Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial Intelligence for Treatment-Response Prediction in GLP-1 Therapies: A Systematic Review and Translational Readiness Framework (AI-TRF)

Tahsien Al-Quraishi
Melbourne Institute of Technology https://orcid.org/0000-0003-3814-0431
O. S. Albahri
o.albahrey@federation.edu.au
Federation University
Jemal Abawajy
Deakin University
Raza Nowrozy
Victoria University https://orcid.org/0000-0002-9886-0978

Systematic Review

Keywords: GLP-1 receptor agonists, treatment-response prediction, machine learning, obesity management, type 2 diabetes, clinical decision support

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11763328/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Artificial Intelligence for Treatment-Response Prediction in GLP-1 Therapies: A Systematic Review and Translational Readiness Framework (AI-TRF)

Tahsien Al-Quraishi[1,2], O. S. Albahri[3*], Jemal Abawajy[4], Raza Nowrozy[5]

[1]School of Business, Melbourne Institute of Technology, La Trobe St, Melbourne, Victoria, Australia.
[2]College of Science, Al-Ayen Iraqi University, Nasiriyah, Iraq.
[3*]Institute of Innovation, Science and Sustainability, Federation University, Melbourne, Victoria, Australia.
[4]Faculty of Science, Engineering and Built Environment, Deakin University, Melbourne, Victoria, Australia.
[5]College of Art, Business, Law, Education & IT (CoABLEIT), Victoria University, Melbourne, Victoria, Australia.

*Corresponding author(s). E-mail(s): o.albahrey@federation.edu.au;
Contributing authors: tquraishi@mit.edu.au;
jemal.abawajy@deakin.edu.au; rnowrozy@mit.edu.au;

Abstract

Glucagon-like receptor agonists and incretin-based therapies improve glycaemic control and weight-related outcomes, but individual treatment responses vary. This review examines treatment-response prediction and the clinical evidence relevant to personalised treatment selection. The current evidence table contains 44 records identified through searches in PubMed, Scopus, and Web of Science. These records comprise a mixture of prediction and predictor analyses, clinical efficacy studies, and other evidence of response to treatment; they should not be interpreted as 44 independently validated AI prediction systems. The synthesis describes therapeutic targets, data sources, and analytical approaches. However, incomplete study-level extraction and unresolved classification issues

prevent reliable estimates of prediction-model validation practices, explainability uptake, or comparative algorithm performance. We also propose the Artificial Intelligence Translational Readiness Framework (AI-TRF), comprising external validation, explainability, temporal modelling, prospective evaluation, and clinical implementation. AI-TRF has not been applied to individual studies or empirically validated. Reproducible searches, verified study-level extraction, and independent evaluation remain necessary to establish the clinical readiness of treatment-response prediction systems in this field.

Keywords: GLP-1 receptor agonists, treatment-response prediction, machine learning, obesity management, type 2 diabetes, clinical decision support

1 Introduction

Obesity and type 2 diabetes mellitus (T2DM) are a public health challenge worldwide and one of the most important risk factors for cardiovascular disease, chronic kidney disease, hypertension, metabolic dysfunction and poor quality of life [1, 2]. The growing incidence of such chronic metabolic conditions has placed a significant clinical and economic strain on healthcare systems around the world and requires more effective and efficient, individualised, and sustainable approaches to their management [3, 4]. Glucagon-like receptor agonists (GLP-1 RA) and incretin-based therapies have become a major paradigm shift in the field of obesity medicine and diabetes management in recent years [5, 6]. Agents including *Semaglutide*, *Liraglutide*, *Dulaglutide* and the dual GIP/GLP-1 receptor agonist *tirzepatide* have shown a significant positive effect on body weight loss, glycaemic control, HbA1c lowering, cardiovascular risk reduction and long-term metabolic outcomes [7]. They have gained a lot of traction in endocrinology clinics, obesity treatment centres, and primary care settings due to robust evidence from large clinical trials and real-world studies [8].

Nevertheless, these therapeutic developments and outcomes are still highly varied between patient populations [9–11]. Some seem to lose substantial weight and maintain lost weight, gain control over glycaemic parameters, and remain on the drug; others only receive a modest therapeutic effect, suffer from gastrointestinal intolerance, discontinue using the drug, or do not obtain a sufficient improvement of their cardiometabolic parameters [12]. This represents a major clinical challenge in the current approach to obesity and diabetes care, since it is based on what can be described as escalation and empirical prescribing approaches rather than targeted stratification. Computational prediction methods are increasingly being used to guide personalised treatment strategies for GLP-1 receptor agonists and incretin-based therapies [9–11]. Recent real-world machine learning studies also showed that predictive modelling using longitudinal electronic health record data can help identify those who are likely to respond to GLP-1 pathway therapies in more heterogeneous groups of patients with obesity and type 2 diabetes [13, 14]. Several studies investigated pharmacogenomic variability in the context of genotype-driven stratification schemes to better estimate the response to different drugs and maximise incretin therapy allocation [15, 16].

Within GLP-1 therapies, pharmacogenomic factors are emerging as key contributors to the glycaemic response, the heterogeneity of weight loss and the persistence of treatment [17].

Predictive models can help identify those who are a good fit for treatment, predict non-treatment responses, assess adherence to treatment, predict negative results and choose between semaglutide and tirzepatide [18, 19]. In addition, XAI methods are increasingly being investigated with the purpose of making the model more transparent and interpretable in clinical decision support settings. By contrast, although there has been a surge in data-driven research in metabolic medicine, the literature is methodologically scattered, and there are significant differences between the data sets used, the validation techniques used, the modelling approaches, the explainability methods, and the maturity of clinical implementation [9, 10]. A big hurdle in the prediction of the response to treatment is the longitudinal variability of the outcome of treatment. Metabolic adaptation, adverse events, an increase in treatment, and behaviour of adherence are all factors that affect the therapeutic effects of GLP-1 over time [20]. These temporal features make the performance of traditional static machine learning models learned from the historical dataset suboptimal. In addition, the structures of the EHRs, follow-up periods, and treatment protocols vary, which also hinders the generalisability and reproducibility of the prediction systems from one healthcare setting to another.

Hence, a limited systematic synthesis, specifically focusing on clinically integrated prediction systems [21], remains to respond and not respond to GLP-1 therapies. The methodological designs, data sources, prediction goals, and assessment methods used in existing studies differ widely and make it difficult to directly compare them and translate them into the assessment. Therefore, a systematic review is needed to critically evaluate current clinical prediction trends, methodological limitations, explainability practices, validation strategies, and the translational readiness of predictive models for clinical implementation. This review provides a comprehensive assessment of clinically integrated prediction systems developed for GLP-1 receptor agonist and incretin-based therapies. Particular attention is given to commonly used modelling techniques, key predictors of treatment success and non-response, explainability approaches, validation practices, and barriers to clinical deployment. The review also identifies important research gaps and future opportunities for the development of AI-enabled treatment-planning and therapy-selection systems for the management of obesity and type 2 diabetes.

1.1 Research Gaps and Contributions of This Review

Table 1 provides an overview of recent systematic reviews focusing on the use and implementation of AI, ML and predictive analytics for diabetes prediction, obesity prediction, and diabetes-related clinical decision-support systems. Although these reviews offer valuable perspectives on predictive approaches and healthcare applications, they focus primarily on predictive methods used for the diagnosis of diseases, risk prediction, and/or complication forecasting but do not focus on the use of treatment-response modelling for GLP-1 receptor agonists and incretin-based therapies. De Silva et al. [22] provided a review of ML models for the prediction of type

2 diabetes in a community healthcare setting and found promising predictive performance in various algorithms. However, the review was based on disease prediction and acknowledged issues in external validation and clinical implementation. In that same vein, Fregoso-Aparicio et al. studied ML and DL methods to predict T2D, including predictive models, validation strategies, feature engineering methods, and the size of the datasets. Although the review provided a comprehensive overview of predictive methodologies, it did not focus on treatment-response prediction or treatment optimisation for GLP-1-based interventions.

Although these reviews [22–31] provide information on diabetes prediction, obesity prediction, and AI-enabled clinical decision-support systems, the main focus is on modelling disease diagnosis, risk stratification, and clinical complication prediction, but not the treatment-response modelling of GLP-1 receptor agonists and incretin-based therapies. Limited synthesis of existing reviews is provided regarding the use of computational approaches for predicting response and non-treatment response, treatment stratification guided by pharmacogenomic data, adherence modelling, treatment discontinuation forecasting, and selection of personalised therapy. [9–11, 32, 33]. In addition, external validation, calibration reliability, temporal robustness, explainability, and translational readiness of treatment-response prediction systems have received little attention. Thus, key considerations remain about whether these predictive models can be generalised, interpreted, and applied clinically in other diverse healthcare settings and different cohorts of patients [9, 10].

The present review, instead of subjects, specifically concerns adaptive prediction systems for patient-level therapeutic benefit, non-response detection, adherence prediction, discontinuation analysis, and therapy optimisation in GLP-1 receptor agonist and incretin-based therapies. Specific focus is placed on clinically relevant agents (semaglutide, tirzepatide, liraglutide, and dulaglutide) and outcome parameters (weight-loss response, improvement of HbA1c, persistence of treatment, prediction of adverse events, and long-term sustainability of treatment). This review, in addition to evidence synthesis, proposes the Artificial Intelligence Translational Readiness Framework (AI-TRF), a conceptual framework to characterise the translational maturity and clinical applicability of predictive models for GLP-1 receptor agonist therapies. However, unlike existing tools such as PROBAST, PROBAST-AI, and TRIPOD-AI, which focus on risk-of-bias assessment or reporting quality, AI-TRF offers added value in the evolution of predictive models to clinically interpretable, externally validated, and implementation-orientated predictive model-based decision-support systems [34–36].

The framework has resulted from a recurring pattern of methodology categorised in included studies, including integration of explainability, validation robustness, temporal modelling capacity, and clinical implementation readiness. AI-TRF is designed to enhance current reporting, quality assessment, and risk of bias frameworks and not supplant them [34]. Conceptually and as a hypothesis-generating tool, it aims to enable evaluation and discussion of the 'translational maturity' of predictive modelling studies. Future research should include formal assessment of the framework by expert consensus, inter-rater reliability, and a variety of clinical AI applications to

Table 1 Comparison of existing systematic reviews and the contributions of this study

Ref	Study	Year	Focus	Main contribution	Limitation relative to this study
[1]	De Silva et al.[22]	2020	T2DM prediction	Evaluated ML models in community healthcare settings	Focused on disease-risk prediction rather than GLP-1 treatment-response modelling.
[2]	Fregoso-Aparicio et al.[23]	2021	ML/DL for T2DM	Compared models, datasets, validation metrics, and modelling approaches	Focused on disease prediction rather than therapeutic optimisation.
[3]	Tuppad and Patil[24]	2022	Clinical decision support	Identified clinical and technical implementation gaps	Limited coverage of GLP-1 therapeutic prediction.
[4]	Olusanya et al.[25]	2022	T2DM classification	Meta-analysis of machine-learning classification models	Focused on disease prediction rather than treatment outcomes.
[5]	Ferreras et al.[26]	2023	Obesity prediction	Reviewed machine-learning methods for obesity prediction	Focused on obesity-risk prediction rather than pharmacological response and treatment-response modelling.
[6]	Petridis et al.[27]	2024	Diabetes prediction	Reviewed recent machine-learning approaches for diabetes prediction and management	General prediction focus without GLP-1 treatment-response modelling.
[7]	Khokhar et al.[28]	2024	AI for diabetes	Analysed datasets, algorithms, and emerging artificial intelligence techniques for diabetes prediction	Did not address GLP-1 therapies, treatment-response prediction, or personalised treatment optimisation.
[8]	Haghighathoseini et al.[29]	2025	AI for obesity prevention	Reviewed artificial intelligence applications for obesity prevention and prediction	Focused on obesity-risk prediction and prevention rather than therapeutic-response modelling.
[9]	Li et al.[30]	2025	Diabetic kidney disease prediction	Meta-analysis of machine-learning models for diabetic kidney disease prediction	Focused on complication prediction rather than treatment-response optimisation.
[10]	Masood et al.[31]	2026	Data-driven diabetes	Reviewed machine learning, deep learning, multimodal data integration, and explainable AI for diabetes applications	Focused primarily on diagnosis, risk prediction, and disease management rather than GLP-1 treatment-response modelling.
[11]	**This study**	**2026**	**GLP-1 response modelling**	**Reviews treatment response, adherence, discontinuation, and adverse events associated with GLP-1 therapies.**	**Integrates predictive modelling, explainable AI, and the AI Translational Readiness Framework to support AI-assisted therapeutic decision-making.**

validate its reliability, reproducibility, and generalisability. A framework is introduced and conceptually shown in Figure 1.

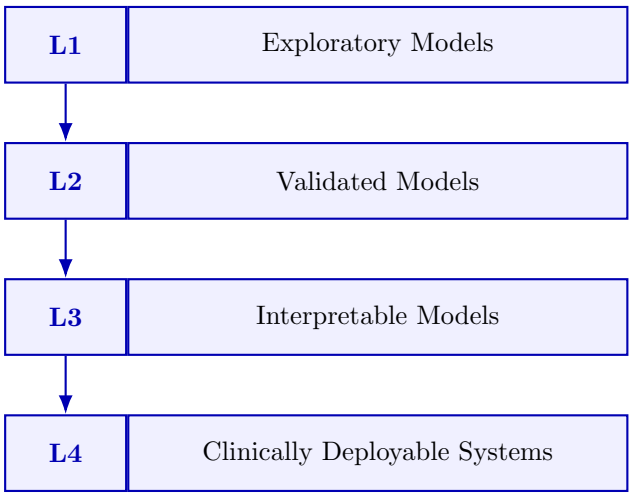


Fig. 1 Artificial Intelligence Translational Readiness Framework (AI-TRF)
Note: L = Translational readiness level

The majority of the assessed studies were in the exploratory and validated phases of the proposed Artificial Intelligence Translational Readiness Framework (AI-TRF), and only a few exhibited features of interpretable or clinically deployable systems. This discovery reflects the ongoing issue that exists between the development of predictive models and regular clinical use.

2 Methods

A systematic review approach is used to examine the use of AI, ML, DL and predictive analytics in the field of receptor agonists similar to glucagon-1 peptide-1 (GLP-1 RA) and incretin-based therapies. The review has been carried out according to the Guidelines for Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) to maintain transparency, methodological rigour, and reproducibility in the process [34, 37]. The process involved formulating research questions, searching databases, selecting studies, assessing quality, extracting data, and evaluating bias. Special attention was given to studies examining the possibility of prediction of treatment response, the detection of non-response, treatment persistence and discontinuation modelling, explainable decision-support systems and personalised treatment strategies related to GLP-1 therapies.

2.1 Review Scope

This systematic review is intended to identify, analyse, and critically evaluate the use of AI, ML, DL, and predictive analytics in GLP-1 RA and incretin-based therapy.

The main objectives of the review are prediction of treatment response, non-response detection, adherence estimation, treatment-discontinuation analysis, adverse-event prediction, and optimisation of therapy in obesity and type 2 diabetes management. Specific focus is placed on clinically important outcomes such as weight-loss response, glycaemic improvement (HbA1c reduction), persistence of treatment, sustainability of adherence, cardiometabolic outcomes, and occurrence of adverse events. In addition to predictive performance, the review covers modelling methodologies, approaches for explainability, validation strategies, characteristics of the datasets used, and clinical translation readiness. It also outlines the methodological gaps, research gaps, and translation challenges that affect the design of precision systems to prescribe GLP-1 therapy for clinical deployment. This scope allows the study to approach the synthesis of existing evidence for personalised treatment planning in obesity and diabetes healthcare in a structured way.

2.2 Research Questions

Research questions were formulated to inform the systematic review and facilitate a systematic assessment of computational analytics systems used in glucagon-like peptide-1 receptor agonist (GLP-1 RA) and incretin-based therapies:

1. **Research Question 1 (RQ1):** What types of data set, clinical variables, and patient characteristics are commonly used to develop prediction systems for metabolic improvement and non-response in GLP-1 receptor agonist and incretin-based therapies?
2. **Research Question 2 (RQ2):** Which ML and DL approaches are applied most frequently, and which clinical decision-support systems demonstrate the strongest performance for the effectiveness, adherence, persistence, and prediction of discontinuation of therapy?
3. **Research Question 3 (RQ3):** What validation strategies, evaluation metrics, and explainable decision-making approaches are used to assess the reliability, interpretability, and external validity of the predictive model in clinical settings?
4. **Research Question 4 (RQ4):** To what extent do existing studies incorporate cross-institutional validation, interpretable AI approaches, and real-world translational readiness considerations within analytical modelling approaches?
5. **Research Question 5 (RQ5):** What are the major methodological limitations, research gaps, and future opportunities to develop clinically integrated ML-based personalised therapy optimisation systems for GLP-1 therapies?

The literature search aimed to comprehensively capture studies exploring treatment-response prediction in GLP-1 receptor agonist and incretin-based treatments using AI. Table 2 presents the exact search strings applied in each database. The databases chosen for the searches were PubMed, Scopus and Web of Science, as they cover a wide range of biomedical, clinical, and computational research. Combining therapy-related language, treatment-response outcomes, and artificial intelligence/predictive modelling terms, database-specific search strategies were developed. Controlled vocabulary (MeSH) and free text keywords were used in PubMed, if

Table 2 Search strings used in the literature review

Database	Search String
PubMed	(("GLP-1 receptor agonist" OR semaglutide OR tirzepatide OR liraglutide OR dulaglutide OR "incretin-based therapy")) AND (predict* OR response OR "non-response" OR adherence OR discontinuation OR "HbA1c" OR "weight loss") AND ("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network" OR "random forest" OR XGBoost OR "logistic regression")
Web of Science	TS=(("GLP-1 receptor agonist" OR semaglutide OR tirzepatide OR liraglutide OR dulaglutide OR incretin)) AND TS=(predict* OR response OR "non-response" OR adherence OR discontinuation OR "HbA1c" OR "weight loss") AND TS=("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network" OR "random forest" OR XGBoost OR "logistic regression")
Scopus	TITLE-ABS-KEY (("GLP-1 receptor agonist" OR semaglutide OR tirzepatide OR liraglutide OR dulaglutide OR incretin)) AND TITLE-ABS-KEY (predict* OR response OR "non-response" OR adherence OR discontinuation OR "HbA1c" OR "weight loss") AND TITLE-ABS-KEY ("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network" OR "random forest" OR XGBoost OR "logistic regression")

available, to maximise retrieval sensitivity. Reference lists of eligible studies and relevant review articles were hand-searched to enhance search completeness [38]. Before title and abstract screening, duplicate records were eliminated. Title and abstract screening, full-text eligibility assessment, and data extraction were performed by Tahsien Al-Quraishi. Independent duplicate screening and extraction were not performed. The study identification and screening, eligibility assessment and inclusion process is outlined in Figure 2. Eighty full-text articles were assessed for eligibility, and 30 were excluded because they did not meet the eligibility criteria. Six additional records were subsequently removed because a matching primary source could not be verified, leaving 44 studies in the final synthesis. The review was carried out according to the PRISMA 2020 guidelines [37]. The review was not prospectively registered, and a protocol was not finalised before screening commenced. This limits the ability to distinguish prespecified methods from subsequent methodological decisions.

2.3 Eligibility Criteria

A process of examining the retrieved records was conducted to ensure methodological consistency, clinical relevance, and suitability for evidence synthesis based on predefined inclusion/exclusion criteria. Only studies published in English from 2018 to 2026 were considered. Only peer-reviewed journal articles that studied GLP-1 receptor agonists and incretin-based therapies were included. Review articles, conference abstracts, editorials/ comments, study protocol, and duplicate records were excluded. The study selection process was conducted in two phases. The titles and abstracts were first selected to determine whether the article was relevant to GLP-1 receptor agonists and incretin-based therapies. Studies that were not based on incretins were excluded. Secondly, the methodological relevance and applicability for the clinical prediction of full-text articles were assessed.

Studies in which researchers used AI, ML, DL or predictive analytics were included to investigate prediction of treatment response, non-response, prediction of weight

loss, improvement in HbA1c, adherence, persistence, discontinuation, prediction of adverse events and/or optimisation of therapy. Additional studies considered were selected clinical trials and observational studies that provided clinically relevant data on treatment-response heterogeneity, responder characteristics, adherence patterns, persistence behaviour, or risk of discontinuation of treatment or other determinants of therapeutic benefit that informed the development, interpretation or evaluation of predictive models. Studies that provided only data on the diagnosis of disease, diabetes risk prediction, descriptive statistical data or non-predictive clinical reports and did not provide data on the assessment of treatment response were excluded. Studies that did not provide sufficient methodological detail, validation procedures, or performance evaluation metrics were also excluded to improve the comparability and reliability of the included predictive frameworks [39]. Screening and eligibility assessment were performed by a single reviewer. This introduces a potential risk of selection and extraction errors.

For clarity, the included literature comprised two complementary evidence categories: (i) studies of artificial intelligence and machine learning that explicitly developed predictive models, and (ii) clinically informative treatment-response investigations that provided evidence regarding responder characteristics, adherence behaviour, persistence of treatment, discontinuation patterns, and therapeutic heterogeneity. Although the latter studies did not implement predictive modelling frameworks, they were retained because they provide clinically relevant outcome definitions, benchmark therapeutic endpoints, and patient-level determinants that inform the development and interpretation of future prediction systems [40]. Consequently, the review combines computational prediction studies with clinically informative treatment-response evidence to provide a more comprehensive assessment of the translational landscape to optimise personalised GLP-1 therapy.

2.4 Methodological Quality Assessment

After selecting studies and eligibility assessment, a structured quality assessment (QA) framework was used to evaluate the methodological quality, clinical relevance of patient stratification systems, robustness of validation, and clinical applicability of included studies. Quality assessment criteria were developed to address predefined research questions and comprised three main domains of clinical decision-support evaluation: (i) quality and clinical relevance of the data set, (ii) clinical decision-support frameworks and (iii) validation and performance evaluation practices. Although frameworks such as PROBAST, PROBAST-AI, QUADAS-AI, and TRIPOD-AI offer guidance to assess the quality, risk of bias, and reporting standards of prediction models, [34, 36, 41], these instruments address different purposes and should be selected according to study design and assessment objective. PROBAST+AI addresses prediction-model quality, risk of bias and applicability, whereas TRIPOD+AI addresses reporting completeness. A heterogeneous evidence set requires design-appropriate appraisal rather than replacement of established instruments with a single unvalidated score [42]. Therefore, a customised assessment framework was created to allow a uniform evaluation of the methodological quality, validation rigour, interpretability, and clinical applicability of included studies [43]. Individual studies

in each quality assessment criterion were categorised into three levels, indicating the degree of adherence to the methodological requirement.

2.4.1 QA1. Clinical Relevance and Dataset Quality

This criterion evaluates whether the included studies used clinically relevant datasets suitable for the prediction of the response to treatment in therapies based on the GLP-1 receptor agonist and incretin.

QA1.1 The study includes limited clinical variables and lacks treatment-specific, longitudinal, or patient-level information relevant to the prediction of treatment response.
QA1.2 The study incorporates clinically relevant variables such as BMI, HbA1c, age, or medication history, but provides limited treatment-response, adherence, or longitudinal follow-up information.
QA1.3 The study uses comprehensive patient-level datasets, including electronic health records (EHRs), treatment histories, laboratory measurements, longitudinal results, adherence patterns, and routine clinical follow-up data.

2.4.2 QA2. Predictive Modelling Methodology

This criterion evaluates the use of appropriate approaches to artificial intelligence, machine learning, deep learning, and/or predictive analytics to predict treatment-related outcomes.

QA2.1 No predictive modelling framework is implemented and the study is based primarily on descriptive statistical analysis.
QA2.2 Basic statistical or predictive approaches are applied without advanced machine learning implementation or comparative modelling analysis.
QA2.3 Advanced therapeutic prediction systems, including machine learning, deep learning, ensemble learning, survival analysis, or explainable artificial intelligence approaches, are implemented to predict treatment response, non-response, adherence, persistence, discontinuation, or adverse events.

2.4.3 QA3. Validation Quality and Performance Evaluation

This criterion evaluates the robustness of validation procedures and the adequacy of predictive performance assessment.

QA3.1 The study lacks appropriate validation procedures or uses insufficient performance evaluation strategies.
QA3.2 Limited validation is performed using restricted evaluation metrics, single train–test splits, or incomplete comparative assessment.
QA3.3 Robust validation strategies are implemented using multiple performance metrics, including accuracy, precision, recall, F1-score, ROC-AUC, calibration assessment, cross-validation, or cross-institutional validation.

2.4.4 QA4. Explainability and Interpretability

This criterion assesses the extent to which the study incorporates explainability techniques to improve model transparency and clinician understanding.

QA4.1 No explainability or interpretability mechanisms are reported.
QA4.2 Limited interpretability is provided through feature-importance analysis or basic model interpretation techniques.
QA4.3 Advanced explainability approaches, including SHAP, LIME, attention mechanisms, or interpretable modelling frameworks, are integrated to support transparent clinical decision-making.

2.4.5 QA5. Clinical Translation and Translational Readiness

This criterion evaluates the potential of the predictive framework for clinical implementation and real-world deployment.

QA5.1 The study is exploratory and provides limited discussion of clinical applicability or implementation.
QA5.2 The study discusses potential clinical utility, but lacks external validation, workflow integration, or deployment considerations.
QA5.3 The study demonstrates characteristics consistent with clinically deployable systems, including external validation, real-world evaluation, workflow integration, and translational readiness.

The Quality Assessment Findings were applied to interpret the results and make comparisons between the included studies in terms of methodological strength, validity of the results, interpretability, and clinical applicability. Although PROBAST, PROBAST-AI, QUADAS-AI and TRIPOD-AI provide valuable guidance to evaluate prediction-model quality, reporting standards, and risk of bias [44], the included literature consisted of a heterogeneous mixture of machine-learning studies, explainable AI investigations, observational treatment-response studies, and clinical trial evidence. Consequently, a tailored assessment framework was developed to enable a consistent comparison of dataset quality, predictive methodology, validation robustness, interpretability, and translational readiness across diverse study designs. The proposed framework should therefore be interpreted as a complementary synthesis tool rather than a replacement for established quality-assessment methodologies [45].

2.5 Data Extraction

The information was systematically retrieved through a structured data extraction process that was relevant to the predefined research questions and the objectives of evidence synthesis. The framework for extracting clinical, methodological, and predictive features specific to AI research on GLP-1 receptor agonists and incretin-based therapies. First, studies in the included publications were classified into the following categories at the level of the primary predictive goal: treatment-response prediction, non-response prediction, adherence study, persistence study, discontinuation prediction, adverse events prediction, and prediction of other clinical effectiveness (such

as weight loss response and improvement of HbA1c). This classification allowed for inter-study comparison of studies with different therapeutic and/or predictive endpoints. Then each included study was applied using a predetermined data extraction framework, applying the following.

- **Type of therapy: type** of GLP-1 receptor agonist or incretin-based therapy investigated, including agents such as *Semaglutide*, *Tirzepatide*, *ligandulosic* acid, and *Dulaglutide*.
- **Predictive modelling approach:** AI, ML, DL, statistical, or ensemble learning methods implemented within the study, including logistic regression, Random Forest, gradient boost machines, XGBoost, Support Vector Machines, deep neural networks, and survival analysis models.
- **Clinical outcomes:** Treatment-related outcomes predicted by the model, including prediction of treatment-response, non-response, weight-loss outcomes, reduction in HbA1c, adherence, persistence, discontinuation, cardiometabolic outcomes and occurrence of adverse-events.
- **Validation strategy:** Validation procedures used to assess model robustness and generalisability, including train-test split validation, cross-validation, bootstrapping, temporal validation, and external or multicenter validation.
- **Evaluation metrics:** Performance measures reported within the study, including accuracy, precision, recall, F1-score, ROC-AUC, sensitivity, specificity, calibration measures, and survival-related metrics where applicable.
- **Dataset characteristics:** Data source, type of data set, sample size, and population characteristics, including electronic health records (EHRs), clinical trials, registries, observational cohorts, and genomic datasets.
- **Interpretability and complementary techniques:** Additional modelling and explainability approaches, including feature selection, hyperparameter optimisation, ensemble learning, SHAP, LIME, and other explainable AI techniques.

The resulting information was consolidated to identify commonalities across computational prediction systems, modelling approaches, validation methods, explainability frameworks, and clinical hurdles for the application of AI-driven treatment optimisation for GLP-1 therapy.

2.6 Risk of Bias Assessment

A structured bias assessment was performed to highlight methodological limitations and potential sources of bias that can affect the reliability, interpretability, reproducibility, and clinical applicability of predictive studies of GLP-1 receptor agonists and incretins [46]. Sources of bias within and between studies were assessed, including the quality of datasets, modelling methods, validation processes and practices, performance evaluation methods and practices, and reporting transparency. Within-study bias assessment was performed regarding limited sample size, class imbalance, inadequate description of preprocessing steps, inadequate transparency of feature engineering, potential overfitting, missing external validation, and limited performance evaluation [47]. Special focus was placed on studies that used only one split between

trains and tests or only internally validated datasets, as these results might not lead to robust models that are generalisable to other independent populations [48].

A cross-study bias assessment has evaluated the heterogeneity in data sources, patient populations, treatment outcomes, predictive objectives, modelling frameworks, and evaluation methodologies. Variability in the structure of the EHRs, observational cohort designs, clinical trial populations, and reporting standards was also taken into account, as they could impact comparability and reproducibility between healthcare settings [49]. The assessment also assessed the possibility of reporting bias due to an inadequate description of the methods, selective reporting of performance measures, inadequate explainability analysis, and inadequate discussion of implementation challenges [50]. Researchers who were able to produce studies with a higher degree of methodological transparency, with robust validation procedures and detailed reporting of performance measurements, were deemed to have a lower risk of bias and therefore a higher potential for clinical implementation. Logical transparency, rigorous validation procedures, and complete descriptions of performance criteria were perceived to have less general bias and greater potential to be applied in clinical practice.

Detailed findings from this assessment, reported separately at the level of individual studies and across the body of evidence, are presented in Section 3.6.

3 Results

This section presents the findings of the systematic review, including the results of the study selection, the results of the quality assessment, the characteristics of the included studies, and the comparative synthesis of predictive models applied to glucagon-like receptor agonist (GLP-1 RA) and incretin-based therapies. The reviewed studies demonstrated substantial methodological diversity across predictive objectives, datasets, modelling approaches, validation strategies, interpretability techniques, and therapeutic outcome measures. The results further examine dataset characteristics, predictive architectures, validation practices, explainable AI integration, and clinically relevant outcomes related to treatment-response prediction, non-response, adherence, persistence, discontinuation, and optimisation of therapy in the management of obesity and type 2 diabetes.

3.1 Study Selection

A search of the PubMed, Web of Science and Scopus literature yielded 100 potentially relevant records. After the removal of duplicate studies, 80 unique studies remain from the original studies. Records not related to GLP-1 receptor agonists, incretin-based therapies, or predictive modelling methodologies were excluded during the screening process. Additional exclusions included studies focused solely on disease diagnosis, non-predictive statistical analyses, non-clinical investigations, conference abstracts, and peer-reviewed publications. Subsequently, Full-text articles were assessed for eligibility. Studies were excluded if they did not implement AI, machine learning, deep learning, or predictive analytics approaches for the prediction of treatment-related outcomes or if they lacked clinically relevant results, sufficient methodological detail, validation procedures, or performance evaluation reporting. Targeted searches and

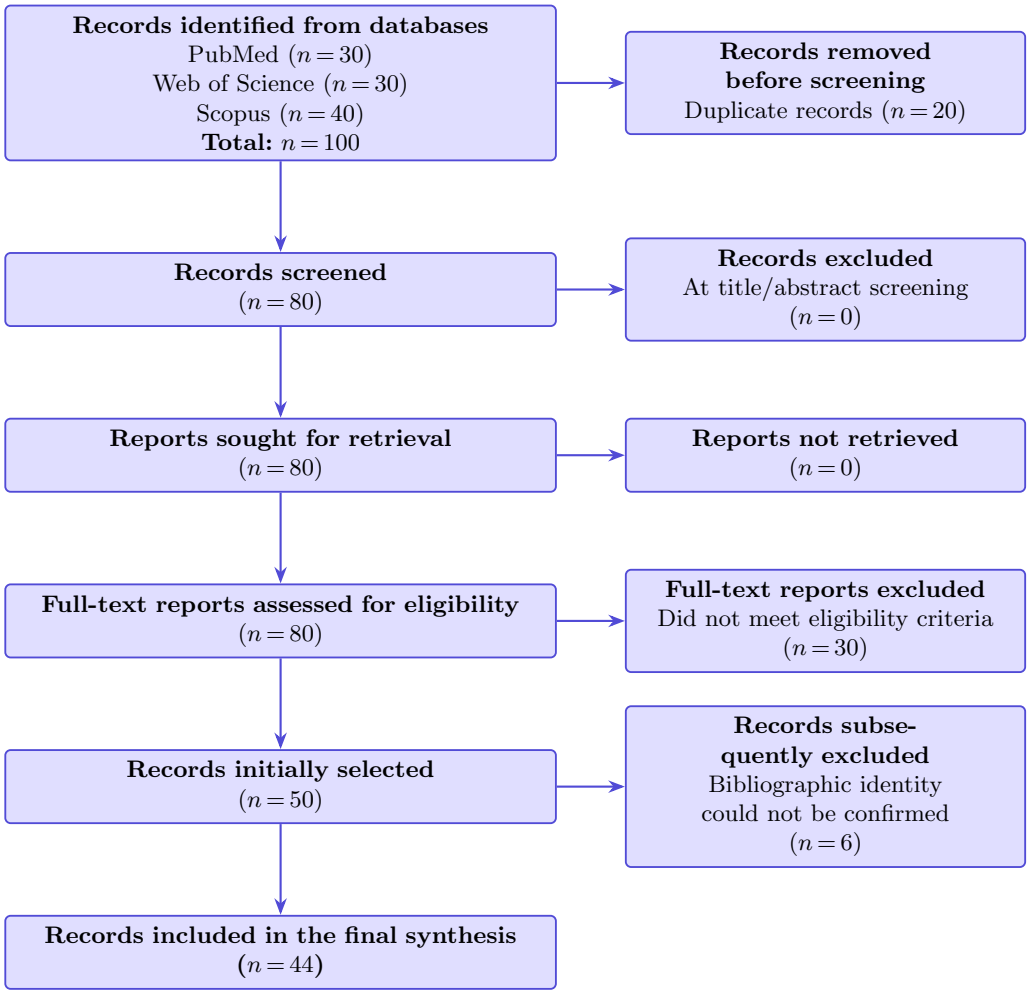


Fig. 2 Study-selection flow diagram

reference-list screening were also reported. However, the number of additional records identified and their contribution to the final included set were not separately documented, limiting the reproducibility of this component of the search. Fifty records were initially selected after eligibility assessment. Subsequently, five records were excluded because a matching primary source could not be confirmed. A further record attributed to Apostolopoulou et al. was excluded after its supplied DOI was found to identify a different publication, and the intended source could not be established. The final synthesis therefore contains 44 records. Table 9 summarises the included studies.

3.2 Characteristics of Included Studies

The included studies were evaluated using the predefined quality assessment framework, which examined the quality and clinical relevance of the dataset (QA1), the

predictive modelling methodology (QA2), validation robustness and performance evaluation (QA3), interpretability (QA4), and clinical translation readiness (QA5). In QA1, the majority of studies demonstrated strong clinical relevance through the use of patient-level datasets, including electronic health records, clinical records, treatment histories, laboratory measurements, and longitudinal follow-up data [9–11]. A smaller subset of studies was assigned lower QA1 categories due to limited clinical variables, restricted follow-up periods, or incomplete treatment-specific information.

The included records used heterogeneous analytical approaches. Clinical efficacy analyses should be distinguished from studies that developed or evaluated individual-level prediction models. The current extraction does not establish that advanced predictive modelling was used in the majority of the 44 records. Interpretability practices varied substantially in the reviewed literature. Although explainability techniques such as SHAP, LIME, and feature-importance analysis were increasingly incorporated, their application remained inconsistent and lacked standardisation in all studies [51]. Consequently, the level of transparency and clinician interpretability differed considerably between the predictive frameworks. Similarly, relatively few investigations demonstrated characteristics consistent with advanced clinical translation readiness. Most studies remained in the exploratory or internally validated stages of development, while only a limited number reported prospective evaluation, workflow integration, multicenter implementation, or ddeployment-orientedassessment [43].

In general, quality assessment findings indicate that the current literature is characterised by clinically relevant datasets and increasingly sophisticated predictive modelling approaches. However, there remains important variability in validation practices, interpretability integration, methodological transparency, and translational maturity. These findings highlight the need for greater standardisation of evaluation frameworks, reporting practices, external validation procedures, and clinically robust benchmarking strategies to facilitate the successful translation of predictive models into routine clinical practice [52]. All studies that met the predefined eligibility criteria were retained for qualitative synthesis and comparative evaluation. Consequently, the quality assessment framework was used to support the interpretation of methodological robustness, validation reliability, interpretability, and clinical maturity rather than as a basis for study exclusion.

3.3 Risk of Bias Assessment Results

At a high level, the risk-of-bias evaluation identified the same recurring weaknesses that are broken down in detail in Section 3.6: limited external validation, heavy reliance on retrospective and internally validated datasets, inconsistent reporting of calibration, and uneven integration of explainability methods. These patterns directly motivate the Artificial Intelligence Translational Readiness Framework (AI-TRF) introduced in Section 4.3, which structures external validation, explainability integration, temporal modelling, prospective evaluation, and clinical implementation readiness as five dimensions of translational maturity [34–36].

3.4 Study Characteristics

The studies included in the analyses consisted of three main types: (i) studies using artificial intelligence or machine learning to predict treatment response; (ii) clinical predictor and treatment response studies; and (iii) clinical trial and observational effectiveness studies that yield evidence on treatment response heterogeneity, responder characteristics, adherence patterns, persistence behaviour, and determinants of therapeutic benefit. Some of the included investigations did not produce predictive AI models, but were still included because they identified clinically relevant outcomes, treatment-response frameworks, benchmark therapeutic outcomes, and patient-level determinants that are critical to the development, validation and interpretation of therapeutic prediction systems for GLP-1 receptor agonist therapies [53]. For this reason, the review provides a more complete translational perspective of the data, including computational prediction data and clinically relevant treatment-response data [54].

Table 9 summarises the principal characteristics of the included studies investigating artificial intelligence, machine learning, deep learning, and predictive analytics approaches in glucagon-like peptide-1 receptor agonist (GLP-1 RA) and incretin-based therapies. The information extracted included therapy type, data set source, predictive modelling approach, clinical results, validation strategies, performance metrics, and key findings. The reviewed studies demonstrated substantial methodological diversity across predictive objectives, datasets, modelling approaches, and evaluation procedures [9–11]. Considerable variation was also observed in the therapies investigated, with studies based on semaglutide and tirzepatide representing the approaches reported most frequently (Table 3). Most studies focused on clinically relevant results, including treatment-response prediction, non-response, weight-loss prediction, HbA1c improvement, adherence, persistence, discontinuation and prediction of advadverse eventsite-Davies2021STEP2,Buse2022,McGovern2018Persistence,DelPrato2021PIONEER10. These findings reflect the growing emphasis on personalised treatment strategies and predictive decision-support systems for the treatment of obesity and type 2 diabetes.

Table 3 Distribution of therapies investigated across included studies

Therapy	**Frequency**
Semaglutide (including oral semaglutide)	17
Tirzepatide	10
GLP-1 receptor agonists (generic)	10
Dulaglutide	2
Multiple or mixed therapies	5

As shown in Table 3, semaglutide and tirzepatide were the therapies investigated most frequently, reflecting their widespread clinical adoption and the growing availability of clinical trial datasets and real-world datasets. Compared to others, fewer studies focused on incretin-based liraglutide, dulaglutide, exenatide, and other

therapies. This trend indicates a growing focus on newer-generation GLP-1 receptor agonists for weight management and glycaemic control. Considerable variation was also observed in the validation methodologies used in the included studies. Table 4 summarises the distribution of study-design and evidence-source categories recorded for the included studies.

Table 4 Distribution of study-design and evidence-source categories recorded across included studies. Categories combine study design, data setting, and validation terminology as originally extracted and should not be read as verified prediction-model validation types (see accompanying text).

Study-Design / Evidence-Source Category	**Frequency**
Randomised clinical trial validation	25
Clinical validation	4
Observational validation	2
Clinical trial validation	2
Real-world validation	3
Retrospective validation	3
Other validation approaches	5
Total	**44**

Table 4 reproduces the labels recorded in the original extraction. These labels combine study design, data setting, and validation terminology. An efficacy trial does not, by itself, validate a prediction model. Accordingly, these frequencies cannot establish the prevalence of internal, external, temporal, or prospective model validation. Such estimates require a separate extraction restricted to eligible prediction studies. In contrast, the real-world, retrospective, and independent validation approaches remained relatively rare. This finding highlights a persistent translational gap between model development and clinical implementation, as many predictive models continue to rely on internally validated datasets and fail to demonstrate reproducibility in independent healthcare settings. XAI approaches, including SHAP and LIME, have been increasingly incorporated into AI-enabled prediction systems to improve model transparency and support the interpretation of treatment-response predictions. However, meaningful clinical integration remained limited, with many studies applying the explainability retrospectively rather than embedding it directly within clinical decision-support workflows [9, 10].

A broad range of predictive methodologies was identified in the reviewed literature, including LR, RF, GBM, XGBoost, survival-analysis models, deep learning architectures and ensemble learning approaches [9–11]. Several studies also incorporated feature-selection procedures, hyperparameter optimisation, and explainability techniques to improve predictive performance and model transparency. Ensemble learning approaches, particularly RF, GBM and XGBoost, were among the most frequently implemented computational frameworks due to their ability to model complex nonlinear relationships between demographic, behavioural, physiological, and

treatment-related variables [9–11]. These methods demonstrated a particular suitability for obesity and type 2 diabetes prediction environments characterised by heterogeneous patient populations and multifactorial treatment-response mechanisms.

Longitudinal metabolic datasets introduced additional modelling complexity because prediction of treatment-response, adherence behaviour, HbA1c progression, weight-loss trajectories, occurrence of adverse events, and discontinuation behaviour evolve dynamically over time [11, 14, 55, 56]. Several investigations highlighted important differences between early glycaemic improvement and sustained long-term weight-loss response following GLP-1 therapies, emphasising the importance of modelling temporal treatment dynamics and longitudinal outcome trajectories. Traditional statistical approaches, including proportional hazards regression of LR and Cox, remained important due to their transparency, interpretability, and clinical applicability [9, 11, 56]. Survival-based methodologies demonstrated particular utility to model persistence in treatment, risk of discontinuation, adherence behaviour, and other day-to-day outcomes associated with GLP-1 therapies.

Substantial cohort heterogeneity was observed in all reviewed studies. Variability in the obesity phenotype, cardiometabolic severity, adherence to treatment, comorbidity burden, and pharmacological response may influence predictive reliability and generalisability of the model in healthcare settings [9–11]. Differences in electronic health record structures, treatment protocols, and outcome definitions further complicate independent validation and cross-cohort comparison of predictive models. Although ensemble learning and deep learning approaches frequently demonstrated stronger predictive performance than conventional statistical models, these improvements were often accompanied by reduced interpretability and increased implementation complexity [9, 10, 14]. In contrast, conventional statistical methodologies generally provided greater transparency and easier integration into routine clinical workflows [9, 11, 56]. This trade-off between predictive performance and clinical interpretability remains a central challenge in the development of clinically applicable therapeutic decision-support systems.

Different predictive architectures demonstrated varying suitability depending on the clinical objective. Models with complex nonlinear relationships between metabolic and behavioural variables were found to be particularly successful for weight loss response and prediction of glycaemic improvement (ensemble learning approaches such as RF and XGBoost). However, when persistence of adherence, discontinuation of treatment, and time to event were examined, survival-based approaches proved to be more appropriate, but LR was an appropriate method for situations where a focus was placed on model transparency and interpretability. Despite the challenges of explanation and usage in clinical practice, deep learning architectures showed promise in terms of multimodal fusion of EHR and longitudinal modelling of treatment response [9–11, 14, 56].

Table 5 shows the relative suitability of the main AI models based on the most prevalent clinical objectives identified in the reviewed literature and should be seen as a synthesis of the literature, based on these objectives, rather than a definitive ranking of the superiority of individual algorithms. Considerable methodological variability was observed between the structure of the data set, the validation methodology

Table 5 Clinical suitability of prediction models across GLP-1 therapeutic outcomes

Clinical Outcome	Candidate Approaches	Rationale
Weight-loss response	Random Forest, XGBoost	Capture nonlinear metabolic and behavioural interactions.
HbA1c improvement	Random Forest, XGBoost, GBM	Strong discrimination across heterogeneous metabolic datasets.
Adherence prediction	Survival Models, Cox Regression	Explicit modelling of longitudinal persistence behaviour.
Discontinuation prediction	Survival Models, Cox Regression	Designed for time-to-event and discontinuation analysis.
Longitudinal response prediction	Deep Learning, Temporal Models	Capture evolving treatment trajectories and sequential patterns.
Interpretable decision support	Logistic Regression, Explainable RF	Greater transparency and clinician interpretability.

and the analytical design based on models [9–11]. The included studies used biologically diverse data sources, including electronic health records, observational records, clinical trial cohorts, pharmacy refill databases, and longitudinal metabolic datasets, with significant differences in patient demographics, treatment duration, follow-up intervals, outcome definitions, and preprocessing strategies [10, 11]. These variations substantially affected reproducibility, calibration stability, and comparability between studies across therapeutic prediction systems.

Baseline HbA1c severity, variability in obesity phenotype, cardiometabolic burden, adherence behaviour to treatment, pharmacogenomic variability, and longitudinal metabolic adaptation emerged as recurrent determinants of therapeutic-response heterogeneity in glucagon-like receptor agonist therapies [9, 10, 14, 32, 33, 57, 58]. Several studies also suggested that cardiometabolic burden, gastrointestinal intolerance, and persistence of treatment substantially influenced both the sustainability of weight-loss and the trajectories of glycaemic response [10, 14, 56]. Variability in behavioural adherence and discontinuation patterns further complicated predictive stability in extended therapeutic follow-up periods [11, 14, 56]. Real-world observational analyses also highlighted substantial variability in long-term adherence, persistence behaviour, and discontinuation trajectories between populations of glucagon-like receptor agonists [11, 13, 56]. These findings highlight the biological and behavioural complexity underlying the prediction of the response to glucagon receptor agonist treatment and reinforce the need for therapeutic decision-support systems capable of integrating dynamic clinical, metabolic, and patient-specific treatment characteristics.

From a machine learning perspective, the findings help explain the widespread adoption of ensemble learning approaches in the reviewed literature. The prediction of response in patients receiving GLP-1 receptor agonist therapies is influenced by complex nonlinear interactions among demographic, metabolic, behavioural, pharmacological, and genetic factors that are often difficult to capture using conventional linear statistical models. Consequently, ensemble-learning approaches such as Random Forest and gradient-boosting algorithms frequently demonstrate superior predictive performance because they can automatically model higher-order feature interactions and nonlinear decision boundaries without requiring explicit specification of interaction terms [9, 10].

RF models also maximise robustness by applying bootstrap aggregation, which helps reduce variance compared to the individual decision tree [59]. Similarly, XGBoost and other boosting methods make a series of improvements by persisting in the direction of minimising residual prediction error, which makes them very useful for a set of metabolic data that are very heterogeneous, with different treatment trajectories and different patient-specific response patterns [60]. However, good prediction accuracy does not imply real-world applications [61]. Discrimination metrics, such as ROC-AUC, precision, precision, recall, and F1-score, were reported more often, while calibration evaluation and clinical utility evaluation were reported less often [62]. In addition, several predictive models focused mainly on baseline variables that were static, and response to treatment, adherence behaviour, chance of discontinuation, and metabolic adaptation are characteristics that change over time [63].

Although foundation models, multimodal transformer architectures, and clinical large language models (LLMs) were not identified among the included studies, recent advances in these technologies may represent promising future directions for therapeutic prediction research. Beyond conventional machine-learning and ensemble-learning methodologies, future predictive systems may increasingly benefit from foundation models, clinically adapted LLMs, multimodal transformer architectures, and generative artificial intelligence. Emerging foundation-model frameworks have demonstrated the potential to integrate heterogeneous healthcare data sources [64], including structured electronic health records, laboratory measurements, clinical narratives, medical imaging, genomic information, and patient-reported results within unified predictive environments.

Similarly, multimodal transformer architectures can facilitate improved modelling of longitudinal treatment trajectories and patient-specific treatment-response patterns through simultaneous analysis of various streams of clinical information [65]. Generative artificial intelligence and clinically adapted LLMs can also support clinical decision-support systems [66], patient stratification, automated evidence synthesis, and clinically assisted therapeutic recommendation workflows. Although these technologies are still in an early stage of clinical translation, they represent promising avenues for future research and the development of next-generation therapeutic prediction systems for the management of obesity and type 2 diabetes [67, 68].

In particular, the widespread reliance on retrospective institution-specific datasets and internally validated train-test frameworks remains a major limitation that affects the reliability of clinical implementation [9, 10]. Although several studies

Table 6 Semi-quantitative comparison of ML approaches

Model Type	Common Outcomes	Key Strengths	Limitations
Random Forest	Weight loss, HbA1c	Captures nonlinear relationships	Reduced interpretability
XGBoost	Weight loss, glycaemic response	Flexible modelling of nonlinear relationships	Parameter complexity
Logistic Regression	Response classification	High transparency and interpretability	Limited nonlinear modelling
Survival Models	Adherence/ discontinuation	Time-to-event analysis	Limited interaction modelling
Deep Learning	Longitudinal EHR modelling	Temporal pattern learning	Computational complexity; reduced explainability

reported favourable predictive performance, relatively few incorporated independent cohort validation, prospective clinical investigation, or cross-institutional benchmarking strategies to assess robustness in clinically heterogeneous healthcare environments [10, 11, 13]. Only a limited proportion of the investigations incorporated external cohort validation or cross-institutional benchmarking procedures to evaluate predictive reproducibility in independent healthcare settings [9, 10, 13]. Consequently, many currently reported glucagon-like peptide-1 therapeutic prediction systems remain closer to retrospective exploratory frameworks than clinically generalisable therapeutic decision-support systems suitable for routine metabolic disease management. To further synthesise the evidence extracted, additional analyses were conducted to examine the distribution of computational prediction approaches, target clinical outcomes, validation practices, interpretability integration, and data set characteristics among included studies.

The synthesis is descriptive. Table 6 summarises this semi-quantitative comparison of the most commonly reported model types, their typical clinical outcomes, and their principal strengths and limitations. The extracted information does not support a pooled estimate of predictive performance or a reliable ranking of algorithms. Comparative assessment requires study-level information on prediction targets, evaluation datasets, performance estimates, uncertainty, calibration, and validation procedures. Therefore, general methodological comparisons should be distinguished from empirical findings derived from the included studies. Random Forest and XGBoost emerged as the most frequently implemented machine learning approaches [9–11]. However, substantial methodological heterogeneity precluded formal meta-analysis and pooled estimation of predictive performance metrics. Considerable variability was observed in the definitions of the outcomes, the evaluation metrics, the characteristics of the data set, the composition of the cohort, the duration of follow-up, the therapeutic agents and the predictive objectives. In addition, studies reported various performance measures, including ROC-AUC, precision, hazard ratios, concordance indices, and survival-based outcomes, which limits direct statistical comparison. Consequently, the present review adopts a semi-quantitative and comparative methodological synthesis to evaluate modelling trends, validation practices, interpretability integration, and clinical outcome targets across included studies [9–11, 14, 56].

Table 7 Quantitative summary of included studies

Characteristic	Frequency
Included studies	44
Semaglutide studies	17
Tirzepatide studies	10
GLP-1RA studies	10
Dulaglutide studies	2
Mixed therapies	5
Randomised clinical trial validation	25
Clinical validation	4
Observational validation	2
Clinical trial validation	2
Real-world validation	3
Retrospective validation	3
Other validation approaches	5

3.5 Quantitative Summary of Included Studies

A quantitative descriptive summary of the reviewed literature is presented in Table 7. The most common therapeutic areas in the 44 included studies were those that included semaglutide or tirzepatide, and the most popular predictive modelling methods were ensemble learning models such as RF, GBMs, and XGBoost. There was significant heterogeneity between the validation strategies used, with independent cohort validation, real-world assessment, and prospective studies still relatively rare. The quantitative summary demonstrates the gradual evolution of the sophistication of therapeutic prediction systems and the ongoing challenges in the robustness of validation and ultimate clinical use [69].

As shown in Table 7, semaglutide and tirzepatide-based investigations dominated the reviewed literature, while ensemble learning approaches represented the predictive modelling paradigm reported the most frequently. In contrast, independent validation and prospective clinical assessment remained relatively limited in all included studies.

Table 8 Distribution of analysis and modelling categories across the 44 included studies. Categories were derived from the modelling descriptions reported in Table 9.

Analysis/Modelling Category	Frequency
Comparative/clinical outcome	31
Predictive/ML/stratification	11
Longitudinal/persistence	1
Pharmacogenomic/genotype	1
Total	**44**

Table 8 summarises the broad categories of analysis assigned in the current extraction: 31 comparative or clinical-outcome records, 11 predictive, machine-learning or stratification records, one longitudinal or persistence record, and one pharmacogenomic or genotype record. These broad categories do not identify individual algorithms and cannot establish the frequency or superiority of Random Forest, XGBoost, or other ensemble methods. Despite the growing integration of explainable AI approaches, relatively few studies incorporated external validation, longitudinal modelling, or future implementation study frameworks [9, 10]. As shown in Table 4, these limitations continue to represent important barriers to clinical translation and implementation of predictive models for GLP-1 therapies.

In all reviewed studies, a progressive shift was observed from retrospective exploratory modelling to more clinically orientated therapeutic decision-support systems [10, 11]. Earlier investigations were predominantly characterised by institution-specific datasets, internal validation, and limited deployment considerations, whereas more recent studies increasingly incorporated independent cohort validation, explainability techniques, extended follow-up analyses, and cross-institutional benchmarking procedures [9, 10]. Nevertheless, relatively few studies demonstrated characteristics consistent with routine clinical deployment in diverse healthcare settings.

The qualitative patterns observed across the included studies – limited external validation, inconsistent explainability integration, predominantly static (non-temporal) modelling, and near-absence of prospective evaluation – motivate the five-dimension AI-TRF structure introduced formally in Section 4.3 and Table 11.

To further synthesise the clinical implementation pathways identified in the reviewed literature, Figure 3 presents a conceptual workflow for AI-driven GLP-1 therapeutic prediction systems [10, 11]. The framework illustrates the progression from clinical data acquisition and feature engineering to predictive modelling, explainability integration, validation, clinical decision support, and optimisation of personalised therapy. In addition, it reflects the transition from population-level outcome prediction to patient-specific therapeutic decision-support systems that integrate longitudinal metabolic, behavioural, and treatment-response characteristics [9, 10].

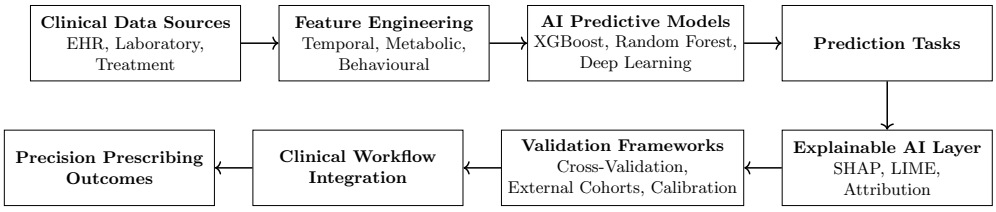


Fig. 3 Conceptual framework for clinical implementation of AI-driven GLP-1 therapeutic prediction systems, from clinical data acquisition and feature engineering through predictive modelling, explainability integration, validation, workflow integration, and personalised prescribing outcomes.

Table 9 Characteristics of the 44 included studies examining treatment-response prediction and clinical outcomes in GLP-1 receptor agonist and incretin-based therapies. Reference numbers correspond to entries in the bibliography.

No.	Ref.	Author(s)	Year	Therapy	Dataset source	Cohort size	Model	Validation	Performance	Main finding
1	[70]	Zhu et al.	2024	GLP-1 pathway drugs	Real-world clinical dataset	NR	Machine learning prediction model	Real-world validation	NR	Predicted responsiveness to GLP-1 pathway drugs using real-world clinical data and machine learning.
2	[71]	Vellikudathil et al.	2024	GLP-1 therapy	Clinical, genomic and proteomic dataset	NR	Machine learning algorithms	Exploratory validation	NR	Investigated clinical, genomic and proteomic markers associated with GLP-1 treatment-response prediction.
3	[72]	Gavigan and Donner	2023	GLP-1RA	Insulin-treated T2DM cohort	NR	Predictor analysis	Clinical validation	NR	Identified predictors of responsiveness to GLP-1 receptor agonists in insulin-treated patients.
4	[73]	Franch-Nadal et al.	2022	GLP-1RA	Real-world T2DM population	NR	Determinants-of-response analysis	Real-world validation	NR	Evaluated determinants of response to GLP-1 receptor agonists in routine clinical practice.
5	[74]	Dennis et al.	2025	Multiple diabetes drug classes	Routine clinical dataset	NR	Prediction model	Development and validation study	NR	Developed a five-drug class prediction model to optimise prescribing in type 2 diabetes.
6	[75]	Cardoso et al.	2024	SGLT2i / GLP-1RA	Type 2 diabetes cohort	NR	Phenotype-based treatment model	Clinical validation	NR	Applied phenotype-based targeted treatment selection for SGLT2 inhibitors and GLP-1 receptor agonists.
7	[76]	Pearson et al.	2025	Tirzepatide	Type 2 diabetes cohort	NR	Predictor analysis	Clinical validation	NR	Examined predictors of initial and sustained glycaemic and weight response to tirzepatide.
8	[58]	Małecki et al.	2023	Tirzepatide	SURPASS 1–4 trial cohorts	NR	Clinical predictor analysis	Trial validation	NR	Identified predictors of $\geq$15% weight reduction and associated cardiometabolic improvements.
9	[57]	Squire et al.	2025	GLP-1 analogues	Obesity treatment cohort	NR	Retrospective cohort analysis	Retrospective validation	NR	Identified factors associated with weight-loss response to GLP-1 analogue therapy.
10	[77]	Powell et al.	2023	Semaglutide	Real-world clinical dataset	NR	Real-world predictor analysis	Observational validation	NR	Evaluated medications and clinical conditions associated with semaglutide-related weight loss.

NR = Not Reported

Table 9 Summary of treatment-response prediction studies for GLP-1 therapies (continued)

No.	Ref.	Author(s)	Year	Therapy	Dataset source	Cohort size	Model	Validation	Performance	Main finding
11	[78]	Pratley et al.	2020	Semaglutide	SUSTAIN-7 post hoc cohort	1201	Post hoc treatment-response analysis	Clinical trial validation	Glycaemic outcomes reported	Baseline characteristics influenced comparative response to semaglutide and dulaglutide.
12	[79]	Ludvik et al.	2021	Tirzepatide	SURPASS-3 trial cohort	1444	Comparative efficacy analysis	Randomised clinical trial validation	HbA1c and weight outcomes reported	Tirzepatide demonstrated superior glycaemic control and weight reduction compared with insulin degludec.
13	[32]	Dawed et al.	2019	GLP-1RA	Pharmacogenomic cohort	NR	Pharmacogenomic analysis	Internal validation	Response variability	Genetic variability contributed to response heterogeneity.
14	[11]	McGovern et al.	2018	GLP-1RA	Real-world T2DM	NR	Persistence analysis	Retrospective validation	Persistence outcomes	Persistence varied across GLP-1RA users.
15	[80]	SURPASS-4 Investigators	2023	Tirzepatide	SURPASS-4 follow-up	NR	Longitudinal response analysis	Extension validation	Long-term outcomes	Follow-up showed sustained treatment effectiveness.
16	[81]	Garvey et al.	2022	Semaglutide	STEP-5 obesity cohort	304	Long-term obesity outcome analysis	Randomised clinical trial validation	Weight-loss outcomes reported	Semaglutide maintained clinically meaningful weight reduction over two years in adults with obesity.
17	[82]	Zinman et al.	2019	Oral Semaglutide	PIONEER-8 trial cohort	731	Comparative efficacy analysis	Randomised clinical trial validation	HbA1c and weight outcomes reported	Oral semaglutide improved glycaemic control and body weight in insulin-treated patients with type 2 diabetes.
18	[83]	Garvey et al.	2023	Tirzepatide	SURMOUNT-2 obesity and T2DM cohort	938	Comparative obesity outcome analysis	Randomised clinical trial validation	Weight-loss outcomes reported	Tirzepatide produced substantial weight reduction in adults with obesity and type 2 diabetes.
19	[84]	Wharton et al.	2022	GLP-1RA	Obesity treatment cohort	NR	Clinical outcome analysis	Clinical validation	Weight-loss outcomes reported	Clinical characteristics were associated with variability in treatment-response prediction among patients receiving GLP-1 receptor agonists.
20	[85]	Wadden et al.	2023	Tirzepatide	SURMOUNT-3 obesity cohort	806	Lifestyle-intervention analysis	Randomised clinical trial validation	Weight-loss outcomes reported	Tirzepatide enhanced weight reduction following intensive lifestyle intervention.

RCT = Randomised Clinical Trial

Table 9 Summary of treatment-response prediction studies for GLP-1 therapies (continued)

No.	Ref.	Author(s)	Year	Therapy	Dataset source	Cohort size	Model	Validation	Performance	Main finding
21	[86]	Aronne et al.	2024	Tirzepatide	SURMOUNT-4 obesity cohort	783	Weight-maintenance analysis	Randomised clinical trial validation	Weight-maintenance outcomes reported	Continued tirzepatide treatment supported long-term weight-loss maintenance.
22	[55]	Yabe et al.	2020	Oral Semaglutide	PIONEER-10 trial cohort	458	Comparative efficacy analysis	Randomised clinical trial validation	Glycaemic outcomes reported	Oral semaglutide demonstrated superior glycaemic improvement compared with dulaglutide in Japanese patients with type 2 diabetes.
23	[87]	Rosenstock et al.	2019	Oral Semaglutide	PIONEER-3 trial cohort	1864	Comparative efficacy analysis	Randomised clinical trial validation	HbA1c and weight outcomes reported	Oral semaglutide demonstrated superior glycaemic control compared with sitagliptin in patients with type 2 diabetes.
24	[88]	Rodriguez et al.	2025	GLP-1RA	Real-world clinical cohort	NR	Discontinuation risk assessment	Real-world validation	Treatment continuation outcomes reported	Factors associated with treatment discontinuation and persistence were identified.
25	[89]	Husain et al.	2019	Oral Semaglutide	PIONEER-6 cardiovascular cohort	3183	Cardiovascular outcome analysis	Randomised clinical trial validation	Cardiovascular safety outcomes reported	Oral semaglutide demonstrated cardiovascular safety in high-risk patients with type 2 diabetes.
26	[90]	Ahrén et al.	2017	Semaglutide	SUSTAIN-2 trial cohort	1231	Comparative efficacy analysis	Randomised clinical trial validation	HbA1c and weight outcomes reported	Semaglutide achieved superior glycaemic control and weight reduction compared with sitagliptin in patients with type 2 diabetes.
27	[33]	Dawed et al.	2023	GLP-1RA	Genotype-informed diabetes cohort	NR	Genotype-informed stratification model	Retrospective validation	Response heterogeneity reported	Genotype-informed variables supported personalised therapeutic stratification.
28	[91]	Frias et al.	2021	Dulaglutide	AWARD-11 trial cohort	1842	Dose-response efficacy analysis	Randomised clinical trial validation	HbA1c and weight outcomes reported	Higher-dose dulaglutide achieved greater glycaemic and weight benefits than standard dosing.
29	[10]	Wilding et al.	2021	Semaglutide	STEP-1 obesity cohort	1961	Comparative treatment-response analysis	Randomised clinical trial validation	Weight-loss outcomes reported	Semaglutide produced substantial weight reduction with variability across patient subgroups.
30	[92]	Perkovic et al.	2024	Semaglutide	FLOW kidney outcomes cohort	3533	Renal outcome analysis	Randomised clinical trial validation	Kidney outcomes reported	Semaglutide reduced kidney disease progression and cardiovascular risk in patients with type 2 diabetes and chronic kidney disease.

Table 9 Summary of treatment-response prediction studies for GLP-1 therapies (continued)

No.	Ref.	Author(s)	Year	Therapy	Dataset source	Cohort size	Model	Validation	Performance	Main finding
31	[93]	Ji et al.	2021	Semaglutide	SUSTAIN China trial cohort	868	Comparative efficacy analysis	Randomised clinical trial validation	HbA1c and weight outcomes reported	Semaglutide demonstrated significant glycaemic improvement and weight reduction compared with sitagliptin in patients with type 2 diabetes.
32	[94]	Wadden et al.	2021	Semaglutide	STEP-3 obesity cohort	611	Comparative obesity outcome analysis	Randomised clinical trial validation	Weight-loss outcomes reported	Semaglutide combined with intensive behavioural therapy produced substantial weight reduction in adults with obesity.
33	[95]	Rubino et al.	2021	Semaglutide	STEP-4 obesity cohort	803	Treatment-maintenance analysis	Randomised clinical trial validation	Weight-maintenance outcomes reported	Continued semaglutide therapy supported long-term weight-loss maintenance compared with treatment withdrawal.
34	[96]	Dahl et al.	2022	Tirzepatide	SURPASS-5 trial cohort	475	Comparative efficacy analysis	Randomised clinical trial validation	HbA1c and weight outcomes reported	Tirzepatide significantly improved glycaemic control and weight reduction when added to insulin glargine therapy.
35	[97]	SURPASS-4 Investigators	2021	Tirzepatide	SURPASS-4 trial cohort	2002	Comparative clinical outcome analysis	Randomised clinical trial validation	Glycaemic and cardiovascular outcomes reported	Tirzepatide demonstrated sustained glycaemic improvement and cardiometabolic benefit.
36	[98]	Frias et al.	2021	Semaglutide	SUSTAIN FORTE trial cohort	961	Comparative efficacy analysis	Randomised clinical trial validation	HbA1c and weight outcomes reported	Higher-dose semaglutide achieved greater glycaemic control and weight reduction than standard-dose therapy.
37	[99]	Pratley et al.	2019	Oral Semaglutide	PIONEER clinical trial cohort	711	Comparative efficacy analysis	Randomised clinical trial validation	Glycaemic outcomes reported	Oral semaglutide demonstrated significant glycaemic improvement and supported expansion of GLP-1 therapy options.
38	[14]	Davies et al.	2021	Semaglutide	STEP-2 obesity and T2DM cohort	1210	Comparative treatment-response analysis	Randomised clinical trial validation	Weight-loss and glycaemic outcomes reported	Semaglutide demonstrated clinically meaningful weight reduction in adults with obesity and type 2 diabetes.
39	[100]	Frías et al.	2021	Tirzepatide / Semaglutide	SURPASS-2 clinical trial cohort	1879	Comparative efficacy analysis	Randomised clinical trial validation	HbA1c and weight outcomes reported	Tirzepatide demonstrated superior glycaemic control and weight reduction compared with semaglutide in patients with type 2 diabetes.
40	[101]	Jastreboff et al.	2022	Tirzepatide	SURMOUNT-1 obesity cohort	2539	Comparative obesity outcome analysis	Randomised clinical trial validation	Weight-loss outcomes reported	Tirzepatide produced substantial and clinically meaningful weight reduction in adults with obesity.

Table 9 Summary of treatment-response prediction studies for GLP-1 therapies (continued)

No.	Ref.	Author(s)	Year	Therapy	Dataset source	Cohort size	Model	Validation	Performance	Main finding
41	[102]	Gerstein et al.	2019	Dulaglutide	REWIND cardiovascular cohort	9901	Cardiovascular outcome analysis	Randomised clinical trial validation	Cardiovascular outcomes reported	Dulaglutide reduced major cardiovascular events in patients with type 2 diabetes and cardiovascular risk factors.
42	[103]	Lincoff et al.	2023	Semaglutide	SELECT cardiovascular outcomes cohort	17604	Cardiovascular outcome analysis	Randomised clinical trial validation	Cardiovascular outcomes reported	Semaglutide significantly reduced major adverse cardiovascular events in adults with overweight or obesity and established cardiovascular disease without diabetes.
43	[13]	Rodriguez et al.	2024	Tirzepatide / Semaglutide	Real-world obesity cohort	18,386	Comparative effectiveness analysis	Observational validation	Weight-loss outcomes reported	Tirzepatide demonstrated greater real-world weight reduction than semaglutide during routine clinical follow-up.
44	[9]	Dennis et al.	2019	Diabetes therapies	Clinical trial dataset	NR	Treatment-response subgroup analysis	Clinical trial validation	Treatment-response outcomes reported	Baseline clinical characteristics influenced glycaemic-response variability across diabetes treatment subgroups.

The included studies demonstrated substantial methodological diversity across therapy types, data set sources, predictive objectives, modelling approaches, and clinical outcome measures [9–11]. Most of the investigations focused on the prediction of the response to treatment and the detection of non-response, particularly for *Semaglutide* and *Tirzepatide*, reflecting the growing clinical interest in personalised therapeutic strategies for the treatment of obesity and type 2 diabetes [10]. Frequently investigated outcomes included prediction of weight loss, glycaemic response, reduction in HbA1c, adherence to treatment, persistence, risk of discontinuation and forecast of adverse events [11, 14, 56]. A broad range of predictive models was identified in the reviewed literature. Ensemble learning methods, including Random Forest, Gradient Boosting Machines, and XGBoost, were among the most frequently implemented architectures due to their ability to model complex nonlinear relationships among demographic, behavioural, physiological, and treatment-related variables [9–11]. These approaches demonstrated strong predictive performance across heterogeneous healthcare datasets and reflected increasing demand for computational systems capable of modelling multifactorial treatment-response trajectories.

Interpretable statistical and survival-based methodologies also remained clinically relevant due to their transparency, temporal modelling ability, and ease of integration into clinical decision-support environments [11, 14, 56]. More recent investigations have increasingly incorporated explainability approaches such as SHAP and LIME to improve model transparency, attribute of characteristics, and trust of clinicians [9, 10]. This trend reflects the growing recognition that predictive performance alone is insufficient for integration into clinical workflows and that explainable AI is essential for reliable therapeutic decision support. The reviewed studies utilised diverse data sources, including electronic health records, clinical registries, pharmacy refill databases, clinical trial cohorts, observational healthcare datasets, and multi-omics resources [10, 11]. Although these datasets provide valuable information on longitudinal treatment behaviour and treatment-response prediction, substantial variability remained in dataset structure, follow-up duration, preprocessing procedures, feature engineering strategies and reporting practices, thereby limiting reproducibility and cross-study comparability [9].

Despite generally favourable predictive performance, relatively few studies incorporated external validation, temporal validation, or forward-looking clinical assessment frameworks [10, 11]. Consequently, many predictive models demonstrating strong retrospective performance may not generalise consistently across independent healthcare settings. Therefore, the limited adoption of robust validation strategies remains a major barrier to the development of clinically reliable and deployable AI-assisted therapeutic prediction systems for GLP-1 therapies.

To further synthesise the evidence extracted, the included studies were analysed according to the AI models used and the primary clinical outcomes investigated [9–11]. This analysis provides insight into dominant methodological trends and emerging research priorities within AI-driven studies of GLP-1 receptor agonist and incretin-based therapies. As shown previously in Table 3, semaglutide and tirzepatide-based investigations dominated the therapy categories represented in the reviewed literature. The most widely used predictive frameworks were ensemble-based methods, such as

RF, GBM, and XGBoost [9–11]. Their widespread adoption reflects the suitability of these approaches for modelling complex nonlinear relationships within heterogeneous clinical datasets.

Treatment-response prediction, non-response detection, weight-loss prediction, adherence modelling, and therapy optimisation were the most frequently investigated clinical outcomes [10, 11, 14, 56]. This distribution highlights the increasing emphasis on personalised treatment strategies and predictive decision-support systems for the management of obesity and type 2 diabetes. The reviewed studies also demonstrated a growing adoption of explainability approaches, including SHAP, LIME, and feature-importance analysis, to improve model transparency and support clinical interpretation of predictive outcomes [9, 10]. In general, the therapy distribution presented in Table 3 indicates that current research is increasingly focused on the optimisation of treatment, long-term therapeutic outcomes, and clinically interpretable therapeutic prediction systems for GLP-1 therapies. As shown in Table 4, validation practices varied considerably among included studies. Randomised clinical trial validation was the most frequently represented validation category, while real-world, retrospective, and observational validation approaches were less frequent. These findings highlight the continued need for independent, prospective and cross-institutional validation of predictive systems prior to routine clinical implementation.

To further evaluate the methodological robustness and clinical applicability, validation practices, explainability integration, andeployment-orienteded predictive models were examined in the included studies [9, 10]. Particular attention was paid to internal and external validation procedures, temporal and prospective clinical investigation strategies, and interpretable modelling approaches. This analysis provides insight into the methodological maturity and readiness for the adoption of healthcare care of current AI-assisted therapeutic prediction systems for GLP-1 therapies. Table 4 summarises the distribution of study-design and evidence-source categories recorded in the reviewed literature. Most of the investigations relied on retrospective datasets and internally validated modelling frameworks [10, 11]. Common validation approaches included train-test splits and cross-validation, while performance was typically assessed using accuracy, ROC-AUC, precision, recall, sensitivity, specificity and F1-score [9, 10]. Although many studies reported favourable predictive performance, comparatively few incorporated external validation, temporal validation, or future implementation study frameworks [10, 11].

Consequently, predictive models demonstrating strong retrospective performance may not generalise consistently in independent healthcare settings [104]. Therefore, the limited adoption of robust validation strategies remains a major barrier to the development of clinically reliable and deployable AI-assisted therapeutic prediction systems for GLP-1 therapies. The reviewed literature also demonstrated an increase in the adoption of explainability approaches and interpretable modelling methodologies. Several studies incorporated SHAP, LIME, feature-importance analysis, and survival-based modelling frameworks to improve transparency, clinician trust, and interpretability of therapeutic decision-support systems [9, 14, 56].

This trend indicates a growing awareness of the value of predictive AI; however, AI alone is not sufficient for real-world implementation, highlighting the importance

of explainability, as explainable AI plays a crucial role in providing reliable therapeutic decision support. Despite these advances, substantial variability remained in the composition of the data set, validation methodologies, feature-engineering procedures, predictive objectives, and reporting practices [10, 11]. These findings highlight the continued need for more standardised validation frameworks, harmonised reporting practices, and clinically robust benchmarking strategies to support the translation of predictive models into routine clinical care.

Table 10 Distribution of dataset-source categories across the 44 included studies. Dataset categories were derived from the dataset-source descriptions reported in Table 9.

Dataset Source	**Frequency**
Trial / emulated trial data	29
Clinical variables / structured data	6
Real-world data	6
Genomic / multi-omics data	3

To further investigate the data bases that support predictive methodologies in research of glucagon-like receptor agonist (GLP-1 RA) and incretin-based therapy, included studies were analysed according to data sets types and data sources used within predictive modelling frameworks [10, 11]. This analysis provides insight into the diversity and clinical relevance of datasets that support treatment-response prediction, adherence modelling, discontinuation analysis, and AI-assisted therapeutic optimisation. Table 10 summarises the distribution of dataset types in the reviewed literature. Under the current dadatasetlassification, trial or emulated-trial data formed the largest category (29 records), followed by clinical variables or structured data (6), real-world data (6), and genomic or multi-omics data (3). These categories describe the full set of mixed evidence and should not be interpreted as the distribution of the data set specifically between prediction studies. These datasets capture heterogeneous patient populations, routine clinical practice patterns, and long-term treatment outcomes associated with the treatment of obesity and type 2 diabetes. Several studies also incorporated pharmacy refill databases, observational registries, randomised clinical trial cohorts, genomic resources, and multi-omics datasets to support patient-specific therapeutic prediction and personalised treatment strategies [9, 10].

This trend reflects the increasing integration of diverse healthcare data sources within predictive modelling frameworks. Despite these advances, considerable variability remained in the structure of the data set, sample composition, follow-up duration, preprocessing procedures, feature-engineering strategies, and reporting practices [9]. Furthermore, many AI-enabled prediction systems continued to rely on retrospective and institution-specific datasets, with comparatively limited use of large-scale multicenter longitudinal cohorts [10, 11]. These limitations may reduce reproducibility,

external validity, and clinical translation into routine practice. In general, the distribution presented in Table 10 highlights the growing importance of routine clinical datasets in GLP-1 therapeutic prediction research while emphasising the need for more standardised reporting practices, transparent preprocessing methodologies and robust longitudinal healthcare datasets to support clinically reliable predictive models [9, 10].

3.6 Validation and Bias

The included studies were analysed to evaluate validation practices, methodological robustness, and potential sources of bias that can influence predictive performance, reproducibility, and clinical applicability [9, 10]. Particular attention was paid to validation methodologies, dataset characteristics, evaluation metrics, modelling strategies, and reporting transparency. Considerable variability was observed in validation approaches across the reviewed literature. Most studies relied on retrospective datasets together with internal validation procedures, including train-test splits and cross-validation [10, 11]. Although these approaches provide evidence of predictive capacity, the limited use of external, temporal, prospective, and multicenter validation frameworks remains a major barrier to clinical translation and deployment of predictive models for GLP-1 therapies [9–11].

For substantial differences at the outcome-level, substantial heterogeneity is also identified in performance. Frequently reported metrics included accuracy, ROC-AUC, precision, recall, sensitivity, specificity, and F1-score [9, 10]. However, inconsistencies in the selection of metrics, the reporting of calibration, and the sparency of evaluation limited the direct ccomparisons between therapeutic prediction systems ond reduced the consistency of benchmarks between studies. Dataset-related bias represented another important methodological limitation. Many predictive models were developed using institution-specific electronic health record datasets, observational cohorts, or limited population samples that may not adequately represent broader patient populations receiving GLP-1 therapies [10, 11]. Variability in sample size, demographic composition, outcome definitions, preprocessing procedures, and feature-engineering strategies further reduced comparability and external validity between studies [9].

Model-related bias was also evident in studies that implemented complex machine learning and deep learning architectures without sufficiently robust validation procedures or comparison with interpretable baseline models [9, 10, 14]. Limited methodological transparency and incomplete reporting of preprocessing workflows further increased the risk of overfitting and reduced interpretability. The reviewed literature also demonstrated an increase in the adoption of explainable AI approaches, including SHAP, LIME, and feature-importance analysis, to improve transparency and clinician trust in predictive models [9, 10]. Nevertheless, substantial variability in validation strategies, dataset quality, modelling methodologies, and reporting practices highlights the need for more rigorous validation frameworks, standardised evaluation protocols, and clinically generalisable predictive models for the optimisation of GLP-1 therapy.

3.6.1 Risk of Bias in Individual Studies

Several important sources of bias within the study were identified in the included predictive modelling investigations [9, 10]. One of the most frequently reported methodological challenges involved class imbalance, particularly in studies investigating treatment-response prediction versus non-response outcomes. In many datasets, the number of people in responder groups far outnumbered the number of non-responder/discontinuation/adverse-event cohorts, which could lead to a bias in the predictive performance and an overestimate of the accuracy of the methods to be developed. In order to overcome this drawback, a number of studies adopted data balancing techniques such as SMOTE, random oversampling, undersampling, and class-weight adjustment procedures [9]. Another major source of bias was that the cohort was not representative. The performance of the predictors and the generalisability of their results can be affected by demographic differences, level of disease activity, previous therapeutic experience, metabolic state, follow-up period and the presence of other diseases, etc. [10, 11].

Some investigations used a restrictive inclusion criterion to ensure internal consistency, but this may lead to selection bias and affect external generalisability. Many of these studies used electronic health records (EHR), clinical trial cohorts, observational registries, and pharmacy refill databases [9, 11] as a source of their datasets, which can also be problematic by introducing bias. Single-centre or institution-specific studies that examined small or specific subgroups of individuals were at high risk of representativeness bias, as they were subject to possible selection bias from other broader real-world populations receiving GLP-1 treatment [10, 11]. Concerning within-study bias, other factors would be related to insufficient methodological reporting and an insufficient number of validation procedures. Some studies lacked sufficient information on the pre-processing workflow, missing data treatment, feature engineering, or hyperparameter optimisation, or did not benchmark the calibration.

Some studies lacked enough details about the pre-processing process, how missing values were handled, feature engineering, hyperparameter optimisation, or calibration tests. In addition, large-scale and complex machine learning and deep learning architectures lacked robust external validation and comparisons with interpretable baseline models, which means that there was an increased possibility of overfitting and the likely overestimation of predictive performance [10, 14]. The second methodological issue was the interpretability of the model: Several studies focused on the predictive performance of AI models, but did not include explainability methods such as SHAP, LIME, or feature importance analysis [9, 10]. The lack of clarity in decision support processes can affect the use of predictive processes in clinical practice by decreasing the trust of clinicians.

3.6.2 Risk of Bias Across Studies

Several sources of bias were identified throughout the body of evidence included in this review. A major source of bias between studies was methodological heterogeneity. The included studies varied considerably in the sources of the data sets, patient populations, therapeutic agents, duration of follow-up, definition of outcomes and predictive

objectives [9–11]. This variability limits direct comparison of predictive performance and complicates evidence synthesis. Substantial variation was also observed in the validation strategies and performance evaluation practices. Although some studies used cross-validation, temporal validation, or external cohort validation, others relied solely on internal train-test splits [10]. Similarly, predictive performance was reported using various metrics, including accuracy, precision, recall, F1-score, ROC-AUC, hazard ratios and survival-based measures, reducing comparability between studies.

Dataset heterogeneity represented an additional source of bias between studies. The reviewed investigations used electronic health records, clinical trial datasets, observational registries, pharmacy refill databases, and genomic data sources. Differences in sample size, demographics of the patient, disease severity, treatment protocols, and follow-up periods may have influenced model performance and contributed to variability in reported outcomes [10, 11]. Across-study bias was further influenced by inconsistent reporting standards. Several studies provided limited information on preprocessing procedures, missing-data handling, feature engineering strategies, calibration assessment, and model deployment considerations [9].

These reporting inconsistencies reduce reproducibility and hinder rigorous comparison between therapeutic prediction systems. Finally, limited external validation remained a recurring issue across the literature. Although many studies reported favourable predictive performance, relatively few evaluated their models using independent cohorts or multi-centre datasets [10, 13]. Consequently, the generalisability and clinical transferability of many reported predictive models remain uncertain. Overall, these findings highlight the need for greater methodological standardisation, transparent reporting practices, harmonised evaluation frameworks, and robust external validation to improve the reliability, reproducibility, and clinical applicability of predictive models for GLP-1 therapy optimisation.

4 Discussion

This systematic review demonstrates the growing role of AI, ML, deep learning (DL), and predictive analytics in supporting personalised treatment strategies for glucagon-like peptide-1 receptor agonist (GLP-1 RA) and incretin-based therapies. Across the reviewed literature, predictive models were increasingly applied to treatment-response prediction, non-response detection, adherence modelling, treatment persistence, discontinuation forecasting, adverse-event prediction, and therapy optimisation. Collectively, these developments reflect a broader shift towards precision medicine approaches in obesity and type 2 diabetes management, where therapeutic decisions are increasingly informed by patient-specific clinical, behavioural, and metabolic characteristics rather than population-level treatment assumptions [9–11].

Despite these advances, the review identified several methodological and translational challenges that continue to limit routine clinical implementation. Although many studies reported favourable predictive performance, substantial heterogeneity remained across datasets, modelling methodologies, validation practices, outcome definitions, and reporting standards. In particular, limited use of external validation,

clinical effectiveness studies, and multi-centre assessment frameworks raises important concerns regarding generalisability and real-world applicability across diverse healthcare environments [10, 11, 13].

The 44 included studies encompassed a broad spectrum of predictive modelling investigations, real-world observational cohorts, clinical trial analyses, pharmacogenomic studies, adherence and persistence frameworks, explainability-focused approaches, and translational validation studies. These investigations evaluated semaglutide, tirzepatide, liraglutide, dulaglutide, and other incretin-based therapies across obesity and type 2 diabetes populations [9–11, 13, 14, 32, 33, 55–58, 77, 84, 100–102]. Consequently, the findings should be interpreted as a comprehensive synthesis of predictive treatment-response research rather than evidence derived from a single methodological paradigm. The diversity of study designs and clinical objectives provides a broad perspective on the current state of AI-assisted therapeutic prediction while simultaneously highlighting the need for greater methodological standardisation and clinically oriented validation strategies.

4.1 Dataset Characteristics and Feature Engineering

The reviewed literature demonstrated substantial diversity in dataset composition, feature-engineering strategies, and clinical variable selection across AI-enabled prediction systems for GLP-1 therapies. Most studies relied on structured clinical variables routinely available within electronic health records (EHRs), including body mass index (BMI), HbA1c levels, age, sex, baseline metabolic characteristics, medication history, treatment duration, and comorbidity profiles [9–11]. These variables formed the foundation of most predictive models because of their clinical relevance and widespread availability within routine healthcare systems. The predominance of retrospective EHR-derived datasets reflects the increasing availability of large-scale digital healthcare repositories and the growing demand for data-driven therapeutic decision support. However, reliance on retrospective and institution-specific datasets introduces important limitations, including restricted demographic diversity, heterogeneous treatment protocols, incomplete adherence information, and variable follow-up duration [10, 11, 13]. These factors may affect model calibration, reproducibility, external validity, and generalisability across healthcare environments.

Several studies highlighted the value of longitudinal real-world datasets for understanding treatment-response heterogeneity across obesity and type 2 diabetes populations. Longitudinal clinical records, pharmacogenomic variability, and patient-specific treatment characteristics were identified as important contributors to therapeutic stratification and treatment-response prediction [10, 13, 14, 57, 58]. Extended follow-up periods were particularly important for evaluating weight-loss sustainability, glycaemic improvement, treatment persistence, and discontinuation behaviour. Considerable variability was observed across preprocessing and feature-engineering methodologies. Common approaches included missing-value imputation, feature scaling, dimensionality reduction, feature-selection techniques, SMOTE-based class balancing, and hyperparameter optimisation workflows [9, 11]. Although these procedures frequently improved predictive performance, they also complicated direct

benchmarking because performance gains may partly reflect preprocessing strategies rather than modelling architectures alone.

The reviewed studies further demonstrated increasing interest in multimodal predictive models that integrate EHR-derived variables with laboratory biomarkers, pharmacy refill records, behavioural information, wearable-device data, and longitudinal treatment trajectories [10, 11, 14]. This trend reflects a broader movement towards personalised therapeutic prediction systems capable of modelling complex metabolic interactions and patient-specific treatment behaviour. Despite these advances, no standardised feature-engineering framework currently exists for GLP-1 treatment-response prediction. Variability in cohort selection, feature extraction, outcome definitions, temporal follow-up structures, and preprocessing methodologies continues to limit reproducibility and cross-study comparability. Importantly, the findings suggest that dataset quality and representativeness may exert a greater influence on predictive reliability than modelling complexity alone. Future research should therefore prioritise large-scale longitudinal datasets, multi-centre validation, pharmacogenomic integration, and multi-omics therapeutic stratification frameworks to improve patient-specific treatment selection and treatment-response prediction [32, 33].

4.2 Predictive Modelling Approaches

The reviewed literature demonstrated substantial diversity in predictive modelling architectures applied to GLP-1 receptor agonist therapeutic prediction. Ensemble learning approaches, particularly Random Forest (RF), Gradient Boosting Machines (GBM), and Extreme Gradient Boosting (XGBoost), emerged as the most frequently implemented methodologies and were consistently associated with strong predictive performance across treatment-response prediction, glycaemic-response modelling, adherence estimation, persistence forecasting, and personalised prescribing applications [9–11]. The predominance of ensemble learning methods likely reflects their ability to capture complex nonlinear relationships among demographic, behavioural, metabolic, pharmacological, and treatment-related variables. Compared with conventional statistical models, ensemble approaches are generally better suited to heterogeneous healthcare datasets characterised by missing values, variable follow-up periods, and multifactorial treatment-response mechanisms. Their flexibility and robustness make them particularly valuable for modelling the complex biological and behavioural interactions that influence therapeutic outcomes in obesity and type 2 diabetes management.

Despite these advantages, improved predictive performance often comes at the expense of transparency. Many ensemble learning frameworks operate as partially opaque decision systems, limiting clinician understanding of how individual predictions are generated. This trade-off between predictive accuracy and clinical transparency was a recurring theme across the reviewed literature and represents a key challenge in the development of clinically deployable therapeutic decision-support systems. Consequently, recent studies increasingly incorporated explainable AI approaches such as SHAP, LIME, and feature-importance analysis to improve transparency, support clinical interpretation, and enhance trust in AI-assisted treatment recommendations [9, 10]. Traditional statistical approaches, including LR and

Cox proportional hazards regression, also remained clinically relevant because of their transparency, interpretability, and suitability for time-to-event analyses [11, 14, 56]. Although these methods generally provide lower modelling flexibility than ensemble learning approaches, they remain valuable in healthcare environments where interpretability and ease of implementation are prioritised.

The reviewed studies also highlighted the growing interest in deep learning and multimodal prediction frameworks capable of integrating electronic health records, laboratory biomarkers, behavioural information, genomic data, and longitudinal treatment trajectories. These approaches offer considerable potential for capturing complex treatment-response patterns and supporting personalised therapeutic decision-making. However, challenges related to explainability, robustness of validation, workflow integration, and translational readiness continue to limit widespread clinical adoption. Collectively, the findings suggest that future progress in the prediction of GLP-1 therapeutics will depend not only on improving predictive performance, but also on achieving an effective balance between precision, interpretability, robustness of validation and clinical maturity. Such advances will be essential for translating predictive modelling frameworks into reliable and clinically actionable decision-support systems for the management of obesity and type 2 diabetes.

4.2.1 Ensemble Learning Approaches

Ensemble learning approaches demonstrated strong flexibility and robustness when applied to retrospective electronic health record datasets characterised by missing values, heterogeneous patient populations, irregular follow-up periods, and multifactorial treatment-response mechanisms. Compared to conventional statistical methodologies, the ensemble models were generally better able to capture complex nonlinear interactions among demographic, behavioural, metabolic, pharmacological, and clinical variables associated with GLP-1 receptor agonist therapeutic outcomes [9–11]. Several studies also incorporated feature-selection procedures, hyperparameter optimisation, class-balancing techniques, and Synthetic Minority Over-sampling Technique (SMOTE) approaches to improve predictive robustness and classification performance in imbalanced healthcare datasets [9, 11]. The increasing dominance of XGBoost and related gradient boosting methodologies likely reflects their ability to balance predictive performance, computational efficiency, and relative interpretability. These models appeared particularly effective in predicting treatment-response prediction, weight-loss outcomes, and glycaemic improvement in clinically heterogeneous populations of obesity and type 2 diabetes [9, 10]. Their strong performance may also be attributed to their ability to model the considerable variability observed in therapeutic outcomes among patients receiving semaglutide, tirzepatide, and other incretin-based therapies [55, 100, 101].

Despite these advantages, most ensemble learning studies remained dependent on retrospective datasets, institution-specific cohorts, and limited external validation procedures [10, 13]. Consequently, concerns remain regarding reproducibility, generalisability, and maturity of implementation in broader healthcare settings. Nevertheless, the prevalence of ensemble learning approaches in the reviewed literature suggests that they currently represent the most practical and effective modelling paradigm for

the prediction of GLP-1 therapeutics, particularly within structured clinical datasets commonly encountered in routine healthcare practice. The prevalence of ensemble learning methodologies in the reviewed literature also reflects the structured and tabular nature of most healthcare datasets used in obesity and type 2 diabetes prediction research. Compared to deep neural architectures, ensemble models generally required smaller training datasets, demonstrated greater robustness to heterogeneous and partially incomplete clinical variables, and maintained stronger predictive stability within retrospective electronic health record environments [9–11]. Their comparatively lower computational complexity and improved interpretability may further contribute to their practical suitability for clinically orientated therapeutic prediction and decision-support applications.

4.2.2 Deep Learning Architectures

Deep learning methodologies, including deep neural networks (DNNs), recurrent neural networks, and multimodal representation-learning frameworks, were increasingly explored to model high-dimensional electronic health records, behavioural monitoring data, longitudinal treatment trajectories, and multimodal metabolic datasets [9, 14]. These architectures demonstrated considerable potential for capturing complex latent relationships and dynamic treatment-response patterns that may not be adequately represented using conventional machine learning approaches. Despite this potential, deep learning models did not consistently outperform ensemble learning methodologies in all reviewed studies. This finding may reflect limitations associated with relatively small cohort sizes, fragmented longitudinal datasets, heterogeneous outcome definitions, class imbalance, and variable temporal standardisation in retrospective healthcare records [9–11]. In many therapeutic prediction settings, the complexity of deep neural architectures can exceed the informational content available within existing datasets, thereby limiting reproducible performance gains over ensemble learning approaches.

Several investigations reported only modest improvements in predictive performance despite substantially greater computational complexity and reduced interpretability [9, 10, 14]. Consequently, challenges related to explainability, validation robustness, workflow integration, and clinician trust continue to constrain the routine clinical adoption of deep learning frameworks within obesity and type 2 diabetes management environments. The reviewed evidence suggests that deep learning approaches may be particularly beneficial for multimodal and longitudinal prediction tasks that involve behavioural, genomic, physiological, and treatment-response data. Such frameworks are well suited to modelling temporally evolving therapeutic trajectories and complex patient-specific interactions that cannot be readily captured using conventional static prediction methods. However, stronger evidence from independent cohort validation studies, prospective evaluation frameworks, and real-world implementation research remains necessary before deep learning systems can be considered clinically deployable for routine GLP-1 therapeutic decision support [9, 10, 13, 14].

4.2.3 Emerging Methodological Trends

The reviewed studies demonstrated an increasing integration of XAI (XAI) methodologies, including SHAP, LIME, feature-attribution analysis, and interpretable machine learning frameworks, reflecting the growing recognition that predictive precision alone is insufficient for the clinically reliable implementation of AI-assisted therapeutic decision-support systems [9, 10, 57, 77]. These approaches were increasingly used to identify influential metabolic, behavioural, pharmacogenomic and treatment-related variables associated with prediction of treatment response, adherence, persistence of treatment and risk of adverse events [55, 84, 102]. Several investigations also suggested that the integration of pharmacogenomic variability and genotype-informed therapeutic stratification may improve both interpretability and patient-specific treatment optimisation within incretin-based therapeutic prediction systems [10, 32, 33].

Despite increasing methodological interest in explainable AI, routine clinical integration remained comparatively limited in the reviewed literature. In many studies, SHAP and LIME were used primarily as post-hoc interpretation tools rather than as clinician-facing decision-support systems [9–11, 57, 77]. Furthermore, several investigations provided limited discussion on the reliability of explanations, the transparency of the calibration, the usability for clinicians and the integration of the workflow. Consequently, although explainability methodologies improve transparency and support model interpretation, their practical role in routine obesity and type 2 diabetes management remains not sufficiently established. The persistent gap between methodological explanation and clinical implementation highlights the need for more robust clclinician-orientednterpretability frameworks, prospective validation studies, and workflow-integrated decision-support systems. Future research should focus not only on generating explanations, but also on evaluating their reliability, clinical usefulness, and impact on therapeutic decision-making. Such advances will be essential for translating explainable AI methodologies into trustworthy and clinically actionable clinical decision-support systems for GLP-1 therapies.

4.2.4 Methodological Opportunities and Evidence Gaps

The current synthesis identifies questions for future treatment-response prediction research rather than establishing a chronological transition towards more advanced or clinically deployed AI systems. The available extraction does not support reliable estimates of the adoption of explainable AI, temporal architectures or multimodal prediction methods across the included literature.

Longitudinal modelling is a potential research direction for outcomes such as treatment persistence, discontinuation and changes in response over time. However, longitudinal follow-up or analysis of treatment outcomes should not be interpreted as evidence that a study developed a temporal prediction model. Future evaluations should distinguish descriptive trajectory analyses, time-to-event prediction and dynamically updated predictions, with methods selected according to the intended clinical task.

Explainability and multimodal data integration also warrant evaluation. Their contribution should be established through comparisons with appropriate baseline

Table 11 Proposed AI-TRF dimensions and evidence requirements. The framework is conceptual and has not been validated as a scoring instrument.

Dimension	Evidence to examine
External validation	Evaluation of a fixed model in data independent of model development and tuning, including assessment of predictive performance and calibration.
Explainability and interpretability	Description of the explanation approach and evidence of its stability, clinical comprehensibility and usefulness for the intended users.
Temporal modelling	Appropriate handling of prediction time, follow-up, time-varying information and potential temporal leakage, where relevant to the intended prediction task.
Prospective evaluation	Evaluation of a prespecified model on subsequently accrued patients or in a prospective clinical study, distinguished from retrospective analysis of observational data.
Clinical implementation	Evidence concerning workflow integration, usability, monitoring, human oversight, and effects on clinical decisions or patient outcomes.

models and assessment of predictive performance, calibration and clinical usefulness. The use of SHAP, LIME or multiple data modalities alone does not demonstrate improved treatment selection or readiness for clinical implementation.

Clinical translation requires evaluation beyond retrospective model development. Future studies should examine performance in independent populations, prospective use, workflow integration and effects on clinical decisions. The proposed AI-TRF organises these considerations as conceptual assessment dimensions; it does not establish that the included studies have achieved them.

4.3 Artificial Intelligence Translational Readiness Framework (AI-TRF)

The Artificial Intelligence Translational Readiness Framework (AI-TRF) is proposed as a conceptual structure for organising evidence relevant to the clinical translation of treatment-response prediction models. It comprises five dimensions: external validation, explainability and interpretability, temporal modelling, prospective evaluation, and clinical implementation.

AI-TRF was not applied to individual studies in this review. Accordingly, no numerical readiness scores, thresholds or study rankings are reported. Table 11 outlines the proposed dimensions and the evidence relevant to their assessment.

The dimensions are not assumed to form a fixed sequence. Interpretability does not establish external validity, and temporal modelling is not necessary for every clinical prediction task. For example, a static model may be appropriate for treatment selection using baseline patient characteristics. Clinical readiness therefore requires

evidence appropriate to the model's intended use rather than the accumulation of methodological features.

AI-TRF is intended to complement established reporting guidance and risk-of-bias assessment. Its reliability, practical usefulness and incremental contribution have not been established. These properties require evaluation through expert consultation and independent application to verified prediction studies.

4.4 Future Research Priorities and Clinical Translation

Future research should focus on the development of large-scale, longitudinal, and multimodal predictive frameworks that integrate electronic health records, behavioural variables, treatment-adherence patterns, wearable-device monitoring, laboratory trajectories, pharmacy refill records, pharmacogenomic information, and multi-omics datasets [105]. Such integration may improve predictive robustness, temporal modelling capability, and patient-specific therapeutic guidance across obesity and type 2 diabetes management environments [10, 11, 14, 32, 33]. In parallel, more advanced temporal modelling approaches capable of capturing dynamic treatment trajectories, adherence behaviour, discontinuation risk, and longitudinal metabolic adaptation should be explored. Potential methodologies include survival-analysis models, recurrent neural networks, transformer-based architectures, and longitudinal ensemble-learning frameworks, which may provide a more comprehensive representation of evolving therapeutic responses over time [9, 11].

Future methodological progress will require stronger alignment between predictive architectures, dataset quality, validation robustness, and clinical implementation requirements. Although ensemble learning approaches such as Random Forest, Gradient Boosting Machines, and XGBoost consistently demonstrated strong predictive performance across heterogeneous therapeutic prediction tasks [9, 10], future systems should increasingly incorporate explainability, calibration transparency, interoperability, workflow compatibility, and clinician-oriented decision-support functionality. Greater emphasis should also be placed on multi-centre validation, temporal validation, forward-looking clinical assessment, and multi-centre benchmarking frameworks to improve reproducibility, generalisability, and clinical applicability across diverse healthcare environments [106].

Emerging methodologies, including federated learning, causal AI, foundation models, and continuous patient-monitoring frameworks, may further enhance privacy-preserving scalability, transferability, and personalised therapeutic stratification across healthcare systems [9–11]. These approaches offer opportunities to address several data-sharing, interoperability, and generalisability challenges that currently limit large-scale routine healthcare implementation of predictive therapeutic modelling systems. Future research should also move beyond retrospective treatment-response prediction towards adaptive therapeutic guidance frameworks capable of supporting early non-response detection, adherence forecasting, treatment persistence modelling, discontinuation-risk estimation, adverse-event prediction, and personalised incretin-based treatment selection [11, 14, 55]. Importantly, prospective clinical investigation studies are needed to determine whether AI-assisted treatment recommendations

improve clinical outcomes, therapeutic adherence, and healthcare decision-making in real-world practice.

The conceptual Artificial Intelligence Translational Readiness Framework (AI-TRF) proposed in this review highlights several dimensions that may support future evaluation of therapeutic prediction systems, including explainability, validation robustness, temporal modelling capability, future implementation studies, and clinical implementation readiness. Future research should further investigate these dimensions across diverse healthcare settings to improve assessment of translational maturity, enhance model reproducibility and generalisability, and support the development of clinically applicable AI-enabled decision-support systems for GLP-1 receptor agonist therapies [107].

4.5 Limitations

Several limitations should be considered when interpreting the findings of this systematic review. First, substantial methodological heterogeneity existed in all included studies with respect to dataset composition, outcome definitions, predictive objectives, feature-engineering strategies, modelling architectures, validation procedures, and evaluation metrics [9, 10]. This variability limited direct comparison of predictive models and precluded formal meta-analysis or pooled estimation of predictive performance. Furthermore, differences in electronic health record structures, treatment protocols, duration of follow-up and reporting practices may have affected the comparability and reproducibility of the cross-study [108].

Second, many studies relied on retrospective institution-specific datasets with limited demographic diversity and variable follow-up periods [10, 11]. External validation, temporal validation, prospective evaluation, and cross-institutional benchmarking remained relatively uncommon in the reviewed literature, potentially limiting model generalisability and reducing confidence in real-world clinical applicability [34, 36]. Consequently, many prediction systems currently reported remain closer to exploratory research tools than clinically deployable decision-support systems [43].

Third, reporting transparency varied considerably between the included studies. Several investigations provided limited information on data preprocessing pipelines, missing-data handling, feature-selection procedures, calibration assessment, hyperparameter optimisation, and explainability integration [36, 41]. Such reporting limitations may affect reproducibility, independent validation, and detailed methodological comparison. In addition, some recently published studies were included to capture emerging developments in AI-assisted therapeutic prediction despite providing comparatively limited methodological detail.

Another limitation relates to the quality assessment methodology and the proposed Artificial Intelligence Translational Readiness Framework (AI-TRF). A customised assessment framework was adopted because existing tools such as PROBAST, PROBAST-AI, QUADAS-AI, and TRIPOD-AI [34, 36, 41, 42] were not specifically designed to evaluate the broad methodological diversity of machine learning, explainable AI, longitudinal modelling, and translational-readiness studies included in this review. Although this approach facilitated a consistent comparative assessment in

heterogeneous study designs, AI-TRF should be regarded as a conceptual framework rather than a formally validated assessment instrument. Critically, AI-TRF was not applied to score any individual study included in this review: no per-study AI-TRF scores, score distribution, or inter-rater reliability statistic are reported, and the framework is presented here as a proposed structure for future assessment rather than as a result of this review's own analysis. The framework was derived from recurring methodological patterns identified across the reviewed literature and has not yet undergone formal validation through expert consensus, inter-rater reliability assessment, or prospective evaluation. AI-TRF has not been positioned against other translational-readiness and reporting frameworks relevant to clinical AI, including DECIDE-AI, CONSORT-AI, SPIRIT-AI, MI-CLAIM, or FUTURE-AI; a rigorous comparison against these frameworks is necessary before any claim that AI-TRF adds value beyond existing instruments can be substantiated. Future research should therefore assess its reliability, reproducibility, and generalisability across diverse clinical AI applications, including formal per-study scoring with documented inter-rater agreement.

The search strategy also represents a potential limitation. Although multiple databases and reference-list screening were used to maximise coverage, only studies indexed in PubMed, Scopus, and Web of Science were included. Consequently, relevant studies indexed in specialised engineering, computer science, or health informatics databases may not have been captured. In addition, studies using alternative terminology such as prognostic modelling, treatment effectiveness, precision medicine, or personalised medicine may have been missed despite the comprehensive search strategy employed [37]. This review was not prospectively registered (e.g., with PROSPERO); the protocol was not finalised before screening began, which is a limitation independent of whether a comparative evaluation of clinical interventions was performed. Database-specific search dates and per-database yield counts, and a completed PRISMA 2020 checklist are not reported here and should accompany any future submission or revision of this work. The title and abstract screening stage did not result in the exclusion of any records prior to full-text evaluation (80 records were screened, and 80 were subsequently assessed at full text); this is noted here rather than obscured, and readers should interpret the screening numbers accordingly. Screening and data extraction were performed by a single reviewer. Consequently, independent duplicate assessment was unavailable and inter-reviewer agreement could not be estimated.

Six of the 50 initially selected records were excluded during bibliographic verification, leaving 44 records in the final synthesis. The first five exclusions concern a cardiometabolic predictor study attributed to Ma łecki et al. (2022), a UK semaglutide obesity cohort attributed to Capehorn et al. (2020), a longitudinal persistence analysis attributed to Buse et al. (2022), a PIONEER-3 extension analysis attributed to Ahrén et al. (2021), and a real-world observational study of PIONEER attributed to Seufert et al. (2021). A matching primary source could not be established for these records as described. A sixth record, attributed to Apostolopoulou et al. (2023), was excluded because its supplied DOI identified a different publication and the intended source could not be confirmed. These exclusions refer to the specific records as originally described and do not imply that other publications by these authors are invalid.

A further limitation concerns the composition of the evidence base. The 44 included records comprise prediction and predictor analyses in conjunction with clinical efficacy studies and other evidence for treatment-response. These categories have not yet been consistently separated using verified study-level definitions. Consequently, the total should not be interpreted as 44 prediction model studies, and frequencies calculated across the entire evidence set cannot establish the methodological maturity of prediction models specifically. In particular, a randomised clinical trial design does not itself constitute validation of a prediction model. Separate classification and synthesis of prediction studies and supporting clinical evidence are required before conclusions about model performance, validation practices, or clinical readiness can be substantiated.

Finally, this review focused specifically on predictive analytics applications involving GLP-1 receptor agonists and incretin-based therapies. Therefore, more general pharmacological, mechanistic, or non-predictive investigations were excluded. Moreover, the rapidly evolving nature of AI-enabled healthcare research means that newly emerging multimodal learning architectures, foundation models, large language models, and clinically integrated decision-support systems can continue to reshape the evidence landscape beyond the review period considered in this study [67, 68, 109]. Despite these limitations, the review provides a comprehensive synthesis of current AI and machine-learning approaches for the prediction of response to treatment in GLP-1 therapies, while identifying important methodological trends, translational challenges, and future research priorities relevant to the development of robust, interpretable, externally validated, and clinically deployable therapeutic decision-support systems for the management of obesity and type 2 diabetes.

5 Conclusions

This review brings together 44 records on the prediction of response to treatment and related clinical evidence in glucagon-like receptor agonist and incretin-based therapies. The evidence set includes prediction and predictor analyses in conjunction with clinical efficacy studies and other treatment-response investigations. Consequently, the total should not be interpreted as 44 AI prediction systems developed or validated.

The current extraction does not support a reliable ranking of predictive algorithms or quantified conclusions about external validation, explainability uptake, or clinical deployment. Addressing these questions requires verified extraction at the study-level, explicit classification of study objectives, and separate synthesis of prediction studies and supporting clinical evidence.

The Artificial Intelligence Translational Readiness Framework (AI-TRF) is proposed as a conceptual structure that encompasses external validation, explainability, and interpretability, temporal modelling, prospective evaluation, and clinical implementation. It was not applied to individual studies and has not been validated as an assessment instrument. Its usefulness and incremental contribution require evaluation alongside established reporting and appraisal guidance.

Future research should prioritise reproducible evaluation, calibration assessment, independent validation, and prospective assessment of clinical usefulness. These

steps are necessary to determine whether treatment-response prediction systems can support personalised therapeutic decisions in routine practice.

Abbreviations

Abbreviation	Definition	Abbreviation	Definition
AI	Artificial Intelligence	ML	Machine Learning
DL	Deep Learning	DNN	Deep Neural Network
RF	Random Forest	DT	Decision Tree
GBM	Gradient Boosting Machine	XGBoost	Extreme Gradient Boosting
SVM	Support Vector Machine	KNN	k-Nearest Neighbours
LR	Logistic Regression	GBMs	Gradient Boosting Machines
CHR	Cox Hazard Regression	XAI	Explainable Artificial Intelligence
SHAP	SHapley Additive exPlanations	LIME	Local Interpretable Model-agnostic Explanations
SMOTE	Synthetic Minority Oversampling Technique	ROC	Receiver Operating Characteristic
AUC	Area Under the Curve	ROC-AUC	Area Under the Receiver Operating Characteristic Curve
EHRs	Electronic Health Records	BMI	Body Mass Index
HbA1c	Glycated Haemoglobin	GLP-1 RA	Glucagon-Like Peptide-1 Receptor Agonist
GIP	Glucose-Dependent Insulinotropic Polypeptide	T2DM	Type 2 Diabetes Mellitus
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses	QA	Quality Assessment
AI-TRF	Artificial Intelligence Translational Readiness Framework	LLMs	Large Language Models

Acknowledgements. Not applicable.

Declarations

- **Funding:** The authors received no specific funding for this work.
- **Conflict of interest/Competing interests:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- **Ethics approval and consent to participate:** Not applicable.
- **Clinical trial number:** Not applicable.
- **Consent for publication:** Not applicable.

- **Data availability:** All data analysed in this study are included in this published article and its references.
- **Materials availability:** Not applicable.
- **Code availability:** Not applicable.

CRediT authorship contribution statement

Tahsien Al-Quraishi: Conceptualisation, Methodology, Literature search, Data extraction, Formal analysis, Framework development, Visualisation, Writing – original draft. **O. S. Albahri**, **Jemal Abawajy**, and **Raza Nowrozy:** Methodological review, Writing – review & editing. All authors read and approved the final manuscript.

References

[1] Davies, M.J., Lim, S., Slater, T., Goldney, J., Philis-Tsimikas, A., Franco, D.R., Lamptey, R., Yates, T., Tankova, T., Lingvay, I.: Type 2 diabetes mellitus. Nature Reviews Disease Primers **12**(1), 13 (2026)

[2] Fenta, E.T., Eshetu, H.B., Kebede, N., Bogale, E.K., Zewdie, A., Kassie, T.D., Anagaw, T.F., Mazengia, E.M., Gelaw, S.S.: Prevalence and predictors of chronic kidney disease among type 2 diabetic patients worldwide, systematic review and meta-analysis. Diabetology & Metabolic Syndrome **15**(1), 245 (2023)

[3] Fernando, K., Connolly, D., Darcy, E., Evans, M., Hinchliffe, W., Holmes, P., Strain, W.D.: Advancing cardiovascular, kidney, and metabolic medicine: a narrative review of insights and innovations for the future. Diabetes Therapy **16**(6), 1155–1176 (2025)

[4] Sun, P., Wan, Z., Liu, Y.: Towards comprehensive chronic disease control: Prevention, early detection, and integrated management. Medicine Bulletin (2026)

[5] Podder, D., Stala, O., Miah, A., Agyapong, A., Moore, M.E., Hirani, R., Diegisser, D., Garcia, V., Etienne, M.: Incretin-based multi-agonist therapies for type 2 diabetes mellitus and obesity: mechanisms, clinical efficacy, and future directions. Diabetology **7**(3), 46 (2026)

[6] Alharbi, A.G.: Glp-1 receptor agonism: a transformative approach for managing type-2 diabetes and obesity. Saudi Pharmaceutical Journal **33**(5), 34 (2025)

[7] Alluri, A.A., Guntupalli, Y., Suvarna, S.S., Prystupa, Y., Khetan, S.P., Vejandla, B., Babu Swathi, N.L.: Incretin-based therapies: advancements, challenges, and future directions in type 2 diabetes management. Journal of Basic and Clinical Physiology and Pharmacology **36**(2-3), 95–111 (2025)

[8] Laddu, D., Neeland, I.J., Carnethon, M., Stanford, F.C., Mongraw-Chaffin, M., Barone Gibbs, B., Ndumele, C.E., Longenecker, C.T., Chung, M.L., Rao, G., *et al.*: Implementation of obesity science into clinical practice: A scientific statement from the american heart association. Circulation **150**(1), 7–19 (2024)

[9] Dennis, J.M., Shields, B.M., Henley, W.E., Jones, A.G., Hattersley, A.T.: Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. The Lancet Diabetes & Endocrinology **7**(6), 442–451 (2019) https://doi.org/10.1016/S2213-8587(19)30087-7

[10] Wilding, J.P.H., Batterham, R.L., Calanna, S., *et al.*: Once-weekly semaglutide in adults with overweight or obesity. New England Journal of Medicine **384**(11), 989–1002 (2021) https://doi.org/10.1056/NEJMoa2032183

[11] McGovern, A., Tippu, Z., Hinton, W., Munro, N., Whyte, M., Lusignan, S.: Comparison of medication adherence and persistence in type 2 diabetes: a systematic review and meta-analysis. Diabetes, Obesity and Metabolism **20**(4), 1040–1043 (2018) https://doi.org/10.1111/dom.13160

[12] Psaltis, J.P., Marathe, J.A., Nguyen, M.T., Le, R., Bursill, C.A., Marathe, C.S., Nelson, A.J., Psaltis, P.J.: Incretin-based therapies for the management of cardiometabolic disease in the clinic: Past, present, and future. Medicinal Research Reviews **45**(1), 29–65 (2025)

[13] Rodriguez, P.J., Goodwin Cartwright, B.M., Gratzl, S., *et al.*: Semaglutide vs tirzepatide for weight loss in adults with overweight or obesity. JAMA Internal Medicine **184**(9), 1056–1064 (2024) https://doi.org/10.1001/jamainternmed.2024.2525

[14] Davies, M., Faerch, K., Jeppesen, O.K., *et al.*: Semaglutide 2.4 mg once a week in adults with overweight or obesity, and type 2 diabetes (step 2): a randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. The Lancet **397**(10278), 971–984 (2021) https://doi.org/10.1016/S0140-6736(21)00213-0

[15] Thottunkal, S., Spahn, C., Wang, B., Rohatgi, N., Hong, J., Khandelwal, A., Palaniappan, L.: Clinician experiences at the frontier of pharmacogenomics and future directions. Journal of Personalized Medicine **15**(7), 294 (2025)

[16] Thapa, K., Khan, H., Kanojia, N., Kumar, J., Singh, T.G., Grewal, A.K.: Update on pharmacogenomic approaches in the treatment of autism spectrum disorders. Psychopharmacology, 1–12 (2026)

[17] Cardoso, P., Dennis, J.M., Pearson, E.R., et al.: Glp-1ra precision medicine in people with type 2 diabetes: current insights and future prospects. The Journal of Clinical Investigation **136**(2) (2026)

[18] Van Dellen, E.: Precision psychiatry: predicting predictability. Psychological medicine **54**(8), 1500–1509 (2024)

[19] Mengistu, A.K., Yeneakale, K.A., Baykemagn, N.D., Melese, Z.Y., Gedefaw, A.E.: Application of causal forest double machine learning (dml) approach to assess tuberculosis preventive therapy's impact on art adherence. Scientific Reports **15**(1), 29130 (2025)

[20] Jalleh, R.J., Rayner, C.K., Hausken, T., Jones, K.L., Camilleri, M., Horowitz, M.: Gastrointestinal effects of glp-1 receptor agonists: mechanisms, management, and future directions. The lancet Gastroenterology & hepatology **9**(10), 957–964 (2024)

[21] Giddings, R., Joseph, A., Callender, T., Janes, S.M., Schaar, M., Sheringham, J., Navani, N.: Factors influencing clinician and patient interaction with machine learning-based risk prediction models: a systematic review. The Lancet Digital Health **6**(2), 131–144 (2024)

[22] De Silva, K., Lee, W.K., Forbes, A., Demmer, R.T., Barton, C., Enticott, J.: Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. International Journal of Medical Informatics **143**, 104268 (2020) https://doi.org/10.1016/j.ijmedinf.2020.104268

[23] Fregoso-Aparicio, L., Noguez, J., Montesinos, L., Garcia-Garcia, J.A.: Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. Diabetology and Metabolic Syndrome **13**(1), 148 (2021) https://doi.org/10.1186/s13098-021-00767-9

[24] Tuppad, A., Patil, H.: Machine learning for diabetes clinical decision support: A review. Journal of Diabetes Science and Technology (2022) https://doi.org/10.1177/19322968221092087

[25] Olusanya, M.O., *et al.*: Accuracy of machine learning classification models for the prediction of type 2 diabetes mellitus: A systematic survey and meta-analysis approach. International Journal of Environmental Research and Public Health **19**(21), 14280 (2022) https://doi.org/10.3390/ijerph192114280

[26] Ferreras, A., Sumalla-Cano, S., Martinez-Licort, R., *et al.*: Systematic review of machine learning applied to the prediction of obesity and overweight. Journal of Medical Systems **47**(1), 8 (2023) https://doi.org/10.1007/s10916-022-01904-1

[27] Petridis, P.D., *et al.*: A review on trending machine learning techniques for type 2 diabetes mellitus management. Informatics **11**(4), 70 (2024) https://doi.org/10.3390/informatics11040070

[28] Khokhar, P.B., Gravino, C., Palomba, F.: Advances in artificial intelligence for

diabetes prediction: Insights from a systematic literature review. arXiv (2024) 2412.14736

[29] Haghighathoseini, A., *et al.*: A systematic review on applications of artificial intelligence in obesity prevention. Obesity Reviews (2025) https://doi.org/10.1111/obr.70062

[30] Li, Y., Jin, N., Zhan, Q., *et al.*: Machine learning-based risk predictive models for diabetic kidney disease in type 2 diabetes mellitus patients: a systematic review and meta-analysis. Frontiers in Endocrinology **16**, 1495306 (2025) https://doi.org/10.3389/fendo.2025.1495306

[31] Masood, S., Al-Bashrawi, M.A., Khan, M.A., Dwivedi, Y.K.: From data to diagnosis: a systematic review on ai-driven approaches to diabetes prediction. Artificial Intelligence Review (2026) https://doi.org/10.1007/s10462-025-11485-3

[32] Dawed, A.Y., et al.: Pharmacogenomic variability and glycaemic response to glp-1 receptor agonists. Diabetes Care (2019)

[33] Dawed, A.Y., *et al.*: Pharmacogenomics of glp-1 receptor agonists: a genome-wide analysis of observational data and large randomised controlled trials. The Lancet Diabetes & Endocrinology (2023) https://doi.org/10.1016/S2213-8587(22)00340-0

[34] Wolff, R.F., Moons, K.G.M., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S.: Probast: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine **170**(1), 51–58 (2019) https://doi.org/10.7326/M18-1376

[35] Moons, K.G.M., Collins, G.S., Riley, R.D., *et al.*: Probast+ai: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ **388**, 082505 (2025) https://doi.org/10.1136/bmj-2024-082505

[36] Collins, G.S., Dhiman, P., Navarro, C.L.A., Ma, J., Hooft, L., Reitsma, J.B., Moons, K.G.M., Beam, A.L., Peng, L., Van Calster, B., Smeden, M., Riley, R.D.: Tripod+ai statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ (2024)

[37] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., *et al.*: The prisma 2020 statement: An updated guideline for reporting systematic reviews. BMJ **372**, 71 (2021) https://doi.org/10.1136/bmj.n71

[38] Petticrew, M., Roberts, H.: Systematic Reviews in the Social Sciences: A Practical Guide. Blackwell Publishing, ??? (2006)

[39] Rockenschaub, P., Akay, E.M., Carlisle, B.G., Hilbert, A., Wendland, J., Meyer-Eschenbach, F., Näher, A.-F., Frey, D., Madai, V.I.: External validation of ai-based scoring systems in the icu: a systematic review and meta-analysis. BMC Medical Informatics and Decision Making **25**(1), 5 (2025)

[40] Steyerberg, E.W.: Clinical Prediction Models, 2nd edn. Springer, ??? (2019)

[41] Sounderajah, V., Ashrafian, H., Rose, S., Shah, N.H., Ghassemi, M., Golub, R.M., Ting, D.S.W., Karthikesalingam, A., Darzi, A.: Developing specific reporting guidelines for diagnostic accuracy studies assessing ai interventions: The stard-ai and quadas-ai initiatives. Nature Medicine **27**, 1663–1665 (2021)

[42] Sounderajah, V., Ashrafian, H., Golub, R.M., *et al.*: Quadas-ai: An ai-specific extension of the quadas-2 tool for quality assessment of diagnostic accuracy studies. Nature Medicine **30**, 1154–1163 (2024)

[43] Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine **17**(1), 195 (2019) https://doi.org/10.1186/s12916-019-1426-2

[44] Guni, A., Sounderajah, V., Whiting, P., Bossuyt, P., Darzi, A., Ashrafian, H.: Revised tool for the quality assessment of diagnostic accuracy studies using ai (quadas-ai): protocol for a qualitative study. JMIR Research Protocols **13**(1), 58202 (2024)

[45] Booth, A., Fryer, K.: Qualitative evidence synthesis. In: Patient Involvement in Health Technology Assessment, pp. 177–194. Springer, ??? (2026)

[46] Tuccinardi, D., Zilich, R., Ponzani, P., Musacchio, N., Manicardi, V., Rocca, A., Pontremoli, R., Fioretto, P., Muscarà, A., Di Cianni, G., *et al.*: Real-world prescriptions of glp-1ras and sglt2is in type 2 diabetes prioritise bmi and age over cardiorenal risk: a machine learning-based large cohort analysis. Cardiovascular Diabetology–Endocrinology Reports **11**(1), 38 (2025)

[47] Mujahid, M., Kına, E., Rustam, F., Villar, M.G., Alvarado, E.S., De La Torre Diez, I., Ashraf, I.: Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. Journal of Big Data **11**(1), 87 (2024)

[48] Lotterhos, K.E.: Principles in experimental design for evaluating genomic forecasts. Methods in Ecology and Evolution **15**(9), 1466–1482 (2024)

[49] Williams, R., Bolton, T., Jenkins, D., Mizani, M.A., Sperrin, M., Sudlow, C., Wood, A., Heald, A., Peek, N., Consortium, C.-C.-U.-I.: The challenges of replication: a worked example of methods reproducibility using electronic health record data. PLoS One **20**(7), 0326335 (2025)

[50] Aderemi, I.A., Kehinde, T.O., Ugochukwu, D.O., Ahmad, K.H., Adjei, K.Y., Chijioke, C.E.: Beyond the black box: a systematic review of explainable ai for transparent and trustworthy water quality monitoring. IEEE Sensors Reviews (2025)

[51] Ahmed, S., Kaiser, M.S., Hossain, M.S., Andersson, K.: A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions. IEEE access **13**, 37370–37388 (2024)

[52] El Naqa, I., Karolak, A., Luo, Y., Folio, L., Tarhini, A.A., Rollison, D., Parodi, K.: Translation of ai into oncology clinical practice. Oncogene **42**(42), 3089–3097 (2023)

[53] Wang, X., Nguyen, J., Nader, K., Miihkinen, M., Wall, P., Singhal, A., Bedard, P.L., Ideker, T., Aittokallio, T., Haibe-Kains, B.: Discovery of predictive biomarkers for cancer therapy through computational approaches. Nature Reviews Clinical Oncology, 1–21 (2026)

[54] Lin, M., Guo, J., Gu, Z., Tang, W., Tao, H., You, S., Jia, D., Sun, Y., Jia, P.: Machine learning and multi-omics integration: advancing cardiovascular translational research and clinical practice. Journal of Translational Medicine **23**(1), 388 (2025)

[55] Yabe, D., Nakamura, J., Kaneto, H., Deenadayalan, S., Navarria, A., Gislum, M., *et al.*: Safety and efficacy of oral semaglutide versus dulaglutide in japanese patients with type 2 diabetes (pioneer 10): an open-label, randomised, active-controlled, phase 3a trial. The Lancet Diabetes & Endocrinology **8**(5), 392–406 (2020) https://doi.org/10.1016/S2213-8587(20)30074-7

[56] Davies, M.J., Aroda, V.R., Collins, B.S., Gabbay, R.A., Green, J., Maruthur, N.M., Rosas, S.E., Del Prato, S., Mathieu, C., Mingrone, G., Rossing, P., Tankova, T., Tsapas, A., Buse, J.B.: Management of hyperglycemia in type 2 diabetes, 2022. a consensus report by the american diabetes association (ada) and the european association for the study of diabetes (easd). Diabetes Care **45**(11), 2753–2786 (2022) https://doi.org/10.2337/dci22-0034

[57] Squire, P., Naude, J., Zentner, A., Bittman, J., Khan, N.: Factors associated with weight loss response to glp-1 analogues for obesity treatment: a retrospective cohort analysis. BMJ Open **15**(1), 089477 (2025) https://doi.org/10.1136/bmjopen-2024-089477

[58] Małecki, M.T., Batterham, R.L., Sattar, N., Levine, J.A., Rodríguez, Á., Bergman, B.K., Wang, H., Ghimpeteanu, G., Lee, C.J.: Predictors of at least 15 percent weight reduction and associated changes in cardiometabolic risk factors with tirzepatide in adults with type 2 diabetes in surpass 1–4. Diabetes Care **46**(12), 2292–2299 (2023) https://doi.org/10.2337/dc23-1135

[59] Mohammadagha, M.: Hyperparameter optimization strategies for tree-based machine learning models prediction: a comparative study of adaboost, decision trees, and random forest. Decision trees, and random forest (April 11, 2025) (2025)

[60] Mohril, R.S., Solanki, B.S., Kulkarni, M.S., Lad, B.K.: Xgboost based residual life prediction in the presence of human error in maintenance. Neural Computing and Applications **35**(4), 3025–3039 (2023)

[61] Zhao, J., Qu, X., Wu, Y., Fowler, M., Burke, A.F.: Artificial intelligence-driven real-world battery diagnostics. Energy and AI **18**, 100419 (2024)

[62] Imani, M., Joudaki, M., Bagheri, A., Arabnia, H.R.: Why roc-auc is misleading for highly imbalanced data: In-depth evaluation of mcc, f2-score, h-measure, and auc-based metrics across diverse classifiers. Technologies **14**(1), 54 (2026)

[63] Marques, L., Costa, B., Pereira, M., Silva, A., Santos, J., Saldanha, L., Silva, I., Magalhães, P., Schmidt, S., Vale, N.: Advancing precision medicine: a review of innovative in silico approaches for drug development, clinical pharmacology and personalized healthcare. Pharmaceutics **16**(3), 332 (2024)

[64] Li, X., Peng, L., Wang, Y.-P., Zhang, W.: Open challenges and opportunities in federated foundation models towards biomedical healthcare. BioData Mining **18**(1), 2 (2025)

[65] Gharaie Amirabadi, D., Miraki Feriz, A., Safarpour, H.: Multimodal foundation models in colorectal cancer: from prediction to trustworthy clinical insight. Briefings in Bioinformatics **27**(2), 179 (2026)

[66] Elkefi, S., Scheinmann, R., Hao, D., Bhar, S., Tounsi, A., Feiner, S.: The use of large language models and other generative ai tools to support decision-making in cancer care: a mapping review. BMC Artificial Intelligence **2**(1), 7 (2026)

[67] Moor, M., Banerjee, O., Abad, Z.S., *et al.*: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023) https://doi.org/10.1038/s41586-023-05881-4

[68] Singhal, K., Azizi, S., Tu, T., *et al.*: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023) https://doi.org/10.1038/s41586-023-06291-2

[69] Goryanin, I., Goryanin, I., Demin, O.: Revolutionizing drug discovery: Integrating artificial intelligence with quantitative systems pharmacology. Drug Discovery Today, 104448 (2025)

[70] Zhu, X., *et al.*: Predicting responsiveness to glp-1 pathway drugs using real-world data and machine learning. BMC Endocrine Disorders (2024) https://do

i.org/10.1186/s12902-024-01798-9

[71] Vellikudathil, A.T., et al.: Computational approaches for clinical, genomic and proteomic markers of response to glucagon-like peptide-1 therapy in type-2 diabetes mellitus: An exploratory analysis with machine learning algorithms. Diabetes & Metabolic Syndrome: Clinical Research & Reviews (2024)

[72] Gavigan, C., Donner, T.: Predictors of responsiveness to glp-1 receptor agonists in insulin-treated patients with type 2 diabetes. Journal of Diabetes Research **2023**, 9972132 (2023) https://doi.org/10.1155/2023/9972132

[73] Franch-Nadal, J., *et al.*: Determinants of response to the glucagon-like peptide-1 receptor agonists in a type 2 diabetes population in the real-world. Primary Care Diabetes **16**, 810–817 (2022) https://doi.org/10.1016/j.pcd.2022.10.005

[74] Dennis, J.M., Young, K.G., Cardoso, P., Güdemann, L.M., McGovern, A.P., Farmer, A., Holman, R.R., Sattar, N., McKinley, T.J., Pearson, E.R., Jones, A.G., Shields, B.M., Hattersley, A.T.: A five-drug class model using routinely available clinical features to optimise prescribing in type 2 diabetes: a prediction model development and validation study. The Lancet **405**(10480), 701–714 (2025) https://doi.org/10.1016/S0140-6736(24)02617-5

[75] Cardoso, P., Young, K.G., Nair, A.T.N., *et al.*: Phenotype-based targeted treatment of sglt2 inhibitors and glp-1 receptor agonists in type 2 diabetes. Diabetologia **67**(5), 822–836 (2024) https://doi.org/10.1007/s00125-024-06099-3

[76] Pearson, E.R., Del Prato, S., Pavo, I., Franco, D.R., Zheng, J., Nicolay, C., Hemmingway, A., Wiese, R.J., Kahn, S.E.: Predictors of initial and sustained glycemic and weight response to tirzepatide: A post hoc analysis of surpass-4. Diabetes **74**(10), 1850–1862 (2025) https://doi.org/10.2337/db25-0276

[77] Powell, W., Song, X., Mohamed, Y., Walsh, D., Parks, E.J., McMahon, T.M., Khan, M., Waitman, L.R.: Medications and conditions associated with weight loss in patients prescribed semaglutide based on real-world data. Obesity **31**(10), 2482–2492 (2023) https://doi.org/10.1002/oby.23859

[78] Pratley, R.E., Aroda, V.R., Catarig, A.-M., *et al.*: Impact of patient characteristics on efficacy and safety of once-weekly semaglutide versus dulaglutide: Sustain 7 post hoc analyses. BMJ Open **10**(11), 037883 (2020) https://doi.org/10.1136/bmjopen-2020-037883

[79] Ludvik, B., Giorgino, F., Jódar, E., Frias, J.P., Fernández Landó, L., Brown, K., Bray, R., Rodríguez, A.: Once-weekly tirzepatide versus once-daily insulin degludec as add-on to metformin with or without sglt2 inhibitors in patients with type 2 diabetes (surpass-3): a randomised, open-label, parallel-group, phase 3 trial. The Lancet **398**(10300), 583–598 (2021) https://doi.org/10.1016/S014

0-6736(21)01443-4

[80] SURPASS-4 Investigators: Long-term therapeutic persistence and longitudinal treatment-response trajectories in tirzepatide therapy. The Lancet Diabetes & Endocrinology (2023)

[81] Garvey, W.T., Batterham, R.L., Bhatta, M., *et al.*: Two-year effects of semaglutide in adults with overweight or obesity: the step 5 trial. Nature Medicine **28**(10), 2083–2091 (2022) https://doi.org/10.1038/s41591-022-02026-4

[82] Zinman, B., Aroda, V.R., Buse, J.B., *et al.*: Efficacy, safety, and tolerability of oral semaglutide versus placebo added to insulin with or without metformin in patients with type 2 diabetes: The pioneer 8 trial. Diabetes Care **42**(12), 2262–2271 (2019) https://doi.org/10.2337/dc19-0898

[83] Garvey, W.T., Frias, J.P., Jastreboff, A.M., *et al.*: Tirzepatide once weekly for the treatment of obesity in people with type 2 diabetes (surmount-2): a double-blind, randomised, multicentre, placebo-controlled, phase 3 trial. The Lancet **402**(10402), 613–626 (2023) https://doi.org/10.1016/S0140-6736(23)01200-X

[84] Wharton, S., Calanna, S., Davies, M., Dicker, D., Goldman, B., Lingvay, I., *et al.*: Gastrointestinal tolerability of once-weekly semaglutide 2.4 mg in adults with overweight or obesity, and the relationship between gastrointestinal adverse events and weight loss. Diabetes, Obesity and Metabolism **24**(1), 94–105 (2022) https://doi.org/10.1111/dom.14551

[85] Wadden, T.A., Chao, A.M., Machineni, S., *et al.*: Tirzepatide after intensive lifestyle intervention in adults with overweight or obesity: the surmount-3 phase 3 trial. Nature Medicine **29**(11), 2909–2918 (2023) https://doi.org/10.1038/s41591-023-02597-w

[86] Aronne, L.J., Sattar, N., Horn, D.B., *et al.*: Continued treatment with tirzepatide for maintenance of weight reduction in adults with obesity: The surmount-4 randomized clinical trial. JAMA **331**(1), 38–48 (2024) https://doi.org/10.1001/jama.2023.24945

[87] Rosenstock, J., Allison, D., Birkenfeld, A.L., *et al.*: Effect of additional oral semaglutide vs sitagliptin on glycated hemoglobin in adults with type 2 diabetes uncontrolled with metformin alone or with sulfonylurea: The pioneer 3 randomized clinical trial. JAMA **321**(15), 1466–1480 (2019) https://doi.org/10.1001/jama.2019.2942

[88] Rodriguez, P.J., Zhang, V., Gratzl, S., Do, D., Goodwin Cartwright, B., Baker, C., Gluckman, T.J., Stucky, N., Emanuel, E.J.: Discontinuation and reinitiation of dual-labeled glp-1 receptor agonists among us adults with overweight or obesity. JAMA Network Open **8**(1), 2457349 (2025) https://doi.org/10.1001/jamanetworkopen.2024.57349

[89] Husain, M., Birkenfeld, A.L., Donsmark, M., *et al.*: Oral semaglutide and cardiovascular outcomes in patients with type 2 diabetes. New England Journal of Medicine **381**(9), 841–851 (2019) https://doi.org/10.1056/NEJMoa1901118

[90] Ahren, B., Masmiquel, L., Kumar, H., *et al.*: Efficacy and safety of once-weekly semaglutide versus once-daily sitagliptin as an add-on to metformin, thiazolidinediones, or both, in patients with type 2 diabetes (sustain 2): a 56-week, double-blind, phase 3a, randomised trial. The Lancet Diabetes & Endocrinology **5**(5), 341–354 (2017) https://doi.org/10.1016/S2213-8587(17)30092-X

[91] Frias, J.P., Bonora, E., Nevarez Ruiz, L., *et al.*: Efficacy and safety of dulaglutide 3.0 mg and 4.5 mg versus dulaglutide 1.5 mg in metformin-treated patients with type 2 diabetes in a randomized controlled trial (award-11). Diabetes Care **44**(3), 765–773 (2021) https://doi.org/10.2337/dc20-1473

[92] Perkovic, V., Tuttle, K.R., Rossing, P., *et al.*: Effects of semaglutide on chronic kidney disease in patients with type 2 diabetes. New England Journal of Medicine (2024) https://doi.org/10.1056/NEJMoa2403347

[93] Ji, L., Dong, X., Li, Y., *et al.*: Efficacy and safety of once-weekly semaglutide versus once-daily sitagliptin as add-on to metformin in patients with type 2 diabetes in sustain china: A 30-week, double-blind, phase 3a, randomized trial. Diabetes, Obesity and Metabolism **23**(2), 404–414 (2021) https://doi.org/10.1111/dom.14232

[94] Wadden, T.A., Bailey, T.S., Billings, L.K., *et al.*: Effect of subcutaneous semaglutide vs placebo as an adjunct to intensive behavioral therapy on body weight in adults with overweight or obesity: The step 3 randomized clinical trial. JAMA **325**(14), 1403–1413 (2021) https://doi.org/10.1001/jama.2021.1831

[95] Rubino, D., Abrahamsson, N., Davies, M., *et al.*: Effect of continued weekly subcutaneous semaglutide vs placebo on weight loss maintenance in adults with overweight or obesity: The step 4 randomized clinical trial. JAMA **325**(14), 1414–1425 (2021) https://doi.org/10.1001/jama.2021.3224

[96] Dahl, D., Onishi, Y., Norwood, P., *et al.*: Effect of subcutaneous tirzepatide vs placebo added to titrated insulin glargine on glycemic control in patients with type 2 diabetes: The surpass-5 randomized clinical trial. JAMA **327**(6), 534–545 (2022) https://doi.org/10.1001/jama.2022.0078

[97] SURPASS-4 Investigators: Tirzepatide versus insulin glargine in type 2 diabetes and increased cardiovascular risk (surpass-4): a randomised, open-label, parallel-group, multicentre, phase 3 trial. The Lancet **398**(10313), 1811–1824 (2021) https://doi.org/10.1016/S0140-6736(21)02188-7

[98] Frias, J.P., Auerbach, P., Bajaj, H.S., *et al.*: Efficacy and safety of once-weekly semaglutide 2.0 mg versus 1.0 mg in patients with type 2 diabetes (sustain forte):

a double-blind, randomised, phase 3b trial. The Lancet Diabetes & Endocrinology **9**(9), 563–574 (2021) https://doi.org/10.1016/S2213-8587(21)00174-1

[99] Pratley, R., Amod, A., Hoff, S.T., *et al.*: Oral semaglutide versus subcutaneous liraglutide and placebo in type 2 diabetes (pioneer 4): a randomised, double-blind, phase 3a trial. The Lancet **394**(10192), 39–50 (2019) https://doi.org/10.1016/S0140-6736(19)31271-1

[100] Frias, J.P., Davies, M.J., Rosenstock, J., *et al.*: Tirzepatide versus semaglutide once weekly in patients with type 2 diabetes. New England Journal of Medicine **385**(6), 503–515 (2021) https://doi.org/10.1056/NEJMoa2107519

[101] Jastreboff, A.M., Aronne, L.J., Ahmad, N.N., *et al.*: Tirzepatide once weekly for the treatment of obesity. New England Journal of Medicine **387**(3), 205–216 (2022) https://doi.org/10.1056/NEJMoa2206038

[102] Gerstein, H.C., Colhoun, H.M., Dagenais, G.R., *et al.*: Dulaglutide and cardiovascular outcomes in type 2 diabetes (rewind): a double-blind, randomised placebo-controlled trial. The Lancet **394**(10193), 121–130 (2019) https://doi.org/10.1016/S0140-6736(19)31149-3

[103] Lincoff, A.M., Brown-Frandsen, K., Colhoun, H.M., Deanfield, J., Emerson, S.S., Esbjerg, S., Hardt-Lindberg, S., Hovingh, G.K., Kahn, S.E., Kushner, R.F., Lingvay, I., Oral, T.K., Michelsen, M.M., Plutzky, J., Tornøe, C.W., Ryan, D.H., SELECT Trial Investigators: Semaglutide and cardiovascular outcomes in obesity without diabetes. New England Journal of Medicine **389**(24), 2221–2232 (2023) https://doi.org/10.1056/NEJMoa2307563

[104] Vavekanand, R., Kumar, T., Kumar, S., Kumar, G., Laghari, A.A.: Multimodal machine learning approaches in predictive healthcare analytics: a comprehensive survey. Archives of Computational Methods in Engineering, 1–25 (2026)

[105] T. Denton, B.: Frontiers of medical decision-making in the modern age of data analytics. IISE Transactions **55**(1), 94–105 (2023)

[106] Arshad, M.F., Chowdhury, A.T., Sharif, Z., Islam, M.S.B., Sumon, M.S.I., Mohammedkasim, A., Chowdhury, M.E., Pedersen, S.: Artificial intelligence and machine learning in lung cancer: Advances in imaging, detection, and prognosis. Cancers **17**(24), 3985 (2025)

[107] Yang, L., Wang, Y., Wang, T., Zhou, J., Lv, J., Liang, H.: Advances in the application of artificial intelligence-driven multi-modal imaging technologies in the comprehensive diagnosis and treatment of diabetic foot ulcers. Reviews in Endocrine and Metabolic Disorders, 1–25 (2026)

[108] Huang, L., Lu, H., Yang, M., Liu, Y., Han Wang, M., Zhang, K.: Artificial

intelligence in fundus photography for type 2 diabetes: a scoping review of systemic biomarkers and multi-organ risk prediction. Frontiers in Digital Health **8**, 1768780 (2026)

[109] Bommasani, R., Hudson, D.A., Adeli, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)