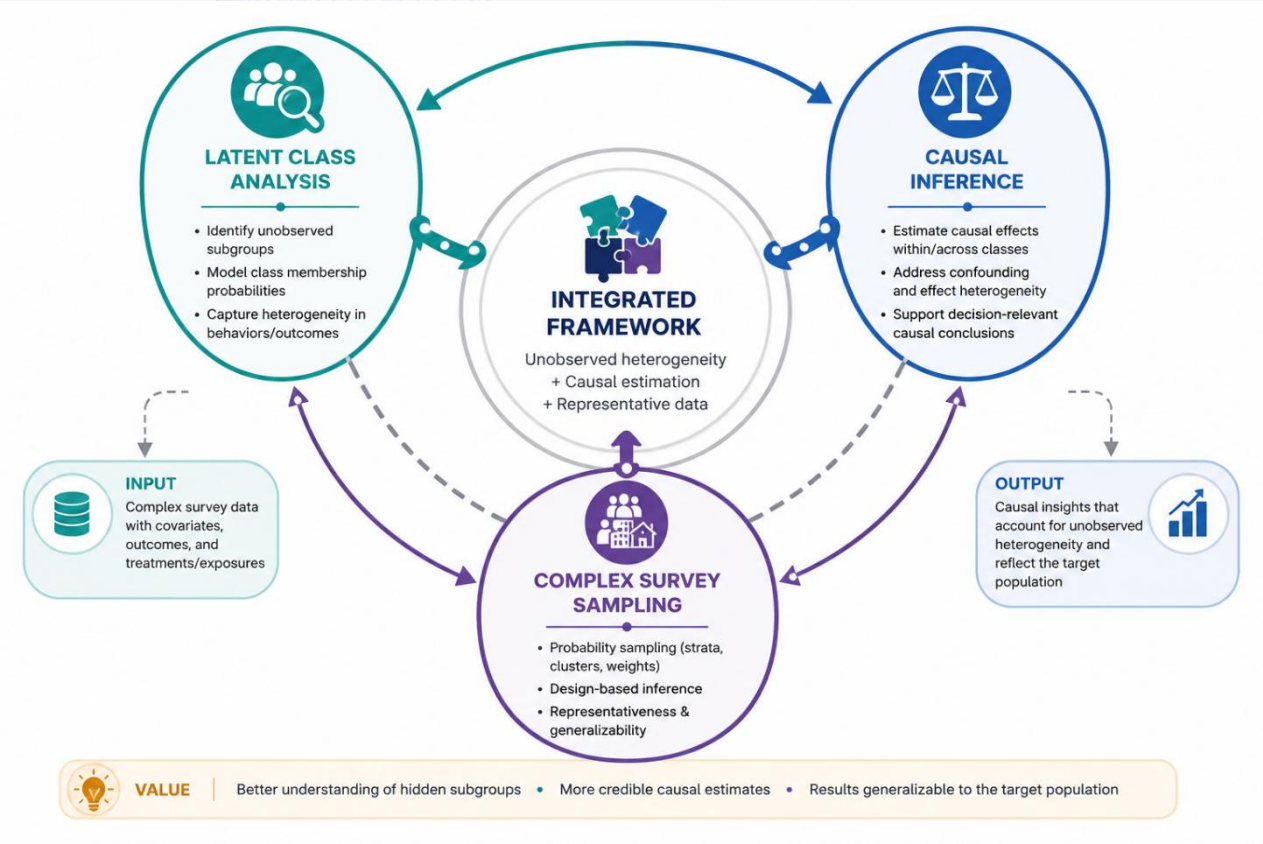
Causal Inference for Latent Class Analysis for Complex Survey Sampling Data

Leila Amorim
Federal University of Bahia, Brazil
03 June 2026
Joint work with Filipe Silva & Marcelo Taddeo

A framework for population-representative causal heterogeneity analysis



Motivation: 2019 National Health Survey (Brazil)

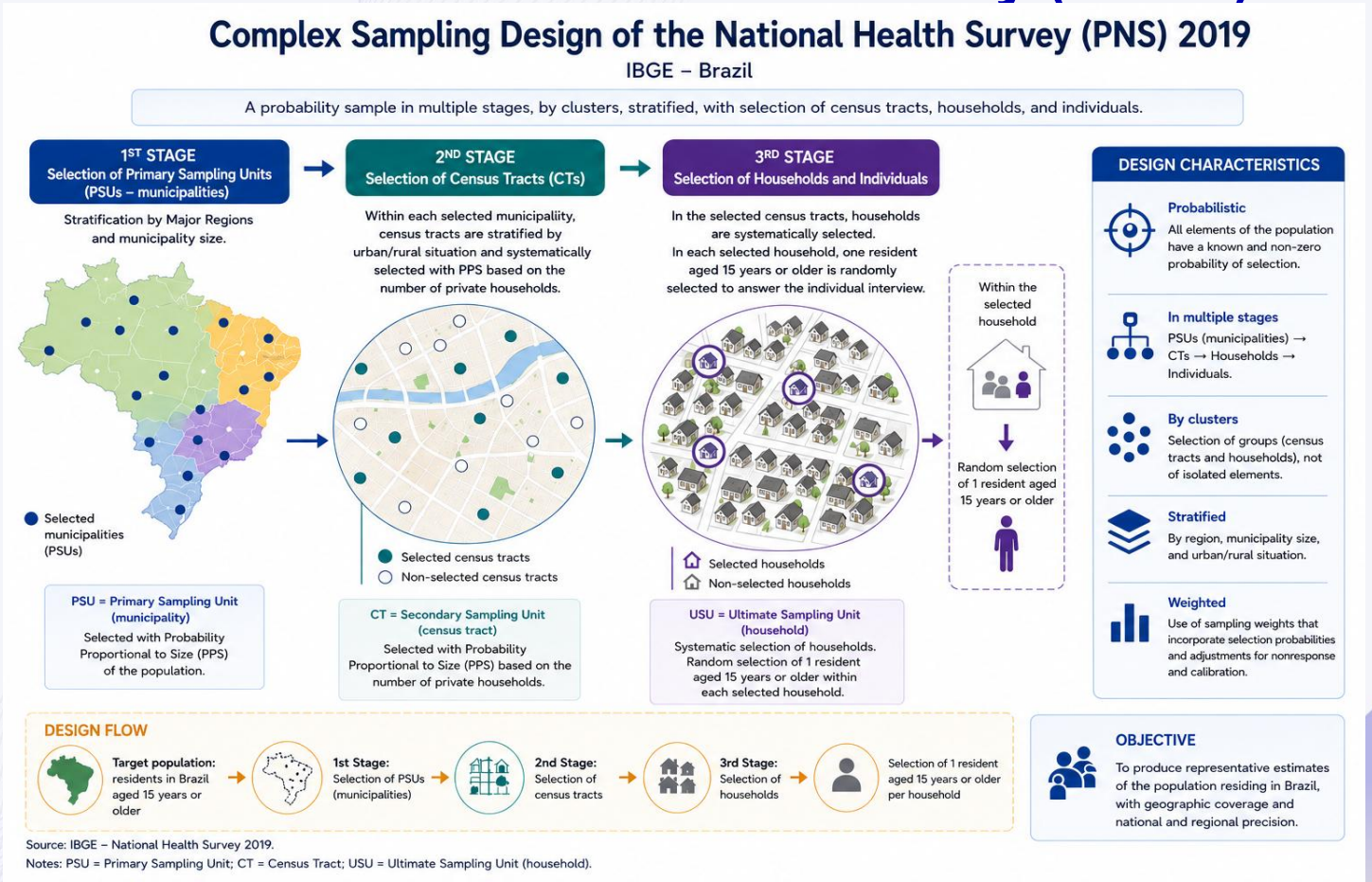


Figure 1. Three-stage cluster sampling design of the 2019 Brazilian National Health Survey (PNS), adapted from Stopa et al. (2020) and IBGE methodological reports.

Motivation: 2019 PNS (Cont.)

PNS

GENERAL AIM

Provide information on the determinants, conditions, and health needs of the Brazilian population to support evidence-based public policies and improve the effectiveness of health interventions.

OUR GOAL: To estimate the causal effect of physical activity on depression.

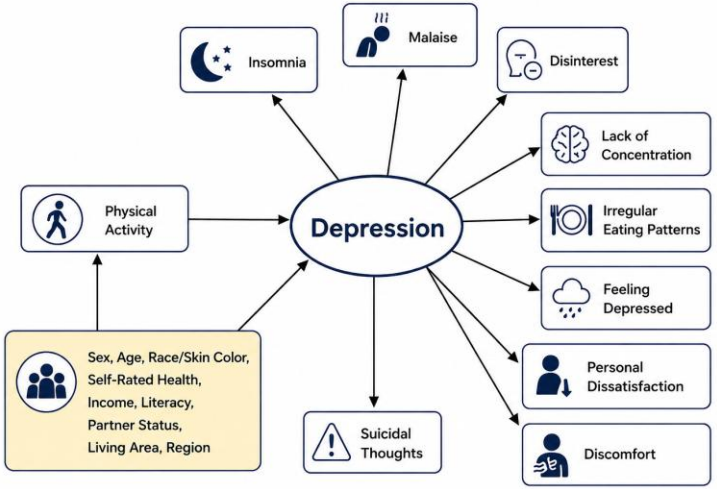


Figure 2. DAG for depression profiles.

(Costa et al, 2021; Lei et al, 2022; Mournet & Kleiman, 2024; Damiano et al, 2026).

Exposure: any physical activity (overall and aerobic exercise/sports) in the last 3 months.

Outcome: depression - measured by the PHQ-9 screening instrument to evaluate the frequency of symptoms in the previous two weeks.

Each symptom was dichotomize: (0=no: not at all/less than half of the days; 1=yes: more than half of the days/almost every day).

Background for population-representative heterogeneity analysis

- Traditional LCA is based on simple random sampling.

$$P(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x})=\sum_{c=1}^{C}\gamma_c(x)\prod_{j=1}^{J}\prod_{r_j=1}^{R_j}\rho_{j,r_j|c}^{I_{(y_i=r_j)}}$$

$$\gamma_c(x)=P(C=c|X=x)=\frac{e^{\beta_0+\beta_1 a+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_{p-1}x_{p-1}}}{1+\sum_{c'=1}^{C}e^{\beta_0+\beta_1 a+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_{p-1}x_{p-1}}}$$

- Survey-weighted LCA with jackknife standard errors enables valid inference for complex survey data (Patterson, Dayton, and Graubard, 2002).

$$\Lambda_w=\sum_{i=1}^{n}w_{i,survey}\ln\sum_{c=1}^{C}\gamma_c(x)\prod_{j=1}^{J}\prod_{r_j=1}^{R_j}\rho_{j,r_j|c}^{I_{(y_i=r_j)}}$$

- Ignoring sample weights resulted in biased parameter estimates.

(Bolck, Croon and Hagenaars, 2004; Collins & Lanza, 2009; Vermunt, 2010; Linzer & Lewis, 2011; Bakk, Tekle et al.,2013; Li et al, 2018)

Background for causal heterogeneity analysis

- Recent studies integrate causal inference with LCA to assess causal relationships based on latent class prevalence:
  - Propensity score matching (PSM)
  - Inverse Probability of Treatment Weighting (IPTW)

$$\hat{p}_i = P(A_i = 1 \mid \tilde{X}_i) = \frac{e^{\tilde{X}_i\beta}}{1 + e^{\tilde{X}_i\beta}}$$

- Many algorithms are available for propensity score matching, with or without replacement:
  - Example: 1:1 genetic matching using the R package *Matching* (Sekhon, 2011).

- Estimator-specific IPTW weights are used for ATE and ATT estimation.
  - Propensity weights are incorporated into the LCA likelihood.
  - Treatment effects on latent class membership are estimated.
  - Only treatment is included in the LCA linear predictor.

(Lanza, Coffman & Xu, 2013; Lanza, Schuler and Bray, 2016 ; Butera, Lanza and Coffman, 2014; Bray, Dziak, Patrick, and Lanza, 2019, Bray et al, 2020; Clouth, F.J., Pauws, S., Mols, F. et al, 2022).

Aims

Extend one-step and bias-adjusted three-step approaches to estimate ATE and ATT on latent classes in complex survey data using propensity scores.

Evaluate the performance of the estimators through simulation studies.

Illustrate the applicability of the proposed methodology.

Methods

- Ridgway et al. (2014) showed that ignoring sampling weights in IPTW estimation may lead to biased causal effect estimates.
  - A simple solution is to combine sampling and IPTW weights, yielding more robust estimates under complex survey designs.

- Incorporation of misclassification errors and propensity scores into the bias-adjusted three-step latent class approach to estimate causal effects (ATE) of LC membership on distal outcomes:
  - Sample weights and jackknife standard errors were used to account for complex survey design (Lê, Clouth, and Vermunt, 2025).

- Integration of these methods may lead to a better understanding of the causal mechanisms related to behaviors or characteristics that cannot be directly measured. Comparison of weighting approaches in LCA with covariates:
  - Survey weights only
  - IPTW weights only
  - Combined survey and IPTW weights

Methods

- Under the potential outcomes framework, we have

$$\mathrm{P}(\boldsymbol{Y}^a = \boldsymbol{y}) = \sum_{c=1}^{C} \gamma_c^a \prod_{j=1}^{J} \prod_{r=1}^{R_j} \left(\rho_{j,r|c}^a\right)^{I(y_j=r)}$$

$$\gamma_c^a = P(L^a = c)$$

$$\rho_{j,r|c}^a = P(Y_j^a = r | L^a = c)$$

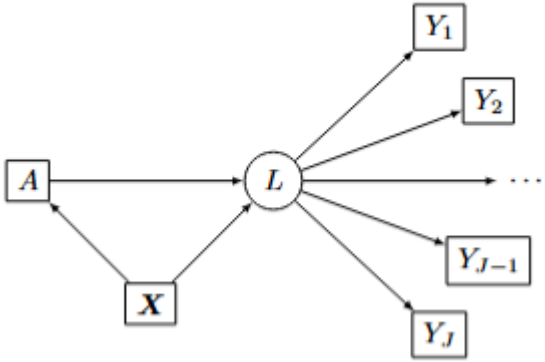


Figure 3 - Theoretical Causal Diagram

- From SUTVA and the ignorability assumption, we show that the parameter vector can be consistently estimated for the one-step approach through the maximization of the weighted log-likelihood

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\rho}) = \sum_{i=1}^{n} w_{a_i}(\boldsymbol{x}_i) \log \left(\sum_{c=1}^{C} \gamma_c(a_i) \prod_{j=1}^{J} \prod_{r=1}^{R_j} \rho_{j,r|c}^{I(y_{ij}=r)} \right)$$

- For three-step approach, measurement error is incorporated into the weights.

$$w_{a_i}(\mathbf{x}_i) = sw_i \times \frac{1}{\pi_{a_i}(\mathbf{x}_i)} \quad \text{for ATE}$$

$$w_{a_i}(\mathbf{x}_i) = sw_i \times \left[I(A=1) + I(A=0) \frac{\pi_{a_i}(\mathbf{x}_i)}{1 - \pi_{a_i}(\mathbf{x}_i)} \right], \quad \text{for ATT}$$

Simulation Studies Design and Data Generation

5th ISI REGIONAL STATISTICS CONFERENCE MALTA 2026

DATA GENERATION

Treatment
A ~ Bernoulli(0.5)

Confounders
X1 ~ Bernoulli(0.50)
X2 ~ N(0,1)

LATENT OUTCOME

4 binary indicators

Measurement quality
- High entropy = 0.90
- Moderate entropy = 0.75

SAMPLING DESIGN

Population = 100,000

Sample sizes (500; 1,500)

Number of Clusters (20; 50)
- Varying cluster sizes

IMPLEMENTATION

R (v4.1.2)
- MplusAutomation
- poLCA.simdata

Mplus (v8.11)
- 3-step LCA estimation

Simulation Studies: Criteria and Results

- 2,000 Monte Carlo replications
- Parameter: causal effect
- Methods: SW, IPW, BW

- Performance Criteria:
  - Relative bias
  - Variance estimation: model-based SE vs empirical SE
  - Mean squared error (MSE)
  - Coverage probability of 95% CIs

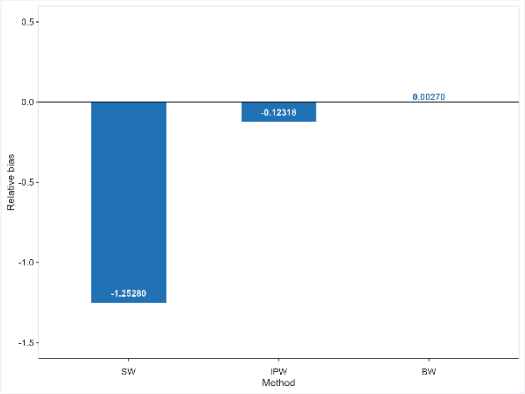
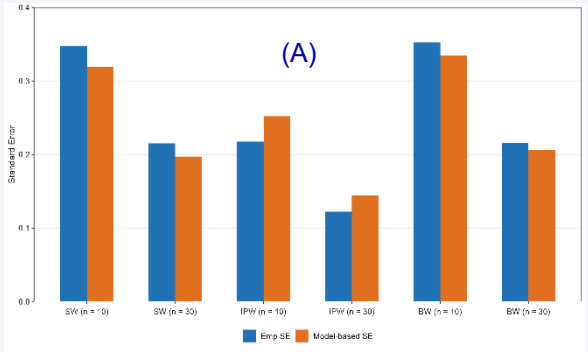


Figure 4 - Relative bias for causal effect with varying weighting strategy for high entropy scenario (m = 50; n = 10).



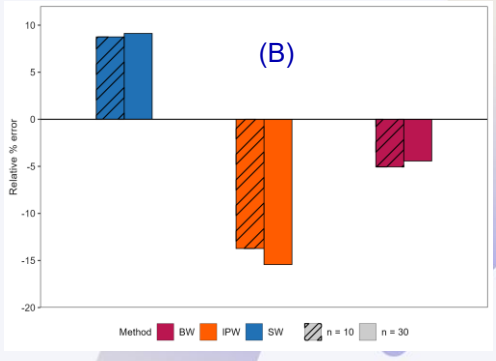


Figure 5 - Variance estimation according to weighting strategy and cluster size (m = 50): (A) Comparison of empirical and model-based SE; (B) Relative % error.

Application: Main Results – PNS2019

Latent Classes Analysis

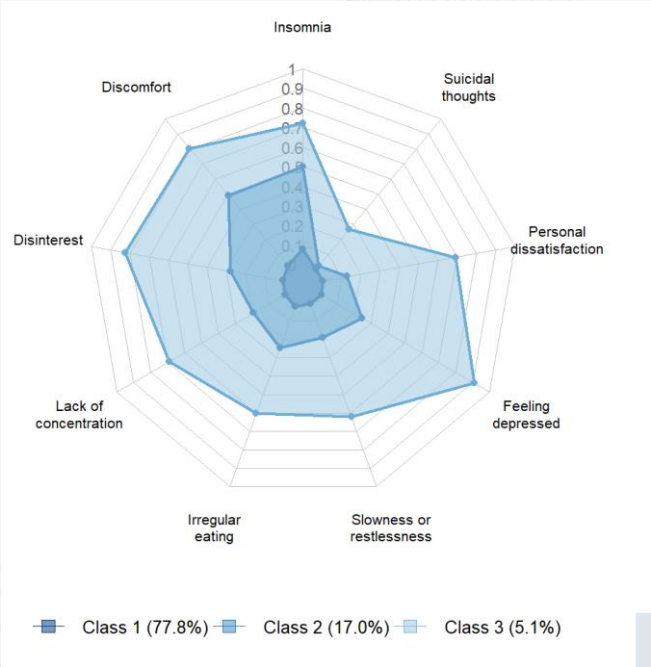


Figure 6 – Distinct Depression Profiles identified by Latent Class Analysis using PNS2019 data.

Entropy: 0.839

Causal Effect of PA on Depression Profiles

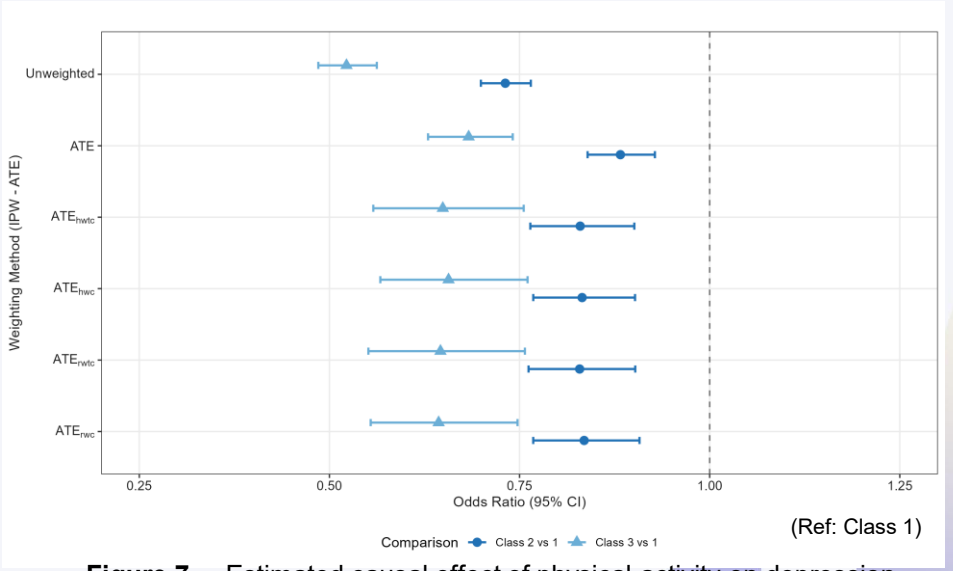


Figure 7 – Estimated causal effect of physical activity on depression profiles using PNS2019 data.

Class 1: Low-Risk Depression Profile
Class 2: Intermediate-Risk Depression Profile
Class 3: High-Risk Depression Profile

Final Remarks

- A unified framework was developed for causal effect estimation on latent classes in complex survey data.
  - The approach accounts for both latent outcome uncertainty and survey design features.
  - Performance depended on weighting strategy, entropy, sample size, and clustering, underscoring the importance of appropriate design-based adjustments.
  - The proposed methods facilitate more reliable population-level causal conclusions from survey data.

- Several applications might benefit from these methodological developments.

5th ISI REGIONAL STATISTICS CONFERENCE MALTA 2026

THANK YOU.

Acknowledgments

fapesb CNPq LINCA