



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Causal proteomic predictors of diabetes in South-Asia: a methodological and applied study using Stability Selection and Mendelian Randomization

LAUREA MAGISTRALE IN MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Author: PAOLO TRIFILIO

Advisor: PROF. FRANCESCA IEVA

Co-advisors: SOLENE CADIOU, NICOLE FONTANA

Academic year: 2025-26

1. Introduction

Type 2 diabetes is a chronic metabolic condition characterized by high blood sugar levels due to the body's ineffective use of insulin. Globally, type 2 diabetes is a growing public health crisis, with the sharpest increases seen in low- and middle-income countries. Bangladesh is particularly affected, currently ranking eighth worldwide in diabetes prevalence. In this context, it is critically important not only to identify early predictors of type 2 diabetes, but also to uncover causal factors that could serve as therapeutic entry points to prevent disease onset. Blood proteins in particular are interesting causal candidates because of both their functional relevance and suitability as drug targets. This thesis addresses both methodological and clinical research objectives. On the methodological side, it assesses the usability of SS as a causal selection method. Although SS is not a formal causal inference method, it's expected to perform well in this regard. The method prioritizes stable predictors, and since these are repeatedly selected across various model configurations, they may more likely represent underlying causal re-

lationships rather than random noise or simple associations. We will therefore apply SS to select proteins related to diabetes onset among a set of 7244 plasma proteins. To assess whether the proteins selected by SS can indeed be considered causal, their associations will be evaluated against existing evidence from the literature as well as through Mendelian Randomization (MR), a well-established causal inference approach. The second, clinical aim, of the thesis is to evaluate causal effects of all available plasma proteins on type 2 diabetes. The data used in this thesis come from the BELIEVE cohort study [1], a South-Asian cohort which contains genetic and clinical data and extensive proteomics measurements for around 10000 individuals.

2. Methods

2.1. Stability Selection

Stability Selection (SS) is a statistical method designed to identify a stable set of predictors, especially for high-dimensional datasets. Introduced by Meinshausen and Bühlmann in 2010 [2], it can be seen as an extension of LASSO (or

any variable selection model that depends on a regularization parameter λ). This method addresses a key limitation of methods like LASSO: small changes in the data often lead to different sets of selected variables, particularly when features are highly correlated or when the number of variables greatly exceeds the number of observations. In Stability Selection, data are perturbed multiple times through subsampling, and the selection frequency of each variable is recorded across a large number of iterations. A variable is considered “stable” if it is selected with a frequency higher than a chosen threshold π , across N repetitions and for a given regularization parameter λ . However, choosing optimal values for the parameters (λ, π) is non-trivial and can lead to suboptimal results. Indeed, in its first version SS lacked an analytical method to derive such values simultaneously and had to find the optimal value of one parameter given a fixed and arbitrary value of the other. To address this calibration problem, Bodinier *et al.* [3] proposed a likelihood-based metric known as the *stability score*. This score quantifies how unlikely it is that the observed pattern of variable selection arises under a null model of random selection. Assuming independence of subsamples and binomially distributed selection counts, they derived the likelihood of observing a certain classification of variables (as stably selected, unstably selected, or stably excluded). The stability score is then defined as the negative log-likelihood under the null hypothesis of equiprobability of selection:

$$S_{\lambda, \pi} = -\log(L_{\lambda, \pi}), \quad (1)$$

where $L_{\lambda, \pi}$ is the likelihood of the observed selection pattern given the model parameters. A higher score indicates greater stability, meaning that the selected set of variables is less likely to result from random noise. Therefore, an optimal pair (λ^*, π^*) can be formally determined as the one maximizing the stability score:

$$(\lambda^*, \pi^*) = \arg \max_{\lambda, \pi} S_{\lambda, \pi}. \quad (2)$$

Although Stability Selection is not a formal causal inference method, it identifies predictors that remain consistently selected across different model specifications, subsamples, or small data perturbations. Indeed, true causal relationships

tend to be stable, whereas spurious associations often disappear or change direction when the data or model changes. Therefore, predictors repeatedly selected by SS are more credible as potential causal candidates.

2.2. Mendelian Randomization

The Mendelian Randomization algorithm [4] is a method that uses genetic data as proxies for the exposure: the method will compute the association between a genetic instrument (called SNP) and both outcome (in our case, type 2 diabetes) and exposure (the blood proteins), and then leverage these effects to estimate the causal relationship between exposure and outcome. The MR algorithm requires the following assumptions (see Figure 1 for a visual representation):

1. The SNP Z is associated with the non genetic exposure X ;
2. The SNP Z is independent on possible confounding factors U that affect both the exposure X and the outcome Y ;
3. The SNP Z is related to the outcome Y only via the association with exposure X .

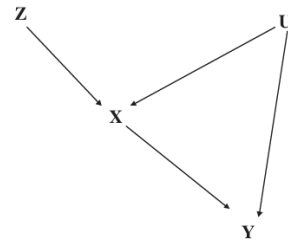


Figure 1: Visual representation of MR’s hypothesis. The arrows represent causal links.

The simplest way of estimating causal effects using MR is the so-called Wald estimator, because it makes use of just a single SNP for each exposure of interest. The estimated effect is computed as

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{ZY}}{\hat{\beta}_{ZX}} \quad (3)$$

where $\hat{\beta}_{ZY}$ is the regression coefficient for the outcome Y on the SNP Z , and $\hat{\beta}_{ZX}$ is the regression coefficient for the exposure X on the SNP Z . In case of binary variables, logistic regressions are being performed, so the results are in the log-odds scale.

Finally, MR can be conducted using either one

sample or two sample designs. The main difference lies in the data source: one sample MR estimates both SNP–exposure and SNP–outcome associations within the same cohort, whereas two sample MR uses two independent datasets. One sample MR is more susceptible to overfitting, which can exaggerate causal estimates and bias them toward the observational association. Consequently, one sample MR results require careful interpretation, especially in studies with small sample sizes or limited statistical power. Two sample MR is also susceptible to bias, as using an external study may involve a broader population, potentially introducing population bias in the genetic data. In this thesis, due to the limited cohort size ($n = 1270$), both one sample and two sample MR approaches were employed to balance the trade-offs between statistical power and potential biases.

3. Data engineering

The data we used came from the BELIEVE (BangladEsh Longitudinal Investigation of Emerging Vascular and nonvascular Events) cohort study, a large-scale research initiative aimed at quantifying the burden of non-communicable diseases in urban, rural, and urban slum areas of Bangladesh. It consisted of 9934 initial patients for which 7244 protein measurements (that were inverse-normal transformed and then scaled) were available, as well as 6 relevant confounding factors (age, sex, smoking status, kidney disease, fasting time, BMI), and information about type 2 diabetes. In particular, for each patient we had both prevalent and incident type 2 diabetes. What this means is that at a certain time point t , protein levels were measured, and each patient’s diabetes status was recorded. If a patient was already diabetic at this time, we referred to this as prevalent diabetes. At a later time point $t^* > t$, diabetes status was assessed again. New cases identified at this stage were classified as incident diabetes. Since causal links could only be established for patients who were non-diabetic at baseline (t), we had to restrict the number of patients to 1391.

To prepare the data for Stability Selection and LASSO modeling, two versions of the dataset were used: the full dataset (D) and a reduced version (\tilde{D}). The reduced dataset was

obtained by conducting a Protein-Wide Association Study (PWAS), selecting 447 proteins that showed a statistically significant association with type 2 diabetes after Bonferroni correction, with the aim of removing "noise" proteins.

To adjust for confounding variables, we computed residual matrices: for each protein, its values were regressed on the covariates, and the residuals were extracted. This process produced two residual matrices, R and \tilde{R} , corresponding to the full and reduced datasets. These matrices capture the portion of protein variation not explained by confounders and are used as inputs for subsequent modeling.

For the MR analyses, two main datasets were required:

- the *outcome_data* contained the SNP–outcome associations. The data was different between one sample MR and two sample MR. To generate the one sample MR outcome data, patients were linked to their genetic data, and associations were computed adjusting for covariates (the same as in the Stability Selection analyses) plus the first ten genetic principal components to account for population structure. Due to incomplete patient data, the cohort was reduced to 1270 individuals, limiting statistical power. The outcome data for two sample MR was obtained using an external outcome data from Loh *et al.* [5], which included a much larger population of 50533 South-Asian individuals.
- the *exposure_data* contained the SNP–exposure associations, and was obtained from a pre-existing study. Since we adopted a single-SNP MR approach, we included only one SNP per protein. To ensure the strongest possible instrument, we retained the most significant *cis*-SNP for each protein. This was done because *cis*-acting SNPs are the genetic variants located close to the gene encoding the protein, therefore making them more suitable as instruments. Moreover, only proteins whose SNP was still present in the outcome data could be used for the analysis. These restrictions reduced the protein set from 7244 to 1477 proteins for one sample MR and from 7244 to 855 for two sample MR.

4. Results

4.1. Stability Selection

The first step in the Stability Selection analysis was to assess the consistency of the model, i.e. understanding the impact of different input sets on the output. To this end, four different models were implemented, each using a different input set:

- Model 1 ($M_{baseline}$) was the reference model. It used the residual matrix \tilde{R} as input and identified 20 proteins, of which 18 were unique, with 2 detected through two separate measurements.
- Model 2 ($M_{perturbed}$) used a modified input set composed of the 20 proteins identified by $M_{baseline}$, combined with 427 randomly selected proteins (for a total of 447). Its purpose was to assess whether adding random noise could affect the stability of the selection.
- Model 3 (M_{random}) used an entirely random set of 447 proteins from the full list. It served to evaluate how many of the proteins identified in $M_{baseline}$ would be recovered when included in the input.
- Model 4 (M_{full}) used the complete residual matrix R , including all 7244 proteins, without any preselection. The aim was to test whether the inter-correlation among all proteins would influence the final selection.

These four configurations allowed for a comprehensive evaluation of Stability Selection’s robustness and its sensitivity to input variations. To fully assess the role of chance in feature selection, all models were run 100 times with different random seeds. This included the four SS models and two LASSO models: one with preselection ($LASSO_{reduced}$) and one using the full set of 7244 proteins ($LASSO_{full}$). The results were overall consistent across runs, with important differences between fixed and variable input configurations:

- Fixed input models ($M_{baseline}$ and M_{full}) consistently produced the same output across all runs. The only difference was a loss of three proteins from $M_{baseline}$ to M_{full} . This variation is expected and not concerning: the more correlated variables are added, the more LASSO may pick one of a correlated group interchangeably across

subsamples. Therefore, selection probabilities tend to decrease.

- Variable input models showed more variability, but still strong consistency: $M_{perturbed}$ successfully recovered all 18 unique baseline proteins in every run (though it also included additional ones due to the noisy input) and M_{random} managed to detect all baseline proteins that happened to be included in the input set in 87% of the runs (again with some extra detections).
- LASSO models proved less reliable for identifying a stable set of predictors. Neither was able to recover all the 18 unique stable proteins of Stability Selection, and the number of selected features varied more significantly compared to Stability Selection.

All these analyses showed that the automatically calibrated SS method is stable when applied to a single dataset, unlike LASSO. The 20 proteins detected by $M_{baseline}$ (which comprehend the 18 unique proteins, with 2 being measured twice) will be therefore denoted as the SS proteins. Regarding the predictive performance of Stability Selection compared to LASSO, we observed that, although stable predictors are not necessarily the most predictive by definition, SS achieved comparable performance to LASSO, with only a minimal reduction in predictive power. Given its additional advantage of consistently identifying stable subsets of variables, Stability Selection can be considered to outperform LASSO overall.

Finally, we performed an extensive literature review on the SS proteins. This process showed that SS is promising when looking for causal predictors, since 8 out of the 18 unique proteins (*N-terminal pro BNP*, *Adiponectin*, *Kallistatin*, *Growth Hormone Receptor*, *Epidermal Growth Factor Receptor*, *Afamin*, *Insulin-like growth factor-binding protein 2* and *Stromal cell-derived factor 1*) have strong evidence of a possible causal role in the development of type 2 diabetes.

4.2. Mendelian Randomization

The Mendelian Randomization analyses were conducted both on the SS proteins (to investigate their potential causal role in type 2 diabetes) and to all available proteins, to capture

additional causal candidates that SS might have missed and to better address our clinical question. Given the limited size of our cohort (1270 patients), one sample MR was expected to suffer from limited statistical power, increasing the risk of false negatives. To address this, two sample MR was performed using external outcome data from Loh *et al.* [5]. However, the external dataset was also relatively small (50533 individuals) compared with the typical standards in genetic studies, and most importantly, it also introduced potential population bias, as participants were drawn from across South Asia rather than exclusively from Bangladesh. To at least mitigate the number of false positives, a Bonferroni correction for multiple testing was applied to every MR analysis. When focusing on the SS proteins, neither the one sample nor the two sample MR identified any significant causal effects after correction. The only notable finding was the *raw* (uncorrected) estimate for *N-terminal pro-BNP*, which appeared as a causal risk factor in the one sample MR, a causal protective factor according to both SS results and the existing literature, and non-causal in the two sample MR. This inconsistency highlights the unreliability of uncorrected estimates, suggesting that the apparent association detected by one sample MR is likely due to chance and cannot be considered credible. Overall, the limited statistical power of the current MR analyses can not fully validate Stability Selection as a tool for causal inference.

Extending the MR analyses to the full set of available proteins yielded similar results, as no causal associations were detected in either the one sample or two sample MR after Bonferroni correction for multiple testing. When considering the *raw* (uncorrected) results, a large number of apparent associations emerged (being 37 for one sample MR and 47 for two sample MR), most of which are likely false positives. Among these, one sample MR identified 10 proteins supported by previous literature as potentially causal, whereas two sample MR identified only 3. A summary of the MR results is reported in Table 1.

	With Bonferroni correction	Without Bonferroni correction
One sample MR (20 SS proteins)	0	1 (reverse effect)
Two sample MR (11 SS proteins)	0	0
One sample MR (1477 proteins)	0	10 (3 reverse effects)
Two sample MR (855 proteins)	0	3 (1 reverse effect)

Table 1: Summary of the MR results.

5. Discussion and conclusion

Type 2 diabetes is a complex, multifactorial disease affecting millions worldwide, and identifying its causal molecular drivers remains a major challenge. This thesis addressed two different research questions: a methodological one, testing the ability of Stability Selection to identify causal predictors, and a clinical one, aiming to uncover proteins potentially causal for type 2 diabetes among the 7244 plasma proteins measured. Stability Selection proved robust and consistent, reliably identifying 18 stable proteins across multiple input configurations. Compared to classical LASSO regression, SS achieved greater model stability while maintaining competitive predictive performance, highlighting its value for prioritizing predictive variables even without formal causal inference.

The proteins identified by SS were further evaluated using both one sample and two sample MR, along with a review of the literature. However, due to limited statistical power, MR did not select any causal candidate. Nevertheless, the literature review pointed out 8 out of the 18 SS proteins as promising causal candidates, suggesting the possible use of SS for causal inference. When all available proteins were considered for MR, several candidates showed partial agreement with prior evidence, but no associations remained significant after correcting for multiple testing. The main limitations were small sample size for one sample MR (and to a certain extent to two sample MR) and cohort heterogeneity for two sample MR, which reduced statistical power despite generally strong instrumental variables. These results indicate that, under current data constraints, MR alone cannot neither robustly validate SS proteins as causal predictors nor find any causal predictor. Nevertheless, SS could be particularly useful when two sample MR is not feasible (e.g., highly specific populations or outcomes), due to

its promising nature in detecting causal predictors being confirmed by the literature. Overlapping findings between SS and MR could provide higher confidence in identifying causal proteins, and the framework is likely to improve with larger cohorts.

References

- [1] <https://www.believestudy-bangladesh.org/>.
- [2] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
- [3] Barbara Bodinier, Sarah Filippi, Therese Haugdahl Nøst, Julien Chiquet, and Marc Chadeau-Hyam. Automated calibration for stability selection in penalised regression and graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2023. Published online: 13 July 2023.
- [4] Debbie A. Lawlor, Roger M. Harbord, Jonathan A. C. Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Wiley InterScience*, 27:1133–1163, 20 September 2007.
- [5] Marie Loh, Weihua Zhang, Hong Kiat Ng, Katharina Schmid, Amel Lamri, Lin Tong, Meraj Ahmad, Jung-Jin Lee, Maggie C Y Ng, Lauren E Petty, Cassandra N Spracklen, Fumihiko Takeuchi, Md Tariqul Islam, Farzana Jasmine, Anuradhani Kasturiratne, Muhammad Kibriya, Karen L Mohlke, Guillaume Paré, Gauri Prasad, Mohammad Shahriar, Miao Ling Chee, H Janaka de Silva, James C Engert, Hertzl C Gerstein, K Radha Mani, Charumathi Sabanayagam, Marijana Vujkovic, Ananda R Wickremasinghe, Tien Yin Wong, Chittaranjan S Yajnik, Salim Yusuf, Habibul Ahsan, Dwaipayan Bharadwaj, Sonia S Anand, Jennifer E Below, Michael Boehnke, Donald W Bowden, Giriraj R Chandak, Ching-Yu Cheng, Norihiro Kato, Anubha Mahajan, Xueling Sim, Mark I McCarthy, Andrew P Morris, Jaspal S Kooner, Danish Saleheen, and John C Chambers. Identification of genetic effects underlying type 2 diabetes in south asian and european populations. *Communication biology*, 5, 7 April 2022.