M2

Same data, same conclusions?

Replicating published diabetes and cardiovascular analyses on synthetic cohorts and measuring inference agreement.

Master's · 30 credits | Biostatistics · Epidemiology | Suits: KI (Epidemiology, Biostatistics & Data Science) · SU (Statistics)

The question

When a researcher runs a standard cardiometabolic analysis, such as diabetes outcomes or heart-failure survival, on a synthetic cohort instead of the real one, do they reach the same conclusions: the same direction of effect, overlapping confidence intervals, the same curves?

Why it matters

The point of synthetic data in research is that inferences carry over. Most evaluations report distance metrics instead of asking whether the scientific answer changes. A clear protocol for inference agreement is what a research group would accept before trusting a synthetic cohort. We start in diabetes and cardiovascular care because the conditions are common and the analyses are well established.

What you would do

01 Select three to five published diabetes and cardiovascular analyses that can be reproduced on public data: risk-factor regression, Kaplan–Meier and Cox survival analysis, comorbidity prevalence.

02 Re-run each analysis on the real seed data and on synthetic releases at several privacy budgets.

03 Measure agreement: effect-size difference, confidence-interval overlap, sign and significance agreement, curve distance.

04 Propose a compact inference-agreement report that could sit on a release certificate.

What you would produce

- A reproducibility protocol other groups can reuse
- Agreement results across analyses and budgets
- A short reporting template for inference agreement

Good fit if you have

R or Python, regression and survival analysis, careful reporting. Epidemiology or biostatistics background.

Data and access

Public seed data (CMS DE-SynPUF) and synthetic releases from day one. Work on real patient records only inside a partner's approved secure environment, with ethics approval. We hold no Swedish patient data.

Timing and place

Spring or autumn 2027, 20 weeks / 30 credits. Stockholm (Solna) or hybrid.

Supervision

Gaurav Raj Chattarki, Founder, Synthetix Health, together with your academic supervisor.

Questions or a proposal of your own? gaurav@synthetixhealth.com · synthetixhealth.com/careers#thesis
All projects run under a confidentiality and IP agreement agreed before work begins. We review drafts only for confidential information, never for results. Negative results are publishable.