TABLE 1 Summary of estimates for the causal effect (continuous outcome).

Situation	Method	Summary statistics			
		Mean	Empirical SE	Bias	RMSE
Valid IVs are in the majority	Proposed (strong NCO)	0.357	0.013	0.057	0.059
	Proposed (weak NCO)	0.480	0.021	0.180	0.181
	IVM	0.576	0.007	0.276	0.276
	BDHB	0.614	0.031	0.314	0.316
	HDB	0.318	0.015	0.018	0.023
	sisVIVE	0.401	0.009	0.101	0.101
Situation	**Method**	**Summary statistics**			
		Mean	**Empirical SE**	**Bias**	**RMSE**
Valid IVs are in the minority	Proposed (strong NCO)	0.487	0.017	0.187	0.188
	Proposed (weak NCO)	0.676	0.013	0.376	0.376
	IVM	0.724	0.005	0.424	0.424
	BDHB	0.781	0.015	0.481	0.481
	HDB	0.788	0.079	0.488	0.494
	sisVIVE	0.732	0.009	0.432	0.432

Note: The true value of the causal effect is 0.3. The sample size and the number of iterations are 10,000 and 1000, respectively. Mean, Empirical SE, Bias, and RMSE of the treatment's outcome coefficient are summarized for each situation and method, detailed in the 'Situation' and 'Method' columns, across 1,000 iterations. Regarding our proposed method, $(K_1, K_{2n}, \tau) = (10^8, 10^{-3}, 10^{-6})$. w is selected under $\alpha = 0.5$ (i.e., the probability that at least one covariance between a valid IV and an NCO exceeds the cutoff point w is 0.5; details are in online Appendix B.7).

Abbreviations: BDHB, median-based estimator (Bowden et al., 2016); HDB: mode-based estimator (Hartwig et al., 2017); IV, instrumental variable; IVM, inverse-variance weighting estimator; NCO, negative control outcomes; RMSE, root mean squared error; SE, squared error.

TABLE 2 Summary of estimates of the causal effects (point estimates (95% CI)).

	Methods		
	Proposed	**GMM**	**Logistic reg.**
Risk ratio	1.906 (1.810, 2.007)	2.194 (2.051, 2.346)	2.368 (2.367, 2.370)

Note: The results are derived as difference from a BMI of 18.5 to a BMI of 25. For the proposed method and GMM, a log-linear model is applied. The 95% CI is derived from a normal approximation for the log relative risk. Regarding our proposed method, $(K_1, K_{2n}, \tau) = (10^8, 10^{-3}, 10^{-6})$. w is selected under $\alpha = 0.5$ (i.e., the probability that at least one covariance between a valid IV and an NCO exceeds the cutoff point w is 0.5; details are in online Appendix B.7). Also, approximation (B.31) from online Appendix B.6 is applied to both GMM and our proposed method.

Abbreviations: BMI, body mass index; CI, confidence interval; GMM, generalized method of moments; NCO, negative control outcomes.

4 | UK BIOBANK DATA ANALYSIS

Next, actual data analysis is performed using our proposed method and previous IV methods on the UK Biobank dataset. We investigate the causal effect of body mass index (BMI) at the time of informed consent (IC) (Data-Field: 21001) on the incidence of diabetes (Data-Field: 2443). More detailed information appears in online Appendix D.

The estimated causal effects are summarized in Table 2. The results of the logistic regression model showed certain obvious biases. In contrast, GMM may provide somewhat conservative results compared to the analysis results of Orihara et al. (2023) (risk ratio (95% confidence interval [CI]) of Limited information Maximum Likelihood estimation (LIML): 1.896 (1.027, 2.407)). Regarding our proposed method, this result maybe more plausible because more plausible IVs were selected by the NCO. Note that our proposed method has a narrower 95% CI than the GMM. Therefore, the proposed method can derive "valid" results in terms of both validity and accuracy by selecting the plausible IVs.

5 | DISCUSSION AND CONCLUSIONS

In this paper, we proposed a novel IV estimator that uses a linear combination of IVs. Even when there are invalid IVs, our proposed method can select valid IVs using NCOs or some auxiliary variables without prior knowledge of the characteristics, number, or proportion of invalid IVs. We showed that our proposed estimator has the same efficiency as the GMM estimator, irrespective of whether valid IVs are selected or not. We confirmed that our proposed method out-

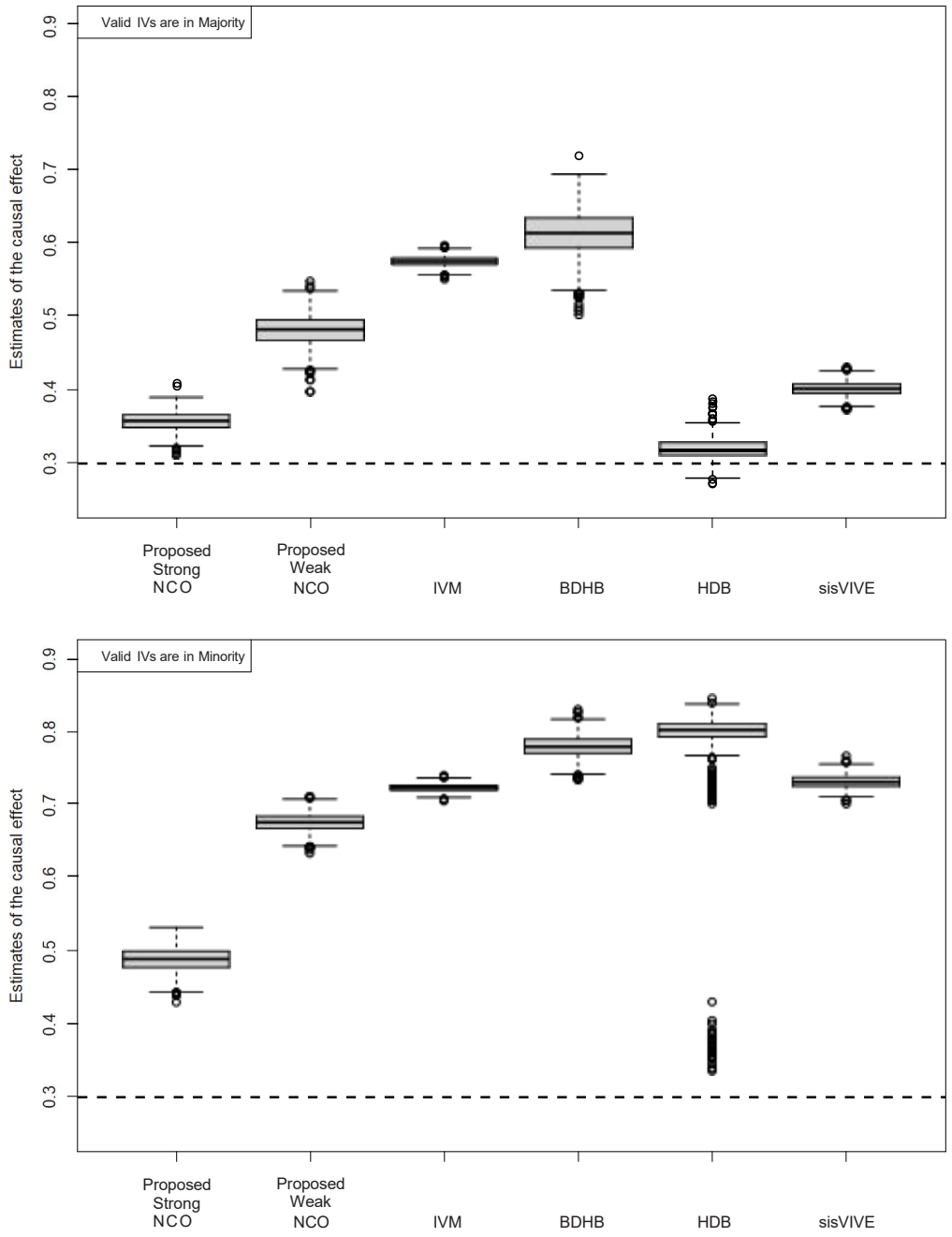

FIGURE 4 Boxplot of estimators for causal effects: The true value of the causal effect is 0.3 (red dashed line). The sample size and the number of iterations are 10,000 and 1000, respectively. Regarding our proposed method, $(K_1, K_{2n}, \tau) = (10^8, 10^{-3}, 10^{-6})$. w is selected under $\alpha = 0.5$ (i.e., the probability that at least one covariance between a valid IV and an NCO exceeds the cutoff point w is 0.5; details are in online Appendix B.7). BDHB, median-based estimator (Bowden et al., 2016); HDB: mode-based estimator (Hartwig et al., 2017); IVM: inverse-variance weighting estimator; IVM: inverse-variance weighting estimator; NCO, negative control outcomes.

performed some previous methods in simulations and analyses of the UK Biobank dataset. Future work is outlined in online Appendix E.

Using an NCO introduces a novel direction for identifying valid IVs. Since unobserved variables cannot be directly observed, observing fluctuations in proxy variables serves as a way to estimate their effects; this is the role of an auxiliary variable. The use of proxy variable-based estimators has gained attention in recent research (Cui et al., 2023; Jiang et al., 2019; Miao et al., 2018). However, our proposed method needs to specify an auxiliary variable related to unobserved variables but not directly related to treatment. In other words, only unobserved variables related to the auxiliary variable are

detected. To solve this problem, we should carefully choose as many valid auxiliary variables as possible. However, the performance of this strategy has not been confirmed in this study, and it remains an important area for future work.

As mentioned in the Introduction, specific methods can be employed in scenarios involving invalid IVs. Approaches such as the "majority rule" (Bowden et al., 2016; Kang et al., 2016), "plurality rule" (Guo et al., 2018; Hartwig et al., 2017), or their extensions (Fan & Wu, 2022) can facilitate the selection of valid IVs when partial IV information is available. This situation is similar to the methodologies proposed by Liao (2013), DiTraglia (2016), and Caner et al. (2018). These methods prove valuable when prior research offers some indications of valid IVs. By contrast, our proposed method remains applicable even in cases where IV-related evidence is absent, provided NCOs can be utilized. Consequently, our approach holds utility in situations where information regarding valid IVs is lacking, but NCOs can be sourced from previous research. In situations where information on both IVs and NCOs is uncertain, the proximal approaches proposed by Miao et al. (2018) and Cui et al. (2023) maybe applicable.

AUTH OR CONTRIBUTIONS

All authors have accepted responsibility for the entire content of this paper and approved its submission.

ACKNOWLEDGMENTS

We would like to express our gratitude to the reviewers for their useful comments. This manuscript is sophisticated by their comments and becomes more useful. We would like to thank Editage (www.editage.com) for English language editing. This work was supported by JSPS KAKENHI Grant Number JP21K10500.

CONFLICT OF INTEREST STATEMENT

The authors state no conflicts of interest.

DATA AVAILABILITY STATEMENT

This study was conducted using the UK Biobank resource under application number 71079. The data that support the findings of this study are available from the UK Biobank, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with the permission of the UK Biobank. The R code implementing the simulation and data analysis is available as a supplement to this paper.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

OR CID

Shunichiro Orihara https://orcid.org/0000-0003-0168-1250

REFERENCES

Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, *33*(13), 2297–2340.

Bowden,J., Davey,S. G., Haycock,P. C., & Burgess,S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, *40*(4), 304–314.

Brookhart, M. A., & Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *The International Journal of Biostatistics*, *3*(1), 1557–4679.

Burgess, S., Small, D. S., & Thompson, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, *26*(5), 2333–2355.

Burgess, S., & Thompson, S. G. (2015). *Mendelian randomization: Methods for using genetic variants in causal estimation*. CRC Press.

Caner, M., Han, X., & Lee, Y. (2018). Adaptive elastic net GMM estimation with many invalid moment conditions: Simultaneous model and moment selection. *Journal of Business & Economic Statistics*, *36*(1), 24–46.

Cui,Y., Pu, H., Shi,X., Miao,W., & TchetgenTchetgen, E. (2023). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 1–12.

Davies, N. M., Holmes, M. V., & Smith, G. D. (2018). Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ*, 362.

DiTraglia, F. J. (2016). Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics*, *195*(2), 187–208.

Fan, Q., & Wu, Y. (2022). Endogenous treatment effect estimation with a large and mixed set of instruments and control variables. *Review of Economics and Statistics*, https://doi.org/10.1162/rest_a_01230

Gkatzionis, A., Burgess, S., Conti, D. V., & Newcombe, P. J. (2021). Bayesian variable selection with a pleiotropic loss function in Mendelian randomization. *Statistics in Medicine*, *40*(23), 5025–5045.

Guo, Z., Kang, H., Tony Cai, T., & Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(4), 793–815.

Hartwig, F. P., Davey, S. G., & Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, *46*(6), 1985–1998.

Hayashi, F. (2011). *Econometrics*. Princeton University Press.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*, *20*(4), 493–506.

Jiang, L., Oualkacha, K., Didelez,V., Ciampi,A., Rosa-Neto, P., Benedet,A. L., Mathotaarachchi, S., Richards, J. B., Greenwood, C. M. T., & for the Alzheimer's Disease Neuroimaging Initiative. (2019). Constrained instruments and their application to Mendelian randomization with pleiotropy. *Genetic Epidemiology*, *43*(4), 373–401.

Kang, H., Zhang, A., Cai, T. T., & Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, *111*(513), 132–144.

Liao, Z. (2013). Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory*, *29*(5), 857–904.

Miao, W., Shi, X., & Tchetgen Tchetgen, E. (2018). *A confounding bridge approach for double negative control inference on causal effects*. arXiv. https://arxiv.org/pdf/1808.04945

Orihara, S., Goto, A., & Taguri, M. (2023). Instrumental variable estimation of causal effects with applying some model selection procedures under binary outcomes. *Behaviormetrika*, *50*(1), 241–262.

Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., Palmer, T., Schooling, C. M., Wallace, C., Zhao, Q., & Smith, G. D. (2022). Mendelian randomization. *Nature Reviews Methods Primers*, *2*(1), 1–21.

Yang,F., Zubizarreta,J. R., Small,D. S., Lorch,S., & Rosenbaum,P. R. (2014). Dissonant conclusions when testing the validity of an instrumental variable. *The American Statistician*, *68*(4), 253–263.

Yang, S., & Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, *105*(2), 487–493.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Orihara, S., Goto, A., & Taguri, M. (2024). Valid instrumental variable selection method using negative control outcomes and constructing efficient estimator. *Biometrical Journal*, *66*, 2300113. https://doi.org/10.1002/bimj.202300113