Study: Uncovering Hidden Risk Factors: A Bernoulli-based Stratification Method for Identifying High-Risk Type 2 Diabetes Subgroups

Research question

Can we identify latent high-risk subgroups for type 2 diabetes and uncover significant interactions using easily accessible features across sub-Saharan Africa to improve screening and clinical decision-making?

Objective 1 — Identify high-risk T2D subgroups in the Agincourt cohort using both comprehensive and clinically accessible feature sets

- **Table A1** All-feature solution set (penalty, literals, n, prevalence, divergence score).
- **Table 1** Easily accessible solution set (same columns)-to be used hereafter.
- **Figure 1** Cumulative-distribution overlays (BMI, age, waist) showing the distribution difference within and outside of each discovered subgroup.

Objective 2 — Validate the discovered subgroups, evaluating the effects of their risk factor cutoffs and benchmark them against established screening thresholds

- **Figure 2A** Forest plot of OR, 95% CI, p-value, comparing our subgroups' ORs with established cut-offs in literature.
- **Figure 2B** Venn diagrams and Jaccard similarity matrix score for overlapping subgroups.
- **Figure 3** Forest plots showing the effect sizes of all subgroups when assessed in a similar population (DIMAMO) before and after accounting for confounders using propensity score matching, and their corresponding significance level.
- **Table 2** Heterogeneity test using Cochran's Q to estimate the similarity of subgroup individuals with the same characteristics in the Agincourt and DIMAMO cohort.

Objective 3 — Assess the cross-regional generalizability of discovered subgroups through reciprocal discovery-transfer analysis

1. **Path 3a:** Assess the consistency of the effects of the discovered Agincourt subgroups in Nairobi and Nanoro.
    - **Figure 4A** Forest plot stacking ORs across the three cohorts.
    - **Figure 4B** Prevalence shifts (Nairobi/Nanoro relative to Agincourt).
2. **Path 3b:** Discover in Nairobi, and assess the consistency of the effects in Agincourt and Nanoro.
    - **Figure 5A** Forest plot stacking ORs across the three cohorts.
    - **Figure 5B** Prevalence shifts (Agincourt/Nanoro relative to Nairobi).
3. **Path 3c:** Discover in Nanoro, and assess the consistency of the effects in Agincourt and Nairobi.
    - **Figure 6A** Parallel forest plot.
    - **Figure 6B** Prevalence shifts (Agincourt/Nairobi relative to Nanoro).
4. **Predictive impact in targets**

- **Figures 7, 8, and 9** ROC curves and ΔAUC for each transfer path.

Objective 4 — Determine which risk-factor cut-offs are common across regions and which remain context-specific

- **Figure 10** Heat-map (or Upset plot) marking the presence of every literal cut-off across Agincourt-accessible, Nairobi, and Nanoro discovery runs.

Objective 5 — Discover and validate sex-specific high-risk subgroups across the study regions

- **Table A2** Male-specific and female-specific literals with ORs and prevalence (discovered in pooled Nairobi + Agincourt).
- **Figure 6** Side-by-side forest plots showing sex-stratified effect sizes across regions.
- **Supplement S5** Interaction model outputs (sex × subgroup term, CIs, p-values).