EJoHP 1(4)/2025

DOI: https://doi.org/10.21697/ejhp.0106.02
e-ISSN 2957-2002

2025, 4, 1: 17-28

Machine Learning Models for Patient Screening Using Routinely Collected Data in Primary Care

Agnieszka Mazurek, MD
School of Public Health
Centre of Postgraduate Medical Education
ORCID: 0009-0002-1176-4076
ul. Kleczewska 61/63, 01-826 Warszawa, Polska
Email: agnieszka.mazurek@cmkp.edu.pl
Phone: +48 22-5601-150

Refceived: 3 Jun 2025; Revised: 18 Jun 2025; Accepted: 27 Jun 2025

Abstract

Screening is crucial to preventing the health consequences associated with undiagnosed diseases. Electronic health records (EHRs) from primary care can be leveraged with machine learning (ML) techniques to create new tools for patient screening in general practice. The aim of this narrative review is to discuss the recent literature on the development and validation of predictive ML models designed for the early detection of health conditions using readily available patient data. The PubMed, Web of Science, Scopus, and IEEE Xplore databases were searched for studies published within the last five years. Twenty-one studies were found, covering a variety of health conditions. ML-based tools can function as independent screening tests or can enhance existing screening methods. Moreover, ML models can be employed to screen for conditions for which screening approaches have not yet been developed. However, primary care EHRs alone are not always a sufficient source of data for effective screening. Poor data quality can result in erroneous or biased predictions. Despite these limitations, the application of ML for screening has shown promising results, and further research in this area is warranted.

Keywords: machine learning, primary care, screening, electronic health records

INTRODUCTION

Screening is a fundamental public health intervention for secondary prevention of disease. The goal of screening is to identify a condition before symptoms develop so as to accelerate the initiation of treatment or supportive care to either cure the condition or slow its progression (WHO Regional Office for Europe, 2020). Screening can be universal (population-level and high-risk-independent), targeted (population-level and high-risk-dependent), or opportunistic (patient-level and high-risk-independent) (Eriksen et al., 2021).

Primary care provides an excellent setting for administering screening, as it is a regular point of contact with the healthcare system for the general population, irrespective of health status. It is estimated that the primary care enrollment rate in developed countries exceeds 70% (Lin et al., 2025), with some countries, such as the UK, having nearly 98% of the population registered in primary care (Nadarajah et al., 2023). Primary care encounters generate vast volumes of data that often reflect longitudinal medical history. This data, stored in electronic health records (EHRs), could be utilized to uncover subclinical indicators of numerous health conditions, which could then be used to develop new digital screening tools.

Currently, machine learning (ML) is one of the leading approaches to analyzing large amounts of data. Hence, the aim of this narrative review is to discuss recently published research in this field and identify emerging trends in the use of ML to develop screening tools for use in primary care.

METHODS

Search Strategy

This work is a narrative review of the recent literature describing patient screening in primary care using machine learning algorithms. PubMed, Web of Science, Scopus, and IEEE Xplore were searched for relevant articles published between January 1, 2020 and May 28, 2025. The following keywords were used in the search: machine learning, screening, and primary care. The inclusion criteria were the use of data readily available in primary care offices and the objective of screening, or aiding screening, for a health condition. Only studies written in English were considered.

Information Extraction

Each article was carefully reviewed to extract key information, including the authors, year of publication, data source, screening target, the algorithms evaluated, the best-performing algorithm and its performance metrics, the total number and types of features, the validation method, and any non-ML comparators used.

DISCUSSION

Summary of Search Results

A total of 21 studies fulfilled the inclusion criteria. They covered a range of conditions. Mental and behavioral disorders were the most frequently investigated (n=6), followed by diseases of the circulatory system (n=5), diseases of the digestive system (n=3), and endocrine, nutritional, and metabolic diseases (n=3).

Nine distinct algorithm classes were evaluated across all studies. The most prevalent model was logistic regression (LR), which was used in 15 studies. Random forest (RF) was the second model of choice, appearing in 14 studies. Five studies tested only one algorithm. Of the studies that tested more than one, four identified eXtreme Gradient Boosting (XGBoost) and three identified neural networks (NN) as the best-performing models. Other top performers were RF, LR, and Light Gradient Boosting Machine (LightGBM).

All studies employed internal validation to assess model performance, but only six carried out additional external validation to evaluate the generalizability. Furthermore, in five cases, ML performance was compared against standard screening tools.

A complete summary of the most important characteristics and findings of each study is presented in Table 1. The performance metrics reported in the table and throughout this work are the highest values achieved by the best-performing model on any validation set unless otherwise specified.

Table 1: Summary of Studies on Machine Learning Models for Screening in Primary Care

Authors, Year, Country	Target	Models	Validation		AUC	Sensitivity/ Recall	Specificity	PPV/ Precision	NPV	Accuracy	F1
Kimura et al., 2025, Japan	PET Aβ-positivity	**EN, LR, SVM**	I	LR	0.76 (0.01)	0.64 (0.03)	0.75 (0.03)	0.62 (0.03)	0.77 (0.01)	0.70 (0.02)	–
Eder et al., 2025, Germany	Depression	**XGB**, XGB+LR, SVM	I		–	0.878	0.886	–	–	0.882[a]	–
Wei et al., 2024, China	Carotid Artery Plaques	**LightGBM,** LR, NB, MLP, RF, SVM, XGB	I		0.854	0.595	0.892	0.729	0.817	0.795	0.655
Lu et al., 2024, Canada	Prediabetes	**DNN**, KNN, LR, NB, RF, SVM, XGB	I		0.76	0.60	–	0.69	–	0.72	0.64
Dabbah et al., 2024, Israel	Advanced Liver Fibrosis	LR, NN, RF, SVM, **XGB**	E		0.91 [0.88–0.97]	0.91 [0.84–0.96]	0.76 [0.72–0.80]	0.31 [0.26–0.34]	0.99 [0.98–1.00]	–	–
Szlejf et al., 2023, Brazil	Cognitive Impairment	CatBoost, LightGBM, LR, NN, **XGB**	I		0.873 [0.839–0.906]	0.316	0.969	0.298	0.972	–	0.307
Qin et al., 2023, China	MASLD	DT, RF, **SVM**, XGB	I		0.850 [0.840–0.850]	–	–	0.795 [0.781–0.795]	–	0.801 [0.789–0.801]	0.795 [0.781–0.795]
Nadarajah et al., 2023, UK	AF	LR, **RF**	I		0.824 [0.814–0.834]	0.781 [0.731–0.829]	0.731 [0.693–0.771]	0.025 [0.023–0.027]	0.998 [0.998–0.998]	–	–
Onishchenko et al., 2022, USA	IPF	PFSA + LightGBM	E	Men	0.88 (0.07)	0.68 (0.01)	0.95	0.50 (0.01)	0.98 (0.00)	–	–
				Women	0.94 (0.06)	0.83 (0.02)		0.38 (0.01)	0.99 (0.00)	–	–
Liu et al., 2022, China	Diabetes	CDKNN, KNN, **LGBM**, LR, NN, RF, SVM	I		0.697	–	–	–	–	–	–
Lin et al., 2022, China	Primary Aldosteronism	LR	E		0.839 [0.790–0.890]	0.582	0.892	0.716	0.820	0.793	–
Lee and Pak, 2022, South Korea	SI; SpoA	LR, RF, SVM, **XGB**	I	SI	0.861	0.853	0.869	0.819	0.895	0.863	–
				SPoA	0.880	0.861	0.900	0.861	0.900	0.884	–
Sekelj et al., 2021, UK	AF	CoxR, LR, **NN**, RF, SVM	E		0.87	0.500	0.926	0.169	0.984	–	–
Bennis et al., 2022, Netherlands	HF	**LR**, RF, XGB	I		0.772 [0.759–0.785]	0.761	0.653	–	–	0.655	–
Yu et al., 2021, China	Carotid Atherosclerosis	DT, **MLP**, RF, SVM, XGB	I		0.766 [0.754–0.769]	–	–	0.743	–	0.748	0.742
Souza Filho et al., 2021, Brazil	Depression	AB, CART, GB, KNN, LR, **RF**, SVM, XGB	I		0.87 (0.08)	0.90 (0.03)	–	0.88 (0.04)	–	0.89 (0.03)	0.89 (0.03)
Malhotra et al., 2021, UK	Pancreatic Cancer	**LR, RF**	I	15–60 y (20 mo)[b]	0.656	0.725	0.587	–	–	–	–
				61–99 y (17 mo)[c]	0.609	0.651	0.568	–	–	–	–
Amit et al., 2021, UK	Postpartum Depression	XGB	E		0.844 [0.830–0.857]	0.764 [0.735–0.791]	0.80	–	–	–	–
van Mens et al., 2020, Netherlands	Suicidality	RF	I		0.82 [0.78–0.86]	0.39 [0.32–0.47]	0.98 [0.97–0.98]	0.05 [0.04–0.06]	–	0.68[a]	–

Authors, Year, Country	Target	Models	Validation		AUC	Sensitivity/ Recall	Specificity	PPV/ Precision	NPV	Accuracy	F1
Rosenfeld et al., 2020, UK	Barret's Esophagus	DT, **LR**, NB, RF, SVM	E		0·81 [0·74–0·84]	0.90	0.58	0.77	0.77	0.769	0.77
Doyle et al., 2020, UK	NTM Lung Disease	XGB	I		0.94	0.135	–	0.091	–	–	–

Notes: Values in parentheses ( ) are standard deviations. Values in square brackets [ ] are 95% confidence intervals. If more than one model was tested, the best-performing model is shown in bold. If all are bolded, the performance was similar for all models and specific metrics are reported for only one model.

[a] Balanced accuracy; [b] Patients aged 15 to 60 years at 20 months before diagnosis; [c] Patients aged 61 to 99 years at 17 months before diagnosis.

Abbreviations: *AUC* – area under receiver operating characteristic curve; *PPV* – positive predictive value; *NPV* – negative predictive value; *PET* – positron emission tomography; *AB* – amyloid beta; *MASLD* – metabolic dysfunction-associated steatotic liver disease; AF – atrial fibrillation; *IPF* – idiopathic pulmonary fibrosis; *SI* – suicidal ideation; *SpoA* – suicide planning and attempt; *HF* – heart failure; *NTM Lung Disease* – nontuberculous mycobacterial lung disease; *EN* – elastic net; *LR* – logistic regression; *SVM* – support vector machine; *XGB* – eXtreme Gradient Boosting; *LightGBM* – Light Gradient Boosting Machine; *NB* – naïve Bayes; *RF* – random forest; *MLP* – multilayer perceptron; *DNN* – deep neural network; *KNN* – K-nearest neighbors; *NN* – neural networks; *DT* – decision tree; *PFSA* – probabilistic finite automata; *CDKNN* – centroid-displacement-based KNN; *CoxR* – Cox regression; *AB* – adaptive boosting; *CART* – classification and regression tree; *GB* – gradient boosting; *I* – internal; *E* – external.

Defining Machine Learning

Machine learning (ML) is a subfield of artificial intelligence (AI). Notably, the proliferation of natural language processing (NLP) and large language models (LLMs) has popularized the colloquial use of the term *AI* to refer to conversational chatbots. However, NLP and ML are distinct in their underlying computational methods and practical applications. ML refers to applying predictive algorithms to data to "learn" from existing patterns to solve classification or regression problems (Google Cloud, n.d.).

Standard Performance Metrics

Assessing the real-world utility of predictive ML models relies on being able to interpret their performance in a clinical context. The performance of classification models can be evaluated with familiar metrics, such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy.

In the field of ML, the four most commonly reported metrics are area under the receiver operating characteristic curve (AUC or AUROC), precision, recall, and F1 score (Google Developers, n.d.). AUC quantifies how well the model discriminates between positive and negative cases; it ranges from 0.5, indicating no discrimination, to 1, indicating perfect discrimination. The other three metrics assume values between 0 (worst score) and 1 (best score). Recall is equivalent to sensitivity and indicates the proportion of positive cases identified. Precision is equivalent to the PPV and indicates the proportion

of correct positive predictions. The F1 score combines precision and recall, indicating the proportion of false positives and false negatives.

Unfortunately, there are no universal cut-offs for these metrics that could be used to definitively deem a model good or bad for screening, as performance expectations are highly use-case-dependent. Higher values are favored. For AUC, values above 0.6, 0.7, 0.8, and 0.9 correspond to acceptable, good, very good, and excellent performance, respectively (Hanna et al., 2023; White et al., 2023).

Model Validation

Generalizability is the ability of an ML model to make useful predictions on unseen data. Validation is crucial in this assessment and can be internal or external (Steyerberg et al., 2001). Internal validation involves randomly splitting the dataset into training and testing subsets. A simple test-train split, cross-validation, and bootstrapping are commonly used internal validation methods (Steyerberg and Harrell, 2016). Internal validation is straightforward and should be considered the bare minimum for evaluating model performance. External validation uses a separate, fully independent dataset to test the model. Strategies for external validation include geographical validation, in which the validation data is obtained from different locations (i.e., clinical sites or countries), and temporal validation, which uses data from different time periods (Steyerberg and Harrell, 2016).

All studies performed internal validation and six conducted external validation. Most models showed sustained performance. The model for advanced fibrosis in metabolic dysfunction-associated steatosis liver disease (MASLD) (Dabbah et al., 2024), which was trained using data from tertiary care, demonstrated comparable performance when tested on data from primary care. This finding highlights that models do not necessarily need to be trained on primary care data to be useful in those settings.

Novel Screening Tools

Perhaps the most promising application of ML for the development of screening tools is to target diseases for which no known screening methods exist. One such disease is idiopathic pulmonary fibrosis (IPF), which is characterized by an insidious onset and poor prognosis. A model, the zero-burden comorbidity risk score for IPF (ZCoR-IPF) (Onishchenko et al., 2022), was developed to predict the risk of IPF at 1 and 4 years prior to a formal clinical diagnosis. Comorbidity codes were used as the sole input. At 1 year, the model achieved an AUC of 0.88 (0.07), with moderate sensitivity at 95% specificity for men and 0.94 (0.06) for women with high sensitivity at 95% specificity. The NPV

was 0.98 (0.00) for men and 0.99 (0.00) for women. These results indicate that ZCoR-IPF would be a valuable tool for screening IPF in general practice.

Enhanced Screening Tools

Another application of ML for screening is improving existing screening methods. ML models could supplement proven screening strategies with information derived from primary care records. The model to screen for depression (Eder et al., 2025) – using a combination of 15 deliberately chosen items from the World Health Organization Well-Being Index (WHO-5), the Patient Health Questionnaire-9 (PHQ-9), and the World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) – achieved better results than the PHQ-9 alone. Another model to identify women at risk of postpartum depression (Amit et al., 2021) based on demographics, medical history, and labor complications performed slightly worse on its own than the Edinburgh Postnatal Depression Scale (EPDS). However, incorporating the EDPS score into the model improved performance over the individual components.

Fasting plasma glucose (FPG) is one of the recommended screening modalities for diabetes, but it can miss cases with incidentally normal results at the time of screening. To address this issue, a model for screening for diabetes (Liu et al., 2022) that incorporates demographic and anthropometric features along with an FPG measurement was developed. It performed better than the model without FPG, even if FPG was below the diagnostic threshold, showing an ability to improve standard screening strategies.

Alternative Screening Tools

In some cases, ML may be utilized to develop predictive models that could serve as an alternative to existing screening tools. This could be motivated by the need to deliver improved results, reduce reliance on specialized equipment or additional laboratory testing, and eliminate the need for invasive testing procedures.

In screening for advanced liver fibrosis in MASLD, the model of Dabbah et al. (2024) outperformed established tools such as the Fibrosis-4 Index (FIB-4) and the NAFLD Fibrosis Score (NFS), offering a markedly higher PPV while maintaining an NPV of 99% [98–100]. The Future Innovations in Novel Detection of Atrial Fibrillation model (FIND-AF) (Nadarajah et al., 2023) for identifying patients at risk of atrial fibrillation showed better performance than the C2HEST and CHA2DS2-VASc scores. The FIND-AF model also identified high-risk patients under 65 years old, whom these traditional approaches could otherwise overlook.

A model to screen for Alzheimer's disease (Kimura et al., 2025) by predicting the presence of intracerebral amyloid β plaques demonstrated moderate effectiveness at excluding individuals unlikely to show amyloid accumulation on positron emission tomography (PET), thereby limiting unnecessary scans. Two models for screening carotid atherosclerosis (Wei et al., 2024; Yu et al., 2021) showed that they could be used to assess atherosclerosis in asymptomatic adults using demographics, physical examination, and laboratory data. This is particularly important, given that carotid duplex sonography is neither economically feasible nor recommended for this population.

The prediabetes screening model of Lu et al. (2024) was built without incorporating any glycemia-related laboratory results, yet still managed good discriminatory performance. However, the recall was poor, suggesting limited potential to eliminate the need for additional blood tests.

Barrett's esophagus, a precursor to esophageal adenocarcinoma, is challenging to screen for due to its low incidence and the reliance on invasive endoscopy with biopsy. A model based solely on demographic and reflux-related symptoms (Rosenfeld et al., 2020) showed that it could be used as a low-burden alternative to identify at-risk patients who may require endoscopic evaluation.

Suboptimal Results

It should be acknowledged that primary care data may sometimes be insufficient to train ML models that could produce clinically actionable results. A model to identify patients at high risk of pancreatic cancer 17 to 20 months prior to diagnosis (Malhotra et al., 2021) was trained using data on demographics, comorbidities, symptoms, pharmacotherapy, and frequency of clinical encounters. Notably, no biomarkers or imaging results were included. The model achieved an AUC of 0.656 for patients aged 60 or younger at 20 months and 0.609 for patients older than 60 at 17 months before diagnosis. Specificity did not exceed 0.59 for either group. The authors proposed that the model would likely benefit from integrating biomarker assays to improve usability.

Model Limitations

One of the most frequently addressed limitations of the use of ML models in medicine is their explainability. This refers to the extent to which the features that most influenced the model's predictions agree with the medical knowledge explaining the pathophysiology of the target condition. Certain algorithms, such as RF, produce feature importance rankings that can be evaluated by healthcare professionals for clinical justifiability. For the algorithms that do not offer such solutions, the Shapley Additive exPlanations

framework (SHAP) (Lundberg & Lee, 2017) can be used to attempt to explain the model's reasoning. Nearly all studies included in this review provided some analysis of feature importance, which should mitigate the risk of their models being dismissed by clinicians due to a lack of trust.

Poor-quality primary care EHRs could also affect the reliability of ML-based screening tools. The quality of the data found in EHRs is not uniform due to reporting inconsistencies between healthcare providers (van Mens et al., 2020). On the one hand, models trained on datasets affected by data missingness or inadequate or erroneous reporting would likely not achieve satisfactory clinical results. On the other hand, incomplete patient records, such as when patients underreport sensitive information regarding addiction, education, and income level (Malhotra et al., 2021), could negatively impact the accuracy of predictions made by even those models that were trained on high-quality data.

Ethical Considerations

Models trained on narrowly defined subpopulations do not represent the general population, which may limit their generalizability and affect the health outcomes of minority groups. For instance, the model used to screen for cognitive impairment (Szlejf et al., 2023) was trained on the medical records of Brazilian government workers, who are more likely than the general population to have higher education. Consequently, primary education, as opposed to higher education, was found to be one of the most important predictors of cognitive impairment. This could mean that the model was biased and could lead to inaccurate predictions and stigmatization of patients from different social classes if implemented in clinical practice.

Future Research Directions

A successful deployment of ML-based screening tools in primary care practice must be preceded by a thorough consideration and elimination of various implementation barriers. Clinical trials should be conducted not only for medical validation, but also to test the most appropriate approaches to seamlessly incorporate ML-based screening tools into the clinical workflow. Particular attention should be paid to designing solutions for the complete automation of the ML-based screening process to avoid placing an additional burden on physicians and other healthcare workers. Furthermore, surveys should be administered to physicians regarding their expectations of model explainability so that these expectations can be accounted for at the earliest stages of model development to ensure acceptability (Ahluwalia et al., 2025). Additionally, cost-effectiveness should be analyzed to justify the use of ML for screening to all stakeholders (Liu et al., 2022).

CONCLUSION

In conclusion, this review demonstrated that the use of ML to develop screening tools is a budding area of medical research. Primary care EHRs could be leveraged to enable screening for conditions that were previously considered medically and economically unfeasible to detect at scale. In addition, ML-based screening tools could replace or supplement existing screening strategies to improve patient outcomes and optimize resource utilization in primary care. The data used for model training should be carefully selected to prevent the incorporation of social bias into the predictions and to ensure equitable care for all patients. Future research should address potential implementation challenges at all stages of model development.

REFERENCES

Ahluwalia, V. S., Schapira, M. M., Weissman, G. E., & Parikh, R. B. (2025). Primary Care Provider Preferences Regarding Artificial Intelligence in Point-of-Care Cancer Screening. *MDM Policy & Practice*, *10*(1). https://doi.org/10.1177/23814683251329007

Amit, G., Girshovitz, I., Marcus, K., Zhang, Y., Pathak, J., Bar, V., & Akiva, P. (2021). Estimation of Postpartum Depression Risk From Electronic Health Records Using Machine Learning. *BMC Pregnancy and Childbirth*, *21*(1), 630. https://doi.org/10.1186/s12884-021-04087-8

Dabbah, S., Mishani, I., Davidov, Y., & Ben Ari, Z. (2024). Implementation of Machine Learning Algorithms to Screen for Advanced Liver Fibrosis in Metabolic Dysfunction-Associated Steatotic Liver Disease: An In-Depth Explanatory Analysis. *Digestion*, 189–202. https://doi.org/10.1159/000542241

Eder, J., Dong, M. S., Wöhler, M., Simon, M. S., Glocker, C., Pfeiffer, L., Gaus, R., Wolf, J., Mestan, K., Krcmar, H., Koutsouleris, N., Schneider, A., Gensichen, J., Musil, R., & Falkai, P. (2025). A Multimodal Approach to Depression Diagnosis: Insights From Machine Learning Algorithm Development in Primary Care. *European Archives of Psychiatry and Clinical Neuroscience* [Preprint]. https://doi.org/10.1007/s00406-025-01990-5

Eriksen, C. U., Rotar, O., Toft, U., & Jørgensen, T. (2021). *What Is the Effectiveness of Systematic Population-Level Screening Programmes for Reducing the Burden of Cardiovascular Diseases?* WHO Regional Office for Europe (WHO Health Evidence Network Synthesis Reports). Retrieved June 2, 2025 from http://www.ncbi.nlm.nih.gov/books/NBK567843

Google Cloud. (n.d.). AI Vs. Machine Learning: How Do They Differ?. Retrieved June 2, 2025 from https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning

Google Developers. (n.d.). Classification: Accuracy, Recall, Precision, and Related Metrics. Retrieved June 2, 2025 from https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

Hanna, M. G., Olson, N. H., Zarella, M., Dash, R. C., Herrmann, M. D., Furtado, L. V., Stram, M. N., Raciti, P. M., Hassell, L., Mays, A., Pantanowitz, L., Sirintrapun, J. S., Krishnamurthy, S., Parwani, A., Lujan, G., Evans, A., Glassy, E. F., Bui, M. M., Singh, R., Souers, R. J., de Baca, M. E., & Seheult, J. N. (2023). Recommendations for Performance Evaluation of Machine Learning in Pathology: A Concept Paper From the College of American Pathologists. *Archives of Pathology & Laboratory Medicine*, *148*(10), e335–e361. https://doi.org/10.5858/arpa.2023-0042-CP

Kimura, N., Sasaki, K., Masuda, T., Ataka, T., Matsumoto, M., Kitamura, M., Nakamura, Y., & Matsubara, E. (2025). Machine Learning Models for Dementia Screening to Classify Brain Amyloid Positivity on Positron Emission Tomography Using Blood Markers and Demographic Characteristics: A Retrospective Observational Study. *Alzheimer's Research & Therapy*, *17*(1), 25. https://doi.org/10.1186/s13195-024-01650-1

Lin, J., Bates, S., Allen, L. N., Wright, M., Mao, L., Chomik, R., Dietz, C., & Kidd, M. (2025). Uptake of Patient Enrolment in Primary Care and Associated Factors: A Systematic Review and Meta-Analysis. *BMC Primary Care*, *26*(1) 76. https://doi.org/10.1186/s12875-025-02779-0

Liu, X., Zhang, W., Zhang, Q., Chen, L., Zeng, T., Zhang, J., Min, J., Tian, S., Zhang, H., Huang, H., Wang, P., Hu, X., & Chen, L. (2022). Development and Validation of a Machine Learning-Augmented Algorithm for Diabetes Screening in Community and Primary Care Settings: A Population--Based Study. *Frontiers in Endocrinology*, 13. https://doi.org/10.3389/fendo.2022.1043919

Lu, K., Sheth, P., Zhou, Z. L., Kazari, K., Guergachi, A., Keshavjee, K., Noaeen, M., & Shakeri, Z. (2024). Identifying Prediabetes in Canadian Populations Using Machine Learning. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4. https://doi.org/10.1109/EMBC53108.2024.10782174

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv. https://doi.org/10.48550/arXiv.1705.07874

Malhotra, A., Rachet, B., Bonaventure, A., Pereira, S. P., & Woods, L. M. (2021). Can We Screen for Pancreatic Cancer? Identifying a Sub-Population of Patients at High Risk of Subsequent Diagnosis Using Machine Learning Techniques Applied to Primary Care Data. *PLOS ONE*, *16*(6). https://doi.org/10.1371/journal.pone.0251876

van Mens, K., Elzinga, E., Nielen, M., Lokkerbol, J., Poortvliet, R., Donker, G., Heins, M., Korevaar, J., Dückers, M., Aussems, C., Helbich, M., Tiemens, B., Gilissen, R., Beekman, A., & de Beurs, D. (2020). Applying Machine Learning on Health Record Data From General Practitioners to Predict Suicidality. *Internet Interventions*, 21, 100337. https://doi.org/10.1016/j.invent.2020.100337

Nadarajah, R., Wu, J., Hogg, D., Raveendra, K., Nakao, Y. M., Nakao, K., Arbel, R., Haim, M., Zahger, D., Parry, J., Bates, C., Cowan, C., & Gale, C. P. (2023). Prediction of Short-Term Atrial Fibrillation Risk Using Primary Care Electronic Health Records. *Heart*, *109*(14), 1072–1079. https://doi.org/10.1136/heartjnl-2022-322076

Onishchenko, D., Marlowe, R. J., Ngufor, C. G., Faust, L. J., Limper, A. H., Hunninghake, G. M., Martinez, F. J., & Chattopadhyay, I. (2022). Screening for Idiopathic Pulmonary Fibrosis Using Comorbidity Signatures in Electronic Health Records. *Nature Medicine*, *28*(10), 2107–2116. https://doi.org/10.1038/s41591-022-02010-y

Rosenfeld, A., Graham, D. G., Jevons, S., Ariza, J., Hagan, D., Wilson, A., Lovat, S. J., Sami, S. S., Ahmad, O. F., Novelli, M., Justo, M. R., Winstanley, A., Heifetz, E. M., Ben-Zecharia, M., Noiman, U., Fitzgerald, R. C., Sasieni, P., Lovat, L. B., Coker, K., Zhao, W., Brown, K., Haynes, B., Grant, T. N., Pietro, M. di, Dewhurst, E., Alias, B., Mills, L., Wilson, C., Bird-Lieberman, E., Bornschein, J., Lim, Y., Shariff, K., Lopez, R. C., Udarbe, M., Shaw, C., Rose, G., Sargeant, I., Al-Izzi, M., Schimmel, R., Green, E., Moorghen, M., Kanani, R., Baulf, M., Butcher, J., Butt, A., Bown, S., Lipman, G., Sweis, R., Sehgal, V., Banks, M., Haidry, R., Louis-Auguste, J., Kohoutova, D., Kerr, S., Eneh, V., Butter, N., Miah, H., Butawan, R., Adesina, G., Holohan, S., Idris, J., Hayes, N., Wahed, S., Houghton, N. K., Hopton, M., Eastick, A., Majumdar, D., Manuf, K., Fieldson, L., Bailey, H., Ortiz, J. F.-S., Patel, M., Henry, S., Warburton, S., White, J., Gadeke, L., Longhurst, B., Abeseabe, R., Basford, P., Bhattacharyya, R., Elliot, S., Bevan, R., Brown, C., Laverick, P., Clifford, G., Gibbons, A., Ingmire, J., Mawas, A., Harvey, J., & Cave, S. (2020). Development and Validation of a Risk Prediction Model

to Diagnose Barrett's Oesophagus (MARK-BE): A Case-Control Machine Learning Approach. *The Lancet Digital Health*, *2*(1), e37–e48. https://doi.org/10.1016/S2589-7500(19)30216-X

Steyerberg, E. W. & Harrell, F. E. (2016). Prediction Models Need Appropriate Internal, Internal-External, and External Validation. *Journal of Clinical Epidemiology*, 69, 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J. J. M., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal Validation of Predictive Models: Efficiency of Some Procedures for Logistic Regression Analysis. *Journal of Clinical Epidemiology*, *54*(8), 774–781. https://doi.org/10.1016/S0895-4356(01)00341-9

Szlejf, C., Batista, A. F. M., Bertola, L., Lotufo, P. A., Benseãor, I. M., Chiavegatto Filho, A. D. P., & Suemoto, C. K. (2023). Data-Driven Decision Making for the Screening of Cognitive Impairment in Primary Care: A Machine Learning Approach Using Data From the ELSA-Brasil Study. *Brazilian Journal of Medical and Biological Research*, 56, e12475. https://doi.org/10.1590/1414-431X2023e12475

Wei, Y., Tao, J., Geng, Y., Ning, Y., Li, W., & Bi, B. (2024). Application of Machine Learning Algorithms in Predicting Carotid Artery Plaques Using Routine Health Assessments. *Frontiers in Cardiovascular Medicine*, 11. https://doi.org/10.3389/fcvm.2024.1454642

White, N., Parsons, R., Collins, G., & Barnett, A. (2023). Evidence of Questionable Research Practices in Clinical Prediction Models. *BMC Medicine*, *21*(1), 339. https://doi.org/10.1186/s12916-023-03048-6

WHO Regional Office for Europe. (2020). Screening Programmes: A Short Guide. Increase Effectiveness, Maximize Benefits and Minimize Harm. WHO Regional Office for Europe.

Yu, J., Zhou, Y., Yang, Q., Liu, X., Huang, L., Yu, P., & Chu, S. (2021). Machine Learning Models for Screening Carotid Atherosclerosis in Asymptomatic Adults. *Scientific Reports*, *11*(1), 22236. https://doi.org/10.1038/s41598-021-01456-3