scientific reports

Check for updates

OPEN

Using machine learning to estimate the incidence rate of intimate partner violence

Zhuo Chen[1], Wen Ma[1], Ying Li[1], Wei Guo[1], Senhu Wang[2]✉, Wansu Zhang[3]✉ & Yunsong Chen[1]✉

It is difficult to accurately estimate the incidence rate of intimate partner violence (IPV) using traditional social survey methods because IPV victims are often reluctant to disclose their experiences, leading to an underestimation of the incidence rate. To address this issue, we applied machine learning algorithms to predict the incidence rate of IPV in China based on data from the Third Wave Survey on the Social Status of Women in China (TWSSSCW 2010). Specifically, we examined five unbalanced sample-processing methods and six machine learning algorithms, choosing the random under-sampling ensemble method and the random forest algorithm to impute the missing data. Analysis of the complete data showed that the incidence rates of physical violence, verbal violence, and cold violence were 7.10%, 13.74%, and 21.35%, respectively, which were higher than the incidence rates in the original dataset (4.05%, 11.21%, and 17.95%, respectively). The robustness of our findings was further confirmed by analysis using different training sets. Overall, this study demonstrates that better tools need to be developed to accurately estimate the incidence rates of IPV. It also serves as a useful guide for future research that imputes missing data using machine learning.

Domestic violence refers to violence or other abusive behavior that occurs in a domestic environment, including violence against partners, children, adolescents, and the elderly. Recently, intimate partner violence (IPV), a specific type of domestic abuse, has received growing attention in global scholarship[1,2]. IPV is typically defined as any conduct that harms another person physically, psychologically, or sexually while they are involved in an intimate relationship—usually a marriage[3,4]. The CDC (Centers for Disease Control and Prevention)'s report reflects that, in the US, the proportion of women suffering from all forms of IPV in 2016/2017 was 47.3%, and the proportion of men was not far behind at 44.2%[5]. According to the World Health Organization, in 2018, approximately 30% of women worldwide had experienced physical and/or sexual violence by an intimate partner or husband in the previous 12 months[6]. Studies in China using local samples estimated that the incidence rate of IPV was between 20 to 25%[7], but the national incidence rate of IPV remains unknown due to the absence of reliable nationwide survey data.

Studies have considered IPV as a sensitive issue and suggested a degree of underreporting in the survey data. Specific findings report that women in the US underreport IPV by roughly 50%, while men are more likely than women to underreport it[8]. Regarding specific explanations for underreporting, victims may feel ashamed of being involved in violence or fear that associating with violent partners will humiliate them[9]. This sense of shame deprives victims of their dignity[10], causing self-blame, self-deprecation, and a self-imposed sense of isolation[11], which makes them more likely to remain silent[12]. Underreporting due to survey methodology should also be considered, as the context in which the survey is conducted and the characteristics of the surveyor can have a significant impact on the likelihood of a response. Research also shows that women are more likely to downplay their IPV experiences when the interviewer is a man, when other people are around, or when they feel their privacy is threatened[13,14].

In particular, the emphasis on harmony and tight family ties in traditional Asian values may deter Asian women from reporting IPV, resulting in higher rates of response avoidance[15]. Chinese women are more likely to underreport their partner's violence against them because the society may stigmatize victims more than perpetrators[16]. Since China has long owned the "face" culture, where self-esteem, social expectations, and relationships are important[17], Chinese people tends to act to avoid humiliation, especially when it comes to immoral

[1]School of Social and Behavioral Sciences, Nanjing University, Nanjing 210023, China. [2]Department of Sociology and Anthropology, National University of Singapore, Singapore 117573, Singapore. [3]School of Law, Nanjing University, Nanjing 210023, China. ✉email: socsw@nus.edu.sg; zhangwansu@nju.edu.cn; yunsong.chen@nju.edu.cn

behavior such as IPV. Due to these factors, it is challenging for researchers and policymakers to estimate the true prevalence of IPV, and discussions of the mechanisms based on IPV statistics may also be questioned.

Due to response avoidance, surveys on IPV may have a proportion of non-response (missing values), making the scientific prediction of the proportion of non-response to IPV becomes a critical issue in estimating the incidence of IPV and in the subsequent analysis of its potential effects. Is there a method to accurately predict the answers of non-respondents on this socially sensitive indicator? The assumption of complete random missingness for IPV is not valid for the psychology of some respondents who deliberately conceal the existence of IPV, which means that traditional methods to handling missing values, such as deleting missing value cases and mean interpolation, would be inaccurate. Though more complex interpolation methods, such as regression interpolation, maximum likelihood estimation, and multiple interpolation, can predict IPV based on more information, they are statistical models of parameter estimation and are not good at prediction because of their limitations on the number of variables, the distribution of variables, and the patterns of relationships (e.g., linear relationships)[18]. In this instance, complex machine learning methods with greater computational power offer a novel way of predicting IPV. They can overcome the limitations of model form and variable selection, capture non-linear relationships, and fully account for various interactions between variables. Researchers have demonstrated the excellent performance of machine learning in predicting and imputing missing values[19–21]. To obtain a more accurate estimate of the prevalence of IPV in China, we adopted six machine learning algorithms to analyze data from the Third Wave Survey on the Social Status of Women in China (TWSSSCW 2010) to estimate the incidence rates of three types of IPV: physical violence, verbal violence, and cold violence. This study can remind IPV researchers to be more attentive to the missing proportion in the survey before empirical analysis, and provide a methodological reference for not only IPV prediction but also other sensitive survey questions.

Method

Survey and sample.

The data were obtained from the TWSSSCW 2010, the most recent nationally representative survey. This survey was conducted and approved by the All-China Women's Federation and the National Bureau of Statistics of China. All the data are anonymous and were collected following relevant guidelines and regulations. It included a wide range of questions on family, gender relations, health and well-being, and domestic violence (including IPV). The TWSSSCW 2010 used a stratified probability-proportional-to-size multi-stage random sampling design to ensure that the sample was representative of the general population at the provincial/municipal, street, and neighborhood levels[22]. The survey included a sample of 26,166 respondents, but due to the research question of our study, we restricted our sample to women and men who were married. Thus, the sample we used for our analysis involved 23,597 married women and men. The Ethics Committee of the School of Social Sciences of Nanjing University approved this study. Informed consent was obtained from all subjects.

Measures.

Three types of IPV were used as the dependent variables, namely physical violence, verbal violence, and cold violence. In the TWSSSCW 2010, respondents were asked "In your whole married life, has your partner ever hit you?", "In your whole married life, has your partner ever insulted you?", and "In your whole married life, has your partner ever constantly ignored you?" and gave responses on a 4-point scale ("never," "occasionally," "sometimes," or "often"). We dichotomized the three variables to measure whether respondents had suffered from IPV. As the frequencies expressed as "occasionally," "sometimes," or "often" are similar and vague in meaning, we collapsed these frequencies into one category. Table 1 shows that 1.22% of men and 2.83% of women experienced physical violence; 4.88% of men and 6.32% of women experienced verbal violence and 8.46% of men and 9.49% of women experienced cold violence. The total percentage of missing data was 6.95%. Figure S1 in the supplementary material shows the demographic distribution of missing data on respondents' experiences of IPV. Of the total missing values for IPV, approximately 65% were female, 60% were urban, and 70% were less educated. Compared to men (35%), rural (40%), and higher educated respondents (30%), the

	All sample	Men	Women
Physical violence (%)			
No	95.95	46.57	49.38
Yes	4.05	1.22	2.83
Verbal violence (%)			
No	88.79	42.90	45.90
Yes	11.21	4.88	6.32
Cold violence (%)			
No	82.05	39.34	42.71
Yes	17.95	8.46	9.49
Proportion of missing cases (%)	6.95	2.41	4.54
Observations (n)	23,597	11,063	12,534

Table 1. Descriptive statistics of IPV and missing cases.

rates of refusal to answer questions about IPV are significantly higher among female, urban, and lower educated respondents.

To predict IPV missing values, we adopted 611 variables from the questionnaire as features that would be provided as input to the machine learning algorithm used to train the model. These feature variables cover a wide range of topics related to aspects such as respondents' health, educational level, employment status, family background, gender ideology, leisure activities, political participation, religious affiliation, housework hours, and insurance enrolment status, providing rich clues for IPV prediction.

Imbalanced data problem.

Table 1 demonstrates how unbalanced the dataset was because only a small percentage of respondents reported having experienced IPV. Since most classifiers work with data drawn from the same distribution as the training set, using a standard prediction algorithm may lead to an underprediction of the percentage of respondents who had experienced IPV. Thus, it would be challenging to prepare appropriate data for training and testing, resulting in incorrect predictions. For example, if 99% of respondents reported that they had not experienced IPV, a standard machine learning algorithm, such as a naïve Bayesian classifier or a decision tree, would struggle to make accurate predictions for the minority group due to low variation. The two main types of approaches applied to address this problem are over-sampling the minority group and under-sampling the majority group, while each approach has its strengths and limitations. We compared five different resampling methods, including two over-sampling methods (the random over-sampling method[23] and the synthetic minority over-sampling technique, SMOTE[24]), two under-sampling methods (the random under-sampling ensemble method[25] and the *K*-means method[26]), and the SMOTE-edited nearest neighbor method[27] (SMOTE-ENN, which combines over- and under-sampling techniques). We divided the 21,956 non-missing cases into a training set (70%) and a validation set (30%), which had the same distribution of IPV cases as the original dataset. Next, we re-sampled the training dataset, used the random forest method to train the classifier, and finally tested the classifier on the validation set. After comparing the results of the different re-sampling methods, we found that the random under-sampling ensemble method gave the best performance in terms of the accuracy rate, the recall rate, the receiver operating characteristic (ROC) curve, and detection error tradeoff (DET) curve (as detailed in Table S1 and Figure S2 in the supplementary material).

Algorithm and parameter selection.

We compared six different machine learning algorithms: random forest algorithm[28], adaptive boost algorithm[29], Gaussian naïve Bayes algorithm[30], support-vector machine[31] (SVM), logistic regression algorithm[32], neural network[33], and one traditional interpolation method: multiple imputation[34]. Table S2-S4 in the supplementary material lists the accuracy and recall rates in the training and test sets for these algorithms. Generally, accuracy is crucial. However, when predicting IPV, given that the number of victims of IPV is quite small compared to the number of non-victims, it is unreasonable to blindly assume that all 21,956 of the sample would be classified as Category 0 who have never been hit by their spouse and easily achieve 96% accuracy. In this context, recall (the percentage of victims who are correctly predicted) is of great significance. The random forest algorithm can balance the prediction accuracy and recall for "physical violence" with 0.71 and 0.75, respectively.

The random forest algorithm is a supervised learning algorithm that performs classification by constructing multiple decision trees based on training datasets and predicts classification or average scores of individual decision trees (more details on the random forest algorithm are given in the supplementary material). We used the grid search method to search the hyperparameter space for the best cross-validation score, specifying the maximum feature range as 20 to 45 and the maximum depth range as 5 to 30 (see Figure S4 in the supplementary material). We then adopted the random forest algorithm to train the models 500 times, and the majority voting method to determine the final classification. To evaluate the models, we first used tenfold cross-validation to divide the resampled data into 10 consecutive folds and then applied one fold as the test set and the remaining nine folds as the training set. The training result was evaluated in terms of the area under the curve and the ROC and DET curves.

Robustness and heterogeneity analyses.

To ensure the robustness of our findings, we examined whether using different training sets would lead to similar results. Respondents' propensity to report IPV is associated with their gender ideology. A more patriarchal ideology holds that men are more capable than women and therefore men should predominate in roles of leadership and authority[35–37]. Consequently, men and women who hold a patriarchal gender ideology are more likely to conceal having experienced IPV: men due to fear of stigma[38], while women due to their desire to protect their partner[39,40]. We adopted the following question to measure patriarchal gender ideology: "Do you think that women are less capable than men?" We assumed that respondents who answered in the affirmative were more likely to have a patriarchal gender ideology; therefore, they tended to provide unreliable answers to the IPV questions. By excluding this particular group of respondents, we were able to replicate our analysis with the random forest algorithm.

Results

Table 2 reveals that the random forest classifier analyzed 1,641 missing cases of physical violence, of which 854 and 787 were classified as "No" and "Yes," respectively. Thus, the incidence rate of physical violence was 7.10% in the data imputed by the random forest algorithm, higher than the rate (4.05%) in the original raw dataset (i.e., without imputation). Similarly, the random forest classifier analyzed 1,643 missing cases of verbal violence and classified 860 and 783 cases as "No" and "Yes," respectively. This equated to an incidence rate of verbal violence of 13.47%, higher than the original incidence rate of 11.21%. Finally, the random forest classifier analyzed 1,645 missing cases of cold violence and classified 546 and 1,099 cases as "No" and "Yes," respectively. This equated

	Physical violence	Verbal violence	Cold violence
T-Acc	0.75	0.77	0.86
V-Acc	0.71	0.68	0.63
T-Rec	1	1	1
V-Rec	0.75	0.72	0.68
AUC	0.81	0.78	0.72
Predicted frequencies			
No	854	860	546
Yes	787	783	1099
Proportions in non-missing data	4.05%	11.21%	17.95%
Proportions in predicted missing data	47.96%	47.65%	66.81%
Proportions in full data	7.10%	13.74%	21.35%

Table 2. Fitness statistics and predicted outcomes of the random forest algorithm. *T-Acc* accuracy in training set; *V-Acc* accuracy in validation set; *T-Rec* recall in training set; *V-Rec* recall rate in test set; *AUC* area under the ROC curve.

to an incidence rate of cold violence of 21.35%, which was greater than the original incidence rate of 17.95%. Reassuringly, these analyses produced similar results in robustness analyses using different training data (see Table S5 in the supplementary material).

Figure 1 shows the joint distributions of IPV in different sex groups, which supports the random forest algorithm's inference of the missing data that there were much higher incidence rates of IPV than previously reported, especially among women. This implies that the incidence rates of IPV may be underestimated if missing data are not taken into account. When considering the dataset that included imputed missing data, the incidence rate of physical violence increased from 2.83% to 4.99% for women and from 1.22% to 2.11% for men; the incidence rate of verbal violence increased from 6.32% to 8.05% for women and from 4.88% to 5.69% for men; and the incidence rate of cold violence increased from 9.49% to 11.92% for women and from 8.46% to 9.43% for men.

Similarly, Fig. 2 illustrates the joint distributions of IPV in both rural and urban areas. According to the imputed missing data, more rural cases (30.35% and 31.77%, respectively) than urban cases (17.61% and 15.89%, respectively) were expected to have experienced physical and verbal violence. Conversely, the urban sample was predicted to have a larger percentage of people encountering cold violence (37.26% in urban and 29.54% in rural areas). When adding the imputed data to the database, the prevalence of physical violence increased from 1.21% to 2.35% for urban residents and from 2.84% to 4.75% for rural residents, the prevalence of verbal violence increased from 3.46% to 4.33% for urban residents and from 7.74% to 9.41% for rural residents, and the prevalence of cold violence increased from 8.5% to 10.51% for urban residents and from 9.44% to 10.85% for rural residents. We also constructed joint probability distributions of IPV in different education groups (see Fig. 3) and found that respondents with lower levels of education consistently had higher incidence rates of the three types of IPV, both before and after imputing missing data.

Conclusions

So far, the problem of domestic violence, particularly IPV, has received increasing attention from academics and policymakers worldwide, as it is considered to be an important predictor of couples' health, well-being, and quality of life. However, missing data on certain types of domestic violence poses a substantial challenge to researchers conducting IPV studies. The pattern of these missing data is missing not at random because these missing data are largely related to IPV. For instance, people may not report having experienced IPV because of privacy concerns, fear of reprisal or stigma, or a desire to protect the perpetrator. Traditional approaches used to deal with datasets with missing data, such as listwise deletion and multiple imputation, are not suitable for handling with datasets with missing IPV data: simple interpolation is not appropriate for sensitive indicators such as IPV because it violates the assumption of missing completely at random; traditional model interpolation such as multiple interpolation does not perform well in prediction because of its limitations on the number of variables, the distribution of variables, and pre-defined patterns of data relationships. The aforementioned methods are also ineffective for interpolating data with a high degree of imbalance. Indeed, we found that the incidence rates of physical, verbal, and cold violence were underestimated when we performed listwise deletion or multiple interpolation with missing values. Thus, we applied novel machine learning methods to predict the missing data on the incidence rates of three types of IPV: physical violence, verbal violence, and cold violence.

Therefore, we compared the fitness statistics of different machine learning methods and selected the optimal method—the random forest algorithm—to impute missing data. We discovered that analysis of datasets with imputed data resulted in increased incidence rates of physical violence (7.10%), verbal violence (13.74%), and cold violence (21.35%), compared with analysis of datasets without imputed data (4.05%, 11.21%, and 17.95%, respectively). These results reveal the incidence rate of IPV in China, reduce measurement bias due to intentional concealment by respondents, and provide data and methodological support for correcting this sensitive social survey indicator. We suggest that when IPV researchers work with IPV data with high missing values, they should first predict these indicators before conducting further analyses. At the same time, we hope that our efforts in the analysis method will provide a reference for interpolating missing values and predicting social indicators

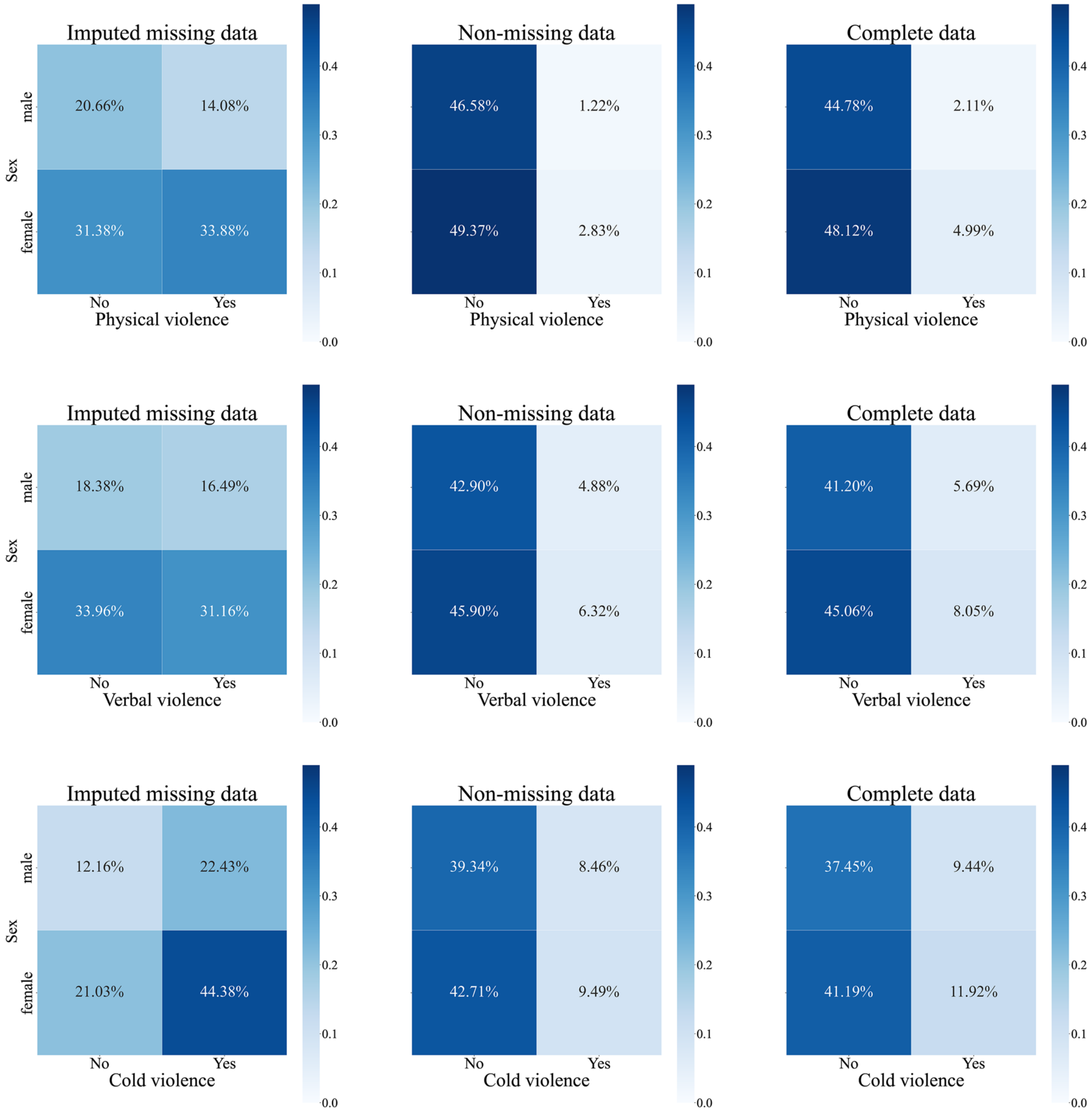

Figure 1. Joint probability distributions of sex and IPV.

in the future. In addition to the need for social governance, our results will help government policymakers to understand the true incidence and predict highly sensitive and subjective hidden indicators such as drug use, sexual orientation[41], AIDS[42], surrogacy, extramarital affairs, and crime, and thus better prevent and address the social problems related to marginal groups.

However, there are some limitations to this study. First, the data we used was from 2010. We hope that subsequent studies will be able to use more recent data. Second, for machine learning, the original structure and quality of the data determine the upper limit of accuracy. Although we try to minimize the error, any algorithm or resampling effort is less precise than collecting a larger number of minority class samples. Third, predicting whether an individual had experienced IPV is a probabilistic prediction. Our prediction aims to be as close to reality as possible but does not represent the actual state of the sample. Fourth, the prevalence of IPV predicted in this paper is based on married groups only. Although unmarried or extramarital IPV is common, we did not analyze it due to the limitations of the secondary data. Fifth, we noted that several recent studies focused more on the overreporting issue of IPV in Western societies[43,44]. Given the various sociocultural environments in different regions, we suggest that this study's analytical tools and findings based on the underreporting issue of IPV should be carefully considered before applying in other regions.

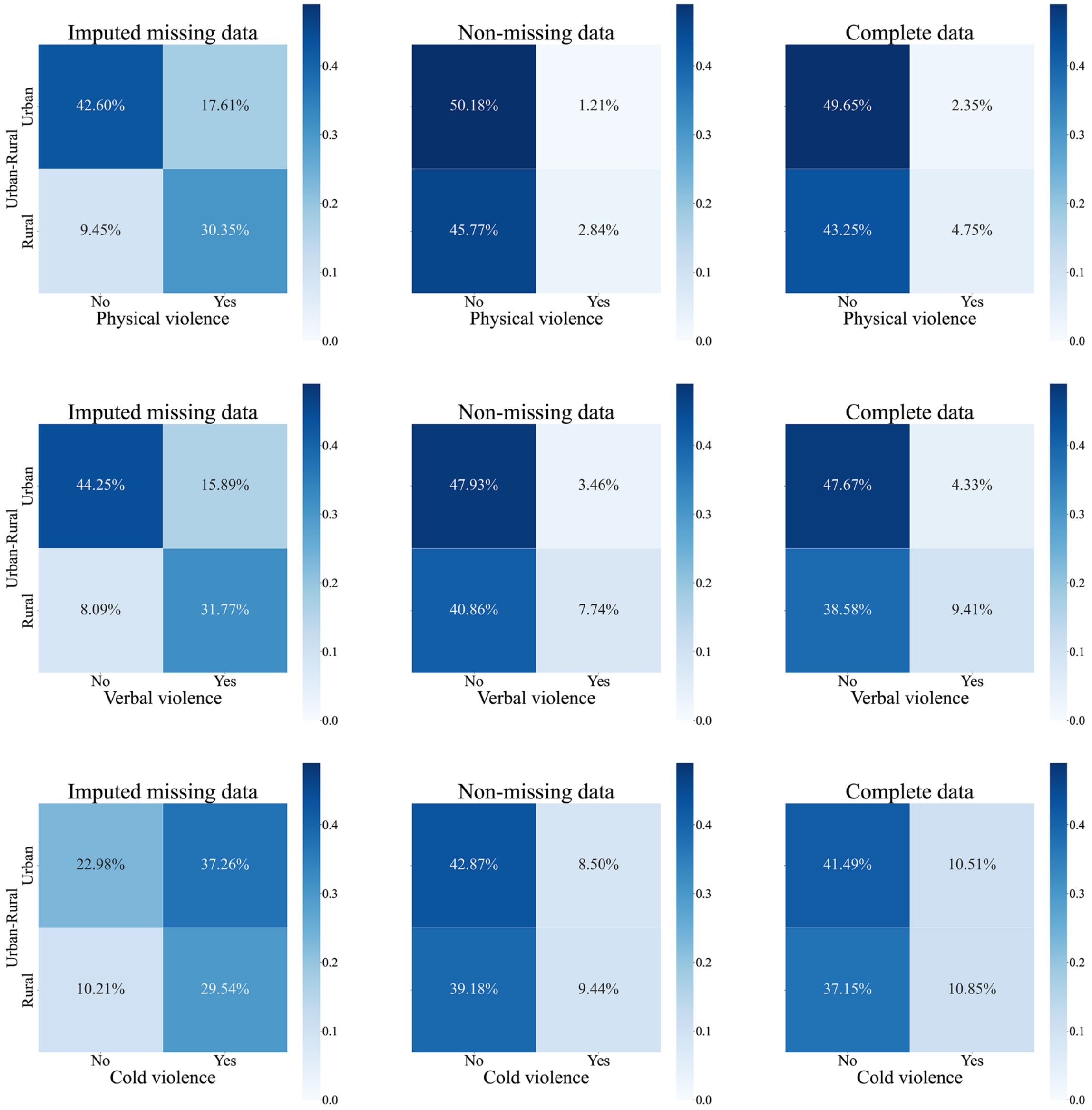

Figure 2. Joint probability distributions of residence and IPV.

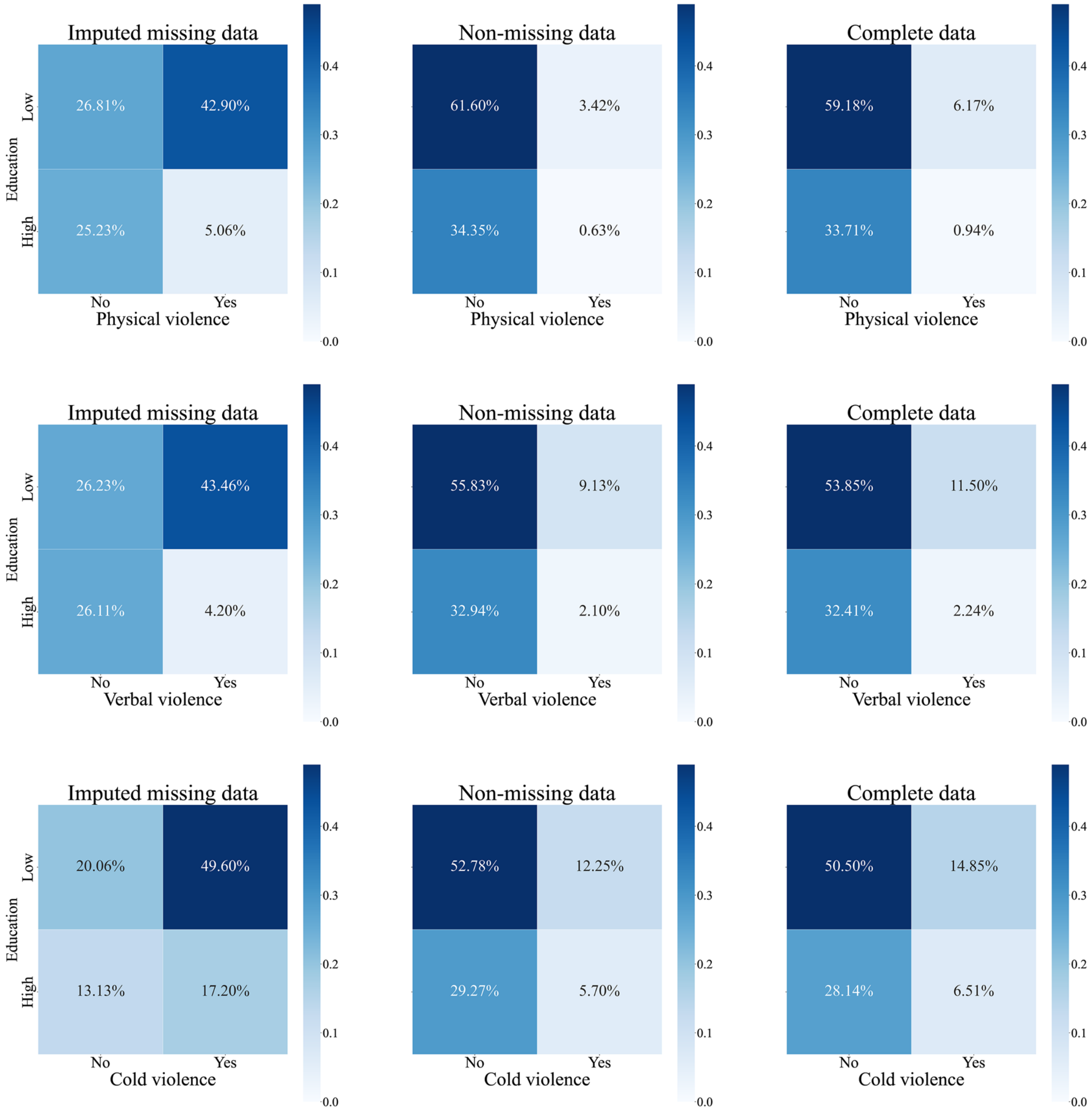

Figure 3. Joint probability distributions of education and IPV. *Note Low* Junior high school or below, *High* High school or above.

Data availability

Given that the data we applied in the study, the Third Wave Survey on the Social Status of Women in China (2010) is second-hand, and the unauthorized disclosure of a third party is not allowed under the data processing agreement, we are afraid that we are not able to show these data in the link. Basic information about the data can be found at http://www.china.com.cn/zhibo/zhuanti/ch-xinwen/2011-10/21/content_23687810.htm. Readers can request the data from the China Women's Federation.

Code availability

Code related to this paper are available on GitHub (https://github.com/UdvIPV/Unspeakable-domestic-violence).

Received: 29 August 2022; Accepted: 17 March 2023
Published online: 04 April 2023

References

1. Garcia-Moreno, C., Jansen, H. A., Ellsberg, M., Heise, L. & Watts, C. H. Prevalence of intimate partner violence: findings from the WHO multi-country study on women's health and domestic violence. *The Lancet.* **368**(9543), 1260–1269 (2006).
2. Adams, A. E., Tolman, R. M., Bybee, D., Sullivan, C. M. & Kennedy, A. C. The impact of intimate partner violence on low-income women's economic well-being: The mediating role of job stability. *Violence Against Women* **18**(12), 1345–1367 (2012).
3. Krug, E. G., Mercy, J. A., Dahlberg, L. L. & Zwi, A. B. The world report on violence and health. *The Lancet* **360**(9339), 1083–1088 (2002).

4. Stewart, D. E. & Chandra, P. S. WPA international competency-based curriculum for mental health providers on intimate partner violence and sexual violence against women. *World Psychiatry* **16**(2), 223 (2017).
5. Leemis, R. W., *et al. The National Intimate Partner and Sexual Violence Survey: 2016/2017 Report on Intimate Partner Violence.* Atlanta, GA: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention (2022).
6. World Health Organization. Violence against women prevalence estimates, 2018: global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women (2021).
7. Xu, X. *et al.* Prevalence of and risk factors for intimate partner violence in China. *Am. J. Public Health* **95**(1), 78–85 (2005).
8. Szinovacz, M. E. & Lance, C. E. Comparing one-partner and couple data on sensitive marital behaviors: The case of marital violence. *J. Marriage Fam.* **57**(4), 995–1010 (1995).
9. Felson, R. B., Steven, F. M., Anthony, W. H. & Glenn, D. Reasons for reporting and not reporting domestic violence to the police. *Criminology* **40**(3), 617–648 (2002).
10. Negrao, C. *et al.* Shame, humiliation, and childhood sexual abuse: Distinct contributions and emotional coherence. *Child Maltreat.* **10**(4), 350–363 (2005).
11. Wilson, J. P., Boris, D. & Silvana, T. Posttraumatic shame and guilt. *Trauma Violence Abuse* **7**(2), 122–141 (2006).
12. Holmes, S. T. & Holmes, R. M. *Sex Crimes: Patterns and Behavior.* Sage Publications (2008).
13. Ellsberg, M., Lori, H., Rodolfo, P., Sonia, A. & Anna, W. Researching domestic violence against women: Methodological and ethical considerations. *Stud. Fam. Plann.* **32**(1), 1–16 (2001).
14. Walby, S. & Andrew, M. Comparing the methodology of the new national surveys of violence against women. *Br. J. Criminol.* **41**(3), 502–552 (2001).
15. Tjaden, P. G. *Extent, Nature, and Consequences of Intimate Partner Violence* (National Institute of Justice, 2000).
16. Parish, W. L., Tianfu, W., Edward, O. L., Suiming, P. & Ye, L. Intimate partner violence in china: national prevalence, risk factors and associated health problems. *Int. Fam. Plan. Perspect.* **30**(4), 174–181 (2004).
17. Ho, D. Y. On the concept of face. *Am. J. Sociol.* **81**(4), 867–884 (1976).
18. Breiman, L. Statistical modeling: The two cultures. *Stat. Sci.* **16**(3), 199–231 (2001).
19. Deb, R. & Liew, A. W. C. Missing Value Imputation for the Analysis of Incomplete Traffic Accident Data. *Inf. Sci.* **339**, 274–289 (2016).
20. He, H., Yuan, C., Yi, C. & Jinyu, W. Ensemble learning for wind profile prediction with missing values. *Neural Comput. Appl.* **22**(2), 287–264 (2013).
21. Pantanowitz, A. & Tshilidzi, M. Missing data imputation through the use of the random forest algorithm. *Advances in Intelligent and Soft Computing* 53–62 (Springer Verlag, 2009).
22. Project Group of the 3rd Survey on the Status of Chinese Women. Executive Report of the 3rd Survey on the Status of Chinese Women. *J. Chin. Women's Stud.* **6**, 5–15 (2011).
23. Menardi, G. & Torelli, N. Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Disc.* **28**(1), 92–122 (2014).
24. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
25. Liu, X. Y. Wu, J. & Zhou, Z. H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B* **2**, 539–550 (2008).
26. Likas, A., Vlassis, N. & Verbeek, J. J. The global k-means clustering algorithm. *Pattern Recognit.* **36**(2), 451–461 (2003).
27. Batista, G. E., Prati, R. C. & Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **6**(1), 20–29 (2004).
28. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
29. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**(1), 119–139 (1997).
30. Zhang, H. The optimality of naive Bayes. *Aa* **1**(2), 3 (2004).
31. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Appl.* **13**(4), 18–28 (1998).
32. Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M. & Klein, M. *Logistic Regression*. New York: Springer-Verlag (2002).
33. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**(6088), 533–536 (1986).
34. Rubin D. B. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons (2004).
35. Morris, E. W. & Ratajczak, K. Critical masculinity studies and research on violence against women: An assessment of past scholarship and future directions. *Violence Against Women.* **25**(16), 1980–2006 (2019).
36. Wang, S. The role of gender role attitudes and immigrant generation in ethnic minority women's labor force participation in Britain. *Sex Roles* **80**(3–4), 234–245 (2019).
37. Wang, S., & Li, L. Z. Double Jeopardy: The roles of job autonomy and spousal gender ideology in employed women's mental health. *Appl. Res. Qual. Life* 1–18(2022).
38. Felson, R. B. & Pare, P. P. The reporting of domestic violence and sexual assault by nonstrangers to the police. *J. Marriage Fam.* **67**(3), 597–610 (2005).
39. Bograd, M. Why we need gender to understand human violence. *J. Interpers. Violence* **5**(1), 132–135 (1990).
40. Kimmel, M. S. "Gender symmetry" in domestic violence: A substantive and methodological research review. *Violence Against Women* **8**(11), 1332–1363 (2002).
41. Chen, Y., He, G. & Ju, G. The hidden sexual minorities: Machine learning approaches to estimate the sexual minority orientation among Beijing college students. *J. Soc. Comput.* **3**(2), 128–138 (2022).
42. Chen, Y., He, G., & Yan, F. *Understanding China Through Big Data: Applications of Theory-Oriented Quantitative Approaches.* Routledge (2021).
43. Ackerman, J. & Love, T. P. Ethnic group differences in police notification about intimate partner violence. *Violence Against Women* **20**(2), 162–185 (2014).
44. Ackerman, J. Assessing conflict tactics scale validity by examining intimate partner violence overreporting. *Psychol. Violence*. **8**(2), 207 (2018).

Acknowledgements

The authors thank the support of the National Social Science Fund of China (grant No. 19ZDA149), the National University of Singapore Assistant Professor Start-up Grant and Singapore Ministry of Education Academic Research Fund Tier 1 Grant (FY2022-FRC2-001). The authors acknowledge the China Women's Federation and the National Bureau of Statistics for their data support.

Author contributions

Y.C. and Z.C. conceived and designed the study; Z.C. and W.M. and Y.L. performed the data collection, data analysis and data interpretation; Z.C., S.W., W.G., Y.L. and W.Z. drafted and revised the manuscript.

Funding

This research has received funding from the Major Project of the National Social Science Fund of China (Grant No. 19ZDA149), the National University of Singapore Assistant Professor Start-up Grant and Singapore Ministry of Education Academic Research Fund Tier 1 Grant (FY2022-FRC2-001).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-31846-8.

Correspondence and requests for materials should be addressed to S.W., W.Z. or Y.C.