Appendix 1: Prediction Model and Internal Validation

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

Antonio Maria Labate and Provvidenza Villari

Reference numbers refer to the main manuscript.

Outcomes

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. 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.

Prediction methods and validation

Predictors and comparator models

The consolidated clinical set comprised baseline age, diabetes duration, weight, BMI, HbA1c, MDRD eGFR [18], log1p albuminuria, sex, smoking, cardiovascular history, retinopathy, an indicator of concomitant insulin and the recorded regimen. A smoking entry labelled SI was recoded as active smoking for analysis. Insulin was inferred from recorded abbreviations GLA, IGLA, DEG, IDEG, ICODEC or INS; the indicator therefore reflects recorded medication abbreviations rather than an independently adjudicated insulin-treatment history. The selected predictors and outcomes had no missing values in the working file. Missing lipid-lowering and antihypertensive medication fields were not used in the consolidated model. No follow-up variable, identifier or cohort label was used as a predictor.

Three models were compared: ordinary linear regression with the baseline value of the outcome and regimen; ordinary linear regression with the complete consolidated clinical set; and histogram gradient boosting with the same complete set. The identical predictors in the latter two models allow an algorithm comparison. Numeric scaling and one-hot category encoding were fitted within training partitions. Gradient boosting used learning rate 0.05, 150 iterations, minimum 20 observations per leaf and no early stopping; inner validation selected 7 or 15 maximum leaves and L2 regularisation 1 or 10 by mean absolute error (MAE). Gradient boosting provides a flexible approach to modelling nonlinear relationships [19]. Models and preprocessing were implemented in scikit-learn [20].

Grouped internal validation and uncertainty

All 559 participants had unique record identifiers. There were 542 distinct signatures across sex, age, diabetes duration, weight, BMI, waist circumference, glucose, HbA1c, creatinine and albuminuria at baseline; 17 pairs, each involving one oral and one subcutaneous semaglutide record, shared these measurements. The investigator confirmed that these were distinct participants. The previously specified conservative validation procedure was retained: matching signatures were kept together in five outer GroupKFold partitions, and three inner grouped partitions selected gradient boosting parameters. Each participant’s outer prediction was obtained without using that participant or the corresponding signature group in training. Separating tuning from outer evaluation reduces optimism arising from model selection [21, 22].

MAE and R^2 were calculated on pooled outer predictions. Paired 95% confidence intervals for MAE and its relative reduction were obtained from 1,000 bootstrap resamples of signature groups, using the same out-of-fold predictions (seed 20261004). These intervals do not include model refitting and are conditional on the cross-validation procedure. They are exploratory and unadjusted for multiple model and outcome comparisons. Observed-versus-predicted regression slope, intercept, mean prediction bias and RMSE were additionally summarised descriptively. No external validation or clinical decision-utility analysis was performed. Resampling was used rather than a single random development/test split to make efficient use of the available sample, consistent with the general rationale for internal validation [23]. Calibration was assessed alongside prediction error because a favourable average error alone does not establish agreement between predicted and observed outcomes [24].

Sensitivity and interpretation

All three models were refitted and retuned after excluding the 59 dapagliflozin records, leaving 500 records and 483 signature groups. This sensitivity was motivated by an unusually narrow HbA1c-change distribution in that group, rather than by duplicate participant records. Held-out permutation importance was calculated with 20 repeats for each block; weight/BMI and eGFR/albuminuria were permuted jointly. Results were averaged equally across outer partitions. Correlations across blocks can make permuted combinations atypical; importance is neither a causal effect nor an exclusive share of explained variance.

Simple rules were explored using decision trees on residuals from the baseline-outcome-plus-regimen model, with baseline clinical predictors but no regimen as tree inputs. Depth 2 or 3 and minimum leaf size 40 or 60 were selected by three inner grouped partitions, refitting both stages in each partition. All rules were evaluated in the outer hold-out partition. Their purpose was to test whether additional predictive information could be represented by a small number of clinical combinations.

Secondary clustering of observed responses

K-means used the three standardised observed changes, with no baseline feature, sex or regimen included. Candidate k values were 2–6; at least 10% of records per cluster and bootstrap Jaccard ≥0.75 were required for eligibility, followed by the largest eligible silhouette. Bootstrap stability used 200 signature-group resamples, restandardisation, reclustering and matched cluster labels. K=2 was the only eligible solution. These rules were specified before that clustering calculation but after initial prediction work; they are not a prospectively registered study protocol. Cluster-internal outcome comparisons are descriptive because these outcomes defined the clusters. Silhouette and cluster-wise resampling stability were used as complementary descriptive assessments [25, 26].

Software and implementation

The analytical environment was Python 3.12.14, pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0 and SciPy 1.17.0. Supplementary Code 1 supplies the consolidated prediction, interpretation and response-clustering programs, dependencies and execution instructions. Only file routing and removal of participant-identifier outputs were changed when packaging the original programs.

The analysis evaluates model-building procedures by nested internal validation. It does not supply a single final deployable fitted model or a clinically validated treatment-selection tool. Participant-level datasets and prediction outputs are not included.

Supplementary Table S1. Data consistency and quality checks

Check	Result
Working file vs retained source data	Exact match across 559 records and 52 columns
Record identifiers	559 unique IDs; no repeated record ID
Matching baseline signatures	542 signatures; 17 pairs involving 34 records
Selected predictor and outcome missingness	No missing selected fields
Other medication fields	40 lipid medication and 100 pressure medication entries missing
Albuminuria zeros	0 at baseline; 3 at T12; handled with log1p
Dapagliflozin HbA1c-change SD	0.05193; other groups 0.47949–0.69143
Exact visit window	Not retained in the analytical dataset; not recoverable here

The audit establishes spreadsheet consistency. The investigator confirmed the single-centre setting, index-treatment timing and distinct participant identities, including dapagliflozin records and matching semaglutide baseline pairs. No value or row was changed during manuscript revision. The narrower dapagliflozin HbA1c-change distribution is documented and addressed by sensitivity analysis.

Supplementary Table S2. Earlier exploratory predictor models

Earlier wider-model outcome	MAE	R^2
HbA1c percentage points	0.255	0.624
Relative weight %	1.932	0.103
Albuminuria log1p	0.201	0.220

Earlier exploratory models used 18 numerical baseline predictors and eight categorical predictors, including waist, glucose, lipids, liver enzymes, platelets and unharmonised medication strings. Their results preceded the consolidated 13-variable clinical set and are disclosed to avoid presenting the latter as a prospectively selected model. The present draft uses the consolidated model for interpretability and a same-predictor comparison with linear regression, not because it had the best observed performance. Neither analysis has external validation.

Partition sizes

Population	Fold	Outer train	Outer test	Inner train range	Inner validation range
All	1	447	112	298–298	149–149
All	2	447	112	298–298	149–149
All	3	447	112	298–298	149–149
All	4	447	112	298–298	149–149
All	5	448	111	298–299	149–150
No dapa	1	400	100	266–267	133–134
No dapa	2	400	100	266–267	133–134
No dapa	3	400	100	266–267	133–134
No dapa	4	400	100	266–267	133–134
No dapa	5	400	100	266–267	133–134

Counts are numbers of participants. Signature grouping is identical for all three continuous outcomes within each population. Inner ranges reflect the three tuning partitions within the corresponding outer training set.