COPULA MODELS FOR ADDRESSING SAMPLE SELECTION IN THE EVALUATION OF PUBLIC HEALTH PROGRAMMES: AN APPLICATION TO THE LEEDS LET'S GET ACTIVE STUDY

Short title

Copula models for sample selection in public health

Paolo Candio[1,2], Andrew J. Hill[2], Stavros Poupakis[3], Anni-Maria Pulkki-Brännström[3,4], Chris Bojke[2], Manuel Gomes[5]

[1]Health Economics Research Centre, University of Oxford; [2]Leeds Institute of Health Sciences, University of Leeds; [3] UCL Institute for Global Health, University College London; [4] Department of Epidemiology and Global Health, Umeå University; [5]Department of Applied Health Research, University College London

Address correspondence to:

Paolo Candio, Health Economics Research Centre, University of Oxford

Richard Doll Building, Old Road Campus, Oxford, OX3 7LF United Kingdom

ORCID: 0000-0003-1521-088X

Email: paolo.candio@ndph.ox.ac.uk

ABSTRACT

Sample selectivity is a recurrent problem in public health programmes and poses serious challenges to their evaluation. Traditional approaches to handle sample selection tend to rely on restrictive assumptions. The aim of this paper is to illustrate a copula-based selection model to handle sample selection in the evaluation of public health programmes. Motivated by a public health programme to promote physical activity in Leeds (England), we describe the assumptions underlying the copula selection, and its relative advantages compared to commonly used approaches to handle sample selection, such as inverse probability weighting and Heckman's selection model. We illustrate the methods in the Leeds Let's Get Active programme and show the implications of method choice for estimating the effect on individual's physical activity. The programme was associated with increased physical activity overall, but the magnitude of its effect differed according to adjustment method. The copula selection model led to a similar effect to the Heckman's approach but with relatively narrower 95% confidence intervals. These results remained relatively similar when different model specifications and alternative distributional assumptions were considered. The copula selection model can address important limitations of traditional approaches to address sample selection, such as the Heckman model, and should be considered in the evaluation of public health programmes, where sample selection is likely to be present.

KEY POINTS FOR DECISION-MAKERS

- Evaluations of public health programmes are prone to sample selection, in which the outcome of interest is observed for a non-random subset of the programme participants.
- Selection models are typically used to address sample selection as they can control for both observed and unobserved factors associated with both selection and outcome of interest.
- The proposed copula selection model can address common limitations of traditional selection approaches and allow for a more adequate characterisation of the uncertainty associated with sample selection.

BACKGROUND

Population-level programmes that aim to improve health-related behaviours (e.g. physical activity, healthy eating) play an important role in improving population health [1-2]. Public health decision makers are increasingly interested in the evaluation of such programmes to inform resource allocation [3]. However, evaluating the effects of such programmes is challenging [4-5], not least because studies tend to be poorly designed and data may not always be collected for primary research purposes [6]. One recurrent challenge is sample selection [4], whereby the outcome of interest (e.g. physical activity, diet) is observed for a non-random subset of participants initially registered in the programme (baseline). For example, this may arise due to resource constraints (e.g. convenience sampling), or individuals self-selecting themselves into the study. Ideally, this problem should be addressed at the design stage by carefully planning the study design prior to data collection. However, this is rarely the case with public health programmes, and evaluation is often faced with sample selection problems due to poor study designs [4].

The major concern with sample selection is that the individuals selected into the study tend to be systematically different from those eligible to participate who end up not being selected. These differences are often intrinsically related to the outcome of interest, hence giving rise to misleading conclusions about the effect of the programme. For example, studying the effects of a smoking cessation programme based on individuals who successfully quit smoking will overestimate the effect of the intervention [7].

Many evaluations of public health programmes focus their analyses on the subset of participants for whom they observe the outcome of interest [8]. Such analyses are based on

the strong assumption that differences between those selected and those not selected into the study can be explained by the observed data. For example, studies often control for key observed factors that may help explain selection within a regression framework. This assumption is often denoted by 'selection on observables' [9]. However, the real challenge with sample selection is that the chances of participating in the study will depend on unobserved data. For example, in the 'Be Active' study [10], individuals with healthier lifestyles (unmeasured factors) may have been more likely to attend the gyms, and hence had greater chances of participating in the study. Therefore, in these circumstances, methods assuming 'selection on observables' may lead to misleading inferences.

Selection models can make more plausible assumptions about sample selection by controlling for both observed and unobserved factors that predict the probability of being selected into the sample and the outcome of interest, i.e. they allow for 'selection on unobservables' [11]. Among this category of models, the Heckman's selection model [12] has been particularly popular as it can be readily implemented in a regression framework in standard software [13-14]. However, this approach has been shown to be particularly sensitive to departures from assumptions about: i) the availability of exclusion restriction variables, i.e. one or more variables that are predictive of non-response, but are independent of the outcome; ii) the joint normal distribution of the selection and outcome [15-16].

A more flexible selection approach that can address these limitations is the copula framework. Copula-based selection models can make less restrictive parametric

assumptions [17], and thanks to recent software development [18], they have become easy to implement, therefore, having the potential for being adopted more widely.

This study therefore aims to illustrate the copula selection approach for addressing sample selection in the evaluation of public health programmes. The paper describes the assumptions underlying the copula selection model and discusses the relative advantages of this approach compared to more traditional methods. Motivated by a real-life population-level physical activity promotion programme in Leeds (England), we illustrate the implications of method choice for estimating the impact of the Leeds Let's Get Active (LLGA) programme on the individuals' physical activity while also providing software code for implementing the proposed copula framework.

MOTIVATING EXAMPLE

Intervention

LLGA was a City Council-led initiative to promote physical activity in adults [19]. LLGA programme offered free universal access to off-peak exercise sessions (e.g. use of free weights area) held in 17 operating City-Council leisure centres located in the most deprived areas of the city. All residents in Leeds were eligible to register to the programme.

Data collection

Baseline

Residents could sign up at any time during the programme duration (i.e. 39 months, September 2013 to December 2016). At the time of registration (i.e. baseline), individuals were asked to provide basic information on age, gender and residential postcode which,

however, was not shared with the research team due to data processing restrictions. Instead, the programme manager provided information on Index of Multiple Deprivation (IMD) status in a binary form (top 20% IMD score or not). The IMD is a neighbourhood-level composite measure that includes seven weighted domains of deprivation: income, employment, education, health, crime, barriers to housing and services and living environment. IMD provides a generic measure of relative deprivation for small areas in the UK and has been widely used by local public health departments [20].

At baseline, all participants were also asked to self-report their current level of physical activity. This was based on a single-item question derived from the short-form International Physical Activity questionnaire [21], which asked the number of active days (NAD) they had per week. An active day was defined within the questionnaire as a day with at least 30 minutes of at least moderate physical activity.

Follow up

The organisers carried out a number of "survey follow-up weeks", roughly every six months, to obtain a second physical activity measurement post exposure (i.e. registration to the programme) from a convenience sample of participants. The mode of data collection changed during the programme. In the first 18 months (cohort 1), programme staff and volunteers conducted face to face surveys in the hosting City-Council leisure centres, collecting outcome data on the present participants. From April 2015 (cohort 2), individuals registered to the programme were instead surveyed using web-based tools and email reminders. However, no record of the number of participants contacted or who contacted them to provide follow-up measurements was kept.

Participants

The LLGA programme enrolled 51,874 individuals who reported information on baseline socio-demographics and NAD outcome. Of these, only 547 (around 1%) individuals were followed up and included in the sample. **Table 1** below summarises the baseline data for all individuals who registered to the programme, comparing the whole sample of 51,874 with the sub-sample of 547 participants for whom also follow-up NAD data were available.

Most of the individuals who signed up for the LLGA programme were female (62%), aged between 16 and 40 years (61%) and not living in the most deprived areas of the city (80% of these were outside the top quintile). Almost a third (29%) reported no physical activity (0 active days) at baseline. Over 50% of participants reported between 1 to 3 active days. Overall, the sub-set of LLGA participants (n=547) were comparable to the whole group of individuals registered to the programme, except for being older, slightly less likely to be totally inactive (NAD=0) or live in the most deprived city areas, and more likely to have registered with the programme during cohort 2.

Figure 1 compares the distribution of the NAD outcome among LLGA participants (n=547) before and after registering to the programme. A shift from a right-skewed to an almost normal distribution was observed, showing that the programme increased the average NAD per week. A marked change was observed in the lower levels of physical activity, particularly in terms of proportion of totally inactive individuals (NAD=0) which reduced from a baseline of 29% to a 7.3% at follow-up (Table A in **Appendix I**).

The LLGA programme has been the subject of economic evaluation [22]. Base case results, which were based on a complete case analysis, showed that the programme was cost-effective, although as acknowledged by the authors, the risk of sample selection was of concern.

METHODS

Substantive model

The outcome of interest is the effect of the LLGA programme (β_1) on NAD six months after registration (denoted here as Y). For the purpose of illustrating the sample selection methods, we will model the outcome on the continuous scale, and assume it is a linear function of the intervention (A) and covariates (X), say:

$$Y = \beta_0 + \beta_1 A + \beta_2 X + e \qquad e \sim N(0, \sigma^2) \qquad (1)$$

where A is a time dummy for before and after registration to the LLGA programme (exposure), and the covariates X include demographic and socio-economic variables, that is age, sex and socio-economic status (IMD).

Selection on observables

Applying model 1 to the observed sample can allow for 'selection on observables' by controlling for some key observed factors (included in X) that may help explain selection. A popular approach under the 'selection on observables' assumption is the Inverse Probability Weighting (IPW) [23]. The IPW approach essentially involves creating a pseudo-population where each individual is weighted by the inverse of the probability of participating in the study. The weights are typically taken from the predicted probabilities of being selected into

the sample conditional on the observed data. This is often obtained through a logistic regression:

$$logit\ P(S) = \alpha_1 X \quad (2)$$

where $S = 1$ if the individual participates in the study, 0 otherwise. The parameters of interest, say the effect of the intervention, can be estimated by applying the substantive model to the re-weighted sample. IPW estimators tend to be imprecise as the estimated weights often vary considerably [24].

Selection on unobservables

Selection models address sample selection typically by jointly modelling the outcome and the selection mechanism [11].The validity of this approach depends on the plausibility of two main (untestable) assumptions: 1) the error terms of the two equations are assumed to follow a particular joint distribution, typically bivariate Normal, and 2) availability of variables that are predictive of selection, but unrelated to the outcome of interest, also known as the exclusion restriction assumption.

Heckman model

This model can be described as:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 A + \beta_2 X + e \\ Y_2 &= \gamma_0 + \gamma_1 X + \gamma_2 Z + u \end{aligned} \qquad \binom{e}{u} \sim BVN\left(\mathbf{0}, \mathbf{\Omega} = \begin{pmatrix} \sigma_e^2 & \rho\sigma_u\sigma_\varepsilon \\ & \sigma_u^2 \end{pmatrix}\right) \quad (3)$$

Where Y_2 is a latent variable such that $Y_2 > 0$ if individual has been selected to the sample ($S = 1$), $Y_2 \leq 0$ otherwise ($S = 0$), and Z is the set of variables that satisfy the exclusion restriction. Model 3 can be easily estimated using either maximum likelihood (ML) or

Heckman's two-step estimator [11]. This approach can be severely biased, either if the bivariate Normality does not hold (particularly when using ML estimators), or when the exclusion restriction is not met (particularly when using Heckman's two-step approach) [15,16].

Copula selection models

Copula models can address these two limitations of the Heckman model by providing a flexible approach to jointly model the outcome and selection. Essentially, copula is a function that parameterizes the dependence between 2 or more univariate marginal distributions to form a joint distribution.

Let $H(Y_1, Y_2)$ be the joint distribution of the outcome Y_1 and selection Y_2, and $F_1(Y_1)$ and $F_2(Y_2)$ be the corresponding univariate marginal cumulative distribution functions (CDFs). Sklar's theorem [25] shows that there exists a joint distribution that binds $F_1(Y_1)$ and $F_2(Y_2)$ to form a joint distribution:

$$H(Y_1, Y_2) = C(F_1(Y_1),\ F_2(Y_2); \theta) \tag{4}$$

Where $C(.)$ is the copula function, and θ incorporates the correlation between the two margins.

There are alternative ways of constructing a copula, but a popular one is the inversion method, which uses the inverted distribution function in a known multivariate distribution [26]. For example, the Gaussian copula can be described as $C\big(F_1(Y_1),\ F_2(Y_2)\big) = \Phi_2(\Phi^{-1}(F_1), \Phi^{-1}(F_2); \theta)$, where Φ_2 is the CDF of a standard bivariate normal distribution. This would be equivalent to Heckman's model 3.

Unlike the Heckman's approach, the copula selection model provides further flexibility in terms of distributional assumptions, model specification (e.g. link function) and correlation structure between the selection and outcome models [27]. In addition, the copula selection model has been shown to be less sensitive to the exclusion restriction assumption [17].

Analysis

We illustrate the copula selection model for addressing sample selection in the LLGA case study and explore the implications for estimating the effect of the programme using it compared to the Heckman model, IPW and standard regression analysis. For the Heckman, we considered the two-step estimator, and used the 'survey mode' as the exclusion restriction variable. The survey mode (face-to-face versus online) in the LLGA example was a strong predictor of whether the individual participated in the study (Table B in Appendix I), but was anticipated to affect the outcome (NAD) only through the LLGA programme.

For the copula approach, we explored different combinations of i) distributional assumptions, ii) copula model, and iii) link function that provided the best fit to the data. Both the Heckman model and IPW were implemented in STATA 15 [28], whereas the copula approach was implemented using the GJRM package in R [17] (Appendix II).

RESULTS

Standard regression analysis, which uses complete case analysis as a default, suggested that the LLGA programme was associated, on average, with 1.067 more active days compared to no intervention [21]. The estimates provided by IPW suggested a relatively lower (10.5%) effect, and a larger standard error compared to the standard regression approach. Overall,

older, male participants, living in less deprived areas, engaged in higher levels of physical activity (**Table 2**).

Both selection on unobservables approaches led to a stronger effect of the LLGA programme compared to the methods assuming selection on observables. In addition, both the Heckman and the copula models suggested that age and socioeconomic status were not significantly associated with physical activity levels. The estimated levels of correlation between selection and outcome models were relatively small in both selection approaches (e.g. rho=-0.014, Heckman model). However, the copula selection model led to smaller standard errors compared to the Heckman model.

Capitalising on the flexibility of the method, we allowed alternative specifications of the selection model and different copulas for the copula selection model, but the results remained relatively consistent. We also considered the NAD outcome as a count variable and estimated the parameters of interest using Poisson models (Appendix III).

In addition, **Figure 2** compares the quantiles of the data against the quantiles of the desired distribution for the four copulas that provided the best fit to the data (based on AIC / BIC scores, Appendix III). From left to right: Joe, Plackett, Clayton and Frank copulas. This figure indicates that both the Plackett and Frank copulas can better represent the dependence structure between the selection and outcome models, compared to the other two copulas. Overall, the net effect of a more appropriate dependence structure and joint distribution for the error terms, and lower dependence on the strength of the exclusion restriction, is a more precise estimate of LLGA's effect, compared to the Heckman approach.

DISCUSSION

This paper is concerned with the recurrent problem of bias due to sample selection in the evaluation of public health programmes. Motivated by a real-life population-level programme to promote physical activity in the general population, we illustrated the application of a copula-based selection approach which addresses some of the limitations common to traditional approaches to correct for sample selection. *While this is an area where sample selection issues are particularly concerning, in principle, this framework can be used to address sample selection in the evaluation of other health interventions or policies [29]*.

The copula selection model provides ample flexibility to i) choose a plausible joint distribution for the error terms, ii) explore alternative correlation structures (copulas), iii) allow for more complex model specification of the outcome and selection, iv) provide estimates that are less sensitive to the lack of strong exclusion restriction variables. Overall, the copula approach can allow for a more adequate characterisation of the uncertainty associated with sample selection in the evaluation of public health programmes, and help future studies provide more sound evidence to inform decision making [30]. Thanks to its flexibility in modelling outcomes jointly, the proposed copula approach could also be particularly useful in economic evaluation studies, by jointly modelling costs and effects, as well as sample selection (trivariate model) [30].

The results from applying the alternative methods to correct for sample selection show that LLGA was associated with increased physical activity overall, but the magnitude of its effect differed according to adjustment method. This study finds that an approach that ignored

sample selection have under-estimated the effect of the programme. This was expected as in our case-study the more active individuals were relatively more likely to participate in the study after registration, and hence the effect of LLGA would be smaller for these individuals compared to the other participants who had lower levels of physical activity and benefited relatively more from the LLGA programme. The two selection on unobservables methods aligned with this expectation, with the copula model leading to a slightly stronger effect and relatively lower standard errors compared to the Heckman's approach. These results remained relatively similar when different model specifications and alternative distributional assumptions were considered for the copula approach.

The recurrent problem of sample selection in the evaluation of public health programmes emphasises the need for a shift from retrospective to prospective evaluations. In addition, it highlights the need to take active steps towards minimising the extent of selection bias. These may include: i) planning the study design at early stages before any implementation has taken place, ii) developing a careful plan for data collection, including clear definition of the outcomes of interest, specification of follow up times, and strategies to engage the participants and maximise follow up participation, and iii) collecting information on the reasons for individuals to decline to participate or drop out of the study.

This study presented some limitations. The paucity of the data available , particularly on covariate information, limited the extent of statistical analysis for adequately addressing the potential confounding. The analyses relied on a strong assumption that individuals would maintain the same level of physical activity had they not participated in the programme. This is analogous to the parallel trend assumption, where the baseline values act as the

'control group'. While the plausibility of this assumption may be questionable, this limitation was common to both the Heckman and Copula selection models, and was not anticipated to exacerbate any differences between the two approaches. As a result, this study does not attempt to make any causal claims about the effect of the LLGA programme.

In addition, assessing the relative performance of the different methods was beyond the scope of this paper. A recent study investigated the statistical properties of the Heckman selection model and copula approach [17], and found that the copula approach provided the lowest biases and mean squared error across a wide range of typical scenarios with sample selection. Therefore, we would anticipate those tangible benefits (less biased and more precise estimates) to apply to this study as well. However, the differences across methods in our case study are small, and hence the benefits of the copula model lie mostly in its flexibility.

Despite the broad flexibility provided by the copula selection model, a potential limitation of this approach is that it is less straightforward to implement compared to simpler approaches, such as the Heckman model or IPW. To encourage the uptake of the methods [31], we provided implementation code in freely available software (R). In addition, the use of copula selection models in settings with missing covariates is more challenging as it would require an additional selection equation for the missing covariates.

CONCLUSIONS

This study illustrates the flexibility and relative merits of the copula selection models to address sample selection in the evaluation of public health programmes. Off-the-shelf

traditional approaches to address sample selection, such as IPW and the Heckman model, continue to be widely used in practice. However, these approaches often rely on restrictive assumptions. The copula selection model can address these limitations and should be considered in evaluation studies of public health programmes, where sample selection is likely to occur.

REFERENCES

1. World Health Organization. Health Promotion. 2020 [cited 2020 18 July]; Available from: https://www.who.int/health-topics/health-promotion#tab=tab_1.

2. Centers for Disease Control and Prevention. Promoting Healthy Behaviors. 2020 [cited 2020 18 July]; Available from: https://www.cdc.gov/healthyschools/healthybehaviors.htm.

3. House of Lords Science and Technology Select Committee, Behaviour change. 2nd report of session 2010–12., The Stationery Office, Editor. 2011: London.

4. Craig, P., et al., Developing and evaluating complex interventions: the new Medical Research Council guidance. Int J Nurs Stud, 2013. 50(5): p. 587-92.

5. Fletcher, A., et al., Realist complex intervention science: Applying realist principles across all phases of the Medical Research Council framework for developing and evaluating complex interventions. Evaluation (Lond), 2016. 22(3): p. 286-303.

6. Skivington, K.M., Lynsay; Craig, Peter; Simpson, Sharon; Moore, Laurence. Developing and evaluating complex interventions: updating Medical Research Council guidance to take account of new methodological and theoretical approaches. 2018.

7. Adda, J. and F. Cornaglia, Taxes, Cigarette Consumption, and Smoking Intensity. Am Econ Rev, 2006. 96(4): p. 1013-28.

8. Raghunathan, T.E., What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. Annual Review of Public Health, 2004. Vol. 25:99-117.

9. Craig P, C.C., Gunnell D, et al, Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. J Epidemiol Community Health, 2012. 66:1182-1186.

10. Frew, E.J., et al., Cost-effectiveness of a community-based physical activity programme for adults (Be Active) in the UK: an economic analysis within a natural experiment. British Journal of Sports Medicine, 2014. 48(3): p. 207.

11. Molenberghs, G., et al., eds. Handbook of Missing Data Methodology. 2014, Chapman and Hall/CRC: New York.

12. Heckman, J.J., Sample selection bias as a specification error. Econometrica: Journal of the econometric society, 1979: p. 153-161.

13. Bärnighausen, T., et al., Correcting HIV Prevalence Estimates for Survey Nonparticipation Using Heckman-type Selection Models. Epidemiology, 2011. 22(1).

14. Koné, S., et al., Heckman-type selection models to obtain unbiased estimates with missing measures outcome: theoretical considerations and an application to missing birth weight data. BMC Medical Research Methodology, 2019. 19(1): p. 231.

15. Puhani, P., The Heckman Correction for Sample Selection and Its Critique. Journal of Economic Surveys, 2000. 14(1): p. 53-68.

16. Gomes, M., et al., Estimating treatment effects under untestable assumptions with nonignorable missing data. Statistics in Medicine, 2020. 39(11): p. 1658-1674.

17. Gomes, M., et al., Copula selection models for non-Gaussian outcomes that are missing not at random. Statistics in Medicine, 2019. 38(3): p. 480-496.

18. Marra, G. and R. Radice, GJRM: generalised joint regression modelling. R package version 0.1-1. 2017.

19. Active Leeds. Leeds Let's Get Active. Available from: https://active.leeds.gov.uk/classesandactivities/leeds-lets-get-active

20. Fairburn, J. Maier, W. and Braubach M. Incorporating Environmental Justice into Second Generation Indices of Multiple Deprivation: Lessons from the UK and Progress Internationally. Int J Environ Res Public Health. 2016. https://doi.org/10.3390/ijerph13080750.

21. Craig, C.L., et al., International Physical Activity Questionnaire: 12-Country Reliability and Validity. Medicine & Science in Sports & Exercise, 2003. 35(8).

22. Candio, P., et al., Cost-effectiveness of a proportionate universal offer of free exercise: Leeds Let's Get Active. Journal of Public Health, 2020. https://doi.org/10.1093/pubmed/fdaa113.

23. Wooldridge, J.M., Inverse probability weighted M-estimators for sample selection, attrition, and stratification. Portuguese Economic Journal, 2002. 1(2): p. 117-139.

24. Seaman, S.R. and I.R. White, Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res, 2013. 22(3): p. 278-95.

25. Sklar, A., Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 1959. 8, 229–231.

26. Nelsen, R.B., Methods of Constructing Copulas, in An Introduction to Copulas, Springer, Editor. 2006, Springer, New York, NY.

27. Smith, M.D., Modelling sample selection using Archimedean copulas. The Econometrics Journal, 2003. 6(1): p. 99-123.

28. StataCorp, Stata Statistical Software: Release 15. 2017, StataCorp LLC: College Station, TX.

29. Tamakloe, R., Hong, J., Park, D. A copula-based approach for jointly modeling crash severity and number of vehicles involved in express bus crashes on expressways considering temporal stability of data. Accid Anal Prev. 2020;146:105736

30. Briggs, A., Claxton, K., & Sculpher, M., Decision modelling for health economic evaluation. 2006, Oxford: Oxford University Press.

31. Incerti, D., Thom H., Baio G., Jansen J. P., *R You Still Using Excel? The Advantages of Modern Software Tools for Health Technology Assessment.* Value in Health, 2019(ISSUE 5): p. P575-579.

DECLARATIONS

Ethical approval

Analysis of anonymised data did not require ethical approval.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

PC was supported through the White Rose PhD Studentship Network scheme as part of the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber.

Conflicts of interest

The authors declare no conflict of interest.

Availability of data and material

No data are available. Programme data have been provided by the local City Council under a Data Processing Agreement.

Code availability

Software code for implementing the proposed copula framework using the R package GJRM is provided.

Authors' contributions

PC and MG were responsible for designing the study and drafting the manuscript. AJH, SP, AP and CB revised the paper critically for intellectual content. All the authors read and approved the final version of the manuscript.