than half of the algorithms (61, 52.6%) were developed using rule-based methods, such as the presence of any tobacco-related ICD code or a procedure/intervention code in any data source, or the presence of any smoking-related information in inpatient records and/or outpatient medical claims within a defined period of time. The model-based algorithms (n=55) were almost exclusively (53, 96.4%) developed using EMR data. Largest number of the model-based algorithms (24, 43.6%) were developed using natural language processing methods. Specifically, these algorithms were

developed by extracting smoking-related information from EMRs and constructing features relevant to smoking status (e.g., former smoker, current smoker), frequency of smoking (e.g., number of cigarettes per day), and temporal information relevant to date or duration (e.g., smoked for 10 years). A total of 16 (29.1%) model-based algorithms were developed using support vector machine models. The remainder (15, 27.3%) used statistical or machine-learning models, such as logistic regression, naïve Bayes, Bayesian networks, neural networks, deep learning methods, and decision trees.

2.5.4 Validity Measures

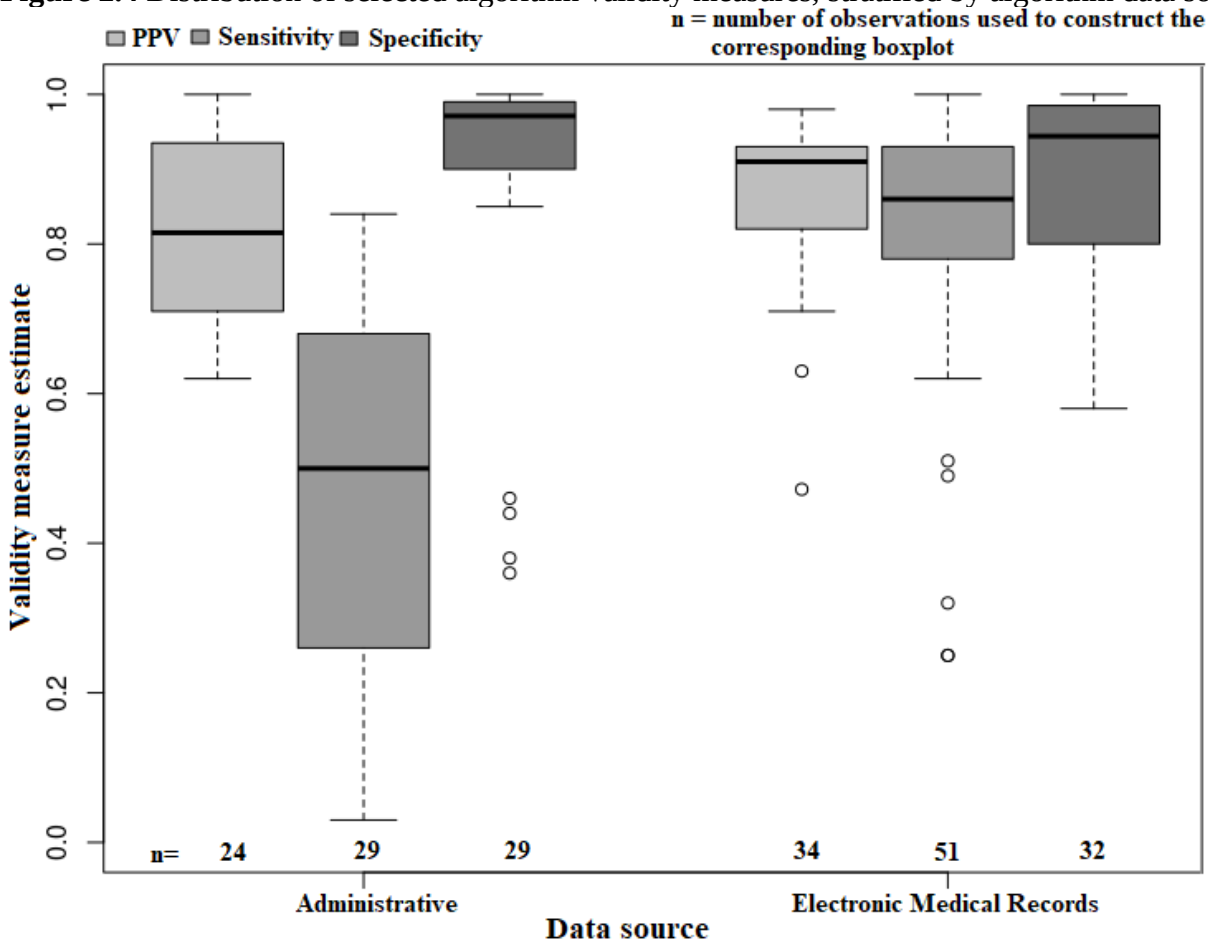
Figure 2.3 Percent (%) of smoking status algorithms characterized by validity measures (n=116)



Algorithm validity measures reported are depicted in Figure 2.3. The number of validation measures reported per algorithm had a median value of 3.0 (IQR=2). The STARD recommendation of four measures was met for slightly less than half (45.7%) of the identified algorithms. The three most

common validity measures were sensitivity (80, 68.9%), specificity (61, 52.6%), and PPV (58, 50.0%). Area under the receiver operating characteristic (ROC) curve (10, 8.6%), true positives (12, 10.3%), and true negatives (14, 12.1%) were the least reported validity measures. The median (IQR) for PPV, sensitivity, and specificity were 88.3% (14.5%), 77.5% (36.5%), and 97.0% (12.0%) respectively.

Figure 2.4 Distribution of selected algorithm validity measures, stratified by algorithm data source

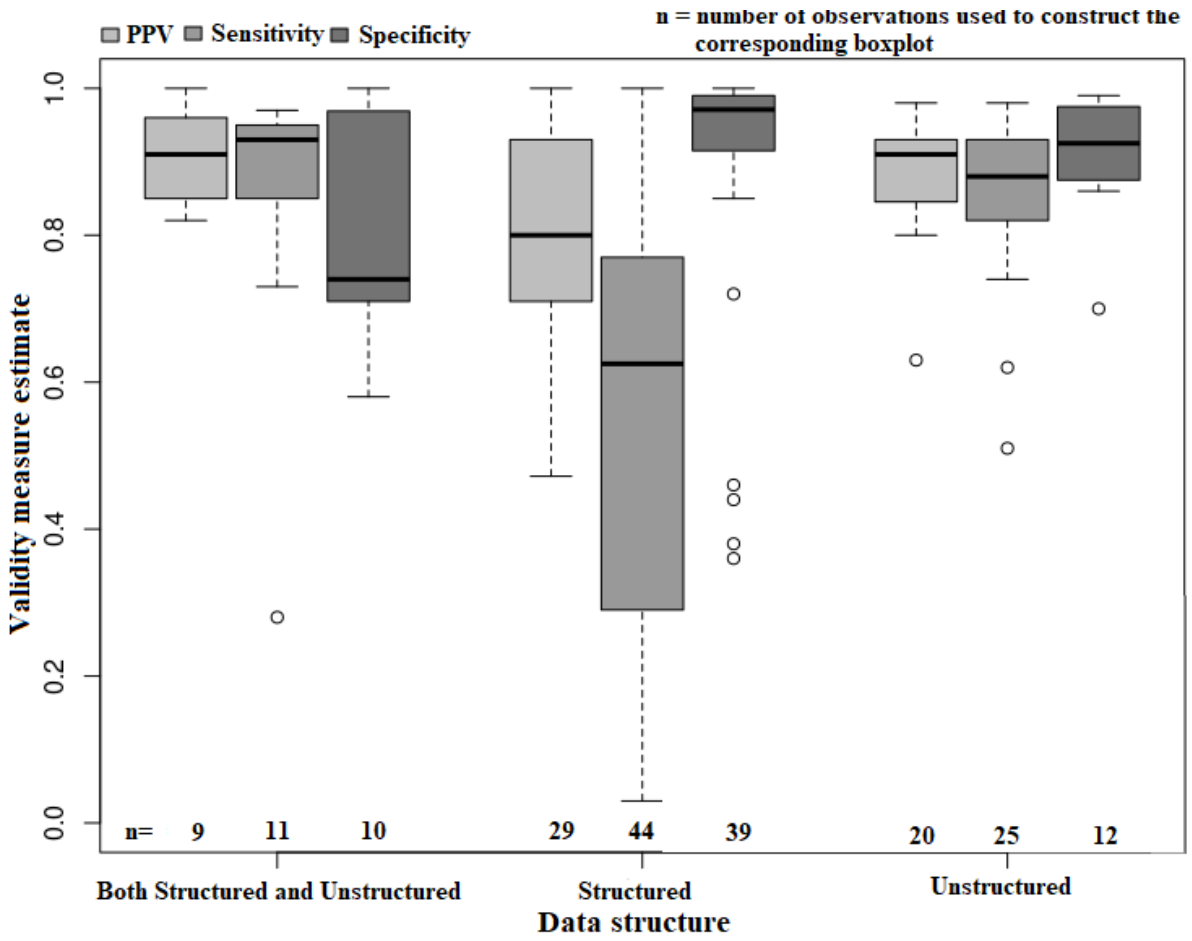


Note: The centre horizontal line within the box represents the median (50th percentile); upper and lower bounds of the box indicate the 25th and 75th percentiles; dashed lines connect the maximum and minimum values; circles represent outliers. PPV=positive predictive value

Figure 2.4 indicates a negatively skewed distribution of specificity estimates for algorithms constructed using administrative data or EMRs, but not for sensitivity and PPV. The median PPV

(91.0%) and sensitivity (86.0%) of the algorithms based on EMR data was higher than for administrative data (PPV=81.5%, sensitivity=50.0%). However, median specificity (94.4%) estimate was lower for EMR data than for administrative data (97.1%). For EMR data, variability in estimates of PPV (IQR=10.5%) and sensitivity (IQR=15.0%) were lower than for administrative data (IQR for PPV=23.0% and sensitivity=42.0%). However, variation in estimates of specificity for EMR data was about twice that of administrative data (IQR 15.2% and 9.0%, respectively).

Figure 2.5 Distribution of selected algorithm validity measures, stratified by algorithm data structure

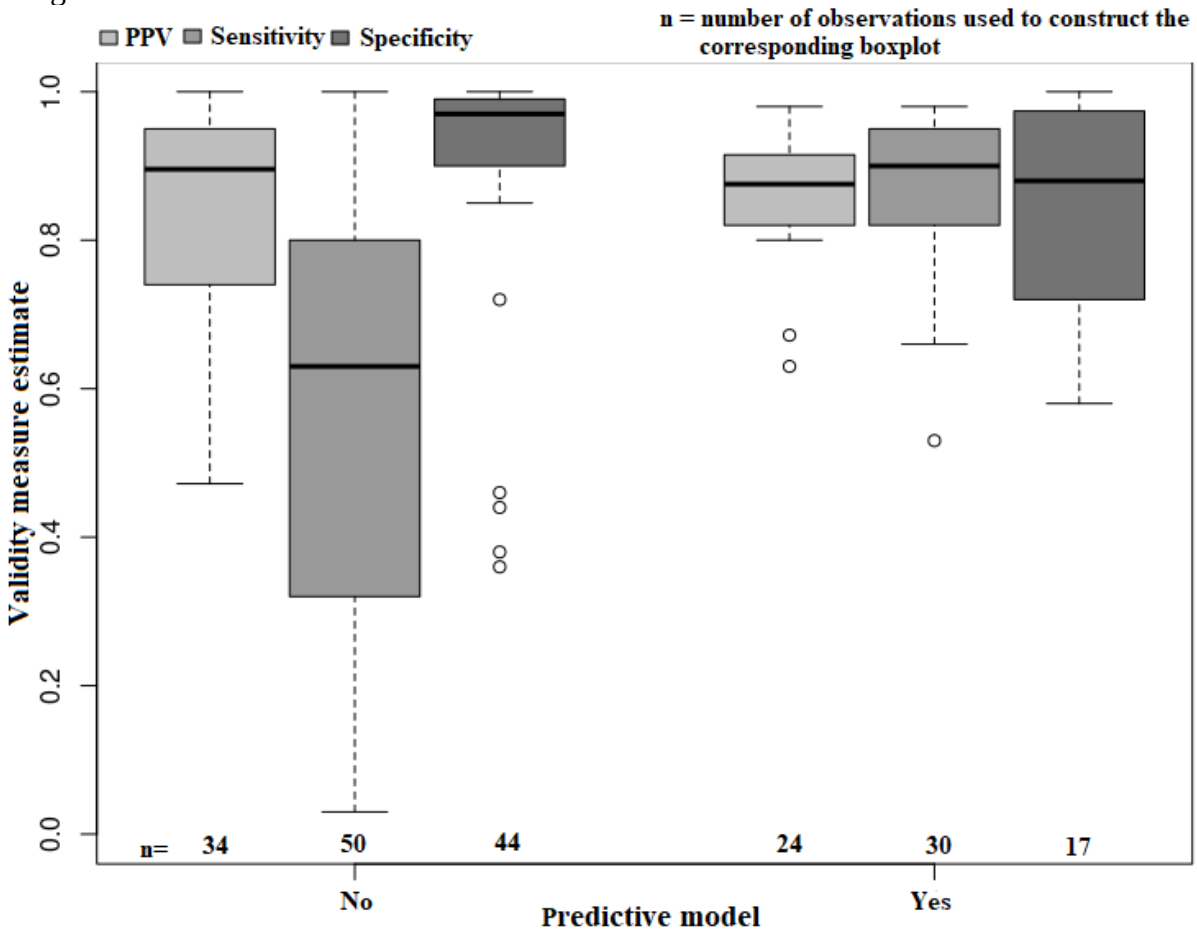


Note: The centre horizontal line within the box represents the median (50th percentile); upper and lower bounds of the box indicate the 25th and 75th percentiles; dashed lines connect the maximum and minimum; circles represent outliers. PPV=positive predictive value

Figure 2.5 shows that the distribution of sensitivity was negatively skewed irrespective of data structure, while this was not the case for PPV and specificity. Median PPV and sensitivity estimates for algorithms based on unstructured data (PPV=91.0%, sensitivity=88.0%, specificity=92.5%) were higher than the estimates for algorithms based on structured data (PPV=80.0%, sensitivity=62.5%, specificity=97.1%). Algorithms developed using both structured and unstructured data had PPV and sensitivity estimates similar to algorithms based on unstructured data alone (PPV=91.0%, sensitivity=93.0%, specificity=74.0%). The estimated validity of algorithms based on unstructured data showed less variation (IQR for PPV=8.2%, sensitivity=11.0%, and specificity=9.5%) than algorithms based on structured data (IQR for PPV=22.0%, sensitivity=46.5%, and specificity=7.5%) or that used a combination of structured and unstructured data (IQR for PPV=11.0%, sensitivity=10.0%, and specificity=23.9%).

Regardless of the approach (model-based or rule-based) used to create an algorithm, the distributions of sensitivity and specificity estimates were highly skewed; most of the observations were below their respective average estimates (Figure 6). Mean sensitivity was 87.3% for model-based algorithms and 57.1% for rule-based algorithms, while mean specificity was 83.9% for model-based algorithms and 90.5% for rule-based algorithms. The median sensitivity for model-based algorithms and rule-based algorithms were 90.0% and 63.0%, respectively. The median PPV and specificity of the algorithms based on predictive models were 87.5% and 88.0% respectively, while algorithms based on deterministic methods had median PPV and specificity of 89.5% and 97.0%, respectively. Variation in PPV (IQR=9.2%) and sensitivity (IQR=12.8%) estimates of model-based algorithms was lower than for rule-based algorithms (PPV IQR = 20.5%, sensitivity IQR=47.5%). Variation in specificity estimates (IQR=25.4%) was higher for model-based algorithms than rule-based algorithms (IQR=9.0%).

Figure 2.6 Distribution of selected algorithm validity measures, stratified by use of predictive model in algorithm construction



Note: The centre horizontal line within the box represents the median (50th percentile); upper and lower bounds of the box indicate the 25th and 75th percentiles; dashed lines connect the maximum and minimum with the box; circles represent outliers. PPV=positive predictive value

Estimates of PPV (47 from 20 articles), sensitivity (69 from 34 articles), and specificity (57 from 28 articles) with necessary information to calculate standard error (e.g., CI) were included in the meta-analysis. Three-level model provided significantly better fits compared to two-level models for all three measures PPV (LR=7.75; p=0.005), sensitivity (LR=10.180; p=0.001), and specificity (LR=7.63; p=0.006). The pooled estimates from the null three-level meta-regression models were 84.0% (95% [CI]: 79.0-90.0), 67.0% (95% [CI]: 59.0-76.0), and 92.0% (95% [CI]:87.0-97.0) for PPV,

sensitivity, and specificity, respectively. The estimated variances of PPV, sensitivity, and specificity were 0.02, 0.07, and 0.03, respectively. The largest portion of estimated variance was attributable to level 2 in case of sensitivity (64.1%) and specificity (79.7%) followed by level 3 (sensitivity=35.8%, specificity=20.3%). However, level 3 (54.8%) accounted for the largest share of estimated variance in PPV followed by level 2 (45.2%). The residual regressions tests (deviation of the intercept from zero) did not indicate the presence of publication bias in the reported estimates of PPV (p=0.753), sensitivity (p=0.471), and specificity (0.124).

Table 2.2 Meta-regression model parameter estimates (SE) for validity measures of PPV, sensitivity, and specificity

Variable	Validity measure					
	PPV		Sensitivity		Specificity	
	Parameter estimate (SE)	P-value	Parameter estimate (SE)	P-value	Parameter estimate (SE)	P-value
Data source						
EMR	-0.076 (0.073)	0.304	0.251 (0.076)	0.001	-0.057 (0.054)	0.297
Administrative	Ref		Ref		Ref	
Predictive model						
Yes	-0.017 (0.076)	0.826	0.195 (0.069)	0.006	-0.094 (0.052)	0.078
No	Ref		Ref		Ref	
Reference standard						
Chart review/clinical data	0.095 (0.072)	0.198	0.086 (0.082)	0.297	-0.025 (0.059)	0.672
Self-report	Ref		Ref		Ref	
Clinical population						
Yes	0.108 (0.063)	0.144	-0.096 (0.081)	0.239	-0.140 (0.058)	0.020
No	Ref		Ref		Ref	
Study population age						
Restricted	0.133(0.067)	0.054	-0.016(0.062)	0.793	0.035(0.045)	0.441
All ages	Ref		Ref		Ref	
Country of data origin						
US	-0.006(0.072)	0.937	-0.083(0.079)	0.304	0.195(0.059)	0.002
non-US	Ref		Ref		Ref	

Note: boldface font denotes a statistically significant estimate; EMR = electronic medical record; SE=standard error; PPV=positive predictive value

Algorithms for EMRs had significantly higher (p=0.001) sensitivity estimates than algorithms for administrative data (Table 2). The model-based algorithms produced significantly higher sensitivity estimates than their rule-based counterparts (p=0.006). The articles including only individuals with certain clinical conditions (e.g., HIV patients) had lower estimates of specificity than articles that used data for general populations (p=0.020). The articles that used data from the US reported significantly larger specificity estimates than the articles based on non-US data (p=0.002). None of the algorithm characteristics and article characteristics included in the model were significantly associated with PPV estimates.

The tests of residual heterogeneity in the models for all three measures suggest statistically significant amounts of unexplained heterogeneity in their estimates (p<0.0001). The models with predictors had lower variances compared to the initial random-effects pooling model by 41.1% and 16.0%, respectively, for sensitivity and specificity. However, the estimated PPV variance remained unchanged in the null model and the model with the predictors. The predictors included in our models were able to explain the variability in sensitivity and specificity better than the respective null models. However, the predictors did not add any value when explaining variation in the PPV estimates.

Table 2.3 Risk of bias (ROB) assessment results

Article Reference	**ROB domains**			
	Patient selection	Index test	Reference standard	Flow and timing
45,32,48,43,51,36,37,68,69,38,64, 72,9,70,73,74,40,63,42,77,78,41, 57,60,33,12,53, 80, 83	⊕	⊕	⊕	⊕
44,55,58,75	⊕	⊕	?	⊕

34,66	?	+	+	+
50	+	−	+	+
54	?	+	+	−
10	+	+	+	−
			Judgment on bias	
			+ Low risk − High risk ? Unclear risk	

The results of the ROB assessment are summarized in Table 3. In total, 38 articles from the meta-regression models were evaluated for ROB using the QUADAS-2 tool. The majority of these articles (29, 76.3%) were deemed to have a low ROB across all four domains. High ROB was observed in three articles, with one in the index test domain and two in the flow and timing domain. Additionally, ROB was unclear in four articles in the reference standard domain and three articles in the patient selection domain.

The PPV, sensitivity, and specificity meta-regression models, respectively, had two articles (containing 4 algorithms), nine articles (containing 14 algorithms), and five articles (containing 5 algorithms) with high/unclear ROB. A sensitivity analysis that removed these articles with potential bias produced results consistent with the primary meta-analysis (Additional file 3).

2.6 Discussion

Smoking status is an important covariate in many disease risk prediction models and trends in smoking status are of interest in epidemiologic studies (85–87). Electronic health databases can be leveraged to develop prediction models (88) and surveillance estimates that include smoking status

information. Validation studies are important to assess the quality of electronic health data to ascertain a variety of individual characteristics, including smoking status (89). Validation studies of electronic health data sources have been synthesized for chronic diseases such as diabetes (90), cancer (91), and social determinants of health (e.g., ethnicity, occupation) (92), but the validity of smoking status algorithms constructed from electronic health databases is a gap in the literature.

We found that a large number of validation studies for smoking status algorithms used electronic health data from the US; a similar trend was reported in systematic reviews for other validation studies, such as for comorbidity indices (93) and kidney disease (94). Canadian provinces and territories collect comprehensive administrative health data (95). Additionally, many provinces collect EMR data through the Canadian Primary Care Sentinel Surveillance Network (96) or other activities (97). However, this study identified only two articles that used Canadian data. Similarly, we found very few articles used data from European countries, including the Scandinavian countries, which also have comprehensive administrative and EMR data (98,99). The vast majority of the algorithms were constructed with data from a single jurisdiction, potentially limiting the transportability of these algorithms across jurisdictions. At the same time, this finding is not unexpected, given that it can be challenging to find a comparable validation data source in more than one jurisdiction. A large number of studies used medical chart review to validate the algorithms, a result comparable to a systematic review of algorithms for obesity (100).

The median number of validity measures reported per algorithm was below the STARD recommendation of at least four measures. Overall, the sensitivity estimates were lower than the estimates of PPV and specificity. This finding is in line with a study that reviewed validation studies of algorithms to identify obesity in administrative databases (100). Similar to other reviews focusing on chronic conditions (101), infections (102), and nonmedical opioid use (103); considerable variation

was found in the estimates for the reported algorithm validity measures such as PPV, sensitivity, and specificity overall and by selected algorithm characteristics. The model-based algorithms tended to have less variable PPV and sensitivity estimates than the rule-based algorithms. This suggests the model-based algorithms had more consistent performance in terms of accurately predicting positive cases (i.e., PPV) and capturing true positives (i.e., sensitivity). However, the model-based algorithms were more variable in specificity estimates compared to the rule-based algorithms. This may suggest the complexity of model-based algorithms affected specificity in different study-specific scenarios. Model-based smoking status ascertainment algorithms had better performance than rule-based algorithms, which is in contrast to findings from a rheumatoid arthritis validation study that utilised these two approaches and found similar predictive performance (104). Nevertheless, another study demonstrated that model-based algorithms can improve sensitivity and specificity estimates over the estimates from rule-based algorithms for classifying carotid endarterectomy indication using physician claims and hospital discharge data (105). Residual confounding due to data source may remain in the summarized relationship between validity measures and the use of a predictive model. This is particularly noteworthy, because a majority of model-based algorithms were constructed using EMR data. These data capture more detailed and comprehensive clinical information compared to administrative data, resulting in better performance than the rule-based algorithms that relied on administrative data. The model for sensitivity suggested significant difference in algorithms for EMRs and administrative data, in contrast to findings from an earlier study to predict binary outcomes such as one year mortality and hospital readmission (106). The specificity model indicated considerably lower estimates for algorithms belonging to articles focusing only on clinical population and higher estimates for articles using US data. Class imbalance may be partially responsible for these findings on specificities (107). No significant difference was detected in the results of PPV model. Overall, the

fitted models with predictors did not adequately explain heterogeneity in validity measure estimates. This finding may be attributed to factors identified as sources of heterogeneity, such as the use of alternative coding methods and misclassified diagnosis by examining medications prescribed for different purposes, in the studies included in our meta-analysis (108). Many articles in our meta-analysis had low ROB, but a review of algorithms for neurodevelopmental disorders had contrasting results (109). The key findings from the meta-analysis did not change after a sensitivity analysis that excluded articles considering the ROB. A similar result was observed in a meta-analysis of articles using machine-learning to predict the spread of breast cancer to armpit lymph nodes (110). Our analysis did not find strong evidence of publication bias in the pooled estimates of validity measures from null models. This finding should be interpreted with caution considering the limitations of linear regression (111). For example, the relationship between residuals and variances may be non-linear, which can lead to inaccurate assessment of publication bias.

2.6.1 Strengths and Limitations

The strengths of this systematic review and meta-analysis include the breadth of citation databases that we searched, the wide variety of article characteristics, and the detailed analysis of extracted information at the algorithm level. We identified statistically significant sources of variation in the estimates of sensitivity (data source and use of predictive model) and specificity (patient characteristics and country of data origin). However, we recognize that this study is not without limitations. Articles published in languages other than English and grey literature, such as government reports and graduate dissertations, were excluded. These exclusions may affect the generalizability of our findings. To mitigate this gap, the reference lists of the included articles were searched for additional articles. A large portion of the variation in the reported estimates of PPV, sensitivity, and specificity remained unexplained in the meta-regression models. To evaluate the reliability of the

results of these models, a sensitivity analysis was performed incorporating the findings from the ROB assessment.

2.6.2 Conclusions

Evidence syntheses of algorithm validation studies have often focused on chronic or infectious disease case ascertainment and social determinants of health (90–92). This study contributes to the body of literature about validation studies and examines a relatively unexplored area of behavioral risk factor algorithms for electronic health databases.

We found that numerous algorithms have been developed to identify smoking status in electronic health databases. The identified algorithms vary in terms of data source, data structure, and methods of construction. In general, the algorithms had high specificity and low sensitivity when predicting smoking status, although there is evidence that sensitivity can be improved by using EMR data and predictive models to construct the algorithms.

A number of opportunities exist to develop algorithms to measure smoking status using population-based electronic health data. For example, combining multiple data sources, including EMR and administrative data, may produce algorithms with high sensitivity (112). The breadth of the longitudinal information (22) available in the electronic health databases can be utilized to develop algorithms. Methods such as longitudinal discriminant analysis (113) and semiparametric mixed-effects model (114) can be used to construct algorithms based on longitudinal data. The application of ensemble machine learning classification models and use of large language models (LLMs) remained unexplored in this line of research. Ensemble machine learning involves combining individual models to improve overall predictive performance. For example, random forest ensemble classifiers (115,116) may be used to identify smoking status in electronic health databases. These classifiers may have

reduced potential over-fitting of the model and improve performance measures relative to decision trees (117). LLMs are trained deep-learning models that understands and generates text in a human-like fashion (118). These models can be deployed to identify smoking status from text-based unstructured EMR data.

2.7 References

1. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M, Michel A. Electronic health records to facilitate clinical research. Clinical Research in Cardiology. 2017;106:1-9.
2. Lee S, Xu Y, D'Souza AG, Martin EA, Doktorchik C, Zhang Z, Quan H. Unlocking the potential of electronic health records for health research. International Journal of Population Data Science. 2020;5(1).
3. Kierkegaard P. Electronic health record: Wiring Europe's healthcare. Computer Law & Security Review. 2011;27(5):503-15.
4. Harbaugh CM, Cooper JN. Administrative databases. In: Seminars in Pediatric Surgery. 2018 (Vol. 27, No. 6, pp. 353-360).
5. World Health Organization. Tobacco fact sheet from WHO providing key facts and information on surveillance. https://www.who.int/news-room/fact-sheets/detail/tobacco. Accessed 10 Apr 2022.
6. Canadian Lung Association. Smoking and tobacco statistics. https://www.lung.ca/lung-health/lung-info/lung-statistics/smoking-and-tobacco-statistics. Accessed 10 Apr 2022.
7. Barrett JK, Sweeting MJ, Wood AM. Dynamic risk prediction for cardiovascular disease: an illustration using the ARIC study. In: Handbook of Statistics. 2017 (Vol. 36, pp. 47-65).
8. Kelsey JL, Kelsey C, Whittemore AS, Whittemore P, Evans AS, Thompson WD, et al. Methods in Observational Epidemiology. Oxford University Press; 1996. p.458.
9. Desai RJ, Solomon DH, Shadick N, Iannaccone C, Kim SC. Identification of smoking using Medicare data—a validation study of claims-based algorithms. Pharmacoepidemiology and Drug Safety. 2016;25(4):472-5.

10. Chen LH, Quinn V, Xu L, Gould MK, Jacobsen SJ, Koebnick C, Reynolds K, Hechter RC, Chao CR. The accuracy and trends of smoking history documentation in electronic medical records in a large managed care organization. Substance Use & Misuse. 2013;48(9):731-42.
11. Chowdhury M, Cervantes EG, Chan WY, Seitz DP. Use of machine learning and artificial intelligence methods in geriatric mental health research involving electronic health record or administrative claims data: A Systematic Review. Frontiers in Psychiatry. 2021;12:738466.
12. Groenhof TK, Koers LR, Blasse E, de Groot M, Grobbee DE, Bots ML, Asselbergs FW, Lely AT, Haitjema S, van Solinge W, Hoefer I. Data mining information from electronic health records produced high yield and accuracy for current smoking status. Journal of Clinical Epidemiology. 2020;118:100-6.
13. Yadav P, Steinbach M, Kumar V, Simon G. Mining electronic health records (EHRs): A survey. ACM Computing Surveys. 2018;50(6):1-40.
14. Caldwell PH, Bennett T. Easy guide to conducting a systematic review. Journal of Paediatrics and Child Health. 2020;56(6):853-6.
15. Deeks JJ, Higgins JP, Altman DG, Cochrane Statistical Methods Group. Analysing data and undertaking meta-analyses. In: Cochrane Handbook for Systematic Reviews of Interventions. John Wiley & Sons, Ltd; 2019. p. 241–84.
16. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. The BMJ. 2015;349.
17. PRISMA Statement Organization. PRISMA Endorsers. http://www.prismastatement.org/Endorsement/PRISMAEndorsers?AspxAutoDetectCookieSupport=1. Accessed 16 May 2023

18. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews. 2016; 5:1-10.

19. Belur J, Tompson L, Thornton A, Simon M. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. Sociological Methods & Research. 2021;50(2):837-65.

20. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22(3):276-82.

21. Lange RT. Inter-rater Reliability. In: Kreutzer JS, DeLuca J, Caplan B, editors. Encyclopedia of Clinical Neuropsychology. New York, NY: Springer; 2011. p. 1348.

22. Feely A, Lim LS, Jiang D, Lix LM. A population-based study to develop juvenile arthritis case definitions for administrative health data using model-based dynamic classification. BMC Medical Research Methodology. 2021;21(1):1-3.

23. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, De Vet HC, Kressel HY. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clinical Chemistry. 2015;61(12):1446-52.

24. Weisz JR, Kuppens S, Ng MY, Eckshtain D, Ugueto AM, Vaughn-Coaxum R, Jensen-Doss A, Hawley KM, Krumholz Marchette LS, Chu BC, Weersing VR. What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. American Psychologist. 2017;72(2):79.

25. Wallis S. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. Journal of Quantitative Linguistics. 2013;20(3):178-208.

26. Glover S, Dixon P. Likelihood ratios: A simple and flexible statistic for empirical psychologists. Psychonomic Bulletin & Review. 2004;11(5):791-806.

27. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, Amin S, Liu H. A clinical text classification paradigm using weak supervision and deep representation. BMC Medical Informatics and Decision Making. 2019;19:1-3.

28. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing meta-analysis with R: A hands-on guide. CRC Press; 2021.

29. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. 2010;36(3):1-48.

30. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine. 2011;155(8):529-36.

31. Doleman B, Freeman SC, Lund JN, Williams JP, Sutton AJ. Funnel plots may show asymmetry in the absence of publication bias with continuous outcomes dependent on baseline risk: presentation of a new publication bias test. Research Synthesis Methods. 2020;11(4):522-34.

32. Wang L, Ruan X, Yang P, Liu H. Comparison of three information sources for smoking information in electronic health records. Cancer Informatics. 2016;15: CIN-S40604.

33. Harris DR, Henderson DW, Corbeau A. Improving the utility of tobacco-related problem list entries using natural language processing. In: American Medical Informatics Association Annual Symposium Proceedings (Vol. 2020, p. 534).

34. Regan S, Meigs JB, Grinspoon SK, Triant VA. Determinants of smoking and quitting in HIV-infected individuals. PLOS One. 2016;11(4):e0153103.

35. Chung WS, Kung PT, Chang HY, Tsai WC. Demographics and medical disorders associated with smoking: a population-based study. BMC Public Health. 2020;20:1-8.

36. Melzer AC, Pinsker EA, Clothier B, Noorbaloochi S, Burgess DJ, Danan ER, Fu SS. Validating the use of veterans affairs tobacco health factors for assessing change in smoking status: accuracy, availability, and approach. BMC Medical Research Methodology. 2018;18:1-10.

37. Huo J, Yang M, Shih YC. Sensitivity of claims-based algorithms to ascertain smoking status more than doubled with meaningful use. Value in Health. 2018;21(3):334-40.

38. Luck J, Larson AE, Tong VT, Yoon J, Oakley LP, Harvey SM. Tobacco use by pregnant Medicaid beneficiaries: Validating a claims-based measure in Oregon. Preventive Medicine Reports. 2020;19:101039.

39. Etzioni DA, Lessow C, Bordeianou LG, Kunitake H, Deery SE, Carchman E, Papageorge CM, Fuhrman G, Seiler RL, Ogilvie J, Habermann EB. Concordance between registry and administrative data in the determination of comorbidity: a multi-institutional study. Annals of Surgery. 2020;272(6):1006-11.

40. McVeigh KH, Lurie-Moroni E, Chan PY, Newton-Dame R, Schreibstein L, Tatem KS, Romo ML, Thorpe LE, Perlman SE. Generalizability of indicators from the New York city macroscope electronic health record surveillance system to systems based on other EHR platforms. Journal of Electronic Health Data and Methods. 2017;5(1).

41. Marrie RA, Tan Q, Ekuma O, Marriott JJ. Development of an indicator of smoking status for people with multiple sclerosis in administrative data. Multiple Sclerosis Journal–Experimental, Translational and Clinical. 2022;8(1).

42. Floyd JS, Blondon M, Moore KP, Boyko EJ, Smith NL. Validation of methods for assessing cardiovascular disease using electronic health data in a cohort of Veterans with diabetes. Pharmacoepidemiology and Drug Safety. 2016;25(4):467-71.

43. Calhoun PS, Wilson SM, Hertzberg JS, Kirby AC, McDonald SD, Dennis PA, Bastian LA, Dedert EA, VA Mid-Atlantic MIRECC Workgroup, Beckham JC. Validation of Veterans Affairs electronic medical record smoking data among Iraq- and Afghanistan-era veterans. Journal of General Internal Medicine. 2017;32:1228-34.

44. Mu Y, Chin AI, Kshirsagar AV, Bang H. Data concordance between ESRD medical evidence report and Medicare claims: is there any improvement?. PeerJ. 2018;6:e5284.

45. LeLaurin JH, Gurka MJ, Chi X, Lee JH, Hall J, Warren GW, Salloum RG. Concordance between electronic health record and tumor registry documentation of smoking status among patients with cancer. JCO Clinical Cancer Informatics. 2021;5:518-26.

46. Caccamisi A, Jørgensen L, Dalianis H, Rosenlund M. Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records. Upsala Journal of Medical Sciences. 2020;125(4):316-24.

47. Palmer EL, Higgins J, Hassanpour S, Sargent J, Robinson CM, Doherty JA, Onega T. Assessing data availability and quality within an electronic health record system through external validation against an external clinical data source. BMC Medical Informatics and Decision Making. 2019;19(1):1-9.

48. Golden SE, Hooker ER, Shull S, Howard M, Crothers K, Thompson RF, Slatore CG. Validity of Veterans Health Administration structured data to determine accurate smoking status. Health Informatics Journal. 2020;26(3):1507-15.

49. Atkinson MD, Kennedy JI, John A, Lewis KE, Lyons RA, Brophy ST. Development of an algorithm for determining smoking status and behaviour over the life course

from UK electronic primary care records. BMC Medical Informatics and Decision Making. 2017;17(1):1-2.

50. Reps JM, Rijnbeek PR, Ryan PB. Supplementing claims data analysis using self-reported data to develop a probabilistic phenotype model for current smoking status. Journal of Biomedical Informatics. 2019;97:103264.

51. Ni Y, Bachtel A, Nause K, Beal S. Automated detection of substance use information from electronic health records for a pediatric population. Journal of the American Medical Informatics Association. 2021;28(10):2116-27.

52. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. Journal of Biomedical Informatics. 2015;58:S128-32.

53. Urbain J. Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models. Journal of Biomedical Informatics. 2015;58:S143-9.

54. McVeigh KH, Newton-Dame R, Chan PY, Thorpe LE, Schreibstein L, Tatem KS, Chernov C, Lurie-Moroni E, Perlman SE. Can electronic health records be used for population health surveillance? Validating population health metrics against established survey data. Journal of Electronic Health Data and Methods. 2016;4(1).

55. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. Journal of Biomedical Informatics. 2015;58:S111-9.

56. Gauthier MP, Law JH, Le LW, Li JJ, Zahir S, Nirmalakumar S, Sung M, Pettengell C, Aviv S, Chu R, Sacher A. Automating access to real-world evidence. JTO Clinical and Research Reports. 2022;3(6):100340.

57. O'Brien EC, Mulder H, Jones WS, Hammill BG, Sharlow A, Hernandez AF, Curtis LH. Concordance Between Patient-Reported Health Data and Electronic Health Data in the ADAPTABLE Trial. JAMA Cardiology. 2022;7(12):1235-43.

58. Alhaug OK, Kaur S, Dolatowski F, Småstuen MC, Solberg TK, Lønne G. Accuracy and agreement of national spine register data for 474 patients compared to corresponding electronic patient records. European Spine Journal. 2022;31(3):801-11.

59. Teng A, Wilcox A. Simplified data science approach to extract social and behavioural determinants: A retrospective chart review. BMJ Open. 2022;12(1).

60. McGinnis KA, Skanderson M, Justice AC, Tindle HA, Akgün KM, Wrona A, Freiberg MS, Goetz MB, Rodriguez-Barradas MC, Brown ST, Crothers KA. Using the biomarker cotinine and survey self-report to validate smoking data from United States Veterans Health Administration electronic health records. JAMIA Open. 2022;5(2).

61. McGinnis KA, Justice AC, Tate JP, Kranzler HR, Tindle HA, Becker WC, Concato J, Gelernter J, Li B, Zhang X, Zhao H. Using DNA methylation to validate an electronic medical record phenotype for smoking. Addiction Biology. 2019;24(5):1056-65.

62. Maier B, Wagner K, Behrens S, Bruch L, Busse R, Schmidt D, Schühlen H, Thieme R, Theres H. Comparing routine administrative data with registry data for assessing quality of hospital care in patients with myocardial infarction using deterministic record linkage. BMC Health Services Research. 2016;16(1):1-9.

63. Nickel KB, Wallace AE, Warren DK, Ball KE, Mines D, Fraser VJ, Olsen MA. Modification of claims-based measures improves identification of comorbidities in non-elderly women undergoing mastectomy for breast cancer: a retrospective cohort study. BMC Health Services Research. 2016;16:1-2.

64. Havard A, Jorm LR, Lujic S. Risk adjustment for smoking identified through tobacco use diagnoses in hospital data: a validation study. PLOS One. 2014;9(4).

65. Lujic S, Watson DE, Randall DA, Simpson JM, Jorm LR. Variation in the recording of common health conditions in routine hospital data: study using linked survey and administrative data in New South Wales, Australia. BMJ Open. 2014;4(9).

66. Wiley LK, Shah A, Xu H, Bush WS. ICD-9 tobacco use codes are effective identifiers of smoking status. Journal of the American Medical Informatics Association. 2013;20(4):652-8.

67. McGinnis KA, Brandt CA, Skanderson M, Justice AC, Shahrir S, Butt AA, Brown ST, Freiberg MS, Gibert CL, Goetz MB, Kim JW. Validating smoking data from the Veteran's Affairs Health Factors dataset, an electronic data source. Nicotine & Tobacco Research. 2011;13(12):1233-9.

68. Kim HM, Smith EG, Stano CM, Ganoczy D, Zivin K, Walters H, Valenstein M. Validation of key behaviourally based mental health diagnoses in administrative data: suicide attempt, alcohol abuse, illicit drug abuse and tobacco use. BMC Health Services Research. 2012;12(1):1-9.

69. Lee JD, Delbanco B, Wu E, Gourevitch MN. Substance use prevalence and screening instrument comparisons in urban primary care. Substance Abuse. 2011;32(3):128-34.

70. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. Annals of Internal Medicine. 1993;119(8):844-50.

71. Steffen MW, Murad MH, Hays JT, Newcomb RD, Molella RG, Cha SS, Hagen PT. Self-report of tobacco use status: comparison of paper-based questionnaire, online

questionnaire, and direct face-to-face interview—implications for meaningful use. Population Health Management. 2014;17(3):185-9.

72. Borzecki AM, Wong AT, Hickey EC, Ash AS, Berlowitz DR. Identifying hypertension-related comorbidities from administrative data: what's the optimal approach?. American Journal of Medical Quality. 2004;19(5):201-6.

73. Bui DD, Zeng-Treitler Q. Learning regular expressions for clinical text classification. Journal of the American Medical Informatics Association. 2014;21(5):850-7.

74. Khor R, Yip WK, Bressel M, Rose W, Duchesne G, Foroudi F. Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements. Journal of the American Medical Informatics Association. 2014;21(1):27-30.

75. DeJoy S, Pekow P, Bertone-Johnson E, Chasan-Taber L. Validation of a Certified Nurse-Midwifery Database for Use in Quality Monitoring and Outcomes Research. Journal of Midwifery & Women's Health. 2014;59(4):438-46.

76. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Medical Informatics and Decision Making. 2006;6(1):1-9.

77. Longenecker JC, Coresh J, Klag MJ, Levey AS, Martin AA, Fink NE, Powe NR. Validation of comorbid conditions on the end-stage renal disease medical evidence report: the CHOICE study. Journal of the American Society of Nephrology. 2000;11(3):520-9.

78. Meystre SM, Deshmukh VG, Mitchell J. A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations. In: American Medical Informatics Association Annual Symposium Proceedings (Vol. 2009, p. 442).

79. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. Journal of the American Medical Informatics Association. 2008;15(1):36-9.

80. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. Journal of the American Medical Informatics Association. 2008;15(1):25-8.

81. Mant J, Murphy M, Rose P, Vessey M. The accuracy of general practitioner records of smoking and alcohol use: comparison with patient questionnaires. Journal of Public Health. 2000;22(2):198-201.

82. Yeager DS, Krosnick JA. The validity of self-reported nicotine product use in the 2001–2008 National Health and Nutrition Examination Survey. Medical Care. 2010;1128-32.

83. Liu M, Shah A, Jiang M, Peterson NB, Dai Q, Aldrich MC, Chen Q, Bowton EA, Liu H, Denny JC, Xu H. A study of transportability of an existing smoking status detection module across institutions. In: American Medical Informatics Association Annual Symposium Proceedings (Vol. 2012, p. 577).

84. Figueroa RL, Soto DA, Pino EJ. Identifying and extracting patient smoking status information from clinical narrative texts in Spanish. In: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2014 (pp. 2710-2713).

85. Teramukai S, Okuda Y, Miyazaki S, Kawamori R, Shirayama M, Teramoto T. Dynamic prediction model and risk assessment chart for cardiovascular disease based on on-treatment blood pressure and baseline risk factors. Hypertension Research. 2016;39(2):113-8.

86. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, Schlüssel MM. Prediction models for cardiovascular disease risk in the general population: systematic review. The BMJ. 2016;353.

87. Chang JT, Meza R, Levy DT, Arenberg D, Jeon J. Prediction of COPD risk accounting for time-varying smoking exposures. PLOS One. 2021;16(3):e0248535.

88. Cadarette SM, Wong L. An introduction to health care administrative data. The Canadian Journal of Hospital Pharmacy. 2015;68(3):232.

89. Hoeven LR, Bruijne MC, Kemper PF, Koopman MM, Rondeel JM, Leyte A, Koffijberg H, Janssen MP, Roes KC. Validation of multisource electronic health record data: an application to blood transfusion data. BMC Medical Informatics and Decision Making. 2017;17(1):1-10.

90. Rahimi AK, Canfell OJ, Chan W, Sly B, Pole JD, Sullivan C, Shrapnel S. Machine learning models for diabetes management in acute care using electronic medical records: a systematic review. International Journal of Medical Informatics. 2022; 162:104758.

91. Conderino S, Bendik S, Richards TB, Pulgarin C, Chan PY, Townsend J, Lim S, Roberts TR, Thorpe LE. The use of electronic health records to inform cancer surveillance efforts: a scoping review and test of indicators for public health surveillance of cancer prevention and control. BMC Medical Informatics and Decision Making. 2022;22(1):1-3.

92. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. Journal of the American Medical Informatics Association. 2022;29(1):187-96.

93. Sharabiani MT, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. Medical care. 2012;1109-18.

94. Vlasschaert ME, Bejaimal SA, Hackam DG, Quinn R, Cuerden MS, Oliver MJ, Iansavichus A, Sultan N, Mills A, Garg AX. Validity of administrative database coding for kidney disease: a systematic review. American Journal of Kidney Diseases. 2011;57(1):29-43.

95. Lucyk K, Lu M, Sajobi T, Quan H. Administrative health data in Canada: lessons from history. BMC Medical Informatics and Decision Making. 2015;15(1):1-6.

96. Birtwhistle R, Keshavjee K, Lambert-Lanning A, Godwin M, Greiver M, Manca D, Lagacé C. Building a pan-Canadian primary care sentinel surveillance network: initial development and moving forward. Journal of the American Board of Family Medicine. 2009;22(4):412-22.

97. Tu K, Mitiku TF, Ivers NM, Guo H, Lu H, Jaakkimainen L, Kavanagh DG, Lee DS, Tu JV. Evaluation of Electronic Medical Record Administrative data Linked Database (EMRALD). American Journal of Managed Care. 2014;20(1):e15-21.

98. Hess DT. The Danish National Patient Register. Surgery for Obesity and Related Diseases. 2016;12(2):304.

99. Rusk N. The UK Biobank. Nature Methods. 2018;15(12):1001.

100. Samadoulougou S, Idzerda L, Dault R, Lebel A, Cloutier AM, Vanasse A. Validated methods for identifying individuals with obesity in health care administrative databases: a systematic review. Obesity Science & Practice. 2020;6(6):677-93.

101. McBrien KA, Souri S, Symonds NE, Rouhi A, Lethebe BC, Williamson TS, Garies S, Birtwhistle R, Quan H, Fabreau GE, Ronksley PE. Identification of validated case definitions for medical conditions used in primary care electronic medical record

databases: a systematic review. Journal of the American Medical Informatics Association. 2018;25(11):1567-78.

102. Barber C, Lacaille D, Fortin PR. Systematic review of validation studies of the use of administrative data to identify serious infections. Arthritis Care & Research. 2013;65(8):1343-57.

103. Canan C, Polinski JM, Alexander GC, Kowal MK, Brennan TA, Shrank WH. Automatable algorithms to identify nonmedical opioid use using electronic data: a systematic review. Journal of the American Medical Informatics Association. 2017;24(6):1204-10.

104. Kroeker K, Widdifield J, Muthukumarana S, Jiang D, Lix LM. Model-based methods for case definitions from administrative health data: Application to rheumatoid arthritis. BMJ Open. 2017;7(6).

105. Van Gaal S, Alimohammadi A, Yu AY, Karim ME, Zhang W, Sutherland JM. Accurate classification of carotid endarterectomy indication using physician claims and hospital discharge data. BMC Health Services Research. 2022;22(1):1-9.

106. Zeltzer D, Balicer RD, Shir T, Flaks-Manov N, Einav L, Shadmi E. Prediction accuracy with electronic medical records versus administrative claims. Medical Care. 2019;57(7):551-559.

107. Van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. Journal of the American Medical Informatics Association. 2022;29(9):1525-34.

108. Coleman N, Halas G, Peeler W, Casaclang N, Williamson T, Katz A. From patient care to research: a validation study examining the factors contributing to data quality

in a primary care electronic medical record database. BMC Family Practice. 2015;16(1):1-8.

109. O'Donnell S, Palmeter S, Laverty M, Lagacé C. Accuracy of administrative database algorithms for autism spectrum disorder, attention-deficit/hyperactivity disorder and fetal alcohol spectrum disorder case ascertainment: a systematic review. Health Promotion & Chronic Disease Prevention in Canada: Research, Policy & Practice. 2022;42(9).

110. Chen C, Qin Y, Chen H, Zhu D, Gao F, Zhou X. A meta-analysis of the diagnostic performance of machine learning-based MRI in the prediction of axillary lymph node metastasis in breast cancer patients. Insights into Imaging. 2021;12:1-2.

111. Furuya-Kanamori L, Xu C, Lin L, Doan T, Chu H, Thalib L, Doi SA. P value–driven methods were underpowered to detect publication bias: analysis of Cochrane review meta-analyses. Journal of Clinical Epidemiology. 2020;118:86-92.

112. Al-Azazi S, Singer A, Rabbani R, Lix LM. Combining population-based administrative health records and electronic medical records for disease surveillance. BMC Medical Informatics and Decision Making. 2019;19(1):1-2.

113. Hughes DM, El Saeiti R, García-Fiñana M. A comparison of group prediction approaches in longitudinal discriminant analysis. Biometrical Journal. 2018;60(2):307-22.

114. Arribas-Gil A, De la Cruz R, Lebarbier E, Meza C. Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators. Biometrics. 2015;71(2):333-43.

115. Miled ZB, Haas K, Black CM, Khandker RK, Chandrasekaran V, Lipton R, Boustani MA. Predicting dementia with routine care EMR data. Artificial Intelligence in Medicine. 2020;102:101771.

116. Jauk S, Kramer D, Großauer B, Rienmüller S, Avian A, Berghold A, Leodolter W, Schulz S. Risk prediction of delirium in hospitalized patients using machine learning: An implementation and prospective evaluation study. Journal of the American Medical Informatics Association. 2020;27(9):1383-92.

117. James G, Witten D, Hastie T, Tibshirani R. Tree-Based Methods. In: James G, Witten D, Hastie T, Tibshirani R, editors. An Introduction to Statistical Learning: with Applications in R. New York, NY: Springer; 2013. p. 303–35.

118. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nature Medicine. 2023; 29(8):1930-40.

2.8 SUPPLEMENTARY MATERIAL

Additional file 1

Search strategy

Ovid interface: EMBASE

Search Date: 11/22/2022

Search Line	Search Terms	Results
1	(valid* or quality or accuracy or sensitivity or specificity).ti,ab,kf.	4721416
2	("electronic medical record*" or "electronic health record*" or "administrative health data" or "administrative data" or "health care data" or "administrative billing record*" or "administrative claims data" or "claims data" or "hospital data" or "hospital discharge data" or "medicare data" or "medicaid data").ti,ab,kf.	142981
3	(smok* or Tobacco).ti,ab,kf.	556655
4	1 AND 2 AND 3	1330
5	Limit 4 to (English language and yr=”1990-Current”)	1283

Ovid Interface: MEDLINE

Search Date: 11/22/2022

Search Line	Search Terms	Results
1	(valid* or quality or accuracy or sensitivity or specificity).ti,ab,kf.	3554243
2	("electronic medical record*" or "electronic health record*" or "administrative health data" or "administrative data" or "health care data" or "administrative billing record*" or "administrative claims data" or "claims data" or "hospital data" or "hospital discharge data" or "medicare data" or "medicaid data").ti,ab,kf.	83536
3	(smok* or Tobacco).ti,ab,kf.	386090
4	1 AND 2 AND 3	581
5	Limit 4 to (English language and yr=”1990-Current”)	555

Scopus

Search Date: 11/22/2022

Search Line	Search Terms	Results
1	TITLE-ABS-KEY(valid* or quality or accuracy or sensitivity or specificity)	8299452
2	TITLE-ABS-KEY("electronic medical record*" or "electronic health record*" or "administrative health data" or "administrative data" or "health care data" or "administrative billing record*" or "administrative claims data" or "claims data" or "hospital data" or "hospital discharge data" or "medicare data" or "medicaid data")	109444
3	TITLE-ABS-KEY(smok* or Tobacco)	539714
4	1 AND 2 AND 3	1394
5	Limit 4 to (Language=English AND Pubyear=1990 to Present)	1275

Web of Science
Search Date: 11/22/2022

Search Line	**Search Terms**	**Results**
1	TS=(valid* or quality or accuracy or sensitivity or specificity)	8087727
2	TS=("electronic medical record*" or "electronic health record*" or "administrative health data" or "administrative data" or "health care data" or "administrative billing record*" or "administrative claims data" or "claims data" or "hospital data" or "hospital discharge data" or "medicare data" or "medicaid data")	102195
3	TS=(smok* or Tobacco)	620314
4	1 AND 2 AND 3	1263
5	Refine 4 by (Languages=English and publication Years=1990-Current)	1222

Additional file 2

Supplemental Table 2.1 Article Characteristics

Article Reference	Authors, Publication year, Country	Journal discipline	Population characteristics	Geographic location	Study period	Study data source	Reference standard
42	Mu et al.,2018,USA	Medicine	Men an women aged ≥67 years diagnosed with end-stage renal disease	California	2011-2013	Medicare	Self-reported
43	LeLaurin et al.,2018,USA	Medical Informatics	Men and women aged ≥18 years diagnosed with cancer	Florida	2015-2018	University of Florida health system	Manual review
64	Wiley et al.,2013,USA	Medical Informatics	Cancer patients matched to cancer free patients on the basis of age and sex.	Tennessee	Not reported	Vanderbilt Biorepository(BioUV)	Manual review
65	Chen et al.,2013,USA	Substance-Related Disorders	Men aged 45-69 years	California	2003-2010	Kaiser Permanente Southern California (KPSC) integrated health care system	Self-reported
44	Caccamisi et al.,2020,Sweden	Medicine	Men and women diagnosed with chronic obstructive pulmonary disease/cardiovascular disease	Stockholm	1999-2012	Swedish primary care centres	Manual review
45	Palmer et al.,2019,USA	Medical Informatics	Aged ≥18 recieved colonoscopy	New Hampshire	2016	Dartmouth-Hitchcock epic data warehouse	Self-reported
29	Wang et al.,2016,USA	Biomedical Research	Lung cancer patients aged 15-45 years	Minnesota	1997-2011	Mayo Clinic (a nonprofit American academic medical center)	Manual review

46	Golden et al.,2020,USA	Medical Informatics	Veterans received chest computed tomography	Oregon	2012-2017	VHA	Manual review
41	Calhoun et al.,2017,USA	Internal Medicine	Veterans participated in the Study of Post-Deployment Mental Health	North Carolina	2001-2016	VHA	Self-reported
49	Ni et al.,2021,USA	Medical Informatics	Pediatric patients (10-20 years of age)	Ohio	2012-2017	Cincinnati Children's Hospital	Manual review
66	McGinnis et al.,2011,USA	Public Health	Veterans aged≥45 years	Pittsburgh	1999-2009	Veterans Health Administration (VHA)	Self-reported
34	Melzer et al.,2018,USA	Medicine	Veterans aged 18-80 years	Minneapolis	2009-2011	VHA	Self-reported
35	Huo et al.,2017,USA	Public Health	Men and women aged ≥18 years	Florida	2009-2014	MarketScan Commercial Claims & Encounters (CC&E) MarketScan Health Risk Assessment (HRA)	Self-reported
67	Kim et al.,2012,USA	Health Services Research	Vetrans in depression treatment	Michigan	1999-2004	VHA	Manual review

31	Regan et al.,2016,USA	Medicine	HIV-infected patients matched to HIV-uninfected patients on the basis of age, gender, race, and the number of medical encounters in a 3:1 ratio. Aged>18 years	Massachusetts	2005-2008	Partners HealthCare System (PHS) Research Patient Data Registry (RPDR)	Manual review
68	Lee et al.,2011,USA	Substance-Related Disorders	Men and women aged ≥18 years	New York	2004	Bellevue Hospital Center electronic medical record (EMR)	Self-reported
47	Atkinson et al.,2017,UK	Medical Informatics	Aged ≥16	Wales	2007-2015	General Practitioners database from the Secure Anonymised Information Linkage (SAIL)	Self-reported
36	Luck et al.,2020,USA	Preventive Medicine	Pregnant women aged 18-44 years	Oregon	2008-2013	Oregon birth certificates Oregon Medicaid eligibility and claims files	Self-reported
62	Havard et al.,2014,Australia	Medicine	Men and women aged ≥45 years	New South Wales (NSW)	2000-2011	NSW Admitted Patients Data Collection (APDC) and the NSW Register of Births Deaths and Marriages (RBDM).	Self-reported
71	Borzecki et al.,2004,USA	Health Services	Veterans	Massachusetts	1998-1999	VHA	Manual review
32	Desai et al.,2016,USA	Epidemiology	Women diagnosed with Rheumatoid Arthritis	Massachusetts	2006-2010	Medicare Brigham and Women's Hospital	Self-reported

69	Jollis et al.,1993,USA	Internal Medicine	Men and women had cardiac catheterization	North Carolina	1985-1990	Duke University Medical Center	Manual review
47	Reps wt al.,2019,USA	Medical Informatics	Men and women of all ages	New Jersey	2000-2016	IBM MarketScan	Self-reported
33	Chung et al.,2020,Taiwan	Public Health	Men and women aged ≥40 years	Taiwan	2000-2012	National Health Insurance Research Database	Self-reported
70	Steffen et al.,2014,USA	Disease Management	Men and women aged ≥20 years	Minnesota	2010	Mayo Clinic (a nonprofit American academic medical center)	Self-reported
72	Bui et al.,2013,USA	Medical Informatics	Veterans	Utah	Not reported	VHA	Self-reported
73	Khor et al.,2013,Australia	Medical Informatics	Cancer patients	Melbourne	Not reported	Peter MacCallum Cancer Centre	Manual review
39	Marrie et al.,2021,Canada	Internal Medicine	Multiple sclerosis patients	Manitoba	1984-2017	Discharge Abstract Database Medical Services Drug Program Information Network	Self-reported
50	Khalifa et al.,2015,USA	Medical Informatics	Not reported	Utah	2006	Informatics for Integrating Biology and the Bedside (i2b2)	Manual review
38	McVeigh et al.,2017,USA	Electronic Data Processing	Men and women aged ≥20 years	New York	2011-2014	New York city Macroscope electronic health record surveillance system	Self-reported

61	Nickel et al.,2016,USA	Health Services Research	Women aged 18-64 years and went through mastectomy operation	Washington	2004-2011	HealthCore Integrated Research Database	Self-reported
40	Floyd et al.,2016,USA	Epidemiology	Veterans with type 2 diabetes	Washington	2004-2010	VHA	Manual review
63	Lujic et al.,2014,Australia	Medicine	Men and women aged ≥45 years	NSW	2000-2010	NSW APDC	Self-reported
74	DeJoy et al.,2014,USA	Women's Health	Women with mean age 23.4	Massachusetts	2001-2008	Certified nurse-midwives (CNM) birth database	Self-reported
52	McVeigh et al.,2016,USA	Electronic Data Processing	Men and Women	New York	2013-2014	The New York City Macroscope	Self-reported
76	Longencker et al.,2000,USA	Nephrology	Men and Women End Stage Renal Disease (ESRD) patients with mean age 57.8	Maryland	1995-1998	Choices for Healthy Outcomes in Caring for ESRD data	Self-reported
53	Roberts et al,2015,USA	Medical Informatics	Not reported	Maryland	2014	Informatics for Integrating Biology and the Bedside (i2b2)	Manual review
77	Meystre et al.,2009,USA	Medical Informatics	Asthma patients aged≥18 years	Utah	Not reported	Informatics for Integrating Biology and the Bedside (i2b2) Hive 1.2	Manual review
60	Maier et al.,2016,Germany	Health Services Research	Myocardial infarction patients aged ≥ 20 years	Berlin	2009-2011	Nongovernmental health insurance (AOK Nordost)	Clinical data
54	Gauthier et al.,2022,Canada	Thoracic Neoplasms	Male and Female patients with advanced lung cancer	Ontario	2015-2017	Princess Margaret Cancer Centre	Manual review

55	O'Brien et al.,2022,USA	Medical Informatics	Men and women with a mean age 66.8 Atherosclerotic cardiovascular disease patients aged≥21 years	North Carolina	2016-2020	National Patient-Centered ClinicalResearch Network	Self-reported
56	Alhaug et al.,2021,Norway	Orthopedics	Male and Female patients (mean age= 66.3 years) operated for spinal stenosis	Brumunddal	2015-2016	National spine register (NORspine) Unstructured electronic patient records (EPRs)	Self-reported
57	Teng et al.,2022,USA	Medicine	Men and women mostly aged between 19-84 years	Washington	2015-2019	Harborview Medical Center	Manual review
58	McGinnis et al.,2022,USA	Medical Informatics	Male and female veterans	Connecticut	2018-2019	Veterans Health Administration (VHA)	Clinical data
59	McGinnis et al.,2019,USA	Substance-Related Disorders	HIV-infected veterans matched to HIV-uninfected by region, age, race/ethnicity, and gender	Connecticut	1996-2015	Veterans Health Administration (VHA)	Clinical data
30	Harris et al.,2021,USA	Medical Informatics	Not reported	Kentucky	2017-2019	University of Kentucky's network of hospitals and clinics	Manual review
36	Etzioni et al.,2019,USA	General surgary	Men and women experienced surgical hospitalization with length of stay>1 day	Arizona	2013-2015	Academic hospitals participated in the American College of Surgeons' National Surgical Quality Improvement Program (NSQIP)	Self-reported

10	Groenhof et al.,2020,Netherlands	Epidemiology	Men and women Median age-60	Utrecht	2013-2019	University Medical Center Utrecht	Self-reported
24	Wang et al.,2019,USA	Medical Informatics	Lung cancer patients aged 15-45 years Not reported	Minnesota	1997-2011 2006	Mayo Clinic (a nonprofit American academic medical center) Informatics for Integrating Biology and the Bedside (i2b2)	Manual review
75	Zeng et al.,2006,USA	Medical Informatics	Asthma and Chronic Obstructive Pulmonery Disease(COPD) Patients	Massachusetts	Not reported	Partners' Health Care Research Patient Data Repository (RPDR)	Manual review
51	Urbain et al.,2015,USA	Medical Informatics	Not reported	Wisconsin	2014	Informatics for Integrating Biology and the Bedside (i2b2)	Manual review
79	Savova et al.,2008,USA	Medical Informatics	Not reported	Minnesota	2006	Informatics for Integrating Biology and the Bedside (i2b2)	Manual review
78	Clark et al.,2008,USA	Medical Informatics	Not reported	Massachusetts	2006	Informatics for Integrating Biology and the Bedside (i2b2)	Manual review
81	Mant et al.,2000,UK	Public health Medicine	Men and women with mean age 62 years	Oxford	1993-1996	General practitioner (GP) records from Oxfordshire	Self-reported

80	Yeager et al.,2010,USA	Healthcare administration	Asthma patients aged≥20 years	California	2001-2008	Mobile Examination Centre(MEC) data from a portion of the participants of National Health and Nutrition Examination Survey	Clinical data
82	Liu et al.,2012,USA	Medical Informatics	Not reported	Massachusetts	Not reported	Vanderbilt Synthetic Derivative (SD) database	Manual review
83	Figueroa et al.,2014,Chile	Biomedical Engineering	Not reported	Concepción	Not reported	Guillermo Grant Benavente hospital (HGGB)	Manual review

Additional file 3

Sensitivity analysis

Supplemental Table 2.2 Meta-regression model parameter estimates (SE) for validity measures of PPV, sensitivity, and specificity

Variable	Validity measure					
	PPV		Sensitivity		Specificity	
	Parameter estimate (SE)	P-value	Parameter estimate (SE)	P-value	Parameter estimate (SE)	P-value
Data source						
EMR	-0.096(0.095)	0.319	**0.257(0.092)**	0.007	-0.063(0.059)	0.299
Administrative	Ref		Ref		Ref	
Predictive model						
Yes	-0.003(0.112)	0.979	**0.187(0.086)**	0.034	-0.118(0.062)	0.064
No	Ref		Ref		Ref	
Reference standard						
Chart review/clinical data	0.088(0.077)	0.261	0.088(0.092)	0.342	-0.039(0.063)	0.540
Self-report	Ref		Ref		Ref	
Clinical population						
Yes	0.175(0.097)	0.079	-0.045(0.099)	0.648	**-0.152(0.065)**	0.025
No	Ref		Ref		Ref	
Study population age						
Restricted	-0.112(0.072)	0.131	-0.018(0.077)	0.814	0.050(0.052)	0.344
All ages	Ref		Ref		Ref	
Country of data origin						
US	-0.037(0.095)	0.702	-0.066(0.100)	0.511	**0.208(0.063)**	0.002
non-US	Ref		Ref		Ref	

Note: boldface font denotes a statistically significant estimate; EMR = electronic medical record; SE=standard error; PPV= positive predictive value

Results from the null model

The null three-level meta-regression models produced pooled estimates for PPV, sensitivity, and specificity of 0.849 (95% [CI]: 0.788-0.911), 0.658 (95% [CI]: 0.557-0.758), and 0.915 (95% [CI]: 0.857-0.973) respectively. The estimated variances for these measures were 0.018 (PPV), 0.075 (sensitivity), and 0.027 (specificity). Level 2 accounted for most of the estimated variance in sensitivity (61.2%) and specificity (79.2%), with level 3 contributing to a lesser extent (sensitivity=38.7%, specificity=20.8%). In contrast, level 3 accounted for the largest share of estimated variance in PPV (62.4%), followed by level 2 (37.6%). The residual regressions tests (deviation of the intercept from zero) suggest the

absence of publication bias in the reported estimates of PPV (p=0.646), sensitivity (p=0.552), and specificity (0.117). The introduction of predictors in the models led to a reduction in heterogeneity variance for PPV, sensitivity, and specificity by 5.3%, 37.3%, and 18.5%, respectively, compared to the null model.

Additional file 4

Supplemental Table 2.3 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist

TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	30
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	31-32
INTRODUCTION/ BACKGROUND			
Rationale	3	Describe the rationale for the review in the context of what is already known.	33
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	34
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	-
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	34-35
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	34-35
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	(Additional file 1)
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	35-36
Data collection	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any	36-37

process		processes for obtaining and confirming data from investigators.	
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	36-38
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	39
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	37
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	37-39
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	39
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	37-39
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	39-40
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	40-42
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	49-50
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	-
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	47-49
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	48
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	50 Additional file 3
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	50-53

Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	53
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	53-54
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	iii, 240

CHAPTER 3: MODEL-BASED ALGORITHMS TO ASCERTAIN SMOKING IN ADMINISTRATIVE HEALTH DATA: A REGISTRY-BASED VALIDATION STUDY

3.1 Overview

This chapter evaluates the validity of administrative health data (AHD) for identifying current smoking status through the development and evaluation of model-based algorithms (MBAs). Based on a clinical registry-based cohort, the study applies machine learning models, random forest, and least absolute shrinkage and selection operator logistic regression to enhance the ascertainment of smoking by integrating clinical, demographic, and socioeconomic indicators across AHD sources. The methodological innovation lies in the comparative evaluation of MBAs and rule-based algorithms (RBAs) within the same cohort, offering insights into performance trade-offs across sensitivity, specificity, and predictive values. By demonstrating improved sensitivity and balanced accuracy of MBAs relative to RBAs, this work contributes to a growing body of literature seeking to progress behavioral risk factor surveillance using routinely collected data. Stratified analyses found performance variations by sex and residence location, underscoring the importance of sociodemographic context in algorithm validation. The study supports the use of linked data and data-driven models to strengthen public health monitoring while recommending methodological choices tailored to specific research objectives.

Publication Details

Haque MA, Nickel N, Turgeon M, Lix LM. Model-based algorithms to ascertain smoking in administrative health data: a registry-based validation study. Revised and Resubmitted, *BMC Medical Research Methodology*. *January, 2025.*

3.2 Abstract

Background: Accurate measurement of smoking in population-based administrative health data (AHD) poses challenges due to the indirect nature of smoking-related information collection. While most studies use rule-based algorithms (RBAs) based on diagnosis codes, model-based algorithms (MBAs) utilizing machine learning (ML) with diverse data features might have improved sensitivity and accuracy. We developed ML model-based algorithms (MBAs) for ascertaining smoking in AHD and compared them to RBAs.

Methods: We conducted a retrospective cohort study using AHD (hospital abstracts, medical claims, and prescription drug records) from April 1, 2012, to March 31, 2020, from Manitoba, Canada. The study included adults (≥18 years) from a clinical registry containing self-reported current smoking. Clinical data were linked with up to five years of hospital records, physician billing claims, and prescription medication records. RBAs were based on diagnosis codes for tobacco use and nicotine dependence medication. MBAs, constructed using Random Forest (RF) and Least Absolute Shrinkage and Selection Operator (LASSO) models, included smoking indicators, comorbid condition, and sociodemographic factors. Training and test datasets were used to develop and evaluate the MBAs, respectively. Sensitivity, specificity, positive and negative predictive values (PPV, NPV), balanced accuracy, and their 95% confidence intervals (CIs) were estimated.

Results: The cohort comprised 24,718 individuals (88.6% female); prevalence of current smokers was 10.0%. A comprehensive RBA had sensitivity 23.3% (95% CI: 20.3-26.5), specificity 98.9% (95% CI: 98.7-99.2), and PPV 70.9% (95% CI: 65.1-76.1). An MBA based on RF had sensitivity 66.8% (95% CI: 63.3-70.2), specificity 77.8% (95% CI: 76.8-78.8), and PPV 25.1% (95% CI: 23.8-26.4). NPV was consistently above 90.0%. MBAs had higher balanced

accuracy than RBAs. Stratified analyses by sex and residence location revealed differences in estimates for MBAs and the RBAs. The number of years of AHD did not affect the MBA results. While MBAs had better sensitivity, RBAs had better specificity.

Conclusions: Our study highlights the potential of comprehensive data integration and ML methods to improve the sensitivity and accuracy of smoking identification in AHD. Balancing accurate smoker identification with the risk of false positives is crucial when choosing an algorithm to ascertain current smokers using AHD.

3.3 Background

Administrative health data (AHD) are a rich source of information collected and maintained for the purpose of managing and supporting the healthcare system (1). These data typically capture hospital visits, physician billing claims, and drug dispensation records. The wide coverage of these data sources provides opportunities to conduct population-based studies. Smoking is an important variable in population-based studies for understanding the burden of smoking-related diseases (2), designing effective public health interventions (3), evaluating policy interventions (4), and addressing health disparities (5). However, because smoking is not directly recorded in AHD, obtaining valid information is challenging. Therefore, it is important to evaluate the validity of AHD for accurately measuring smoking behaviors.

Current approaches to identify smoking status in AHD often rely on rule-based algorithms (RBAs) or deterministic case definitions that involve measuring the type and frequency of specific codes, such as diagnosis codes in the records of an individual over a specific timeframe (6). However, studies have consistently shown that RBAs for smoking ascertainment in electronic health data, including AHD, have low sensitivity (7,8) and accuracy (9). To illustrate, a study evaluating RBAs to identify tobacco use in US Medicare claims reported sensitivity as low as 9.0%, highlighting substantial under-detection of smokers (8). Model-based methods offer an alternative for algorithm development (10) with the capacity to incorporate a diverse set of data elements, including demographics, diagnosis codes, billing codes, and prescription medications associated with smoking, and to assess their relative importance in case ascertainment. By leveraging a broader spectrum of information, machine-learning (ML) model-based methods can potentially improve the sensitivity and accuracy (11) of smoking ascertainment algorithms over their rule-based counterparts.

A recent systematic review that assessed the validity of smoking measurement in AHD and electronic medical records (EMR) found that the majority of the algorithms were rule-based and that model-based algorithms were primarily developed for EMR data (12). Algorithms to identify smoking status in AHD are frequently validated using non-clinical data sources prone to misclassification errors, such as survey data (13), birth certificates (14), and medical chart reviews (15). Moreover, constructing model-based algorithms (MBAs) for a single AHD source is challenging due to a limited number of relevant indicators for model development. However, the availability of multiple smoking indicators in different linkable AHD sources and its population-based nature suggest the viability of using ML models to construct smoking ascertainment algorithms in linked data.

This study aimed to assess the validity of AHD to ascertain smoking for a clinical registry-based cohort using multiple linked data sources. The objectives were to 1) develop ML MBAs to identify current smoking status and 2) investigate whether the MBAs result in higher validity estimates than RBAs.

3.4 Methods

We conducted a retrospective cohort study with three linked AHD data sources from April 1, 2012, to March 31, 2020. We followed the Reporting of Studies Conducted Using Observational Routinely-collected Data guidelines (16) (Additional file 1).

3.4.1 Data Sources

Study data were from the Population Research Data Repository housed at the Manitoba Centre for Health Policy (MCHP) (17). Study data sources included the Manitoba health insurance registry, hospital abstracts, medical (i.e., physician billing) claims, prescription medication dispensations, and the Statistics Canada Census. These data were used to develop

algorithms using rule-based and ML methods. The Manitoba Health Insurance Registry contains individual-level demographic and residential information for everyone registered to receive health services in Manitoba. Hospital abstracts encompass records from hospital stays, while medical claims include documentation of physician visits in both offices and outpatient departments. Diagnoses in hospital abstracts and physician billing claims are coded using the International Classification of Diseases (ICD) codes. Specifically, hospital abstracts use the 10th version with Canadian adaptations (ICD-10-CA) and physician billing claims use the 9th version with clinical modifications (ICD-9-CM). Prescription medication dispensations from community-based pharmacies are captured in the Drug Program and Information Network (DPIN) database. The Anatomical Therapeutic Chemical (ATC) Classification System (18) codes for classifying pharmaceutical substances are used to capture information in DPIN. The Canada Census (19) is a nationwide population-based data source that includes aggregated socio-demographic information that can be used to construct income quintiles based on postal codes in the Manitoba Health Insurance Registry.

The Manitoba bone mineral density (BMD) clinical registry (20) was used as the validation data source, considering the availability of information on smoking status and its completeness and accuracy. A BMD test (21) is a diagnostic technique employed to quantify the mineral density, particularly calcium, within bone structures. These tests play a pivotal role in identifying individuals vulnerable to fractures resulting from diminished bone density and are crucial for the timely detection and treatment of osteoporosis. These tests are usually advised for postmenopausal women, older individuals, those with specific medical conditions or medications impacting bone health, and individuals with a history of fractures. The BMD registry contains BMD tests performed within Manitoba; it includes men and women, many of

whom are 50 years of age and older. The data on smoking status documented in this registry was obtained directly from patients during clinical visits, ensuring greater reliability than survey responses. This registry's self-reported smoking status information is frequently used as an independent variable in fracture risk prediction models (22).

3.4.2 Study Cohort

The study cohort included adults (age≥18 years) captured in the BMD registry with a service date between April 1, 2017, and March 31, 2020, and linkable to administrative data through a scrambled personal health identification number (PHIN) at the individual level. The PHIN is a unique nine-digit number given by Manitoba Health to individuals registered for health insurance in Manitoba to receive treatment at facilities submitting electronic claims (23). If there was more than one record for a cohort member, then a single record was selected randomly to ascertain smoking. The consistency of current smoking status across multiple records for a cohort member was also assessed.

3.4.3 Target Measure

The target measure (dependent variable in models) was self-reported structured information on current smoking status (1=Yes, 0=No) from the BMD registry. The registry data do not capture information on whether an individual had previously smoked or the amount that an individual currently smokes. The service date (i.e., date of appointment for BMD test) in the BMD registry was the index date.

3.4.4 Indicators of Smoking Status in Administrative Health Data

Independent variable identification used up to five years of AHD before each individual's index date. This inclusion criterion (Figure 1) was used to ensure that every individual in the validation cohort was followed for the same length of time during the study period. Information

from the AHD sources included diagnosis codes for smoking-related conditions, such as a history of tobacco use. The use of prescribed medications such as nicotine dependence drugs was measured from the DPIN data. The Manitoba Health Insurance Registry was used to obtain age (in years), biological sex, and place of residence (urban, rural). Income quintiles, measured at the level of a dissemination area (consisting of 400-700 individuals) based on total household income, were obtained from the Statistics Canada Census (24). Indicators from administrative data were used as independent variables in models for constructing current smoking status identification algorithms (i.e., variables related to nicotine dependence, chronic obstructive pulmonary disease [COPD], asthma, and smoking cessation drugs; biological sex, place of residence, age, and income quintiles). Descriptions of the indicators is provided in Additional file 2.

3.4.5 Rule-based Algorithm (RBA)

Two rule-based algorithms were developed using the diagnosis codes from AHD. The distinction between these two algorithms lies in ATC code for Bupropion, which is a smoking cessation aid, and is also used to treat depression.

RBA1: One or more hospital separation records with at least one of the following diagnosis codes (25): ICD-10-CA code F17.2 (nicotine dependence), ICD-10-CA code Z72.0 (problems related to lifestyle: tobacco use) OR one or more physician claims with at least one of the following diagnosis codes: ICD-9-CM code 305.1(nicotine dependence), ICD-9-CM code V15.82 (personal history of tobacco use), OR one or more pharmaceutical dispensation records with at least one of the following ATC drug codes (26): N07BA03 (Varenicline), N07BA01 (Nicotine).

RBA2: The ATC drug code (27): N06AX12 (Bupropion), was added to RBA1.

3.4.6 Statistical Analysis

Cohort characteristics were reported using frequencies and percentages or mean and standard deviation stratified by the individuals' current smoking status. Descriptive analysis was performed in SAS® version 9.4 (SAS Institute, Cary, North Carolina, United States). Model-based analyses were conducted using R (28). We conducted analyses separately for one to five years of administrative data preceding the index date to facilitate comparisons across years since previously validated algorithms used different lengths of time preceding the index date to identify smoking status (25–27). MBAs were constructed using ML models developed with and without sociodemographic variables: age, biological sex, place of residence, and income quintile. The missing category for the place of residence, named 'Not Found', was combined with the rural category in the MBAs. Similarly, the 'Not Found' category for the income quintile was merged into the highest income quintile category (Q5) in the MBAs.

3.4.7 Model-based Algorithm (MBA)

We used two ML models, random forest (RF) and least absolute shrinkage and selection operator (LASSO) logistic regression to predict current smoking status. We selected these two models for comparison since model performance has been shown to vary with the characteristics of the dataset, the magnitude of imbalance in the target variable, and the specific task or domain in which they are applied (29,30). The ability to avoid overfitting and identify important variables also motivated the selection of these two ML models. The RF is an ensemble learning method that operates by constructing a large number of decision trees, where each tree is built by recursively partitioning the data into nodes and making splits based on specific criteria, such as the minimum number of observations in a node. The results from individual trees are then aggregated with a majority vote. RF overcomes the main limitations of single decision trees by

reducing the variance and decorrelating the trees with bootstrapping and aggregating, respectively (31). LASSO extends logistic regression by favoring a sparse model, setting less impactful variable coefficients to zero through a penalty term proportional to the coefficients (32).

The study cohort was randomly stratified into a training set (70% of the cohort) and a test set (30%), ensuring that the prevalence of current smoking was consistent in both sets. An optimization set (30%) from the training data was randomly chosen to find the optimal value for hyperparameters using the grid search approach (33). The hyperparameters for the RF model were the number of variables randomly sampled as candidates at each split and the minimum number of observations required to split an internal node during the construction of a decision tree for the model. The number of trees was set at 1500 for each RF model (34). The single hyperparameter for LASSO logistic regression was the multiplier of the penalty term, commonly known as lambda (the higher the value of lambda, the stronger the penalty, and the more coefficients are set to zero). The hyperparameter estimates that resulted in the highest area under the curve (AUC) of the receiver operating characteristic (ROC) curve were chosen to train the model. The relative importance of variables was measured with the contribution to the reduction in impurity and the magnitude of coefficients, respectively, for RF and LASSO logistic regression models using the VarImp function from the caret R package (35).

The trained models were applied to the test dataset to estimate the predicted probability of current smoking. When choosing the classification threshold from predicted probabilities, we aimed for a balance between accurately identifying both positive and negative cases by giving equal importance to sensitivity and specificity. Hence, the cut-off probabilities corresponding to

the highest Youden's J index from ROC curve analyses were chosen for an individual to be classified as a current smoker according to ML model-based algorithms (36).

3.4.8 Performance Evaluation

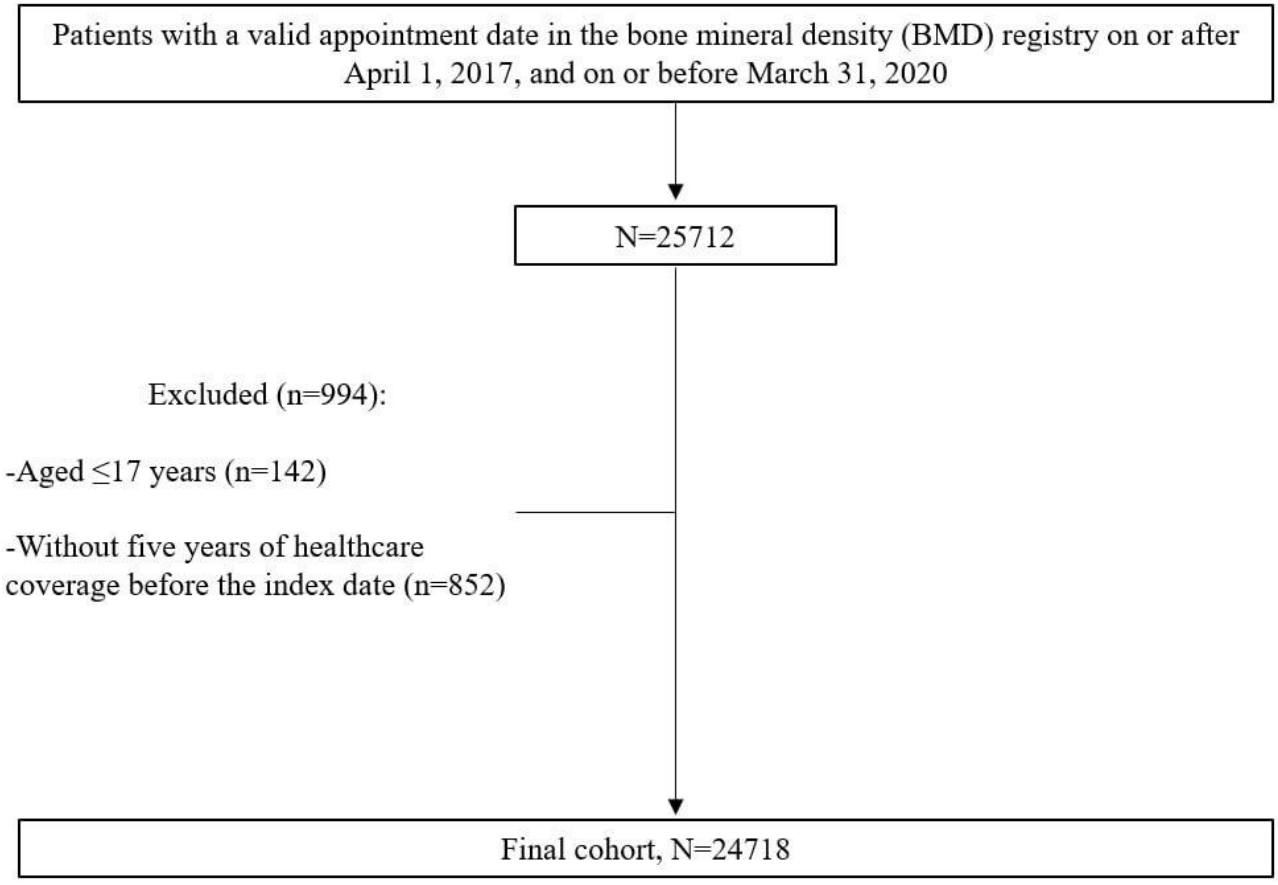
For each MBA and RBA, performance was assessed using the test set based on validity measure estimates. The validity measures were sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and balanced accuracy (37). Sensitivity estimates the proportion of true smokers correctly identified by an algorithm, while specificity estimates the proportion of true non-smokers accurately identified as non-smokers. PPV estimates the probability that an individual classified by an algorithm as a smoker is a true smoker. In contrast, NPV estimates the probability that an individual identified as a non-smoker is a true non-smoker. Balanced accuracy is a summary measure that integrates sensitivity and specificity to estimate an algorithm's classification performance. By equally weighting the correct identification of smokers and non-smokers, it offers a balanced evaluation of accuracy across classes (smokers and non-smokers). However, the relevance of balanced accuracy may vary depending on the study context, as some scenarios may prioritize one measure over the other, such as focusing on high specificity when reducing false positives is crucial. For each algorithm, we reported both overall and stratified estimates of validity measures by biological sex and location of residence. The corresponding 95% confidence interval (CI) was also reported. DeLong's non-parametric test was used to compare estimates for MBAs (38) using a nominal $\alpha = 0.05$.

3.5 Results

We identified 25,712 individuals in the BMD registry, of which 24,718 (96.1%) met the inclusion criteria (Figure 3.1). Only 465 (1.9 %) cohort members had more than one record in the registry, for whom a single record was randomly selected. For the vast majority (97.0 %) of

those having more than one registry record, current smoking status was consistent across all records.

Figure 3.1 Cohort flow diagram



3.5.1 Cohort Characteristics

The prevalence of current smokers within the study cohort was 10.0% (Table 3.1). The majority of cohort members were female (88.6%), residing in urban areas (70.2%), and older adults (average age=67.0 years). Significant differences were observed in the percentages of current smokers and non-smokers by biological sex, residence location, and income quintile. Notably, within the lowest income quintile, the prevalence of current smokers exceeded that of non-smokers (26.4% vs. 18.4%). In contrast, within the highest income quintile, the percentages of current smokers and non-smokers were 12.5% and 22.2%, respectively. On average, current

non-smokers tended to be older than smokers. Additionally, the prevalence of current smokers exceeded that of non-smokers across various smoking-related indicators, including diagnoses of tobacco use, smoking-related comorbid conditions, and prescription medication dispensations. For instance, with five years of AHD usage, the proportion of current smokers was 22.1 times that of non-smokers for the diagnosis of nicotine dependence in physician claims.

Table 3.1 Characteristics of the study cohort stratified by self-reported current smoking status (N = 24,718)

Characteristic	**Current smoker** n=2475 (10.0%)	**Non-smoker** n=22,243 (90.0%)
Sex		
Male	360 (14.5)	2452 (11.0)
Female	2115 (85.5)	19,791 (89.0)
Residence location at index date		
Urban	1541 (62.3)	15,813 (71.1)
Rural	905 (36.5)	6257 (28.1)
Not found	29 (1.2)	173 (0.8)
Income quintile		
Q5 (highest)	309 (12.5)	4942 (22.2)
Q4	415 (16.7)	4506 (20.3)
Q3	507 (20.5)	4596 (20.6)
Q2	562 (22.7)	3942 (17.7)
Q1 (lowest)	653 (26.4)	4084 (18.4)
Not found	29 (1.2)	173 (0.8)
Age at index date, mean (SD)	62.4 (10.3)	67.6 (10.9)
Diagnoses		
Nicotine dependence or tobacco use problems related to lifestyle in hospital abstracts / personal history of tobacco use in medical claims	30 (1.2)	24 (0.11)
Nicotine dependence diagnosis in medical claims	258 (10.4)	105 (0.47)
COPD /asthma diagnosis in hospital abstracts	140 (5.6)	484 (2.2)
COPD in medical claims	675 (27.3)	1718 (7.7)
Asthma in medical claims	460 (18.6)	2949 (13.3)
Use of respiratory system treatment drug (Salbutamol) in prescription drug records	598 (24.3)	3214 (14.5)
Use of medication primarily recommended as an antidepressant	159 (6.4)	552 (2.5)

and for smoking cessation in prescription drug records (Bupropion)		
Use of medication to quit smoking in prescription drug records (Varenicline)	276 (11.2)	186 (0.84)
Use of medication to aid in smoking cessation in prescription drug records (Nicotine)	123 (4.9)	59 (0.27)

Note: Diagnoses are based on five years of administrative health data; Standard deviation (SD); Chronic obstructive pulmonary disease (COPD); Quintile (Q); Values reported are frequency (%) unless otherwise stated

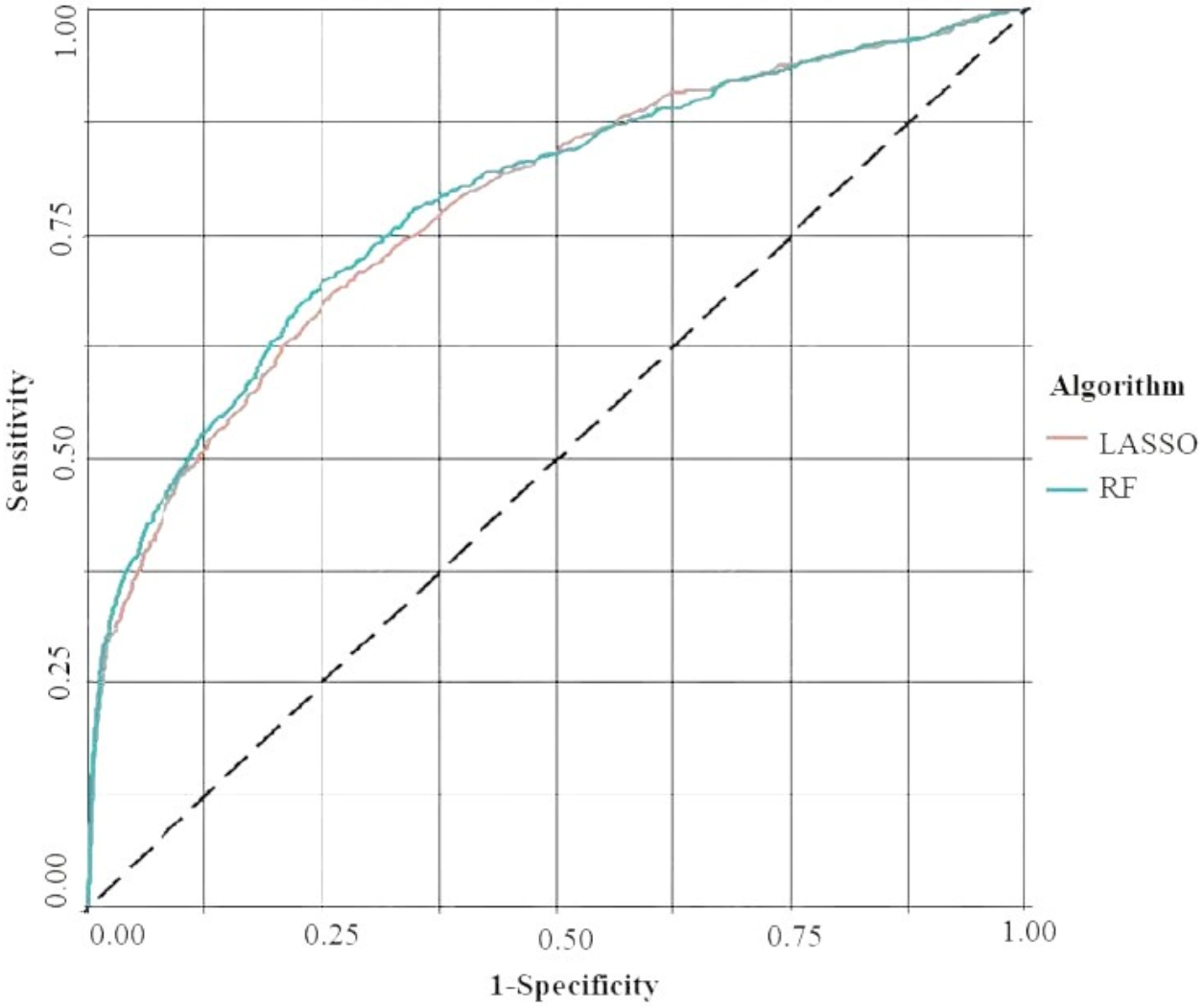
3.5.2 Overall Performance of Algorithms

For the MBAs, the AUC estimates for the RF models were slightly higher than the AUC estimates for the LASSO models (Figure 3.2). The AUC for the RF models ranged from 0.74 to 0.79, while for the LASSO models the range was 0.72 to 0.77 for different numbers of years of AHD for independent variable identification. The comparisons of AUC estimates for the algorithms (Additional file 3) using DeLong's test indicated statistically significant differences between the pair of MBAs except for the LASSO model versus the RF model using five years of AHD.

The hyperparameter estimates for RF models did not vary substantially across the number of years of AHD for independent variable identification (number of variables randomly sampled as candidates at each split: 2, 3; minimum number of observations required to split: 31, 35). For the LASSO models, the estimates of lambda ranged between 0.0004 and 0.0046. The estimated optimized cut-off for predicted probability varied from 0.088 to 0.1146.

Figure 3.2 Comparison of model-based algorithms (independent variables identified based on

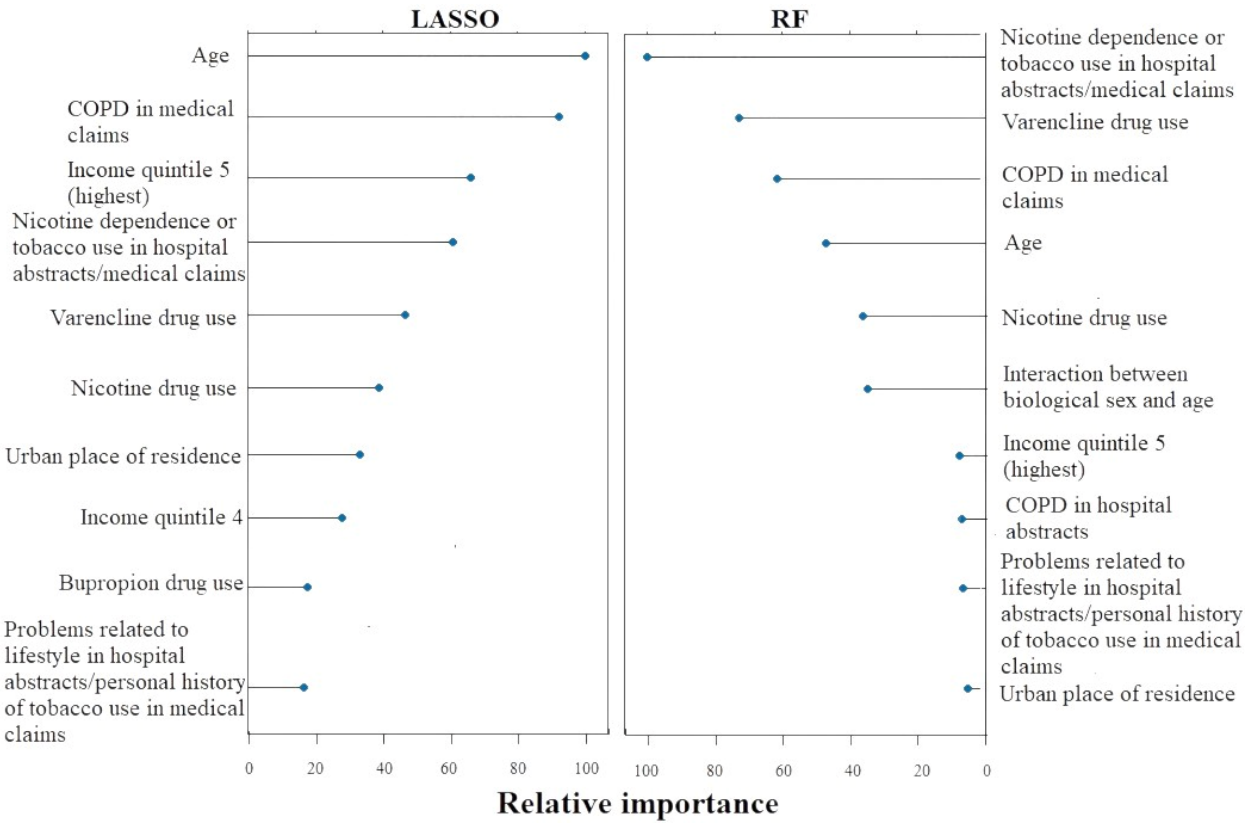
five years of administrative health data) for area under the receiver operating characteristic curve in test data



Note: Random forest (RF); Least Absolute Shrinkage and Selection Operator (LASSO)

Figure 3.3 depicts the ranking of the importance of variables resulting from the RF and LASSO models. Age, diagnosis of nicotine dependence in medical claims, and diagnosis of chronic obstructive pulmonary disease in medical claims consistently appeared in this list of variables. The income quintiles also contributed notably to smoking ascertainment in AHD using MBAs.

Figure 3.3 Top ten most important variables for Random Forest (RF) and Least Absolute Shrinkage and Selection Operator (LASSO) models using independent variables identified based on five years of administrative health data



Estimates of sensitivity, NPV, and balanced accuracy estimates for MBAs were higher than the estimates for RBAs for all years (Table 2). RBAs performed better than MBAs for specificity and PPV estimates. MBAs based on models excluding sociodemographic variables had low sensitivity and high specificity.

Table 3.2 Random forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression, and rule-based algorithm (RBA) performance for test data (n=7414) by number of years of administrative health data used for identifying independent variables

Validity measure	**Algorithm**	**Number of years of administrative data before cohort index date**				
		1	2	3	4	5
Sensitivity	RBA1	7.5 (5.7, 9.7)	11.7 (9.5, 14.3)	15.6 (13.1, 18.4)	20.3 (17.5, 23.4)	23.3 (20.3, 26.5)
	RBA2	9.7 (7.7, 12.1)	14.7 (12.2, 17.4)	19.3 (16.5, 22.3)	23.9 (20.9, 27.2)	27.3 (24.2, 30.7)
	RF	63.6 (60.0, 67.1)	66.8 (63.3, 70.2)	62.5 (58.9, 66.0)	64.3 (60.7, 67.7)	66.8 (63.3, 70.2)
	LASSO	61.7 (58.1, 65.2)	**70.1 (66.6, 73.4)**	65.6 (62.1, 69.0)	65.8 (62.2, 69.2)	68.5 (64.9, 71.8)
	RF*	0.7 (0.0, 1.2)	0.4 (0.0, 0.8)	0.6 (0.0, 1.0)	0.4 (0.0, 0.8)	0.9 (0.2, 1.6)

	LASSO*	1.2	0.1	0.3	0.4	0.4
		(0.4, 2.0)	(0.0, 0.5)	(0.0, 0.7)	(0.0, 0.8)	(0.0, 0.8)
	RBA1	**99.8**	99.6	99.4	99.3	98.9
		(99.6, 99.9)	(99.5, 99.8)	(99.2, 99.6)	(99.0, 99.5)	(98.7, 99.2)
	RBA2	98.6	98.1	97.6	97.2	96.6
		(98.3, 98.9)	(97.8, 98.4)	(97.2, 98.0)	(96.8, 97.6)	(96.1, 97.0)
	RF	74.4	72.3	78.5	78.4	77.8
		(73.3, 75.4)	(71.2, 73.4)	(77.5, 79.5)	(77.3, 79.3)	(76.8, 78.8)
Specificity	LASSO	73.3	66.7	72.9	74.8	73.7
		(72.3, 74.4)	(65.5, 67.8)	(71.8, 74.0)	(73.7, 75.8)	(73.6, 74.8)
	RF*	99.6	99.6	99.5	99.8	99.3
		(99.5, 99.8)	(99.5, 99.8)	(99.3, 99.7)	(99.7, 99.9)	(99.1, 99.5)
	LASSO*	99.1	99.6	99.7	99.8	99.8
		(98.9, 99.3)	(99.5, 99.7)	(99.6, 99.8)	(99.7, 99.9)	(99.7, 99.9)
	RBA1	**78.9**	78.3	74.3	75.9	70.9
		(67.9, 86.7)	(69.9, 85.0)	(67.1, 80.5)	(69.6, 81.2)	(65.1, 76.1)
	RBA2	44.1	46.6	47.3	48.7	47.2
		(36.9, 51.6)	(40.5, 52.7)	(42.1, 52.6)	(44.0, 53.5)	(42.9, 51.5)
	RF	21.7	21.2	24.4	24.8	25.1
		(20.5, 22.8)	(20.1, 22.2)	(23.1, 25.8)	(23.5, 26.1)	(23.8, 26.4)
PPV	LASSO	20.5	18.9	21.2	22.5	22.4
		(19.4, 21.6)	(18.1, 19.8)	(20.1, 22.3)	(21.4, 23.7)	(21.4, 23.6)
	RF*	17.8	10.3	13.5	27.3	13.7
		(3.7, 32.0)	(0.0, 21.4)	(2.5, 24.5)	(0.1, 53.6)	(4.2, 23.2)
	LASSO*	13.2	20.0	20.0	23.1	20.0
		(5.2, 21.3)	(0.0, 55.5)	(0, 44.7)	(0.2, 45.9)	(0.0, 40.0)
	RBA1	90.6	91.0	91.3	91.8	92.1
		(90.5, 90.8)	(90.8, 91.2)	(91.1, 91.6)	(91.5, 92.1)	(91.8, 92.3)
	RBA2	90.7	91.2	91.5	91.9	92.3
		(90.6, 91.0)	(90.9, 91.4)	(91.3, 91.8)	(91.7, 92.3)	(92.0, 92.6)
	RF	94.8	95.1	94.9	95.2	**95.5**
		(94.3, 95.3)	(94.7, 95.6)	(94.5, 95.4)	(94.7, 95.6)	**(95.0, 95.9)**
NPV	LASSO	94.5	95.2	95.0	95.2	95.5
		(94.0, 95.0)	(94.7, 95.7)	(94.5, 95.5)	(94.7, 95.6)	(95.0, 95.9)
	RF*	90.0	89.9	90.0	90.0	90.0
		(89.3, 90.7)	(89.3, 90.7)	(89.3, 90.7)	(89.3, 90.7)	(89.3, 90.7)
	LASSO*	90.0	89.9	90.0	90.0	90.0
		(89.3, 90.7)	(89.3, 90.7)	(89.3, 90.7)	(89.3, 90.7)	(89.3, 90.7)
	RBA1	53.7	55.7	57.5	59.8	61.1
		(52.5, 54.8)	(54.6, 56.8)	(56.3, 58.6)	(58.6, 60.9)	(59.9, 62.2)
	RBA2	54.2	56.4	58.4	60.6	61.9
		(53.1, 55.3)	(55.3, 57.9)	(57.3, 59.5)	(59.5, 61.7)	(59.9, 62.1)
	RF	68.9	69.6	70.5	71.3	**72.3**
Balanced		(67.8, 69.9)	(68.5, 70.6)	(69.5, 71.5)	(70.3, 72.3)	**(71.3, 73.3)**
accuracy	LASSO	67.5	68.4	69.3	70.3	71.1
		(66.4, 68.6)	(67.3, 69.4)	(68.3, 70.3)	(69.3, 71.3)	(70.1, 72.1)
	RF*	50.2	50.0	50.0	50.1	50.1
		(49.8, 51.1)	(49.4, 50.4)	(49.4, 50.4)	(49.7, 50.6)	(49.7, 50.6)
	LASSO*	50.1	50.0	50.1	50.1	50.1
		(49.7, 50.6)	(49.4, 50.4)	(49.7, 50.6)	(49.7, 50.6)	(49.7, 50.6)

Note: * excluded sociodemographic variables of age, biological sex, income quintile, and place of residence; boldface font denotes the highest estimates for each validity measure; Positive predictive value (PPV); Negative predictive value (NPV)

Overall, the estimates for the MBAs using the RF model and LASSO model were similar. The number of years of AHD increased the sensitivity estimates for RBAs considerably. Estimates of sensitivity, NPV, and balanced accuracy were slightly higher for RBA2 than for RBA1. Sensitivity estimates ranged from 7.5% (95% CI: 5.7-9.7) to 68.5% (95% CI: 63.3-70.2), with the highest and lowest values derived from the LASSO model (five years of AHD) and RBA1 (one year of AHD), respectively. Specificity estimates varied between 66.7% (95% CI: 65.5-67.8) and 99.8% (95% CI: 99.6-99.9), with RBA1 (one year of AHD) and LASSO (two years of AHD) yielding the highest and lowest values, respectively. PPV estimates ranged from 18.9% (95% CI: 18.1-19.4) to 78.9% (95% CI: 67.9-86.7), with the RBA1 (one year of AHD) and the LASSO model (two years of AHD) providing the highest and lowest values, respectively. For all algorithms, NPV estimates remained consistent across the years of AHD usage.

3.5.3 Performance of Algorithms by Demographic Characteristics

The MBAs showed substantial differences between females and males (Additional file 4) for sensitivity and specificity estimates. PPV estimates differed considerably between females and males for RBA1. The magnitude of the differences decreased with the number of years of AHD usage. For the RF model based on one year of AHD, estimates of sensitivity for females and males were 59.9% (95% CI: 56.0-63.8) and 85.2% (95% CI: 77.1-91.3), respectively. For females, the specificity estimate for the LASSO model based on one year of AHD was 75.5% (95% CI: 74.4-76.6), and it was 56.2% (95% CI: 52.5-69.8) for males. The RBA1 based on one

year of AHD produced PPV estimates of 80.5% (95% CI: 68.9-88.5) and 66.7 (33.7, 88.7), respectively, for females and males.

Sensitivity and specificity estimates for the MBAs also differed considerably based on the place of residence (Additional file 5). Specifically, sensitivity estimates were higher for rural/not-found places of residence than for urban residents, whereas specificity was relatively higher for urban residents than for residents of non-urban areas. For example, sensitivity estimates from the LASSO model based on five years of AHD for urban and rural/not found places of residence were 62.2% (95% CI: 55.7-67.6) and 78.8% (95% CI: 71.7, 83.4) respectively. For the same algorithm, specificity estimates were 80.1% (95% CI: 76.8, 83.9) and 57.1% (95% CI: 54.5, 60.9), respectively, for urban and rural/not found places of residence.

3.6 Discussion

Smoking is a risk factor of interest in many disease risk prediction models developed using AHD. (39,40) Valid identification of smoking status may improve the accuracy of risk prediction models. Consequently, it is imperative to conduct comprehensive algorithm validation studies to assess the accuracy of smoking ascertainment in AHD (41). While RBAs for smoking status identification in AHD have been subject to numerous validation studies (12), the validation of MBAs has been limited. Moreover, there is a gap in the literature concerning incorporating multiple AHD sources to construct model-based smoking ascertainment algorithms and using reliable validation data sources to validate these algorithms. Furthermore, the smoking algorithm validation studies rarely compare MBAs with rule-based counterparts in the same cohort. This study used ML model-based methods to develop algorithms for smoking ascertainment in AHD, validated them with a clinical data source, and compared MBAs against RBAs.

3.6.1 Performance of RBAs

In line with previous research (7), we estimated low sensitivity and high specificity for RBAs. Even though we had high NPV and low PPV, those studies had low NPV and high PPV. Our study showed that including a less specific diagnosis code in RBA2 improved sensitivity but significantly reduced PPV compared to RBA1, contrasting with previous research on ever-smoking ascertainment (27). Additionally, we noted that the sensitivity estimates of RBAs rose with greater numbers of years of AHD, while specificity remained consistent, aligning with findings from a previous study (25). While PPV estimates dropped with the rise in the number of years of AHD in the aforementioned study, our PPV estimates remained stable.

3.6.2 Performance of MBAs

The specificity estimates for MBAs surpassed their sensitivity estimates, aligning with previous validation studies of MBA-based ascertainment of current smoking (42) and ever-smoking (43) in claims data. Notably, in our study, the difference between sensitivity and specificity estimates was narrower, with relatively higher sensitivity estimates compared to prior investigations, possibly due to our use of Youden's J index for cut-off probability selection. Our MBAs had balanced accuracy and AUC estimates similar to those from earlier studies. However, our PPV estimates were lower than those of an earlier study on current smoking (42). The important variables (e.g., age and drugs used for addictive disorders) identified by our MBAs mirrored those identified in an ever-smoking ascertainment study. The validity measures were comparable between algorithms based on RF and LASSO models in our study, which differs from a previous study on predicting nicotine addiction (e-cigarette and waterpipe) (44). In that study, RF models showed better performance compared to LASSO models when using the

National Youth Tobacco Survey data from the US. These findings emphasize the necessity of refining MBAs to enhance the accuracy of smoking ascertainment algorithms.

3.6.3 Comparison Between RBAs and MBAs

Our findings demonstrate that MBAs generally outperform RBAs in identifying smokers. Specifically, higher sensitivity and balanced accuracy estimates were observed for MBAs compared to RBAs. Despite this superior performance of MBAs, RBAs demonstrated higher specificity and PPV. These findings contrast with those of a validation study on rheumatoid arthritis, where MBA and RBA exhibited similar validity measure estimates (45). The difference between MBAs and RBAs on the estimates of validity measures is likely attributed to the information incorporated into their construction. In particular, RBAs included diagnosis codes that are potential indicators of smoking (25–27), while MBAs also incorporated sociodemographic characteristics and comorbid conditions in addition to these indicators.

3.6.4 Impact of Number of Years of Administrative Data and Sociodemographic Factors

We observed considerable variations in sensitivity estimates across different numbers of years of AHD, highlighting the importance of considering this attribute of algorithm development. Additionally, sensitivity and specificity estimates varied by biological sex and place of residence, particularly for MBAs. The ML models identified age and income as important variables contributing to MBAs. While specificity estimates improved with sociodemographic variable exclusion from the MBAs, sensitivity estimates dropped substantially. These results underscore the necessity of accounting for sociodemographic factors in algorithm refinement and validation (46).

3.6.5 Study Implications

The findings have a number of implications for further research and public health. Developing MBAs utilizing multiple AHD sources represents a scalable and replicable methodology for enhancing the accuracy of smoking ascertainment in population-based studies. Moreover, incorporating sociodemographic variables into MBAs can improve their capacity to identify disparities in smoking-related behaviors across different population subgroups, providing critical insights for the design and implementation of targeted public health interventions.

3.6.6 Strengths and Limitations

The strengths of this study include the use of a reliable clinical registry data source for algorithm validation, which has been repeatedly used to develop risk prediction models for osteoporosis-related fractures (47,48). Moreover, we compared the performance of RBAs and MBAs for smoking ascertainment in the same cohort. Furthermore, we used three population-based AHD sources for algorithm development.

Some limitations should be considered when interpreting our findings. Firstly, our study focused exclusively on a single registry-based cohort, predominantly comprised of elderly females with specific healthcare needs. This may limit the generalizability of our findings across the entire population. Secondly, the ML models were trained on an imbalanced dataset, potentially introducing bias towards the majority class (49). To circumvent potential inaccuracies stemming from this imbalance, we focused on additional metrics beyond accuracy to evaluate the validity of the algorithms. Thirdly, our algorithms only underwent internal validation, which affects the generalizability of the estimates of algorithm validity measures.

3.6.7 Future Directions

Further validation in external datasets is essential to confirm the robustness and reliability of the MBAs. Alternatively, artificially generated synthetic data that have the statistical properties of real-world validation data might be used to produce augmented training data to more accurately evaluate the performance of MBAs. Previous studies have demonstrated the use of synthetic data to address class imbalance, preserve privacy, and enhance model training. For example, studies have used synthetic data to create balanced datasets that better reflect the population characteristics (50,51). Subsequent research could also leverage the use of longitudinal data to uncover subtle shifts in smoking behavior. For example, a longitudinal model (52) might identify individuals who have recently quit smoking or those who have gradually increased their smoking over time, providing insights into evolving patterns of tobacco use. Including additional data sources that capture variables relevant to smoking, such as hospital outpatient data, could be another way to improve the performance of the developed algorithms.

3.7 Conclusions

By leveraging analytic techniques and incorporating a broad set of data elements, our study demonstrates a promising avenue for improving the accuracy of smoker identification in AHD for population-based studies. However, RBAs achieved high specificity and PPV, highlighting their importance for identifying smoking status. Hence, we suggest adopting a nuanced approach. Rather than exclusively relying on model-based or rule-based methods, we recommend assessing the context based on the specific research questions to determine the most suitable method.

3.8 References

1. Harbaugh CM, Cooper JN. Administrative databases. Seminars in Pediatric Surgery. 2018 Dec 1;27(6):353–60.

2. Barrett JK, Sweeting MJ, Wood AM. Chapter 2 - Dynamic Risk Prediction for Cardiovascular Disease: An Illustration Using the ARIC Study. In: Srinivasa Rao ASR, Pyne S, Rao CR, editors. Handbook of Statistics [Internet]. Elsevier; 2017. p. 47–65. (Disease Modelling and Public Health, Part A; vol. 36). Available from: https://www.sciencedirect.com/science/article/pii/S0169716117300044

3. Tobias MI, Cavana RY, Bloomfield A. Application of a system dynamics model to inform investment in smoking cessation services in New Zealand. American Journal of Public Health. 2010 Jul;100(7):1274–81.

4. McGill E, Er V, Penney T, Egan M, White M, Meier P, et al. Evaluation of public health interventions from a complex systems perspective: A research methods review. Social Science & Medicine. 2021 Mar 1;272:113697.

5. McGowan VJ, Buckner S, Mead R, McGill E, Ronzi S, Beyer F, et al. Examining the effectiveness of place-based interventions to improve public health and reduce health inequalities: an umbrella review. BMC Public Health. 2021 Oct 19;21(1):1888.

6. Lix L, Puchtinger R, Reimer K, Robitaille C, Smith M, Svenson L, et al. The Canadian Chronic Disease Surveillance System: A model for collaborative surveillance. International Journal of Population Data Science [Internet]. 2018 Oct 5;3(3). Available from: https://ijpds.org/article/view/433

7. Desai RJ, Solomon DH, Shadick N, Iannaccone C, Kim SC. Identification of smoking using Medicare data--a validation study of claims-based algorithms. Pharmacoepidemiology and Drug Safety. 2016 Apr;25(4):472–5.

8. Mu Y, Chin AI, Kshirsagar AV, Bang H. Data concordance between ESRD Medical Evidence Report and Medicare claims: Is there any improvement? PeerJ. 2018;6:e5284.

9. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, et al. A clinical text classification paradigm using weak supervision and deep representation. BMC Medical Informatics and Decision Making. 2019 Jan 7;19(1):1.

10. Pang X, Forrest CB, Lê-Scherban F, Masino AJ. Prediction of early childhood obesity with machine learning and electronic health record data. International Journal of Medical Informatics. 2021 Jun 1;150:104454.

11. van Vuuren CL, van Mens K, de Beurs D, Lokkerbol J, van der Wal MF, Cuijpers P, et al. Comparing machine learning to a rule-based approach for predicting suicidal behavior among adolescents: Results from a longitudinal population-based survey. Journal of Affective Disorders. 2021 Dec 1;295:1415–20.

12. Haque MA, Gedara MLB, Nickel N, Turgeon M, Lix LM. The validity of electronic health data for measuring smoking status: A systematic review and meta-analysis. BMC Medical Informatics and Decision Making. 2024 Feb 2;24(1):33.

13. McVeigh KH, Newton-Dame R, Chan PY, Thorpe LE, Schreibstein L, Tatem KS, et al. Can electronic health records be used for population health surveillance? Validating population

health metrics against established survey data. The Journal of Electronic Health Data and Methods. 2016;4(1):1267.

14. Luck J, Larson AE, Tong VT, Yoon J, Oakley LP, Harvey SM. Tobacco use by pregnant Medicaid beneficiaries: Validating a claims-based measure in Oregon. Preventive Medicine Reports. 2020 Sep;19:101039.

15. Kim HM, Smith EG, Stano CM, Ganoczy D, Zivin K, Walters H, et al. Validation of key behaviourally based mental health diagnoses in administrative data: suicide attempt, alcohol abuse, illicit drug abuse and tobacco use. BMC Health Services Research. 2012 Jan 23;12(1):18.

16. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLOS Medicine. 2015 Oct;12(10):e1001885.

17. Smith M, Roos LL, Burchill C, Turner K, Towns DG, Hong SP, et al. Health services data: Managing the data warehouse: 25 years of experience at the Manitoba Centre for Health Policy. In: Levy A, Goring S, Gatsonis C, Sobolev B, van Ginneken E, Busse R, editors. Health Services Evaluation [Internet]. New York, NY: Springer US; 2019. p. 19–45. (Health Services Research). Available from: https://doi.org/10.1007/978-1-4939-8715-3_3

18. WHO. Anatomical Therapeutic Chemical (ATC) Classification [Internet]. 2024. Available from: https://www.who.int/tools/atc-ddd-toolkit/atc-classification

19. Government of Canada SC. Census of Population [Internet]. Available from: https://www12.statcan.gc.ca/census-recensement/index-eng.cfm

20. Leslie WD, MacWilliam L, Lix L, Caetano P, Finlayson GS. A population-based study of osteoporosis testing and treatment following introduction of a new bone densitometry service. Osteoporos Int. 2005 Jul 1;16(7):773–82.

21. Lewiecki EM. Update on bone density testing. Current Osteoporosis Reports. 2005 Dec 1;3(4):136–42.

22. Goel H, Binkley N, Hans D, Leslie WD. Bone density and trabecular bone score to predict fractures in adults aged 20–39 years: a registry-based study. Osteoporosis International. 2023 Jun 1;34(6):1085–91.

23. MCHP Metadata Repository. Term: Personal Health Identification Number (PHIN) [Internet]. Available from: http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=103292

24. MCHP Metadata Repository. Term: Income Quintiles / Income Quintile [Internet]. Available from: http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=102882

25. Havard A, Jorm LR, Lujic S. Risk adjustment for smoking identified through tobacco use diagnoses in hospital data: A validation study. PLOS One. 2014 Apr 15;9(4).

26. Roper L, Tran DT, Einarsdóttir K, Preen DB, Havard A. Algorithm for resolving discrepancies between claims for smoking cessation pharmacotherapies during pregnancy and smoking status in delivery records: The impact on estimates of utilisation. PLOS One. 2018 Aug 30;13(8):e0202999.

27. Marrie RA, Tan Q, Ekuma O, Marriott JJ. Development of an indicator of smoking status for people with multiple sclerosis in administrative data. Multiple Sclerosis Journal – Experimental, Translational and Clinical. 2022;8(1):20552173221074296.

28. tidymodels - Tidymodels packages [Internet]. Available from: https://www.tidymodels.org/packages/

29. Allam A, Nagy M, Thoma G, Krauthammer M. Neural networks versus Logistic regression for 30 days all-cause readmission prediction. Scientific Reports. 2019 Jun 26;9:9277.

30. Uddin S, Imam T, Hossain ME, Gide E, Sianaki OA, Moni MA, et al. Intelligent type 2 diabetes risk prediction from administrative claim data. Informatics for Health and Social Care. 2022 Jul 3;47(3):243–57.

31. Hastie T, Tibshirani R, Friedman J. Random Forests. In: Hastie T, Tibshirani R, Friedman J, editors. The elements of statistical learning: Data mining, inference, and prediction [Internet]. New York, NY: Springer; 2009. p. 587–604. (Springer Series in Statistics). Available from: https://doi.org/10.1007/978-0-387-84858-7_15

32. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58(1):267–88.

33. Feurer M, Hutter F. Hyperparameter Optimization [Internet]. 2018. Available from: https://www.semanticscholar.org/paper/Chapter-1-Hyperparameter-Optimization-Feurer-Hutter/1a90075b632c8cb7762a337431f143c6fe682bad

34. Khan Z, Gul A, Perperoglou A, Miftahuddin M, Mahmoud O, Adler W, et al. Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification. 2020 Mar 1;14(1):97–116.

35. Kuhn M. Building predictive models in R using the caret package. Journal of Statistical Software. 2008 Nov 10;28:1–26.

36. Fluss R, Faraggi D, Reiser B. Estimation of the youden index and its Associated cutoff Point. Biometrical Journal. 2005;47(4):458–72.

37. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, et al. Foundational statistical principles in medical research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. Medicina (Kaunas). 2021 May 16;57(5):503.

38. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. 1988;44(3):837–45.

39. Teramukai S, Okuda Y, Miyazaki S, Kawamori R, Shirayama M, Teramoto T. Dynamic prediction model and risk assessment chart for cardiovascular disease based on on-treatment blood pressure and baseline risk factors. Hypertension Research. 2016 Feb;39(2):113–8.

40. Chang JT, Meza R, Levy DT, Arenberg D, Jeon J. Prediction of COPD risk accounting for time-varying smoking exposures. PLOS One. 2021 Mar 10;16(3):e0248535.

41. Hoeven LR van, Bruijne MC de, Kemper PF, Koopman MMW, Rondeel JMM, Leyte A, et al. Validation of multisource electronic health record data: An application to blood transfusion data. BMC Medical Informatics and Decision Making. 2017 Jul 14;17(1):107.

42. Reps JM, Rijnbeek PR, Ryan PB. Supplementing claims data analysis using self-reported data to develop a probabilistic phenotype model for current smoking status. Journal of Biomedical Informatics. 2019 Sep;97:103264.

43. Ali S, Na R, Waterhouse M, Jordan SJ, Olsen CM, Whiteman DC, et al. Predicting obesity and smoking using medication data: A machine-learning approach. Pharmacoepidemiology and Drug Safety. 2022 Jan;31(1):91–9.

44. Choi J, Jung HT, Ferrell A, Woo S, Haddad L. Machine learning-based nicotine addiction prediction models for youth e-cigarette and waterpipe (Hookah) users. Journal of Clinical Medicine. 2021 Mar 2;10(5):972.

45. Kroeker K, Widdifield J, Muthukumarana S, Jiang D, Lix LM. Model-based methods for case definitions from administrative health data: Application to rheumatoid arthritis. BMJ Open. 2017 Jun 23;7(6):e016173.

46. Tate C, Kumar R, Murray JM, Sanchez-Franco S, Montgomery SC, Montes F, et al. Socio-environmental and psychosocial predictors of smoking susceptibility among adolescents with contrasting socio-cultural characteristics: A comparative analysis. BMC Public Health. 2021 Dec 9;21(1):2240.

47. Leslie WD, Morin SN, Martineau P, Bryanton M, Lix LM. Factors associated with bone density monitoring while on antiosteoporosis treatment in routine clinical practice: A registry-based cohort study. Journal of Clinical Densitometry. 2020 Oct 1;23(4):568–75.

48. Agarwal A, Baleanu F, Moreau M, Charles A, Iconaru L, Surquin M, et al. External validation of FRISBEE 5-year fracture prediction models: A registry-based cohort study. Archives of Osteoporosis. 2022 Dec 23;18(1):13.

49. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: Bagging, Boosting, and Hybrid-Based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2012 Jul;42(4):463–84.

50. Pezoulas VC, Zaridis DI, Mylona E, Androutsos C, Apostolidis K, Tachos NS, et al. Synthetic data generation methods in healthcare: A review on open-source tools and methods. Computational and Structural Biotechnology Journal. 2024 Dec 1;23:2892–910.

51. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns. 2024 Apr 12;5(4):100946.

52. Feely A, Lim LS, Jiang D, Lix LM. A population-based study to develop juvenile arthritis case definitions for administrative health data using model-based dynamic classification. BMC Medical Research Methodology. 2021 May 16;21:105.

3.9 SUPPLEMENTARY MATERIAL

Additional file 1

Supplemental Table 3.1 The Reporting of studies Conducted using Observational Routinely-collected Data (RECORD) statement – checklist of items, extended from the Strengthening the Reporting of Observational Studies in Epidemiology (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	87-89	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.	87-89
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	90-91		
Objectives	3	State specific objectives, including any prespecified hypotheses	91		
Methods					
Study Design	4	Present key elements of study design early in the paper	91-92		

Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	91-93		

Participants	6	*(a) Cohort study* - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *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	91-93	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If 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.	91-93
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	93-94, Additional file 2	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.	93-94, Additional file 2

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	93		

Bias	9	Describe any efforts to address potential sources of bias	N/A		
Study size	10	Explain how the study size was arrived at	98; Figure 3.1		
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	93-99		
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	99-101		

Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population.	91-93
				RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	
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.	92
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) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram	98, Figure 3.1	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.	98, Figure 3.1
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)	99-104, Table 3.1		

Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report numbers in each exposure	99-104, Table 3.1		
		category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures			
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	101-109, Figure 3.2, Figure 3.3, Table 3.2		
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	Additional file 3, Additional file 4, Additional file 5,		
Discussion					
Key results	18	Summarise key results with reference to study objectives	109-113		

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	112	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.	111-113
Interpretation	20	Give a cautious overall interpretation of results considering objectives,	112-114		
		limitations, multiplicity of analyses, results from similar studies, and other relevant evidence			
Generalisability	21	Discuss the generalisability (external validity) of the study results	113		
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	iii		
Accessibility of protocol, raw data, and programming code		..		RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	242

Additional file 2

Supplemental Table 3.2 List of indicators used to develop rule-based and model-based algorithms for current smoking status by data source

Name	Description	Data source
Diagnosis of nicotine dependence*	1=Presence of ICD-10-CA code F17.2; 0=Else	Hospital Abstracts
Diagnosis of problems related to lifestyle: tobacco use*	1=Presence of ICD-10-CA code Z72.0; 0=Else	Hospital Abstracts
Diagnosis of nicotine dependence*	1=Presence of ICD-9-CM code 305.1; 0=Else	Medical Claims
Diagnosis of personal history presenting hazards to health*	1=Presence of ICD-9-CM code V15.82; 0=Else	Medical Claims
Diagnosis of chronic obstructive pulmonary disease (COPD)	1=Presence of at least one code from ICD-10-CA codes: J41 (Simple chronic bronchitis), J42 (Unspecified chronic bronchitis), J43 (Emphysema), J44 (Other COPD); 0=Else	Hospital Abstracts
Diagnosis of COPD	1=Presence of at least one code from ICD-9-CM codes: 490 (Bronchitis, not specified as acute or chronic), 491 (Chronic bronchitis), 492 (Emphysema), 496 (Chronic airway obstruction); 0=Else	Medical Claims
Diagnosis of asthma	1=Presence of ICD-10-CA code J45 (Asthma); 0=Else	Hospital Abstracts
Diagnosis of asthma	1=Presence of ICD-9-CM code 493 (Asthma); 0=Else	Medical Claims
Diagnosis of the use of respiratory system treatment drug	1=Presence of ATC code R03AC02 (Salbutamol); 0=Else	Prescription Drug Records

Diagnosis of the use of medication primarily recommended as an antidepressant and for smoking cessation*	1=Presence of ATC code N06AX12 (Bupropion); 0=Else	Prescription Drug Records
Diagnosis of the use of medication to quit smoking*	1=Presence of ATC code N07BA03 (Varenicline); 0=Else	Prescription Drug Records
Diagnosis of the use of medication to aid in smoking cessation*	1=Presence of ATC code B07BA01 (Nicotine); 0=Else	Prescription Drug Records
Urban place of residence	1=urban place of residence; 0=rural place of residence	Health Insurance Registry
Income quintile 2	1=Belongs to income quintile group 2; 0=Belongs to income quintile group 1 (lowest)/ not found	Statistics Canada Census
Income quintile 3	1=Belongs to income quintile group 3; 0=Belongs to income quintile group 1/ not found	Statistics Canada Census
Income quintile 4	1=Belongs to income quintile group 4; 0=Belongs to income quintile group 1/ not found	Statistics Canada Census
Income quintile 5 (highest)	1=Belongs to income quintile group 5; 0=Belongs to income quintile group 1/ not found	Statistics Canada Census
Biological sex	1=Female; 0=Male	Health Insurance Registry
Age	Age on index date in years	Health Insurance Registry

Note: *Represents indicators used in rule-based algorithms

Additional file 3

Supplemental Table 3.3 Area under the curve (AUC) of the receiver operating characteristic curve estimates (95% confidence intervals [CI]) for model-based algorithms by the number of years of administrative health data used for independent variable identification (%)

Number of years of administrative data before cohort index date	Algorithm	AUC estimate (95% CI)	Difference (95% CI)
1	LASSO	72.5 (70.5, 74.5)	**-1.3** (-2.2, -0.05)
	RF	73.8 (71.8, 75.8)	
2	LASSO	73.9 (71.9, 75.8)	**-1.2** (-1.9, -0.05)
	RF	75.1 (73.1, 77.1)	
3	LASSO	75.1 (73.1, 77.0)	**-1.1** (-1.7, -0.02)
	RF	76.0 (74.1, 78.0)	
4	LASSO	76.6 (74.6, 78.5)	**-0.8** (-1.5, -0.1)
	RF	77.4 (75.4, 79.3)	
5	LASSO	77.9 (76.1, 79.9)	-0.8 (-1.4, 0.0)
	RF	78.7 (76.8, 80.6)	

Note: boldface font indicates the difference is statistically significant (P value from DeLong's test < 0.05); Least Absolute Shrinkage and Selection Operator (LASSO); Random forest (RF)

Additional file 4

Supplemental Table 3.4 Sex-stratified validity measure estimates (95% confidence intervals) for algorithms by number of years of administrative health data used for independent variable identification

Validity measure	Algorithm	Number of years of administrative data before cohort index date									
		1		2		3		4		5	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Sensitivity	RBA1	7.9 (5.9, 10.3)	5.6 (2.1, 11.7)	12.0 (9.6, 14.8)	10.2 (5.2, 17.5)	16.1 (13.1, 19.2)	13.0 (7.3, 20.8)	21.0 (17.9, 24.4)	16.7 (10.2, 25.1)	24.1 (20.1, 27.7)	18.5 (11.7, 27.1)
	RBA2	10.4 (8.1, 13.0)	5.6 (2.1, 11.7)	15.5 (12.7, 18.5)	10.2 (5.2, 17.5)	20.3 (19.2, 25.0)	13.0 (7.3, 20.8	25.2 (21.9, 28.8)	16.7 (10.2, 25.1)	28.7 (25.2, 30.1)	19.4 (16.3, 28.6)
	RF	59.9 (56.0, 63.8)	85.2 (77.1, 91.3)	64.0 (60.2, 67.8)	83.3 (74.9, 89.8)	60.2 (56.3, 64.1)	75.9 (66.7, 83.6)	62.3 (58.4, 66.1)	75.9 (66.7, 88.6)	65.1 (64.6, 69.1)	76.8 (75.2, 79.0)
	LASSO	59.4 (55.5, 63.3)	75.0 (65.7, 82.8)	68.6 (64.8, 72.2)	78.7 (69.8, 86.0)	64.5 (60.6, 68.2)	72.2 (62.8, 80.4)	64.8 (61.0, 68.5)	71.3 (61.8, 79.6)	68.1 (65.7, 69.8)	70.3 (69.7, 73.5)
Specificity	RBA1	99.8 (99.6, 99.9)	99.6 (98.8, 99.9)	99.7 (99.5, 99.8)	99.3 (98.4, 99.8)	99.5 (99.2, 99.6)	98.9 (97.9, 99.5)	99.3 (99.1, 99.5)	98.8 (97.7, 99.4)	99.0 (98.7, 99.3)	98.2 (97.0, 99.1)
	RBA2	98.6 (98.3, 98.9)	98.9 (97.9, 99.5)	98.1 (97.7, 98.5)	98.1 (96.8, 99.0)	97.9 (96.5, 98.9)	97.6 (96.2, 98.5)	97.2 (96.7, 97.6)	97.2 (95.7, 98.2)	96.6 (96.3, 97.1)	96.3 (95.2, 96.9)
	RF	77.7 (76.7, 78.8)	47.0 (43.3, 50.6)	75.3 (74.2, 76.4)	48.4 (44.8, 52.1)	81.1 (80.1, 82.1)	57.5 (53.8, 61.1)	80.9 (79.8, 81.9)	58.2 (54.5, 61.8)	80.6 (79.1, 81.5)	56.6 (54.3, 57.8)
	LASSO	75.5 (74.4, 76.6)	56.2 (52.5, 59.8)	68.3 (67.1, 69.4)	54.0 (50.3, 57.6)	74.7 (73.6, 75.8)	58.3 (54.7, 61.9)	76.6 (75.5, 77.7)	60.2 (56.6, 63.8)	75.5 (73.8, 77.9)	59.4 (57.5, 60.1)
PPV	RBA1	80.5 (68.9, 88.5)	66.7 (33.7, 88.7)	79.9 (70.8, 86.7)	68.7 (43.8, 86.1)	76.0 (68.2, 82.4)	63.6 (42.9, 80.3)	77.2 (70.5, 82.7)	66.7 (48.0, 81.3)	72.5 (66.3, 77.9)	60.6 (44.1, 75.0)
	RBA2	44.1 (36.6, 51.9)	42.8 (21.0, 68.0)	46.7 (40.4, 53.2)	44.0 (26.8, 62.8)	47.6 (42.1, 53.2)	43.7 (28.5, 60.3)	48.9 (43.9, 53.9)	46.1 (32.1, 60.9)	47.6 (42.4, 50.9)	43.7 (40.5, 47.3)
	RF	22.2 (20.9, 23.6)	19.0 (17.5, 20.7)	21.6 (20.4, 22.9)	19.1 (17.5, 20.9)	25.3 (23.8, 26.9)	20.7 (18.6, 23.0)	25.7 (24.2, 27.2)	21.0 (18.8, 23.3)	26.3 (22.3, 27.1)	20.5 (18.9, 21.6)
	LASSO	20.6 (19.2, 21.8)	20.0 (17.9, 22.3)	18.7 (17.7, 19.7)	20.0 (18.1, 22.1)	21.3 (20.1, 22.6)	20.2 (18.0, 22.6)	22.8 (21.5, 24.1)	20.7 (18.4, 23.3)	22.9 (21.7, 26.2)	20.2 (17.5, 21.0)
NPV	RBA1	91.1	87.8	91.4	88.3	91.8	88.6	92.2	89.0	92.4	89.2

		(90.9, 91.3)	(87.3, 88.3)	(91.2, 91.6)	(87.6, 89.0)	(91.5, 92.0)	(87.5, 89.3)	(91.9, 92.5)	(88.2, 89.8)	(92.1, 92.7)	(88.3, 90.0)
	RBA2	91.2 (90.9, 91.4)	87.7 (87.2, 88.2)	91.6 (91.3, 91.9)	88.2 (87.5, 88.8)	92.0 (91.7, 92.3)	88.5 (87.7, 89.2)	92.4 (92.1, 92.8)	88.9 (88.0, 89.7)	92.7 (92.2, 93.3)	89.1 (88.0, 90.0)
	RF	94.8 (94.3, 95.3)	95.6 (93.2, 97.2)	95.2 (94.7, 95.6)	95.2 (92.8, 96.8)	95.1 (94.6, 95.5)	94.2 (92.1, 95.8)	95.3 (94.8, 95.7)	94.3 (92.2, 95.9)	95.7 (94.9, 95.9)	94.3 (94.1, 95.2)
	LASSO	94.6 (94.1, 95.1)	93.9 (91.7, 95.5)	95.3 (94.8, 95.8)	94.5 (92.3, 96.2)	95.2 (94.7, 95.7)	93.5 (91.3, 95.1)	95.3 (94.9, 95.8)	93.5 (91.4, 95.1)	95.6 (95.0, 96.0)	93.2 (92.9, 94.0)
Balanced accuracy	RBA1	53.8 (52.1, 57.6)	52.5 (49.0, 53.9)	55.8 (52.7, 56.5)	54.7 (52.0, 55.2)	57.8 (56.3, 59.4)	55.9 (54.1, 57.3)	60.2 (59.0, 62.2)	57.7 (55.6, 59.4)	53.5 (50.0, 54.9)	58.4 (57.0, 60.3)
	RBA2	54.5 (52.3, 56.2)	52.2 (49.4, 54.1)	56.8 (54.8, 58.4)	54.1 (51.4, 54.6)	58.1 (56.7, 59.9)	55.3 (53.4, 56.7)	61.2 (60.0, 63.0)	56.9 (55.7, 58.5)	54.1 (52.6, 56.3)	57.9 (55.9, 8.7)
	RF	68.8 (66.7, 70.1)	66.1 (63.3, 69.1)	69.7 (68.9, 71.0)	65.9 (63.0, 68.6)	70.7 (69.1, 72.0)	66.7 (65.8, 68.3)	71.5 (70.3, 73.4)	67.1 (64.6, 68.8)	72.8 (71.2, 75.6)	66.7 (64.3, 67.9)
	LASSO	67.5 (64.9, 69.0)	65.6 (62.7, 68.3)	68.4 (67.0, 70.0)	66.3 (63.7, 68.8)	69.6 (68.0, 71.1)	65.3 (62.4, 67.9)	70.7 (69.1, 70.9)	65.7 (62.8, 68.3)	71.8 (69.7, 73.0)	64.9 (62.8, 66.3)

Note: Rule-based algorithm (RBA); Random forest (RF); Least Absolute Shrinkage and Selection Operator (LASSO)

Additional file 5

Supplemental Table 3.5 Place of residence-stratified validity measure estimates (95% confidence intervals) for algorithms by number of years of administrative health data used for independent variable identification

Validity measure	**Algorithm**	**Number of years of administrative data before cohort index date**									
		1		2		3		4		5	
		Urban	Rural/ Not found	Urban	Rural/ Not found	Urban	Rural/ Not found	Urban	Rural/ Not found	Urban	Rural/ Not found
Sensitivity	RBA1	9.7 (5.5, 16.7)	7.9 (4.7, 11.1)	11.8 (8.9, 14.8)	11.5 (7.7, 15.3)	15.3 (13.1, 19.2)	16.2 (7.1, 19.7)	19.6 (16.8, 23.2)	21.6 (16.3, 27.1)	22.4 (20.1, 27.7)	24.8 (19.6, 32.2)
	RBA2	10.4 (8.1, 13.0)	9.7 (5.6, 16.8)	15.1 (12.1, 17.9)	14.3 (9.2, 25.4)	19.4 (15.1, 23.2)	19.6 (10.2, 27.7	23.9 (21.9, 28.8)	24.1 (20.2, 35.1)	27.1 (24.2, 29.1)	27.7 (21.3, 37.5)
	RF	57.8 (50.2, 61.8)	73.3 (67.1, 81.3)	61.2 (57.2, 65.6)	76.2 (64.8, 80.1)	57.3 (50.3, 62.1)	71.2 (61.7, 78.5)	59.1 (54.4, 64.1)	73.0 (65.7, 84.6)	62.5 (56.6, 69.0)	74.1 (69.2, 81.0)
	LASSO	53.4 (54.5, 62.3)	75.5 (66.2, 83.3)	62.5 (57.8, 67.2)	82.7 (73.8, 89.0)	57.9 (52.6, 62.3)	78.4 (67.7, 84.3)	58.4 (53.0, 65.5)	78.1 (67.8, 86.5)	62.2 (55.7, 67.6)	78.8 (71.7, 83.4)
Specificity	RBA1	99.8 (99.6, 99.9)	99.7 (99.4, 99.9)	99.6 (99.4, 99.8)	99.6 (99.3, 99.9)	99.4 (99.1, 99.5)	99.4 (99.1, 99.5)	99.3 (99.1, 99.5)	99.3 (99.1, 99.5)	98.9 (98.7, 99.3)	99.1 (97.2, 99.8)
	RBA2	98.6 (98.3, 98.9)	98.8 (97.8, 99.4)	98.0 (97.7, 98.5)	98.3 (96.8, 99.0)	97.5 (96.1, 98.5)	97.9 (96.4, 98.7)	97.1 (96.6, 97.5)	97.4 (95.9, 98.4)	96.5 (96.2, 97.1)	96.8 (95.7, 97.6)
	RF	79.6 (70.5, 83.2)	61.6 (53.3, 70.1)	78.4 (75.2, 84.4)	57.7 (43.8, 62.2)	83.7 (79.1, 87.1)	65.9 (60.8, 71.2)	81.3 (79.5, 88.6)	66.7 (54.5, 71.0)	82.7 (77.1, 91.0)	66.0 (59.3, 73.8)
	LASSO	82.5 (78.3, 86.6)	52.4 (50.5, 57.8)	75.8 (68.1, 79.4)	44.6 (40.3, 52.6)	80.5 (77.6, 85.8)	54.7 (50.5, 59.9)	76.6 (75.5, 77.7)	59.1 (52.6, 65.8)	80.1 (76.8, 83.9)	57.1 (54.5, 60.9)
PPV	RBA1	79.1 (66.9, 91.2)	78.6 (63.4, 93.8)	77.4 (67.7, 87.1)	80.0 (67.6, 92.4)	71.7 (68.2, 82.4)	78.9 (72.9, 90.1)	77.2 (70.5, 82.7)	81.1 (74.6, 87.0)	66.2 (66.3, 77.9)	79.3 (65.1, 85.3)
	RBA2	39.8 (32.6, 45.7)	54.0 (29.9, 70.1)	43.2 (37.4, 51.2)	54.2 (21.8, 71.8)	43.2 (34.1, 49.2)	56.4 (38.4, 69.3)	44.5 (39.9, 51.7)	57.7 (42.1, 70.7)	43.1 (40.4, 51.1)	55.7 (50.5, 59.9)
	RF	21.8 (18.9, 26.6)	21.3 (19.5, 26.7)	21.6 (20.4, 22.9)	20.3 (16.5, 25.9)	25.7 (23.8, 27.9)	22.8 (20.6, 26.8)	25.7 (24.2, 27.2)	23.7 (19.8, 26.9)	26.3 (22.3, 27.1)	23.6 (20.9, 26.9)
	LASSO	22.6 (21.2, 22.8)	18.4 (16.9, 21.8)	20.3 (17.7, 22.3)	17.5 (14.1, 19.1)	22.6 (20.1, 26.1)	19.7 (16.0, 21.6)	23.5 (21.0, 26.7)	21.3 (19.4, 23.9)	24.1 (22.7, 28.1)	20.7 (18.0, 22.1)

NPV	RBA1	91.6 (90.8, 92.4)	88.4 (87.1, 89.7)	91.9 (91.2, 92.7)	88.9 (87.5, 90.1)	92.3 (91.5, 92.0)	89.3 (87.2, 90.2)	92.6 (91.8, 93.0)	89.9 (88.6, 90.7)	92.3 (92.0, 92.6)	90.2 (87.3, 89.0)
	RBA2	91.7 (90.9, 92.4)	88.5 (87.0, 89.2)	92.1 (90.5, 93.4)	89.0 (88.5, 89.8)	92.5 (92.1, 93.2)	89.5 (86.7, 88.2)	92.8 (92.3, 93.6)	90.0 (89.0, 91.0)	92.7 (92.2, 93.3)	90.4 (89.1, 90.9)
	RF	95.0 (94.0, 96.0)	94.2 (93.9, 95.0)	95.2 (94.7, 95.6)	94.4 (92.3, 95.8)	95.2 (94.7, 95.6)	94.1 (92.0, 95.7)	95.4 (94.9, 95.8)	94.5 (92.3, 96.0)	95.7 (94.9, 95.9)	94.7 (94.5, 95.6)
	LASSO	94.7 (94.2, 95.2)	93.7 (92.7, 94.5)	95.3 (94.8, 95.8)	94.7 (92.5, 96.4)	95.1 (94.6, 95.6)	94.7 (92.3, 95.8)	95.2 (94.8, 95.9)	95.0 (94.0, 96.0)	95.6 (95.0, 96.0)	95.0 (93.9, 95.8)
Balanced accuracy	RBA1	53.1 (52.0, 56.5)	52.5 (49.0, 53.9)	55.7 (52.6, 56.4)	54.7 (55.5, 59.1)	57.3 (56.0, 59.1)	55.9 (54.1, 57.3)	59.4 (57.0, 62.1)	60.4 (58.6, 65.4)	60.6 (53.5, 66.9)	61.9 (59.0, 70.3)
	RBA2	54.1 (48.3, 60.0)	54.3 (49.2, 59.1)	56.6 (54.9, 61.9)	56.2 (54.7, 59.9)	58.4 (53.7, 60.9)	58.5 (53.9, 62.7)	60.5 (57.1, 62.3)	60.8 (57.7, 63.5)	61.8 (58.3, 69.3)	62.3 (57.9, 68.7)
	RF	68.6 (66.9, 70.3)	67.5 (62.3, 71.1)	69.8 (64.8, 71.0)	66.9 (62.0, 69.6)	70.5 (68.1, 73.0)	68.5 (61.8, 70.3)	71.1 (60.2, 73.0)	69.8 (64.1, 72.8)	72.6 (71.0, 75.4)	70.0 (68.3, 75.8)
	LASSO	67.7 (65.1, 69.2)	63.9 (60.6, 67.3)	69.2 (65.0, 72.2)	63.6 (59.7, 67.8)	69.2 (66.5, 71.1)	66.5 (60.4, 68.8)	69.9 (69.0, 71.9)	68.5 (63.8, 71.6)	71.5 (69.4, 72.9)	67.9 (65.8, 70.3)

Note: Rule-based algorithm (RBA); Random forest (RF); Least Absolute Shrinkage and Selection Operator (LASSO)

CHAPTER 4. LEVERAGING SYNTHETIC DATA TO MITIGATE COHORT IMBALANCE IN DIAGNOSTIC ALGORITHM VALIDATION STUDIES FOR ADMINISTRATIVE HEALTH DATA

4.1 Overview

This chapter addresses the challenges of cohort imbalance in the validation of diagnostic algorithms using administrative health data, with a particular focus on identifying smoking status. The analysis centers on the use of synthetic data to improve representativeness within a clinical registry-derived validation cohort, thereby enhancing the generalizability of algorithm validity across demographic subgroups. Through a comparative evaluation of synthetic data generation models applied to underrepresented population subgroups, specifically males and younger adults, the chapter explores how training data augmentation for model-based algorithms can influence key performance metrics such as sensitivity, specificity, and positive predictive value. The findings demonstrate that different synthetic data generation models have distinct effects on algorithm performance, identifying trade-offs and highlighting the relevance of methodological choices in shaping validation outcomes. By integrating perspectives from epidemiology and machine learning, this study contributes to ongoing efforts to improve the generalizability and fairness of diagnostic algorithms for routinely collected data. It highlights the significance of cohort composition in shaping model-based algorithm performance and suggests that synthetic data, when applied thoughtfully, can serve as a valuable tool for mitigating bias and improving subgroup equity in health data research.

Publication Details

Haque MA, Nickel N, Turgeon M, Lix LM. Leveraging synthetic data to mitigate cohort imbalance in diagnostic algorithm validation studies for administrative health data. Under review, *JMIR Medical Informatics. 2025.*

4.2 Abstract

Background: While administrative health data (AHD) are a valuable resource for healthcare system management, using AHD for research requires validation to ensure data accuracy. Diagnostic algorithm validation studies assess the accuracy of case definitions to identify health conditions and behaviors for AHD. Imbalances in the number of cases or non-cases in the validation cohort by covariates can affect algorithm generalizability.

Objective: We assessed the effect of data augmentation on the performance of a diagnostic algorithm to identify smoking status in AHD using different synthetic data generation models. We compared algorithm validity estimates across augmented and non-augmented training datasets for underrepresented groups in the validation cohort.

Methods: Population-based AHD for 2012 to 2020 were linked with clinical registry data that contained current smoking status. Cohort members were predominantly older females, leading to an underrepresentation of males and younger adults. A model-based algorithm for smoking status identification was developed using Least Absolute Shrinkage and Selection Operator logistic regression. The algorithm's performance was evaluated in a test dataset including 30% of cohort members. Synthetic data to augment the validation cohort were generated using Synthetic Minority Oversampling Technique (SMOTE), Conditional Tabular Generative Adversarial Networks (CTGAN), and Classification and Regression Trees (CART). Sensitivity, specificity, positive and negative predictive values (PPV and NPV) with 95% confidence intervals (CIs) were estimated for the algorithm trained on augmented and non-augmented data. Fairness of the algorithm for demographic subgroups was evaluated using the ratio of validity estimates.

Results: The cohort consisted of 24,718 individuals with a mean age of 67.0 years who were predominantly female (88.6%). SMOTE for algorithm training data augmentation produced the highest specificity and PPV for underrepresented males and younger adults, while CTGAN produced the highest sensitivity in the test dataset. SMOTE increased specificity to 84.3% (95% CI: 67.8-90.8) and PPV to 26.7% (95% CI: 19.0-28.2) among males. In contrast, the use of non-augmented training data showed lower specificity at 59.4% (95% CI: 20.3-26.5) and a PPV of 20.2% (95% CI: 17.9-22.7). Additionally, while CTGAN improved sensitivity, CART exhibited improvements across sensitivity, specificity, and PPV for males compared to non-augmented training data. For individuals aged 18-45 years, specificity and PPV were less than 15.0% for non-augmented training data; when SMOTE was used to augment training data, specificity increased to 80.9% (95% CI: 69.3-88.8) and PPV to 25.4% (95% CI: 22.2-30.0). NPV remained consistently high (≥90%) for augmented and non-augmented training data. SMOTE resulted in greater fairness than CTGAN and CART, with validity estimate ratios for the former near one for sex and age group.

Conclusions: This study demonstrates the value of synthetic data to augment AHD validation cohorts by addressing imbalances arising from covariate distribution. However, trade-offs among algorithm evaluation metrics require context-specific consideration in validation studies using synthetic data.

4.3 Background

Validation studies are frequently conducted to assess the accuracy of algorithms (i.e., rules) to identify individuals with specific health conditions or behaviors in administrative health data (AHD) [1]. Numerous studies have evaluated the validity of AHD for identifying individuals with chronic conditions, infectious diseases, and behavioral risk factors in different populations and healthcare settings [2–7]. Validation studies are important because AHD are primarily designed to support healthcare system management rather than research, potentially leading to biased or erroneous study findings without thorough validation [8].

Clinical registries and surveys are primary data sources for AHD validation study cohort formation [3,9]. However, these data sources may under- or over-represent certain groups or categories of individuals relative to their appearance in the general population [10]. These imbalances can impact the accuracy and generalizability of the results from AHD validation studies [11].

Methods originally developed to address imbalances in binary response variables for machine learning models [12] can be repurposed to mitigate the effects of covariate imbalances in validation studies. These methods include pre-analysis rebalancing, algorithmic adjustments during analysis, and post-hoc adjustments after analysis [13]. Pre-analysis rebalancing aims to create datasets that reflect the population by correcting cohort imbalances through data augmentation. Commonly used methods include undersampling, oversampling, and the Synthetic Minority Oversampling Technique (SMOTE), which adjusts sample proportions to enhance representativeness [14]. Synthetic data generation methods for data augmentation, such as generative adversarial networks (GAN) and tree-based synthesizers, also offer rebalancing options and may be useful when cohorts are small in size or severely imbalanced [15]. During

analysis, algorithmic methods, such as ensemble-based and cost-sensitive learning, mitigate imbalance by incorporating penalties that prevent models from overfitting to overrepresented groups [16,17]. However, rebalancing data before analysis provides a simpler way to achieve balance, especially in AHD validation studies, by ensuring cohort representativeness and easier evaluation, unlike algorithmic methods that integrate imbalance correction during training and require additional parameter estimation. When imbalances cannot be resolved at earlier stages, as a last resort, post-hoc adjustment methods provide a corrective measure after analysis. For example, bootstrap resampling has improved effect size estimation in gene-based rare variant studies [18].

Synthetic data generation for augmentation involves creating additional observations to increase the dataset size [19]. In recent years, tabular health data-based studies have used rebalancing with synthetic data generation for data augmentation to address the imbalance. Studies have shown that rebalancing methods applied to clinical datasets and AHD can enhance predictive model performance [20]. Specifically, these methods mitigate data scarcity, which occurs when a limited number of cohort members from certain groups impair effective model training [21]. However, imbalance is rarely addressed in AHD validation studies of diagnostic algorithms despite the potential benefits of a balanced dataset in improving the accuracy and generalisability of findings across population subgroups.

This study assessed how data augmentation impacts the validity of a smoking status identification algorithm for AHD, using a clinical registry cohort for validation. The objectives were to (1) compare algorithm validity estimates across augmented and non-augmented training

datasets for underrepresented groups in the validation cohort and (2) evaluate the effectiveness of different synthetic data generation models in improving these estimates.

4.4 Methods

We conducted a validation study using a retrospective cohort design. AHD were linked to clinical registry data from April 1, 2012, to March 31, 2020. We followed the Reporting of Studies Conducted Using Observational Routinely-collected Data guidelines [22] (Supplementary file 1).

4.4.1 Data Sources

Study data were from the Population Research Data Repository housed at the Manitoba Centre for Health Policy (MCHP) at the University of Manitoba [23] in the province of Manitoba, Canada. Manitoba has a system of universal healthcare, which means the AHD sources contain information on everyone who gets service from this provincial system. The province has a total population of 1.3 million, according to the most recent Statistics Canada Census.

The Manitoba bone mineral density (BMD) clinical registry [24] was the validation data source; it contains high-quality information about smoking status. A BMD test [25] is a diagnostic technique used to identify individuals vulnerable to fractures from diminished bone density. It is usually advised for postmenopausal women, older adults, individuals with specific medical conditions or medications that impact bone health, and individuals with a history of fractures. The data on smoking status documented in the registry was obtained directly from patients during clinical visits. Smoking status information has been used previously to construct covariates for fracture risk prediction models [26].

Other study data sources included the Manitoba Health Insurance Registry, hospital

discharge abstracts, physician billing claims, prescription medication dispensations, and the Statistics Canada Census. These data sources were used to generate synthetic data and develop algorithm to identify smoking status in AHD. The Manitoba Health Insurance Registry contains individual-level demographic and residential information for everyone registered to receive health services in Manitoba. Hospital abstracts encompass records from hospital stays, while medical claims include documentation of physician visits in both offices and outpatient departments. Diagnoses in hospital abstracts and physician billing claims are coded using the International Classification of Diseases (ICD) of the World Health Organization (WHO). Specifically, hospital abstracts use the 10th revision with Canadian adaptations (ICD-10-CA), and physician billing claims use the 9th revision with clinical modifications (ICD-9-CM). Prescription medication dispensations from community pharmacies are captured in the Drug Program and Information Network (DPIN) database. The WHO's Anatomical Therapeutic Chemical Classification System [27] codes for classifying pharmaceutical substances are used to capture information in the DPIN database. The Statistics Canada Census [28] is a nationwide population-based aggregated data source used to construct area-level income quintiles.

4.4.2 Population and Validation Cohort

The target population for this study consisted of adults (≥18 years) captured in the Manitoba Health Insurance Registry from April 1, 2012, to March 31, 2020. The validation cohort included adults in the BMD registry with a test date between April 1, 2017, and March 31, 2020. BMD registry records were linked to AHD using a scrambled personal health identification number [29].

4.4.3 Response Variable

The response variable was self-reported, clinically-confirmed current smoking status from the BMD registry. This information was recorded from the patient's response to the closed question, "Do you currently smoke? (1=Yes, 2=No)". The BMD test date was the study index date.

4.4.4 Indicators of Smoking in AHD

Diagnosis codes for smoking-related conditions, such as a history of tobacco use, were identified from hospital discharge abstracts and physician billing claims (Supplementary file 2). Prescriptions for nicotine dependence drugs were identified from DPIN data. Along with defining health insurance coverage (Supplementary file 3), the Manitoba Health Insurance Registry was used to obtain age (in years), sex, and place of residence (urban, rural). Income quintile, measured at the level of a dissemination area (consisting of 400-700 persons) based on total household income, was obtained from the Statistics Canada Census using postal code of residence from the Insurance Registry [30]. All diagnosis and medication code-based indicators were constructed using five years of AHD before the index date [31]. The indicators (i.e., diagnoses related to nicotine dependence, chronic obstructive pulmonary disease, asthma, and smoking cessation drugs; sex, place of residence, age, and income quintiles) were used in synthetic data generation for data augmentation and as covariates in the smoking status algorithm.

4.4.5 Synthetic Data Generation

The validation cohort was randomly split into a training set (70% of the cohort) and a test set (30% of the cohort), ensuring that the prevalence of current smoking was consistent in both sets. The synthetic data were generated so that when integrated with the non-augmented training data, the augmented training dataset reflected the Manitoba population distribution of sex (male,

female) and age group (18 years to the mean age of the population, above the mean age of the population). Training data were augmented separately for the marginal and joint distributions of sex and age group. These two variables were selected because they influence smoking behavior and related health outcomes [32] and, most importantly, introduce imbalance in the study cohort.

The frequency of synthetic observations for an under-represented group was determined considering sex and age group distributions in the population and non-augmented training data [33]. To illustrate, if the male-to-female ratio in the population was 1.0 (meaning perfectly balanced with 50% male and 50% female), but in the training data, it was 0.1 (10% male and 90% female), the number of synthetic males was generated to ensure the male-to-female ratio in the augmented training data matched the population ratio.

We augmented the training data with synthetic observations to enable the models to learn from a representative population sample [19]. We did not include synthetic observations in the test set, as they could artificially inflate performance metrics due to overfitting [34,35]. To account for the stochastic nature of the data generation models, we generated 500 synthetic datasets, each with a random seed consistent with established best practices in synthetic data research, to strengthen the validity of algorithm comparisons between the models trained on augmented and non-augmented data [36,37].

SMOTE, GAN, and Classification and Regression Trees (CART) models were used to generate synthetic data. These models were chosen considering their demonstrated effectiveness in handling imbalanced datasets and generating realistic synthetic data in various applications, including healthcare and finance [15,38].

Synthetic Minority Oversampling Technique (SMOTE)

SMOTE has been extensively used in previous studies to address class imbalances by generating synthetic observations through interpolation between existing minority class instances [39]. This method has been effective in enhancing model performance in situations where specific categories are underrepresented; it increases the representation of these categories without simply duplicating existing data. SMOTE's ability to reduce the risk of overfitting to majority classes while improving the representation of minority classes in the dataset makes it a valuable tool for our research [40]. We used a modified version of SMOTE for mixed data types (numeric and categorical), SMOTENC. This modification generated synthetic observations for underrepresented sociodemographic categories by interpolating between their k-nearest neighbors to enhance the diversity of the training data.

Conditional Tabular Generative Adversarial Networks (CTGAN)

CTGAN is a specialized variant of GAN for generating synthetic tabular data. CTGAN has been successfully used in prior studies to address the challenge of generating high-quality synthetic data, including categorical and continuous variables [41]. GANs consist of two neural networks, the generator and the discriminator, that engage in an adversarial process [42]. This competitive framework allows the networks to refine their output iteratively, producing synthetic data that closely mimics the original dataset. CTGAN is particularly advantageous due to its capacity to handle imbalances in categorical variables [43,44].

Classification and Regression Trees (CART)

In synthetic data generation, tree-based synthesizers use a sequential approach wherein each variable from the empirical training dataset is regarded as a potential response, with the remaining variables functioning as covariates of that response variable [45]. The operational

mechanism of CART involves the recursive partitioning of the data into subsets, guided by the observed values of the covariates. Prior research has shown that CART generates synthetic tabular health data while preserving its analytical value for researchers [46,47].

Hyperparameter Optimization for Synthetic Data Generation Models

The optimal hyperparameter values were estimated using the grid search approach [48], starting with default values in R (*themis, synthpop*) and Python (*ctgan*) packages. The number of nearest neighbors was the only hyperparameter to optimize for synthetic data generation with SMOTE models [33]. The hyperparameters for CTGAN were epoch and batch size (the number of observations used in each iteration to calculate gradients and update the model's weights) [41]. More epochs give the model more opportunities to learn from the data. However, too many epochs can lead to overfitting, where the model learns the training data too well and performs poorly on unseen data. Large batch sizes can lead to more stable gradient updates with generator and discriminator loss but require more memory. Small batch sizes can introduce noise into the training process, potentially leading to better generalization and greater stability in model parameter estimates. For CART, we optimized the maximum tree depth [45].

The Total Variation Complement (TVComplement) was used to assess the similarity of synthetic and real data, where a value of 1 indicates identical real-world and synthetic data and a value of 0 signifies maximal dissimilarity. TVComplement is the complement of the Total Variation Distance, calculated as the average difference in frequency-based probabilities for each variable category between the real-world and synthetic data [49]. The hyperparameter values corresponding to the highest value of TVComplement were used to train the models. Besides this metric, the proportion of rows in the real-world dataset disclosed in the synthetic dataset (common row proportion) was used to evaluate the quality of the synthetic data [50].

4.4.6 Statistical Analysis

Cohort characteristics were reported using frequencies and percentages or means and standard deviations, stratified by the individuals' current smoking status. Models for synthetic data generation and current smoking status identification were implemented using R 4.4.1 [51] and Python 3.11 [52].

Model-based Algorithm Development

We used the Least Absolute Shrinkage and Selection operator (LASSO) logistic regression model to construct the algorithm for predicting current smoking status with covariates defined from AHD because it is efficient for binary classification problems, simple to implement, and both time and computational resource-efficient [53]. LASSO extends logistic regression by favoring a sparse model, setting less impactful covariate coefficients to zero through a penalty term proportional to the coefficients

Using the grid search approach, an optimization set (30%) from the training data was randomly chosen to find the optimal value for the hyperparameter [48]. The single hyperparameter for this model was the multiplier of the penalty term, commonly known as lambda (the higher the value of lambda, the stronger the penalty and the more coefficients are set to zero). The hyperparameter estimate that resulted (using R packages *tidymodel* and *glmnet*) in the highest area under the receiver operating characteristic (ROC) curve was chosen to train the model.

The trained models were applied to the test dataset to estimate the predicted probability of current smoking. When choosing the classification threshold from predicted probabilities, we aimed for a balance between accurately identifying both positive (smokers) and negative (non-smokers) cases by giving equal importance to sensitivity and specificity. Hence, the cut-off

probabilities corresponding to the highest Youden's J index [54] from ROC curve analyses were chosen for an individual to be classified as a current smoker according to the algorithm.

Performance Evaluation

Algorithm performance was assessed in the test dataset using sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) [55]. Sensitivity and specificity indicate the ability of an algorithm to correctly identify current smokers and non-smokers, respectively. PPV indicates the likelihood that an algorithm-identified current smoker is indeed a current smoker, while NPV shows how likely it is that an algorithm-identified non-smoker truly is a non-smoker. For each algorithm, we reported stratified estimates of validity metrics by sex and age group. The aggregated (mean, standard error) validity metric estimates and quantile-based [56] 95% confidence intervals (CI) were reported for the algorithm trained with augmented data.

The one-sample Wilcoxon sign-rank test [57] compared algorithm performance (validity metric estimates for the underrepresented groups in the validation cohort) based on non-augmented training data with those based on augmented training data. The algorithm trained on augmented data was compared using the Wilcoxon sign-rank paired test at a 5% level of significance across different synthetic data generation models. Standard errors of the validity metric estimates resulting from the algorithm trained on the augmented data were calculated to assess their precision [58]. Algorithm fairness was assessed using a ratio-based metric, with ratios close to one indicating greater fairness, to compare validity metric estimates across categories of sex and age group [59].

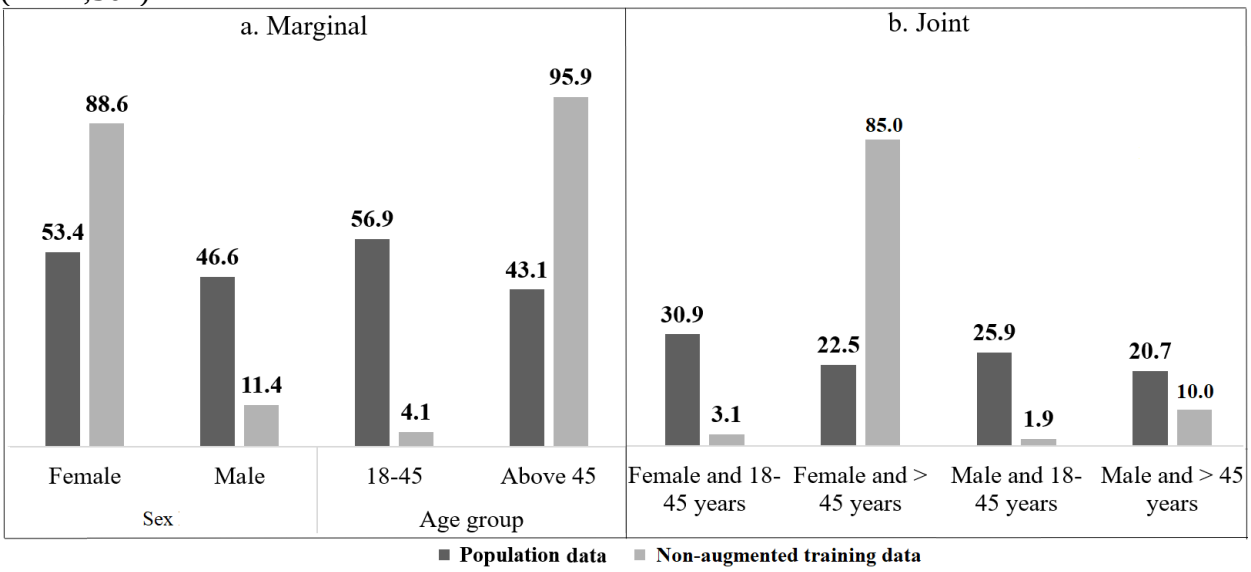
A sensitivity analysis of the performance evaluation was conducted to assess the robustness of the findings based on augmented training data. This analysis compared the effect

of including additional variables (obesity and alcohol use) from the validation data to train the synthetic data generation models on evaluation metrics, given their well-established association with smoking [60].

4.5 Results

We identified 25,712 individuals in the BMD registry, of whom 24,718 (96.1%) satisfied the inclusion criteria (Supplementary file 3). A small subset of the cohort (465 individuals) had multiple BMD test records in the registry, from which a single record was randomly selected. Smoking status was consistent across all records for the majority (97.0%) of individuals with multiple records (1.9%).

Figure 4.1 Marginal and joint distribution (%) for demographic characteristics of sex and age group in the Manitoba population data (N=435,120) and non-augmented training data (N=17,304)



The cohort was predominantly female (88.6%) and comprised primarily older adults, with a mean age of 67.0 years (Supplementary file 4). Consequently, males and younger individuals were underrepresented in this cohort based on their marginal distribution (Figure

4.1a), and the three underrepresented groups based on the joint distribution were males aged 18-45 years, females aged 18-45 years, and males greater than 45 years (Figure 4.1b).

4.5.1 Estimates of Model Hyperparameters

The optimal number of nearest neighbors for SMOTE-based synthetic observation generation considering the marginal distribution were 5 and 15 for the underrepresented sex (male) and age groups (18-45 years), respectively. For all three underrepresented groups from the joint distribution, using 15 nearest neighbors resulted in the highest TVComplement (0.98) and lowest common row proportion (0.17). The optimal parameters for generating synthetic observations for the underrepresented groups using CTGAN based on marginal distributions were 300 epochs and a batch size of 60. When generating synthetic observations considering the joint distribution, the estimated optimal parameters were 100 epochs and a batch size of 30. For CART, the estimated optimal value for maximum depth remained consistent at 10 across synthetic observation generations, considering marginal and joint distributions. The resulting TVComplement was above 0.96, and the common row proportion was below 0.19 in all cases of synthetic observation generations. In the case of the LASSO regression models for predicting current smoking status, lambda estimates ranged between 0.003 and 0.005.

4.5.2 Estimates of Validity Metrics

The results in Table 4.1 indicate variability in validity metrics across sex and age groups, with younger individuals showing higher sensitivity but lower specificity compared to older individuals. For males, sensitivity was 70.3%, specificity was 59.4%, PPV was 20.2%, and NPV was 93.2%. Females showed slightly lower sensitivity (68.1%) but higher specificity (75.5%), PPV (22.9%), and NPV (95.6%). In the 18-45 age group, sensitivity and NPV were both 100%,

while specificity was notably low at 4.4%. For individuals > 45 years, sensitivity was 66.6%, specificity was 76.4%, PPV was 23.9%, and NPV was 95.4%.

Table 4.1 Validity metric estimates (95% confidence intervals; CI) by demographic characteristics of sex and age group in test dataset (N=7414; Female=6567, Male=847; 18-45 years=293, >45 years=7121) for Least Absolute Shrinkage and Selection Operator logistic regression model based on non-augmented training data

Demographic characteristic	**Validity metric**	**Estimate, % (95% CI)**
Male	Sensitivity	70.3 (60.8,78.8)
	Specificity	59.4 (55.7,62.9)
	PPV	20.2 (17.9,22.7)
	NPV	93.2 (92.9,94.0)
Female	Sensitivity	68.1 (64.3,71.7)
	Specificity	75.5 (74.4,76.6)
	PPV	22.9 (21.7,24.2)
	NPV	95.6 (95.0,96.0)
18-45 years	Sensitivity	100.0 (91.6,100.0)
	Specificity	4.4 (2.2,7.7)
	PPV	14.9 (14.6,15.2)
	NPV	100.0 (71.5,100.0)
>45 years	Sensitivity	66.6 (62.9,70.0)
	Specificity	76.4 (75.4,77.5)
	PPV	23.9 (22.3,24.8)
	NPV	95.4 (94.9,95.9)

Note:
positive predictive value (PPV); negative predictive value (NPV)

Results in Tables 4.2 and 4.3 show that CTGAN achieved the highest sensitivity (>82.0%) estimates for marginal distribution-based data augmentation. In contrast, SMOTE resulted in the highest estimates of specificity (>64.0%) and PPV (>21.0%) in the underrepresented groups.

Table 4.2 Sex-stratified validity metric estimates (95% confidence intervals; CI) in test dataset (N=7414; Female=6567, Male=847) for Least Absolute Shrinkage and Selection Operator logistic regression model trained with augmented data in 500 synthetic data generation experiments

Distribution to generate synthetic observations	Validity metric	Synthetic data generation model for training data augmentation	Estimate, % (95% CI)	
			Male	Female
Marginal	Sensitivity	SMOTE	66.7 (62.1,72.2)	66.0 (61.7,68.9)
		CTGAN	82.4 (76.1,89.3)	57.5 (53.3,65.7)
		CART	64.6 (66.7,77.8)	65.3 (59.4,69.7)
	Specificity	SMOTE	64.6 (58.4,70.0)	78.6 (74.9,81.9)
		CTGAN	21.1 (18.3,25.9)	83.3 (76.2,87.0)
		CART	61.3 (54.5,65.6)	77.1 (72.7,82.9)
	PPV	SMOTE	21.6 (20.3,23.1)	24.6 (22.6,26.8)
		CTGAN	13.4 (9.5,20.1)	27.6 (22.8,30.7)
		CART	21.2 (20.0,22.3)	23.5 (21.3,27.0)
	NPV	SMOTE	93.0 (92.4,93.6)	95.4 (95.2,95.7)
		CTGAN	92.0 (87.1,93.8)	94.8 (94.5,95.4)
		CART	93.7 (93.1,94.4)	95.4 (95.0,95.7)
Joint	Sensitivity	SMOTE	52.8 (44.7,58.3)	54.2 (49.7,65.8)
		CTGAN	87.0 (86.3,91.7)	69.0 (65.6,72.3)
		CART	73.9	60.2

			(68.5,79.6)	(55.9,64.1)
	Specificity	SMOTE	84.3 (67.8,90.8)	81.8 (70.0,86.3)
		CTGAN	36.8 (32.7,41.6)	72.0 (68.9,75.5)
		CART	53.2 (46.8,59.4)	78.2 (74.3,82.2)
	PPV	SMOTE	26.7 (19.0,28.2)	24.5 (19.0,27.9)
		CTGAN	16.8 (15.4,17.5)	20.9 (19.8,25.2)
		CART	18.8 (17.9,19.6)	22.9 (20.9,24.5)
	NPV	SMOTE	90.8 (90.5,91.6)	94.4 (94.1,95.0)
		CTGAN	94.9 (94.4,95.6)	95.6 (95.3,95.9)
		CART	93.3 (92.7,94.0)	94.8 (94.6,95.1)

Notes:
[1]Synthetic Minority Oversampling Technique (SMOTE); Conditional Tabular Generative Adversarial Networks (CTGAN); Classification and Regression Trees (CART); positive predictive value (PPV); negative predictive value (NPV)
[2] Estimates are mean for 500 experiments; 95% confidence intervals are quantile-based

Conversely, SMOTE yielded the lowest sensitivity (<67.0%), and CTGAN produced the lowest specificity (<23.0%) and PPV (<16.0%). CTGAN and CART had high sensitivity estimates (>81.0%) compared to their counterparts when the joint distribution was used for data augmentation. Using SMOTE-generated observations in augmented data resulted in the highest specificity (>80.0%) and PPV (>24.0%) estimates. The lowest sensitivity (<55.0%) and specificity (<50.0%) estimates were produced using SMOTE and CTGAN, respectively. The lowest PPV (<17.0%) estimates were produced using CTGAN. Across all scenarios, NPV remained consistently high (≥90%) for males and individuals aged 18-45 years. The validity metric estimates for females and those aged >45 were relatively stable compared to the underrepresented subgroups.

Table 4.3 Age group-stratified validity estimates (95% confidence intervals) in test dataset (N=7414; 18-45 years=293, > 45 years=7121) for Least Absolute Shrinkage and Selection Operator logistic regression models trained with augmented data in 500 synthetic data generation experiments

Distribution to generate synthetic observations	Validity metric	Synthetic data generation model for training data augmentation	Estimate, % (95% CI)	
			18-45 years	>45 years
Marginal	Sensitivity	SMOTE	57.1 (49.5,70.9)	56.5 (48.3,65.5)
		CTGAN	95.2 (97.6,100.0)	65.3 (62.3,68.7)
		CART	89.0 (85.7,97.6)	65.2 (54.4,69.9)
	Specificity	SMOTE	60.1 (52.6,66.9)	69.9 (61.8,88.7)
		CTGAN	12.3 (10.3,15.5)	77.5 (74.2,80.3)
		CART	41.1 (23.5,38.9)	75.8 (69.8,85.4)
	PPV	SMOTE	22.8 (21.5,23.6)	26.9 (20.8,34.4)
		CTGAN	15.1 (14.6,15.6)	24.1 (22.4,25.7)
		CART	21.8 (17.6,19.1)	22.8 (20.2,28.9)
	NPV	SMOTE	92.2 (90.8,94.3)	94.5 (94.1,95.1)
		CTGAN	99.4 (95.1,100.0)	95.3 (95.1,95.6)
		CART	93.5 (93.7,98.5)	95.1 (94.5,95.5)
Joint	Sensitivity	SMOTE	40.8 (31.4,56.0)	52.1 (46.5,63.8)
		CTGAN	69.0 (66.2,100.0)	70.1 (67.0,73.5)
		CART	83.3 (80.7,100.0)	60.5 (55.9,64.4)
	Specificity	SMOTE	80.9 (69.3,88.8)	82.2 (70.1,87.5)
		CTGAN	50.2 (47.8,63.9)	70.1 (66.9,73.3)
		CART	30.4 (23.1,35.1)	77.1 (73.0,81.4)
	PPV	SMOTE	25.4 (22.2,30.0)	24.6 (18.9,28.9)

		CTGAN	18.3 (14.7,18.6)	20.4 (19.4,21.5)
		CART	18.0 (17.5,18.6)	22.4 (20.6,24.7)
	NPV	SMOTE	88.7 (88.2,90.5)	94.0 (93.7,94.7)
		CTGAN	94.9 (90.4,100.0)	95.7 (95.3,95.8)
		CART	95.7 (93.6,100.0)	94.7 (94.4,94.9)

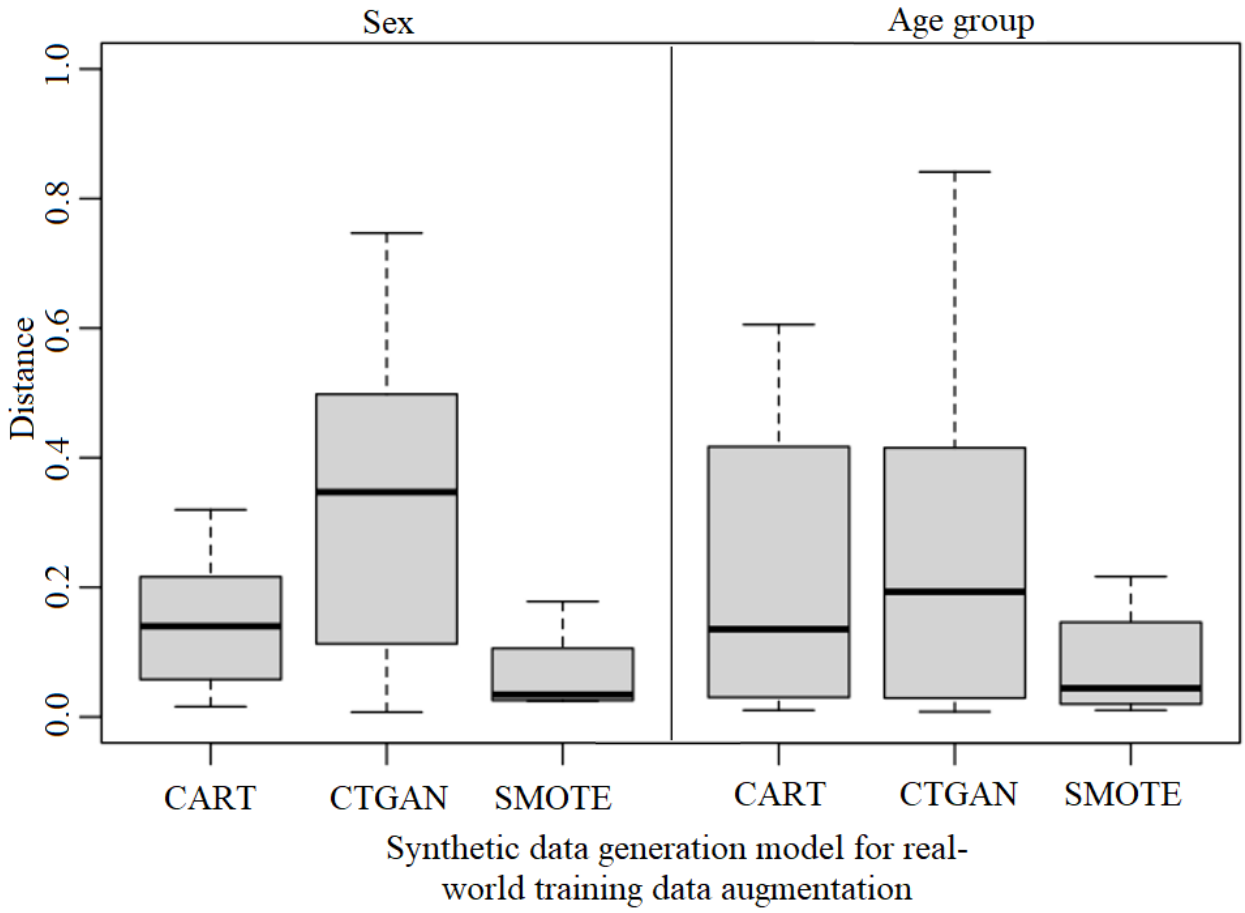
Notes:
[1]Synthetic Minority Oversampling Technique (SMOTE); Conditional Tabular Generative Adversarial Networks (CTGAN); Classification and Regression Trees (CART); positive predictive value (PPV); negative predictive value (NPV)
[2] Estimates are mean for 500 experiments; 95% confidence intervals are quantile-based

4.5.3 Comparison of Performance

One-sample Wilcoxon sign-rank tests indicated significant improvements in specificity and PPV for males and individuals aged 18-45 with training data augmentation using SMOTE compared to the algorithm based on non-augmented training data. Training data augmentations using CTGAN and CART showed higher sensitivity for males, with CART additionally improving specificity and PPV.

The paired Wilcoxon sign-rank tests indicated that SMOTE consistently outperformed CTGAN and CART in producing higher specificity and PPV estimates for both marginal and joint distribution-based synthetic data for males. CART demonstrated superior sensitivity and NPV estimates compared to CTGAN for males. For the 18-45 years age group, SMOTE yielded higher specificity, PPV, and NPV than CTGAN across both distribution types, while CART outperformed SMOTE in sensitivity and NPV. CART also surpassed CTGAN in sensitivity, specificity, and PPV for this age group.

Figure 4.2 Boxplot of the absolute distance of fairness ratio from perfection (1.0) by sex and age group



Notes:
[1] Synthetic Minority Oversampling Technique (SMOTE), Conditional Tabular Generative Adversarial Networks (CTGAN), Classification and Regression Trees (CART)
[2] The center horizontal line within the box represents the median (50th percentile); the upper and lower bounds of the box indicate the 25th and 75th percentiles; dashed lines connect the maximum and minimum values.

Irrespective of the distribution used to generate synthetic observations, the difference in the estimates of validity metrics across sex and age groups was lower for SMOTE (indicating greater fairness with ratios closer to one) and higher for CTGAN (Figure 4.2). For example, the ratio (male to female) of sensitivity estimates was 1.02 when SMOTE was used to augment training data based on marginal distribution. However, in this case, for CTGAN, the ratio was 1.43. Standard errors of estimates of validity metrics were higher for the algorithm trained with synthetic observations generated based on joint distributions of sex and age group compared to

their marginal distribution counterparts. In the sensitivity analysis incorporating obesity and alcohol use in synthetic observation generation, the findings did not differ considerably from the primary analysis (Supplementary file 5).

4.6 Discussion

This study applied SMOTE, CTGAN, and CART to generate synthetic observations to augment training data for testing a model-based algorithm to identify smoking status in AHD. Specifically, we focused on how these synthetic data generation models impact the estimates of common diagnostic algorithm validity metrics, including sensitivity, specificity, PPV, and NPV. The underrepresentation of males and younger individuals in our validation cohort reflects a well-documented challenge of imbalance in clinical data sources used to validate algorithms in AHD [61]. These imbalances can affect the generalizability of validity estimates, highlighting the importance of addressing them using proper mitigation strategies [11]. Even though imbalance in the response variable is commonly addressed in developing disease and health risk factor prediction models, mitigation of covariate imbalance has been limited [62]. In particular, implementing covariate imbalance mitigation strategies remained relatively unexplored in validation studies of diagnostic algorithms for AHD.

Overall, the algorithm trained on augmented data performed better than the algorithm trained on non-augmented data. The validity metric estimates also differed across models used to generate synthetic observations for data augmentation. The algorithm trained with data augmented using SMOTE consistently achieved high specificity and PPV for males and individuals aged 18-45, aligning with prior findings that balanced training data [13]. However, this came at the expense of sensitivity, a known trade-off in class balancing [63]. A similar performance of SMOTE regarding specificity was evident in identifying diabetes in a clinical

dataset using diagnostic measures, but the sensitivity estimates also improved in that study [64]. Conversely, CTGAN produced the highest sensitivity, particularly for males, likely due to its ability to capture rare patterns through learned distributions. However, the use of training data that were augmented with CTGAN resulted in decreased specificity and PPV, suggesting the possibility of overfitting. CART excelled in sensitivity and NPV, making it suitable for applications prioritizing these metrics, but it underperformed in specificity and PPV compared to SMOTE. Similar results were reported in a comparison between CART and CTGAN in terms of the accuracy of synthetic patient survival time generated by these models for multiple clinical data sources [37].

SMOTE effectively balanced sample distributions across sex and age groups, producing fairness metric estimates close to one, indicating equitable validity metric estimates across subgroups. This suggests that SMOTE was particularly effective in mitigating fairness challenges arising from imbalances in sex and age groups. These findings align with a study that evaluated the impact of age and gender on classifying systemic inflammatory response syndrome using electronic health records [65]. In contrast, CTGAN exhibited large disparities, particularly in sensitivity, potentially introducing biases to certain groups with high validity estimates. By leveraging decision tree-based learning, CART demonstrated robust sensitivity and NPV performance but lacked SMOTE's fairness consistency due to hierarchical splits that could amplify subgroup differences.

The study highlights the importance of considering both marginal and joint distributions when augmenting data. Our results indicate that data augmentation based on marginal distributions provided more precise estimates (with small standard errors) of validity metrics

than joint distribution-based augmentation [58]. This likely occurred because data augmentation based on the joint distribution of covariates required more synthetic observations to account for the interaction between demographic factors than when these factors were considered separately. In the sensitivity analysis, the incorporation of additional variables in synthetic observation generation for data augmentation did not change the pattern of findings. However, a study on multiple clinical datasets revealed conflicting results, suggesting the variables used for data augmentation with synthetic observations may impact diagnostic algorithm accuracy [66].

4.6.1 Strengths and Limitations

One of the strengths of this study is the accuracy of current smoking status as captured in the clinical registry used for algorithm validation [67]. A second strength is that a comprehensive data augmentation strategy was employed, accounting for both marginal and joint distributions of key demographic variables. Furthermore, the evaluation of different synthetic data generation models facilitated a comparative analysis, offering valuable guidance for model selection tailored to specific applications.

However, the study has some limitations. Although the algorithm was trained using data balanced through augmentation, the evaluation was conducted on real-world data from the imbalanced validation cohort. This approach restricted the evaluation to commonly employed internal validation with the test set [13,66] that was not representative of the broader population, potentially raising concerns regarding the robustness of our estimates. Additionally, our focus on smoking status also constrains the ability to generalize the effectiveness of synthetic data in AHD validation studies to diagnostic algorithms for other health conditions and behaviors, such as obesity [5] and social determinants of health [7].

4.6.2 Future Directions

The study findings underscore the need for further research on covariate imbalance mitigation strategies in diagnostic algorithms for AHD validation studies to ensure equitable and reliable results for all population subgroups. Separate augmentation of training and test datasets with synthetic observations and external validation with real-world data reflective of the population distribution could be considered in subsequent studies [68]. Considering data augmentation in the context of missing value in AHD validation studies presents an important avenue for future research [69]. For instance, an experimental study could investigate the impact of varying levels of missingness (e.g., 5-10%) in a covariate within augmented training data on the validity metric estimates. Future research could consider more than two covariates for synthetic data generation that may influence health conditions and risk factors, such as place of residence.

4.6.3 Study Implications

This study offers methodological insights and practical guidance for advancing the accuracy and equity of AHD validation studies. The findings underscore the potential of synthetic data to improve the generalizability of validity assessments of algorithms for AHD, which are widely used in surveillance and research. By addressing data imbalances and improving representativeness, synthetic data can contribute to more equitable and reliable outcomes across diverse population subgroups.

4.6.4 Conclusions

This study demonstrates the potential of data augmentation using synthetic data generation models, particularly SMOTE, to improve the accuracy and equity of validity estimates from AHD validation studies for diagnostic algorithms. By generating synthetic

observations for underrepresented groups, researchers can potentially mitigate biases inherent in real-world datasets and enhance the generalizability of their findings. However, as the results suggest, careful consideration must be given to the synthetic data generation model choice and the trade-offs between different validity metrics.

4.7 References

1. Lo Re V. Chapter 15 - Validation of health outcomes of interest in healthcare databases. In: Girman CJ, Ritchey ME, editors. Pragmatic Randomized Clinical Trials [Internet]. Academic Press; 2021. p. 207–18. Available from: https://www.sciencedirect.com/science/article/pii/B9780128176634000222

2. Nc H, Ra M, D J, P I, L L. Building and validating trend-based multiple sclerosis case definitions: a population-based cohort study for Manitoba, Canada. BMJ Open. 2024 Aug 15;14(7). PMID: 39151935

3. Lix LM, Renoux C, Moriello C, Choi KL, Dormuth CR, Fisher A, et al. Validity of diagnoses of SARS-CoV-2 infection in Canadian administrative health data: a multiprovince, population-based cohort study. Canadian Medical Association Open Access Journal. 2023 Sep 1;11(5):E790–8. PMID: 37669811

4. McGinnis KA, Skanderson M, Justice AC, Tindle HA, Akgün KM, Wrona A, et al. Using the biomarker cotinine and survey self-report to validate smoking data from United States Veterans Health Administration electronic health records. Journal of the American Medical Informatics Association Open. 2022 Jul 1;5(2):ooac040. PMID: 37252267

5. Samadoulougou S, Idzerda L, Dault R, Lebel A, Cloutier AM, Vanasse A. Validated methods for identifying individuals with obesity in health care administrative databases: a systematic review. Obesity Science & Practice. 2020;6(6):677–93. PMID: 33354346

6. Conderino S, Bendik S, Richards TB, Pulgarin C, Chan PY, Townsend J, et al. The use of electronic health records to inform cancer surveillance efforts: a scoping review and test of

indicators for public health surveillance of cancer prevention and control. BMC Medical Informatics and Decision Making. 2022 Apr 6;22(1):91. PMID: 35387655

7. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. Journal of American Medical Informatics Association. 2021 Oct 19;29(1):187–96. PMID: 34664641

8. Harbaugh CM, Cooper JN. Administrative databases. Seminars in Pediatric Surgery. 2018 Dec 1;27(6):353–60. PMID: 30473039

9. Haque MA, Gedara MLB, Nickel N, Turgeon M, Lix LM. The validity of electronic health data for measuring smoking status: a systematic review and meta-analysis. BMC Medical Informatics and Decision Making. 2024 Feb 2;24(1):33. PMID: 38308231

10. Panzeri S, Treves A. Analytical estimates of limited sampling biases in different information measures. Network. 1996;7(1):87-107. PMID: 29480146

11. Kang S. Model validation failure in class imbalance problems. Expert Systems with Applications. 2020 May 15;146:113190.

12. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Computing Survey. 2019 Aug 30;52(4):79:1-79:36.

13. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns. 2024 Apr 12;5(4):100946. PMID: 38645766

14. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. Journal of Big Data. 2018 Nov 1;5(1):42.

15. Pezoulas VC, Zaridis DI, Mylona E, Androutsos C, Apostolidis K, Tachos NS, et al. Synthetic data generation methods in healthcare: a review on open-source tools and methods. Computational and Structural Biotechnology Journal. 2024 Dec 1;23:2892–910. PMID: 39108677

16. Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. Frontiers of Computer Science. 2020 Apr 1;14(2):241–58.

17. Lomax S, Vadera S. A survey of cost-sensitive decision tree induction algorithms. ACM Computing Survey. 2013 Mar 12;45(2):16:1-16:35.

18. Grinde KE, Arbet J, Green A, O'Connell M, Valcarcel A, Westra J, et al. Illustrating, quantifying, and correcting for bias in post-hoc analysis of gene-based rare variant tests of association. Frontiers in Genetics. 2017 Sep 14;8. PMID: 28959274

19. Mumuni A, Mumuni F. Data augmentation: Data augmentation: a comprehensive survey of modern approaches. Array. 2022 Dec 1;16:100258.

20. Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, et al. A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access. 2021;9:109960–75.

21. Papadaki E, Vrahatis AG, Kotsiantis S. Exploring innovative approaches to synthetic tabular data generation. Electronics. 2024 Jan;13(10):1965.

22. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The reporting of studies conducted using observational routinely-collected health Data (RECORD) statement. PLOS Medicine. 2015 Oct;12(10):e1001885. PMID: 26440803

23. Smith M, Roos LL, Burchill C, Turner K, Towns DG, Hong SP, et al. Health services data: managing the data warehouse: 25 years of experience at the Manitoba Centre for Health Policy. In: Levy A, Goring S, Gatsonis C, Sobolev B, van Ginneken E, Busse R, editors. Health Services Evaluation [Internet]. New York, NY: Springer US; 2019. p. 19–45. Health Services Research. Available from: https://doi.org/10.1007/978-1-4939-8715-3_3

24. Leslie WD, MacWilliam L, Lix L, Caetano P, Finlayson GS. A population-based study of osteoporosis testing and treatment following introduction of a new bone densitometry service. Osteoporosis International. 2005 Jul 1;16(7):773–82. PMID: 15580480

25. Lewiecki EM. Update on bone density testing. Current Osteoporosis Reports. 2005 Dec 1;3(4):136–42. PMID: 16303113

26. Goel H, Binkley N, Hans D, Leslie WD. Bone density and trabecular bone score to predict fractures in adults aged 20–39 years: a registry-based study. Osteoporosis International. 2023 Jun 1;34(6):1085–91. PMID: 37000209

27. WHO. Anatomical Therapeutic Chemical (ATC) Classification [Internet]. 2024. Available from: https://www.who.int/tools/atc-ddd-toolkit/atc-classification

28. Government of Canada SC. Census of Population [Internet]. Available from: https://www12.statcan.gc.ca/census-recensement/index-eng.cfm

29. MCHP Metadata Repository. Term: Personal Health Identification Number (PHIN) [Internet]. Available from: http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=103292

30. MCHP Metadata Repository. Term: Income Quintiles / Income Quintile [Internet]. Available from: http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=102882

31. Havard A, Jorm LR, Lujic S. Risk adjustment for smoking identified through tobacco use diagnoses in hospital data: a validation study. PLOS One. 2014 Apr 15;9(4):e95029. PMID: 24736621

32. Schäfer AA, Santos LP, Quadra MR, Dumith SC, Meller FO. Alcohol consumption and smoking during Covid-19 pandemic: association with sociodemographic, behavioral, and mental health characteristics. Journal of Community Health. 2022 Aug 1;47(4):588–97. PMID: 35334032

33. Hvitfeldt E, Software P, PBC. themis: extra recipes steps for dealing with unbalanced data [Internet]. 2023. Available from: https://cran.r-project.org/web/packages/themis/index.html

34. Pan X, Zhao J, Xu J. Conditional generative adversarial network-based training sample set improvement model for the semantic segmentation of high-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. 2021 Sep;59(9):7854–70.

35. Offenhuber D. Shapes and frictions of synthetic data. Big Data & Society. 2024 Jun 1;11(2):20539517241249390.

36. Khadka K, Chandrasekaran J, Lei Y, Kacker RN, Richard Kuhn D. Synthetic data generation using combinatorial testing and variational autoencoder. In: 2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW) [Internet]. 2023. p. 228–36. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/10132195

37. Akiya I, Ishihara T, Yamamoto K. Comparison of synthetic data generation techniques for control group survival data in oncology clinical trials: simulation study. JMIR Medical Informatics. 2024 Jun 18;12(1):e55118. PMID: 38889082

38. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: a systematic review. Neurocomputing. 2022 Jul 7;493:28–45.

39. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002 Jun 1;16:321–57.

40. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research. 2018 Apr 20;61:863–905.

41. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems [Internet]. 2019. Available from: https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

42. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Communications of the Association for Computing Machinery. 2020 Oct 22;63(11):139–44.

43. Jia J, Wu P, Zhang K, Zhong J. Imbalanced disk failure data processing method based on CTGAN. In: Huang DS, Jo KH, Jing J, Premaratne P, Bevilacqua V, Hussain A, editors. Intelligent Computing Theories and Application. Cham: Springer International Publishing; 2022. p. 638–49.

44. Wang J, Yan X, Liu L, Li L, Yu Y. CTTGAN: traffic data synthesizing scheme based on conditional GAN. Sensors. 2022 Jul 13;22(14):5243. PMID: 35890921

45. Nowok B, Raab GM, Dibben C. synthpop: bespoke creation of synthetic data in R. Journal of Statistical Software. 2016 Oct 28;74:1–26.

46. CanPath - Canadian Partnership for Tomorrow's Health [Internet]. CanPath Student Dataset. Available from: https://canpath.ca/student-dataset/

47. Sallier K. Toward more user-centric data access solutions: producing synthetic data of high analytical value by data synthesis. Statistical Journal of the International Association for Official Statistics. 2020 Jan 1;36(4):1059–66.

48. Feurer M, Hutter F. Hyperparameter optimization [Internet]. 2018. Available from: https://www.semanticscholar.org/paper/Chapter-1-Hyperparameter-Optimization-Feurer-Hutter/1a90075b632c8cb7762a337431f143c6fe682bad

49. Moore K, Thomas A, Taylor S, Jackson J, White K. Exploring synthetic data generation: leveraging the SDV library for realistic dataset creation. 2023.

50. Fössing E, Drechsler J. An evaluation of synthetic data generators implemented in the Python library Synthcity. International Conference on Privacy in Statistical Databases 2024 Sep 13 (pp. 178-193).

51. R: The R Project for Statistical Computing [Internet]. Available from: https://www.r-project.org/

52. Python.org [Internet]. 2024. Welcome to Python.org. Available from: https://www.python.org/

53. Tibshirani R. Regression shrinkage and selection via the LASSO. The Journal of the Royal Statistical Society, Series B (Statistical Methodology). 1996;58(1):267–88.

54. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biometrical Journal. 2005;47(4):458–72. PMID: 16161804

55. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. Medicina. 2021 May 16;57(5):503. PMID: 34065637

56. Ialongo C. Confidence interval for quantiles and percentiles. Biochemia Medica. 2019 Feb 15;29(1):0–0. PMID: 30591808

57. Vierra A, Razzaq A, Andreadis A. Chapter 27 - Continuous variable analyses: t-test, Mann–Whitney U, Wilcoxon sign rank. In: Eltorai AEM, Bakal JA, Newell PC, Osband AJ, editors. Translational Surgery [Internet]. Academic Press; 2023. p. 165–70. Handbook for Designing and Conducting Clinical and Translational Research. Available from: https://www.sciencedirect.com/science/article/pii/B9780323903004000458

58. Soori H. Precision, validity, and repeatability of measurements and diagnostic tests. In: Soori H, editor. Errors in Medical Science Investigations [Internet]. Singapore: Springer Nature; 2024. p. 73–96. Available from: https://doi.org/10.1007/978-981-99-8521-0_4

59. Yeh MH, Metevier B, Hoag A, Thomas P. Analyzing the relationship between difference and ratio-based fairness metrics. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency [Internet]. New York, NY, USA: Association for Computing Machinery; 2024. p. 518–28. (FAccT '24). Available from: https://dl.acm.org/doi/10.1145/3630106.3658922

60. Oh HS, Seo WS. The compound relationship of smoking and alcohol consumption with obesity. Yonsei Medical Journal. 2001 Oct;42(5):480–7. PMID: 11675675

61. Afrose S, Song W, Nemeroff CB, Lu C, Yao D (Daphne). Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. Communications Medicine. 2022 Sep 1;2(1):1–14. PMID: 36059892

62. Powers S, McGuire V, Bernstein L, Canchola AJ, Whittemore AS. Evaluating disease prediction models using a cohort whose covariate distribution differs from that of the target

population. Statistical Methods in Medical Research. 2019 Jan 1;28(1):309–20. PMID: 28812439

63. Shao Y, Sun Y, Guan H. Dual self-paced SMOTE for imbalanced data. In: 2022 26th International Conference on Pattern Recognition (ICPR) [Internet]. 2022. p. 3083–9. Available from: https://ieeexplore.ieee.org/document/9956517

64. Aubaidan BH, Kadir RA, Ijab MT. A comparative analysis of Smote and CSSF techniques for diabetes classification using imbalanced data. Journal of Computational Science. 2024 Jul 22;20(9):1146–65.

65. Das PP, Mast M, Wiese L, Jack T, Wulff A. Algorithmic fairness in healthcare data with weighted loss and adversarial learning. In: Arai K, editor. Intelligent Systems and Applications. Cham: Springer Nature Switzerland; 2024. p. 264–83.

66. Al–Qerem A, Ali AM, Attar H, Nashwan S, Qi L, Moghimi MK, et al. Synthetic generation of multidimensional data to improve classification model validity. Journal of Data and Information Quality. 2023 Sep 28;15(3):37:1-37:20.

67. Leslie WD, Morin SN, Martineau P, Bryanton M, Lix LM. Factors associated with bone density monitoring while on antiosteoporosis treatment in routine clinical practice: a registry-based cohort study. Journal of Clinical Densitometry. 2020 Oct 1;23(4):568–75. PMID: 31003744

68. Rebuffi SA, Gowal S, Calian DA, Stimberg F, Wiles O, Mann TA. Data augmentation can improve robustness. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2021. p. 29935–48. Available from:

https://proceedings.neurips.cc/paper/2021/hash/fb4c48608ce8825b558ccf07169a3421-Abstract.html

69. Hoogland J, van Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. Handling missing predictor values when validating and applying a prediction model to new patients. Statistics in Medicine. 2020 Nov 10;39(25):3591–607. PMID: 32687233

4.8 SUPPLEMENTARY MATERIAL

Supplementary file 1

Supplemental Table 4.1 The RECORD statement – checklist of items, extended from the STROBE statement, to report 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	131-133	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.	131-133
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	134-136		
Objectives	3	State specific objectives, including any prespecified hypotheses	135-136		
Methods					
Study Design	4	Present key elements of study design early in the paper	136-138		

Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	136-138		
Participants	6	*(a)* Cohort study - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up 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	136-138, Supplementary file 3	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If 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.	136-138, Supplementary file 3
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	137-138, Supplementary file 2	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.	137-138, Supplementary file 2
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			
Bias	9	Describe any efforts to address potential sources of bias	N/A		
Study size	10	Explain how the study size was arrived at	14; Supplementary file 3		
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	138		
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	142-144		

Data access and cleaning methods				RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population.	136-137
				RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	
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.	137
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) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram	14, Supplementary file 3	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.	144, Supplementary file 3
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)	144, Supplementary file 4		

Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report numbers in each exposure	144-145, Supplementary file 4		
		category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures			
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	144-152		
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	146-150, Table 4.1, Table 4.2, Table 4.3		
Discussion					
Key results	18	Summarise key results with reference to study objectives	152-154		

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	154-155	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.	154-155
Interpretation	20	Give a cautious overall interpretation of results considering objectives,	155		
		limitations, multiplicity of analyses, results from similar studies, and other relevant evidence			
Generalisability	21	Discuss the generalisability (external validity) of the study results	155-156		
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	iii		
Accessibility of protocol, raw data, and programming code		..		RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	242

*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.

*Checklist is protected under Creative Commons Attribution (CC BY) license.

Supplementary file 2

Supplemental Table 4.2 Indicators used to develop rule-based and model-based algorithms for current smoking status by data source

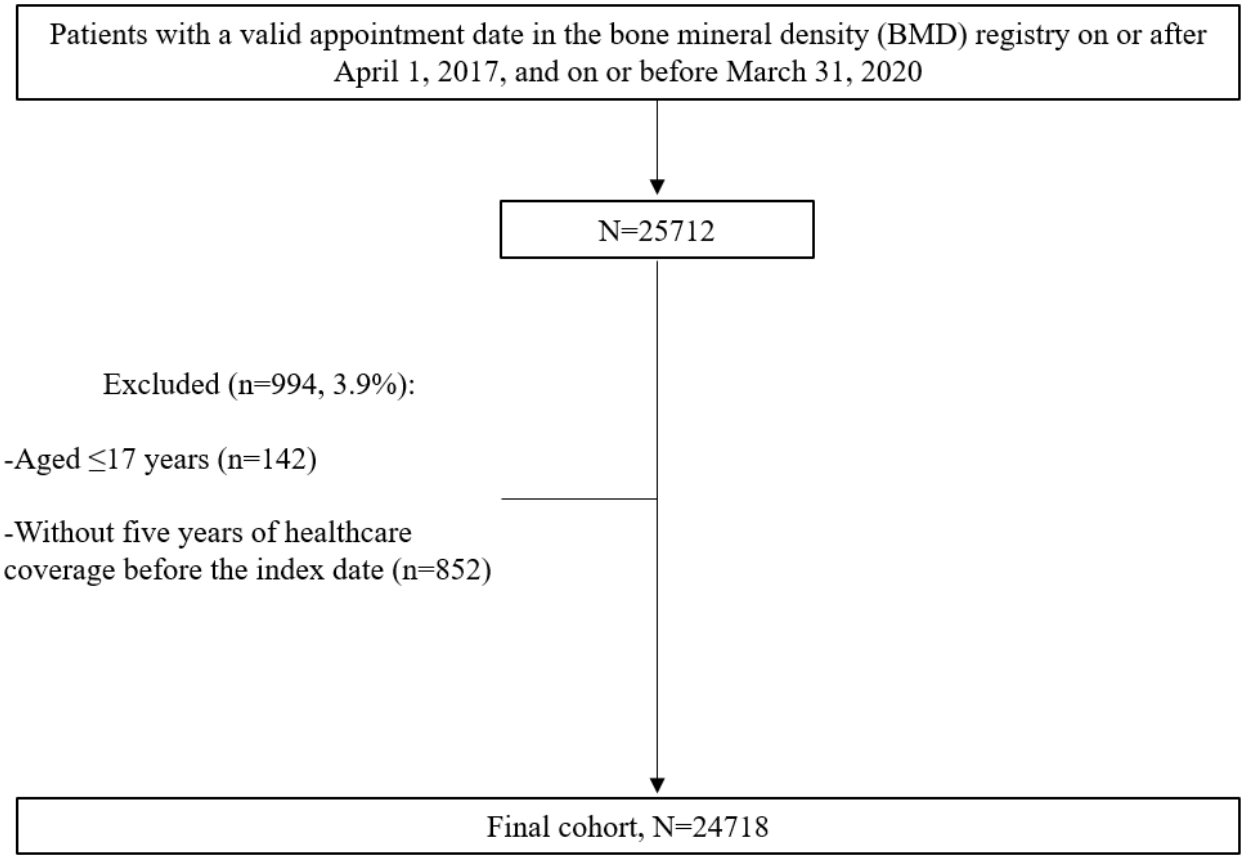
Indicator	Description	Data source
Nicotine dependence diagnosis	1=Presence of ICD-10-CA code F17.2; 0=Else	Hospital Abstracts
Nicotine dependence diagnosis	1=Presence of ICD-9-CM code 305.1; 0=Else	Medical Claims
Problems related to lifestyle: tobacco use diagnosis	1=Presence of ICD-10-CA code Z72.0; 0=Else	Hospital Abstracts
Personal history presenting hazards to health diagnosis	1=Presence of ICD-9-CM codes V15.82; 0=Else	Medical Claims
COPD diagnosis	1=Presence of at least one code from ICD-10-CA codes: J41 (Simple chronic bronchitis), J42 (Unspecified chronic bronchitis), J43 (Emphysema), J44 (Other COPD); 0=Else	Hospital Abstracts
COPD diagnosis	1=Presence of at least one code from ICD-9-CM codes: 490 (Bronchitis, not specified as acute or chronic), 491 (Chronic bronchitis), 492 (Emphysema), 496 (Chronic airway obstruction); 0=Else	Medical Claims
Asthma diagnosis	1=Presence of ICD-10-CA code J45 (Asthma); 0=Else	Hospital Abstracts
Asthma diagnosis	1=Presence of ICD-9-CM code 493 (Asthma); 0=Else	Medical Claims
Respiratory system treatment drug	1=Presence of ATC code R03AC02 (Salbutamol); 0=Else	Prescription Drug Records
Medication primarily recommended as an antidepressant and for smoking cessation	1=Presence of ATC code N06AX12 (Bupropion); 0=Else	Prescription Drug Records

Medication to quit smoking	1=Presence of ATC code N07BA03 (Varenicline); 0=Else	Prescription Drug Records
Medication to aid in smoking cessation	1=Presence of ATC code B07BA01 (Nicotine); 0=Else	Prescription Drug Records
Urban residence	1=The place of residence is an urban area; 0=The place of residence is a rural area/not found	Health Insurance Registry
Income quintile 2	1=Belongs to income quintile group 2; 0=Belongs to income quintile group 1 (lowest)/ not found	Statistics Canada Census
Income quintile 3	1=Belongs to income quintile group 3; 0=Belongs to income quintile group 1/ not found	Statistics Canada Census
Income quintile 4	1=Belongs to income quintile group 4; 0=Belongs to income quintile group 1/ not found	Statistics Canada Census
Income quintile 5 (highest)	1=Belongs to income quintile group 5; 0=Belongs to income quintile group 1/ not found	Statistics Canada Census
Biological sex	1=Female; 0=Male	Health Insurance Registry
Age	Age on index date in years	Health Insurance Registry

COPD: chronic obstructive pulmonary disease

Supplementary file 3

Supplemental Figure 4.1 Cohort flow diagram



Supplementary file 4

Supplemental Table 4.3 Characteristics of the study cohort stratified by self-reported current smoking status (N = 24,718)

Characteristic	Current smoker n=2475 (10.0%)	Non-smoker n=22,243 (90.0%)
Sex		
Male	360 (14.5)	2452 (11.0)
Female	2115 (85.5)	19791 (89.0)
Residence location at index date		
Urban	1541 (62.3)	15813 (71.1)
Rural	905 (36.5)	6257 (28.1)
Unknown	29 (1.2)	173 (0.8)
Income quintile		
Q5 (highest)	309 (12.5)	4942 (22.2)
Q4	415 (16.7)	4506 (20.3)
Q3	507 (20.5)	4596 (20.6)
Q2	562 (22.7)	3942 (17.7)
Q1 (lowest)	653 (26.4)	4084 (18.4)
Unknown	29 (1.2)	173 (0.8)
Age at index date, mean (SD)	62.4 (10.3)	67.6 (10.9)
Diagnoses		
Nicotine dependence or tobacco use problems related to lifestyle in hospital abstracts / personal history of tobacco use in medical claims	30 (1.2)	24 (0.11)
Nicotine dependence diagnosis in medical claims	258 (10.4)	105 (0.47)
Chronic obstructive pulmonary disease (COPD)/asthma diagnosis in hospital abstracts	140 (5.6)	484 (2.2)
COPD diagnosis in medical claims	675 (27.3)	1718 (7.7)
Asthma diagnosis in medical claims	460 (18.6)	2949 (13.3)
Use of respiratory system treatment drug (Salbutamol) in prescription drug records	598 (24.3)	3214 (14.5)
Use of medication primarily recommended as an antidepressant and for smoking cessation in prescription drug records (Bupropion)	159 (6.4)	552 (2.5)
Use of medication to quit smoking in prescription drug records (Varenicline)	276 (11.2)	186 (0.84)

Use of medication to aid in smoking cessation in prescription drug records (Nicotine)	123 (4.9)	59 (0.27)

Note: Diagnoses are based on five years of administrative health data, Standard Deviation (SD), values reported are frequency (%) unless otherwise stated

Supplementary file 5

Results from the sensitivity analysis

Supplemental Table 4.4 Sex-stratified validity measure estimates (95% confidence intervals) in test dataset (N=7414; Female=6567, Male=847) for Least Absolute Shrinkage and Selection Operator logistic regression models trained with augmented data in 500 synthetic data generation experiments

Distribution to generate synthetic observations	Validity metric	Synthetic data generation model for training data augmentation	Estimate, % (95% CI)	
			Male	Female
Marginal	Sensitivity	SMOTE	59.9 (56.1,69.3)	64.1 (59.5,67.3)
		CTGAN	81.4 (74.1,86.5)	58.2 (53.7,64.2)
		CART	69.6 (65.7,78.3)	64.9 (58.3,68.8)
	Specificity	SMOTE	66.7 (59.1,71.8)	78.1 (74.3,81.0)
		CTGAN	22.4 (19.6,27.3)	80.8 (77.4,85.9)
		CART	58.5 (52.4,66.1)	74.6 (70.5,83.1)
	PPV	SMOTE	19.1 (16.4,21.2)	23.2 (20.5,25.9)
		CTGAN	16.1 (13.7,23.4)	26.4 (21.5,30.9)
		CART	21.0 (19.8,22.2)	26.1 (23.5,27.0)
	NPV	SMOTE	93.0 (92.4,93.6)	95.4 (95.2,95.7)
		CTGAN	91.0 (88.1,93.6)	94.7 (93.2,95.8)
		CART	89.7 (88.1,90.4)	95.5 (94.9,96.1)
Joint	Sensitivity	SMOTE	54.1 (47.5,59.1)	57.2 (50.3,63.8)
		CTGAN	87.1 (85.3,91.8)	69.0 (65.6,72.3)
		CART	75.6 (69.1,78.7)	60.0 (55.0,64.3)

	Specificity	SMOTE	83.7 (74.5,92.2)	85.4 (73.2,89.5)
		CTGAN	38.6 (34.2,45.1)	71.7 (67.2,74.7)
		CART	50.1 (43.4,57.1)	79.7 (75.8,84.6)
	PPV	SMOTE	24.6 (17.7,29.7)	23.8 (18.2,29.5)
		CTGAN	16.8 (15.4,17.5)	20.9 (19.8,25.2)
		CART	18.1 (17.2,19.9)	23.3 (20.9,25.7)
	NPV	SMOTE	92.3 (90.9,92.7)	95.5 (94.4,96.2)
		CTGAN	94.9 (94.5,95.8)	95.6 (95.3,95.9)
		CART	93.1 (92.9,94.5)	94.6 (93.7,95.3)

Notes:
[1]Synthetic Minority Oversampling Technique (SMOTE); Conditional Tabular Generative Adversarial Networks (CTGAN); Classification and Regression Trees (CART); positive predictive value (PPV); negative predictive value (NPV)
[2] Estimates are mean over 500 experiments; 95% confidence intervals are quantile-based

Supplemental Table 4.5 Age group-stratified validity measure estimates (95% confidence intervals) in test dataset (N=7414; 18-45 years=293, >45 years=7121) for Least Absolute Shrinkage and Selection Operator logistic regression trained with augmented data in 500 synthetic data generation experiments

Distribution to generate synthetic observations	Validity metric	Synthetic data generation model for training data augmentation	Estimate, % (95% CI)	
			18-45 years	>45 years
Marginal	Sensitivity	SMOTE	57.7 (49.8,71.5)	58.1 (47.7,65.1)
		CTGAN	95.2 (97.6,100.0)	65.3 (62.3,68.7)
		CART	67.8 (65.7,97.9)	64.5 (54.6,70.0)
	Specificity	SMOTE	60.7 (54.3,66.7)	70.6 (66.6,88.7)
		CTGAN	14.4 (12.6,18.5)	76.5 (71.3,82.5)
		CART	41.0 (23.4,38.7)	76.8 (69.9,86.3)
	PPV	SMOTE	22.4 (21.0,23.9)	27.1 (22.8,36.7)
		CTGAN	15.7 (14.8,15.9)	24.5 (22.6,26.3)
		CART	21.6 (19.7,23.0)	22.9 (21.0,29.1)
	NPV	SMOTE	92.2 (90.8,94.3)	94.5 (94.1,95.1)
		CTGAN	99.4 (95.1,100.0)	95.6 (95.2,96.0)
		CART	93.7 (93.0,98.8)	95.2 (94.4,95.6)
Joint	Sensitivity	SMOTE	41.3 (32.1,56.5)	50.1 (45.3,64.4)
		CTGAN	70.0 (69.1,100.0)	70.0 (67.0,73.0)
		CART	85.4 (81.3,100.0)	61.1 (56.0,64.5)
	Specificity	SMOTE	80.9 (69.3,88.8)	83.5 (72.2,88.4)
		CTGAN	52.7 (48.9,64.4)	70.7 (67.2,74.2)

		CART	30.5 (24.5,35.5)	77.2 (73.4,82.1)
	PPV	SMOTE	25.1 (23.2,30.2)	25.7 (19.7,29.8)
		CTGAN	18.8 (14.9,20.1)	20.6 (19.4,21.7)
		CART	18.1 (17.6,18.8)	23.2 (21.5,26.4)
	NPV	SMOTE	88.7 (88.2,90.5)	94.0 (93.7,94.7)
		CTGAN	95.2 (91.4,100.0)	94.6 (93.3,95.4)
		CART	95.6 (93.5,100.0)	94.8 (94.5,95.1)

Notes:
[1]Synthetic Minority Oversampling Technique (SMOTE); Conditional Tabular Generative Adversarial Networks (CTGAN); Classification and Regression Trees (CART); positive predictive value (PPV); negative predictive value (NPV)
[2] Estimates are mean over 500 experiments; 95% confidence intervals are quantile-based

CHAPTER 5: SOCIAL DETERMINANTS OF HEALTH DOCUMENTATION IN ADMINISTRATIVE HEALTH DATA: A POPULATION-BASED STUDY IN MANITOBA, CANADA

5.1 Overview

This chapter examines social determinants of health (SDOH) diagnosis code documentation in Manitoba's administrative health data (AHD). SDOH documentation, while uncommon in AHD, was associated with both the type of data and patient characteristics. For example, individuals in the lowest income quintile (Q1) had higher odds of SDOH documentation in hospital discharge abstracts than individuals in the highest quintile (Q5). For physician billing claims, the converse was true. SDOH documentation was also associated with health region of residence and age. The odds of SDOH documentation were positively associated with healthcare use and presence of comorbidities.

Temporal trends over an 11-year period revealed that coding policy changes (i.e., Canadian Institute for Health Information's 2018 update mandating the use of codes for homelessness in hospital discharge abstracts) and external events (i.e., COVID-19) affected documentation of SDOH in AHD.

Diagnosis codes in AHD that capture information about an individual's social, education, housing and other related circumstances are associated with patient characteristics as well as health system factors. These associations might affect future development of valid code-based SDOH algorithms.

Publication Details

Haque MA, Nickel N, Turgeon M, Lix LM. Social determinants of health documentation in administrative health data: A population-based study in Manitoba, Canada. Manuscript in preparation, 2025.

5.2 Abstract

Background: Social determinants of health (SDOH) are associated with health inequities. Administrative health data (AHD) captures SDOH through diagnosis codes primarily for management purposes, not research, raising questions about the validity of this documentation. We evaluated the face validity of AHD to identify SDOH by examining whether coding patterns exhibited expected sociodemographic associations and temporal trends.

Methods: We conducted a retrospective population-based study (2012-2023) using linked AHD from the Manitoba Population Research Data Repository, including hospital discharge abstracts, physician billing claims, and health insurance registrations. SDOH were identified using International Classification of Diseases codes from multiple categories, such as social, economic, and psychosocial circumstances. Multivariable logistic regression tested the association of SDOH documentation in hospital and physician records, or in either data source, with patient socio-demographic, healthcare use, and comorbid characteristics. Segmented regression analysis tested for temporal trends in the prevalence of SDOH documentation. Odds ratios (ORs) from logistic regression and slope coefficients from segmented regression models and 95% confidence intervals (CIs) were reported.

Results: The prevalence of at least one SDOH code in either hospital or physician records was 1.1%. Males were more likely than females to have SDOH documentation in hospital discharge abstracts (OR 2.0, 95% CI 1.9–2.1), but less likely in physician claims. In hospital discharge

abstracts, SDOH documentation was higher in the lowest than the highest income quintile (OR=5.5; 95% CI 4.9-6.2), but the converse was true in physician claims (OR 0.8, 95% CI 0.7-0.8). SDOH documentation was associated with high healthcare utilization and presence of comorbidities. The odds of SDOH documentation were higher for Winnipeg than non-Winnipeg health region residents. Segmented regression analysis identified an increasing trend in economic circumstances codes in hospital records that began in 2017. Documentation of psychosocial circumstances began to increase during the COVID-19 pandemic (i.e., 2020/21), followed by a decline.

Conclusions: The observed patterns, such as the socioeconomic gradient and trends responsive to external events such as the COVID-19 pandemic, support the face validity of using AHD to identify SDOH. However, the prevalence of ADH documentation is low, suggesting substantial misclassification.

5.3 Background

The social determinants of health (SDOH) refer to the conditions in which individuals are born, grow, live, work, and age, shaped by social, economic, and political systems (1). These determinants significantly contribute to health inequities observed within and between populations. For example, individuals experiencing economic hardships or housing instability are at increased risk of chronic illness, mental health disorders, and are less likely to access healthcare services (2).

Administrative health data (AHD), such as hospital discharge abstracts and physician billing claims, capture information on SDOH through specific International Classification of Diseases (ICD) diagnosis codes (3). These codes are intended to document non-medical factors

influencing health status and contact with health services (4–6). However, since AHD was primarily developed for administrative rather than research purposes, the validity of these codes requires rigorous assessment (7).

Two types of validity are criterion and face (8). Criterion validity can be evaluated by comparing AHD to an external reference standard, such as clinical data, to estimate AHD accuracy, sensitivity, and specificity (9). Face validity involves using multiple techniques or methods to determine whether a data element (e.g., SDOH documentation) appears to be a reasonable representation of the intended concept; it does not use an external reference standard data source (10). Since obtaining a reference standard for SDOH is often challenging due to the limited availability of external data sources for direct comparisons (11), examining patterns of SDOH documentation in AHD is a potential approach to assess the face validity of this documentation.

While some studies have assessed the validity of ICD codes for specific SDOH, such as homelessness (12) or food insecurity (13), studies about SDOH categories in AHD can aid in assessing system-wide documentation patterns and might be useful to inform the development of algorithms for measuring social risk in populations. Moreover, research on this topic has often not focused on the relationship of individual-level factors, such as sex, age, and healthcare utilization, with SDOH documentation (12,14). This knowledge gap limits the ability of researchers and policymakers to use AHD to accurately identify population-level social risks, which is a necessary first step for designing effective, targeted interventions (15,16). Face validity is further supported when temporal trends in specific SDOH codes change is associated with policy changes or external events (17).

This study assessed the face validity of SDOH documentation in AHD. The objectives were to 1) describe the presence of SDOH documentation in AHD with socio-demographic, healthcare use, and comorbid characteristics of individuals, and 2) test for temporal trends in SDOH documentation in AHD.

5.4 Methods

A retrospective cohort study was conducted using AHD from April 1, 2012 to March 31, 2023. The study followed the Reporting of Studies Conducted Using Observational Routinely Collected Data (RECORD) guidelines (Additional file 1) (18).

5.4.1 Data Sources

Study data were from the Population Research Data Repository housed at the Manitoba Centre for Health Policy (MCHP) at the University of Manitoba (19). The data were linked at the individual level using a scrambled personal health identification number. Study data sources were the Manitoba Health Insurance Registry, hospital discharge abstracts, and physician billing claims. These data sources were used to define the study cohort and its characteristics, and to ascertain the presence of SDOH documentation.

The Manitoba Health Insurance Registry contains health insurance coverage, demographic, and residence information for individuals eligible to receive health services from the provincial ministry of health. Hospital discharge abstracts contain records from inpatient hospital stays. Physician billing claims include documentation of physician visits in outpatient settings. Hospital discharge abstracts contain diagnosis codes recorded using the 10th revision of ICD, Canadian adaptation (i.e., ICD-10-CA), while Manitoba physician billing claims employ the 9th revision of ICD with clinical modifications (i.e., ICD-9-CM). The Statistics Canada Census (20) was also utilized to derive area-level income quintiles.

5.4.2 Population and Study Cohort

The study population consisted of individuals with health insurance coverage, as defined from the Manitoba Health Insurance Registry, between April 1, 2012, and March 31, 2023. In addition, individuals had to have at least one hospital discharge abstract with a discharge date in the study observation period or at least one physician billing claim with a service date in the study observation period to be retained in the cohort. Individuals were excluded from the cohort if they had only one or more physician billing claims with a service date prior to April 1, 2015; this exclusion was applied because in Manitoba, diagnoses in physician billing claims prior to this date were recorded only to three digits, while as of April 1, 2015, diagnoses in physician billing claims were recorded to five digits. We required five-digit codes to identify specific categories of SDOH documentation.

The study index date was assigned as the date of the most recent healthcare encounter. If an individual had multiple healthcare encounters, the index date was the most recent date among them. For individuals with records in both data sources, the index date was defined as the later date between the hospital discharge date and the physician service date.

5.4.3 Study Outcome

The outcome was the presence (yes vs. no) of SDOH documentation for each individual. The outcome was defined based on the presence of at least one code from a set of ICD codes (Additional file 2) used in prior research to identify SDOH in AHD (21,22). These codes fall under the category “Factors influencing health status and contact with health services” in ICD-10-CA and from the chapter on “Supplementary classification of factors influencing health status and contact with health services” in ICD-9-CM (4–6). These codes provide supplementary information regarding the reason for hospitalization or physician consultation.

The ICD codes used to define SDOH documentation were categorized as: social, economic, environmental, and health education in ICD-10-CA (21). In ICD-9-CM they were categorized as: alcoholism and drug addiction in family, mental health services for victim and perpetrator of abuse or history of abuse, other problems related to primary support group, including family circumstances, problems related to care provider dependency, problems related to education and literacy, problems related to employment and unemployment or occupational maladjustment, problems related to housing and economic circumstances, problems related to legal circumstances, and problems related to other psychosocial circumstances (22). The categorization differs between data sources because the structure and available codes in ICD-10-CA (at least four digits) and ICD-9-CM (at least five digits) are inherently different, with ICD-9-CM offering a more granular classification system.

5.4.4 Covariates

Socio-demographic characteristics of the study cohort, as defined from the Manitoba Health Insurance Registry, included age (years), sex, place of residence (urban or rural, as defined by Statistics Canada's standard geographic classification and based on postal code of residence), and health region (Winnipeg, Prairie Mountain, Southern, Interlake-Eastern, and Northern). Area-level income quintile, measured at the level of a dissemination area (consisting of 400-700 persons) based on total household income, was obtained from the Statistics Canada Census using postal code from the Registry (23).

Other characteristics of the study cohort were defined from hospital discharge abstracts and physician billing claims. They included the number of hospital visits and physician visits measured over the study period. The Charlson Comorbidity Index (CCI) score was calculated for each individual over the study period using both hospital discharge abstracts and physician

billing claims (24). This index is based on a list of 17 conditions identified from diagnoses in hospital discharge abstracts and physician billing claims (25). Each condition is assigned a weight from 1 to 6, and the index score is calculated as the sum of the weights for all identified conditions. A score of zero represents the absence of comorbidities, while higher scores indicate greater numbers of comorbid conditions.

5.4.5 Statistical Analysis

The cohort was described using frequencies and percentages for categorical variables and median with interquartile range (IQR) for continuous variables, by the outcome of SDOH documentation status (documented vs. not documented). Stratified analyses were conducted by the source of SDOH documentation: hospital discharge abstracts only, physician billing claims only, and either source (i.e., SDOH documentation in either hospital discharge abstracts or physician billing claims). The analyses were conducted in SAS 9.4 (26) and R 4.4.1 (27).

Multivariable logistic regression models were used to test the association between the outcome (i.e., presence of SDOH documentation) and the following covariates: sex, age, health region, income quintile, number of hospital visits, number of physician visits, and Charlson Comorbidity Index score (28). Age, number of hospitalizations, and physician billing claims were dichotomized based on median values. Firth's penalized version of the logistic regression model was used to avoid biased parameter estimates due to low prevalence of the outcome (29,30). Odds ratios (ORs) and 95% confidence intervals (CIs) were estimated. Model performance was assessed using measures of discrimination and calibration. Discrimination, a model's ability to distinguish those with and without the outcome, was measured using the C-statistic, where a value of 0.5 indicates no discrimination and 1.0 indicates perfect discrimination

(31). Calibration, the agreement between predicted and observed probabilities, was measured using quantile calibration error (QCE), where a value closer to 0 indicates better calibration (32).

The prevalence of SDOH documentation per 10,000 inpatient hospital stays or physician visits was calculated by fiscal year quarter: April to June, July to September, October to December, and January to March. Prevalence was calculated separately for hospital discharge abstracts and physician billing claims (22). Overall prevalence was also calculated; it was defined as the presence of any SDOH ICD code in a healthcare encounter. Category-specific prevalence was also calculated, defined as the presence of any SDOH ICD code belonging to a specific category in a healthcare encounter.

Segmented linear regression models were used to identify statistically significant changes in trend direction and estimate their magnitude (33). Slope coefficients with 95% CIs were reported for each segment (33). The presence of multiple segments of change in a trend can indicate periods of acceleration, deceleration, or stability.

5.5 Results

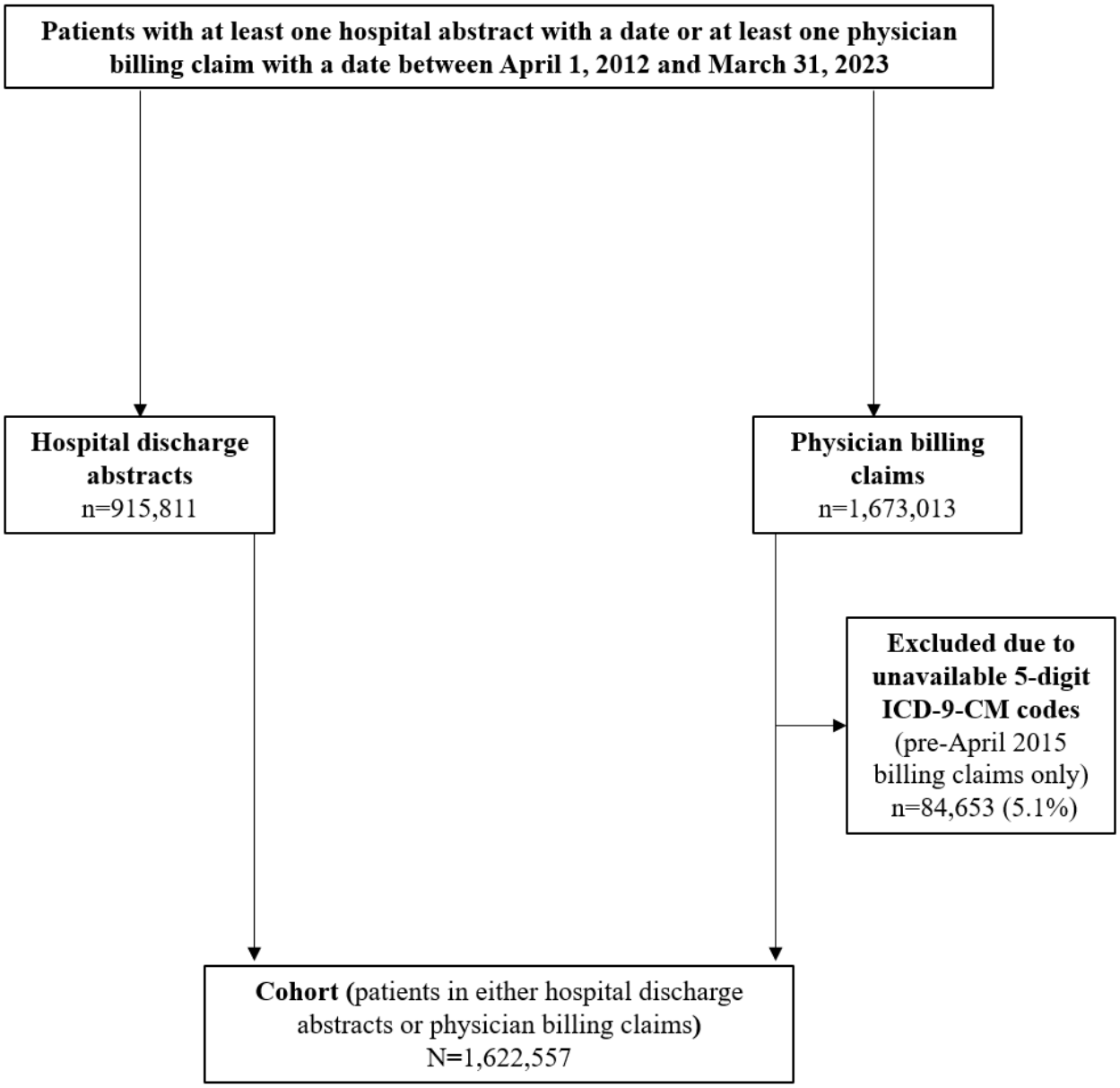
5.5.1 Study Cohort Characteristics

Table 5.1 summarizes the characteristics of the study cohort, stratified by the presence of SDOH documentation and data source. Over half of the cohort (56.4%) had at least one healthcare encounter captured in hospital discharge abstracts (i.e., an inpatient stay), while nearly the entire cohort (>97.9%) had at least one encounter captured in physician billing claims (i.e., a billed physician visit) (Figure 5.1).

Overall, 1.1% (n = 17,895) of the cohort had at least one documented SDOH code in either hospital discharge abstracts or physician billing claims, and <0.01% (n=157) had SDOH documentation in both data sources in the study period. The prevalence of SDOH documentation

was 0.7% in hospital discharge abstracts only and 0.8% in physician billing claims only. Among those with at least one SDOH code, only a small percentage had more than one code: 2.8% in hospital discharge abstracts (Additional file 2) and 4.6% in physician billing claims (Additional file 3).

Figure 5.1 Study cohort based on administrative health data from Manitoba (2012-2023)



Notable sex-based differences were observed in the results from the analyses stratified by data source. Females had a higher prevalence of SDOH documentation in physician billing claims (63.7%) and in the combined cohort (56.9%). Conversely, SDOH documentation in hospital discharge abstracts was more prevalent among males (57.0%). The median age was consistently higher among individuals with SDOH documentation than among those without it. A greater percentage of individuals with SDOH documentation resided in urban areas compared to those without, particularly in physician billing claims (74.3%) and in either hospital discharge abstracts or physician billing claims (69.9%). Individuals residing in the Winnipeg health region accounted for the majority of SDOH-documented cases, with 69.2% of cases identified through hospital discharge abstracts, 74.1% from physician billing claims, and 72.4% from either source. This is despite these individuals constituting less than two-thirds (58.0%) of the total study cohort. The Northern region consistently had the lowest percentage (≤5.3%) of SDOH documentation in different data sources, comparable to its representation within the overall study cohort. Individuals in the lowest income quintile (Q1) accounted for the highest percentage of SDOH-documented cases: 39.9% of cases from hospital discharge abstracts and 25.3% of cases from either data source. Median hospital and physician visits were notably higher among individuals with SDOH documentation. For example, in either source (i.e., SDOH documentation in either hospital discharge abstracts or physician billing claims), those with SDOH documentation had a median of 3.0 hospital visits and 85.0 physician visits, compared to 2.0 and 36.0, respectively, among those without. A higher proportion of individuals with non-zero CCI scores were observed among the SDOH-documented group across all data sources (e.g., 55.5% in hospital discharge abstracts).

Table 5.1 Characteristics of the study cohort stratified by presence of social determinants of health (SDOH) documentation and administrative data source

Characteristic	**Hospital discharge abstracts only** (N = 915,811)		**Physician billing claims only** (N = 1,588,360)		**Either data source** (N = 1,622,557)	
	No SDOH documentation n=909,823 (99.3%)	**SDOH documentation** n=5988 (0.7%)	**No SDOH documentation** n=157,6296 (99.2%)	**SDOH documentation** n=12,064 (0.8%)	**No SDOH documentation** n=160,4662 (98.9%)	**SDOH documentation** n=17,895 (1.1%)
Sex						
Male	402,521 (44.2)	3413 (57.0)	783,628 (49.7)	4374 (36.3)	797,132 (49.7)	7710 (43.1)
Female	503,302 (55.8)	2575 (43.0)	792,668 (50.3)	7690 (63.7)	807,530 (50.3)	10,185 (56.9)
Age, median (IQR)	43.0 (46.0)	49.0 (35.0)	38.0 (40.0)	46.0 (22.0)	39.0 (41.0)	47.0 (25.0)
Residence location						
Urban	500,270 (55.0)	3646 (60.9)	960,796 (60.9)	8957 (74.3)	975,656 (60.8)	12,510 (69.9)
Rural	370,201 (40.7)	1421 (23.7)	592,920 (37.6)	2949 (24.4)	604,922 (37.7)	4340 (24.3)
Unknown	39,352 (4.3)	921 (15.4)	22,580 (1.4)	158 (1.3)	24,084 (1.5)	1045 (5.8)
Health region						
Winnipeg	483,304 (53.1)	4143 (69.2)	909,864 (57.7)	8928 (74.1)	924,524 (57.6)	12,949 (72.4)
Prairie Mountain	129,165 (14.2)	771 (12.9)	197,452 (12.5)	646 (5.3)	202,173 (12.6)	1407 (7.9)
Southern	143,147 (15.7)	389 (6.5)	238,721 (15.1)	933 (7.7)	242,746 (15.1)	1315 (7.4)
Interlake-Eastern	95,141 (10.5)	375 (6.3)	147,837 (9.4)	945 (7.8)	150,637 (9.4)	1310 (7.3)
Northern	59,066 (6.5)	310 (5.2)	82,422 (5.2)	612 (5.1)	84,582 (5.3)	914 (5.1)
Income quintile						
Q1 (lowest)	182,028 (20.0)	2389 (39.9)	324,999 (20.6)	2979 (18.1)	330,577 (20.6)	4526 (25.3)
Q2	178,654 (19.6)	1025 (17.1)	306,191 (19.4)	2236 (18.5)	312,397 (19.5)	3231 (18.1)
Q3	173,631 (19.1)	703 (11.7)	307,251 (19.5)	2194 (18.2)	313,148 (19.5)	2872 (16.1)

Q4	173,451 (19.1)	640 (10.7)	310,006 (19.7)	2403 (19.9)	314,775 (19.6)	3017 (16.9)
Q5 (highest)	162,707 (17.9)	310 (5.2)	305,269 (19.4)	2894 (24.0)	309,681 (19.3)	3204 (17.9)
Unknown	39,352 (4.3)	921 (15.4)	22,580 (1.4)	158 (1.3)	24,084 (1.5)	1045 (5.8)
Number of hospital visits, median (IQR)	2.0 (2.0)	6.0 (7.0)	N/A	N/A	2.0 (2.0)	3.0 (4.0)
Number of physician visits, median (IQR)	N/A	N/A	76.0 (87.0)	102.0 (93.0)	36.0 (64.0)	85.0 (105.0)
Charlson Comorbidity Index score						
0	672,972 (74.0)	2662 (44.5)	930,812 (59.1)	5631 (46.7)	934,095 (58.2)	7271 (40.6)
≥1	236,851 (26.0)	3326 (55.5)	645,484 (40.9)	6433 (53.3)	670,567 (41.8)	10,624 (59.4)

Note: Inter-quartile range (IQR); values reported are frequency (%) unless otherwise stated; not applicable (N/A)

5.5.2 Association Between Cohort Characteristics and SDOH Documentation

The multivariable logistic regression models tested associations between the presence of SDOH documentation and patient characteristics. In hospital discharge abstracts, demographic characteristics showed significant associations with SDOH documentation. Female patients had lower odds of SDOH documentation compared to males (OR 0.5, 95% CI 0.5-0.6). Patients older than the median age (43.0 years) had lower odds of documentation (OR 0.4, 95% CI 0.3-0.4) compared to their younger counterparts. Documentation likelihood increased substantially with healthcare utilization, as patients with hospital visits above the median (2.0) showed 9.0 times greater odds (95% CI 8.4-9.7) than patients below the median. Patients with CCI scores ≥1 had 2.0 times greater odds of documentation (95% CI 1.9-2.2) compared to those with no comorbidities.

Table 5.2 Odds ratios (ORs) and 95% confidence intervals (CIs) for the association of cohort characteristics with social determinants of health (SDOH) documentation, stratified by data sources

Characteristic	Hospital discharge abstracts only	Physician billing claims only	Either data source
Sex			
Female	0.5 (0.5-0.6)	1.6 (1.5-1.7)	1.0 (1.0-1.1)
Male	-	-	-
Age			
>Median (43, 38, 39)	0.4 (0.3-0.4)	1.9 (1.8-2.0)	1.1 (1.1-1.2)
<=Median	-	-	-
Health region			
Winnipeg	1.8 (1.6-2.0)	1.1 (1.0-1.2)	0.9 (0.9-1.1)
Prairie Mountain	1.0 (0.9-1.1)	0.4 (0.3-0.4)	0.4 (0.4-0.5)
Southern	0.6 (0.5-0.7)	0.5 (0.4-0.5)	0.4 (0.3-0.5)
Interlake-Eastern	0.9 (0.7-1.0)	0.7 (0.6-0.8)	0.6 (0.5-0.7)
Northern	-	-	-
Income quintile			
Q1 (lowest)	5.5 (4.9-6.2)	0.8 (0.7-0.8)	1.4 (1.3-1.5)
Q2	2.6 (2.3-2.9)	0.8 (0.7-0.8)	1.0 (0.9-1.1)

Q3	1.9 (1.7-2.2)	0.8 (0.7-0.8)	0.9 (0.8-1.0)
Q4	1.8 (1.6-2.1)	0.8 (0.8-0.9)	0.9 (0.9-1.0)
Q5 (highest)	-	-	-
Unknown	10.9 (9.5-12.4)	0.5 (0.4-0.6)	2.7 (2.5-2.9)
Number of hospital visits			
>Median (2, N/A, 2)	9.0 (8.4-9.7)	N/A	1.9 (1.8-2.0)
<=Median	-	-	-
Number of physician visits			
>Median (N/A, 76, 36)	N/A	1.8 (1.7-1.9)	3.1 (3.0-3.3)
<=Median	-	-	-
Charlson Comorbidity Index score			
≥1	2.0 (1.9-2.2)	2.7 (2.5-3.0)	1.9 (1.8-2.0)
0	-	-	-
Model performance			
C statistic	0.83	0.68	0.73
Quantile calibration error	<0.01	<0.01	<0.01

Note: - represents reference category

In physician billing claims, female patients had a greater odds of SDOH documentation (OR 1.6; 95% CI 1.5-1.7) than males. Older patients (above the median age of 38 years) had a greater odds of SDOH documentation than younger patients (OR 1.9, 95% CI 1.8-2.0). Healthcare utilization measures were strongly associated with SDOH documentation, particularly for physician visits above the median (76.0) (OR 1.8, 95% CI 1.7-1.9). Patients with a higher number of comorbid conditions, based on the Charlson Comorbidity Index had a higher odds of SDOH documentation in physician billing claims (OR 2.7, 95% CI 2.5-3.0).

Among individuals with SDOH documentation in hospital discharge abstracts, those residing in Winnipeg had 1.8 times higher odds of documentation (95% CI 1.6-2.0) than those in the Northern region. In contrast, individuals from the Southern region had 40% lower odds (OR 0.6, 95% CI 0.5-0.7). Among individuals with SDOH documented in physician billing claims, regional variation was less pronounced than that observed in hospital discharge abstracts. The odds of SDOH documentation for individuals in Winnipeg were similar to those in the Northern

region (OR 1.1, 95% CI 1.0-1.2), while individuals in the Prairie Mountain health region had 60% lower odds (OR 0.4, 95% CI 0.3-0.4).

For individuals with SDOH documentation in hospital discharge abstracts, there was a notable linear relationship where lower income was associated with higher odds of documentation. Compared to individuals in the highest income quintile (Q5), those in the lowest quintile (Q1) had 5.5 times higher odds of documentation (95% CI 4.9-6.2). This association was strongest for individuals with a missing income quintile (OR 10.9, 95% CI 9.5-12.4). Conversely, among individuals with documentation in physician billing claims, lower income was associated with slightly lower odds of documentation (e.g., Q1 OR 0.8, 95% CI 0.7-0.8).

In the multivariable logistic regression model for the patients with SDOH documentation in either hospital discharge abstracts or physician billing claims, healthcare utilization measures were positively associated with the odds of SDOH documentation. Specifically, individuals with a greater number of physician visits had a higher odds of SDOH documentation that individuals with fewer physician visits (OR 3.1, 95% CI 3.0-3.3). Individuals in the lowest income quintile (Q1) had a higher odds of documentation (OR 1.4, 95% CI 1.3-1.5) than individuals in the highest income quintile. The odds of SDOH documentation for residents of all health regions outside of Winnipeg, specifically the Prairie Mountain, Southern, and Interlake-Eastern regions, were 40-60% lower than for residents of the Northern health region.

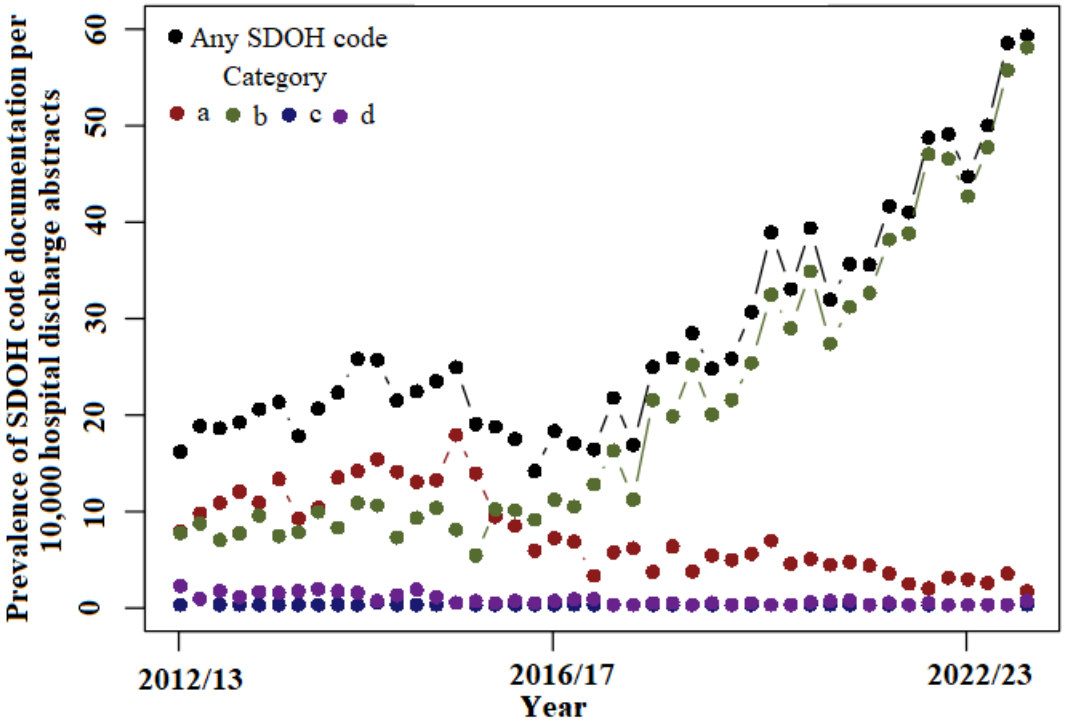
The multivariable logistic regression model that tested the association of SDOH documentation in hospital discharge abstracts with patient characteristics had the highest discriminative performance; the C-statistic had a value of 0.83 (95% CI 0.82-0.84). The model for individuals in physician billing claims had the lowest discriminative performance with a C-statistic estimate of 0.68 (0.66-0.71). The QCE remained below 0.01 for all three models (Table

5.2). These small QCE values indicate high agreement between predicted probabilities and observed outcomes.

5.5.3 Trends in SDOH Documentation

The quarterly prevalence of SDOH documentation per 10,000 healthcare encounters showed substantial variation over time. The changes in the SDOH documentation pattern in hospital discharge abstracts were primarily derived from the "Economic" category of codes (Figure 5.2).

Figure 5.2 Temporal trends in social determinants of health (SDOH) documentation in hospital discharge abstracts

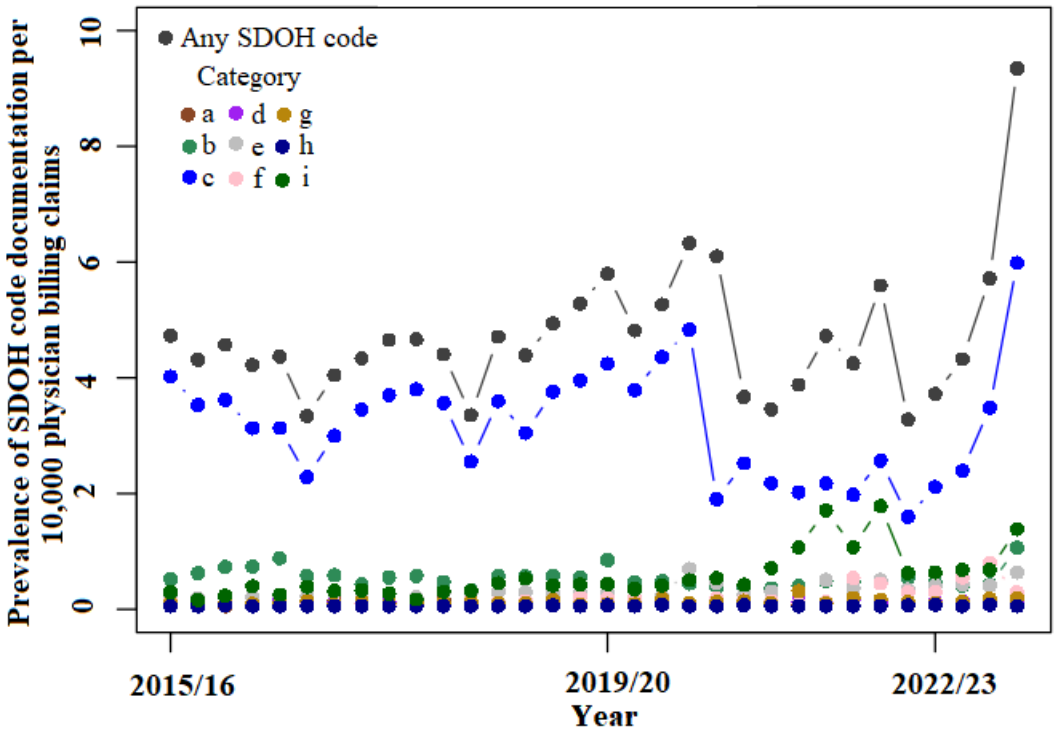


Note: categories (a=social, b=economic, c=environmental, and d=health education) for hospital discharge abstracts

In physician billing claims, the codes corresponding to "other problems related to primary support group, including family circumstances" and "problems related to other psychosocial

circumstances" categories contributed the most to the notable changes in the trend of SDOH documentation prevalence (Figure 5.3).

Figure 5.3 Temporal trends in social determinants of health (SDOH) documentation in physician billing claims

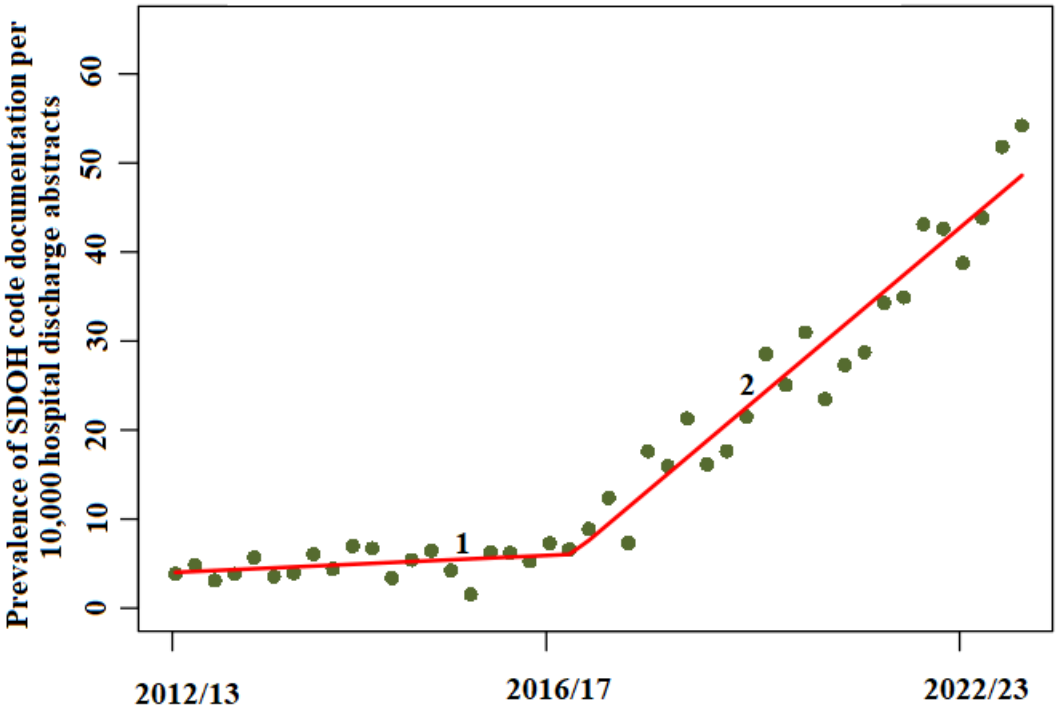


Note: categories (a=alcoholism and drug addiction in family, b=mental health services for victim and perpetrator of abuse or history of abuse, c=other problems related to primary support group, including family circumstances, d=problems related to care provider dependency, e=problems related to education and literacy, f=problems related to employment and unemployment or occupational maladjustment, g=problems related to housing and economic circumstances, h=problems related to legal circumstances, and i=problems related to other psychosocial circumstances) for physician billing claims.

Segmented regression analysis for the "Economic" category in hospital discharge abstracts identified a single join-point, initiating a phase characterized by an upward trend (slope of segment 2 = 1.9, 95% CI 1.7–2.1) beginning in the second quarter of fiscal year 2017/18 (Figure 5.4). In physician billing claims, documentation pertaining to the category "issues with the primary support group experienced" demonstrated a decline in the first quarter of 2020/21

(slope of segment 3 = -0.2, 95% CI -0.3–-0.1), followed by a substantial increase in the second quarter of 2022/23 (slope of segment 4 = 1.8, 95% CI 0.9–2.8) (Figure 5.5).

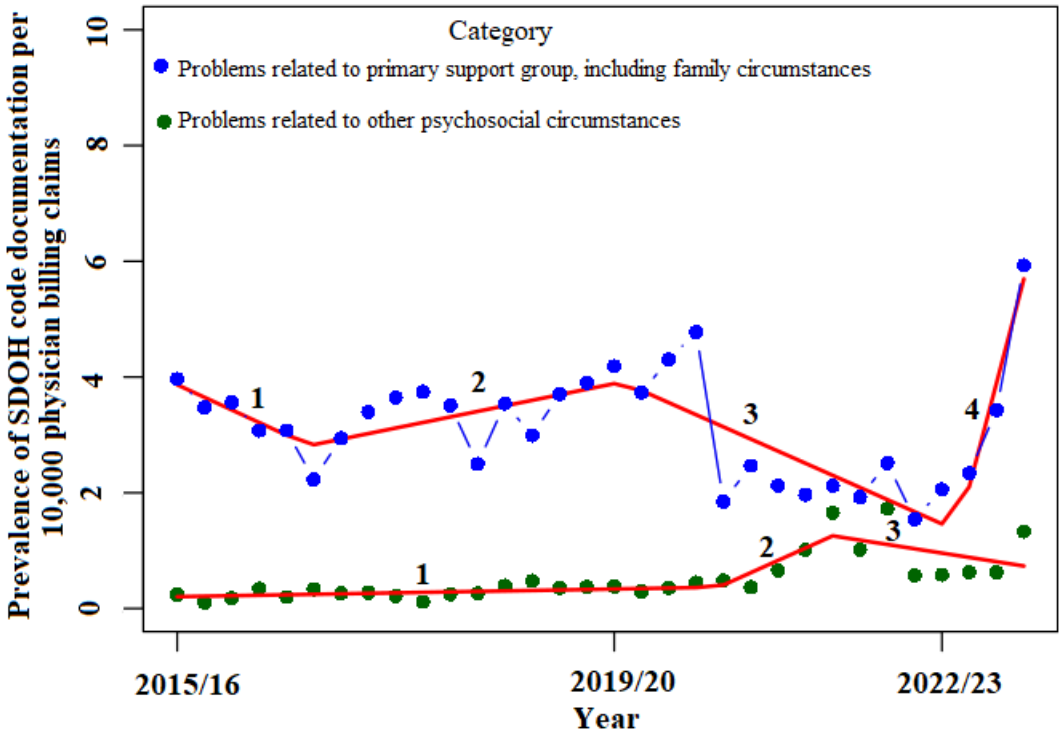
Figure 5.4 Segmented regression models for trends in types of social determinants of health (SDOH) documentation in hospital discharge abstracts (any code corresponding to the "Economic" category of codes)



Note: numbers on the fitted regression line (in red) represent segments in the trend

Moreover, SDOH documentation related to the category of other psychosocial circumstances began to rise in the second quarter of 2020/21 (slope of segment 2 = 0.2, 95% CI 0.1–0.4), with a subsequent decline observed by the end of the fiscal year 2021/22.

Figure 5.5 Segmented regression models for trends in types of social determinants of health (SDOH) documentation in hospital discharge abstracts (any code corresponding to the "Economic" category of codes)



Note: numbers on the fitted regression line (in red) represent segments in the trend

5.6 Discussion

This study assessed the face validity of SDOH documentation in AHD by examining the distribution of this documentation in terms of a broad spectrum of ICD codes, from economic hardship and housing instability to family circumstances and psychosocial stress, across data sources, patient characteristics, and over time. This contextual understanding is essential for researchers aiming to develop SDOH ascertainment algorithms for AHD. Overall, the low prevalence of SDOH documentation in AHD suggests potential under-documentation of social determinants in routinely-collected healthcare records. This aligns with previous research

indicating that structural and systemic barriers such as time constraints during healthcare encounters and the sensitive nature of SDOH-related disclosure may limit the identification and coding of SDOH within administrative datasets (14). Despite the low overall prevalence, the associations we observed support their face validity for the cases that are documented.

The observed differences in documentation by sex and geography highlight important sociodemographic patterns. Females had a higher prevalence of SDOH documentation in physician billing claims, whereas males had a higher prevalence in hospital discharge abstracts. This divergence may be attributed to differences in healthcare-seeking patterns and the nature of clinical encounters that lead to coding in each setting. For example, the act of a physician selecting a diagnostic code for a billing claim may capture different patient-provider interactions than the subsequent translation of clinical notes into codes by a hospital coder (35). The variation across health regions likely reflects differences in healthcare accessibility, socioeconomic conditions, and provider awareness of SDOH documentation. Individuals residing in urban areas, particularly in Winnipeg, had a higher prevalence of SDOH documentation than individuals in rural and Northern regions. This geographic variation suggests that region-specific factors, such as diagnostic practices and population demographics, may influence documentation (36).

5.6.1 Face Validity Assessment

Several findings supported the face validity of SDOH documentation in AHD. The socioeconomic gradient observed in hospital discharge abstracts showed a monotonic increase in the prevalence of SDOH documentation from the highest (Q5) to the lowest (Q1) income quintile, aligning with theoretical expectations about the relationship between poverty and social risk factors. Additionally, the finding that this gradient was most pronounced in hospital discharge abstracts, with individuals in the lowest income quintile showing greater odds of

documentation than individuals in the highest income quintile, suggests that inpatient settings may be particularly sensitive to capturing socioeconomic disparities, possibly because hospitalization often represents a crisis point where social vulnerabilities become more apparent to providers (37). Moreover, higher healthcare utilization (measured by hospital and physician visits) was positively associated with having at least one SDOH documentation, suggesting an expected pattern: patients with frequent healthcare interactions are more likely to have documented social determinants in AHD (38). Clinical context further supports face validity, as the strong association between SDOH documentation and comorbidity burden (CCI scores). This pattern is particularly important for health services research, as it suggests that AHD-based SDOH measures may be especially useful for studying populations where social and medical complexity intersect (39).

Temporal trends provide additional evidence for face validity, as they demonstrate documentation patterns responding to real-world events. The increase in economic-related SDOH codes beginning in the second quarter of fiscal year 2017/18 suggests that the introduction of mandatory homelessness documentation (Z590) in hospital discharge abstracts in 2018, as per the Canadian Institute for Health Information's updated coding guidelines, played a role in this trend (7). Similarly, the pandemic-related fluctuations in psychosocial code documentation, including the initial decline in primary support group codes in early 2020, followed by a substantial increase in mid-2022, mirror the documented impact of COVID-19 on interpersonal relationships, mental health, and stressors within household and family environments (40,41). These responsive patterns suggest that, when documented, SDOH codes related to psychosocial conditions in AHD reflect real social risk fluctuations.

5.6.2 Strengths and Limitations

A major strength of this study is the population-based design that captured SDOH documentation across Manitoba's healthcare system over an 11-year period. Unlike studies focused on a single SDOH category, we investigated multiple ICD codes representing different categories of SDOH. Moreover, the use of up to 5-digit ICD-9-CM codes, particularly for physician billing claims, allowed for a detailed examination of SDOH codes.

However, the study has some limitations. The study focused on AHD from a single Canadian province, which may limit the generalizability of the findings to other jurisdictions with different population characteristics. The descriptive analysis limits the ability to establish causal relationships between SDOH documentation and the observed associations. Marginalized individuals such as those experiencing homelessness, undocumented immigrants, or people living in remote communities may face barriers accessing the healthcare system (42). These individuals might not be represented in AHD.

5.6.3 Future Directions

Future studies could explore longitudinal SDOH documentation, examining how patient characteristics, such as age, sex, comorbidity burden, and frequency of healthcare utilization, influence the likelihood and consistency of SDOH coding over time (43). In addition, validation studies comparing SDOH codes in AHD to external sources, such as population health surveys, patient-reported outcome measures, or information captured in structured clinical screening tools, could be used to estimate the sensitivity, specificity, and predictive validity of coded data (44,45). Qualitative research could clarify how and when providers document SDOH in AHD (46).

5.6.4 Conclusions

This study examined the face validity of SDOH diagnosis code documentation in AHD. SDOH documentation was associated with patient characteristics, including socio-demographics, healthcare use, and comorbid conditions. The direction and magnitude of these associations were not always consistent for SDOH documentation found in hospital discharge abstracts and physician billing claims, suggesting the healthcare setting (acute care versus primary care) influences documentation practices.

The findings could support further studies about the relationship between SDOH and both healthcare use and health outcomes. The identified patterns, such as the socioeconomic gradient and the sex-based differences between data sources, provide information that could be helpful in developing and refining SDOH algorithms. Specifically, these findings can inform the selection of the optimal data source (hospital abstracts vs. billing claims) for measuring social risks within specific patient subgroups, thereby enhancing the accuracy of algorithms developed for AHD.

5.7 References

1. World Health Organization. Social determinants of health [Internet]. [cited 2025 Mar 24]. Available from: https://www.who.int/health-topics/social-determinants-of-health
2. Singh A, Daniel L, Baker E, Bentley R. Housing disadvantage and poor mental health: a systematic review. American Journal of Preventive Medicine. 2019 Aug;57(2):262-72.
3. Cadarette SM, Wong L. An introduction to health care administrative data. Canadian Journal of Hospital Pharmacy. 2015;68(3):232-7.
4. Find-A-Code. ICD-9-CM - Medical Codes [Internet]. [cited 2025 May 1]. Available from: https://www.findacode.com/code-set.php?set=ICD9&i=2432
5. Centers for Disease Control and Prevention. International Classification of Diseases, Clinical Modification (ICD–CM) - Health, United States [Internet]. [cited 2025 Mar 21]. Available from: https://www.cdc.gov/nchs/hus/sources-definitions/icd-cm.htm
6. Canadian Institute for Health Information. Version 2022 ICD-10-CA/CCI classifications, Canadian Coding Standards and related products | CIHI [Internet]. [cited 2025 Mar 21]. Available from: https://www.cihi.ca/en/version-2022-icd-10-cacci-classifications-canadian-coding-standards-and-related-products
7. Hinds A, Lix LM, Smith M, Quan H, Sanmartin C. Quality of administrative health databases in Canada: a scoping review. Canadian Journal of Public Health. 2016 Jan;107(1):e56-61.
8. Lim WM. A typology of validity: content, face, convergent, discriminant, nomological and predictive validity. Journal of Trade Science. 2024 Jul 17;12(3):155-79.
9. Lix LM, Renoux C, Moriello C, Choi KL, Dormuth CR, Fisher A, et al. Validity of diagnoses of SARS-CoV-2 infection in Canadian administrative health data: a multiprovince,

population-based cohort study. Canadian Medical Association Open Access Journal. 2023 Sep 1;11(5):E790-8.

10. Cunningham CT, Cai P, Topps D, Svenson LW, Jetté N, Quan H. Mining rich health data from Canadian physician claims: features and face validity. BMC Research Notes. 2014 Oct 1;7(1):682.
11. Ajao Y. Improving social determinants of health data collection for patient-centered care. COPE Health Solutions. Available from: https://copehealthsolutions.com/wp-content/uploads/2020/09/Improving-Social-Determinants-of-Health-Data-Collection-for-Patient-Centered-Care.pdf
12. Prophetis ED, Beck K, Ridgeway D, Chuang J, Richard L, Durbin A, et al. Trends in hospital coding for people experiencing homelessness in Canada, 2015–2020: a descriptive study. Canadian Medical Association Open Access Journal. 2023 Nov 1;11(6):E1188-96.
13. Gupta N, Sheng Z. A population-based study of the association between food insecurity and potentially avoidable hospitalization among persons with diabetes using linked survey and administrative data. International Journal of Population Data Science. 4(1):1102.
14. Agarwal AR, Prichett L, Jain A, Srikumaran U. Assessment of use of ICD-9 and ICD-10 codes for social determinants of health in the US, 2011-2021. JAMA Network Open. 2023 May 9;6(5):e2312538.
15. Jack L. Advancing health equity, eliminating health disparities, and improving population health. Preventing Chronic Disease. 2021 Aug 12;18:E79.
16. Koh HK, Oppenheimer SC, Massin-Short SB, Emmons KM, Geller AC, Viswanath K. Translating research evidence into practice to reduce health disparities: a social determinants approach. American Journal of Public Health. 2010 Apr;100(S1):S72-80.

17. Kneipp SM, Schwartz TA, Drevdahl DJ, Canales MK, Santacroce S, Santos HPJ, et al. Trends in health disparities, health inequity, and social determinants of health research: a 17-year analysis of NINR, NCI, NHLBI, and NIMHD funding. Nursing Research. 2018 Jun;67(3):231.
18. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Medicine. 2015 Oct;12(10):e1001885.
19. Smith M, Roos LL, Burchill C, Turner K, Towns DG, Hong SP, et al. Health services data: managing the data warehouse: 25 years of experience at the Manitoba Centre for Health Policy. In: Levy A, Goring S, Gatsonis C, Sobolev B, van Ginneken E, Busse R, editors. Health services evaluation. New York, NY: Springer US; 2019. p. 19-45. (Health services research). Available from: https://doi.org/10.1007/978-1-4939-8715-3_3
20. Government of Canada SC. Census of population. Statistics Canada. Available from: https://www12.statcan.gc.ca/census-recensement/index-eng.cfm
21. Friedman NL, Banegas MP. Toward addressing social determinants of health: a health care system strategy. Permanente Journal. 2018 Oct 22;22:18-095.
22. Torres JM, Lawlor J, Colvin JD, Sills MR, Bettenhausen JL, Davidson A, et al. ICD social codes: an underutilized resource for tracking social needs. Medical Care. 2017 Sep;55(9):810.
23. Manitoba Centre for Health Policy. Term: Income Quintiles / Income Quintile [Internet]. [cited 2024 Feb 11]. Available from: http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=102882

24. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Diseases. 1987 Jan 1;40(5):373-83.
25. Manitoba Centre for Health Policy. Concept: Charlson Comorbidity Index [Internet]. [cited 2025 Mar 4]. Available from: http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1098
26. SAS: Maintenance Releases and Product Upgrades [Internet]. [cited 2025 Mar 4]. Available from: https://support.sas.com/software/updates/index.html
27. R: The R Project for Statistical Computing [Internet]. [cited 2025 Feb 26]. Available from: https://www.r-project.org/
28. Kleinbaum DG, Klein M. Modeling strategy guidelines. In: Kleinbaum DG, Klein M, editors. Logistic regression: a self-learning text. New York, NY: Springer; 2010. p. 165-202. (Statistics for biology and health). Available from: https://doi.org/10.1007/978-1-4419-1742-3_6
29. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993 Mar 1;80(1):27-38.
30. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Statistics in Medicine. 2002 Aug 30;21(16):2409-19.
31. Nattino G, Pennell ML, Lemeshow S. Assessing the goodness of fit of logistic regression models in large samples: a modification of the Hosmer-Lemeshow test. Biometrics. 2020 Jun 1;76(2):549-60.

32. Song H, Diethe T, Kull M, Flach P. Distribution calibration for regression. In: Proceedings of the 36th International Conference on Machine Learning. PMLR; 2019. p. 5897-906. Available from: https://proceedings.mlr.press/v97/song19a.html
33. Feder PI. On asymptotic distribution theory in segmented regression problems-- identified case. Annals of Statistics. 1975;3(1):49-83.
34. Tzeng IS, Chien KL, Tu YK, Chen JY, Ng CY, Chien CY, et al. Segmented regression analysis of emergency departments patient visits from septicemia in Taiwan. Health Policy and Technology. 2018 Jun 1;7(2):149-55.
35. Crear-Perry J, Correa-de-Araujo R, Lewis Johnson T, McLemore MR, Neilson E, Wallace M. Social and structural determinants of health inequities in maternal health. Journal of Women's Health. 2021 Feb;30(2):230-5.
36. Kepper MM, Walsh-Bailey C, Prusaczyk B, Zhao M, Herrick C, Foraker R. The adoption of social determinants of health documentation in clinical settings. Health Services Research. 2023;58(1):67-77.
37. Hatef E, Kharrazi H, Nelson K, Sylling P, Ma X, Lasser EC, et al. The association between neighborhood socioeconomic and housing characteristics with hospitalization: results of a national study of veterans. Journal of the American Board of Family Medicine. 2019;32(6):890-903.
38. Alderwick H, Gottlieb LM. Meanings and misunderstandings: a social determinants of health lexicon for health care systems. Milbank Quarterly. 2019 Jun;97(2):407-19.
39. Schillinger D. The intersections between social determinants of health, health literacy, and health disparities. In: Health literacy in clinical practice and public health. IOS Press; 2020. p. 22-41. Available from: https://ebooks.iospress.nl/doi/10.3233/SHTI200020

40. Boden M, Zimmerman L, Azevedo KJ, Ruzek JI, Gala S, Abdel Magid HS, et al. Addressing the mental health impact of COVID-19 through population health. Clinical Psychology Review. 2021 Apr 1;85:102006.
41. Baptiste D, Commodore-Mensah Y, Alexander KA, Jacques K, Wilson PR, Akomah J, et al. COVID-19: shedding light on racial and health inequities in the USA. Journal of Clinical Nursing. 2020 Aug;29(15-16):2734-6.
42. Aborode AT, Lawal L, Agwuna FO, Adewunmi RO, Ubechu SC, Badri R. Health disparities among illegal immigrants and homeless people in the USA: a struggle within. IJS Global Health. 2023 Nov;6(6):e0362.
43. Hughes AE, Tiro JA, Balasubramanian BA, Skinner CS, Pruitt SL. Social disadvantage, healthcare utilization, and colorectal cancer screening: leveraging longitudinal patient address and health records data. Cancer Epidemiology, Biomarkers & Prevention. 2018 Dec;27(12):1424-32.
44. Wray CM, Vali M, Abraham A, Zhang A, Walter LC, Keyhani S. Validation of administrative measures of social and behavioral risk in veterans affairs medical records. Journal of General Internal Medicine. 2019 Jun 1;34(6):796-8.
45. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. Journal of the American Medical Informatics Association. 2021 Oct 19;29(1):187-96.
46. Sokol RL, Ammer J, Stein SF, Trout P, Mohammed L, Miller AL. Provider perspectives on screening for social determinants of health in pediatric settings: a qualitative study. Journal of Pediatric Health Care. 2021 Nov 1;35(6):577-86.

5.8 SUPPLEMENTARY MATERIAL

Supplemental Table 5.1 The Reporting of studies Conducted using Observational Routinely-collected Data (RECORD) statement – checklist of items, extended from the Strengthening the Reporting of Observational Studies in Epidemiology (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	184-186	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.	184-186
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	187-188		
Objectives	3	State specific objectives, including any prespecified hypotheses	188		
Methods					

Study Design	4	Present key elements of study design early in the paper	188-189		
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	188-190		
Participants	6	Cohort study - Give the l methods of selection of ow-up 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 Cohort study - For matched nber of exposed and unexposed Case-control study - For matched studies, give matching criteria and the number of controls per case	188-190	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If 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.	188-190
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	189-190, Additional file 2, 3	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	189-190, Additional file 2, 3

				should be provided.	
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	189		

Bias	9	Describe any efforts to address potential sources of bias	N/A		
Study size	10	Explain how the study size was arrived at	193, Figure 5.1		
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	189-190		

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	199-201		
Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population.	188-190
				RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	

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.	189
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) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram	193, Figure 5.1	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.	193, Figure 5.1
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)	194-195, Table 5.1		

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	193-194, Table 5.1		
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	191-199, Figure 5.2, Figure 5.3, Table 5.2		
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	Additional file 2, Additional file 3,		
Discussion					
Key results	18	Summarise key results with reference to study objectives	199-203		

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	202	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.	201-203
Interpretation	20	Give a cautious overall interpretation of results considering objectives,	202-204		
		limitations, multiplicity of analyses, results from similar studies, and other relevant evidence			
Generalisability	21	Discuss the generalisability (external validity) of the study results	203		
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	iii		
Accessibility of protocol, raw data, and programming code		..		RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	242

Additional file 2

Supplemental Table 5.2 Social determinants of health (SDOH) code documentation in hospital discharge abstracts, 2012/13 to 2022/23 (n=5988)

SDOH category	ICD-10-CA code	n (%)
Social		1591 (26.6)
Family/Caregiver stress	Z638	360 (6.0)
Insufficient social insurance or welfare support	Z597	13 (0.2)
Need assistance with community resources	Z748	519 (8.6)
Unavailability or inaccessibility of other helping agencies	Z754	717 (11.9)
Social isolation	Z624	**
Problems related to other legal circumstances/release from prison	Z625	**
Economic		4387 (73.3)
Financial problem/Housing or economic circumstance/Inadequate material resources	Z599	102 (1.7)
Food insecurity,	Z594	10 (0.2)
Homelessness	Z590	4264 (71.2)
Intentional underdosing of medication due to financial stressors	Z91120	**
Unemployment	Z560	32 (0.5)
Low income	Z596	22 (0.4)
Environmental		**
Fall risk	Z9181	**
Stressful work schedule	Z563	**
Foster care status	Z6221	**
Problem related to social environment	Z629	**
Health education		112 (1.8)
Dental well-care counseling	Z7189	**
Referral to county mental health agency	Z6881	**
Illiteracy and low-level literacy	Z550	**
Nutrition and exercise counseling	Z713	105 (1.8)

Note: ** indicates suppressed numbers due to small cell size

Supplemental Table 5.3 Social determinants of health (SDOH) code documentation in hospital discharge abstracts by SDOH category (n = 5,988)

Any SDOH code (20 codes)	**n (%)**
1	5819 (97.2)
≥2	169 (2.8)
Social code (6 codes)	
1	1566 (26.1)
≥2	25 (0.4)
Economic code (6 codes)	
1	4344 (72.5)
≥2	43 (0.7)
Environmental code (4 codes)	**
1	**
≥2	**
Health education code (4 codes)	
1	112 (1.9)
≥2	**

Note: ** represents suppressed numbers due to small cell size

Additional file 3

Supplemental Table 5.4 Social determinants of health (SDOH) code documentation in physician billing claims, 2015/16 to 2022/23 (n=12,064)

SDOH Category	ICD-9-CM Code	n (%)
Alcoholism and drug addiction in family		**116 (1.0)**
Alcoholism in family	V61.41	37 (0.3)
Substance abuse in family	V61.42	80 (0.7)
Mental health services for victim and perpetrator of abuse or history of abuse		**1265 (10.5)**
Counselling for marital and partner problems, unspecified	V61.10	1130 (9.4)
Counselling for victim of spousal and partner abuse	V61.11	49 (0.4)
Counselling for victim of child abuse	V61.21	**
Counselling for perpetrator of parental child abuse	V61.22	25 (0.2)
Counselling for perpetrators of physical/sexual abuse	V62.83	**
History of physical abuse	V15.41	14 (0.1)
History of emotional Abuse	V15.42	37 (0.3)
History of other psychological trauma	V15.49	**
Other problems related to primary support group, including family circumstances		**8160 (67.6)**
Person living alone	V60.3	88 (0.7)
Family disruption due to family member on military deployment	V61.01	10 (0.1)
Family disruption due to return of family member from military deployment	V61.02	**
Family disruption due to divorce or legal separation	V61.03	**
Family disruption due to parent-child estrangement	V61.04	75 (0.6)
Low income	Z59.6	292 (2.4)
Family disruption due to child welfare custody	V61.05	16 (0.1)
Family disruption due to child in foster care or in care of nonparental family member	V61.06	17 (0.1)
Counselling for parent-biological child problem	V61.23	547 (4.5)
Counselling for parent-adopted child problem	V61.24	30 (0.3)
Counselling for parent (guardian)- foster child problem	V61.25	15 (0.1)

Foster care (status)	V60.81	19 (0.2)
Other parent-child problems	V61.29	335 (2.8)
Problems with aged parents or in-laws	V61.3	15 (0.1)
Counselling for parent-child problem, unspecified	V61.20	6992 (57.9)
Problems related to care provider dependency		**169 (1.4)**
No other household member able to render care	V60.3	149 (1.2)
Holiday relief care	V60.4	13 (0.1)
Person living in residential institution	V60.5	**
Problems related to education and literacy		**692 (5.7)**
Educational circumstances	V62.3	692 (5.7)
Problems related to employment and unemployment or occupational maladjustment		**375 (3.1)**
Unemployment	V62.0	23 (0.2)
Adverse effects of work environment	V62.1	84 (0.7)
Other occupational circumstances or maladjustment	V62.29	196 (1.6)
Personal current military deployment status	V62.21	**
Personal history of return from military deployment	V62.22	72 (0.6)
Problems related to housing and economic circumstances		**205 (1.7)**
Lack of housing	V60.0	76 (0.6)
Inadequate housing	V60.1	49 (0.4)
Inadequate material resources	V60.2	15 (0.1)
Other specified housing or economic circumstances	V60.89	35 (0.3)
Unspecified housing or economic circumstance	V60.9	30 (0.2)
Problems related to legal circumstances		**11 (0.1)**
Legal circumstances	V62.5	11 (0.1)
Problems related to other psychosocial circumstances		**1344 (11.1)**
Interpersonal problems, not elsewhere classified	V62.81	86 (0.7)
Other psychological or physical stress	V62.89	599 (4.9)
Unspecified psychosocial circumstance	V62.9	591 (4.9)
Social maladjustment	V62.4	72 (0.6)

Note: ** represents suppressed numbers due to small cell size

Supplemental Table 5.5 Social determinants of health (SDOH) code in physician billing claims, by SDOH category (n=12,064)

Any SDOH code (44 codes)	**n (%)**
1	11,514 (95.4)
≥2	550 (4.6)
Alcoholism and drug addiction in family (2 codes)	
1	115 (0.9)
≥2	**
Mental health services for victim and perpetrator of abuse or history of abuse (8 codes)	
1	1261 (10.4)
≥2	**
Other problems related to primary support group, including family circumstances (15 codes)	
1	7870 (65.2)
≥2	289 (2.4)
Problems related to care provider dependency (3 codes)	
1	169 (1.4)
≥2	**
Problems related to education and literacy (1 code)	
1	692 (5.7)
Problems related to employment and unemployment or occupational maladjustment (6 codes)	
1	375 (3.1)
≥2	**
Problems related to housing and economic circumstances (5 codes)	
1	189 (1.6)
≥2	**

Problems related to legal circumstances (1 code)	
1	11 (0.1)
Problems related to other psychosocial circumstances (4 codes)	
1	1340 (11.1)
≥2	**

Note: maximum=4, ** represents suppressed numbers due to small cell size

CHAPTER 6: SUMMARY

6.1 Overview

This thesis assessed the validity of administrative health data (AHD) to identify modifiable risk factors, particularly smoking and social determinants of health (SDOH), through a series of four interrelated studies. These data are not intended for research use, and have only been explored in a limited way as a potential source of information about behavioral or social exposures. The validity of using these data for research and surveillance requires further exploration. The research contained in this thesis is grounded in Phase (b) of the Behavioral Epidemiology Framework (BEF), which emphasizes the development of valid and reliable behavioral measures (1). Within this framework, each study contributes to building, evaluating, or contextualizing algorithms to measure health behaviors using real-world, routinely-collected electronic health data.

6.2 Summary and Interpretation of Research Findings

The table below shows how each study contributes to Phase (b) of the BEF. This table highlights the focus of the thesis research on improving the identification of modifiable risk factors in routinely-collected electronic health data.

Table 6.1 Summary of novel contributions of the thesis research

Objective	**Contribution Theme**	**Previously Known**	**Novel Findings**
Objective 1: Evidence synthesis of smoking	Algorithm Development: Identified gaps in the literature amenable to	The accuracy of smoking algorithms varies depending on the data source and	MBAs, especially those based on ML algorithms, have higher sensitivity than RBAs. However, their

algorithms (Study 1)	improvement	method used for their development.	accuracy is often low.
Objective 2: Development and validation of smoking algorithms (Study 2)	Algorithm Development and Validation: Developed and validated MBAs and RBAs	RBAs are often used to identify smoking in AHD but have low sensitivity and are not often compared directly with MBAs in the same validation cohort.	MBAs have higher sensitivity than RBAs if they are based on multiple AHD sources and sociodemographic information. RBAs retain higher specificity and PPV than MBAs, if they are based on a limited set of diagnosis codes.
Objective 2: Development and validation of smoking algorithms (Study 3)	Algorithm Validation: Introduced the use of synthetic data for MBA training to address covariate imbalance in algorithm validation studies	Data imbalance in AHD validation cohorts, especially across sex and age, may limit the generalizability of diagnostic algorithm validation results.	Training data augmentation for MBAs with synthetic observations, particularly SMOTE improves specificity, and CTGAN improves sensitivity of model-based smoking algorithms for AHD by mitigating representation bias due to covariate imbalance in a validation

			cohort.
Objective 3: SDOH documentation in AHD (Study 4)	Algorithm Development: inform future algorithm development	SDOH are often under-documented in AHD.	Variation in SDOH documentation was observed across income quintile (i.e., income gradient) and varied with documentation policy changes. The sensitivity of SDOH documentation to these factors provides evidence about the face validity of using AHD to identify SDOH.

Note: administrative health data (AHD); model-based algorithms (MBAs); machine learning (ML); rule-based algorithms (RBAs); synthetic minority oversampling technique (SMOTE); conditional tabular adversarial networks (CTGAN); social determinants of health (SDOH)

To address objective 1, a systematic review and meta-analysis evaluated the validity of algorithms for ascertaining smoking status using electronic health data. Of the 116 algorithms identified from 57 studies, the majority (71.6%) were derived from electronic medical records (EMR), while 28.4% were derived from administrative data. Rule-based approaches accounted for 52.6% of algorithms, slightly exceeding model-based methods (47.4%), which predominantly employed machine learning (ML) techniques such as natural language processing and support vector machines. Pooled

performance estimates for all algorithms were: positive predictive value (PPV) 84.3%, sensitivity 67.2%, and specificity 91.8%. Model-based algorithms (MBAs) had significantly greater sensitivity than rule-based counterparts, and those based on EMR data outperformed administrative data on sensitivity. Heterogeneity in performance estimates was observed; sensitivity estimates ranged from 3% to 100% and specificity estimates ranged from 36% to 100%. Algorithms based on unstructured EMR data, such as clinical notes, showed higher sensitivity (median = 88.0%) compared to algorithms based on structured data like diagnosis codes (median = 62.5%). Specificity was lower in studies on clinical populations and higher in studies utilizing US data. Risk of bias assessment indicated that 76.3% of studies were at low risk of bias, although several exhibited high or unclear bias, particularly in the patient selection and reference standard domains of the quality assessment tool (2). The review identified several gaps in the development of smoking status algorithms, including limited representation of multi-jurisdictional and non-US data (14% of studies) and limited application of advanced ML models such as ensemble learning and large language models to create sensitive and specific algorithms. The overall low sensitivity of many algorithms raised concerns about the potential under-identification of smokers, which may bias epidemiological analyses and compromise the accuracy of disease risk prediction models.

Two related studies were conducted to address objective 2. Both studies used a retrospective clinical registry-based cohort of 24,718 individuals containing self-reported smoking status linked to AHD sources, including hospital discharge abstracts, physician billing claims, and prescription records. The first study developed and validated MBAs to identify smoking status from AHD and compared their performance against rule-based algorithms (RBAs). Two ML models were used to construct MBAs: random forest (RF) (3) and least absolute shrinkage and selection operator (LASSO) logistic regression (4). Two previously validated RBAs were used: RBA2 (5) builds on RBA1 (6) by adding bupropion prescription dispensations as an algorithm feature. This medication is used for smoking cessation and

depression treatment. The algorithms were developed using smoking-related diagnoses, medication use, and sociodemographic characteristics. The MBAs outperformed RBAs in terms of sensitivity (RF: 66.8% [95% CI: 63.3–70.2] vs. RBA1: 23.3% [20.3–26.5]) and balanced accuracy (LASSO: 71.1% [70.1–72.1] vs. RBA1: 61.1% [59.9–62.2]). However, RBAs demonstrated superior specificity (RBA2: 96.6% [96.1–96.6] vs. RF: 77.8% [76.8–78.8]) and PPV (RBA2: 47.2% [42.9–51.5] vs. RF: 25.1% [23.8–26.4]). All algorithms achieved negative predictive values (NPVs) above 90%. The study results highlighted the potential of MBAs to improve smoking ascertainment in AHD by leveraging different data sources and ML techniques, addressing the limitations of RBAs, which often exhibit low sensitivity. However, the trade-off between sensitivity and specificity underscores the need for context-specific algorithm selection. The second study corresponding to objective 2 evaluated the effectiveness of synthetic data generation models, synthetic minority oversampling technique (SMOTE), conditional tabular generative adversarial networks (CTGAN), and classification and regression trees (CART) to mitigate the bias resulting from covariate imbalances by augmenting training data for an MBA to identify smoking status in AHD (7). The cohort was predominantly female (88.6%) and older (mean age: 67.0 years), leading to underrepresentation of males and younger adults. The algorithm developed using LASSO logistic regression was trained on both augmented and non-augmented data and tested on a real-world imbalanced dataset. The results indicate that the use of SMOTE for training data augmentation improved specificity and fairness, while the use of CTGAN for training data augmentation improved sensitivity but resulted in lower specificity and PPV. CART had similar sensitivity and NPV but lower fairness than SMOTE. The results underscored the importance of selecting synthetic data generation models aligned with study priorities, such as SMOTE for equitable subgroup performance or CTGAN for maximal sensitivity.

To achieve objective 3, the face validity of SDOH documentation in AHD was assessed by analyzing their distribution across data sources, time, and geographic regions of Manitoba. Among more

than 1.6 million individuals in the cohort, 1.1% had SDOH documentation in at least one healthcare record (i.e., hospital discharge abstracts or physician claims records), suggesting under-documentation of social factors in routinely-collected healthcare data. Variation in SDOH documentation was associated with patient socio-demographic characteristics and healthcare utilization, and the data source used to capture this documentation. In hospital discharge abstracts, females (odds ratio (OR) 0.5, 95% CI 0.5–0.6 vs. males) and older patients (OR 0.4, 95% CI 0.3–0.4 vs. young patients) had lower odds of SDOH documentation. Winnipeg residents (OR 1.8, 95% CI 1.6–2.0 vs. Northern region) and those in lower-income quintiles had higher odds (Q1 OR 5.5, 95% CI 4.9–6.2 vs. Q5) of SDOH documentation. In physician claims records, there was a higher odds of SDOH documentation among females (OR 1.6, 95% CI 1.5–1.7 vs. males) and older patients (OR 1.9, 95% CI 1.8–2.0 vs. young patients), with modest geographic variation and slightly lower odds among lower-income groups (Q1 OR 0.8, 95% CI 0.7–0.8 vs. Q5). In the cohort of patients who had SDOH documentation in either hospital discharge abstracts or physician claims records, healthcare utilization was strongly associated with SDOH documentation; it was higher amongst patients with higher physician visit frequency (OR 3.1, 95% CI 3.0–3.3). The income gradient was similar to that for hospital discharge abstracts (Q1 OR 1.4, 95% CI 1.3–1.5). Patients from all regions, except Winnipeg, had a 40% to 60% lower odds of SDOH documentation than patients from the Northern region. A sharp increase in economic SDOH documentation in hospital data was evident post-2017, coinciding with revised coding guidelines. The use of diagnosis codes related to psychosocial circumstances in physician billing claims peaked during the COVID-19 pandemic, likely reflecting increased attention to mental health and healthcare access.

6.3 Thematic Integration and Research Implications

The use of AHD to identify modifiable risk factors for research and surveillance has some challenges, but also benefits (8,9). Across four studies, common themes emerged around the

development and validation of algorithms for identifying modifiable risk factors using AHD. Incorporating model-based approaches and synthetic data into developing and validating risk-factor ascertainment algorithms in this research appears to improve the valid identification of modifiable risk factors in AHD.

This thesis research contributes to algorithm development by demonstrating how methodological choices influence the identification of modifiable risk factors in AHD. The results of the evidence synthesis in Chapter 2 (Study 1) show that considerable variation in the estimates of algorithm validity metrics can partially be attributed to the methods (model-based vs. rule-based) for algorithm development. Results from Chapter 3 (Study 2) showed tradeoffs between validity metrics. For example, algorithms constructed using model-based methods had high sensitivity estimates, while RBAs resulted in high specificity estimates. These results support the application of MBAs in studies that require more comprehensive case identification for confounding adjustment, while RBAs are appropriate for clinical use where avoiding false positives is often a priority. The results of the descriptive analysis in Chapter 5 (Study 4) demonstrated that although SDOH documentation remains sparse in AHD, observed gradients by income and geography support the face validity of using AHD for identifying SDOH. This study about SDOH documentation provides essential groundwork for future algorithm development strategies by identifying which codes are captured, by whom, and for whom.

Chapter 4 (Study 3) introduced the use of synthetic data-based augmentation as a methodological strategy to enhance algorithm validation. Synthetic data generation models can mitigate demographic imbalances in training data, such as the underrepresentation of males and younger individuals. The results showed that training data augmentation for MBAs with synthetic observations improved algorithm performance, but the improvements depended on the model used for synthetic data generation. To illustrate, SMOTE enhanced specificity and PPV for underrepresented groups, while CTGAN improved sensitivity. These findings highlight the importance of data-centric validation approaches and

provide a replicable framework for future algorithm validation in demographically skewed validation cohorts.

6.4 Strengths and Limitations

The four studies in this thesis make several methodological and substantive contributions to data science and population health research. A major strength is the use of population-based data, including a clinical registry and linked administrative health databases. The first study, a systematic review and meta-analysis (10), utilized multi-level meta-regression models to synthesize evidence on the validity of smoking ascertainment algorithms, offering a quantitative summary of algorithm performance across diverse data sources and methods. The second study validated MBAs using self-reported smoking status from a clinical registry, leveraging more than one administrative data source, and exploring the influence of sociodemographic factors on current smoking status. The third study used data augmentation (11) to address cohort imbalance in diagnostic algorithm validation and compared the performance of different synthetic data generation models. The fourth study examined the face validity of SDOH documentation in population-based administrative data across an entire healthcare system (12).

The thesis research has some limitations. First, the generalizability of the results is constrained in multiple ways. Studies were conducted using data from specific cohorts, such as a single registry-based population of predominantly elderly females or data restricted to one Canadian province, which may not reflect the broader population or other healthcare settings (13,14). Moreover, language restrictions (English-only publications) and the exclusion of grey literature in the evidence synthesis may have introduced selection bias in the systematic review process (15). Furthermore, substantial heterogeneity remained unexplained in the meta-regression models for the systematic review, indicating that key sources of variability in algorithm performance remain unidentified, which may further affect the interpretability of pooled estimates. Methodological limitations also impact interpretability. All ML

algorithms developed in these studies were validated internally, instead of being validated in external populations. This limits the robustness and transportability of the estimated algorithm performance to other datasets or populations (16). The third study used synthetic data to validate algorithms. This study was limited to current smoking status, which affects the generalizability of the findings to other modifiable risk factors, such as obesity or alcohol use. Additionally, the descriptive nature of the analysis in the fourth study limits the ability to establish causal relationships between SDOH documentation and patient sociodemographic and healthcare use characteristics.

6.5 Conclusions and Future Research Directions

This thesis contributes to the growing body of research on the validation of algorithms for AHD (17,18), with a particular emphasis on smoking status and SDOH. Across four complementary studies, this research examined algorithm performance and limitations in current approaches to algorithm validation.

The findings highlight that while numerous algorithms exist to identify smoking status in AHD, their sensitivity is often low. However, specificity and PPV were consistently high across multiple algorithms. The finding emphasize the utility of RBAs for conservative identification of smoking status. Integrating different AHD sources and adopting ML-based MBAs can improve sensitivity and overall validity. The findings also demonstrated the potential of training data augmentation for MBAs with synthetic data, especially when generated through models like SMOTE, for mitigating biases due to covariate class imbalance, introducing underrepresentation in validation cohorts. However, the impact of data augmentation on the validity metrics is sensitive to the choice of synthetic data generation model. The exploration of SDOH documentation in AHD identified systematic socioeconomic and temporal variations of these factors. However, the generalizability of these results is constrained by a finding consistent with a repeatedly reported phenomenon of “low prevalence of SDOH documentation”

(19,20). This finding underscores the need for documentation interventions, such as providing financial incentives for adding SDOH diagnosis codes to billing claims and adding information about SDOH in hospital discharge reports, to improve documentation in AHD.

There are a number of opportunities for further research:

I. Data Integration and Model Innovation: Future research should pursue two parallel tracks to advance the measurement of modifiable risk factors. To define the reference standard for behavioral risk factor validation, future work could use EMR as a supplementary data source. Ensemble machine learning models (21) and large language models (22) could be applied to extract information to define a reference standard from structured and unstructured EMR data, respectively. A key challenge is that EMR data are often not population-based, which can limit the generalizability of algorithms validated with these data. This could be mitigated by combining EMRs with external population-level datasets, such as national health surveys. Second, future resaerch should focus on capturing the longitudinal nature of behavior. Techniques such as mixed-effects models (23) and longitudinal discriminant analysis (24) could enhance the identification of behavioral patterns over time.

II. External Validation and Equity: Another opportunity for further research is to assess the generalizability of these algorithms through external validation. This involves testing models developed on one dataset (e.g., from a specific province or health network) on independent datasets from different jurisdictions (25). Such studies are essential to identify and mitigate biases, ensuring the algorithms perform equitably in different populations.

III. Augmentation and Missingness: Future research should evaluate synthetic data as a tool to address limitations in imbalanced validation cohorts. Specifically, experimental studies are needed to evaluate how different synthetic data generation models perform under various

patterns of covariate missingness (e.g., missing completely at random, missing not at random) and to identify optimal strategies for selecting covariates when augmenting training and test datasets for MBAs with synthetic observations (26).

IV. SDOH Documentation Validity Assessment and Policy: An important direction for future work is to assess the criterion validity of SDOH algorithms for AHD, by linking these data with social surveys or registries that accurately capture SDOH information (27). The implementation challenges of such linkage include potential misalignments in variable definitions and collection periods between sources, as well as privacy-preserving linkage methods that maintain data utility while complying with regulations (28). Even with successful linkage of databases, biases may exist from survey non-response or registry coverage gaps. Incorporating mandatory fields and conducting qualitative research with healthcare providers can support the development of a more efficient, improved framework for capturing SDOH in AHD (29).

Collectively, the studies in this thesis demonstrate the potential and complexity of validating modifiable risk factor algorithms for real-world AHD. An approach that explicitly incorporates domain-specific knowledge, socio-environmental context, and methodological transparency, balancing rule-based methods with data-driven models and guided by principles of equity, is essential for developing reliable and generalizable algorithms for public health surveillance, policy planning, and disease risk assessment.

6.6 References

1. Sallis JF, Owen N, Fotheringham MJ. Behavioral epidemiology: A systematic framework to classify phases of research on health promotion and disease prevention. Annals of Behavioral Medicine. 2000;22(4):294-8.
2. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine. 2011 Oct 18;155(8):529-36.
3. Hastie T, Tibshirani R, Friedman J. Random Forests. In: Hastie T, Tibshirani R, Friedman J, editors. The Elements of Statistical Learning: Data Mining, Inference, and Prediction [Internet]. New York, NY: Springer; 2009. p. 587-604. (Springer Series in Statistics). Available from: https://doi.org/10.1007/978-0-387-84858-7_15
4. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58(1):267-88.
5. Marrie RA, Tan Q, Ekuma O, Marriott JJ. Development of an indicator of smoking status for people with multiple sclerosis in administrative data. Multiple Sclerosis Journal - Experimental, Translational and Clinical. 2022;8(1):20552173221074296.
6. Roper L, Tran DT, Einarsdóttir K, Preen DB, Havard A. Algorithm for resolving discrepancies between claims for smoking cessation pharmacotherapies during pregnancy and smoking status in delivery records: The impact on estimates of utilisation. PLOS ONE. 2018 Aug 30;13(8):e0202999.
7. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: A systematic review. Neurocomputing. 2022 Jul 7;493:28-45.
8. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the

American Medical Informatics Association. 2017 Jan 1;24(1):198-208.

9. Sidebottom AC, Johnson PJ, VanWormer JJ, Sillah A, Winden TJ, Boucher JL. Exploring Electronic Health Records as a Population Health Surveillance Tool of Cardiovascular Disease Risk Factors. Population Health Management. 2015 Apr;18(2):79-85.
10. PRISMA Statement Organization. PRISMA Endorsers [Internet]. Available from: http://www.prisma-statement.org/Endorsement/PRISMAEndorsers?AspxAutoDetectCookieSupport=1
11. Mumuni A, Mumuni F. Data augmentation: A comprehensive survey of modern approaches. Array. 2022 Dec 1;16:100258.
12. Prophetis ED, Beck K, Ridgeway D, Chuang J, Richard L, Durbin A, et al. Trends in hospital coding for people experiencing homelessness in Canada, 2015–2020: a descriptive study. Canadian Medical Association Open Access Journal. 2023 Nov 1;11(6):E1188-96.
13. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns. 2024 Apr 12;5(4):100946.
14. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 2012 Jul;42(4):463-84.
15. Stern C, Kleijnen J. Language bias in systematic reviews: you only get out what you put in. JBI Evidence Synthesis. 2020 Sep;18(9):1818.
16. Liu M, Shah A, Jiang M, Peterson NB, Dai Q, Aldrich MC, et al. A study of transportability of an existing smoking status detection module across institutions. AMIA Annual Symposium Proceedings. 2012;2012:577-86.
17. Yousif A, Dault R, Courteau M, Blais L, Cloutier AM, Lacasse A, et al. The validity of diagnostic algorithms to identify asthma patients in healthcare administrative databases: a systematic literature

review. Journal of Asthma. 2022 Jan 2;59(1):152-68.

18. Johnston A, Dancey SR, Tseung V, Skidmore B, Fell DB, Tanuseputro P, et al. A systematic review of validated case-finding definitions to identify hypertensive disorders of pregnancy in administrative healthcare databases. European Heart Journal. 2022 Oct 1;43(Supplement_2):ehac544.2504.
19. Torres JM, Lawlor J, Colvin JD, Sills MR, Bettenhausen JL, Davidson A, et al. ICD Social Codes: An Underutilized Resource for Tracking Social Needs. Medical Care. 2017 Sep;55(9):810-6.
20. Agarwal AR, Prichett L, Jain A, Srikumaran U. Assessment of Use of ICD-9 and ICD-10 Codes for Social Determinants of Health in the US, 2011-2021. JAMA Network Open. 2023 May 9;6(5):e2312538.
21. Polikar R. Ensemble Learning. In: Zhang C, Ma Y, editors. Ensemble Machine Learning: Methods and Applications [Internet]. Boston, MA: Springer US; 2012. p. 1-34. Available from: https://doi.org/10.1007/978-1-4419-9326-7_1
22. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology. 2024 Mar 29;15(3):39:1-39:45.
23. Wernecke KD, Kalb G, Schink T, Wegner B. A Mixed Model Approach to Discriminant Analysis with Longitudinal Data. Biometrical Journal. 2004;46(2):246-54.
24. Hughes DM, El Saeiti R, García-Fiñana M. A comparison of group prediction approaches in longitudinal discriminant analysis. Biometrical Journal. 2018 Mar;60(2):307-22.
25. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiology: Artificial Intelligence. 2022 May;4(3):e210064.
26. Hoogland J, van Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. Handling missing predictor values when validating and applying a prediction model to new

patients. Statistics in Medicine. 2020 Nov 10;39(25):3591-607.

27. Wray CM, Vali M, Abraham A, Zhang A, Walter LC, Keyhani S. Validation of Administrative Measures of Social and Behavioral Risk in Veterans Affairs Medical Records. Journal of General Internal Medicine. 2019 Jun 1;34(6):796-8.
28. Vatsalan D, Sehili Z, Christen P, Rahm E. Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges. In: Zomaya AY, Sakr S, editors. Handbook of Big Data Technologies [Internet]. Cham: Springer International Publishing; 2017. p. 851-95. Available from: https://doi.org/10.1007/978-3-319-49340-4_25
29. Sokol RL, Ammer J, Stein SF, Trout P, Mohammed L, Miller AL. Provider Perspectives on Screening for Social Determinants of Health in Pediatric Settings: A Qualitative Study. Journal of Pediatric Health Care. 2021 Nov 1;35(6):577-86.

Appendix A: Contributions of Authors

For the first paper of this thesis, LML and AH conceived the study. All authors prepared the analysis plan. AH and MLBG performed the literature search, screening for study inclusion/exclusion, data extraction, and risk of bias assessment. AH and LML conducted the analysis and prepared the draft manuscript. All authors reviewed and approved the final version of the manuscript.

For the second paper of this thesis, LML conceived the study. AH and LML conducted the analysis and prepared the draft manuscript. All authors reviewed and approved the final version of the manuscript.

For the third paper of this thesis, LML and AH conceived the study. AH and LML conducted the analysis and prepared the draft manuscript. All authors reviewed and approved the final version of the manuscript.

For the fourth manuscript of this thesis, LML and AH conceived the study. AH and LML conducted the analysis and prepared the draft manuscript. All authors reviewed and approved the draft version of the manuscript.

Appendix B: Ethics Approval and Data Access

The authors acknowledge the Manitoba Centre for Health Policy (MCHP) for using data contained in the Manitoba Population Health Research Data Repository (the Provincial Health Research Privacy Committee # P2023-29). The presented study findings and interpretations reflect the authors' perspectives. No formal endorsement by the Manitoba Centre for Health Policy, Manitoba Health, or any other data providers is implied or should be inferred.

The data was provided to the investigators under specific data-sharing agreements only for approved use at the MCHP. The researchers or MCHP does not own the original source data and, as such, cannot be provided to a public repository. The original source data are available through the submission of appropriate ethics and data access approval forms to the Health Research Ethics Board of the University of Manitoba (see https://umanitoba.ca/research/opportunities-support/ethics-compliance for more details) and the Provincial Health Research Privacy Committee (see https://www.rithim.ca/phrpc-overview for more details).