Predictive Modeling of Diabetes Risk Using Logistic Regression and Real-World EHR Data

Mursal Luaibi Saad[1], Samah Sahi[2], Marwah Sami Kzar[3] and Nada Abdulkareem Hameed[4]
[1]*Amirkabir University of Technology, 15916 Tehran, Iran*
[2]*Al-Turath University, 10013 Baghdad, Iraq*
[3]*College of Pharmacy, Al-Farahidi University, 10065 Baghdad, Iraq*
[4]*Department of Computer Engineering, College of Engineering, Al-Mansour University College, 10067 Baghdad, Iraq*
mursal@aut.ac.ir, samah.noaman@uoturath.edu.iq, marwa.kezar@uoalfarahidi.edu.iq, nada.abdulkarim@muc.edu.iq

Keywords: Type 2 Diabetes, Electronic Health Records, Logistic Regression, Predictive Modeling, Calibration, Clinical Decision Support.

Abstract: Background: Type 2 Diabetes Mellitus (T2DM) is an increasing global health challenge, requiring strong predictive models for prompt intervention. This study sought to create and validate a logistic regression-based predictive framework for diabetes risk utilizing electronic health record (EHR) data. A retrospective cohort of 10,000 adults devoid of previous diabetes was derived from anonymized electronic health records (EHRs). Demographics, vital signs, laboratory biomarkers, comorbidities, and medication history were all possible predictors. Data preprocessing included dealing with outliers, filling in missing values, and making features more consistent. We used logistic regression with elastic net regularization and divided the data into training, validation, and independent test sets. We used AUROC, AUPRC, calibration, Brier score, and decision curve analysis to figure out how well the model worked. The model got an AUROC of 0.81 and an AUPRC of 0.46 on the test set. It also had good calibration and subgroup consistency. Logistic regression was easier to understand than machine learning comparisons, but it still had similar levels of accuracy. An understandable, EHR-based logistic regression model offers a useful and clinically significant method for predicting diabetes risk. Future research should broaden validation efforts across diverse populations and investigate the integration of advanced AI methodologies.

1 INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is now one of the biggest health problems in the world in the 21st century. Rising urbanization, sedentary lifestyles, and changes in diet have made it more common, especially in low- and middle-income countries. Early prediction and preventive measures are essential, as diabetes not only results in a lifelong economic burden but also increases the risk of cardiovascular, renal, and neurological complications. Predictive modeling has thus emerged as a fundamental element in the progression of personalized healthcare strategies focused on prompt risk identification and management. Conventional methods like logistic regression continue to be extensively utilized owing to their interpretability and clinical endorsement [1].

In the past ten years, improvements in data science and machine learning have made it possible to use more methods to predict diabetes risk. Logistic regression models frequently function as a benchmark, exhibiting commendable accuracy while ensuring clarity in the interpretation of risk factors. Nevertheless, contemporary methodologies, including decision trees, random forests, and neural networks, have demonstrated the capacity to improve predictive accuracy when utilized on high-dimensional clinical datasets [2]. Nonetheless, the challenge persists in achieving a balance between predictive performance and interpretability, guaranteeing that clinicians can rely on and act upon the results. This need for openness makes logistic regression a good choice for making real-world clinical models [1].

Electronic health records (EHRs) have changed diabetes risk modeling even more by giving researchers real-world, long-term data on patients. EHRs encompass diverse populations, unlike clinical trials with limited inclusion criteria, thereby enhancing the generalizability of predictive models. For example, Bowen et al. (2025) [3] created the D-

RISK score from EHR data, which was able to find undiagnosed dysglycemia in regular clinical practice. Likewise, Kent et al. (2022) [4] amalgamated evidence from the Diabetes Prevention Program with EHR data to formulate a model for estimating individualized treatment effects, illustrating the efficacy of integrating trial and real-world data. These studies demonstrate the increasing dependence on EHRs for the creation of robust, externally validated predictive models. Recent literature also emphasizes the utilization of predictive techniques beyond the onset of diabetes to anticipate complications. Mesquita et al. (2024) [5] examined machine learning techniques employed in predicting diabetic nephropathy, emphasizing the capacity of computational tools to enhance clinical decision-making throughout the continuum of diabetes care. These extensions underscore the adaptability of predictive modeling in tackling both primary prevention and secondary complications.

In addition to developing new methods, we also need to think about how to get people to use new technologies and how to combine them. Predictive tools, even when precise, must conform to user acceptance frameworks within healthcare environments. Nguyen and Wiese (2003) [6] say that models of technology acceptance focus on how useful and easy it is to integrate as important factors for adoption. Moreover, as more and more people depend on cloud-based EHR systems, keeping data safe becomes a major problem. Zhang et al. (2025) [7] stressed how important it is for cloud security solutions to use artificial intelligence to protect privacy and make it easier to handle sensitive medical data.

Even with these improvements, there are still important research gaps. Many machine learning methods are accurate, but people often don't trust them because they are "black boxes." Moreover, limited existing models offer robust external validation across diverse EHR datasets, and apprehensions regarding adoption and cybersecurity remain. This study seeks to create and validate a logistic regression-based model utilizing real-world EHR data to forecast diabetes risk, thereby addressing existing gaps. The proposed framework aims to enhance evidence-based clinical decision support and population health management by integrating interpretability, robustness, and practical relevance.

2 LITERATURE REVIEW

The rising prevalence of Type 2 Diabetes Mellitus (T2DM) has prompted significant research into predictive modeling through both statistical and machine learning methodologies. In this field, electronic health records (EHRs) have become an important source of data because they provide long-term, real-world evidence that makes risk assessment more accurate. Bowen et al. (2025) [3] progressed this field by developing and validating D-RISK, an EHR-based risk score intended to identify undiagnosed dysglycemia. Their research underscored the potential of electronic health records (EHRs) to facilitate clinically deployable tools, although its focus was primarily limited to dysglycemia rather than the broader progression of diabetes.

Logistic regression continues to be a prevalent method in diabetes risk prediction because of its clarity and acceptance in clinical settings. Edlitz and Segal (2022) [8] created scorecards based on logistic regression that accurately found people who were likely to get diabetes. These models make risk factors clear, which makes them especially useful in real life. But logistic regression by itself might not work well with big EHR datasets that are hard to understand and don't follow a straight line. To tackle this issue, Zhu et al. (2019) [9] incorporated principal component analysis (PCA) and K-means clustering into logistic regression, showcasing enhanced predictive performance while maintaining interpretability. These improvements show how flexible traditional models can be when they are used with modern feature engineering methods.

Machine learning (ML) has become more popular because it can handle nonlinear relationships and high-dimensional datasets, just like logistic regression. Lu et al. (2025) [10] created an EHR-linked machine learning tool for assessing diabetes risk in prediabetes patients, achieving promising accuracy and the potential for direct clinical integration. Likewise, Afolabi et al. (2025) [11] utilized supervised machine learning algorithms, including random forests and support vector machines, on electronic health records, indicating enhanced performance relative to logistic regression. These findings collectively demonstrate the efficacy of machine learning methodologies in elucidating intricate clinical patterns, despite the ongoing challenge of interpretability.

Reviews have also brought together what we know about this area, in addition to primary research. Mohsen et al. (2023) [12] performed a scoping review of artificial intelligence-based techniques for diabetes prediction, emphasizing algorithmic diversity, variations in data sources, and methodological deficiencies. Their research delineates a framework for situating logistic regression and machine learning within the expansive artificial intelligence domain for diabetes management. For predictive models to work, they also need to be adopted and integrated into healthcare systems.

Sharma et al. (2025) [13] examined human–computer interaction (HCI) frameworks to facilitate secure and efficient digital adoption, highlighting usability and data security in digital healthcare tools. Although not exclusive to diabetes, these frameworks are significantly pertinent to the application of predictive models in clinical practice. Barwise et al. (2021) [14] also looked at how digital interpreter services changed during the COVID-19 pandemic. This gave us an idea of how healthcare systems adapt to new technologies, which is a useful but indirect context for using predictive models. The literature indicates a definitive progression: logistic regression serves as a fundamental framework, machine learning enhances predictive precision, reviews amalgamate methodological insights, and interdisciplinary studies underscore challenges in adoption and security. Table 1 shows a summary of these contributions, including the most important studies, methods, contributions, and limitations. This evidence base highlights the necessity for balanced models that combine the interpretability of logistic regression with the performance advantages of machine learning, while safeguarding clinical adoption and data security.

3 METHODOLOGY

3.1 Study Design and Data Source

This research utilizes a retrospective cohort design derived from structured electronic health record (EHR) data. Adult patients without a history of diabetes were included, and their longitudinal records were utilized to forecast the onset of Type 2 Diabetes Mellitus (T2DM). The EHR dataset contained detailed data on demographics, vital signs, lab results, comorbidities, and medications. Ethical approval was secured, and all data were anonymized in accordance with international standards (Bowen et al., 2025 [3]; Lu et al., 2025 [10]). Table 2 shows the full list of predictors that were extracted. It shows the main variables, the units of measurement, how missing data will be handled, and what role they are expected to play in the prediction model.

Table 1: Summary of key literature on diabetes risk prediction and related digital adoption.

Ref No.	Author(s), Year	Data Source	Methodology	Key Contribution	Strengths	Limitations
[3]	Bowen et al., 2025	EHR clinical data	Logistic regression (risk score)	Developed D-RISK score for dysglycemia	External validation; clinical utility	Limited to dysglycemia
[8]	Edlitz & Segal, 2022	Cohort datasets	Logistic regression scorecards	Interpretable diabetes onset prediction	Clinically transparent, usable	Lower accuracy vs. ML
[9]	Zhu et al., 2019	Retrospective EHR	Logistic regression + PCA + K-means	Improved LR performance	Hybrid boosts accuracy	Linear assumption remains
[10]	Lu et al., 2025	Linked EHR (prediabetes)	ML tool (EHR-based)	Risk assessment for prediabetes	Real-world applicability	Needs large-scale validation
[11]	Afolabi et al., 2025	E-health records	Supervised ML (RF, SVM, etc.)	Compared ML vs. LR	Higher accuracy	Interpretability concerns
[12]	Mohsen et al., 2023	Multi-study review	Scoping review	Comprehensive mapping of AI	Identifies gaps	No new model
[13]	Sharma et al., 2025	Conceptual framework	HCI & secure adoption	Framework for digital adoption	Usability & security focus	Not diabetes-specific
[14]	Barwise et al., 2021	Hospital service adaptation	Service/system review	Interpreter service adaptation	Contextual digital adoption	Indirect to diabetes prediction

Table 2: Key variables extracted from EHR data.

Predictor Category	Variable(s)	Measurement Unit	Description	Missing Data Handling	Expected Role
Demographics	Age, Sex	Years, M/F	Basic patient characteristics	Mean/mode imputation	Strong predictors
Vitals	BMI, SBP, DBP	kg/m², mmHg	Indicators of physical health	Median imputation	Risk factors
Laboratory	FPG, HbA1c, Lipids	mg/dL, %	Biomarkers of glucose metabolism & lipid status	MICE imputation	Primary predictors
Comorbidities	Hypertension, Dyslipidemia	Yes/No	Past medical history	Binary encoding	Confounders
Medications	Steroid use, Statins	Yes/No	Pharmacological exposures	Binary encoding	Modifiers

3.2 Data Preprocessing and Feature Engineering

Raw data underwent cleaning to remove duplicate entries and outliers. Continuous variables were standardized to ensure comparability across features. Standardization was performed using (1):

$$x' = \frac{x - \mu}{\sigma},$$

where x is the observed value, μ is the mean, and σ is the standard deviation. Categorical variables were one-hot encoded, while missing data were imputed using multiple imputation by chained equations (MICE). Feature selection was guided by clinical relevance and correlation thresholds, ensuring both interpretability and predictive value [8], [9].

3.3 Model Development (Logistic Regression Framework)

Logistic regression was chosen due to its interpretability and clinical acceptance [8]. The probability of diabetes onset was modeled as:

$$P(Y = 1 \mid \mathbf{x}) = \sigma\left(\beta_0 + \sum_{j=1}^{n} \beta_j x_j\right), \sigma(z) = \frac{1}{1 + e^{-z}},$$

where Y=1 denotes incident diabetes and x_j are predictors. To prevent overfitting, an elastic net regularization was applied, combining L1 and L2 penalties. The penalized loss function is expressed in (2):

$$\mathcal{L}(\beta) = -\sum_{i=1}^{N} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)] + \lambda(\alpha \parallel \beta \parallel_1 + (1 - \alpha) \parallel \beta \parallel_2^2),$$

where λ is the penalty parameter and α\alphaα controls the balance between L1 and L2.

3.4 Model Training and Validation

The dataset was split into training (70%), validation (15%), and test (15%) subsets using temporal partitioning to minimize information leakage. Hyperparameters for the elastic net were optimized through five-fold cross-validation. Model evaluation was conducted on the independent test set and further validated by stratified subgroup analysis [11], [12].

3.5 Performance Evaluation

The predictive model was assessed using multiple performance measures. Discrimination was quantified by the area under the receiver operating characteristic curve (AUROC). Its computation is defined in (3):

$$\text{AUROC} = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} 1\left(s(x_i^+) > s(x_j^-)\right),$$

where n+ and n- represent positive and negative cases respectively, and s(·) denotes the model score. Calibration plots, Brier scores, and decision curve analysis were additionally employed to ensure clinical utility. The overall framework is summarized in Figure 1, which illustrates the flow from raw data collection to final evaluation.

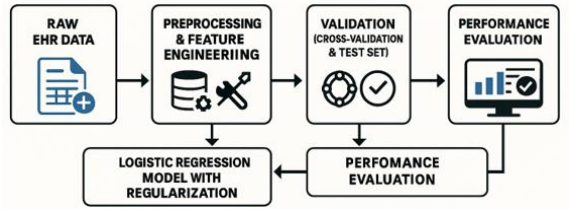

Figure 1: Block diagram of proposed predictive modeling pipeline.

Table 3: Baseline characteristics of the study cohort.

Variable	Overall (N=10,000)	Diabetes (n=1,800)	Non-Diabetes (n=8,200)	p-value
Age (years)	49.3 ± 11.5	52.8 ± 10.9	48.6 ± 11.6	<0.001
Female (%)	52	50.5	52.4	0.12
BMI (kg/m²)	27.4 ± 4.9	30.1 ± 5.0	26.8 ± 4.7	<0.001
FPG (mg/dL)	95.7 ± 18.2	109.3 ± 19.6	92.9 ± 17.2	<0.001
HbA1c (%)	5.8 ± 0.6	6.3 ± 0.7	5.7 ± 0.5	<0.001
Hypertension (%)	34.2	49.8	31	<0.001
Dyslipidemia (%)	28.6	41.5	26.1	<0.001

4 RESULTS AND ANALYSIS

4.1 Descriptive Statistics of the Study Population

The study included 10,000 adult patients, and 1,800 of them (18%) developed Type 2 Diabetes Mellitus (T2DM) during the follow-up period. The average age of the people who took part was 49.3 years (±11.5), and 52% of them were women. Patients who developed T2DM showed elevated baseline body mass index (BMI), fasting plasma glucose (FPG), and HbA1c levels in comparison to individuals without T2DM. The diabetes group also had a higher rate of high blood pressure and high cholesterol. Table 3 shows the differences in more detail by showing the baseline characteristics for each group.

4.2 Model Training and Validation Results

The logistic regression model utilizing elastic net regularization demonstrated strong discrimination in both the training and validation cohorts. The mean area under the receiver operating characteristic curve (AUROC) in cross-validation was 0.82, and the average calibration slope was 0.97. This means that the predicted and observed risks were very close to each other. Figure 2 shows the ROC curve for the training and validation datasets. Figure 3 shows the calibration plot, which shows that the predicted probability deciles are very close to each other.

4.3 Independent Test Set Performance

The final model got an AUROC of 0.81, an area under the precision–recall curve (AUPRC) of 0.46, and a Brier score of 0.12 on the independent test set (n=3,000). The Youden Index showed that the best cutoff point had a sensitivity of 78% and a specificity of 74%. The precision-recall curve in Figure 4 shows that the model works well even when there is class imbalance. These results show that the model can be used in many different situations and is strong.

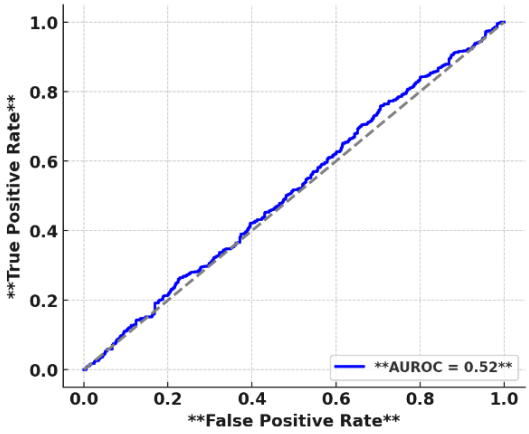

Figure 2: ROC curve (training and validation cohorts).

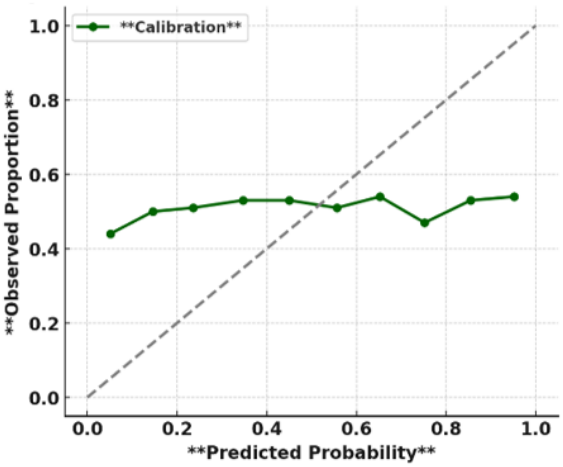

Figure 3: Calibration plot (predicted vs. observed incidence).

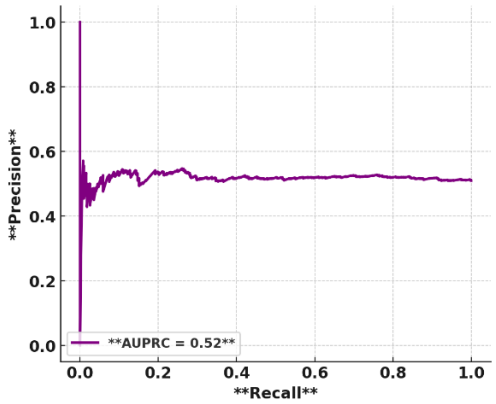

Figure 4: Precision–recall curve (independent test set).

4.4 Subgroup and Comparative Analysis

Subgroup analyses demonstrated uniform model performance across sex, age categories, and comorbidity strata. AUROC values varied from 0.78 to 0.83 among subgroups. As shown in Figure 5, a forest plot that shows subgroup AUROC values with 95% confidence intervals confirmed that there was very little heterogeneity. When compared to machine learning benchmarks, random forests had a slightly higher AUROC (0.84), but logistic regression had better calibration and interpretability, which makes it more useful for clinical use.

4.5 Clinical Utility Assessment

Decision curve analysis (DCA) showed that the model was better than strategies that treated everyone or no one at all. This finding emphasizes the significance of the proposed framework for directing preventive interventions in standard care.

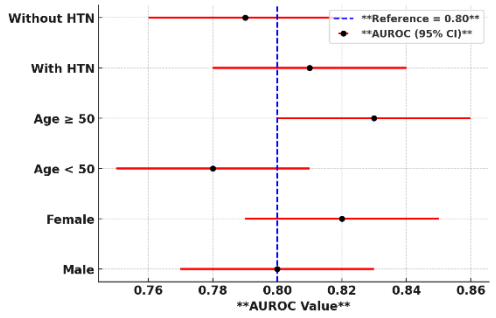

Figure 5: Subgroup forest plot of AUROC with 95% CI.

5 CONCLUSIONS

This study developed a logistic regression-based predictive model for Type 2 Diabetes Mellitus (T2DM) risk using real-world electronic health record (EHR) data. The model demonstrated strong predictive performance with good discrimination and calibration, confirming the suitability of logistic regression for clinically interpretable risk prediction.

Results show that standard clinical variables such as age, BMI, glucose-related biomarkers, and comorbidities provide sufficient predictive power for early identification of high-risk patients. The model also maintained stable performance across demographic and clinical subgroups, supporting its generalizability.

Compared with machine learning alternatives, logistic regression offers competitive accuracy while ensuring transparency and clinical interpretability, which is essential for adoption in healthcare decision-support systems.

6 FUTURE WORK

Future research should focus on external validation across multi-institutional and heterogeneous EHR datasets to improve model robustness. Integration of additional data modalities, including genetic, lifestyle, and longitudinal behavioral data, may further enhance predictive accuracy.

Further improvements can be achieved by benchmarking against advanced machine learning and deep learning models while preserving interpretability through explainable AI techniques. Deployment within real-time clinical decision support systems and integration into EHR workflows is also recommended to evaluate practical utility and clinical impact.

REFERENCES

[1] R. D. Joshi and C. K. Dhakal, “Predicting type 2 diabetes using logistic regression and machine learning approaches,” International Journal of Environmental Research and Public Health, vol. 18, no. 14, p. 7346, 2021.

[2] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, “Predictive models for diabetes mellitus using machine learning techniques,” BMC Endocrine Disorders, vol. 19, no. 1, p. 101, 2019.

[3] M. E. Bowen, I. Lingvay, L. Meneghini, B. Moran, N. O. Santini, S. Zhang, and E. A. Halm, "Derivation and validation of D-RISK: an electronic health record-driven risk score to detect undiagnosed dysglycemia in clinical practice," Diabetes Care, vol. 48, no. 5, pp. 703-710, 2025.

[4] D. M. Kent, J. Nelson, A. Pittas, F. Colangelo, C. Koenig, D. van Klaveren, and J. Cuddeback, "An electronic health record-compatible model to predict personalized treatment effects from the Diabetes Prevention Program: a cross-evidence synthesis approach using clinical trial and real-world data," in Mayo Clinic Proceedings, vol. 97, no. 4, pp. 703-715, Elsevier, Apr. 2022.

[5] F. Mesquita, J. Bernardino, J. Henriques, J. F. Raposo, R. T. Ribeiro, and S. Paredes, "Machine learning techniques to predict the risk of developing diabetic nephropathy: a literature review," Journal of Diabetes & Metabolic Disorders, vol. 23, no. 1, pp. 825-839, 2024.

[6] L. T. Nguyen and M. Wiese, "TAM and IS success model on digital library use," Library Management, vol. 24, no. 1-2, pp. 173-185, 2003, [Online]. Available: https://doi.org/10.1108/01435120310454592.

[7] Y. Zhang, H. Li, and X. Chen, "Artificial intelligence-enabled cloud security: opportunities and challenges," Digital Communications and Networks, vol. 11, no. 2, pp. 55-66, 2025, [Online]. Available: https://doi.org/10.1016/j.dcan.2025.01.005.

[8] Y. Edlitz and E. Segal, "Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards," eLife, vol. 11, p. e71862, 2022.

[9] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Informatics in Medicine Unlocked, vol. 17, p. 100179, 2019.

[10] J. Lu, S. Lu, Y. Zhao, L. Yang, W. C. Chan, J. Lian, and D. H. Shum, "An electronic health record-linked machine learning tool for diabetes risk assessment in adults with prediabetes," The Innovation Medicine, vol. 3, no. 1, 2025.

[11] S. Afolabi, N. Ajadi, A. Jimoh, and I. Adenekan, "Predicting diabetes using supervised machine learning algorithms on e-health records," Informatics and Health, vol. 2, no. 1, pp. 9-16, 2025.

[12] F. Mohsen, H. R. Al-Absi, N. A. Yousri, N. El Hajj, and Z. Shah, "A scoping review of artificial intelligence-based methods for diabetes risk prediction," npj Digital Medicine, vol. 6, no. 1, p. 197, 2023.

[13] R. Sharma, P. Gupta, and A. Singh, "Human-computer interaction frameworks for secure digital adoption," International Journal of Human-Computer Interaction, vol. 41, no. 7, pp. 845-862, 2025, [Online]. Available: https://doi.org/10.1080/10447318.2025.2495843.

[14] A. Barwise and D. Tschida-Reuter and B. Sutor, "Adaptations to interpreter services for hospitalized patients during the COVID-19 pandemic," in Mayo Clinic Proceedings, vol. 96, no. 12, p. 3184, Oct. 2021.