SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2021, Vol. 30(3) 702–716
© The Author(s) 2020

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280220970986
journals.sagepub.com/home/smm

SAGE

Employing a latent variable framework to improve efficiency in composite endpoint analysis

Martina McMenamin[1], Jessica K Barrett[1], Anna Berglind[2] and James MS Wason[1,3]

Abstract

Composite endpoints that combine multiple outcomes on different scales are common in clinical trials, particularly in chronic conditions. In many of these cases, patients will have to cross a predefined responder threshold in each of the outcomes to be classed as a responder overall. One instance of this occurs in systemic lupus erythematosus, where the responder endpoint combines two continuous, one ordinal and one binary measure. The overall binary responder endpoint is typically analysed using logistic regression, resulting in a substantial loss of information. We propose a latent variable model for the systemic lupus erythematosus endpoint, which assumes that the discrete outcomes are manifestations of latent continuous measures and can proceed to jointly model the components of the composite. We perform a simulation study and find that the method offers large efficiency gains over the standard analysis, the magnitude of which is highly dependent on the components driving response. Bias is introduced when joint normality assumptions are not satisfied, which we correct for using a bootstrap procedure. The method is applied to the Phase IIb MUSE trial in patients with moderate to severe systemic lupus erythematosus. We show that it estimates the treatment effect 2.5 times more precisely, offering a 60% reduction in required sample size.

Keywords

Composite endpoint, latent variable model, responder analysis, systemic lupus erythematosus

1 Introduction

Composite endpoints combine a number of individual outcomes in order to determine the effectiveness or efficacy of a treatment for a given disease. A subset of these endpoints are composite responder endpoints in which patients are classed as ‘responders’ or ‘non-responders’ based on whether they cross predefined thresholds in the individual outcomes. These endpoints are common in autoimmune diseases such as systemic lupus erythematosus (SLE), lupus nephritis and sjögrens syndrome. Physicians and health authorities advocate these endpoints as they attempt to capture the effect in multiple dimensions of the disease. For instance, in SLE the composite is used to ensure that as improvement occurs in the SLE Disease Activity Index (SLEDAI), there is no simultaneous worsening in any other organ domains. In other diseases, such as rheumatoid arthritis and myositis, the composite endpoints capture improvements from both a clinical and patient perspective.

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
[2]Late RIA, R&D BioPharmaceuticals AstraZeneca, Gothenburg, Sweden
[3]Institute of Health and Society, Newcastle University, Newcastle, UK

Corresponding author:
Martina McMenamin, MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK.
Email: martina.mcmenamin@mrc-bsu.cam.ac.uk

The composite of interest may be a combination of continuous and discrete outcomes which are typically collapsed in to a single binary responder index and analysed using a logistic regression model, termed the standard binary method. The work by Wason and Seaman[1] showed that analysing in this way solves problems with multiplicity, however, at the expense of large losses in efficiency due to discarding information on how close each patient was to the responder threshold. For a composite containing a single continuous and binary endpoint they proposed the augmented binary method, a likelihood-based approach using the theory of factorisation models. They factorised the joint distribution and fit a univariate model to each component of the factorisation. This accounts for correlations between the outcomes by including one response as a covariate in the model for the other response. In the graphical modelling literature, this has been termed the 'conditional Gaussian distribution'.[2,3] The augmented binary method has been shown to reduce the required sample size in clinical trials by approximately 35% in a range of applications where the composite is formed of one continuous and one binary outcome.[1,4–6] For composites made up of multiple continuous, ordinal and binary outcomes, we hypothesise that we may further increase efficiency due to the additional information in continuous and ordinal components. However, one limitation of these methods beyond the bivariate scenario is the range of possibilities for the factorisations, with no consensus on how this should be determined. In the case of an endpoint with four components, this amounts to 24 possible factorisations, each of which may result in different conclusions.[7,8] To model complex, higher dimensional composite endpoints, we require a more general joint modelling framework.

To achieve this, we propose adopting a correlated Gaussian distribution for the components by assuming that the discrete outcomes are manifestations of underlying continuous variables, subject to some threshold specifications.[9,10] This framework dates back to Pearson (1904)[11] in relation to his generalised theory of alternative inheritance and has received much consideration in the literature since, although the terminology has been largely inconsistent. In the graphical modelling literature, they have been termed 'conditional grouped continuous models (CGCMs)'[12] and elsewhere have been referred to as 'multivariate ordered probit models',[13] 'correlated probit models'[14] and 'generalised multivariate probit models'.[15] The general mixed-data model for mixed nominal, ordinal, and continuous data also reduces to a CGCM in the absence of nominal outcomes.[16] The application of CGCMs for a mixture of continuous and binary outcomes has featured throughout the statistics literature.[17–19] A CGCM has also been proposed in clinical trials to deal with the problem of multiple continuous and binary co-primary endpoints, where a treatment effect must be achieved in all outcomes to conclude it is successful overall.[20,21] Extensions have allowed for modelling continuous and ordinal variables, with applications such as developmental toxicology and the joint modelling of hybrid traits in genetics.[22–27] Other work has combined employing a latent Gaussian distribution for the response variables and introducing latent variables in the model; however, these ideas are most applicable in the longitudinal setting.[8,14,28–30]

The purpose of this work is to employ the CGCM framework to a different end. Rather than using the latent Gaussian distribution to make inference on multivariate outcomes, we will use it to model the multiple components within a composite, while still making inference on the one-dimensional composite endpoint. By employing the latent structure to collapse the multiple outcomes after the model is fitted, rather than before, we aim to greatly improve efficiency whilst still providing the overall treatment effect on the composite.

The paper proceeds as follows. In Section 2 we discuss SLE, the motivating example for the methods. In Section 3 we introduce the latent variable model for our application and discuss how we conduct estimation and inference for the composite endpoint problem. In Section 4 we introduce the comparison methods. In Section 5 we compare the behaviour of the latent variable model with the augmented binary and standard binary methods, including the case when the key assumptions are not satisfied. In Section 6 we apply the methods to the Phase IIb MUSE trial in patients with moderate to severe SLE. Finally, in Section 7 we discuss our findings and make recommendations for use.

2 Motivating example

Table 1 shows examples of composite endpoints combining multiple criteria to define response. Responders in fibromyalgia must respond in two continuous and one ordinal component; however, responders in trials for frailty or soft tissue infections must respond in a total of five continuous and discrete components. In what follows, we will focus specifically on SLE as a motivating example however, the methods introduced will be relevant to other diseases using endpoints with a similar structure.

In the SLE endpoint, a continuous Physician's Global Assessment (PGA) measure, a continuous SLEDAI measure and an ordinal British Isles Lupus Assessment Group (BILAG) measure are combined to form the SLE

Table 1. Examples of diseases that use complex composite endpoints combining multiple discrete and continuous measures to determine effectiveness of a treatment including criteria for response and how each component is measured.

Disease	Responder endpoint	Measured by
Fibromyalgia	• Achieved a 30% improvement in pain	Electronic diary
	• 30% improvement in functional status	Subscale of Fibromyalgia Impact Questionnaire (FIQ)
	• Improved, much improved, or very much improved	7-point Patient Global Impression of Change (PGIC) scale
Frailty	• BMI <18.5 kg/m2 OR >10% weight loss since last wave	Weight and height
	• One positive answer to exhaustion questions	CES-D questionnaire
	• Low grip strength (M < 31.12 kg, F < 17.60 kg)	E.g. Jamar hand dynamometer
	• Gait speed (M < 0.691 m/s, F < 0.619 m/s)	Distance/time
	• Low activity (M < 16.5 activity units, F < 13.5 activity units)	Activity units derived using intensity versus frequency
Necrotising soft tissue infections	• Alive until day 28	Yes/No
	• Day 14 debridements ≤3	Surface area
	• No amputation if debridement	Yes/No
	• Day 14 mSOFA score ≤1	mSOFA score – composite additively
	• Reduction of at least 3 score points in mSOFA score	combining scores in different systems mSOFA score – composite additively combining scores in different systems
Systemic lupus erythematosus	• Change in SLEDAI ≤ −4	SLE Disease Activity Index
	• Change in PGA < 0.3	Physicians Global Assessment
	• No Grade A or more than one Grade B in BILAG	British Isles Lupus Assessment Group
	• Reduction in oral corticosteroids	Medical Notes

Responder Index (SRI).[31] This is combined with a binary measure, which indicates tapering of the oral corticosteroids dose, to form the overall SLE responder endpoint of interest. The BILAG measure is a translational index which measures changing severity of clinical manifestations in nine organ systems. It has five levels for each parallel organ system, labelled Grade A–Grade E.[32] Patients must meet the response criteria in all components in order to be classed as a responder overall. A figure denoting the structure of the SLE responder endpoint is shown in Appendix A of the supplemental material.

3 Methods

3.1 Model

The mean structure for the outcomes is shown in equation (1). The baseline measures y_{10} and y_{20} are included in the model for Y_1 and Y_2, respectively.

$$\begin{aligned} Y_{i1} &= \alpha_0 + \alpha_1 T_i + \alpha_2 y_{i10} + \varepsilon_{i1} \\ Y_{i2} &= \beta_0 + \beta_1 T_i + \beta_2 y_{i20} + \varepsilon_{i2} \\ Y^*_{i3} &= \gamma_1 T_i + \varepsilon^*_{i3} \\ Y^*_{i4} &= \psi_0 + \psi_1 T_i + \varepsilon^*_{i4} \end{aligned} \tag{1}$$

The observed discrete variables are related to the latent continuous variables by partitioning the latent variable space, as shown in equation (2). The lower and upper thresholds for both discrete variables are set at $\tau_{03} = \tau_{04} = -\infty, \tau_{53} = \tau_{24} = \infty$ and the binary cut-point is set at $\tau_{14} = 0$. The intercept term for the ordinal variable is set at

$\gamma_0 = 0$ so that the cut-points $\tau_{13}, \tau_{23}, \tau_{33}, \tau_{43}$ may be estimated. The intercept for the binary outcome ψ_0 may be estimated as $\tau_{14} = 0$.

$$Y_{i3} = \begin{cases} \text{Grade E} & \text{if } \tau_{03} \leq Y_{i3}^* < \tau_{13}, \\ \text{Grade D} & \text{if } \tau_{13} \leq Y_{i3}^* < \tau_{23}, \\ \text{Grade C} & \text{if } \tau_{23} \leq Y_{i3}^* < \tau_{33}, \\ \text{Grade B} & \text{if } \tau_{33} \leq Y_{i3}^* < \tau_{43}, \\ \text{Non} - \text{responder} & \text{if } \tau_{43} \leq Y_{i3}^* < \tau_{53} \end{cases} \quad Y_{i4} = \begin{cases} 0, & \text{if } \tau_{04} \leq Y_{i4}^* < \tau_{14}, \\ 1, & \text{if } \tau_{14} \leq Y_{i4}^* < \tau_{24} \end{cases} \tag{2}$$

Following these assumptions, we can model the error terms in equation (1) as multivariate normal with zero mean and variance–covariance matrix Σ, as shown in (3). Note that the error variances for $\varepsilon_3^*, \varepsilon_4^*$ are $\sigma_3 = 1$ and $\sigma_4 = 1$, however this does not represent a constraint on the model but rather a rescaling required for identifiability.

$$(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}^*, \varepsilon_{i4}^*) \sim N(\mathbf{0}, \Sigma) \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1 & \rho_{14}\sigma_1 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2 & \rho_{24}\sigma_2 \\ \rho_{13}\sigma_1 & \rho_{23}\sigma_2 & 1 & \rho_{34} \\ \rho_{14}\sigma_1 & \rho_{24}\sigma_2 & \rho_{34} & 1 \end{pmatrix} \tag{3}$$

The joint likelihood contribution for patient i with, for instance, $Y_{i3} =$ Grade C and $Y_{i4} = 0$, can be factorised as shown below

$$l(\boldsymbol{\theta}; \mathbf{Y}_{\mathbf{i}}^*) = f(Y_{il}, Y_{i2}; \boldsymbol{\theta}) \int_{\tau_{23}}^{\tau_{33}} \int_{-\infty}^{0} f(Y_{i3}^*, Y_{i4}^* | Y_{i1}, Y_{i2}; \boldsymbol{\theta}) \, \mathrm{d}y_4^* \, \mathrm{d}y_3^* \tag{4}$$

where $\boldsymbol{\theta}$ is a vector which contains all model parameters. Note that it is possible to evaluate the joint likelihood contribution for patient i using $f(Y_{i1}, Y_{i2}, Y_{i3}^*, Y_{i4}^*; \boldsymbol{\theta})$; however, factorising as in equation (4) may reduce computational times, particularly in high-dimensional models. This formulation also allows us to express the observed likelihood as shown in equation (5).

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{Y}) = & \prod_{i=1}^{N} \prod_{w=1}^{5} \prod_{k=1}^{2} f(Y_{i1}, Y_{i2}; \boldsymbol{\theta}) \\ & [pr(Y_{i3} = w, Y_{i4} = k | Y_{i1} = y_{i1}, Y_{i2} = y_{i2}; \boldsymbol{\theta})]^{I\{Y_{i3}=w, Y_{i4}=k\}} \end{aligned} \tag{5}$$

The joint probability of patients having discrete measurements $Y_{i3} = w$ and $Y_{i4} = k$ must be multiplied over the five ordinal levels and two binary levels resulting in 10 combinations of the probabilities to be calculated as shown in equation (6)

$$\begin{aligned} & pr(Y_{i3} = w, Y_{i4} = k | Y_{i1} = Y_{i1}, Y_{i2} = Y_{i2}; \boldsymbol{\theta}) \\ & \quad = \Phi_2(\tau_{w3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}) - \Phi_2(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{k4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}) \\ & \qquad - \Phi_2(\tau_{w3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}) + \Phi_2(\tau_{(w-1)3} - \mu_{3|1,2}, \tau_{(k-1)4} - \mu_{4|1,2}; \Sigma_{3,4|1,2}) \end{aligned} \tag{6}$$

where Φ_2 is the bivariate standard normal distribution function and $\mu_{3|1,2}, \mu_{4|1,2}$ and $\Sigma_{3,4|1,2}$ are derived using the rules of conditional multivariate normality, resulting in equation (7).

$$\begin{aligned} \mu_{3|1,2} &= \mu_3 + \frac{(\rho_{13} - \rho_{12}\rho_{23})}{\sigma_1(1 - \rho_{12}^2)}(Y_{i1} - \mu_1) + \frac{(\rho_{23} - \rho_{12}\rho_{13})}{\sigma_2(1 - \rho_{12}^2)}(Y_{i2} - \mu_2) \\ \mu_{4|1,2} &= \mu_4 + \frac{(\rho_{14} - \rho_{12}\rho_{24})}{\sigma_1(1 - \rho_{12}^2)}(Y_{i1} - \mu_1) + \frac{(\rho_{24} - \rho_{12}\rho_{14})}{\sigma_2(1 - \rho_{12}^2)}(Y_{i2} - \mu_2) \end{aligned} \tag{7}$$

$$\Sigma_{3,4|1,2} = \begin{pmatrix} 1 - \dfrac{\rho_{13}^2 - 2\rho_{12}\rho_{13}\rho_{23} + \rho_{23}^2}{1 - \rho_{12}^2} & \rho_{34} - \dfrac{\rho_{13}\rho_{14} - \rho_{12}\rho_{13}\rho_{24} - \rho_{12}\rho_{14}\rho_{23} + \rho_{23}\rho_{24}}{1 - \rho_{12}^2} \\ \rho_{34} - \dfrac{\rho_{13}\rho_{14} - \rho_{12}\rho_{13}\rho_{24} - \rho_{12}\rho_{14}\rho_{23} + \rho_{23}\rho_{24}}{1 - \rho_{12}^2} & 1 - \dfrac{\rho_{14}^2 - 2\rho_{12}\rho_{14}\rho_{24} + \rho_{24}^2}{1 - \rho_{12}^2} \end{pmatrix}$$

We discuss the intuition for equation (6) in Appendix B of the supplemental material.

3.2 Estimation

As the variance parameters (σ_1, σ_2) are required to be greater than 0, we introduce parameters (ω_1, ω_2) such that $\sigma_1 = exp(\omega_1)$ and $\sigma_2 = exp(\omega_2)$. This transformation ensures that the variance is above 0 whilst allowing the estimated parameters to take any real value. We must also ensure that the correlation parameters $(\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34})$ are estimated within $(-1,1)$ by introducing $(\omega_{12}, \omega_{13}, \omega_{14}, \omega_{23}, \omega_{24}, \omega_{34})$, where

$$\rho_{12} = 2expit(\omega_{12}) - 1, \rho_{13} = 2expit(\omega_{13}) - 1, \rho_{14} = 2expit(\omega_{14}) - 1,$$
$$\rho_{23} = 2expit(\omega_{23}) - 1, \rho_{24} = 2expit(\omega_{24}) - 1, \rho_{34} = 2expit(\omega_{34}) - 1$$

We fit the model in R by coding the likelihood function and probability of response. The bivariate distribution functions in equation (6) are estimated using 'pmvnorm', applying the method of Genz.[33] The likelihood maximisation is conducted using a quasi-Newton method based on port routines and can be implemented using the 'nlminb' function in the 'optimx' package. This is the best performing method in this setting in terms of accuracy and convergence rate however, it is also the slowest. Note that model parameters for these types of models may also be estimated using weighted least squares in the 'lavaan' package in R however, it allows for less flexibility in fitting the model.[34] We use the 'hessian' function in the 'numDeriv' package to calculate the Hessian matrix using Richardson extrapolation[35] and obtain the covariance matrix of the model parameters by inverting the Hessian. We ensure finite sample positive definiteness through solving a constrained optimisation problem,[36] where the nearest correlation matrix projection is used to compute the nearest correlation matrix. We achieve this using the 'near PD' function, which implements the algorithm of Higham.[37]

3.3 Inference

We wish to make inference on the probability of response. Let S_i be an indicator for patient i denoting whether or not they achieved response defined by $S_i = 1$ if $Y_{i1} \le \eta_1, Y_{i2} \le \eta_2, Y_{i3}^* \le \eta_3, Y_{i4}^* \le \eta_4$. Therefore

$$P(S = 1|T, y_{10}, y_{20}) = \int_{-\infty}^{\eta_1} \int_{-\infty}^{\eta_2} \int_{-\infty}^{\eta_3} \int_{-\infty}^{\eta_4} f_{\mathbf{Y}^*}(\mathbf{Y}^*; T, y_{10}, y_{20})\, \mathrm{d}y_4^*\, \mathrm{d}y_3^*\, \mathrm{d}y_2\, \mathrm{d}y_1 \tag{8}$$

where $f_{\mathbf{Y}^*}(\mathbf{Y}^*; .)$ is the multivariate normal density function for the observed and latent continuous measures. We obtain the integrand in equation (8) by using the fitted values of the parameters in the conditional mean and conditional covariance matrix in equation (7). Parameter estimates from these methods are maximum likelihood estimates and so we avail of asymptotic maximum likelihood theory. The integral in equation (8) is evaluated using the 'R2Cuba' package to obtain estimates for each patient, assuming they were treated $\tilde{p}_{i1}$ and not treated $\tilde{p}_{i0}$. The odds ratio treatment effect is then defined as shown in equation (9).

$$\tilde{\delta} = \frac{\left(\dfrac{\sum_{i=1}^{N} \tilde{p}_{i1}}{N - \sum_{i=1}^{N} \tilde{p}_{i1}}\right)}{\left(\dfrac{\sum_{i=1}^{N} \tilde{p}_{i0}}{N - \sum_{i=1}^{N} \tilde{p}_{i0}}\right)} \tag{9}$$

Note that we can easily define a risk difference or risk ratio using these quantities but in what follows we consider $\tilde{\delta}$ to be the effect of interest. The standard error estimates are obtained using the delta method. This requires the covariance matrix of the maximum likelihood estimates $\mathrm{Cov}(\hat{\boldsymbol{\theta}})$ and $''\tilde{\boldsymbol{\delta}}$, the vector of partial derivatives of $\tilde{\delta}$ with respect to each of the parameter estimates. The variance of $\tilde{\delta}$ is obtained as shown in equation (10).

$$Var(\tilde{\delta}) = (''\tilde{\boldsymbol{\delta}})TCov(\hat{\boldsymbol{\theta}})(''\tilde{\boldsymbol{\delta}}) \tag{10}$$

Another important consideration for the model is how to assess goodness-of-fit. We propose an extension to an existing method for application in this case, which is detailed in Appendix C in the supplemental material.

4 Comparison methods

4.1 Standard binary method

The standard binary method is a logistic regression on the overall responder index, as shown in equation (11)

$$logit(Pr(S_i = 1|T_i, y_{i10}) = \alpha_0 + \alpha_1 T_i + \alpha_2 y_{i10} + \alpha_3 y_{i20} \tag{11}$$

The maximum likelihood estimates and the covariance matrix can be used directly to estimate the odds ratio and standard error.

4.2 Augmented binary method

The augmented binary method is a joint modelling approach which retains the information from one continuous component and combines the remaining outcomes to form a binary response outcome. The model is shown below where the baseline measures for Y_{i1} and Y_{i2} are included for comparison, as they are accounted for in the mean structure of the latent variable method. As one time point is modeled, we can use a linear model for Y_{i1} as shown in equation (12)

$$Y_{i1} = \alpha_0 + \alpha_1 T_i + \alpha_2 y_{i10} + \alpha_3 y_{i20} + \varepsilon_i \tag{12}$$

where $\varepsilon \sim N(0, \sigma)$. In this case, the failure time binary indicator will contain information from the remaining three components. F_i is set equal to 0 if $Y_{i2} \leq \eta_2$, Y_{i3} is Grade B–E and $Y_{i4} = 0$, otherwise the patient is labelled a non-responder in these components and $F_{i1} = 1$. F_i is modelled using the logistic regression model in equation (13). Note that as this method retains the additional information contained in only one of the continuous measures, the most informative continuous outcome should be chosen.

$$logit(Pr(F_i = 1|T_i, y_{i10}, y_{i20}) = \beta_0 + \beta_1 T_i + \beta_2 y_{i10} + \beta_3 y_{i20} \tag{13}$$

Maximum likelihood estimates for the parameters are obtained from fitting models (12) and (13). As in the latent variable method, equation (14) is used to obtain probability of response estimates for each patient, assuming they were treated $\tilde{p}_{i1}$ and not treated $\tilde{p}_{i1}$.

$$P(Y_1 \leq \eta_1, F_1 = 0|T, y_{10}, y_{20}) \ = \int_{-\infty}^{\eta_1} P(F_1 = 0|T, y_{10}, y_{20}) f_{Y_1}(y_1; T, y_{10}, y_{20})\, \mathrm{d}y_1 \tag{14}$$

As before, these quantities are used to define an odds ratio, risk ratio or risk difference.

5 Simulation study

5.1 Data generating model

Initially, we investigate the properties of the methods when the assumptions of the latent variable model are satisfied. The parameter values in the 'baseline' scenario are chosen to simulate the settings where composite endpoints are typically recommended for use. Namely, that all components contribute to classifying responders and non-responders and that components are coherent but not perfectly correlated. The parameter values have been informed by the MUSE trial dataset, in particular the correlation structure. The response probability in the control arm is 0.28 and in the treatment arm is 0.38, resulting in an odds ratio approximately equal to 1.60. The parameter values selected for the model in equation (1) are shown in Appendix D of the supplemental material. From this baseline case, we vary parameters to determine how the methods behave under various scenarios of interest. In particular, we vary the treatment effect, the responder threshold and the drivers of response. The parameter values for these data generating models are also included in Appendix D.

5.2 Performance criteria

The methods are evaluated against a range of performance criteria. The bias of the methods is calculated using $\frac{1}{n_{sim}}\sum_{j=1}^{n_{sim}} \hat{\delta}_j - \delta$, where $\hat{\delta}_j$ is the estimated treatment effect in repetition j, δ is the true treatment effect and n_{sim} is the number of simulated datasets. We assess the coverage of the methods using $\frac{1}{n_{sim}}\sum_{j=1}^{n_{sim}} 1(\hat{\delta}_{low,j} \leq \delta \leq \hat{\delta}_{upp,j})$ where $\hat{\delta}_{low,j}$ and $\hat{\delta}_{upp,j}$ are the estimated lower and upper confidence interval limits in repetition j. The power is evaluated using $\frac{1}{n_{sim}}\sum_{j=1}^{n_{sim}} 1(p_j < \alpha)$, where p_j is the p-value returned by the jth repetition and α is the nominal significance level. We are also interested in the relative precision of the methods, which is obtained using $\frac{\hat{Var}(\hat{\delta}_j)_B}{\hat{Var}(\hat{\delta}_j)_A}$, where $\hat{Var}(\hat{\delta}_j)_A$ and $\hat{Var}(\hat{\delta}_j)_B$ are the estimated variances of the estimated treatment effect in repetition j for method A and method B, respectively.

Another useful measure for evaluating performance is the bias-corrected coverage.[38] It is obtained using $\frac{1}{n_{sim}}\sum_{j=1}^{n_{sim}} 1(\hat{\delta}_{low,j} \leq \bar{\delta} \leq \hat{\delta}_{upp,j})$, where $\bar{\delta}$ is the mean $\hat{\delta}_j$. By assessing coverage using $\bar{\delta}$ as the true treatment effect, we can determine whether poor coverage is due to bias in the treatment effect estimate or some other cause. If coverage is not nominal but bias-corrected coverage is nominal then we can attribute poor coverage to bias and attempt to correct this.[38]

5.3 Results

5.3.1 Varying treatment effect

Figure 1 shows the bias estimates for each method. The latent variable method is unbiased for smaller treatment effects but a small bias towards the null is introduced as the treatment effect increases. The augmented binary method is biased away from the null in this setting and the bias increases as the treatment effect increases. The standard binary method is unbiased, as we would expect for a logistic regression in a large sample.

The latent variable method has nominal coverage for smaller treatment effects; however, the coverage probability decreases as the treatment effect increases. The augmented binary method has coverage of approximately 0.91, which also decreases when the treatment effect increases. We find that the binary method has approximately nominal coverage.

Figure 2 shows both the coverage and bias-corrected coverage for the methods. The bias-corrected coverage of the latent variable method is 0.95, which indicates that any under-coverage is due to the bias present for larger treatment effect estimates. The augmented binary method shows small improvements in bias-corrected coverage, indicating that under-coverage is present in this method due to reasons other than bias. This may be due to model misspecification.

The power of the methods is shown in Figure 3. The performance of the binary and augmented binary methods is as we would expect based on previous findings.[1,4] The latent variable method offers much higher power. In this setting it has close to 100% power for odds ratios larger than 1.6, an effect that is plausible to observe in a trial. The mean squared error (MSE) for the standard and augmented binary methods is approximately 6.5 times that of the latent variable method. The MSE plot is shown in Appendix E of the supplemental material.

5.3.2 Varying η_1

To understand more about the precision performance of the augmented binary method in particular, we vary the responder threshold η_1 to change the proportion of responders in that outcome. We find that the precision gains from the augmented binary method diminish as the threshold increases. This is intuitive, as gains in efficiency fall as the continuous component becomes less responsible for driving response. It is interesting to note that all precision gains are lost for any thresholds above -4. Therefore, even when 20% of patients are non-responders, all efficiency gains are lost. The percentage of responders needed to improve efficiency using the

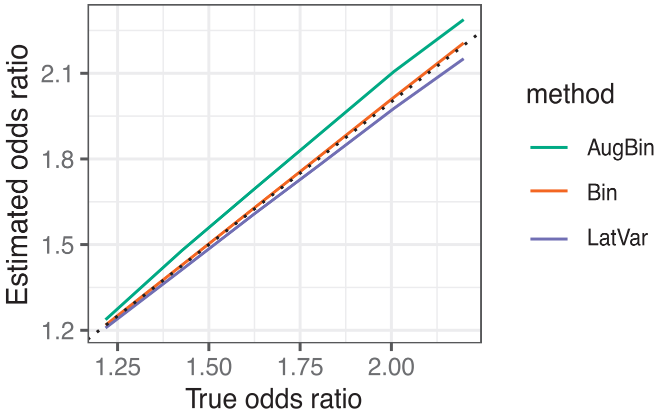

Figure 1. Bias reported from the latent variable method, augmented binary method and standard binary method when $n_{sim} = 5000$, total sample size $n = 300$ for true odds ratio between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components.

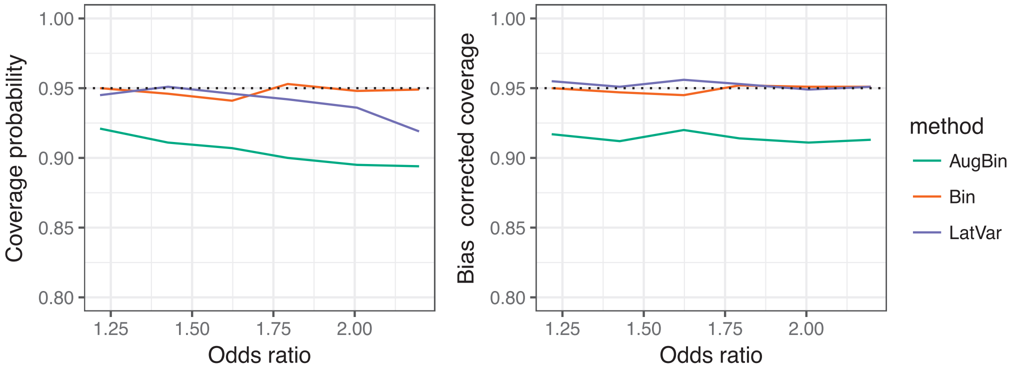

Figure 2. Coverage probability (left) and bias-corrected coverage probability (right) reported from the latent variable method, augmented binary method and standard binary method for $n_{sim} = 5000$, and total sample size $n = 300$ for true odds ratio between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, one binary and treatment effects are present in all four components.

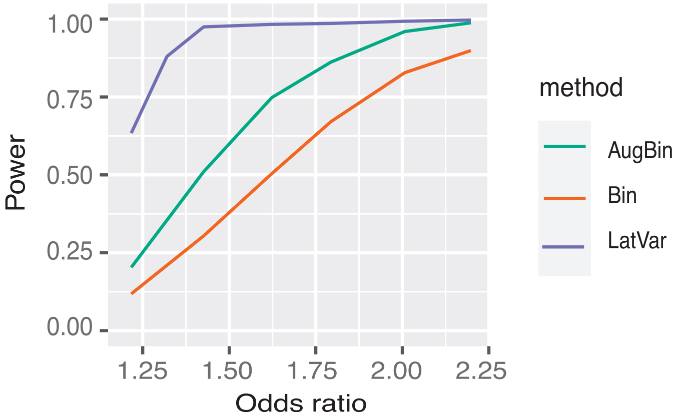

Figure 3. Statistical power reported from the latent variable method, augmented binary method and standard binary method for $n_{sim} = 5000$, and total sample size $n = 300$ for true odds ratio between 1.2 and 2.2. The composite endpoint of interest contains four components: two continuous, one ordinal, and one binary, and treatment effects are present in all four components.

augmented binary method will of course depend on the correlation structure employed. Due to the additional information in the other components, the latent variable method is still five times as precise as the other methods. The results are shown in Appendix E.

5.3.3 Components contributing to response

Figure 4 shows boxplots of the relative precision of the methods for four different response combinations, namely when response is driven by $(Y_1, Y_2, Y_3, Y_4), (Y_1, Y_2, Y_3), (Y_1, Y_4)$ and (Y_4), where Y_1 and Y_2 are observed as continuous variables, Y_3 is ordinal and Y_4 is binary.

When all four components contribute to response, the latent variable method outperforms the other methods, offering large precision gains. The variability in the magnitude of these gains is large, with the median result showing that the latent variable method reports the treatment effect eight times more precisely than the binary method and six times more precisely than the augmented binary method. If response is driven by (Y_1, Y_2, Y_3) then the relative median gains for the latent variable method are larger; however, note that in less than 2% of cases the treatment effect is reported equally or less precisely than from both of the other methods. The findings are similar when response is driven by (Y_1, Y_4), however the median gains are much smaller. The treatment effect is reported five times more precisely from the latent variable method than the binary method in this setting. Note that as the augmented binary method models the relevant components, it still performs well and again better than the latent variable method in a very small number of cases. When binary Y_4 determines response, the augmented binary method offers no improvement in precision whereas the latent variable method is approximately 1.5 times more precise.

5.3.4 Sensitivity analysis

The key assumptions in this model are joint normality of the four components and that discrete variables can be modelled as latent continuous variables. Although it is not possible to test these assumptions in real data, we can investigate how robust the latent variable method is to deviations from these conditions. We do this by drawing from the multivariate skew-normal distribution with different degrees of skew in each of the components. The first

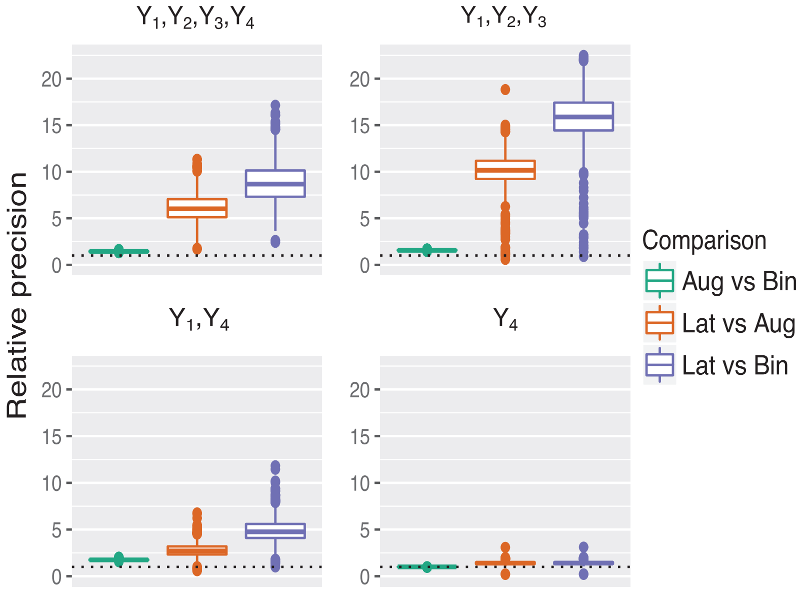

Figure 4. Estimated relative precision gains from augmented binary versus standard binary method, latent variable versus augmented binary method and latent variable versus standard binary method when different combinations of components are driving response. Response driven by (Y_1, Y_2, Y_3, Y_4), (Y_1, Y_2, Y_3), (Y_1, Y_4) and (Y_4) where Y_1 and Y_2 are continuous, Y_3 is ordinal, Y_4 is binary for $n_{sim} = 5000$ and total sample size $n = 300$. The composite endpoint of interest contains four components: two continuous, one ordinal, and one binary, and treatment effects are present in all four components.

scenario investigated considers when all four components are skewed. Scenarios 2–3 consider different magnitudes of skew in the latent continuous components only. Scenario 4 considers skew in the latent continuous components only for the case when there is no treatment effect. The results are shown in Appendix F of the supplemental material.

In summary, scenarios 1–3 have increased bias resulting in under-coverage as the bias-corrected coverage is close to nominal for all scenarios. The coverage of the latent variable method is nominal in the null case. This introduction of potentially large biases in the treatment effect estimate may be problematic for practice; however, it will result in a more conservative treatment effect estimate. The latent variable method still offers large power gains over the other methods and the MSE is the smallest for the latent variable method across all scenarios investigated. The latent variable method estimates the probability of response in the control arm well, however it underestimates the probability of response in the treatment arm. The magnitude of this underestimation is unaffected by the degree of skew or whether the skew is present in the observed continuous components. The relative precision of the methods are consistent with our previous findings indicating that the violation of joint normality only affects the bias and not the variance. The augmented binary and standard binary methods behave similarly to when the joint normality assumptions are satisfied, which is expected given that the assumptions of those models are potentially violated in both contexts.

6 Case study

6.1 Data structure

The real data underpinning this motivation comes from the MUSE study.[39] It was a Phase IIb, randomised, double-blind, placebo-controlled study investigating the efficacy and safety of anifrolumab in adults with moderate to severe SLE. Patients ($n = 305$) were randomised to receive anifrolumab (300 mg or 1000 mg) or placebo, in addition to standard therapy every four weeks for 48 weeks. The primary end point was the percentage of patients achieving an SRI response at week 24 with sustained reduction of oral corticosteroids (<10 mg/day and less than or equal to the dose at week 1 from week 12 through 24). Due to data sharing policy, we conduct the analysis for a subset of the patients, $n = 278$ rather than $n = 305$ reported in the paper, so the results will differ from the original paper. Furthermore, only the anifrolumab 300 mg arm ($n = 95$) and the placebo arm ($n = 87$) will be used to illustrate the methods.

The simulation results have suggested that the structure of the data is important for how the methods will perform; in particular, the magnitude of the precision gains depends highly on which components drive response. In the case of the MUSE study, the components responsible for driving response are the continuous SLEDAI measurement and the binary taper measure. The structure of the data is further explored in Appendix G of the supplemental material.

6.2 Results

The probability of response in the placebo arm is estimated as 0.199 by the latent variable method, 0.211 by the augmented binary method and 0.224 by the standard binary method. A much larger discrepancy between the methods is shown in the treatment arm, where the probability of response is estimated at 0.311, 0.324 and 0.382 in the latent variable, augmented binary and standard binary methods, respectively.

The log-odds treatment effect point estimates and confidence intervals for the MUSE trial are shown in Table 2. Both joint modelling methods estimate the treatment effect more precisely. Although there may be bias present in the point estimates for the joint modelling methods, the confidence intervals entirely overlap with that of the binary method. All three methods indicate that anifrolumab 300 mg performs better than placebo, as in the original findings. The latent variable model fits the data well according to the modified Pearson residuals (see Appendix G, supplemental material).

The simulation results indicated that the latent variable method may report the treatment effect with bias when the effect is large and when the assumption of joint normality is not satisfied. Although the observed continuous outcomes are normally distributed, we were unable to assess joint normality due to the discrete components. As the problems with performance are bias related, we suggest implementing a bootstrap procedure to correct for this. The concept is based on treating the observed sample as the population and sampling with replacement from this. An estimate of the bias is obtained using the difference between treatment effects in the assumed population

Table 2. Log-odds treatment effect estimates and 95% confidence intervals from the latent variable method, augmented binary method and standard binary method in the Phase IIb MUSE trial and the bootstrap sample when $n = 182$ and $n_{\text{boot}} = 1000$

	Log-odds treatment effect	
Method	MUSE trial estimate	Bootstrap estimate
Latent variable	0.641 (0.217, 1.072)	0.682 (0.275, 1.137)
Augmented binary	0.580 (0.139, 1.021)	0.608 (0.096, 1.111)
Binary	0.763 (0.078, 1.449)	0.809 (0.112, 1.561)

and the mean effect from the bootstrap samples.[40] A theoretical justification for this procedure specific to latent variable and structural equation modelling, along with examples, is provided in the literature.[41,42]

In this scenario $n = 182$ and $n_{\text{boot}} = 1000$, therefore the procedure is as follows:

1. Sample with replacement $n = 182$ patients from the MUSE trial.
2. Compute the treatment effect using the latent variable, augmented binary and standard binary methods.
3. Repeat steps 1 and 2 $n_{\text{boot}} = 1000$ times.
4. Estimate bias as the difference in MUSE trial effect and mean of the bootstrap treatment effects.

A 95% bootstrap confidence interval for the treatment effect estimate can be obtained by ordering the 1000 bootstrap estimates of the treatment effect and taking the 25th and 975th estimate. The point estimates and 95% confidence intervals from the MUSE trial and from the bootstrap re-sampling are shown in Table 2.

The log-odds point estimate from the latent variable method has been shifted away from the null by approximately 0.04. This is in agreement with the magnitude of bias suggested by the simulation results. The point estimate for the binary method has also been shifted substantially which is likely due to the large imprecision in the treatment effect reported by the binary method. The latent variable method reports the treatment effect 2.5 times more precisely than the standard binary method in this setting, whilst the augmented binary method is 2.4 times more precise. We would have expected the methods to perform similarly as the augmented binary method models the only components driving response. This increase in precision from the latent variable method compared with the binary method amounts to a 60% reduction in required sample size.

7 Discussion

In this paper we addressed the issue of substantial losses of information when modelling complex composite endpoints. By partitioning latent variable outcome spaces, we could model the observed structure of the composite endpoint, which resulted in large gains in efficiency. Sensitivity analyses showed that a bias is introduced when the assumptions of joint normality were not satisfied; however, similar reductions in variance were observed. When applying the methods to the MUSE trial, we implemented a bootstrap procedure to correct for the presumed bias, as joint normality could not be assessed. The treatment effect was reported 2.5 times more precisely than that reported from the standard binary method.

Bias correction appears to perform well in the real data, where the assumptions cannot be tested. The point estimate is shifted by a magnitude that would have been expected from the simulation results and the estimate of the variance is similar to that obtained in the single trial dataset. Furthermore, the bootstrap confidence interval for the treatment effect is contained within that for the binary method, which offers further reassurance for application. However, more work could be done to investigate different structures and scenarios to ensure that the bias correction is always performing as expected. Ideally, we would investigate this further across a large number of datasets however this is too computationally intensive. To perform this on one replicate where $n_{\text{boot}} = 1000$ currently takes 7 h using 200 cores on a high performance computer (HPC). An alternative in this setting is to model the multivariate dependence between components using copulas.[43] Proceeding in this way would allow us to relax the Gaussian assumption and instead join the multivariate distribution functions to their one-dimensional marginal distribution functions. Given that we have shown the latent variable model to be sensitive to these assumptions, exploring the application of copulas in the composite endpoint setting is an important area for future research.

The precision gains offered by the latent variable method offer justification for the additional complexity, however the magnitude of these gains are highly dependent on the components that drive response. In the baseline scenario where all components drive response, the latent variable method reported the effect 2.5 to 17.5 times more precisely than the standard binary method. However, in practice in SLE trials all four components have not been found to drive response. A review of two phase 3 trials ($n = 2262$) using the SRI-5 index found that the SRI-5 response rate at week 52 for all patients was 32.8%.[44] Non-response due to a lack of SLEDAI improvement, concomitant medication non-compliance or dropout was 31, 16.5 and 19.1%, respectively. Non-response due to deterioration in BILAG or Physician's Global Assessment after SLEDAI improvement, concomitant medication compliance and trial completion was 0.5%. This is in agreement with our findings from the MUSE trial data, which suggests that the precision gains in the baseline case are optimistic. The simulation results show that when one continuous and one binary components drive response, the latent variable method may be anywhere between 1 and 12 times as precise as the binary method and up to seven times as precise as the augmented binary method. In a very small number of cases (<2%), there are no efficiency gains from using the latent variable method in this scenario. However, the potential gains available in 98% of cases ensure that implementing the latent variable method is still very much a worthwhile endeavour for all stakeholders in a clinical trial.

In addition to SLE, we have identified other disease areas that have a similar complex composite structure, meaning the potential to improve efficiency extends well beyond SLE. However, it must be acknowledged that the exact structure of the endpoint may offer different magnitudes of bias and precision, and may require longer computational time. Furthermore, in conditions where longitudinal data is required to sufficiently capture disease activity, trials may include multiple follow-up times and the method will need to be extended to include latent variables in the mean structure to account for this. In terms of scalability to more complex endpoints, the computational time depends on many things, in particular the number of outcomes, the outcome scale and the number of levels in the ordinal variable. In our case, we find the number of ordinal levels to be the most influential factor. This is due to the fact that five levels in the ordinal variable leads to 10 probability calculations in equation (6); however three levels would require the computation of six joint probabilities. Consequently, the run time will be substantially increased if there are multiple ordinal outcomes and decreased if the discrete variables are binary. If the computational time for a particular endpoint is deemed to be too large, then we may reduce the complexity of the endpoint by collapsing the least informative components into a single binary variable. It must be acknowledged that as we have coded the method, the likelihood and probability of response code will have to be tailored specifically to each endpoint. The potential gains in efficiency justify this additional complexity.

Obtaining maximum likelihood estimates from latent variable models has been achieved in different ways throughout the literature. In this paper we use a quasi-Newton algorithm, however these and Newton type algorithms are not without their limitations, such as tending to be slow or intractable in higher dimensions.[14] The EM algorithm has been proposed in this setting as it lends itself well to situations with unknown parameters such as the τ-thresholds, however, conditioning on these parameters as in equation (4) violates regularity conditions. Hence a Parameter-Expanded EM algorithm which transforms the latent variables and expands the parameter space may be more appropriate.[45] For an implementation of this estimation method when identifying genetic factors for comorbid conditions, we refer the reader to Zhang et al.[27] Implementing the method as we have done in this paper is computationally demanding however, we would not expect the Parameter Expanded EM algorithm to rectify this and may actually lead to increased computational time. More work is required to compare estimation methods for these latent variable models.

We have shown that the latent variable method is a powerful tool in composite endpoint analysis and should be considered as a primary analysis method in a trial using these endpoints. In order for implementation in the general case, where the composite contains any number of continuous and discrete outcomes, we have developed a web based Shiny application, as detailed below. Furthermore, in order for patients and investigators to benefit from the efficiency gains, our current work is focused on developing a method for sample size calculation using these models, along with software to implement this.[46] Our future work involves extending the method to include count and time-to-event endpoints for more general application.

Software

A Shiny application for implementing the method is available at https://martinamcm.shinyapps.io/augbin/. Documentation and example data are available at https://github.com/martinamcm/AugBin.

Acknowledgements

The authors would like to thank the two anonymous reviewers whose comments helped to improve the quality of this article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: MMM and JW are supported by funding from the Medical Research Council (MRC), grant code MC_UU_00002/6. The funding body did not have any role in the design of the study, analysis, interpretation of data or writing the manuscript.

Supplemental material

Supplemental material is available online.

ORCID iDs

Martina McMenamin https://orcid.org/0000-0001-7784-2271
Jessica K Barrett https://orcid.org/0000-0003-1889-9803
James MS Wason https://orcid.org/0000-0002-4691-126X

References

1. Wason J and Seaman SR. Using continuous data on tumour measurements to improve inference in phase II cancer studies. *Stat Med* 2013; **32**: 4639–4650.
2. Whittaker J. *Graphical models in applied multivariate statistics*. Hoboken, NJ: Wiley and Sons, 1990.
3. Lauritzen S and Wermuth N. Graphical models for association between variables, some of which are qualitative and some quantitative. *Ann Stat* 1989; **17**: 31–54.
4. Wason J and Jenkins M. Improving the power of clinical trials of rheumatoid arthritis by using data on continuous scales when analysing response rates: an application of the augmented binary method. *Rheumatology* 2016; **55**: 1796–1802.
5. Lin CJ and Wason JM. Improving phase II oncology trials using best observed recist response as an endpoint by modelling continuous tumour measurements. *Stat Med* 2017; **36**: 4616–4626.
6. McMenamin M, Berglind A and Wason J. Improving the analysis of composite endpoints in rare disease trials. *Orphanet J Rare Dis* 2018; **13**: 81.
7. Verbeke G, Fieuws S and Molenberghs G. The analysis of multivariate longitudinal data: a review. *Stat Meth Med Res* 2014; **23**: 42–59.
8. De Leon A and Carriere K (eds.) *Analysis of mixed data methods and applications*. London, UK: Chapman and Hall/CRC, 2013.
9. Ashford J and Sowden R. Multivariate probit analysis. *Biometrics* 1970; **26**: 535–46.
10. Chib S and Greenberg E. Analysis of multivariate probit models. *Biometrika* 1998; **85**: 347–361.
11. Pearson K. Mathematical contributions to the theory of evolution. XII. on a generalised theory of alternative inheritance, with special reference to mendel's laws. *Philos Trans R Soc London, Ser A* 1904; **203**: 53–86.
12. Poon W and Lee S. Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika* 1987; **52**: 409–430.
13. De Leon A. Pairwise likelihood approach to grouped continuous model and its extension. *Stat Probabil Lett* 2005; **75**: 49–57.
14. Gueorguieva R and Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. *J Am Stat Assoc* 2001; **96**: 1102–1112.
15. Boscardin WJ, Zhang X and Belin TR. Modeling a mixture of ordinal and continuous repeated measures. *J Stat Comput Simul* 2008; **78**: 873–886.
16. De Leon AR and Carriégre KC. General mixed-data model: Extension of general location and grouped continuous models. *Can J Stat* 2007; **35**: 533–548.
17. Tate R. The theory of correlation between two continuous variables when one is dichotomised. *Biometrika* 1955; **42**: 205–216.
18. Cox D and Wermuth N. Response models for mixed binary and quantitative variables. *Biometrika* 1992; **79**: 441–61.
19. Catalano P and Ryan L. Bivariate latent variable models for clustered discrete and continuous outcomes. *J Am Stat Assoc* 1992; **87**: 651–658.
20. Sozu T, Sugimoto T and Hamisaki T. Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometrical J* 2012; **54**: 716–729.

21. Wu B and De Leon A. Review of 'sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables'. *Biometrical J* 2013; **55**: 807–812.
22. Catalano P. Bivariate modelling of clustered continuous and ordered categorical outcomes. *Stat Med* 1997; **16**: 883–900.
23. Samani E and Ganjali M. A multivariate latent variable model for mixed continuous and ordinal responses. *World Appl Sci J* 2008; **3**: 294–299.
24. Arminger G and Küsters U. Latent Trait Models with Indicators of Mixed Measurement Level. In: Langeheine R and Rost J (eds) *Latent Trait and Latent Class Models*. Springer, Boston, MA, 1988, pp. 51–73.
25. Regan M and Catalano P. Regression models and risk estimation for mixed discrete and continuous outcomes in developmental toxicology. *Risk Anal* 2000; **20**: 363–376.
26. Faes C, Geys H, Aerts M, et al. Modelling combined continuous and ordinal outcomes from developmental toxicity studies. In: *Proceedings of the 17th international workshop on statistical modelling*, Chania, Crete, 2002.
27. Zhang H, Liu D, Zhao J, et al. Modeling hybrid traits for comorbidity and genetic studies of alcohol and nicotine co-dependence. *Ann Appl Stat* 2018; **12**: 2359–2378.
28. Gueorguieva R and Sanacora G. A latent variable model for joint analysis of repeatedly measured ordinal and continuous outcomes. In: Verbeke G, Molenberghs G, Aerts M, et al. (eds) In: *Proceedings of the 18th international workshop on statistical modelling*. Leuven: Katholieke Universiteit Leuven, pp.171–176.
29. Gueorguieva R and Sanacora G. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Stat Med* 2006; **25**: 1307–1322.
30. Renard D, Geys H, Molenberghs G, et al. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical J* 2002; **44**: 921–935.
31. Luijten K, Tekstra J, Bijlsma J, et al. The systemic lupus erythematosus responder index (sri); a new sle disease activity assessment. *Autoimmun Rev* 2012; **11**: 326–329.
32. Symmons DPM, Coppock JS, Bacon P, et al. Development and assessment of a computerized index of clinical disease activity in systemic lupus erythematosus. *QJM: Int J Med* 1988; **69**: 927–937.
33. Genz A. Numerical computation of multivariate normal probabilities. *J Comput Graph Stat* 1992; **1**: 141–150.
34. Rosseel Y. lavaan: An r package for structural equation modeling. *J Stat Software* 2012; **48**: 1–36.
35. Richardson LF. The approximate arithmetical solution by finite differences of physical problems including differential equations, with an application to the stresses in a masonry dam. *Philos Trans R Soc London, Ser A* 1911; **210**: 307–357.
36. Fan J, Liao Y and Liu H. An overview of the estimation of large covariance and precision matrices. *Econometrics J* 2016; **19**: C1–C32.
37. Higham NJ. Computing the nearest correlation matrix-a problem from finance. *IMA J Numer Anal* 2002; **22**: 329–343.
38. Morris T, White I and Crowther M. Using simulation studies to evaluate statistical methods. *Stat Med* 2019; **38**: 2074–2102.
39. Furie R, Khamashta M, Merrill J, et al. Anifrolumab, an anti interferon alpha receptor monoclonal antibody, in moderate-to-severe systemic lupus erythematosus. *Arthritis Rheumatol* 2017; **69**: 376–386.
40. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat* 1979; **7**: 1–26.
41. Beaujean AA. *Bootstrapping latent variable models appendix to latent variable modeling using R: a step-by-step guide*. New York, NY: Routledge, 2014.
42. SG West PC JF Finch. *Structural equation models with non-normal variables: problems and remedies*. Thousand Oaks, CA: Sage, pp.56–75.
43. Joe H. *Multivariate models and dependence concepts*. London: Chapman and Hall, 1997.
44. Kalunian KC, Urowitz MB, Isenberg D, et al. Clinical trial parameters that influence outcomes in lupus trials that use the systemic lupus erythematosus responder index. *Rheumatology* 2018; **57**: 125–133.
45. Ruud PA. Extensions of estimation methods using the em algorithm. *J Econom* 1991; **49**: 305–341.
46. McMenamin M, Barrett JK, Berglind A, et al. Sample Size Estimation using a Latent Variable Model for Mixed Outcome Co-Primary, Multiple Primary and Composite Endpoints. *arXiv* 2020; arXiv:1912.05258

Appendix

Notation

$\mathbf{Y_i} = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})^T$	vector of observed outcomes for patient $i \in N$
$\mathbf{Y} = (\mathbf{Y}_1, \ldots \mathbf{Y_N})^T$	observed outcomes for all patients
Y_{i1} and Y_{i2}	observed continuous SLEDAI and PGA measures
Y_{i3}	observed ordinal BILAG measure
Y_{i4}	observed binary oral corticosteroid tapering measure

Y_{i3}^*	Latent continuous BILAG measure
Y_{i4}^*	Latent continuous oral corticosteroid tapering measure
$\mathbf{Y}_\mathbf{i}^* = (Y_{i1}, Y_{i2}, Y_{i3}^*, Y_{i4}^*)^T$	vector of observed and latent continuous measures for patient i
$\mathbf{Y}^* = (\mathbf{Y}_1^*, \ldots \mathbf{Y}_\mathbf{N}^*)^T$	Vector of observed and latent continuous measures for all patients
T_i	treatment indicator for patient i
y_{i10} and y_{i20}	baseline measures for Y_{i1} and Y_{i2}, respectively