**Supplementary Materials**

**Causal Analysis with BCAUS**

As described in the main text, the BCAUS model is trained using the joint loss . Here is the scalar ratio of to that is detached from the computation graph. The relative contribution of each loss component is tuned using hyperparameter, . The cross-entropy loss is calculated as:

Here is the treatment given to individual . To compute the bias loss, the propensity score is used to compute the inverse probability weight (IPW):

The mean squared error of the covariates weighted according to Eq. S2 is used to calculate the bias loss:

The two terms in the equation above represent the weighted means of the covariates for the treatment and control groups respectively. To assess balance, the standardized mean difference for each covariate is computed according to:

Here is the weighted mean of and is its weighted variance with weights assigned according to Eq. 3. The standardized mean difference can also be defined for the raw data without the weights, in which case and represent the unweighted mean and variance respectively.

The BCAUS model is implemented in Python using the PyTorch neural networks library. Each BCAUS model consists of two hidden layers with the number of neurons in each layer set to twice the number of input covariates. Rectified Linear Units (ReLU) activation is used for all layers except the last layer consisting of single neuron which uses sigmoid activation. Code for BCAUS definition and training has been published elsewhere. The learning rate was set to 0.001, the hyperparameter was set to 4 and the networks were trained for 1000 epochs. An early-stopping procedure was implemented where training terminated if all covariates remained balanced (i.e. standardized mean difference < 0.1) for more than 10 epochs.

For each clinical subgroup, all treatments with more than 35 treated individuals were chosen and BCAUS models were trained comparing every treatment with every other treatment. For a treatment pair and the estimated ATE values should be antisymmetric i.e. and for treatments, pairwise comparisons should suffice. However, since the propensity scores output by BCAUS are not calibrated probabilities, a small deviation from this asymmetric property (with differences much smaller than the standard error) is observed in practice. Therefore and were computed separately and a total of BCAUS models were trained. Prior to training all continuous covariates in each clinical subgroup were Z-scored to have zero mean and unit standard deviation. Propensity scores trimming was applied at the 0.01 level i.e. propensity scores below 0.01 were set to 0.01 and those above 0.99 were set to 0.99. This ensured that no individual received an inverse propensity weight > 10. A bootstrapping procedure was used to estimate the standard error for the ATE values. Inverse propensity weighted outcomes were picked at random and with replacement from the dataset and ATEs were computed between control and treatment individuals in each draw. The standard deviation of ATE values across 100 draws was reported as the standard error. ATE values and their standard errors for all pairwise treatment combinations were computed for each clinical subgroup and NMA was performed using the procedure described below.

**Network Meta-Analysis**

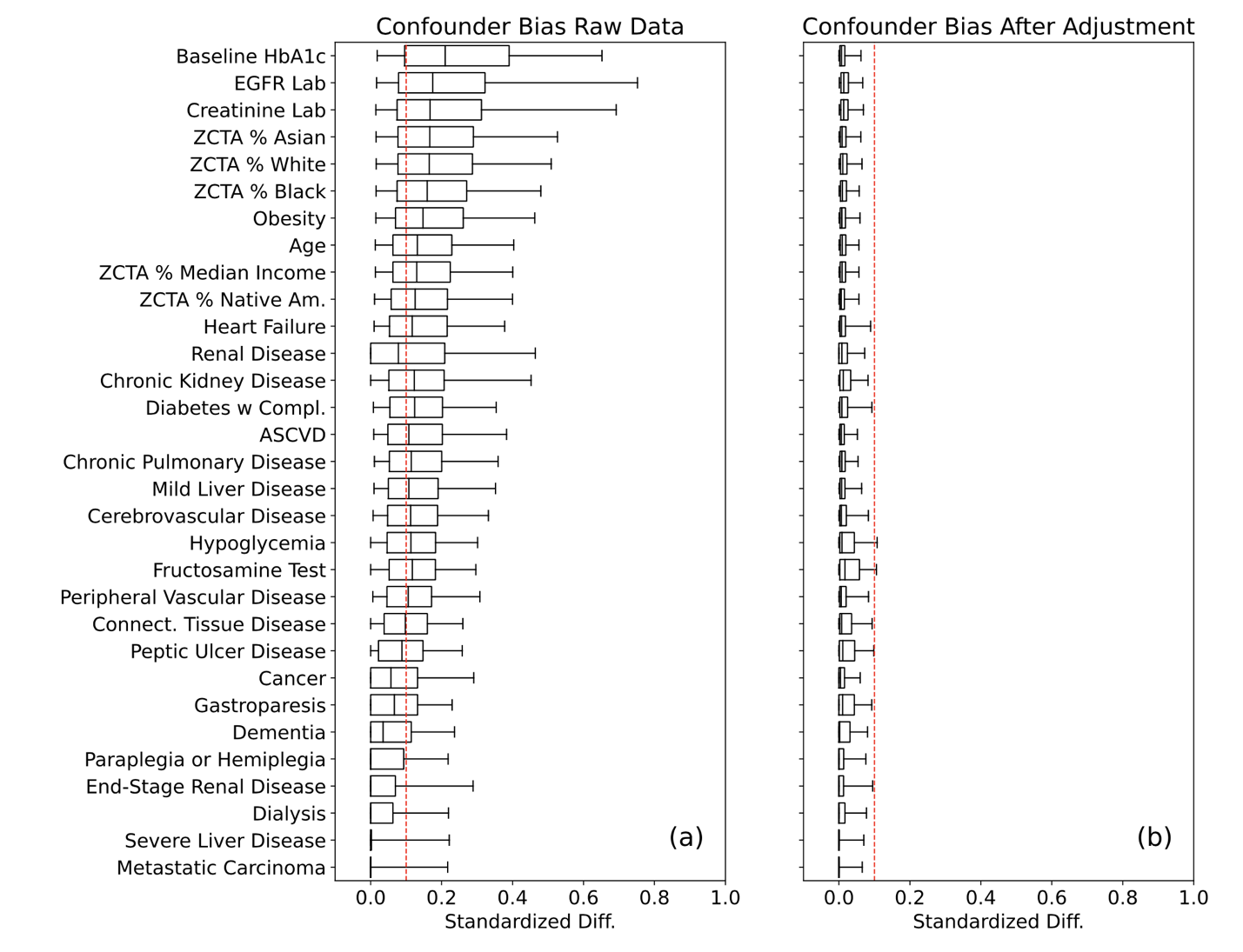
Network Meta-Analysis was performed with Python code developed using the PyMC3 probabilistic programming library. The network graph was encoded as a hierarchical, mixed-effects model:

Here is the ATE value measured by comparing treatment against treatment and is the corresponding standard error, is the ATE of treatment relative to the baseline treatment with . Uninformative priors were set for (the hierarchical standard deviation) and with the standard deviation for the sampling distribution of the latter set to 15 times the maximum absolute value of measured ATEs. A non-centered parameterization was chosen for the model because Markov Chain Monte Carlo (MCMC) samplers have difficulties sampling from the “Neal’s funnel” that can lead to divergent trajectories and biased results. A No U-Turns Sampler (NUTS) was tuned with 10,000 warm-up steps and 100,000 samples were drawn from 4 chains that were run simultaneously. The tuning samples and the first 50,000 samples in each chain were discarded. To compute SUCRA scores, 200,000 samples were drawn from the posterior distribution of and treatments were ranked for each draw. The SUCRA score for treatment was calculated as:

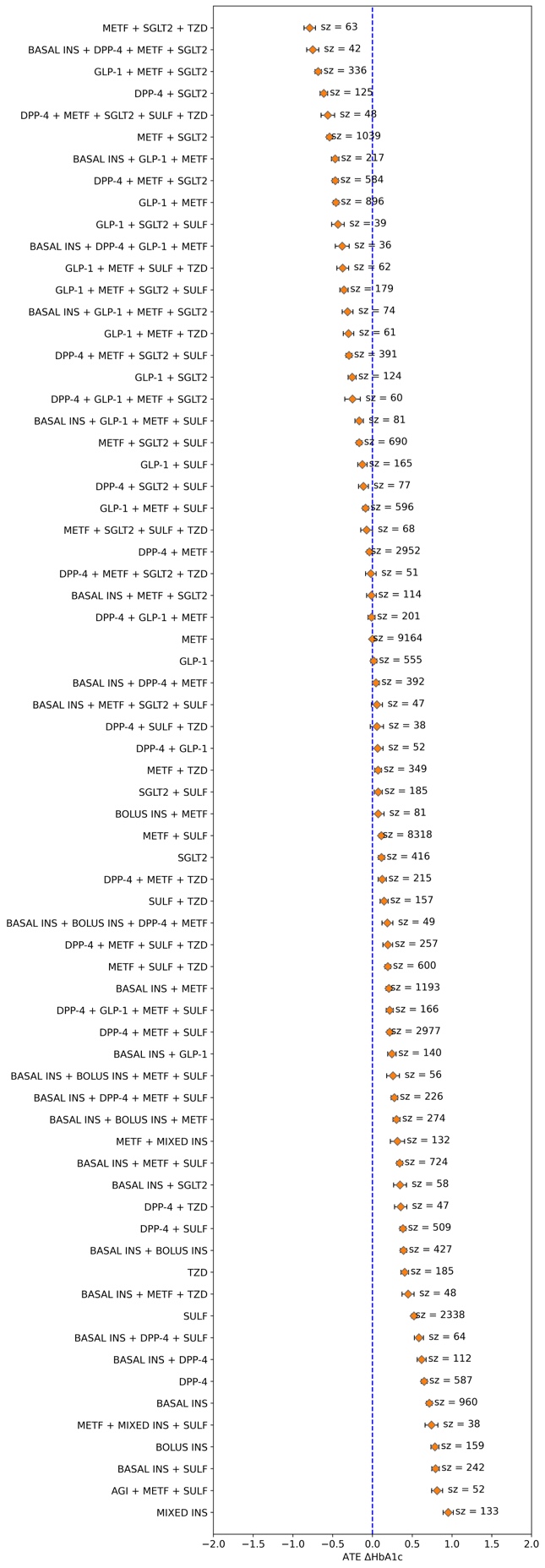
where is the mean rank for treatment across all draws and is the number of treatments (). Posterior samples were used to compute the mean and 94% credible intervals for ATE values of all treatments relative to Metformin, the baseline treatment.

**Supplementary Figures**

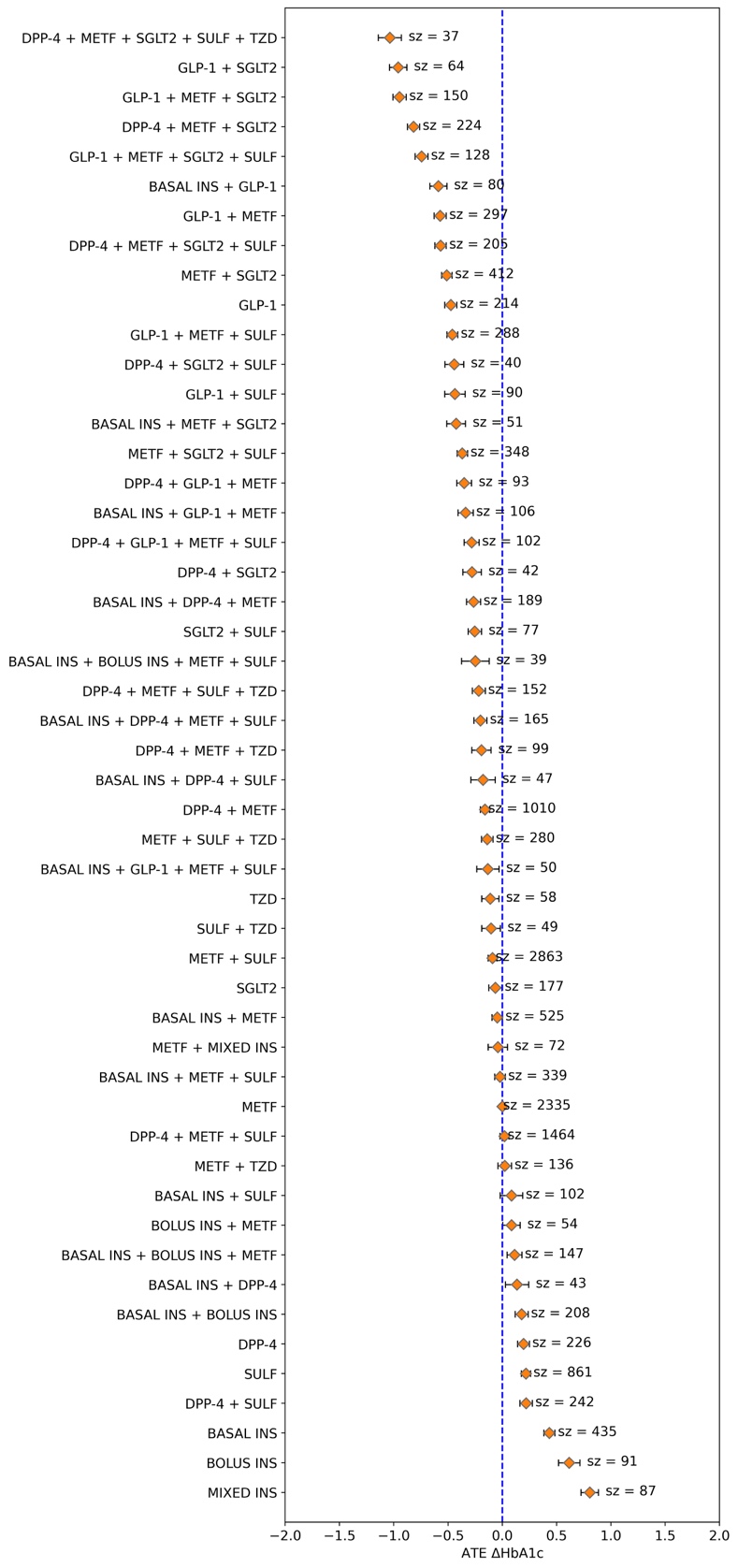
**Figure S1. Confounder bias adjustment with BCAUS** (a) Standardized difference plots for all confounders used in the analysis before (left) and after (right) adjustment using BCAUS. Whiskers represent 5th and 95th percentiles. 15,198 total networks were trained.



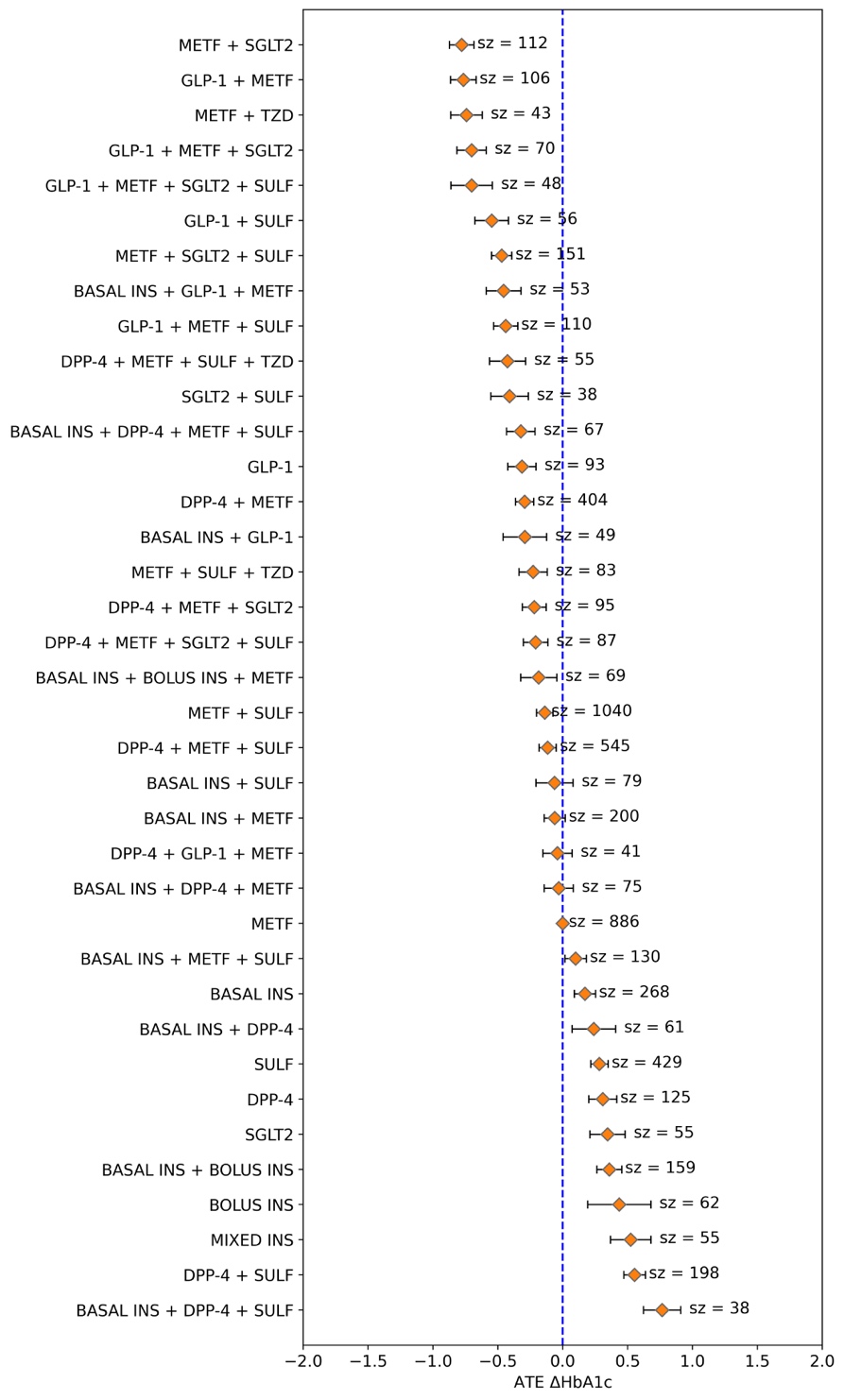
**Figure S2. Forest plot results of NMA for clinical subgroup 1** Average treatment effect of medication combinations for insulin naïve, under 65 and CCI<=2 stratum. Treatment cohort size is annotated (sz).



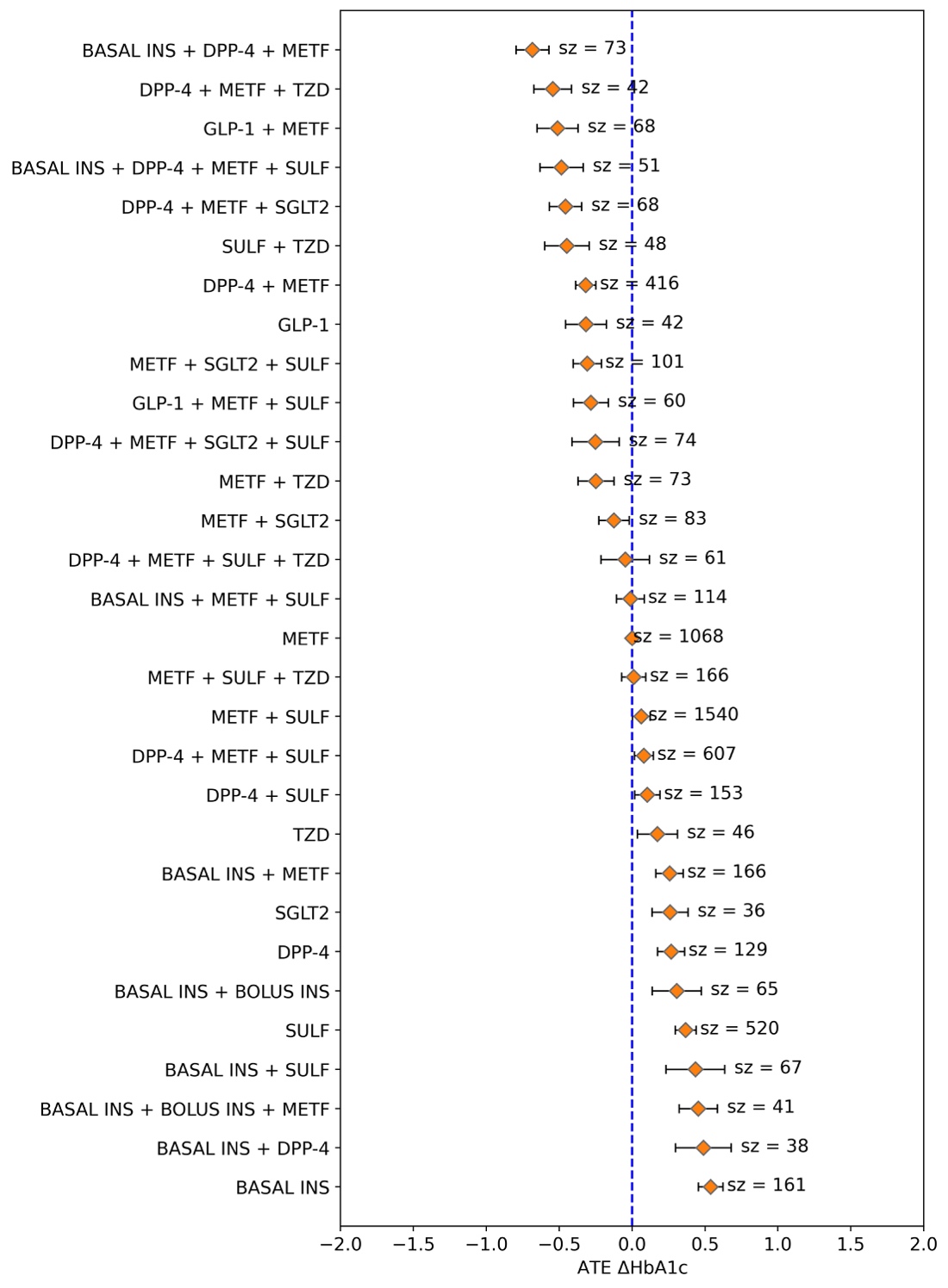
**Figure S3. Forest plot results of NMA for clinical subgroup 2** Average treatment effect of medication combinations for insulin naïve, under 65 and 2<CCI<5 stratum. Treatment cohort size is annotated (sz).



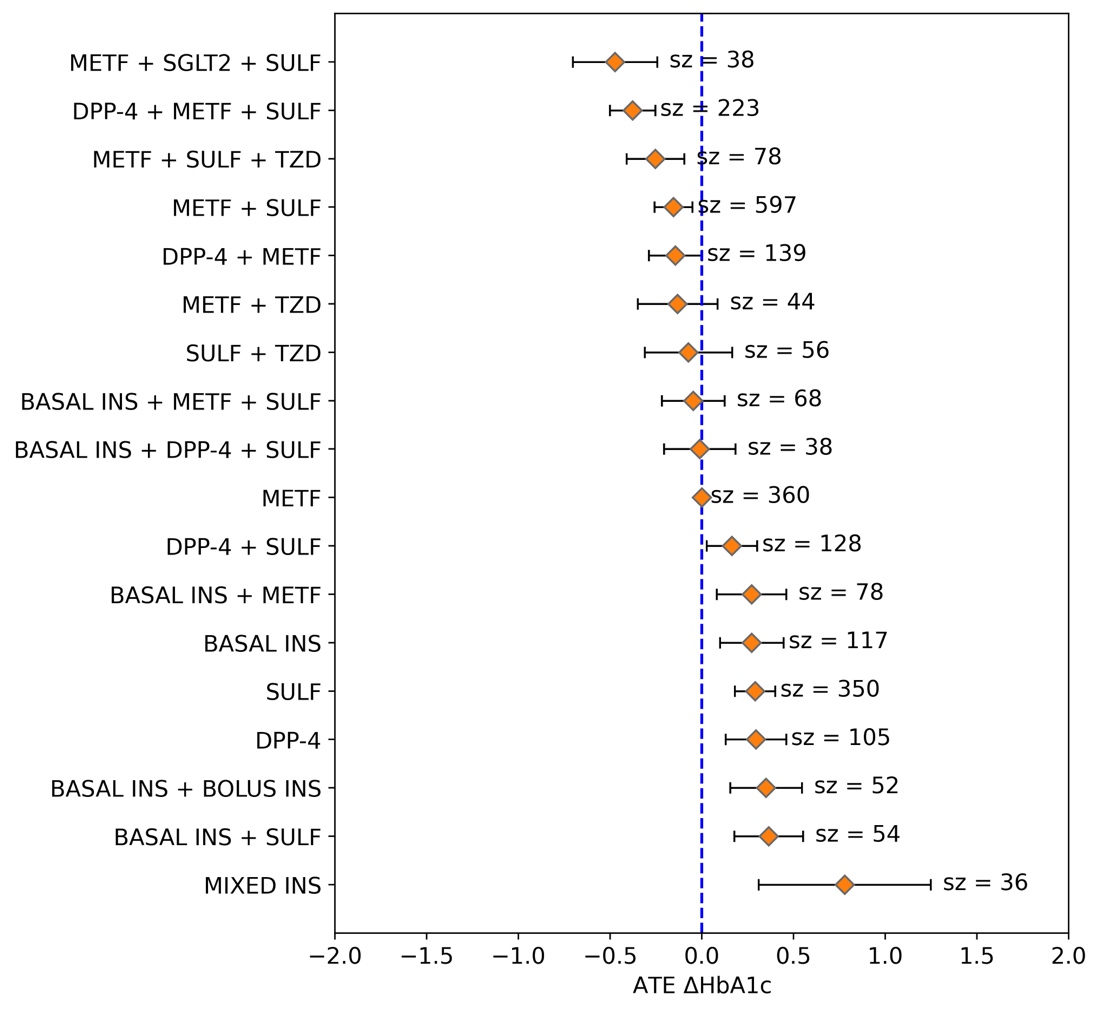
**Figure S4. Forest plot results of NMA for clinical subgroup 3** Average treatment effect of medication combinations for insulin naïve, under 65 and CCI>=5 stratum. Treatment cohort size is annotated (sz).



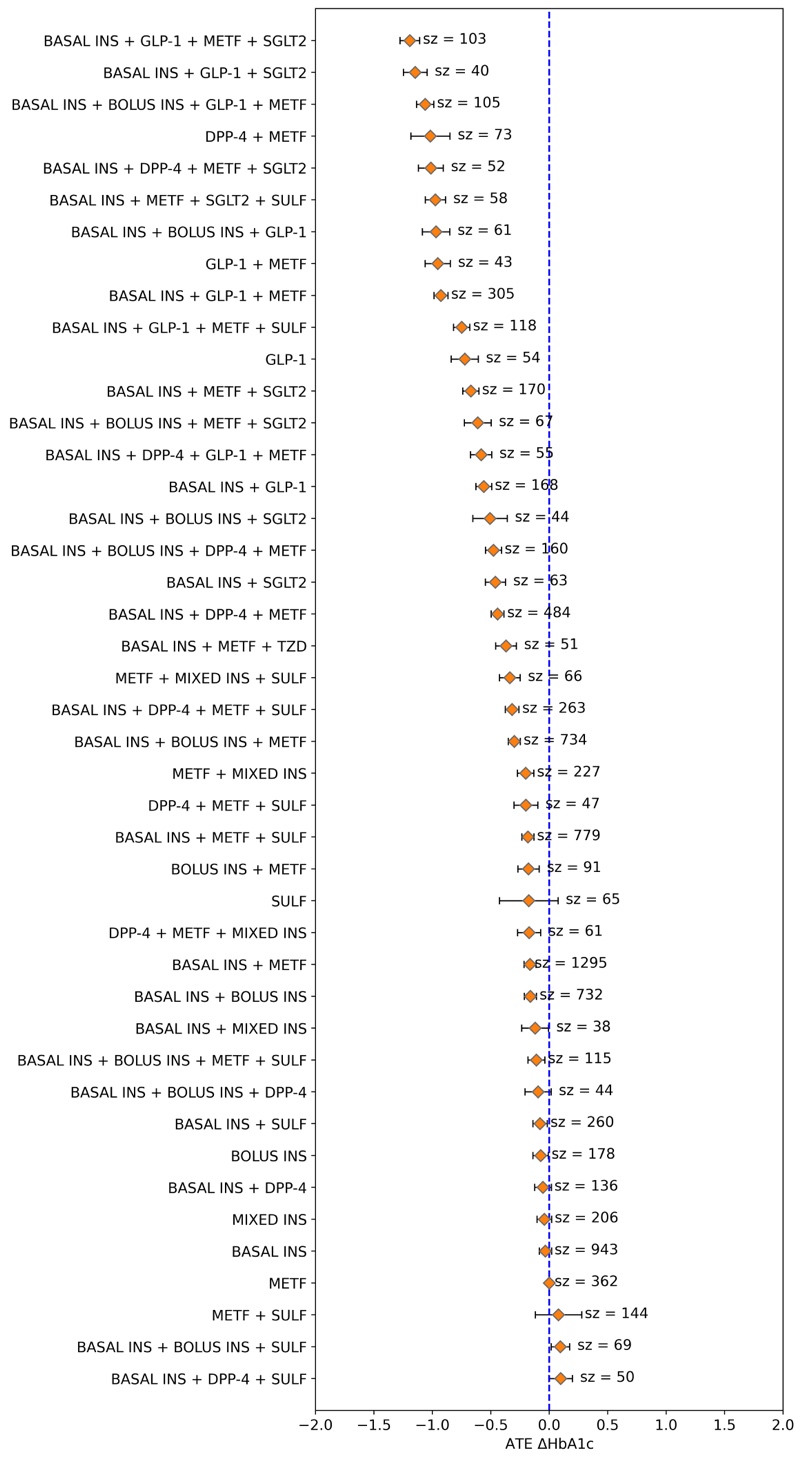
**Figure S5. Forest plot results of NMA for clinical subgroup 4** Average treatment effect of medication combinations for insulin naïve, over 65 and CCI<5 stratum. Treatment cohort size is annotated (sz).



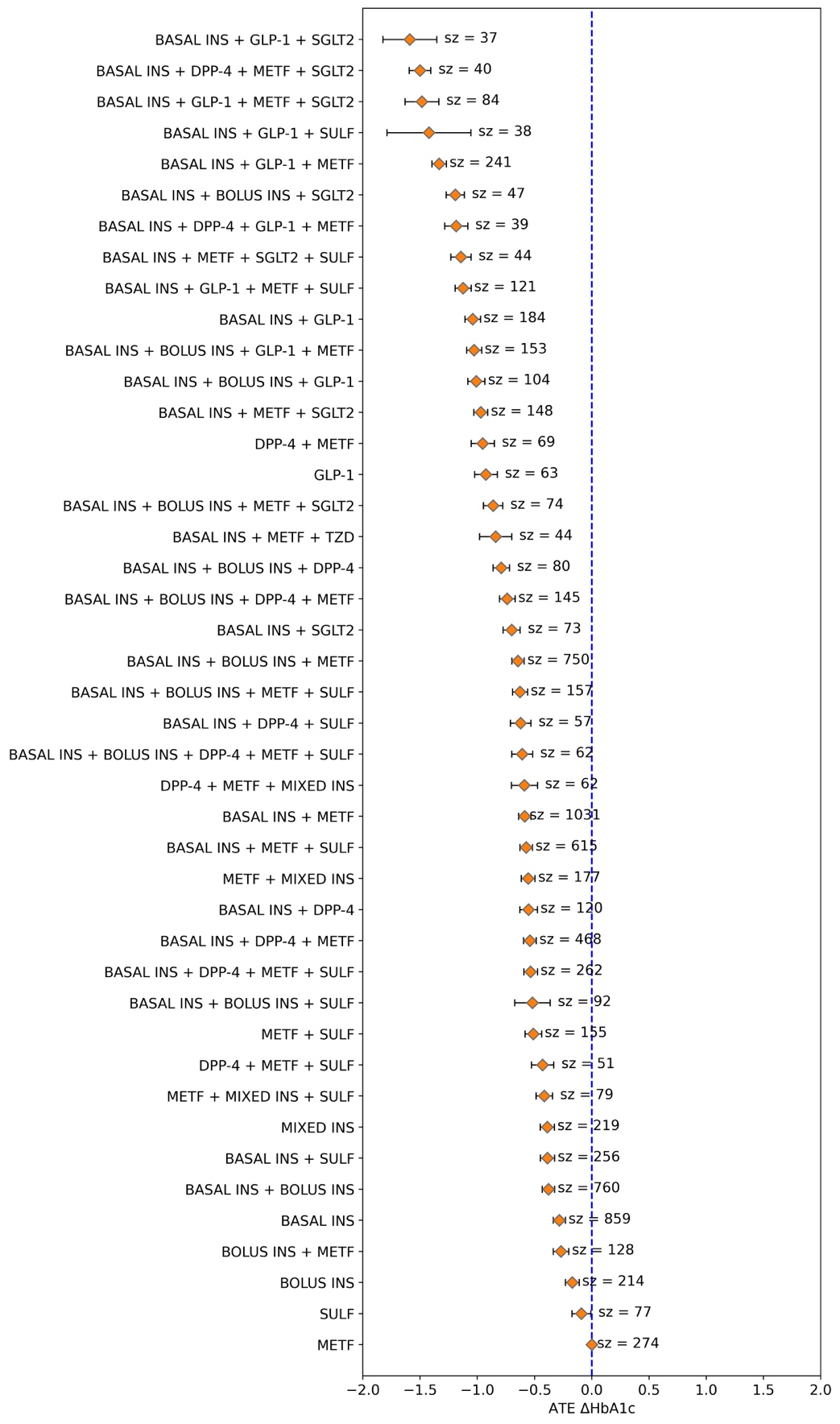
**Figure S6. Forest plot results of NMA for clinical subgroup 5** Average treatment effect of medication combinations for insulin naïve, over 65 and CCI>=5 stratum. Treatment cohort size is annotated (sz).



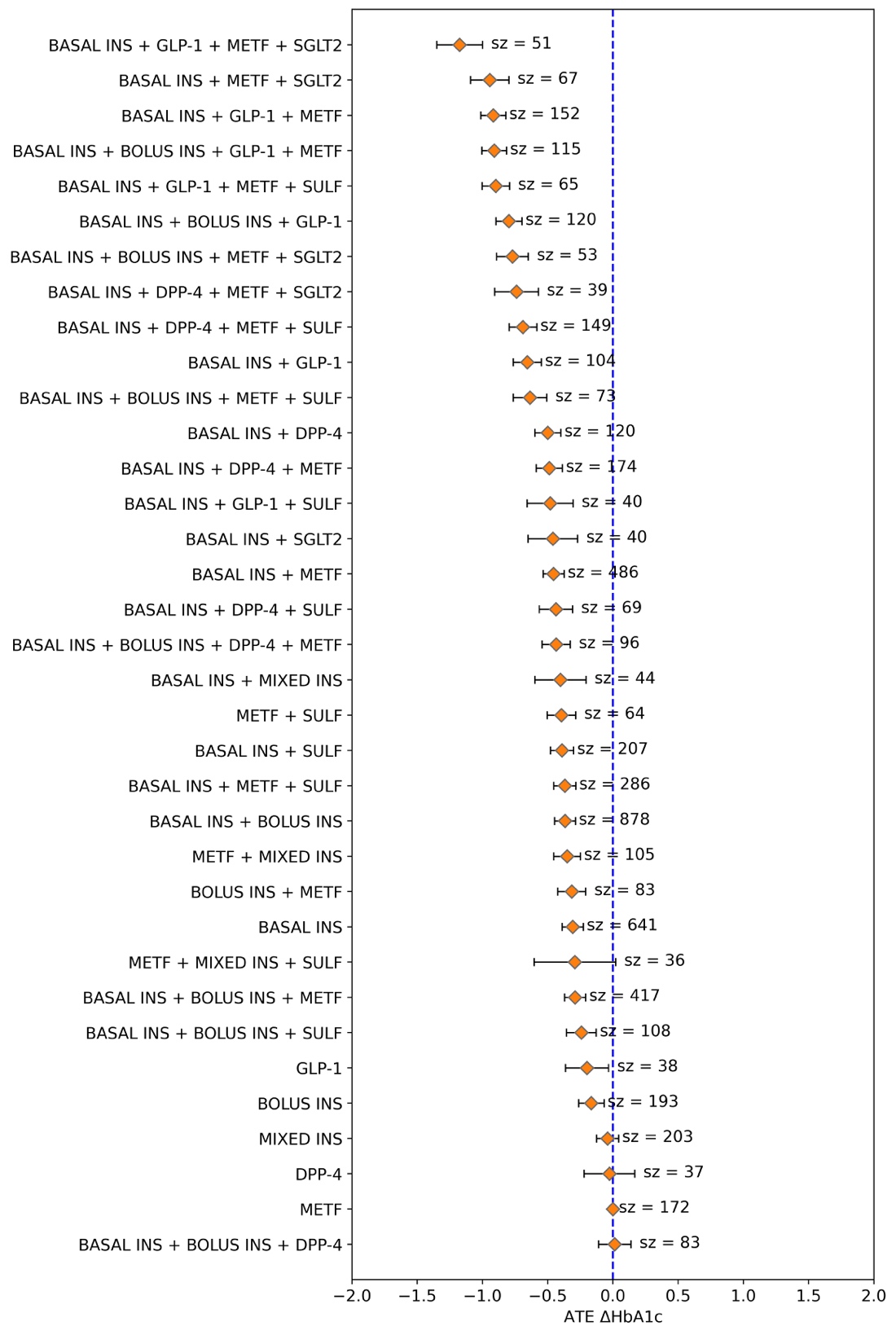
**Figure S7. Forest plot results of NMA for clinical subgroup 6** Average treatment effect of medication combinations for insulin dependent, under 65 and CCI<=2 stratum. Treatment cohort size is annotated (sz).



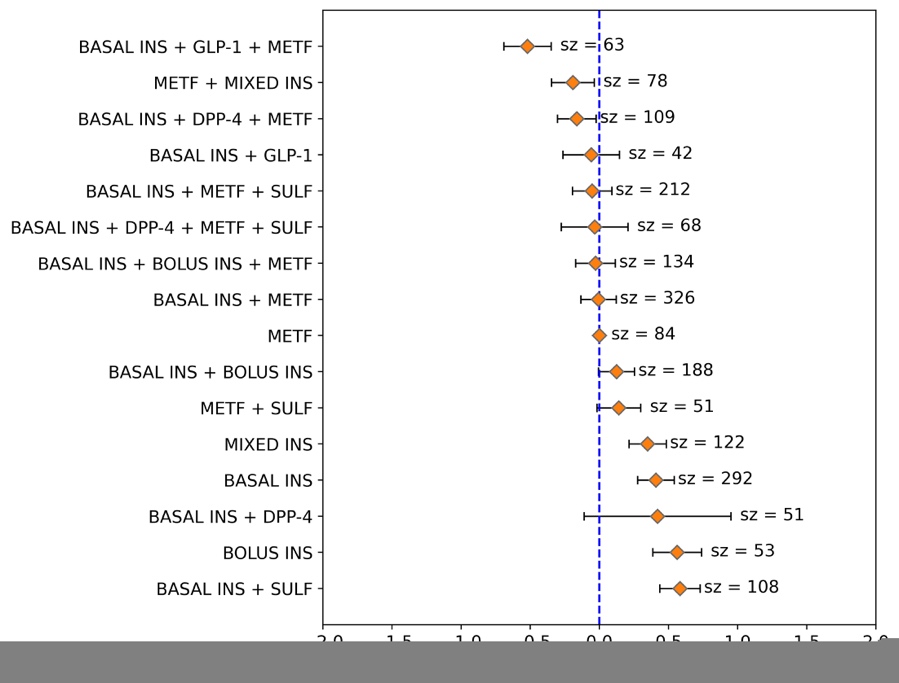
**Figure S8. Forest plot results of NMA for clinical subgroup 7** Average treatment effect of medication combinations for insulin dependent, over 65 and 2<CCI<5 stratum. Treatment cohort size is annotated (sz).



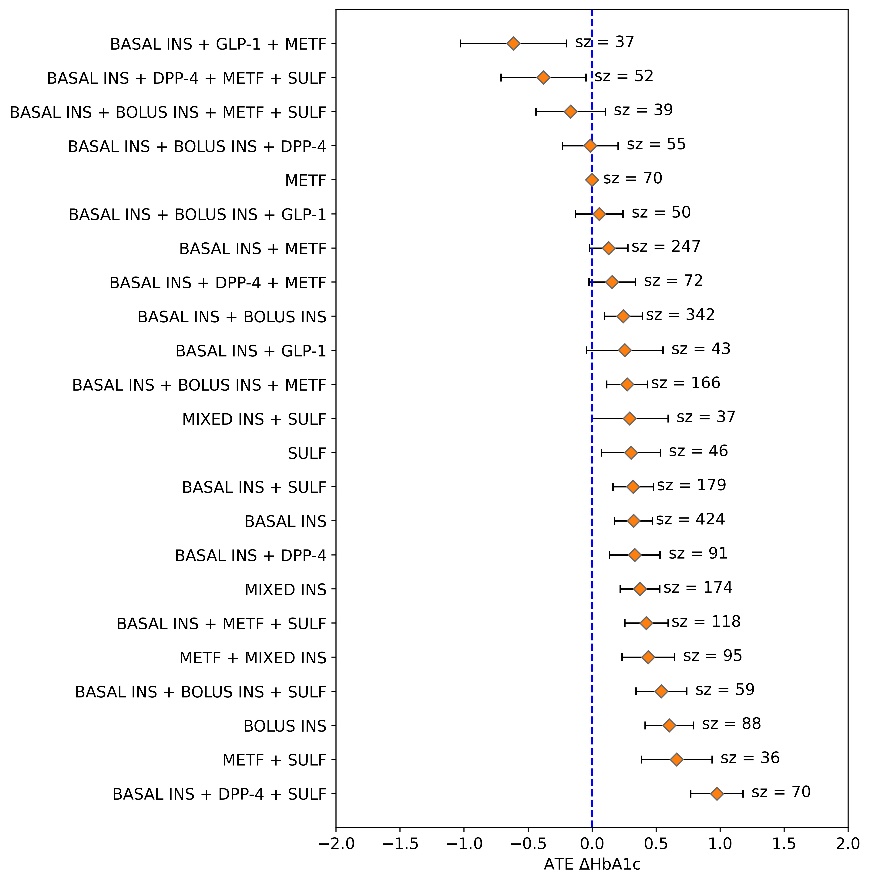
**Figure S9. Forest plot results of NMA for clinical subgroup 8** Average treatment effect of medication combinations for insulin dependent, under 65 and CCI>5 stratum. Treatment cohort size is annotated (sz).



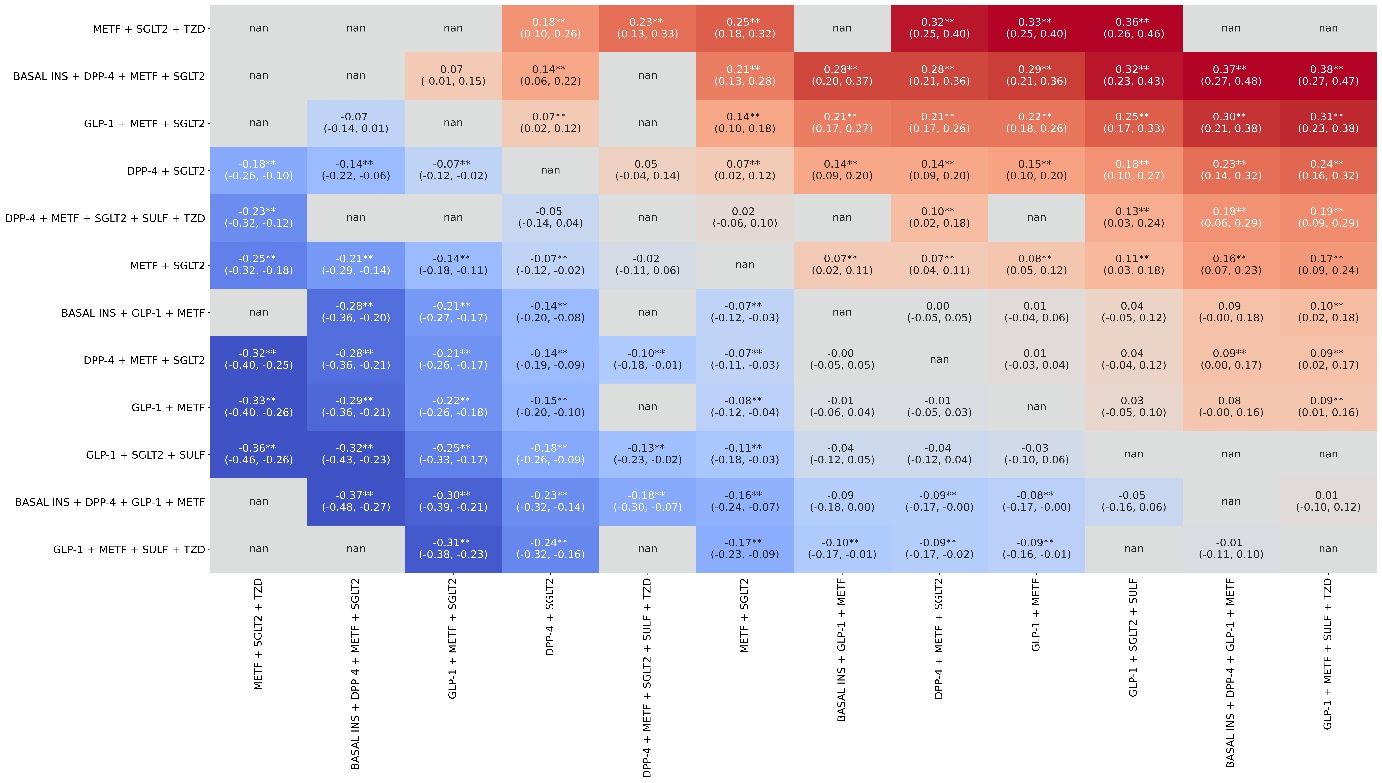
**Figure S10. Forest plot results of NMA for clinical subgroup 9** Average treatment effect of medication combinations for insulin dependent, over65 and CCI<5 stratum. Treatment cohort size is annotated (sz).



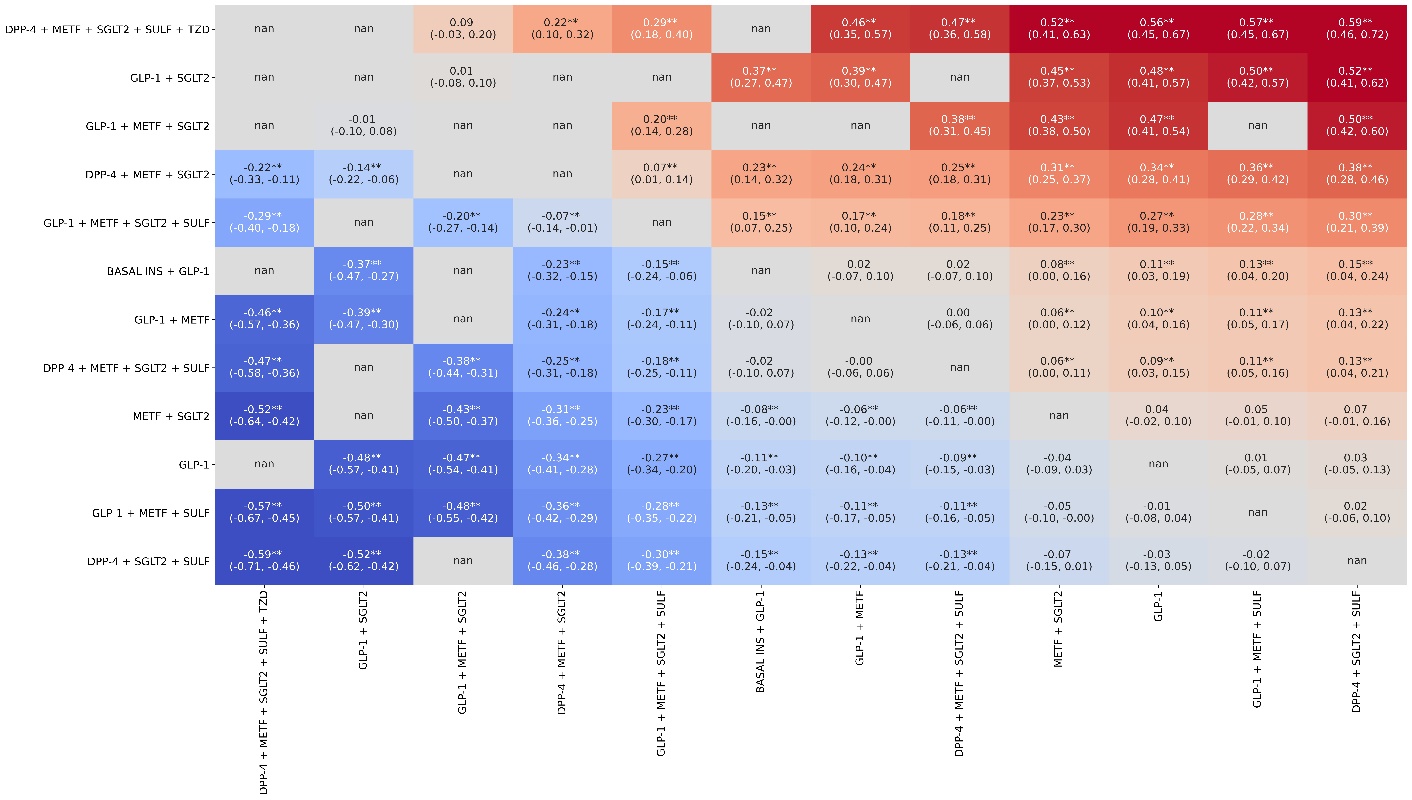
**Figure S11. Forest plot results of NMA for clinical subgroup 10** Average treatment effect of medication combinations for insulin dependent, over 65 and CCI>5 stratum. Treatment cohort size is annotated (sz).



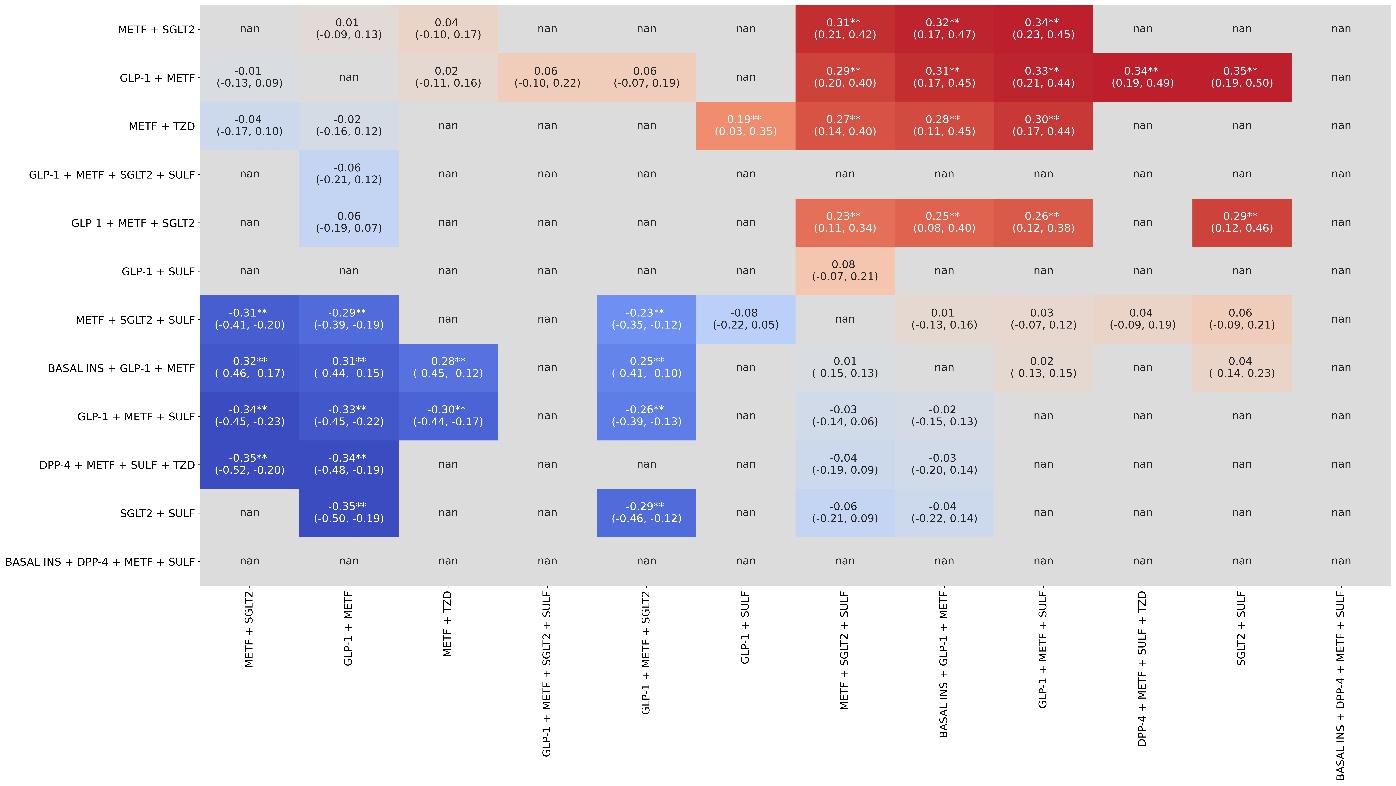
**Figure S12. Sample League Table for clinical subgroup 1**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin naïve, under 65 and CCI<=2 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.



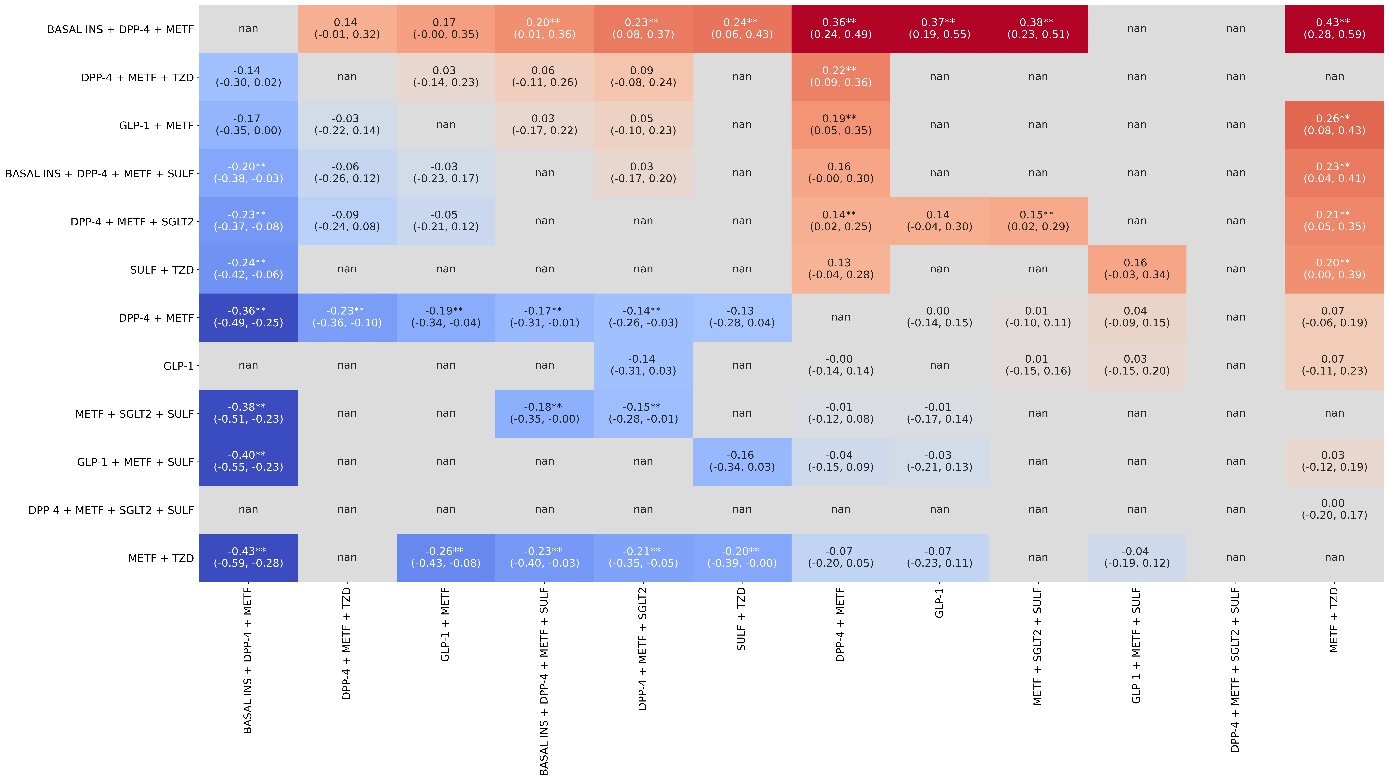
**Figure S13. Sample League Table for clinical subgroup 2**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin naïve, under 65 and 2<CCI<5 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.



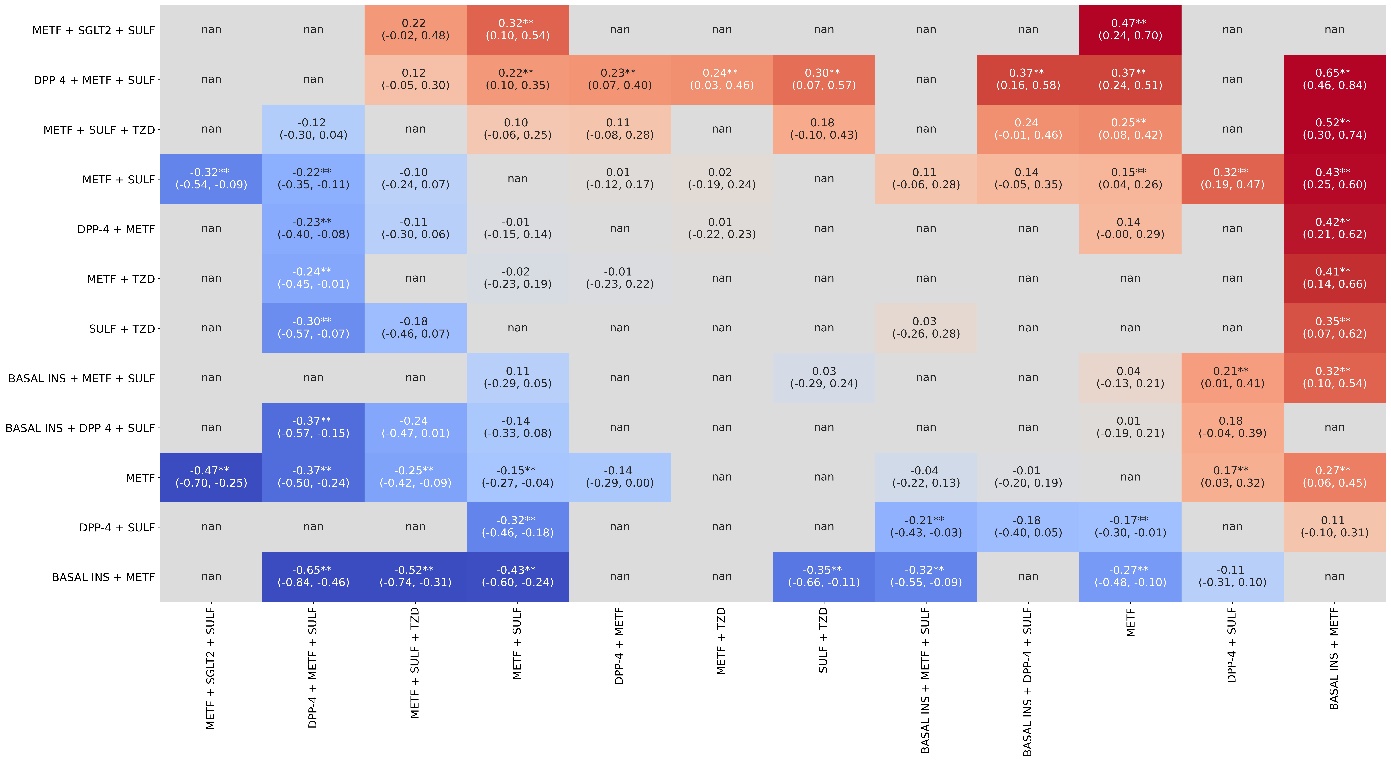
**Figure S14. Sample League Table for clinical subgroup 3**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin naïve, under 65 and CCI>=5 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.



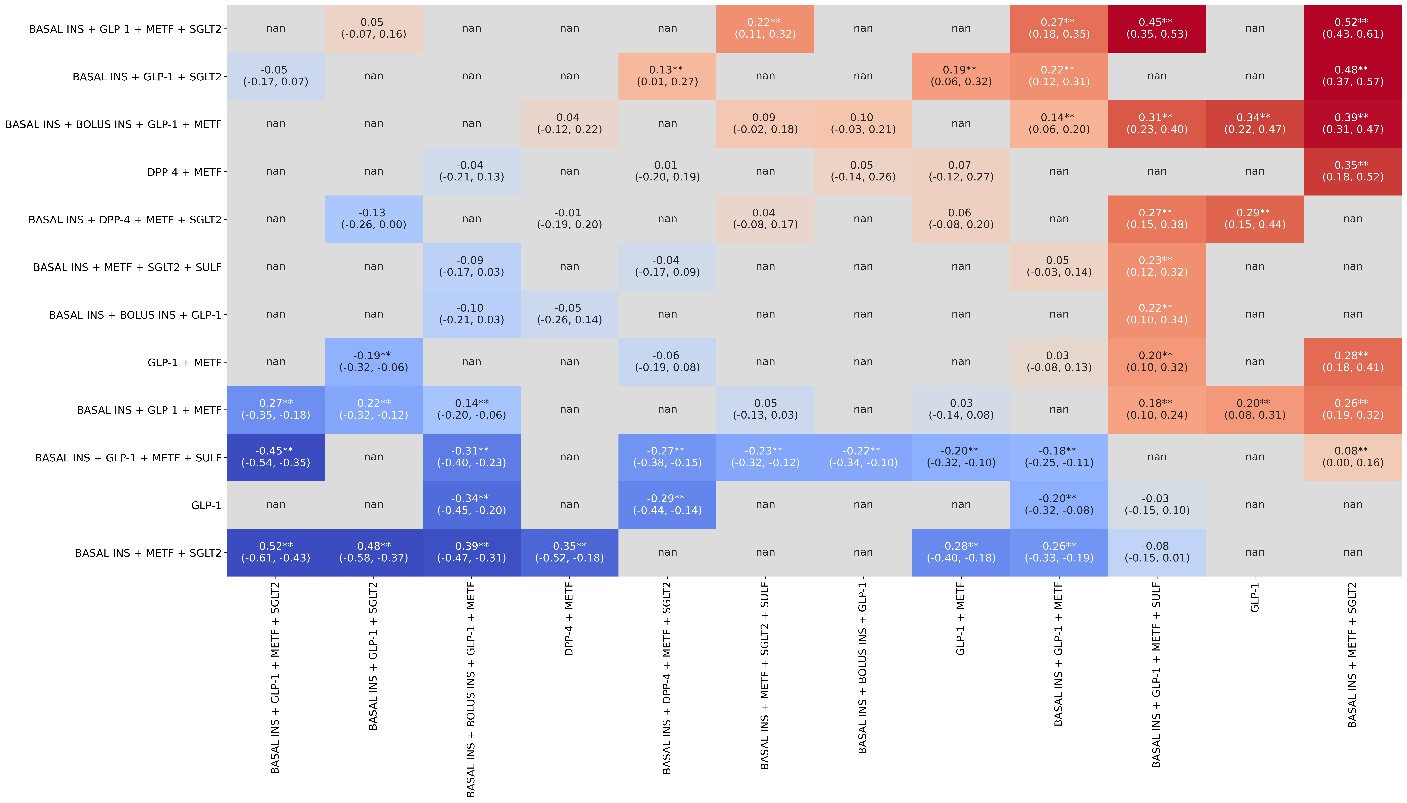
**Figure S15. Sample League Table for clinical subgroup 4**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin naïve, over 65 and CCI<5 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.



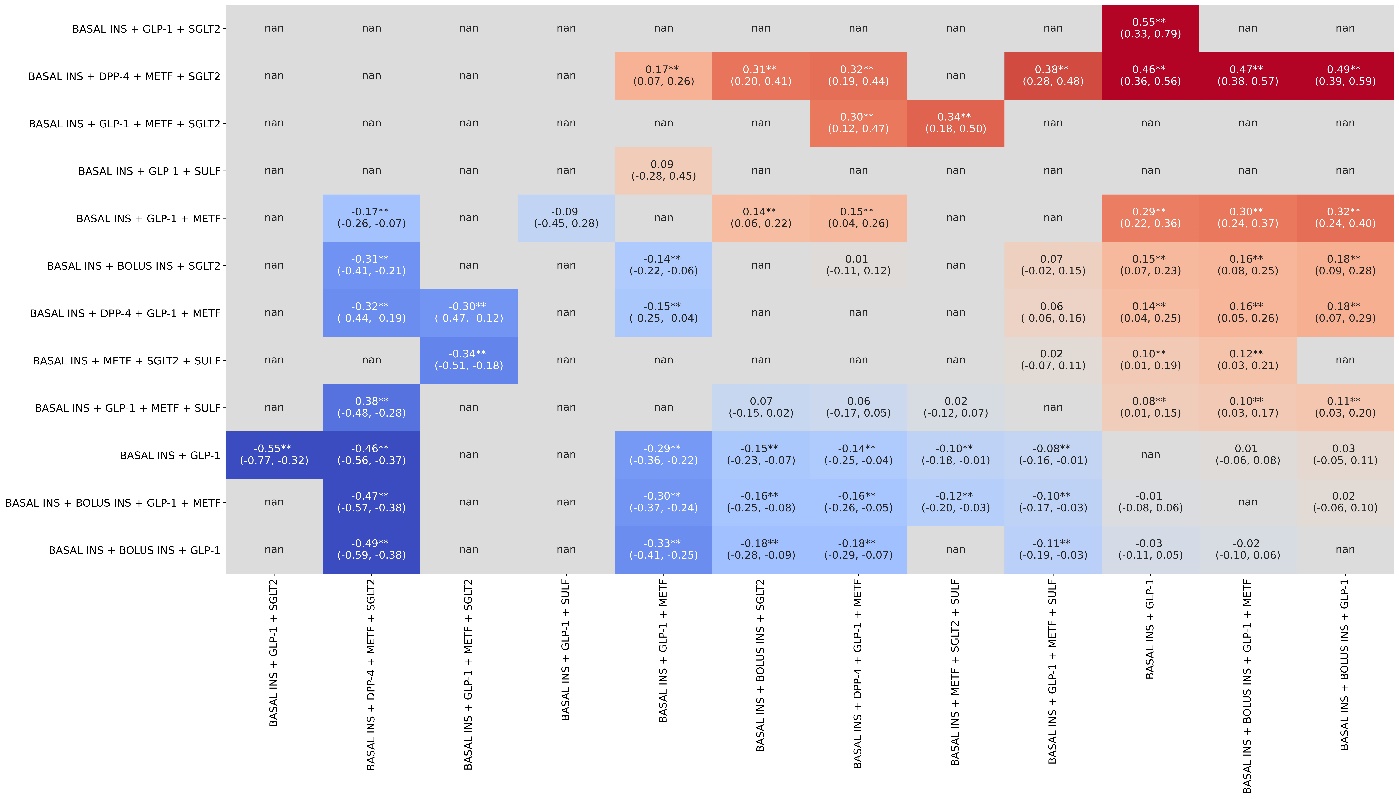
**Figure S16. Sample League Table for clinical subgroup 5**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin naïve, over 65 and CCI>=5 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.



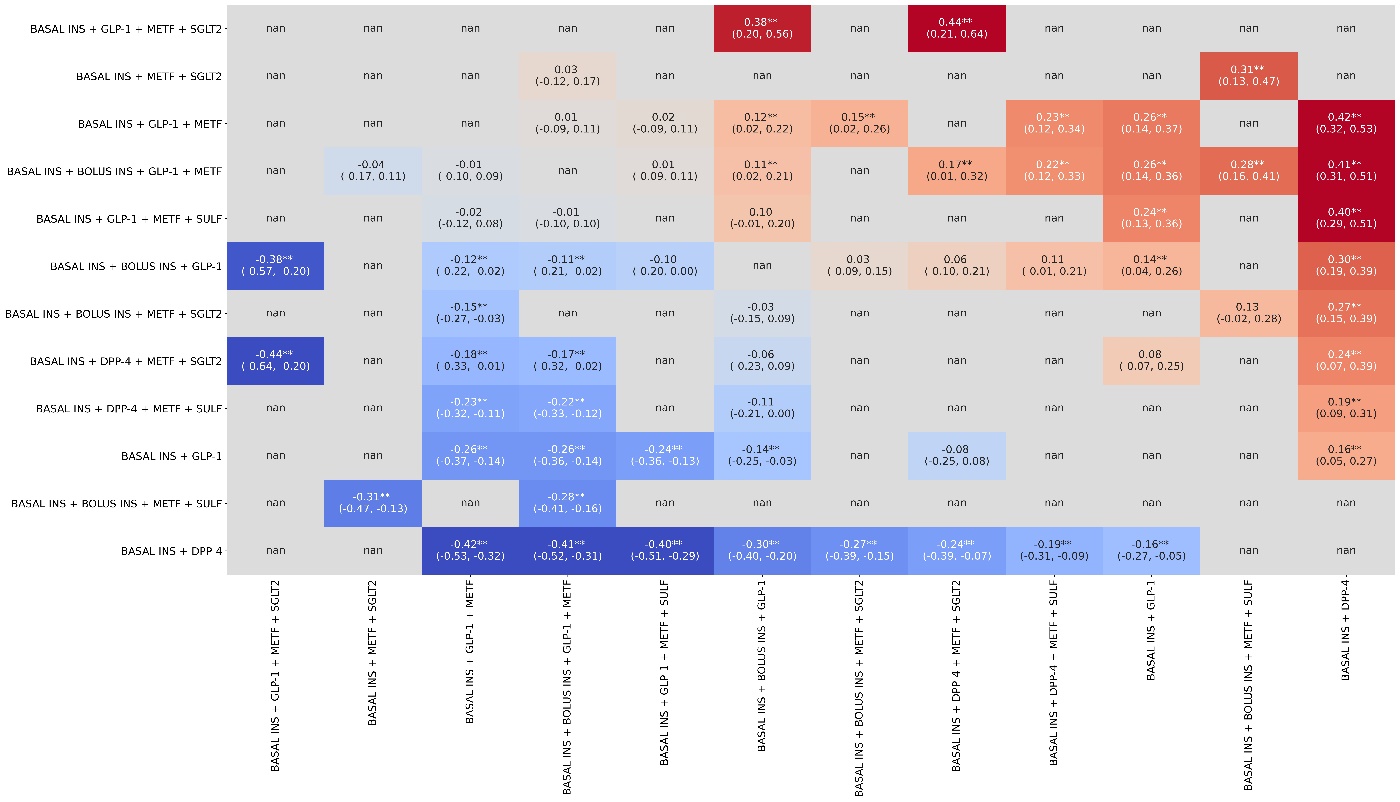
**Figure S17. Sample League Table for clinical subgroup 6**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin dependent, under 65 and CCI<=2 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.



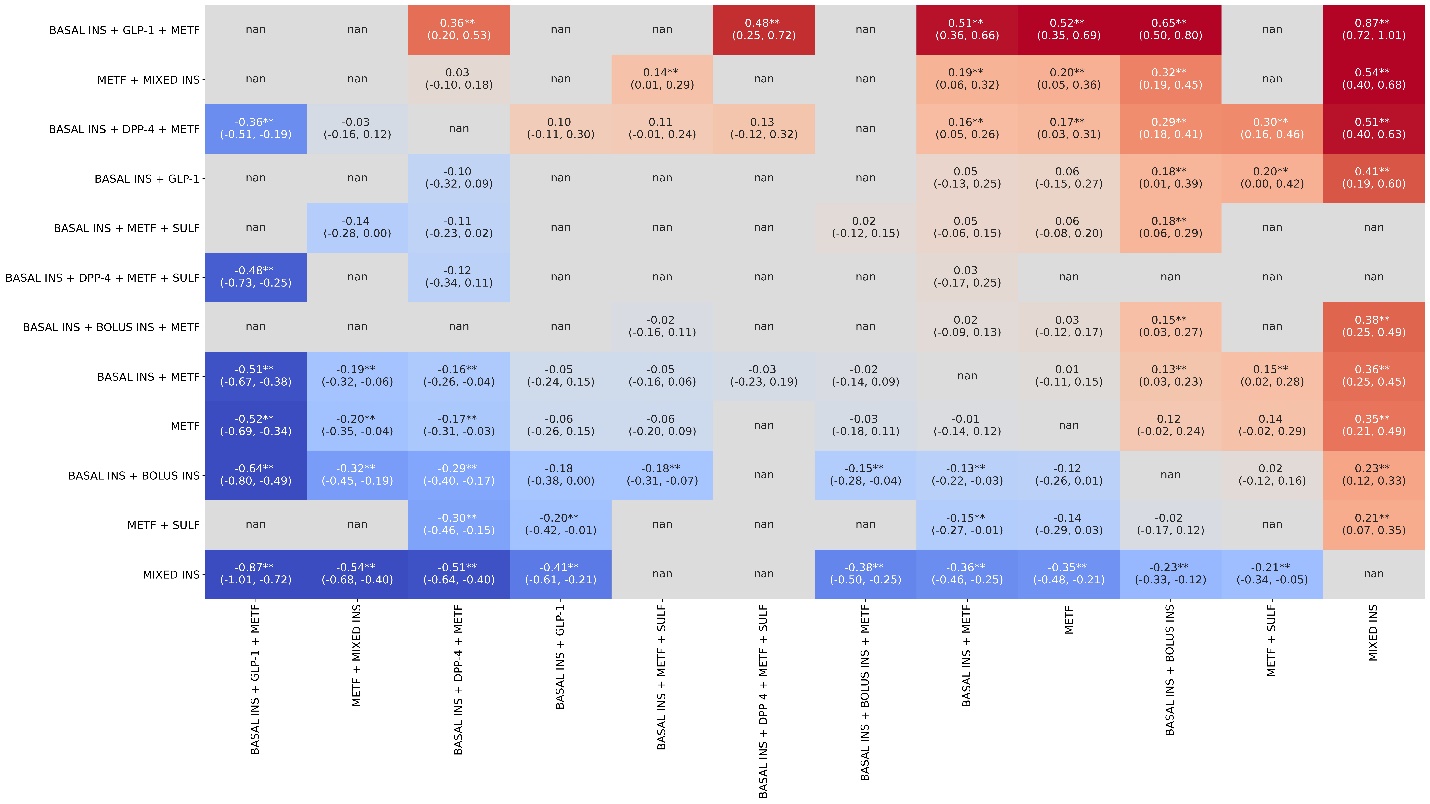
**Figure S18. Sample League Table for clinical subgroup 7**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin dependent, under 65 and 2<CCI<5 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.



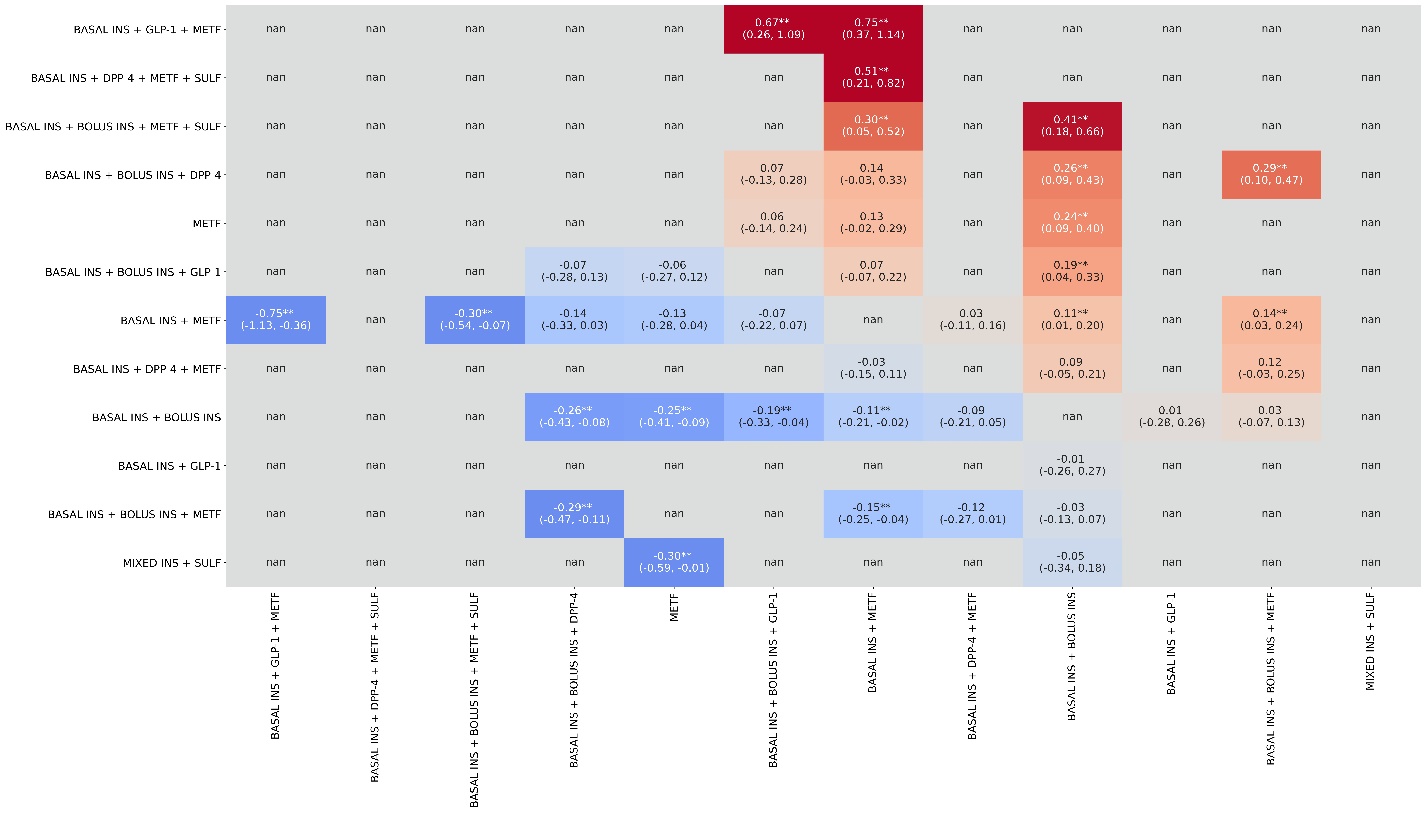
**Figure S19. Sample League Table for clinical subgroup 8**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin dependent, under 65 and CCI>=5 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.



**Figure S20. Sample League Table for clinical subgroup 9**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin dependent, over 65 and CCI<5 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.

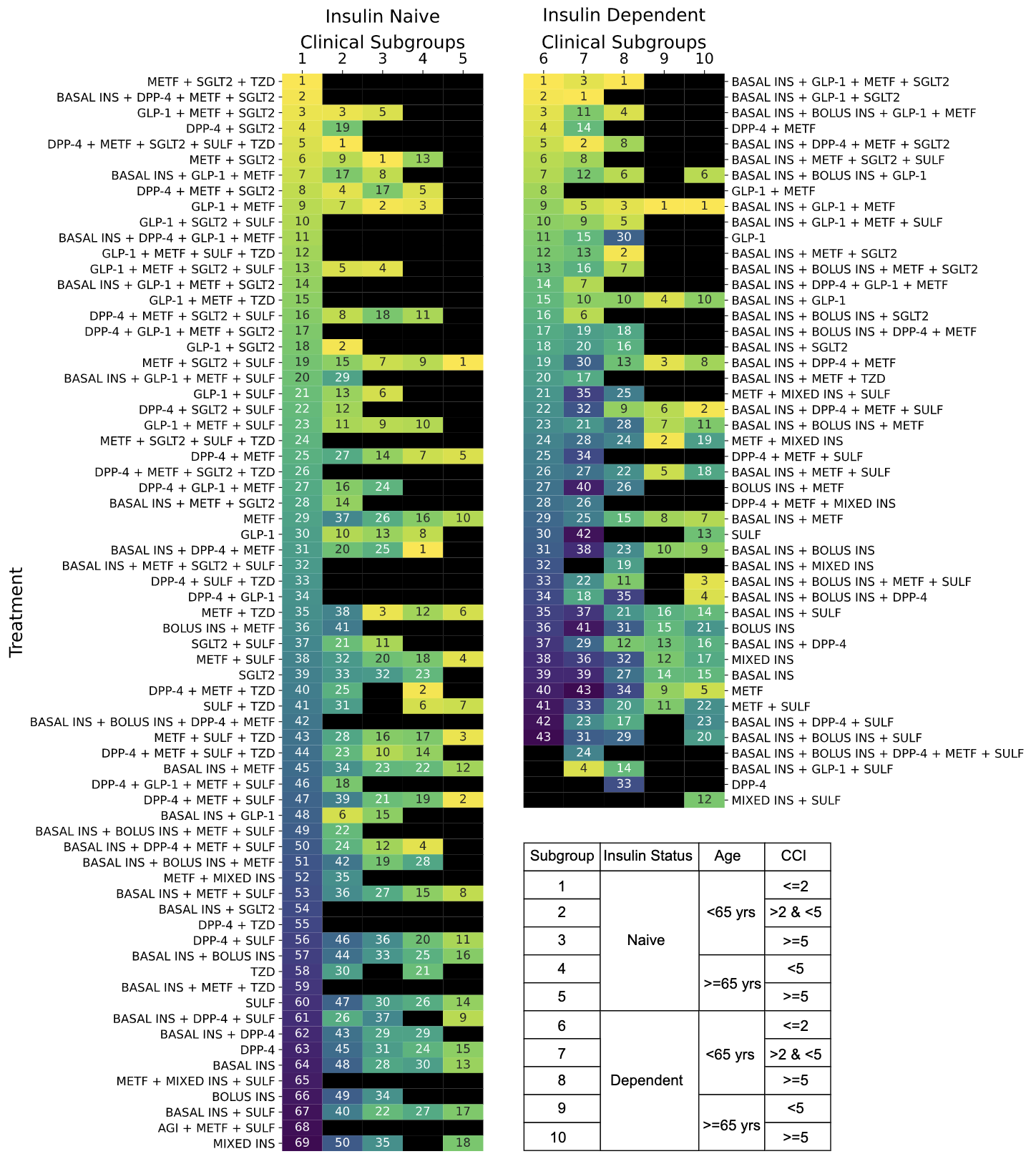


**Figure S21. Sample League Table for clinical subgroup 10**  League table showing pairwise differences for a subset of 12 treatment combinations after estimation with NMA evaluated on insulin dependent, under 65 and CCI>=5 stratum. Cells filled with nan represent trials where effect size estimation did not have all confounders balanced.



**Figure S22: Comprehensive Ranking Results of Causal Effect on Blood Sugar Reduction Attributable to Each Observed Treatment Strategy for Each Clinical Subgroup**

Treatments ranked according to their SUCRA scores for each clinical subgroup. Left panel shows Clinical Subgroups where prior treatment did not contain Insulin. Right panel shows Clinal Subgroups where prior treatment contained Insulin. Lower right table shows subgroup definitions. Abbreviations used: INS= Insulin, GLP-1 = Glucagon-Like Peptide-1 Receptor Agonist; SULF: Sulfonylureas; METF = Metformin; MEGL: Meglitinide; DPP-4 = Dipeptidyl Peptidase 4 Inhibitor; TZD = Thiazolidinedione; SGLT2 = Sodium-Glucose Transport Protein 2 Inhibitor; AGI = Alpha-Glucosidase Inhibitor; CCI = (unweighted) Charlson Comorbidity Index



**Figure S23: Dose Response Comparison Across High, Middle, and Low Ranked Treatment Strategies.** Average treatment effect of the recommender when comparing treatments ranked 1-3 vs those ranked 11 or below (top vs middle) and treatments ranked 4-10 vs 11 or below (middle vs bottom). Each panel represents the outcome for a single clinical subgroup.

Calendar

Description automatically generated with medium confidence