BIOLOGY AND MOLECULAR MEDICINE

Research Article

Machine Learning Prediction of Multidimensional Treatment Response in Type 2 Diabetes: An Internally Validated Observational Study

Antonio Maria Labate[1*] and Provvidenza Villari[2]

[1] ASST Mantova, Servizio di Specialistica Ambulatoriale, Branca Diabetologia, Casa di Comunità di Goito, Mantova, Italy.

[2] ASST Garda, Specialistica Ambulatoriale – Branca Diabetologia, Località Montecroce, 25015 Desenzano del Garda, Brescia, Italy.

[*] Correspondence: Antonio Maria Labate; antonio.labate@asst-franciacorta.it

Abstract

Background: Individual treatment responses in type 2 diabetes vary across glycaemic, weight and albuminuric outcomes. We compared prediction using an expanded clinical gradient-boosting model with a simple linear benchmark under recorded treatment.

Methods: This retrospective single-centre study included 559 participants across five regimens after excluding tirzepatide. Outcomes were nominal 12-month reductions in glycated haemoglobin (HbA1c), relative weight and log1p albuminuria. A baseline-outcome-plus-regimen linear benchmark was compared with clinical linear regression and gradient boosting using identical expanded predictors. Five outer and three inner grouped cross-validation partitions kept matching baseline signatures together. Exploratory analyses assessed exclusion of dapagliflozin, permutation importance, clinical rules and observed-response clustering. The emphasis on HbA1c followed preliminary analyses.

Results: Gradient boosting reduced HbA1c mean absolute error from 0.307 to 0.276 percentage points: a 10.2% improvement (95% confidence interval 4.4–15.3), with out-of-fold coefficient of determination (R^2) 0.576. Albuminuria prediction improved by 13.5% (8.4–18.6), with R^2 0.214. Weight prediction improved by 3.9% (−1.2 to 8.3), with R^2 0.122. Glycaemic and albuminuric gains persisted without dapagliflozin. Baseline HbA1c, renal features and adiposity contributed consistently to glycaemic prediction. Clinical rules did not clearly improve performance. Two internally stable observed-response clusters did not establish validated baseline phenotypes.

Conclusions: The expanded clinical gradient-boosting model improved internally validated response prediction relative to the specified simple linear benchmark, most clearly for HbA1c. This comparison changes both predictors and algorithm and does not isolate the contribution of additional clinical variables. Clinical usefulness, stable responder phenotypes and optimal treatment selection remain unproven. External validation is required.

Keywords: type 2 diabetes; treatment response; machine learning; internal validation; albuminuria.

Abbreviations: **ADA**, American Diabetes Association; **AI**, artificial intelligence; **ALT**, alanine aminotransferase; **BMI**, body mass index; **CI**, confidence interval; **CKD**, chronic kidney disease; **DPP-4**, dipeptidyl peptidase-4; **EASD**, European Association for the Study of Diabetes; **eGFR**, estimated glomerular filtration rate; **GLP-1**, glucagon-like peptide-1; **HbA1c**, glycated haemoglobin; **IQR**, interquartile range; **MAE**, mean absolute error; **MDRD**, Modification of Diet in Renal Disease; **ML**, machine learning; **P1Y**, estimated one-year cardiovascular risk; **R^2**, coefficient of determination; **RMSE**, root mean squared error; **SGLT2**, sodium-glucose cotransporter 2; **T0**, baseline; **T12**, nominal 12-month follow-up; **TRIPOD+AI**, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence.

Editor: Eric H. Rosenn

Received: 10/04/26
Accepted: 11/11/26
Published: 11/26/26

Citation: Labate AM; Villari P. Machine Learning Prediction of Multidimensional Treatment Response in Type 2 Diabetes: An Internally Validated Observational Study. *Biology & Molecular Medicine* 2026, 1, 6.

Copyright: © 2026 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license.

1. Introduction and Study Objective

Individualised treatment in type 2 diabetes requires attention to glycaemic control, weight, comorbidity and patient priorities. The American Diabetes Association (ADA)/European Association for the Study of Diabetes (EASD) consensus places these dimensions within a person-centred framework, including cardiovascular and kidney protection [1]. However, choosing an evidence-based drug class does not determine the magnitude of response that a particular patient will experience.

A recent narrative review by Labate and colleagues framed artificial intelligence (AI) in non-insulin-treated type 2 diabetes as support for recognising deterioration and informing treatment timing [2]. That clinical rationale motivates response prediction, but does not itself establish a validated treatment-selection algorithm.

Randomised trials establish the efficacy of the therapies represented in this study. Subcutaneous and oral semaglutide improve glycaemic control, with accompanying weight reduction in SUSTAIN 1 and PIONEER 1; SUSTAIN 7 demonstrated differences in these outcomes between semaglutide and dulaglutide [3–5]. Empagliflozin reduced cardiovascular events in EMPA-REG OUTCOME, and DAPA-CKD, EMPA-KIDNEY and FLOW established kidney benefits in their respective trial populations [6–9]. These population-level findings support the clinical relevance of glycaemic, weight and renal responses, but do not provide an individual prediction of all three outcomes in routine care.

Heterogeneity in type 2 diabetes has stimulated attempts to identify clinically meaningful subgroups. Ahlqvist and colleagues proposed clusters based on six baseline variables and described different complication risks [10]. In subsequent trial-data analyses, Dennis and colleagues found that simple continuous clinical features could be more useful than cluster membership for stratifying progression and treatment response [11]. This distinction matters when considering whether a discrete responder label adds information beyond the measurements from which it is derived.

Evidence for treatment stratification requires explicit evaluation of differential drug response. TriMaster tested prespecified stratification by body mass index (BMI) and renal function in a randomised crossover design, while an algorithm comparing sodium-glucose cotransporter 2 (SGLT2) and dipeptidyl peptidase-4 (DPP-4) inhibitors was evaluated in independent observational and trial datasets [12, 13]. Other work associated markers of beta-cell failure with reduced glycaemic response to glucagon-like peptide-1 (GLP-1) receptor agonists and clinical markers of insulin resistance with response to DPP-4 inhibitors [14, 15]. Together, these studies motivate evaluation of readily available baseline information while cautioning against extrapolating a predictor from one drug class to another.

The present study compares prediction of subsequent glycated haemoglobin (HbA1c), weight and albuminuria changes using expanded clinical models with a simple linear benchmark containing the starting outcome value and recorded regimen. The principal comparison changes both the predictor set and modelling algorithm; comparisons between the two linear models hold the algorithm constant, whereas comparisons between clinical linear regression and gradient boosting use the same expanded predictors. A secondary question is whether the predictive information can be represented by stable, clinically recognisable combinations of features. Prediction under a recorded regimen and causal selection of the best alternative regimen are different objectives [16]. Our analysis addresses the former.

We consolidated the analysis around continuous response prediction. HbA1c is the leading outcome in this report because preliminary analyses showed the strongest predictive performance; this emphasis was not prospectively registered. Weight and albuminuria remain secondary exploratory outcomes. Observed-response clustering is retained as a descriptive analysis. Reporting is informed by TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence) [17]; the analysis is presented as exploratory development with internal validation rather than a clinically deployable prescribing tool.

2. Methods

2.1. Design and Analytical Population

This retrospective, single-centre observational study was conducted in the outpatient Diabetology Service at Casa di Comunità di Goito, ASST Mantova, Italy. Index treatments were initiated in routine care between 1 January 2023 and 31 December 2024, with the corresponding one-year follow-up completed by 2025. These dates refer to routine-care treatment initiation, not prospective research enrolment. Retrospective research inclusion, extraction of clinical data for research and collection of written informed consent occurred after ethics approval, as detailed below. Baseline T0 corresponded to initiation of the index treatment. In the combination group, oral semaglutide and empagliflozin were initiated together at T0.

The source dataset contained 669 treatment-labelled records. The 110 patients treated with tirzepatide had initiated therapy from February 2025 onwards, outside the protocol-defined treatment-initiation window. Their records were excluded during definition of the analytical population, before the predictive analyses of the present study, leaving 559 participants. Exclusion was based on the treatment-initiation period, not observed response or predictive performance. Recorded regimen groups were subcutaneous semaglutide (184), oral semaglutide (139), oral semaglutide plus SGLT2 inhibitor (92), empagliflozin (85) and dapagliflozin (59). The source dictionary allows concomitant SGLT2 use within the subcutaneous semaglutide category; recorded categories therefore should not be interpreted as verified pharmacological monotherapies.

The first author personally reviewed the individual clinical records of the patients belonging to the 17 pairs with matching baseline measurements. The review covered patient-identifying information and verification of the biochemical values against their transcription in the dataset. It

confirmed that these were distinct patients and that the biochemical values had been transcribed correctly. Blood samples had been collected on different dates. No explanation for the matching baseline values was identified. Matching baseline signatures were kept within the same partition during nested grouped cross-validation. The investigator also confirmed that dapagliflozin records represented distinct patients.

The analytical population comprised patients with type 2 diabetes, an identified index regimen, and available baseline and nominal 12-month measurements for the three outcomes. The working file matched the retained source records across all 52 columns. This outcome-available dataset does not reconstruct the number of all eligible treatment initiators, losses to follow-up or reasons for exclusion before creation of the source file. Exact visit dates, individual dosing and titration, adherence, and subsequent treatment changes were not available in the analytical dataset. No additional eligibility criteria or treatment-stability requirements were inferred from these records.

The present analytical population shares 176 participants with the 332-patient source cohort of the DIAL-STEP machine-learning study (unpublished observations; manuscript under peer review). The related project evaluated two-year changes in estimated cardiovascular risk; the present analysis evaluates continuous 12-month glycaemic, weight and albuminuric responses.

2.2. Outcomes

T0 denotes baseline and T12 the nominal 12-month follow-up. Positive values denote improvement: HbA1c reduction = T0 − T12 in percentage points; weight reduction = 100 × (T0 − T12)/T0; albuminuria reduction = ln(1 + T0) − ln(1 + T12), with albuminuria in mg/24 h as confirmed by the investigator. All baseline and follow-up HbA1c and albuminuria values used to calculate these outcomes were measured; neither outcome was imputed or carried forward from another visit. These measurements are distinct from the follow-up glucose and alanine aminotransferase (ALT) fields described in the source dictionary, which were not used as predictors or outcomes in the consolidated analysis. The log1p transformation accommodates three zero albuminuria measurements at T12. Worsening was defined descriptively as a negative reduction; no minimum clinically important response threshold was imposed. The notation log1p(x) denotes ln(1 + x).

2.3. Statistical Analysis

Baseline clinical predictors and the recorded regimen were evaluated using linear regression and gradient boosting, with nested grouped internal validation. The simple linear benchmark included the baseline outcome value and recorded regimen; clinical linear regression and gradient boosting used identical expanded predictor sets. The principal comparison between expanded gradient boosting and the simple linear benchmark therefore changes both predictors and algorithm and does not isolate the contribution of additional clinical variables. Continuous prediction error, calibration, sensitivity analyses and descriptive response clustering were assessed. Appendix 1 contains the full predictor definitions, algorithms, hyperparameters and validation procedures, together with the methodological references [18–26]. Analyses used Python 3.12.14; executable programs and instructions are supplied as Supplementary Code 1.

3. Results

3.1. Population and Observed Responses

Selected baseline characteristics by recorded regimen are shown in Table 1. The population was heterogeneous across several baseline dimensions; global exploratory p-values are descriptive and do not establish exchangeability between regimens. In particular, age, BMI, albuminuria, estimated glomerular filtration rate (eGFR) and cardiovascular history differed across recorded groups. No causal treatment comparison is inferred from the outcome patterns. The analytical population flow is shown in Figure 1.

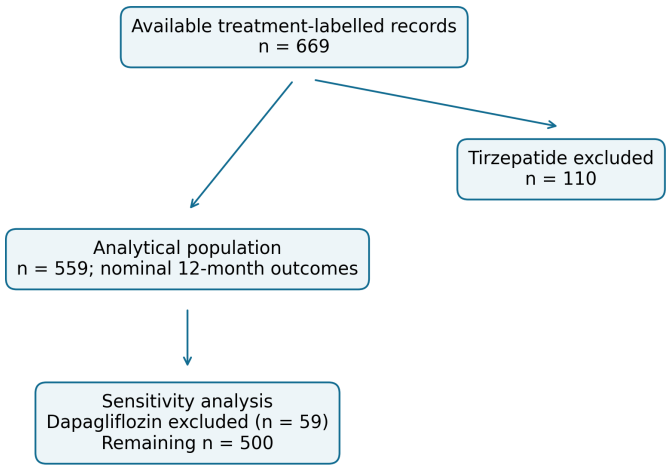

Figure 1. Flow from the available source records to the analytical population and sensitivity analysis. This is an analytical-record flow, not a reconstruction of all eligible treatment initiators or losses to follow-up. The 110 tirzepatide-treated patients were excluded because treatment initiation fell outside the protocol-defined window.

3.2. Internally Validated Prediction

Table 3 presents the model comparisons. Relative to the simple linear benchmark, the expanded clinical gradient-boosting model reduced HbA1c mean absolute error (MAE) by 10.2% (95% confidence interval [CI] 4.4–15.3); relative to clinical linear regression using the same expanded predictors, the reduction was 7.4% (1.6–12.9). For albuminuria, expanded gradient boosting reduced MAE by 13.5% (8.4–18.6) versus the simple linear benchmark; however, its coefficient of determination (R^2) of 0.214 indicates that substantial individual variability remained unexplained. Weight prediction remained weak, with an uncertain 3.9% gain (−1.2 to 8.3) and R^2 0.122.

Table 1. Selected baseline characteristics by recorded regimen.

Characteristic	SC sema n = 184	Oral sema n = 139	Oral sema + SGLT2 n = 92	Empa n = 85	Dapa n = 59	Global p
Age years	66.0 [58.8–73.0]	66.0 [57.5–71.5]	69.0 [60.0–75.2]	75.0 [66.0–79.0]	69.0 [64.0–76.5]	<0.001
Diabetes duration years	9.0 [3.0–14.0]	9.0 [2.0–14.0]	11.0 [4.0–15.0]	9.0 [3.0–15.0]	11.0 [3.5–14.5]	0.605
Weight kg	85.5 [74.0–98.0]	86.0 [74.0–98.0]	89.5 [80.8–100.0]	78.0 [69.0–89.0]	86.0 [76.5–96.5]	0.001
BMI kg/m^2	30.1 [27.7–34.3]	30.1 [27.6–34.3]	31.7 [28.4–35.0]	29.1 [26.1–32.0]	30.8 [28.3–33.9]	0.006
HbA1c %	7.50 [6.98–8.32]	7.60 [7.00–8.53]	7.60 [7.00–8.32]	7.50 [7.09–8.55]	7.50 [7.05–8.75]	0.984
Albuminuria mg/24 h	25.0 [11.0–55.6]	32.0 [15.0–59.0]	50.5 [27.0–90.2]	29.4 [11.0–72.0]	26.0 [9.4–63.9]	<0.001
eGFR mL/min/1.73 m^2	79.8 [66.4–91.8]	79.1 [66.0–91.2]	69.5 [55.9–83.5]	73.6 [55.7–90.2]	74.9 [55.7–89.2]	0.007
Female sex	64 (34.8)	48 (34.5)	25 (27.2)	36 (42.4)	19 (32.2)	0.325
Active smoking	61 (33.2)	13 (9.4)	24 (26.1)	8 (9.4)	26 (44.1)	—
Cardiovascular history	47 (25.5)	38 (27.3)	10 (10.9)	16 (18.8)	22 (37.3)	0.002
Retinopathy	18 (9.8)	11 (7.9)	8 (8.7)	8 (9.4)	7 (11.9)	0.929

Continuous variables: median [interquartile range (IQR)]; categorical variables: n (%). Global p-values from exploratory Kruskal–Wallis or chi-square comparisons, unadjusted for multiplicity. Active smoking is descriptive; the previous global smoking comparison used all three categories and is not reproduced as a test of active smoking alone. eGFR uses the Modification of Diet in Renal Disease (MDRD) equation. Categories identify recorded regimens, not verified monotherapies.

Abbreviations: SC, subcutaneous; sema, semaglutide; Empa, empagliflozin; Dapa, dapagliflozin; BMI, body mass index; HbA1c, glycated haemoglobin; eGFR, estimated glomerular filtration rate; SGLT2, sodium-glucose cotransporter 2.

Table 2. Whole-population observed reductions. A negative reduction denotes worsening. Albuminuria changes on the transformed scale are not equivalent to changes in mg/24 h. HbA1c, glycated haemoglobin; IQR, interquartile range.

Observed reduction	Median [IQR]	Worsening n/N (%)
HbA1c percentage points	0.365 [0.216–0.700]	48/559 (8.6)
Relative weight %	2.632 [1.136–3.876]	66/559 (11.8)
Albuminuria log1p	0.330 [0.107–0.473]	78/559 (14.0)

Table 3. Pooled out-of-fold predictions in all 559 records. Mean absolute error (MAE) is in percentage points for HbA1c and relative weight, and log1p units for albuminuria. Smaller MAE is better. Intervals resample signature groups without refitting. AI, artificial intelligence; CI, confidence interval; HbA1c, glycated haemoglobin; R^2, coefficient of determination.

Outcome	Model	MAE	95% CI MAE	R^2
HbA1c	Baseline + regimen	0.307	0.282–0.333	0.508
HbA1c	Clinical linear	0.298	0.273–0.323	0.528
HbA1c	Clinical AI	0.276	0.252–0.300	0.576
Weight	Baseline + regimen	2.061	1.870–2.245	0.014
Weight	Clinical linear	2.118	1.935–2.289	0.008
Weight	Clinical AI	1.981	1.817–2.155	0.122
Albuminuria	Baseline + regimen	0.232	0.208–0.261	0.112
Albuminuria	Clinical linear	0.238	0.215–0.263	0.120
Albuminuria	Clinical AI	0.201	0.176–0.228	0.214

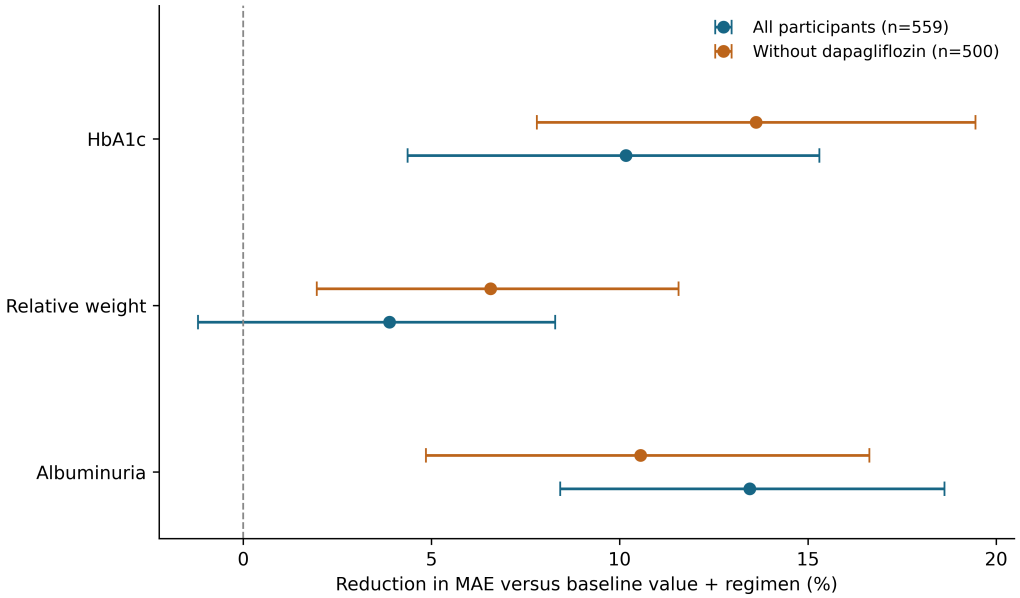

Figure 2. Relative reduction in mean absolute error (MAE) with clinical gradient boosting compared with baseline outcome value plus recorded regimen. Points and bars indicate estimates and paired 95% bootstrap confidence intervals. Positive values favour gradient boosting. Intervals condition on the existing out-of-fold predictions and do not include model refitting.

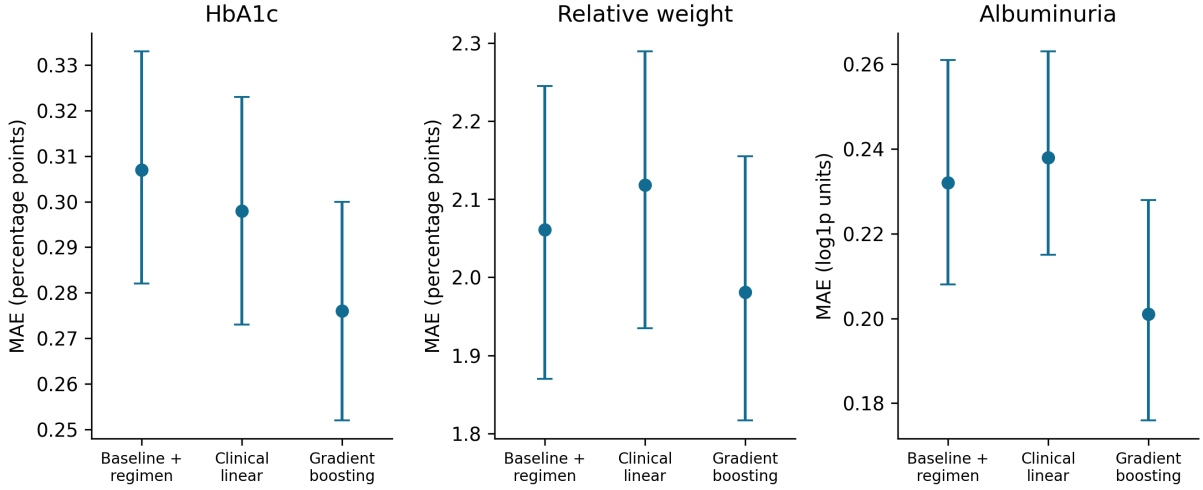

Figure 3. Internally validated prediction error by model and outcome in 559 participants. Bars indicate 95% bootstrap confidence intervals for mean absolute error (MAE). Outcomes have different scales; panel heights must not be compared across outcomes. Lower MAE indicates less prediction error.

At a fixed linear modelling algorithm, expanding the predictor set changed HbA1c MAE from 0.307 to 0.298 percentage points, relative-weight MAE from 2.061 to 2.118 percentage points, and albuminuria MAE from 0.232 to 0.238 log1p units (simple benchmark versus clinical linear regression, respectively). Thus, adding predictors to the linear model did not uniformly improve prediction across outcomes.

Figure 2 summarises relative predictive gains; Figure 3 compares prediction errors by model and outcome.

In the sensitivity analysis excluding dapagliflozin, conducted in 500 patients, the expanded clinical gradient-boosting model reduced HbA1c and albuminuria MAE relative to the simple linear benchmark by 13.6% (95% CI 7.8–19.4) and 10.6% (4.9–16.6), respectively. Weight MAE was reduced by 6.6% (2.0–11.6), but R^2 fell to 0.076. These findings show that improved predictive performance for HbA1c and albuminuria did not depend exclusively on inclusion of the dapagliflozin group; they do not explain that group's narrower outcome distribution or establish transportability.

Table 4. Sensitivity without dapagliflozin: 500 records, 483 signature groups. All models were refitted and retuned. Positive gain denotes smaller MAE. AI, artificial intelligence; MAE, mean absolute error; CI, confidence interval; HbA1c, glycated haemoglobin; R^2, coefficient of determination.

Outcome	AI MAE	AI R^2	MAE gain vs simple [95% CI]
HbA1c	0.261	0.646	13.6% [7.8; 19.4]
Weight	2.138	0.076	6.6% [2.0; 11.6]
Albuminuria	0.216	0.135	10.6% [4.9; 16.6]

Table 5. Observed multidimensional response groups.

Observed group	N	HbA1c reduction (percentage points)	Weight reduction %	Albuminuria reduction log1p
Greater observed response	121	1.20 [0.90–1.78]	4.84 [3.12–7.25]	0.397 [0.281–0.491]
Lower observed response	438	0.30 [0.18–0.50]	2.20 [0.85–3.47]	0.304 [0.069–0.470]

Median [interquartile range (IQR)]. Labels describe observed multidimensional outcomes, not prospectively predictable biological phenotypes. Baseline HbA1c medians were 8.93% and 7.40% in the greater and lower observed-response groups. No dapagliflozin record was in the greater-response cluster. HbA1c, glycated haemoglobin.

3.3. Predictive Information and Clinical Rules

For HbA1c, the largest held-out importance was baseline HbA1c, followed by the renal block and weight/BMI; each had a positive mean importance in all five outer partitions, including the analysis without dapagliflozin. For albuminuria, the renal block had the largest importance. Stable reliance on a feature does not itself define a responder class.

Diabetes duration was the first tree split for HbA1c in all five partitions, both with and without dapagliflozin. In four of the five complete-population partitions, the training-selected branch with the greatest predicted additional reduction included duration $\leq$2.5 years; two of these branches also required higher baseline albuminuria. Their held-out mean residuals were positive. However, the complete rule model improved MAE by only 0.9% (−3.2 to 4.5); the improvement without dapagliflozin was also uncertain at 3.5% (−1.0 to 7.9). These data-derived rules remain candidates, not validated phenotypes.

3.4. Observed Multidimensional Response Groups

Clustering selected two observed-response groups, with 121 and 438 records (Table 5). Their internal stability did not establish baseline predictability or biological endotypes. Baseline HbA1c differed substantially, and no dapagliflozin record belonged to the greater-response group.

Bootstrap Jaccard stability was 0.846 for greater response and 0.954 for lower response; silhouette was 0.369. These stability measures apply to observed-response clustering, not to baseline assignment.

4. Discussion

4.1. Principal Findings

The expanded clinical gradient-boosting model improved internal prediction of continuous treatment response relative to the specified simple linear benchmark, most clearly for HbA1c. It reduced mean absolute error from 0.307 to 0.276 percentage points compared with the starting-value-plus-regimen linear benchmark, an absolute reduction of approximately 0.031 percentage points. This comparison changes both the predictor set and algorithm; the improvement cannot be attributed entirely to additional clinical variables. Gradient boosting also improved on clinical linear regression using the same expanded predictors. Thus, the findings support better predictive performance of the expanded gradient-boosting model within this dataset, while the clinical value of the gain remains undetermined. Albuminuria prediction improved with substantial unexplained variability, whereas individual weight prediction remained weak.

4.2. Glycaemic Response and Baseline Clinical Information

The largest contribution to HbA1c prediction was its baseline value. This is clinically plausible, because a patient's initial level constrains the available reduction, but part of the relation between a starting measurement and its change can also reflect mathematical coupling and regression to the mean. The starting-value-plus-regimen linear comparator was therefore essential. Improvement beyond this comparator indicates that the expanded gradient-boosting model predicted response more accurately than a simple linear representation of the initial outcome and regimen. It does not isolate the effects of added variables from those of the modelling algorithm or establish a biological mechanism for each contributing feature.

In an earlier 12-month observational study of oral semaglutide in patients aged $\geq$65 years, Labate and colleagues reported a glycaemic response largely independent of the assessed baseline characteristics [27]. That study described glycaemic and cardiometabolic treatment outcomes in a distinct population, with no participants shared with the present study. Differences in population, regimens and modelling objectives limit direct comparison. The report is cited as clinical context and does not validate the present prediction models.

Related DIAL-STEP reports address changes in estimated cardiovascular risk, machine-learning prediction of risk trajectories, glycaemic–albuminuric response discordance, and comparisons between treatments. The DIAL-STEP machine-learning project used a source cohort of 332 patients, with 324 included in prediction analyses because they had valid P1Y (estimated one-year cardiovascular risk) estimates at baseline and two years. The present study includes 559 patients across five recorded regimens and predicts continuous 12-month changes in HbA1c, relative weight and log-transformed albuminuria. Although it shares 176 participants with the 332-patient DIAL-STEP ML source cohort, its research question, outcomes and analyses differ. Results from the related projects are not used to validate the present models.

Renal features and adiposity also contributed consistently to glycaemic prediction across held-out partitions. TriMaster demonstrated that renal function and BMI can help stratify responses to specific drug alternatives under a randomised crossover design [12]. The present result is narrower: these blocks contributed to prediction under

recorded treatment, without estimating the counterfactual response to another regimen. The SGLT2-versus-DPP-4 algorithm reported by Dennis and colleagues illustrates the further validation required before translating clinical predictors into a treatment-selection rule [13].

A pathway-focused review linked cardiometabolic residual risk to insulin resistance, inflammation and liver–kidney–vascular interactions [28]. This provides biological context for considering multiple clinical dimensions. However, the present renal and adiposity predictors were routine measurements, not direct assays of these pathways; their predictive contribution cannot establish molecular endotypes.

Shorter diabetes duration appeared repeatedly in candidate tree branches associated with greater additional HbA1c reduction. However, the complete rule model did not clearly improve outer-validation performance. Duration should consequently be interpreted as a candidate feature within an exploratory combination, not as a validated threshold for prescribing. Associations between markers of beta-cell failure and GLP-1 receptor agonist response provide a relevant biological context [14], but C-peptide and islet autoantibodies were unavailable here. Likewise, evidence concerning insulin-resistance markers and DPP-4 inhibitor response [15] cannot be directly transferred to the regimens evaluated in this study.

4.3. Albuminuria and Weight Outcomes

The 13.5% relative reduction in albuminuria prediction error with expanded gradient boosting versus the simple linear benchmark was accompanied by an R^2 of 0.214, leaving most individual variability unexplained. Dapagliflozin reduced albuminuria in the DELIGHT trial, which assessed glycaemic and albuminuric endpoints separately [29]. Our study similarly retained separate continuous outcomes, but cannot use their co-occurrence to identify a common mechanism or to infer comparative kidney protection. The renal predictor block includes both baseline albuminuria and eGFR; its permutation importance does not isolate their individual contributions.

Albuminuria also has substantial within-person biological variability. Repeated-measurement work in chronic kidney disease (CKD) demonstrated high variability in urine albumin and albumin-to-creatinine measurements [30]. Although those sampling methods differ from the 24-hour albumin excretion used here, the findings reinforce the limits of interpreting a single before-and-after change as an individual treatment effect. Trial-level evidence supports albuminuria change as a potential surrogate for kidney disease progression in appropriate settings [31], but our transformed prediction outcome is not a validated surrogate for clinical kidney benefit. The log1p scale accommodates zero values and reduces skewness; it does not provide an immediately interpretable change in mg/24 h.

A stratified synthesis of semaglutide outcome trials emphasised differences between cardiovascular and kidney endpoints and their target populations [32]. Accordingly, the individual albuminuria change predicted here should not be equated with the clinical event reductions demonstrated in those trials.

Weight prediction had low explained variance and an uncertain gain over the simple comparator in the full population. This finding does not challenge the average weight effects established in semaglutide trials [3–5]. Rather, it indicates that the selected baseline information did not accurately describe individual one-year weight response in this routine-care sample. Dose exposure, adherence, dietary changes and treatment tolerability were not available as explanatory variables; their contribution cannot be quantified from the present analysis.

4.4. Observed Profiles and Treatment Selection

The two observed-response groups summarise outcome heterogeneity after treatment. Their internal resampling stability is useful for describing the dataset but does not establish prospective assignment, biological endotypes or an optimal prescription. Outcome-based clustering also guarantees some separation on the variables used to create it. The markedly different baseline HbA1c levels and the absence of dapagliflozin in the greater-response group further caution against interpreting cluster labels as treatment-independent phenotypes.

The distinction between stable clusters and clinically useful stratification is consistent with the comparison of data-driven diabetes subgroups against continuous clinical-feature models [10, 11]. Cluster-wise stability can also depend on the structure imposed by the clustering algorithm [26]. In this exploratory analysis, the clinical rules did not demonstrate a clear improvement in predictive performance. Their clinical usefulness and their ability to identify baseline responder groups therefore remain unestablished.

Predictive associations are insufficient to justify switching an individual patient to a supposedly better regimen [16]. A comparative causal analysis would require a defined treatment contrast, alignment of eligibility and treatment initiation, adequate control of confounding and an appropriate design for estimating alternative outcomes [33]. The nonrandom allocation of therapies and baseline differences in this cohort prevent such an interpretation. Our study therefore complements treatment-stratification research by quantifying internal prediction under observed care, rather than by establishing comparative effectiveness.

4.5. Validation and Future Evaluation

Nested grouped validation separated parameter selection from outer assessment. Nevertheless, choosing the consolidated clinical feature set and highlighting HbA1c after preliminary analyses leaves study-level selection that nested validation cannot fully remove [21, 22]. The earlier model with a broader predictor set is reported in Appendix 1, Supplementary Table S2. The clinical feature set was retained for interpretability and a comparison of algorithms using the same inputs, not because it had the best observed performance.

Descriptive calibration supported evaluation beyond MAE and R^2 [24]. The HbA1c calibration slope was close

to one, but this estimate comes from pooled internal predictions and does not demonstrate calibration at another centre or during another period. Confidence intervals based on fixed outer predictions also omit uncertainty from repeating the full modelling process. They should be read as exploratory uncertainty summaries for this validation run.

Table 6. Descriptive calibration of clinical AI predictions.

Outcome	Mean bias	Slope	Intercept	RMSE
Albuminuria	-0.001	0.858	0.038	0.365
HbA1c	-0.005	0.950	0.031	0.393
Weight	-0.082	0.797	0.620	2.860

Observed outcome regressed on out-of-fold prediction; ideal slope 1 and intercept 0. Bias = prediction − observed outcome. These summaries are descriptive, do not recalibrate the model and are not substitutes for external calibration.

RMSE, root mean squared error. Mean bias, intercept and RMSE are expressed in percentage points for HbA1c reduction and relative weight reduction, and in log1p units for albuminuria reduction. The slope is dimensionless.

AI, artificial intelligence; HbA1c, glycated haemoglobin.

Sample requirements depend on the number of candidate parameters, anticipated performance and the desired precision, rather than on participant count alone [34, 35]. No prospective sample-size calculation was performed for this exploratory analysis. Independent validation, preferably with repeated outcome measurements and documented dose exposure, is required before clinical use. A future study should prespecify the leading outcome, feature set and analysis strategy, and assess quality and applicability using current prediction-model guidance [17, 36].

4.6. Limitations

This exploratory study used a single-centre, outcome-available sample rather than an inception cohort with documented follow-up completeness. The source file did not provide the full denominator of eligible treatment initiators or reasons for attrition, so selection bias cannot be quantified. Exact visit dates, dose trajectories, adherence and subsequent treatment changes were not available. Recorded regimen categories allow some concomitant treatment, limiting attribution of a response to one drug alone. The analysis did not assess safety, treatment discontinuation, cardiovascular events or clinical kidney endpoints.

The selected clinical predictor set and leading emphasis on HbA1c followed preliminary analyses, leaving possible study-level optimism despite nested internal validation. No external validation was performed. Confidence intervals condition on fixed out-of-fold predictions rather than repeating training and model selection. The dapagliflozin HbA1c-change standard deviation was 0.052 percentage points, compared with 0.479–0.691 in other groups. The investigator confirmed that dapagliflozin records represented unique patients. All HbA1c and albuminuria outcome values were measured, with no imputation or carry-forward, but no explanation for the narrow HbA1c-change distribution was identified. The sensitivity analysis excluding dapagliflozin showed that improved HbA1c and albuminuria prediction persisted (Table 4); it does not explain the narrow distribution. The source dictionary states that dapagliflozin follow-up glucose and ALT had been imputed from a later visit; those follow-up fields were not used as predictors or outcomes in the consolidated analysis. Individual review of identifying information and biochemical measurements in the clinical records confirmed that the 17 matching baseline pairs represented distinct patients with correctly transcribed values. No explanation for the matching baseline values was identified; these profiles remained grouped conservatively during validation. Medication abbreviations and historical cardiovascular and retinopathy entries reflect recorded clinical data rather than independent adjudication.

5. Conclusion

The expanded clinical gradient-boosting model improved internal prediction of glycaemic and albuminuric responses relative to a simple starting-value-and-regimen linear benchmark. This comparison changes both predictors and algorithm and does not isolate the contribution of additional clinical variables. Weight prediction remained limited. Stable clinical responder phenotypes and optimal treatment choices were not established. This dataset supports an exploratory prediction study, with phenotype discovery retained as a secondary, unresolved objective.

Ethics Approval and Informed Consent: The study was conducted in accordance with the Declaration of Helsinki and applicable Italian regulations for observational pharmacological research. The Territorial Ethics Committee CET Lombardia 4 approved the retrospective study protocol (CET 48-26) at its meeting on 28 May 2026; the written opinion is dated 29 May 2026. Routine-care clinical measurements had been obtained before ethics approval. Retrospective research inclusion, extraction of clinical data for research and collection of written informed consent occurred after approval. All 559 participants provided written informed consent for study participation and personal-data processing using the approved B1 and B2 documents. Study data were pseudonymised with unique alphanumeric identifiers and access was restricted to authorised investigators.

Author Contributions: Conceptualisation and supervision: Provvidenza Villari. Methodology, investigation, data curation, software, formal analysis, validation, visualisation, project administration and manuscript preparation: Antonio Maria Labate.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

Data and Code Availability: Participant-level clinical data are not publicly available because of patient confidentiality. Aggregate results are reported in the article and supplementary materials. Analysis programs, dependencies and execution instructions are provided as Supplementary Code 1, available with the online version of this article.

Related Unpublished Work: The DIAL-STEP machine-learning study referenced in this article is a separate manuscript from the authors' research group currently undergoing peer review elsewhere. The two studies share 176 participants but address different research questions, outcomes and analytical objectives. The DIAL-STEP findings have not been published at the time of this manuscript's acceptance.

AI Assistance: ChatGPT (OpenAI; GPT-6 model) was used as a support tool for statistical coding and manuscript drafting under the authors' direction. The authors defined the scientific objectives, the analytical approach and the structure of the text, and retain responsibility for the final analyses, interpretation of the results and manuscript content.

Supplementary Materials: Appendix 1: prediction model, internal validation and supplementary methodological tables. Supplementary Code 1: analysis programs, dependencies and execution instructions. TRIPOD+AI checklist: reporting locations and explicitly identified limitations.

References

1. Davies MJ, Aroda VR, Collins BS, et al. 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.* **2022**;*45*:2753–2786. https://doi.org/10.2337/dci22-0034.
2. Labate AM, Cimino E, Giacomelli L, Ettori S, Oladeji OA, Agosti B. Artificial Intelligence in Non-Insulin-Treated Type 2 Diabetes: From Reactive Management to Anticipatory Care. *Endocrines.* **2026**;*7*(3):40. https://doi.org/10.3390/endocrines7030040.
3. Sorli C, Harashima SI, Tsoukas GM, et al. Efficacy and safety of once-weekly semaglutide monotherapy versus placebo in patients with type 2 diabetes (SUSTAIN 1): a double-blind, randomised, placebo-controlled, parallel-group, multinational, multicentre phase 3a trial. *Lancet Diabetes Endocrinol.* **2017**;*5*:251–260. https://doi.org/10.1016/S2213-8587(17)30013-X.
4. Aroda VR, Rosenstock J, Terauchi Y, et al. PIONEER 1: Randomized Clinical Trial of the Efficacy and Safety of Oral Semaglutide Monotherapy in Comparison With Placebo in Patients With Type 2 Diabetes. *Diabetes Care.* **2019**;*42*:1724–1732. https://doi.org/10.2337/dc19-0749.
5. Pratley RE, Aroda VR, Lingvay I, et al. Semaglutide versus dulaglutide once weekly in patients with type 2 diabetes (SUSTAIN 7): a randomised, open-label, phase 3b trial. *Lancet Diabetes Endocrinol.* **2018**;*6*:275–286. https://doi.org/10.1016/S2213-8587(18)30024-X.
6. Zinman B, Wanner C, Lachin JM, et al. Empagliflozin, Cardiovascular Outcomes, and Mortality in Type 2 Diabetes. *N Engl J Med.* **2015**;*373*:2117–2128. https://doi.org/10.1056/NEJMoa1504720.
7. Heerspink HJL, Stefánsson BV, Correa-Rotter R, et al. Dapagliflozin in Patients with Chronic Kidney Disease. *N Engl J Med.* **2020**;*383*:1436–1446. https://doi.org/10.1056/NEJMoa2024816.
8. The EMPA-KIDNEY Collaborative Group. Empagliflozin in Patients with Chronic Kidney Disease. *N Engl J Med.* **2023**;*388*:117–127. https://doi.org/10.1056/NEJMoa2204233.
9. Perkovic V, Tuttle KR, Rossing P, et al. Effects of Semaglutide on Chronic Kidney Disease in Patients with Type 2 Diabetes. *N Engl J Med.* **2024**;*391*:109–121. https://doi.org/10.1056/NEJMoa2403347.
10. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* **2018**;*6*:361–369. https://doi.org/10.1016/S2213-8587(18)30051-2.
11. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT. 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. *Lancet Diabetes Endocrinol.* **2019**;*7*:442–451. https://doi.org/10.1016/S2213-8587(19)30087-7.
12. Shields BM, Dennis JM, Angwin CD, et al. Patient stratification for determining optimal second-line and third-line therapy for type 2 diabetes: the TriMaster study. *Nat Med.* **2023**;*29*:376–383. https://doi.org/10.1038/s41591-022-02120-7.
13. Dennis JM, Young KG, McGovern AP, et al. Development of a treatment selection algorithm for SGLT2 and DPP-4 inhibitor therapies in people with type 2 diabetes: a retrospective cohort study. *Lancet Digit Health.* **2022**;*4*:e873–e883. https://doi.org/10.1016/S2589-7500(22)00174-1.
14. Jones AG, McDonald TJ, Shields BM, et al. Markers of β-Cell Failure Predict Poor Glycemic Response to GLP-1 Receptor Agonist Therapy in Type 2 Diabetes. *Diabetes Care.* **2016**;*39*:250–257. https://doi.org/10.2337/dc15-0258.
15. Dennis JM, Shields BM, Hill AV, et al. Precision Medicine in Type 2 Diabetes: Clinical Markers of Insulin Resistance Are Associated With Altered Short- and Long-term Glycemic Response to DPP-4 Inhibitor Therapy. *Diabetes Care.* **2018**;*41*:705–712. https://doi.org/10.2337/dc17-1827.
16. Shmueli G. To Explain or to Predict? *Stat Sci.* **2010**;*25*:289–310. https://doi.org/10.1214/10-STS330.
17. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* **2024**;*385*:e078378. https://doi.org/10.1136/bmj-2023-078378.
18. Levey AS, Bosch JP, Lewis JB, et al. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Ann Intern Med.* **1999**;*130*:461–470. https://doi.org/10.7326/0003-4819-130-6-199903160-00002.
19. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* **2001**;*29*:1189–1232. https://doi.org/10.1214/aos/1013203451.
20. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* **2011**;*12*:2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html.
21. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics.* **2006**;*7*:91. https://doi.org/10.1186/1471-2105-7-91.

22. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res.* **2010**;*11*:2079–2107. https://www.jmlr.org/papers/v11/cawley10a.html.
23. Steyerberg EW, Harrell FE Jr, Borsboom GJJM, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol.* **2001**;*54*:774–781. https://doi.org/10.1016/S0895-4356(01)00341-9.
24. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **2019**;*17*:230. https://doi.org/10.1186/s12916-019-1466-7.
25. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* **1987**;*20*:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
26. Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data Anal.* **2007**;*52*:258–271. https://doi.org/10.1016/j.csda.2006.11.025.
27. Labate AM, Moretti L, Villari P. Glycaemic and Cardiometabolic Effects of Oral Semaglutide in Patients Aged ≥65 Years with Type 2 Diabetes. *Endocrines.* **2026**;*7*(1):8. https://doi.org/10.3390/endocrines7010008.
28. Labate AM, Cimino E, Giacomelli L, Ettori S, Oladeji OA, Agosti B. Molecular Pathways of Cardiometabolic Residual Risk in Type 2 Diabetes: Insulin Resistance, Metaflammation, and Liver–Kidney–Vascular Crosstalk. *Int J Mol Sci.* **2026**;*27*(14):6170. https://doi.org/10.3390/ijms27146170.
29. Pollock C, Stefánsson B, Reyner D, et al. Albuminuria-lowering effect of dapagliflozin alone and in combination with saxagliptin and effect of dapagliflozin and saxagliptin on glycaemic control in patients with type 2 diabetes and chronic kidney disease (DELIGHT): a randomised, double-blind, placebo-controlled trial. *Lancet Diabetes Endocrinol.* **2019**;*7*:429–441. https://doi.org/10.1016/S2213-8587(19)30086-5.
30. Waikar SS, Rebholz CM, Zheng Z, et al. Biological Variability of Estimated GFR and Albuminuria in CKD. *Am J Kidney Dis.* **2018**;*72*:538–546. https://doi.org/10.1053/j.ajkd.2018.04.023.
31. Heerspink HJL, Greene T, Tighiouart H, et al. Change in albuminuria as a surrogate endpoint for progression of kidney disease: a meta-analysis of treatment effects in randomised clinical trials. *Lancet Diabetes Endocrinol.* **2019**;*7*:128–139. https://doi.org/10.1016/S2213-8587(18)30314-0.
32. Labate AM, Villari P. Semaglutide across cardiovascular and kidney outcome trials: a systematic review and stratified synthesis. *Rom J Diabetes Nutr Metab Dis.* **2026**;*33*(3):416–424. https://rjdnmd.org/index.php/RJDNMD/article/view/2399.
33. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol.* **2016**;*183*:758–764. https://doi.org/10.1093/aje/kwv254.
34. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I—Continuous outcomes. *Stat Med.* **2019**;*38*:1262–1275. https://doi.org/10.1002/sim.7993.
35. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ.* **2020**;*368*:m441. https://doi.org/10.1136/bmj.m441.
36. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ.* **2025**;*388*:e082505. https://doi.org/10.1136/bmj-2024-082505.