Leveraging an External Validation Dataset to Adjust for Missing Confounders in an Enhanced Two-Stage Zero-Inflated Poisson Model Design: A Methodological and Simulation Study

MSR101

David Bin-Chia Wu [1,2,3], Hui-Wen Lin [4]

[1] Market Access (Asia Pacific), Johnson and Johnson International (Singapore) PTE. Ltd, Singapore
[2] Saw Swee Hock School of Public Health, National University of Singapore, Singapore
[3] School of Pharmacy, Faculty of Health & Medical Sciences, Taylor's University Malaysia, Malaysia
[4] Department of Mathematics, Soochow University, Taipei, Taiwan

Key takeaways

- **Enhanced methodology**: The TSC-ZIP model adjusts for missing confounding variables, ensuring unbiased and consistent results with reduced variance and improved power over standard ZIP models.
- **Dual-dataset calibration**: The use of the main and validation datasets that complement each other improves the accuracy and robustness of TSC-ZIP estimates without the need for database linkage.
- **Bridging data and methodological gaps for enhanced decision-making:** The TSC-ZIP model enables policymakers and clinicians to make more robust, reliable, and informed decisions in epidemiological research, comparative effectiveness assessment and health economic evaluations.

Background

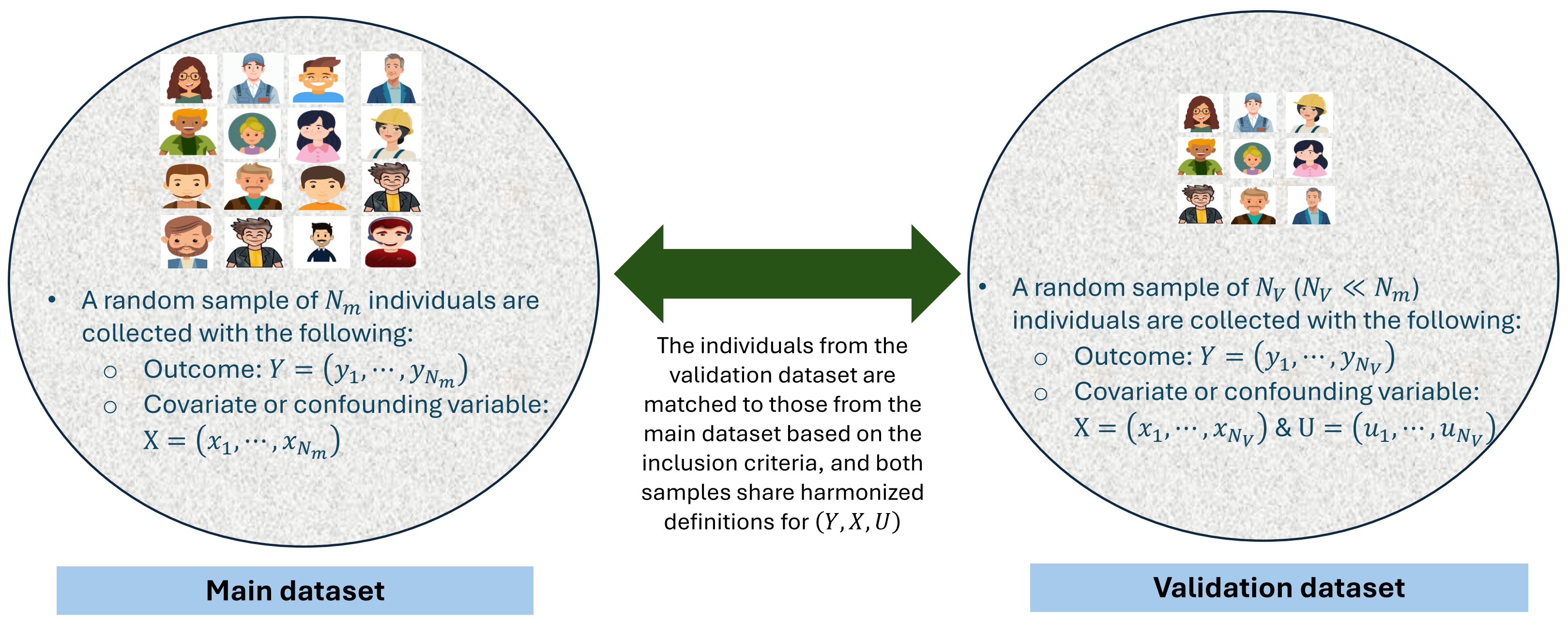
- The use of **real-world evidence (RWE)**, leveraging big data techniques to analyse population-based studies, has grown significantly, particularly in datasets like **administrative claims** and **electronic health records (EHRs)**.
- While national administrative claims datasets offer valuable large-scale insights, they often lack detailed individual-level information, such as **socioeconomic factors** or **laboratory results**, leading to bias observational study.
- Earlier methods, such as the regression calibration method by Stürmer et al. (2005)[1] and Bayesian propensity scores proposed by McCandless et al. (2012)[2], rely on **assumptions about measurement errors** or **independence between the exposure and unobserved confounders**, which are often not met in practice.
- Besides, **excess zero data** is a common challenge in medical databases, such as when patients report zero emergency room visits, no missed medication doses, or no adverse events during clinical trials. These zeros often result from factors like good health, full compliance, or treatment variability. To address this, various **zero-inflated models** have been developed, including the zero-inflated Poisson (Lambert, 1992) [3] and zero-inflated binomial models. However, **these models do not account for missing confounders** — an issue commonly encountered in RWE studies.

Objective

- This study aims to develop an innovative **Two-Stage Calibration Zero-Inflated Poisson (TSC-ZIP)** model to address the issue of missing confounders by leveraging an external validation dataset that complements the primary dataset, which lacks missing confounders.

Method

Dual datasets



- A random sample of N_m individuals are collected with the following:
  - Outcome: $Y = (y_1, \cdots, y_{N_m})$
  - Covariate or confounding variable: $X = (x_1, \cdots, x_{N_m})$

Main dataset

The individuals from the validation dataset are matched to those from the main dataset based on the inclusion criteria, and both samples share harmonized definitions for (Y, X, U)

- A random sample of N_V ($N_V \ll N_m$) individuals are collected with the following:
  - Outcome: $Y = (y_1, \cdots, y_{N_V})$
  - Covariate or confounding variable: $X = (x_1, \cdots, x_{N_V})$ & $U = (u_1, \cdots, u_{N_V})$

Validation dataset

Zero-inflated Poisson (ZIP) model

- The traditional ZIP model is defined as follows:

$$h(y_i|\phi_i, \lambda_i) = \begin{cases} \phi_i + (1-\phi_i)exp(-\lambda_i), & y_i = 0 \\ (1-\phi_i)\,\lambda_i{}^{y_i}/y_i\,!\ exp(-\lambda_i), & y_i > 0 \end{cases} \quad \text{Equation 1}$$

y_i **:** The number of times an event happens
ϕ_i **:** The probability of an observation who contributes to excess zeros
λ_i **:** Expected number of events (e.g. outpatient/ER visits, hospital readmissions, adverse events…etc.) for observations not in the zero-inflation group

Two-stage calibration zero-inflated Poisson (TSC-ZIP) model

Stage 1

- Equation 1 is fitted to the combined $(N_m + N_V)$ observations with (Y, X) from the **main and validation datasets**.
- The parameters ϕ_i and λ_i for the *i^th^* observation can be estimated as $\begin{cases} \phi_i = (1+\exp(-X_i^T\tau))^{-1} \\ \lambda_i = exp(X_i^T\gamma) \end{cases}$ where X_i is a $(k \times 1)$ vector of covariates of the *i^th^* observation and γ and τ are $(k \times 1)$ coefficient vectors of the covariates.
- $\bar{\boldsymbol{\gamma}}$ is a vector coefficients of γ for the observed covariates X can be numerically estimated using maximum likelihood method[4]. However, it is subject to residual bias as the confounding information U is missing.

Stage 2

- The estimate $\hat{\boldsymbol{\gamma}}$ of γ is obtained by fitting equation 1 again to N_V observations with (Y, X) from the **validation dataset.**
- The estimate $\widehat{\boldsymbol{\beta}}$ of β is derived by fitting equation 1 to N_V observations with (Y, X, U) from the **validation dataset**, where $\begin{cases} logit(\phi_i) = (X, U)^T\tau \\ log(\lambda_i) = (X, U)^T\beta \end{cases}$

Development of calibrated statistics of the TSC-ZIP model

- Although $\hat{\beta}$ is free from confounding bias as complete confounding information is incorporated into (X, U), it's solely estimated based on the validation dataset without using information in the main study.
- The closed form of the **TSC-ZIP estimate of $\boldsymbol{\beta}$** can be derived as follows by **fully utilizing information from both main and validation studies** motivated by the double-sampling approach by Chen and Chen[5] :

$$\boxed{\bar{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\Lambda}\boldsymbol{\Theta}^{-1}(\hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}})} \quad \text{Equation 2}$$

- Under regular condition, $\bar{\boldsymbol{\beta}}$ is an **unbiased** estimator of β where $var(\bar{\beta}) = var(\hat{\beta}) - \Lambda^T\Theta^{-1}\Lambda$, implying that $\bar{\boldsymbol{\beta}}$ has **greater statistical power** compared to $\hat{\beta}$.
- $\boldsymbol{\Lambda}$**:** The covariance matrix of $\hat{\beta}$ and $(\hat{\gamma} - \bar{\gamma})$ and $\boldsymbol{\Theta}$**:** The covariance matrix of $(\hat{\gamma} - \bar{\gamma})$ from equation 2 can be derived as follows:

$$\begin{cases} \widehat{\Lambda} = \sum_{i=1}^{N_V} (Q_i(\hat{\beta})Q_i(\hat{\gamma})^T - Q_i(\bar{\beta})Q_i(\bar{\gamma})^T) \\ \widehat{\Theta} = \sum_{i=1}^{N_M+N_V} (Q_i(\bar{\gamma})Q_i(\bar{\gamma})^T) + \sum_{i=1}^{N_V} Q_i(\hat{\beta})Q_i(\hat{\beta})^T - Q_i(\hat{\beta})Q_i(\hat{\gamma})^T - Q_i(\hat{\gamma})Q_i(\bar{\gamma})^T \end{cases}$$

, where $Q_i(.)$ is the efficiency score accounting for the variability and precision of the data

Method (Cont'd)

Simulation study

- A simulation study was conducted to evaluate the performance of the TSC-ZIP model in comparison to the ZIP model (Equation 3) (Table 1) using the specified performance metrics (Table 3).
- 5,000 Monte Carlo replicates were produced and analysed according to the simulation scenario in table 2.
- The simulation was performed using R software version 4.4.1.

$$\text{Simulated model: } \begin{cases} logit(\phi) = a + bX_1 + cU_1 \\ log(\lambda) = d + \beta X_1 + fU_1 \end{cases} \quad \text{Equation 3}$$

Table 1. Baseline characteristics

Covariate / Confounder	Value & distribution
X_1 (continuous variable)	$X_1 \sim Norm(\mu_1 = 0, \sigma_1^2 = 1)$
U_1 (continuous variable)	$U_1 \sim Norm(\mu_2 = 0, \sigma_2^2 = 1)$
(a, b, c, d, f)	$(-0.5, 0.4, 0.4, 0.3, 0.2)$

In a multivariable model, the propensity score can be used in place of U_1

Table 2. Simulation scenarios

Factor	value
Stage-1 sample size $(N = N_m + N_V)$	$N = 500, 1000, 1500$; The sample size of validation study is fixed at 150
The size of β (the parameter of interest)	$\beta = 0.3, 0.4, 0.5$
The association (ρ_{YU_1}) between U_1 and outcome Y	$\rho_{YU_1} = 0.5, 0.7, 0.9$

The sample size of validation study (N_V) is fixed at 150

Table 3. Performance metrics

Performance metric	Objective
Consistency	Measure the **bias** of $\hat{\beta}$ and $\bar{\beta}$, i.e. how close $\hat{\beta}$ and $\bar{\beta}$ are to the true value of β
Precision	Assess the **variance** of $\hat{\beta}$ and $\bar{\beta}$, i.e. how much $\hat{\beta}$ and $\bar{\beta}$ vary from sample to sample
Statistical power	Evaluate the **power** of $\hat{\beta}$ and $\bar{\beta}$, i.e. the probability of detecting an effect when it truly exists

Results

- **Increasing the stage-1 sample size**, while keeping stage-2 samples fixed, enhances the TSC-ZIP method's performance. Figure 2 shows that larger stage-1 samples increased the testing power and reduce variance due to more information for estimating β. The TSC-ZIP model demonstrates up to **23% higher power (0.608, 0.692, 0.644)** compared to the ZIP model (0.460, 0.464, 0.424), while **consistently showing lower variance**.
- As **true $\boldsymbol{\beta}$ increases**, Figure 3 illustrates that the TSC-ZIP model achieves power levels of 0.608, 0.826, and 0.904, compared to 0.460, 0.678, and 0.878 for the ZIP model. This reflects up to a **15% improvement in power**, along with **consistently lower variance**.
- When **the confounding factor $\boldsymbol{U_1}$ has a moderate-to-strong correlation $(\boldsymbol{\rho_{YU_1}})$ with the outcome variable** $(\boldsymbol{Y})$, as shown in Table 4, the stage-1 estimator $\bar{\boldsymbol{\gamma}}$ **exhibits significant bias**, i.e. deviation from the true β, due to missing information of U_1 particularly as ρ_{YU_1} increases**. Both $\widehat{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\beta}}$ maintain low testing sizes** (ranging from 0.036 to 0.062) while **TSC-ZIP estimate $(\bar{\boldsymbol{\beta}})$** has a **much smaller variance than the ZIP estimate, $\widehat{\boldsymbol{\beta}}$** .

Figure 2. Comparison of bias, variance and statistical power between TSC-ZIP $(\bar{\boldsymbol{\beta}})$ and ZIP $(\widehat{\boldsymbol{\beta}})$ models when stage-1 sample size increases (True $\boldsymbol{\beta} = 0.3$)

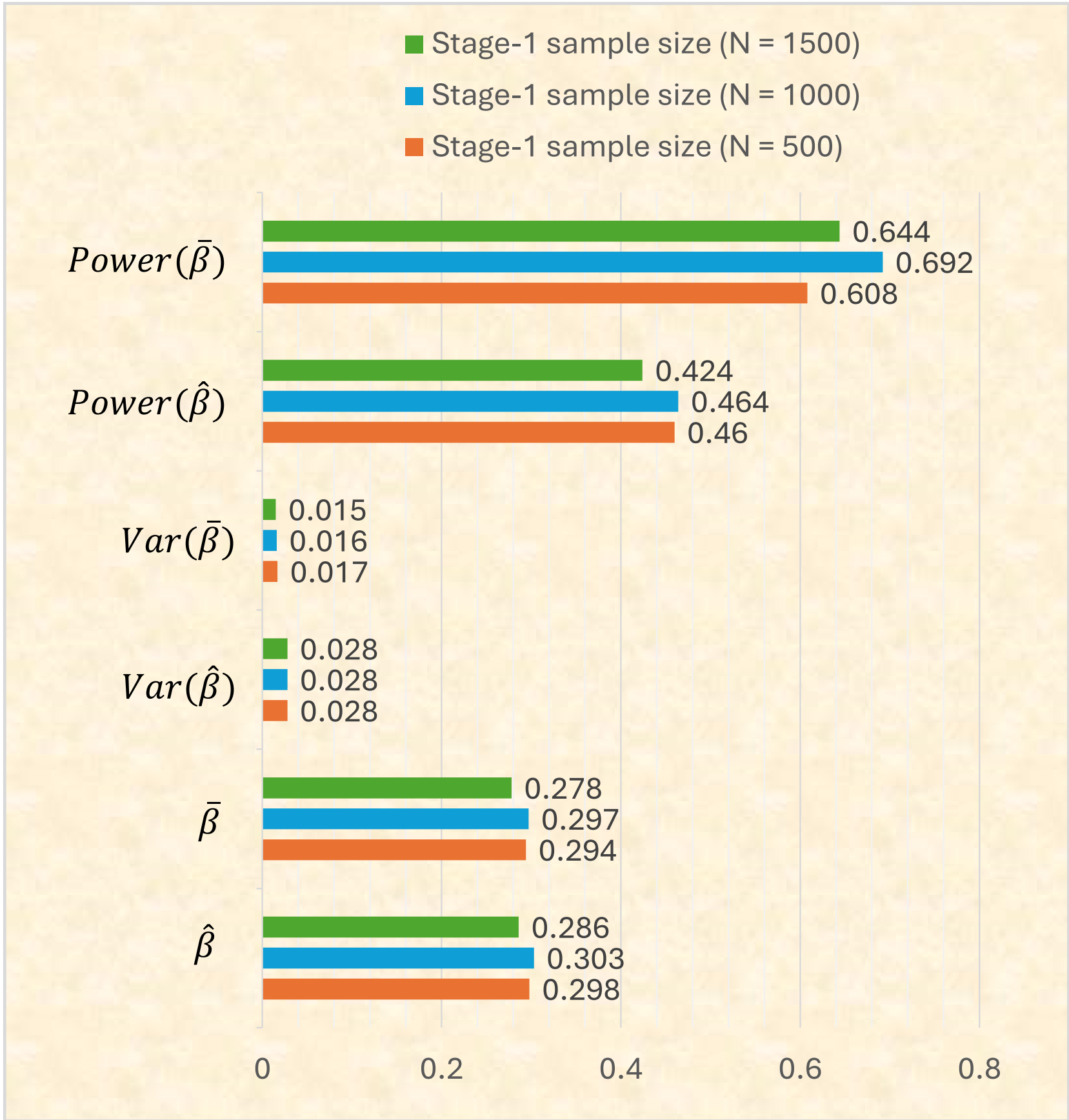


Figure 3. Comparison of bias, variance and statistical power between TSC-ZIP $(\bar{\boldsymbol{\beta}})$ and ZIP $(\widehat{\boldsymbol{\beta}})$ models when true $\boldsymbol{\beta}$ increases (Stage-1 sample size = 500 & Stage-2 sample size = 150)

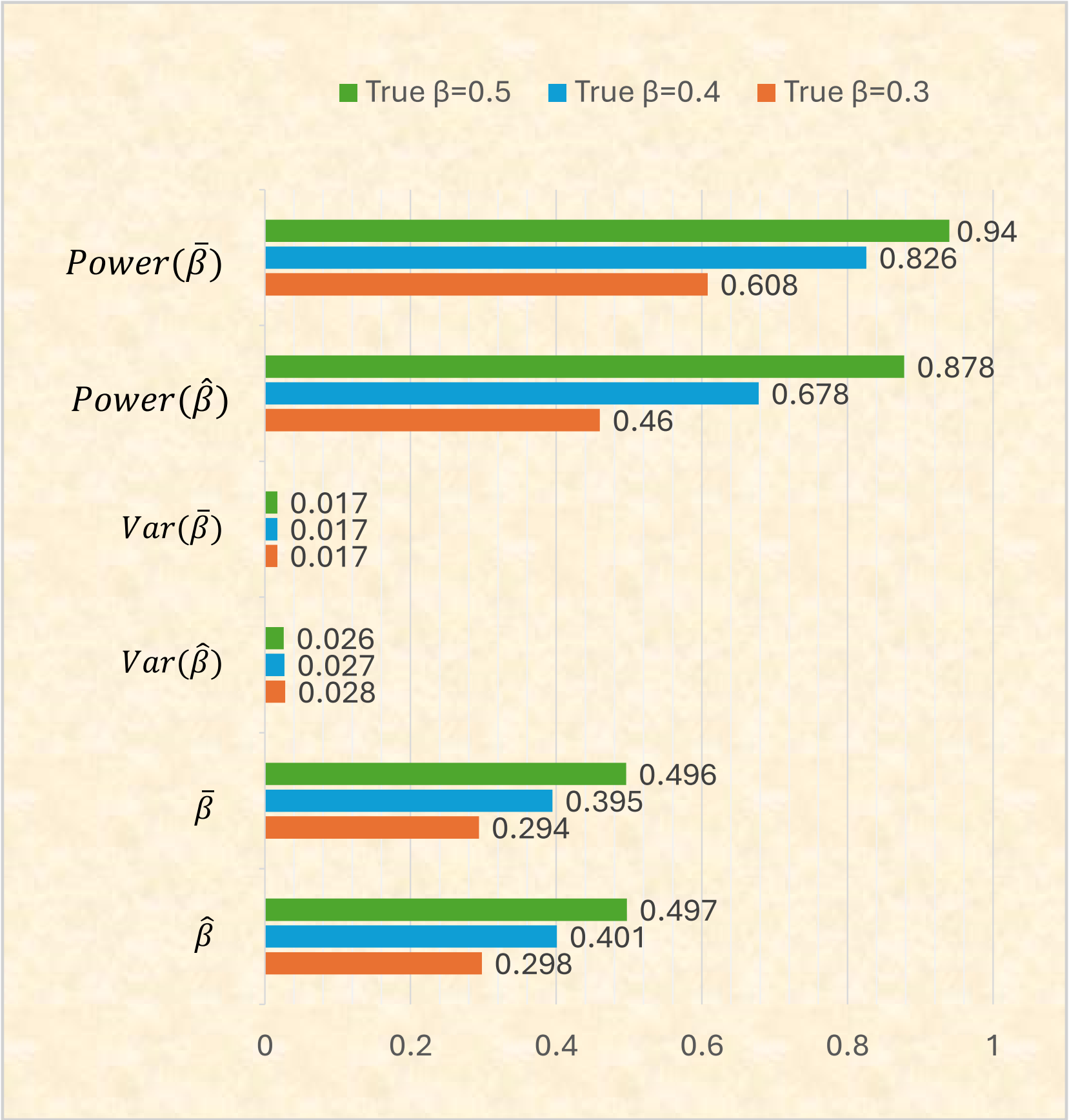


Table 4. Comparison of bias, variance, testing size and statistical power between TSC-ZIP $(\bar{\boldsymbol{\beta}})$ and ZIP $(\widehat{\boldsymbol{\beta}})$ models when the association $(\boldsymbol{\rho_{YU_1}})$ between the confounding factor $(\boldsymbol{U_1})$ and the outcome variable $(\boldsymbol{Y})$ increases (True $\boldsymbol{\beta} = 0$)

$\boldsymbol{\rho_{YU_1}}$ (True $\beta = 0$)	$\bar{\gamma}$	$Var(\bar{\gamma})$	$Size\ of\ \bar{\gamma}$	$\hat{\beta}$	$Var(\hat{\beta})$	$Size\ of\ \hat{\beta}$	$\bar{\beta}$	$Var(\bar{\beta})$	$Size\ of\ \bar{\beta}$
0.5	0.438	0.004	1	-0.017	0.025	0.05	-0.014	0.017	0.062
0.7	0.583	0.003	1	0.002	0.021	0.058	0.001	0.016	0.06
0.9	0.7	0.003	1	-0.011	0.018	0.036	-0.012	0.015	0.062

Size: The proportion of times the null hypothesis (H_0) is rejected under the H_0 over 5000 simulations

Conclusion

- The TSC-ZIP method has proven to be a reliable framework, outperforming the traditional ZIP model by leveraging a large crude dataset to enhance estimation efficiency and incorporating a smaller validation dataset to adjust for missing confounders.
- It can provide robust and reliable insights for policymakers and clinicians in epidemiology, comparative effectiveness, and health economics researches.

Reference

1. Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. Adjusting Effect Estimates for Unmeasured Confounding With Validation Data Using Propensity Score Calibration. American Journal of Epidemiology, 2005;162, 279–289.
2. McCandless LC, Richardson S, Best N. Adjustment for missing confounders using external validation data and propensity scores. J Am Stat Assoc. 2012;107(497):40 – 51.
3. Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 1992;34(1):1-14.
4. Casella, G., & Berger, R. L. (2002). Statistical Inference (2nd ed.). Duxbury.
5. Chen YH, Chen H. A unified approach to regression analysis under double-sampling designs. J R Stat Soc Series B Stat Methodol. 2000;62(3):449–460.

Disclosure

WDBC and LHW conceived and designed the study, as well as developed the algorithm. WDBC performed the statistical simulations and designed the poster. Both authors contributed to the interpretation of the results. Neither author has any conflict of interest to declare.

For any queries, please contact: dwu45@its.jnj.com or david.wu@nus.edu.sg

Presented at ISPOR Europe, Barcelona, Spain, 17-20 November 2024