Gestational Diabetes Diagnoses in Electronic Health Records: A Three-Step Study of Label Accuracy and Its Impact on Machine Learning Models for Early Prediction

Mark Germaine, Amy C O'Higgins, Brendan Egan, Graham Healy

Submitted to: JMIR Medical Informatics
on: February 21, 2025

Table of Contents

Original Manuscript **5**
Supplementary Files **25**
Figures **26**
Figure 1 ..... **27**
Figure 2 ..... **28**
Figure 3 ..... **29**
Figure 4 ..... **30**

Gestational Diabetes Diagnoses in Electronic Health Records: A Three-Step Study of Label Accuracy and Its Impact on Machine Learning Models for Early Prediction

Mark Germaine[1, 2] BSc, MSc; Amy C O'Higgins[3] PhD; Brendan Egan[2] PhD; Graham Healy[1] PhD

[1] School of Computing Dublin City University Dublin IE
[2] School of Health and Human Performance Dublin City University Dublin IE
[3] UCD Centre for Human Reproduction Coombe Women & Infants University Hospital Dublin IE

Corresponding Author:
Mark Germaine BSc, MSc

School of Computing
Dublin City University
Dublin City University, Collins Ave Ext, Whitehall, Dublin 9
Dublin
IE

Abstract

Background: Integration of electronic health records (EHRs) into clinical research offers numerous opportunities for advancing healthcare delivery and patient outcomes, particularly in the era of machine learning (ML). However, EHR data needs to be coded accurately to ensure that models are learning correct representations of diseases.

Objective: This study examines the accuracy of gestational diabetes mellitus (GDM) diagnoses in EHRs compared with a clinical team database (CTD) and their impact on ML models.

Methods: EHRs from 2018-2022 were validated against CTD data to identify true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Logistic regression (LR) models were trained and tested using both EHR and validated labels, whereafter simulated label noise was introduced to increase FP and FN rates. Model performance was assessed using Receiver Operating Characteristic Area Under the Curve (ROC-AUC) and average precision (AP).

Results: Among 3,952 patients, 3,388 (85.7%) were correctly identified with GDM in both databases, while 564 cases lacked a GDM label in EHRs and 771 were missing a corresponding CTD label. Overall, 87.5% of cases were TN, 9.0% TP, 2.0% FP, and 1.5% FN. The model trained and tested with validated labels achieved a ROC-AUC of 0.817 and an AP of 0.450, whereas the same model tested using EHR labels achieved 0.814 and 0.395, respectively. Increased label noise during training led to gradual declines in ROC-AUC and AP, while noise in the test set -- especially elevated FP rates -- resulted in marked performance drops.

Conclusions: Discrepancies between EHR and CTD diagnoses had limited impact on model training but significantly affected performance evaluation when present in the test set, emphasising the importance of accurate data validation.

(JMIR Preprints 21/02/2025:72938)
DOI: https://doi.org/10.2196/preprints.72938

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

ORIGINAL ARTICLE

TITLE

Gestational Diabetes Diagnoses in Electronic Health Records: A Three-Step Study of Label Accuracy and Its Impact on Machine Learning Models for Early Prediction

AUTHOR LIST

Mark Germaine[1,2,5], Amy C O'Higgins[3], Brendan Egan[2,4], Graham Healy[1]

AFFILIATIONS

[1] School of Computing, Dublin City University, Dublin 9, Ireland

[2] School of Health and Human Performance, Dublin City University, Dublin 9, Ireland

[3] UCD Centre for Human Reproduction, The Coombe Hospital, Dublin 8, Ireland

[4] Florida Institute for Human and Machine Cognition, Pensacola FL, USA

[5] SFI Centre for Research Training in Machine Learning, Dublin City University, Dublin 9, Ireland

ORCIDs

Mark Germaine 0000-0002-7862-7714

Amy O'Higgins 0000-0002-2020-1585

Brendan Egan 0000-0001-8327-9016

Graham Healy 0000-0001-6429-6339

CORRESPONDING AUTHOR

Mark Germaine, SFI Centre for Research Training in Machine Learning, Dublin City University, Dublin 9, Ireland mark.germaine2@mail.dcu.ie

ABSTRACT WORD COUNT:	235
WORD COUNT:	2995
REFERENCES:	23
TABLES:	2
FIGURES:	4

ABSTRACT

Objective: This study examines the accuracy of gestational diabetes mellitus (GDM) diagnoses in electronic health records (EHRs) compared with a clinical team database (CTD) and their impact on machine learning (ML) models.

Methods: EHRs from 2018-2022 were validated against CTD data to identify true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Logistic regression (LR) models were trained and tested using both EHR and validated labels, whereafter simulated label noise was introduced to increase FP and FN rates. Model performance was assessed using Receiver Operating Characteristic Area Under the Curve (ROC-AUC) and average precision (AP).

Results: Among 3,952 patients, 3,388 (85.7%) were correctly identified with GDM in both databases, while 564 cases lacked a GDM label in EHRs and 771 were missing a corresponding CTD label. Overall, 87.5% of cases were TN, 9.0% TP, 2.0% FP, and 1.5% FN. The model trained and tested with validated labels achieved a ROC-AUC of 0.817 and an AP of 0.450, whereas the same model tested using EHR labels achieved 0.814 and 0.395, respectively. Increased label noise during training led to gradual declines in ROC-AUC and AP, while noise in the test set -- especially elevated FP rates -- resulted in marked performance drops.

Conclusion: Discrepancies between EHR and CTD diagnoses had limited impact on model training but significantly affected performance evaluation when present in the test set, emphasising the importance of accurate data validation.

Keywords: electronic health records, gestational diabetes, label noise, pregnancy, validation.

INTRODUCTION

Electronic health records (EHRs) are an important source of real-world data, offering detailed, longitudinal patient information historically stored in medical charts, and forming the basis of real-world evidence [1,2]. Together with advancements in artificial intelligence and machine learning (ML), EHRs are increasingly being used to develop models that improve prediction of health and disease outcomes [3].

Integration of EHRs into clinical research offers numerous opportunities for advancing healthcare delivery and patient outcomes. However, EHR data is often stored in unstructured formats like free text, requiring information extraction algorithms to enable to ML applications [4]. This extraction process can introduce data quality concerns due to various issues such as data entry errors and cut and paste errors [5]. The quality and consistency of EHR data is particularly critical when the target variable, i.e. the variable being predicted, is used in ML models.

Inaccuracies in EHRs present challenges for developing and applying ML algorithms in healthcare, primarily due to the dependency on data quality and accuracy of target labels [6]. This “label noise”, which refers to inaccuracies or inconsistencies in the data labels (e.g., diagnosis codes) extracted from EHRs, can significantly impact model performance by introducing errors in the target variable, leading to potentially misleading conclusions [7]. Training ML models on unvalidated EHRs may lead to systematic errors in the model output with the potential for the model to miss, underestimate, or overestimate clinically-significant relationships [8,9].

Accurate diagnosis and recording of gestational diabetes mellitus (GDM) in EHRs is important not only for effective patient management, but also for informing public health strategies and economic forecasting in national healthcare planning [10,11]. EHRs are often used to train ML approaches that support clinical decision-making and care pathways that improve pregnancy outcomes [12]. However, the utility of EHRs remains a concern due to potential discrepancies in data recording practices [8]. When using ML in GDM prediction [13], the accuracy of input data is

paramount because inaccuracies can lead to flawed prediction models, and ineffective or adverse clinical decisions [14].

Several studies have utilized EHRs to build ML models predicting the likelihood of developing GDM later in pregnancy [15], but none have described validation of the GDM 'label' within the EHRs. This study has three primary aims: first, to assess the accuracy of reporting of GDM diagnoses in EHRs by comparing them to a database maintained in real-time by the hospital's clinical team; second to evaluate how discrepancies in GDM reporting impact machine learning models; and third, to examine ML model performance using varying levels of simulated label noise in the dataset. By identifying discrepancies between these data sources, we aim to highlight the importance of data validation for advancing digital health and ML-driven healthcare.

METHODS

Study Design

A retrospective validation design was employed to assess the accuracy of GDM diagnoses recorded in the EHRs of a national maternity hospital (The Coombe Hospital, Dublin). We matched patient identifiers (IDs) between the EHR system and a reference standard established by a real-time clinical team database (CTD) of those formally-diagnosed with and managed for GDM, which served as a ground truth. This approach allowed for direct comparison between the recorded GDM status in the EHRs and the validated GDM status from the CTD, enabling identification of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) in the EHRs. Further, the effect of label noise on ML model performance in predicting the development of GDM (binary classification) was evaluated by firstly, examining its impact in our current EHR dataset, and then secondly, simulating progressively increasing levels of label noise to understand its effect on ML model performance both in terms of training and testing.

Data Source and Validation

The EHR system serves as the repository for patient medical histories, including diagnoses, family history, and outcomes for pregnant women receiving care at the institution. The data is collected routinely from all women by trained midwives using standardized questions and is then computerized onto the electronic system of the hospital, “Euroking K2”. EHRs were collected from 2013 to 2023 and consisted of over 80,000 pregnancies during this time. The dataset from the CTD spanned from 2018 to 2022, thus the timeframe for this analysis spanned from 2018 to 2022 (inclusive). Women aged 18 or above with complete information on GDM status were included in the analysis. Pregnancies with missing or incomplete data for critical variables, women without a recorded GDM status, and pregnancies with pre-existing diabetes were excluded. ML models were trained and tested on pregnancies with complete EHR data up to the 12th week of gestation.

GDM diagnoses were extracted from the EHRs based on information recorded in a column titled "Medical problems during pregnancy." When this column contained the entry "Diabetes developed during pregnancy", the patient was coded as having GDM in a newly created column designated for this study's analysis, referred to hereafter as "EHR-GDM." Patient records not meeting this criterion were coded as not having GDM.

Patient IDs from the EHRs were then matched with a separate database maintained in real-time by the clinical team responsible for diabetes care, with patient details entered each day upon confirmation from the hospital laboratory of a diagnosis of GDM from an oral glucose tolerance test (OGTT) following the International Association of the Diabetes and Pregnancy Study Groups (IADPSG) guidelines. This matching process produced a merged dataset for validating EHR-recorded GDM diagnoses against the CTD database, leading to the creation of two comparison columns: “EHR-GDM” for EHR-identified cases of GDM and “CTD-GDM” for cases of GDM recorded by the CTD.

The validation process involved comparing the GDM diagnosis status in the EHR (“EHR-

GDM") with that in the CTD ("CTD-GDM") to examine the agreement between the two datasets. This comparison allowed for the identification of TPs (positive in EHRs and present in CTD), FPs (positive in EHRs and not present in CTD), TNs (negative in EHRs and not present in CTD), and FNs (negative in EHRs and present in CTD), and thereby enabling evaluation of the accuracy of the reporting of GDM diagnosis in the EHRs. An additional column, VAL-GDM, was created indicating a positive or negative diagnoses of GDM for cases where the EHR-GDM and CTD-GDM labels matched i.e. for TPs and TNs excluding records with FPs and FNs. The true positive rate (TPR), false positive rate (FPR), true negative rate (TNR) and false negative rate (FNR) were calculated for the dataset [16].

Evaluation of Label Noise on ML Modelling

To evaluate the impact of label noise on the performance of ML models in predicting GDM, we employed logistic regression (LR) where the dataset was split into 70% training and 30% test sets to ensure robust evaluation. Default model hyperparameters were used, as the primary objective was to compare performance across different training datasets rather than optimising hyperparameter settings. The training and testing data comprised of EHR data that was available during the first booking visit, typically the 12^{th} week of gestation, and included 79 training features. The target label was GDM. The dataset contained both categorical and numerical features. Categorical features were processed using OneHotEncoder with the 'first' category dropped, and numerical features were standardized using StandardScaler. While the goal of this paper is not to produce an endpoint AI model, a self-assessment checklist for reporting was followed to ensure adequate information about the ML model was present [17].

We trained two ML models: one with the EHR-GDM labels and the other with the VAL-GDM labels. Both models were evaluated using the same test set, which used validated VAL-GDM labels, to facilitate a direct comparison of the effects of label noise during training on a consistent

test set. The year 2020 was excluded from these analyses due emerging research demonstrating reduced detection of diseases during this period [18], something that we confirm in our results below. By using both the 'raw' and 'validated' datasets, the study aimed to demonstrate the impact of label noise on model performance in the prediction of GDM, providing insights into the importance of accurate label validation in developing reliable predictive models using ML.

Additionally, varying levels of label noise were introduced to determine the threshold at which label noise significantly affects model performance. This simulation was performed by progressively increasing the number of FPs and FNs in the VAL-GDM training set from 0% to 90% i.e. changing a percentage of the positive labels to negative labels (creating FNs) and changing a percentage of negative labels to positive labels (creating FPs). This approach resulted in the training of 100 different models. Next, in a separate analysis we applied this progressive noise insertion to the VAL-GDM test set to specifically assess the impact of test set label noise on model evaluation i.e. evaluating these test sets using a model trained with the 'clean' VAL-GDM labels.

Statistical Analysis

The validation findings were quantitatively assessed using accuracy, precision, recall, F1 and overall agreement measured by Cohen's Kappa, between the EHR-recorded GDM diagnoses (EHR-GDM) and the clinical team database (CTD-GDM). The performance of the LR ML models were evaluated using Receiver Operating Characteristic Area Under Curve (ROC-AUC) and the Average Precision score (AP). The statistical and ML analysis were performed using Python version 3.8.8 with libraries including NumPy 1.23.5, pandas 1.2.4, and scikit-learn 1.2.1.

RESULTS

Population Characteristics

The dataset comprised 37,651 EHRs from 31,100 unique patients. The mean±SD patient age

was 32±5 years, and body mass index (BMI) was 26.2±5.5 kg/m^2, with 20.7% exhibiting a BMI greater than 30.0 kg/m^2. The prevalence of GDM according to the EHRs was 11.0%, whereas the prevalence according to the CTD was 10.5%. Patient characteristics for the most important features in the machine learning models are presented in Table 1.

Diagnosis Discrepancies

Of 3,952 patients with matching IDs in both databases, 3,388 were correctly identified with GDM in both EHR-GDM and CTD-GDM (9.0% TP and 85.7% TPR), while 564 lacked a corresponding GDM label in EHR-GDM (1.5% FN and 14.3% FNR) (Figure 1). Additionally, 771 patients were incorrectly identified with GDM in EHR-GDM without matching IDs in CTD-GDM (2.0% FP and 2.3% FPR). In EHRs there were 32,928 (87.5%) TN cases (97.7% TNR) (Figure 1). The accuracy, precision, F1 score and Cohen's kappa are reported in Table 2.

Yearly Data Comparison

Ninety-eight patients identified in CTD lacked corresponding entries in EHRs. Sixty-seven (68%) of these discrepancies were observed in 2020 (Figure 1). Furthermore, GDM prevalence for both EHRs and CTD datasets revealed a notable reduction in 2020 (recorded at 10.0% in EHRs and 7.7% in CTD), indicating a deviation from the trend observed in other years (Figure 2).

Label Noise in EHRs

The performance of LR models trained using the raw (EHR-GDM) and validated (VAL-GDM) labels was evaluated using a test set with VAL-GDM labels only. The model trained using the EHR-GDM labels achieved a ROC AUC of 0.817 and an AP score of 0.451. In comparison, the model trained using the VAL-GDM labels showed a ROC AUC of 0.817 and an AP score of 0.450 (Figure 3), indicating a minor impact of label noise in training the model for this dataset. However,

when the performance of the LR ML model trained using VAL-GDM labels was evaluated on a test set with EHR-GDM labels, a ROC AUC of 0.814 and an AP score of 0.395 was achieved, which demonstrates a greater impact of label noise when it is present in the test set.

Simulated Label Noise

The impact of simulated label noise on model performance was assessed by progressively increasing the number of FNs (pos_noise_level) and FPs (neg_noise_level) in the training set (where 0% noise equates to the original VAL-GDM labels) without modifying the testing set. The results demonstrate a decline in model performance as the level of label noise increases (Figure 4).

Further analysis of noise in the test set showed that model performance metrics, particularly ROC AUC and AP scores, were sensitive to increasing levels of noise, especially FP noise. As the FP rate (neg_noise_level) was increased, the ROC AUC consistently decreased, while the AP score initially decreased before increasing. The introduction of FN (pos_noise_level) into the test set had a less pronounced effect on performance compared to FP, unless both types of noise were combined, which led to a more substantial impact (Figure 4).

DISCUSSION

The This study highlights significant discrepancies be. tween GDM diagnoses recorded in EHRs and those validated by the CTD. Correcting label noise in the training set had a negligible impact on the performance of an LR-based ML model developed from EHRs to predict GDM from early pregnancy data. However, correcting label noise in the test set improved the model's average precision, underscoring the importance of accurate labelling for evaluating model performance accurately. The study also found that increasing label noise in the training set led to a gradual decline in model performance, whilst increasing FPs in the test set had a particularly strong negative impact on ROC-AUC, but counterintuitively increased AP scores. FNs had a less pronounced impact on

ROC-AUC unless combined with FP, which then caused a decline in model performance.

Approximately 14% (564/3,952) of GDM cases were not recorded in the EHRs, while 18.5% (771/4,159) of positive GDM diagnoses in EHRs did not align with CTD records. Overall, there were 32,928 (87.5%) TN, 3,388 (9.0%) TP, 771 (2.0%) FP, and 564 (1.5%) FN. The FPR (2.3%) remained low in comparison to the FNR (14.3%). Similar discrepancies in accuracy of EHRs have been reported in previous studies within Irish maternal hospitals, though with higher agreement in other contexts, such as miscarriage measurements (k=0.92) [19]. More widely across Europe, wide variations exist in the accuracy of reporting in EHRs as it relates to acute cardiovascular outcomes, with sensitivity reported at <66% for heart failure diagnoses in particular [20]. A key challenge in these studies is the absence of a recommended reference standard for validating EHR data, leading to the use of various data sources [8].

The impact of COVID-19 on screening and diagnostic practices, especially in 2020, manifested in a relative reduction of 31% in GDM diagnoses i.e. 11.2% across 2018, 2019, 2021, and 2022 compared to 7.7% in 2020. This observations aligns with research indicating reduced diagnosis rates for various medical conditions during the first year of the pandemic [18], and suggests caution being warranted when utilising EHRs during this year for the purpose of healthcare modelling.

Correcting label noise has been shown to mitigate its adverse effects on model performance, underscoring the importance of ‘clean’ and accurate datasets for training and validating ML algorithms to ensure their efficacy in clinical decision support systems [21]. For example, training a model on a ‘clean’ dataset resulted in an accuracy of 73.6%, whereas with 30% label noise the accuracy fell to 64.1% (-9.5%) [21]. However, the current analysis demonstrated that training a LR model using EHR-GDM labels versus validated VAL-GDM yielded negligible differences in performance metrics, with ROC AUCs of 0.817 and 0.817, respectively (Figure 3). This is presumably due to the low overall representation of FN and FP in the dataset of 3.5% combined.

Previous work has simulated noisy labels with artificial introduction of different levels of

label noise (10%, 20%, and 40%) into the training set, and demonstrated a gradual decline in the accuracy of all models (mean AUC of all models at 10%: 81.3, 20%: 80.2, and 40%: 78.4) as label noise increased [22]. The approach taken in the present study differs in that it introduces systematic label noise using Noise at Random (NAR) [23], increasing both the FP and FN rates linearly. Introducing noise into the training set resulted in a gradual decline in model performance, with both ROC AUC and AP scores decreasing as the level of noise increased. The model was particularly sensitive to FP, which caused a more pronounced decline in performance compared to FN. Introducing noise into the test set also impacted model performance, but the effects were more complex. The ROC AUC consistently decreased as FP rates increased, indicating that the model's ability to distinguish between classes was compromised. However, the AP score showed a different pattern, with an initial decline followed by an increase as noise levels were increased. The introduction of FN in the test set had a less pronounced effect on performance compared to FP, unless FP and FN were combined, which led to a more marked decline in the model's overall performance.

The increase in the AP score as the FP rate in the test set increased can be attributed to the method of calculating AP. AP evaluates the precision-recall trade-off across different thresholds, specifically calculating the proportion of TP to the sum of (TP + FP). When the majority of the negative class in the test set is artificially converted to positive, the opportunity for FP to occur is significantly reduced. This reduction in potential FP leads to an increase in precision, which in turn increases the AP score. Additionally, this manipulation dramatically alters the (e.g. class balance from 90% negative to 90% positive), further influencing the precision-recall dynamics and contributing to the observed ostensible increase in AP.

In conclusion, the identified discrepancies in EHR-recorded GDM diagnoses compared to 'true' GDM diagnoses reflect broader concerns about the accuracy of EHRs for public health and ML applications. Further, the magnitude of inaccuracies may play an important role for maximising the

utility of EHRs in enhancing healthcare outcomes, particularly for conditions such as GDM. However, when these discrepancies remain a small percentage (e.g. <5%) of the dataset, like in the case of the present study, there was no noticeable impact on model training performance. Conversely, the risk of incorrect model evaluation increases when the test set labels are impacted by noise, as this has a more pronounced effect on performance metrics. These observations emphasise the importance of incorporating robust data cleaning, preprocessing, and validation methodologies in the development of ML models for healthcare. Future efforts should aim at developing standardised validation protocols for EHRs to ensure high data quality for training and evaluating ML algorithms.

Abbreviations: (BMI) Body mass index, (CTD) Clinical team database, (EHR) Electronic health records, (GDM) Gestational diabetes mellitus, (IADPSG) International Association of Diabetes and Pregnancy Study Groups, (IDs) patient identifiers, (ML) Machine learning, (OGTT) oral glucose tolerance test

Acknowledgments

The authors are grateful to The Coombe Hospital for their collaboration without whom this analysis would not be possible.

Data and Resource Availability

The datasets analysed during the current study are not publicly available due to use of patient IDs to match the corresponding databases.

Funding and Assistance

This work has emanated from research supported in part by a grant from Science Foundation Ireland under Grant Number 18/CRT/6183.

Conflicts of Interest Disclosure

The authors have no conflicts of interest to declare.

1. Garrison Jr. LP, Neumann PJ, Erickson P, Marshall D, Mullins CD. Using Real-World Data for Coverage and Payment Decisions: The ISPOR Real-World Data Task Force Report. Value in Health 2007;10(5):326–335. doi: 10.1111/j.1524-4733.2007.00186.x

2. Berger ML, Sox H, Willke RJ, Brixner DL, Eichler H-G, Goettsch W, Madigan D, Makady A, Schneeweiss S, Tarricone R, Wang SV, Watkins J, Daniel Mullins C. Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. Pharmacoepidemiology and Drug Safety 2017;26(9):1033–1039. doi: 10.1002/pds.4297

3. Wong J, Murray Horwitz M, Zhou L, Toh S. Using Machine Learning to Identify Health Outcomes from Electronic Health Record Data. Curr Epidemiol Rep 2018 Dec 1;5(4):331–342. doi: 10.1007/s40471-018-0165-9

4. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. Journal of the American Medical Informatics Association 2016 Sep 1;23(5):1007–1015. doi: 10.1093/jamia/ocv180

5. Bowman S. Impact of Electronic Health Record Systems on Information Integrity: Quality and Safety Implications. Perspect Health Inf Manag 2013 Oct 1;10(Fall):1c. PMID:24159271

6. Frénay B, Kaban A. A Comprehensive Introduction to Label Noise. 2014. Available from: https://dial.uclouvain.be/pr/boreal/object/boreal:156597 [accessed Mar 17, 2024]

7. Yang J, Triendl H, Soltan AAS, Prakash M, Clifton DA. Addressing label noise for electronic health records: insights from computer vision for tabular data. BMC Medical Informatics and Decision Making 2024 Jun 27;24(1):183. doi: 10.1186/s12911-024-02581-5

8. Nissen F, Quint JK, Morales DR, Douglas IJ. How to validate a diagnosis recorded in electronic health records. Breathe European Respiratory Society; 2019 Mar 1;15(1):64–68. doi: 10.1183/20734735.0344-2018

9. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. Commun ACM 2021 Feb 22;64(3):107–115. doi: 10.1145/3446776

10. Cebul RD, Love TE, Jain AK, Hebert CJ. Electronic Health Records and Quality of Diabetes Care. New England Journal of Medicine Massachusetts Medical Society; 2011 Sep 1;365(9):825–833. doi: 10.1056/NEJMsa1102519

11. Veinot TC, Zheng K, Lowery JC, Souden M, Keith R. Using electronic health record systems in diabetes care: emerging practices. Proceedings of the 1st ACM International Health Informatics Symposium New York, NY, USA: Association for Computing Machinery; 2010. p. 240–249. doi: 10.1145/1882992.1883026

12. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Internal Medicine 2018 Nov 1;178(11):1544–1547. doi: 10.1001/jamainternmed.2018.3763

13. Mennickent D, Rodríguez A, Farías-Jofré M, Araya J, Guzmán-Gutiérrez E. Machine learning-based models for gestational diabetes mellitus prediction before 24–28 weeks of pregnancy: A review. Artificial Intelligence in Medicine 2022 Oct 1;132:102378. doi: 10.1016/j.artmed.2022.102378

14. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, Aardoom JJ, Debray TPA, Schuit E, van Smeden M, Reitsma JB, Steyerberg EW, Chavannes NH, Moons KGM. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. npj Digit Med Nature Publishing Group; 2022 Jan 10;5(1):1–13. doi: 10.1038/s41746-021-00549-7

15. Zhang Z, Yang L, Han W, Wu Y, Zhang L, Gao C, Jiang K, Liu Y, Wu H. Machine Learning Prediction Models for Gestational Diabetes Mellitus: Meta-analysis. Journal of Medical Internet Research 2022 Mar 16;24(3):e26634. doi: 10.2196/26634

16. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters 2006 Jun 1;27(8):861–874. doi: 10.1016/j.patrec.2005.10.010

17. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. International Journal of Medical Informatics 2021 Sep 1;153:104510. doi: 10.1016/j.ijmedinf.2021.104510

18. Burus T, Lei F, Huang B, Christian WJ, Hull PC, Ellis AR, Slavova S, Tucker TC, Lang Kuhs KA. Undiagnosed Cancer Cases in the US During the First 10 Months of the COVID-19 Pandemic. JAMA Oncology 2024 Feb 22; doi: 10.1001/jamaoncol.2023.6969

19. San Lazaro Campillo I, Meaney S, Harrington M, McNamara K, Verling AM, Corcoran P, O'Donoghue K. Assessing the concordance and accuracy between hospital discharge data, electronic health records, and register books for diagnosis of inpatient admissions of miscarriage: A retrospective linked data study. Journal of Obstetrics and Gynaecology Research 2021;47(6):1987–1996. doi: 10.1111/jog.14785

20. Davidson J, Banerjee A, Muzambi R, Smeeth L, Warren-Gash C. Validity of Acute Cardiovascular Outcome Diagnoses Recorded in European Electronic Health Records: A Systematic Review. Clinical Epidemiology Dove Medical Press; 2020 Oct 14;12:1095–1111. doi: 10.2147/CLEP.S265619

21. Bernhardt M, Castro DC, Tanno R, Schwaighofer A, Tezcan KC, Monteiro M, Bannur S, Lungren MP, Nori A, Glocker B, Alvarez-Valle J, Oktay O. Active label cleaning for improved dataset quality under resource constraints. Nat Commun Nature Publishing Group; 2022 Mar 4;13(1):1161. doi: 10.1038/s41467-022-28818-3

22. Ju L, Wang X, Wang L, Mahapatra D, Zhao X, Zhou Q, Liu T, Ge Z. Improving Medical Images Classification With Label Noise Using Dual-Uncertainty Estimation. IEEE Transactions on Medical Imaging 2022 Jun;41(6):1533–1546. doi: 10.1109/TMI.2022.3141425

23. Sáez JA. Noise Models in Classification: Unified Nomenclature, Extended Taxonomy and Pragmatic Categorization. Mathematics Multidisciplinary Digital Publishing Institute; 2022 Jan;10(20):3736. doi: 10.3390/math10203736

TABLES

Table 1. Patient characteristics for the most important features in the machine learning models, according to the validated dataset.

Characteristics	Mean±SD/Prevalence
Age (years)	32±5
BMI (kg/m^2)	26.2±5.3
Systolic Blood Pressure (mmHg)	111±11
Diastolic Blood Pressure (mmHg)	67±8
Parity	0.9±1.1
Ethnic Origin of Patient	
Caucasian	87.8%
South East Asian	4.9%
Black	2.0%

Asian	1.8%
Middle Eastern	0.6%
Latin American	0.1%
Mixed	0.1%
Other	3.0%
Occupation Skill Level (ISCO)	
Level 0 (Unemployed)	19.0%
Level 1 (Elementary occupations)	1.3%
Level 2 (Clerical and Service)	15.9%
Level 3 (Technicians & Associates)	8.6%
Level 4 (Professionals and Managers)	55.1%
Family history of diabetes mellitus	23.3%
History of GDM	3.9%
Other Endocrine Problems	21.4%
Prevalence of GDM	11.7%

Table 2. Performance metrics for the comparison of GDM diagnoses in electronic health records (EHR) with the real-time clinical team database (CTD).

Year	Cohen's Kappa	Accuracy	Precision	Recall	F1 Score
All Years	0.82	0.96	0.81	0.86	0.84
2018	0.80	0.96	0.78	0.86	0.82
2019	0.82	0.96	0.86	0.82	0.84
2020	0.77	0.96	0.70	0.90	0.79
2021	0.86	0.97	0.89	0.87	0.88
2022	0.82	0.97	0.82	0.86	0.84
All minus 2020	0.82	0.97	0.84	0.85	0.84

Supplementary Table

Table 3. Comparison of receiver operating characteristic area under the curve (ROC AUC) and average precision for machine learning models predicting gestational diabetes mellitus (GDM) trained on the raw data and on the validated data.

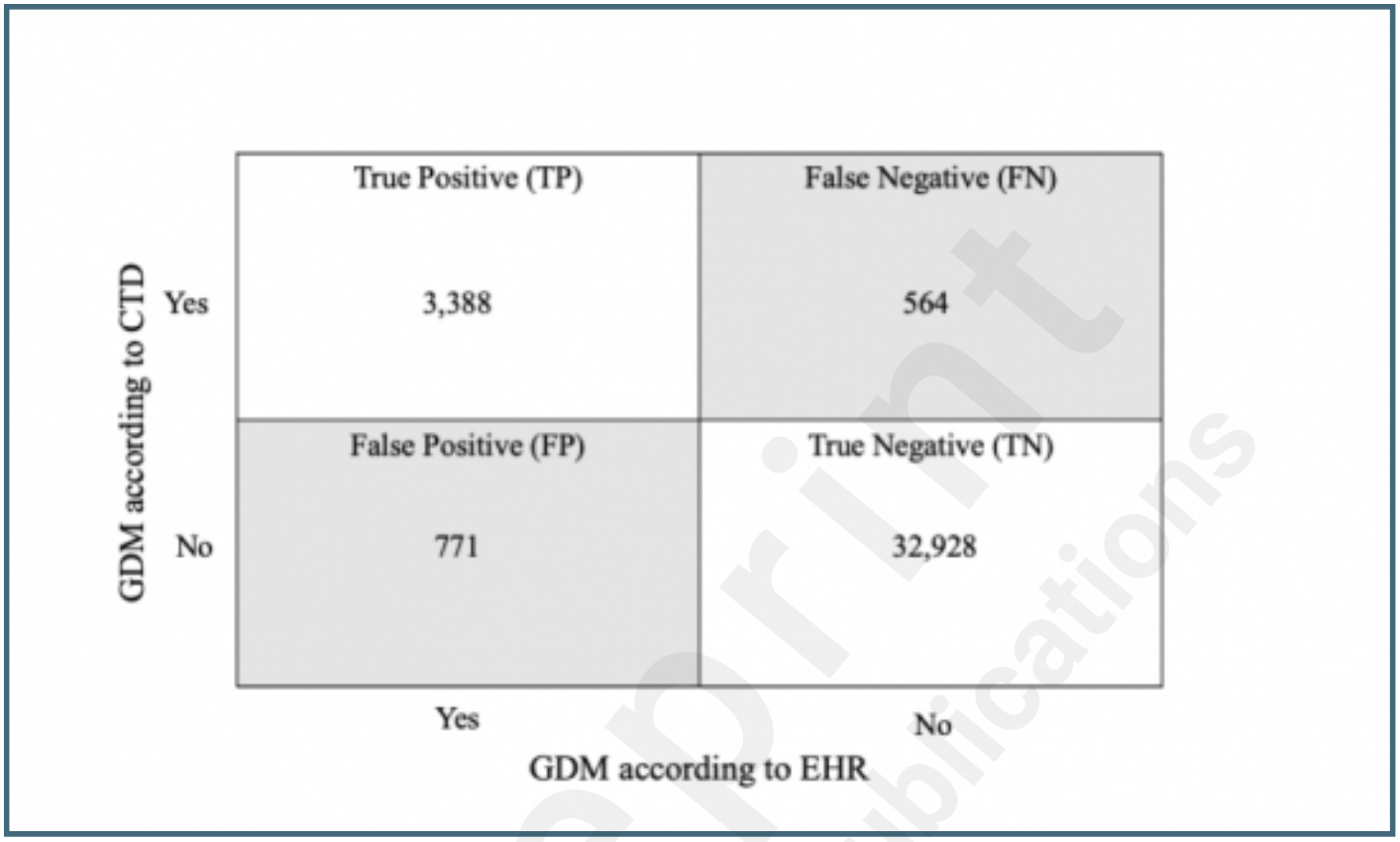
Model	ROC AUC		Average Precision	
	Raw	Val	Raw	Val
Logistic Regression	0.819	0.818	0.455	0.453
Random Forest	0.798	0.801	0.422	0.423

XGBoost	0.787	0.783	0.401	0.386
EBM	0.818	0.816	0.456	0.453

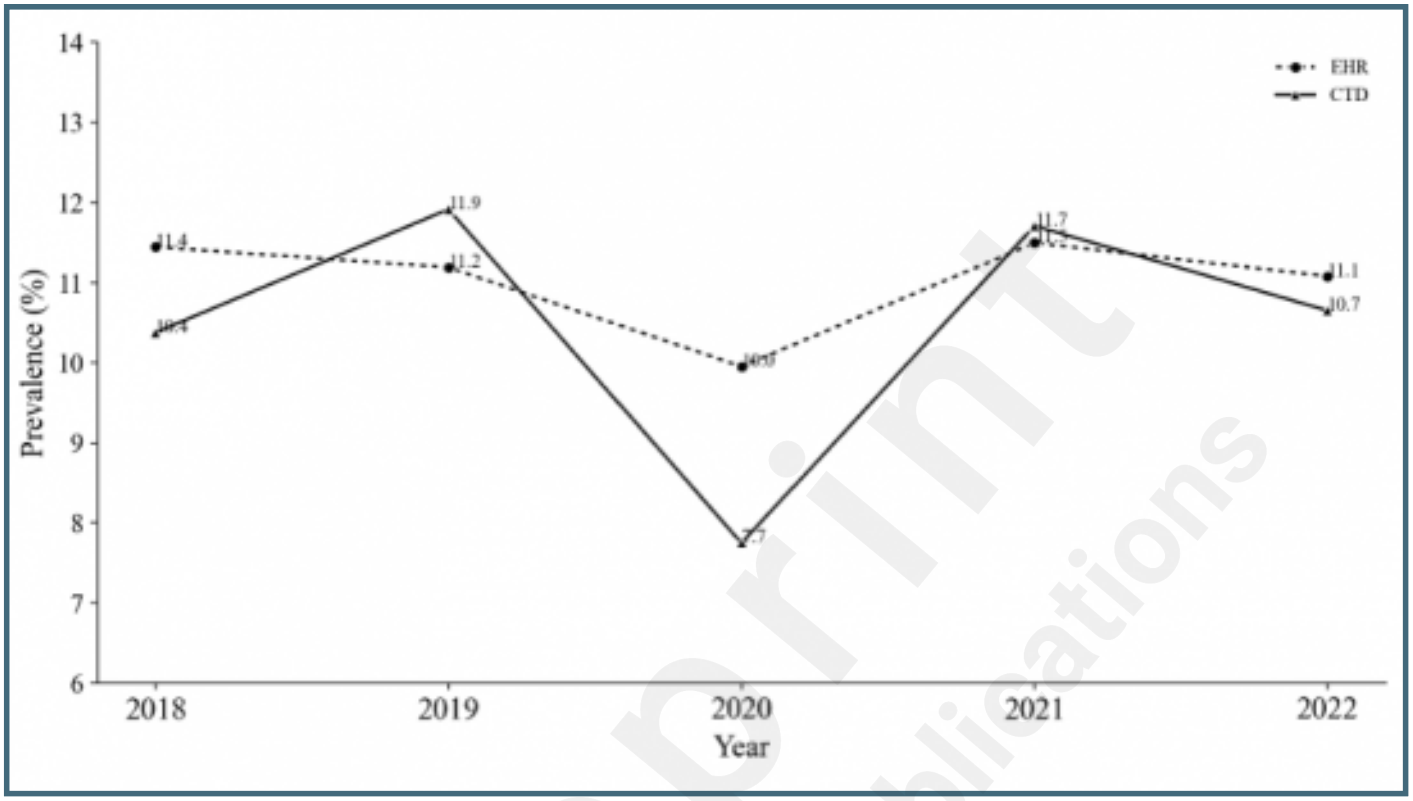
Supplementary Files

Figures

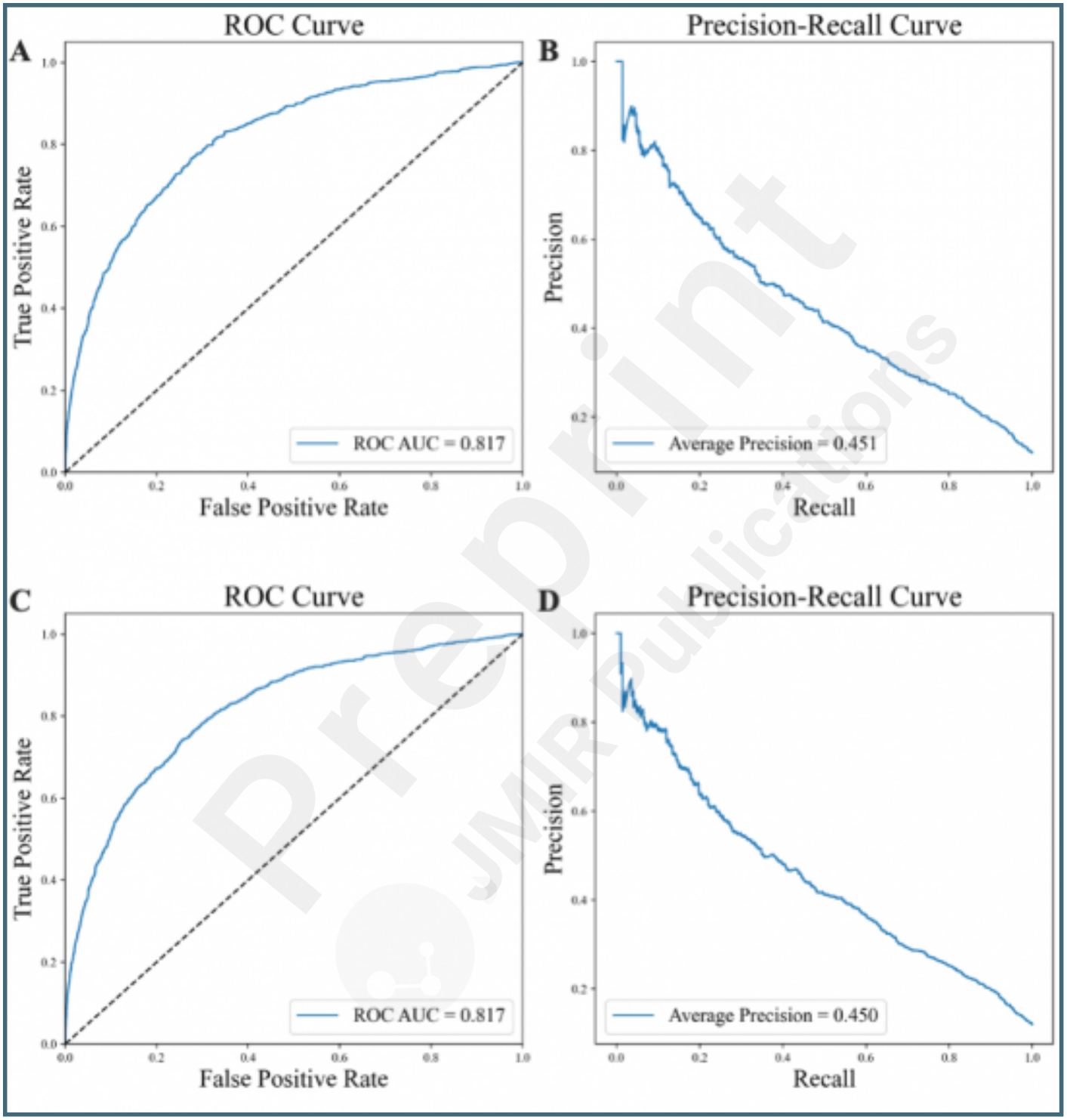
Confusion matrix comparing GDM diagnoses in electronic health records (EHR-GDM) with validated data from the clinical team database (CTD-GDM). The matrix shows 3,388 true positive cases (TP), 564 false negative cases (FN), 771 false positive cases (FP), and 32,928 true negative cases (TN).



Comparison of prevalence rates of GDM diagnosis between electronic health record (EHR-GDM) data and the clinical team database (CTD-GDM) from 2018 to 2022. The solid line represents the CTD data, and the dashed line represents the EHR data.



Receiver operating characteristic (ROC) curves and precision-recall (PR) curves for logistic regression models predicting gestational diabetes mellitus (GDM) trained using the EHR-GDM data (A, B) and on the VAL-GDM validated data (C, D).



Heatmap showing the impact of simulated label noise in the training set on model performance (ROC AUC, A; AP, B), and in the test set on model performance (ROC AUC, C; AP, D) across varying levels of false positive (Neg Noise Level) and false negative (Pos Noise Level) rates.