Predicting Depression in Canadians with or at Risk of Diabetes: A Cross-Sectional Machine Learning Analysis

Konrad Samsel[1,*], Mohammad Noaeen[1,2,*], Amrit Tiwana[1,*], Sarra Ali[1], Aziz Guergachi[2,3], Karim Keshavjee[2], and Zahra Shakeri[2,4,5]

***Abstract*— Depression often goes unrecognized in individuals at risk or living with diabetes, presenting considerable challenges for primary care clinicians. Although large language models and other foundation model approaches are drawing significant attention, we systematically compared six established machine learning algorithms-Logistic Regression, Random Forest, AdaBoost, XGBoost, Naive Bayes, and Artificial Neural Networks-chosen for their reliability, interpretability, and feasibility in everyday clinical settings. By benchmarking their performance under real-world constraints, we identified key factors linked to depression risk in diabetes care, including patient sex, age, osteoarthritis, hemoglobin A1c, and body mass index. Although incomplete demographic information and potential label bias limited predictive power, our results demonstrate that a diverse set of clinical features might help pinpoint high-risk patients. They also indicate a need for longitudinal follow-up and richer clinical data to enhance model accuracy. As a practical benchmark for both clinicians and data scientists, this work suggests that machine learning–based risk stratification can improve early detection of depression and inform targeted interventions in diabetic populations.**

I. INTRODUCTION

Depression is a pervasive and often underrecognized comorbidity among individuals at risk for or living with diabetes mellitus. The bidirectional relationship between depression and diabetes not only complicates metabolic management but also significantly worsens patient outcomes-from deteriorating glycemic control to an overall decline in quality of life [1–4]. Recent studies indicate that patients with diabetes are two to three times more likely to experience depression compared to the general population [5, 6]. This concerning statistic, along with the economic and emotional burden on healthcare systems [7], calls for the need for effective screening and early intervention strategies.

In the current field of advanced artificial intelligence and generative models [8], much attention is given to increasingly complex architectures. However, clinical reality often demands methods that are interpretable, robust, and easily implementable within the constraints of routine primary care data [9–12]. While large-scale deep learning models have indeed pushed the boundaries of pattern recognition, their complexity is not always an asset in heterogeneous, real-world electronic health records (EHRs). Our work deliberately refrains from the pursuit of algorithmic novelty; instead, it focuses on using a diverse suite of established machine learning techniques as a benchmarking tool to reveal clinically actionable insights. Specifically, our study leverages routinely collected clinical data to predict depression risk among Canadians with diabetes or prediabetes. We systematically compare six supervised learning algorithms–including Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), AdaBoost (AB), XGBoost (XGB), and an Artificial Neural Network (ANN)–to identify which clinical and demographic factors best indicate depression risk. Our feature set includes patient sex, age, body mass index (BMI), hemoglobin A1c, and comorbidities such as osteoarthritis and hypertension [13, 14], all chosen for their clinical relevance.

To align technical performance with clinical applicability, we employ SHapley Additive exPlanations (SHAP) [15] to deconstruct each prediction into contributions from key features. This enables us to decompose the risk predictions into meaningful contributions from each feature, thereby translating our technical findings into insights that can inform targeted, patient-centered interventions. Importantly, while the absolute predictive metrics (e.g., AUC or F1 score) achieved by our models are modest–reflecting inherent challenges such as incomplete demographic data and potential label bias [16, 17]–they serve as a transparent benchmark for future research. Therefore, this study demonstrates that traditional and well–established models can effectively elucidate the key determinants of depression risk in diabetic populations when applied to rigorously curated EHR data. This evidence reinforces the value of established methodologies in bridging the gap between technical rigor and clinical applicability, fostering an interdisciplinary dialogue between data scientists and clinicians. The subsequent sections describe our data collection and preprocessing strategy, model development, and interpretability analyses, thereby delineating a framework for more nuanced, equitable, and actionable mental health screening in diabetes care.

II. METHODS

A. Data Collection and Preparation

We obtained a comprehensive dataset ($N = 8{,}602$) from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), a national repository of de-identified electronic

*These authors contributed equally to this work and share first authorship.

[1]Konrad Samsel, Mohammad Noaeen, Amrit Tiwana, and Sarra Ali are with the Dalla Lana School of Public Health, University of Toronto, Canada.

[2]Mohammad Noaeen, Aziz Guergachi, Karim Keshavjee and Zahra Shakeri are with the Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Canada.

[3]Aziz Guergachi is with Ted Rogers School of Information Technology Management, Toronto Metropolitan University, Toronto, Canada; and Department of Mathematics and Statistics, York University, Toronto, Canada.

[4,5]Zahra Shakeri is with the Faculty of Information and the Schwartz Reisman Institute, University of Toronto, Canada.

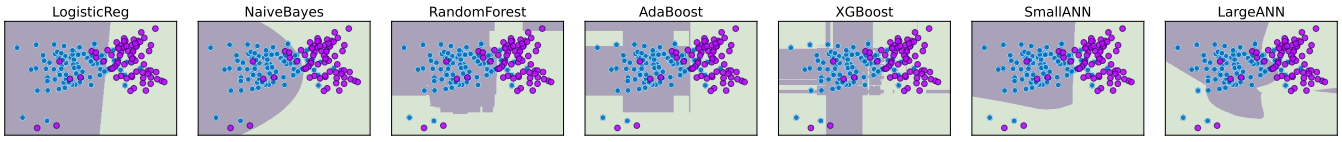

Fig. 1: A synthetic dataset illustrating typical complexities of clinical data: correlated features, overlapping subpopulations, outliers, and label noise. Each subplot shows a learned decision boundary $f : \mathbb{R}^2 \to \{0, 1\}$ for a different model (Logistic Regression, Naive Bayes, Random Forest, AdaBoost, XGBoost, a smaller ANN, and a larger ANN). Simpler or ensemble methods yield stable partitions under moderate sample sizes, whereas the larger ANN overfits to minor clusters and mislabeled points. Axis labels are omitted to highlight differences in boundary shapes.

medical records spanning 2003 to 2015 [10]. This cross-sectional dataset comprises 34 features capturing essential demographic attributes (e.g., age, sex), anthropometric measurements (e.g., body mass index), biomarkers (e.g., hemoglobin A1c), comorbidity diagnoses, and prescription records. To ensure clinical relevance, the cohort was restricted to patients with either a confirmed diagnosis of diabetes mellitus (type 1 or type 2) [14] or an elevated hemoglobin A1c ($\geq 5.7\%$) indicative of prediabetes [18], yielding a final analytic sample of 6,219 individuals. Depression labels were derived from clinician-documented diagnoses, reflecting routine clinical practice.

We evaluated data quality and completeness, noting that fewer than 3% of observations had missing values. Given the minimal scope of missingness, multiple imputation by chained equations (MICE) was applied to the training portion of the data [19]. All continuous variables were normalized to mean zero and unit variance to stabilize subsequent model fitting. An 80:20 train-test split was used, and the training subset exhibited class imbalance (approximately 20% with depression). To manage this, we first removed Tomek links, where a pair $(\mathbf{x}_i, \mathbf{x}_j)$ from opposite classes satisfies each being the other's nearest neighbor in feature space. Removing the majority class instance in each Tomek pair eliminates borderline points that can confuse the model's decision boundary. Next, we employed SMOTE, which synthesizes minority samples using linear interpolations: $\tilde{\mathbf{x}} = \mathbf{x} + \lambda\big(\mathbf{x}^{(k)} - \mathbf{x}\big)$, where $\mathbf{x}^{(k)}$ is a randomly chosen minority class neighbor of $\mathbf{x}$, and $\lambda \in [0, 1]$. This two-step balance strategy addresses the challenge in clinical data where differences between depressed and non-depressed patients can be obscured by the larger control group [20].

New variables were derived to reflect chronic disease burden, which is clinically salient for mental health outcomes. For each documented comorbidity (e.g., osteoarthritis, hypertension), we computed a duration metric by subtracting the date of diagnosis from the most recent clinical encounter, thus representing the approximate years lived with that condition. Undiagnosed cases were set to zero. Medication records were parsed into 45 binary indicators, excluding categories with fewer than ten occurrences to avoid sparse features. These feature engineering steps were motivated by evidence that longer disease duration, multiple comorbidities, and certain prescriptions may compound the risk of depression in patients with metabolic disorders.

B. Model Development

Recognizing the inherent constraints of primary care datasets, such as modest sample sizes and heterogeneous data structures, our approach prioritizes models that strike a balance between interpretability and robust performance. While recent advances in deep learning and generative modeling have expanded the analytical toolkit, the practical limitations of routine EHR data often favor well-established, transparent methods. Accordingly, we conducted a systematic comparison of six supervised learning algorithms: LR, NB, RF, AB, XGB, and an ANN. Each algorithm was chosen based on its unique trade-off between interpretability, computational complexity, and capacity for capturing non-linear relationships. Formally, given N patients $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, with $\mathbf{x}_i \in \mathbb{R}^d$ representing a vector of demographics, comorbidity durations, and medication indicators, and $y_i \in \{0, 1\}$ indicating a depression diagnosis, these methods learn a mapping $f : \mathbb{R}^d \to \{0, 1\}$ that generalizes to new individuals.

LR provides a straightforward linear boundary whose coefficients can serve as intuitive risk indicators, while NB uses feature-independence assumptions to yield a generative view of $P(y \mid \mathbf{x})$. RF, AB, and XGB build ensembles of decision trees, enabling them to adaptively capture interactions among predictors such as age, BMI, and disease history. The ANN leverages a feedforward, multilayer structure to approximate non-linear functions, which can be helpful in capturing more complex relationships. However, as illustrated in Figure 1, larger neural networks may become over-sensitized to noise and outliers when real-world data are limited or contain label inaccuracies. Our intent was to cover a range of baseline architectures suitable for typical EHR scenarios, where advanced deep or generative models often require bigger, more uniform datasets to avoid overfitting or opaque decision boundaries. All source code for data preprocessing, model training, and interpretability analyses is publicly available[1], facilitating replication and further refinement by both clinicians and data scientists.

III. Results and Discussion

Table I presents demographic and clinical attributes for 6,219 individuals who met the criteria for diabetes (62.7%) or prediabetes (37.3%). Of these, 20.7% carried a clinician-diagnosed depression, aligning with previous reports on the elevated comorbidity in metabolic conditions [5]. Notably, 50.4% were female, with hypertension emerging as the most common additional diagnosis (69.2%). To balance interpretability with completeness, we focused on hemoglobin A1c as the primary glycemic marker, given its correlation with fasting glucose and modest missingness (3%). Such choices underscore the routine data limitations in EHRs,

[1] https://github.com/tiwanaam/mlforhealthdata

TABLE I: Patient characteristics are presented as mean [sd] for continuous variables or n (%) for categorical variables.

Variable	Total (N = 6,219)
Age, years	64.6 [12.4]
BMI, kg/m^2	30.9 [6.9]
Sex, female	4,110 (52.3)
Hemoglobin A1c, %	6.5 [0.9]
Fasting blood sugar level, mmol/L	6.6 [1.8]
Total cholesterol, mmol/L	4.6 [1.2]
Missing	149 (1.9)
Depression (Yes)	1620 (20.6)
Hypertension (Yes)	5,439 (69.2)
Years living with hypertension	2.1 [2.5]
Osteoarthritis (Yes)	2,628 (33.4)
Years living with osteoarthritis	2.1 [2.5]
COPD (Yes)	828 (10.5)
Years living with COPD	1.8 [2.0]
Diabetes (Yes)	5,139 (65.4)
Years living with diabetes	2.8 [2.5]
Take at least one hypertension medication	6,042 (76.9)
Take at least one corticosteroid medication	2,298 (29.2)

where high correlation and partial data unavailability frequently shape feature selection [17].

A. Model Benchmarking and Interpretability

Among the six models evaluated, XGBoost emerged as the top performer, achieving an AUC of 0.64 and a weighted average F1 score of 0.72 on the held-out test set. Other methods demonstrated AUCs ranging from 0.54 to 0.64. Although these performance metrics may appear modest relative to more complex deep learning architectures, they are in line with similar studies using routine EHR data for depression screening, where challenges such as imprecise labeling and the absence of detailed psychosocial measures are common [17]. The high precision in identifying non-depressed cases, alongside lower recall for the depressed minority, reflects the intrinsic class imbalance (20.6% depression rate) and the fact that some depressive conditions may remain undocumented or embedded within unstructured clinical notes. Despite these constraints, the consistent detection of several core risk factors, such as BMI, hemoglobin A1c, osteoarthritis, and younger age, suggests that even traditional machine learning algorithms can highlight clinically meaningful patterns in cross-sectional EHR data. This is important since purely deep or generative models may not always yield decisive gains when exposed to the sparse, heterogeneous records found in primary care [21]. Moreover, these reproducible but modest findings provide a transparent benchmark for researchers seeking to refine depression detection methods, underscoring two key points:

Label Reliability and Missing Psychosocial Variables. Depression is frequently underdiagnosed or documented in free-text notes rather than structured fields, suggesting that the actual prevalence may exceed the nominal 20.6%. The absence of standardized patient-reported outcomes or validated screening instruments (e.g., PHQ-9) further limits the model's sensitivity and specificity [22]. These factors highlight the need for enhanced data collection protocols to improve the reliability of automated depression detection.

Cross-Sectional Data and Disease Trajectory. The cross-sectional design of our study precludes tracking temporal changes in glycemic control, comorbidities, or mental health status. Prior research indicates that fluctuations in metabolic

Model	Prec.	Recall	F1 Score	AUC
LR	0.76	0.62	0.65	0.69
NB	0.76	0.65	0.68	0.64
RF	0.75	0.69	0.71	0.69
AB	0.75	0.70	0.72	0.68
XGB	**0.75**	**0.71**	**0.73**	**0.70**
ANN	0.74	0.51	0.54	0.68

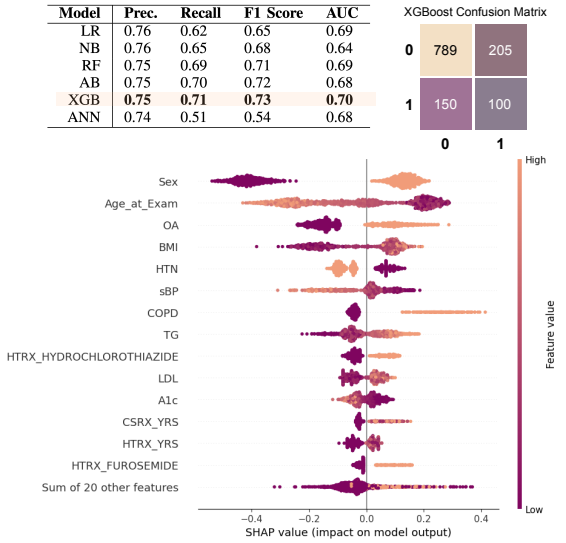

Fig. 2: Performance metrics for six classifiers (left), XGBoost confusion matrix (top right), and SHAP summary plot (bottom). The matrix shows correctly (TN, TP) and incorrectly (FN, FP) classified patients. In the SHAP plot, each dot's horizontal position indicates how much that feature pushes the model toward (positive) or away from (negative) a depression prediction, with purple denoting lower feature values and orange denoting higher values.

parameters may precipitate or exacerbate depressive symptoms [23]. Incorporating longitudinal data could provide a more accurate representation of depression onset and progression, potentially improving predictive performance beyond the intermediate AUC values (around 0.64) observed in our study [21].

B. Feature Importance and Clinical Implications of XGBoost

Figure 2 highlights two essential aspects of the XGBoost model's behavior on the test set. The upper panel confusion matrix (1,244 test instances) shows 795 true negatives (non-depressed) and 96 true positives (depressed), with relatively few false positives (198) corresponding to a precision of 0.33. However, a recall of 0.38 indicates that 150 depressed patients were missed (false negatives). From a clinical standpoint, this balance may be acceptable as an initial risk stratification tool, given the well-documented challenges of accurately capturing mental health conditions in EHRs.

The lower panel is a SHAP summary plot, which explains feature-level contributions to the model's predicted log-odds of depression. Specifically, for each instance $\mathbf{x}_i$ and feature j, the SHAP value $\phi_j(\mathbf{x}_i)$ quantifies how much x_{ij} shifts the model's output relative to a baseline: $\text{logit}\big(f(\mathbf{x}_i)\big) \approx \phi_0 + \sum_{j=1}^{d} \phi_j(\mathbf{x}_i)$, where ϕ_0 is the average prediction on the training set. Positive SHAP values push the prediction toward 'depressed', negative values do the opposite, and the color scale (purple to orange) reflects each patient's actual feature value. 'Sex' (female) and 'Age_at_Exam' (younger) emerged as influential, suggesting that female sex and earlier onset of metabolic conditions may elevate psychosocial burdens [24, 25]. Osteoarthritis ('OA') and BMI also significantly increased depression risk, likely reflecting pain, reduced mobility, and social constraints. Systolic blood pressure ('sBP') and antihypertensives (e.g., 'HTRX_HYDROCHLOROTHIAZIDE') further emphasize the link between cardiovascular strain

and mood disturbances. Although hemoglobin A1c did not dominate on its own, it contributed synergistically with factors like high BMI, reinforcing that depression risk in diabetic and prediabetic populations arises from multifaceted metabolic profiles rather than glycemic control alone.

Despite these clinically coherent findings, the 155 false negatives emphasize a need for improved data capture, through standardized mental health tools or more frequent longitudinal assessments, to enhance recall. Nonetheless, the SHAP interpretation demonstrates how a well-calibrated machine learning model can provide nuanced, patient-level insights into depression risk, thereby informing targeted interventions in real-world primary care settings.

C. Limitations

Although this study leveraged real-world CPCSSN data to demonstrate the feasibility of depression risk modeling in a diverse adult population [10], several limitations remain. The cross-sectional design precludes capturing temporal changes or identifying whether evolving glycemic or comorbidity profiles trigger depressive symptoms. Reliance on clinician-documented diagnoses introduces potential label bias, as subclinical or undiagnosed cases may be overlooked; integrating standardized tools (e.g., PHQ-9) could improve label reliability. Finally, sparse demographic data restricted subgroup analysis, limiting insights into racial or socioeconomic disparities. Despite these constraints, the findings provide a valuable benchmark for future efforts to refine and expand depression detection in patients with or at risk of diabetes.

IV. CONCLUSION

This study established a foundational benchmark for depression detection in individuals with or at risk of diabetes using routinely collected electronic health records. While XGBoost demonstrated the strongest performance (AUC $\approx$ 0.64), our SHAP analyses revealed that non-engineered attributes such as sex, age, osteoarthritis status, A1c, and BMI drove much of the predictive signal. These findings highlight that even standard EHR fields can flag mental health vulnerabilities, yet they also reflect the limited scope of cross-sectional data and potential label bias. Integrating longitudinal observations, richer demographic variables, and validated screening tools would likely enhance both sensitivity and specificity. As machine learning gains traction in clinical workflows, balancing model interpretability, robust data collection, and domain-guided feature selection remains paramount to realizing effective, equitable mental health support for patients facing metabolic risks.

REFERENCES

[1] J. S. Gonzalez, M. Peyrot, L. A. McCarl, *et al.*, "Depression and diabetes treatment nonadherence: A meta-analysis," *Diabetes Care*, vol. 31, no. 12, 2398–2403, 2008.

[2] L. E. Egede, P. J. Nietert, and D. Zheng, "Depression and all-cause and coronary heart disease mortality among adults with and without diabetes," *Diabetes Care*, vol. 28, no. 6, 1339–1345, 2005.

[3] R. I. G. Holt and W. J. Katon, "Dialogue on diabetes and depression: Dealing with the double burden of co-morbidity," *J Affect Disord*, vol. 142, no. Suppl, S1–3, 2012.

[4] R. I. G. Holt, M. de Groot, and S. H. Golden, "Diabetes and depression," *Curr Diab Rep*, vol. 14, no. 6, p. 491, 2014.

[5] S. Bădescu *et al.*, "The association between diabetes mellitus and depression," *Journal of medicine and life*, vol. 9, no. 2, p. 120, 2016.

[6] J.-Y. Lee, D. Won, and K. Lee, "Machine learning-based identification and related features of depression in patients with diabetes mellitus based on the korea national health and nutrition examination survey: A cross-sectional study," *PLoS One*, vol. 18, no. 7, e0288648, 2023.

[7] M. Khaledi, F. Haghighatdoost, A. Feizi, *et al.*, "The prevalence of comorbid depression in patients with type 2 diabetes: An updated systematic review and meta-analysis on a huge number of observational studies," *Acta Diabetol*, vol. 56, no. 6, 631–650, 2019.

[8] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[9] K. Lu *et al.*, "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, pp. 1–4.

[10] S. Garies, R. Birtwhistle, N. Drummond, *et al.*, "Data resource profile: National electronic medical record data from the canadian primary care sentinel surveillance network (cpcssn)," *Int J Epidemiol*, vol. 46, no. 4, 1091–1092f, 2017.

[11] K. Esser *et al.*, "Predicting diabetes in canadian adults using machine learning," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–4.

[12] P. Saha *et al.*, "Predicting time to diabetes diagnosis using random survival forests," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–4.

[13] M. Virtanen, J. E. Ferrie, A. G. Tabak, *et al.*, "Psychological distress and incidence of type 2 diabetes in high-risk and low-risk populations: The whitehall ii cohort study," *Diabetes Care*, vol. 37, no. 8, 2091–2097, 2014.

[14] T. Williamson, M. E. Green, R. Birtwhistle, *et al.*, "Validating the 8 cpcssn case definitions for chronic disease surveillance in a primary care database of electronic health records," *Ann Fam Med*, vol. 12, no. 4, 367–372, 2014.

[15] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, vol. 30, 2017.

[16] Centers for Disease Control and Prevention. "Diabetes and mental health — cdc." Accessed December 8, 2023. (2023), [Online]. Available: `https://tinyurl.com/y9ncfkn2`.

[17] N. R. Winter *et al.*, "A systematic evaluation of machine learning–based biomarkers for major depressive disorder," *JAMA psychiatry*, vol. 81, no. 4, pp. 386–395, 2024.

[18] C. Lorenzo, L. E. Wagenknecht, A. J. G. Hanley, *et al.*, "A1c between 5.7 and 6.4 percent as a marker for identifying pre-diabetes, insulin sensitivity and secretion, and cardiovascular risk factors: The insulin resistance atherosclerosis study (iras)," *Diabetes Care*, vol. 33, no. 9, 2104–2109, 2010.

[19] S. V. Buuren and K. Groothuis-Oudshoorn, "Multivariate imputation by chained equations," *J. Stat. Soft. [electronic article]*, vol. 45, no. 3, 2011.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *MACHINE LEARNING IN PYTHON*, 2011.

[21] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[22] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: Validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[23] B. Mezuk, W. W. Eaton, S. Albrecht, and S. H. Golden, "Depression and type 2 diabetes over the lifespan: A meta-analysis," *Diabetes care*, vol. 31, no. 12, pp. 2383–2390, 2008.

[24] M. M. Esteban y Peña, V. Hernandez Barrera, X. Fernández Cordero, *et al.*, "Self-perception of health status, mental health and quality of life among adults with diabetes residing in a metropolitan area," *Diabetes Metab*, vol. 36, no. 4, 305–311, 2010.

[25] M. Sundaram, J. Kavookjian, J. H. Patrick, *et al.*, "Quality of life, health status and clinical outcomes in type 2 diabetes patients," *Qual Life Res*, vol. 16, no. 2, 165–177, 2007.