STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:3363–3385
Published online 3 September 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3721

Estimating drug effects in the presence of placebo response: Causal inference using growth mixture modeling

Bengt Muthén[1,*,†] and Hendricks C. Brown[2]

[1]*University of California, Los Angeles, CA, U.S.A.*
[2]*University of Miami, Miami, FL, U.S.A.*

SUMMARY

Placebo-controlled randomized trials for antidepressants and other drugs often show a response for a sizeable percentage of the subjects in the placebo group. Potential placebo responders can be assumed to exist also in the drug treatment group, making it difficult to assess the drug effect. A key drug research focus should be to estimate the percentage of individuals among those who responded to the drug who would not have responded to the placebo ('Drug Only Responders'). This paper investigates a finite mixture model approach to uncover percentages of up to four potential mixture components: Never Responders, Drug Only Responders, Placebo Only Responders, and Always Responders. Two examples are used to illustrate the modeling, a 12-week antidepressant trial with a continuous outcome (Hamilton D score) and a 7-week schizophrenia trial with a binary outcome (illness level). The approach is formulated in causal modeling terms using potential outcomes and principal stratification. Growth mixture modeling (GMM) with maximum-likelihood estimation is used to uncover the different mixture components. The results point to the limitations of the conventional approach of comparing marginal response rates for drug and placebo groups. It is useful to augment such reporting with the GMM-estimated prevalences for the four classes of subjects and the Drug Only Responder drug effect estimate. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: principal stratification; latent classes; trajectory types; potential outcomes

1. INTRODUCTION

This paper discusses the assessment of drug effects as evaluated in placebo-controlled randomized trials. To ground the discussion in specific substantive settings, two data sets will be first described.

*Correspondence to: Bengt Muthén, University of California, Los Angeles, CA, U.S.A.
†E-mail: bmuthen@ucla.edu

Contract/grant sponsor: NIAAA; contract/grant number: 1 R21 AA10948-01A1
Contract/grant sponsor: NIMH; contract/grant number: MH40859
Contract/grant sponsor: NIDA; contract/grant number: P30 MH066247

Received 21 March 2009
Accepted 29 July 2009

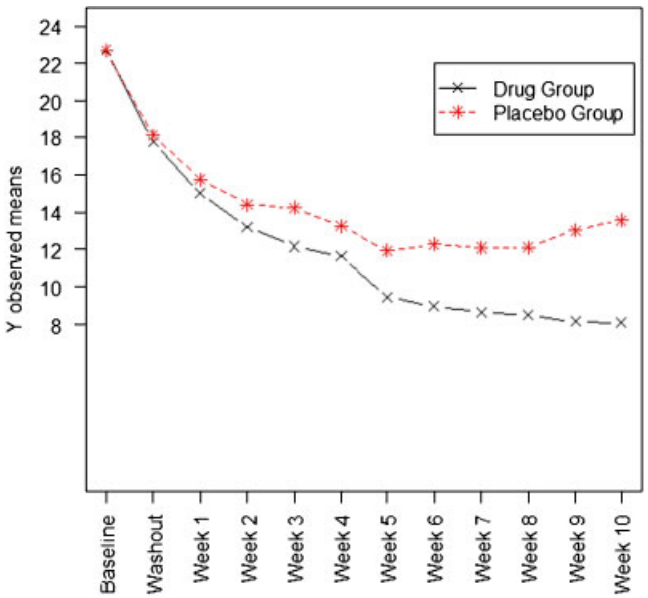

Figure 1. Antidepressant trial.

1.1. Antidepressant trial

Data for this example come from a clinical trial with 154 subjects diagnosed with major depression and randomly assigned to receive fluoxetine, imipramine, or placebo for a 10-week clinical trial [1]. Subjects met criteria for the atypical sub-type of depression. The depression score considered here is the 28-item Hamilton Depression Rating Scale, which includes items suitable for atypical depression not included in the 17-item scale. Depression ratings were also made at the beginning and end of a 1-week, single-blind placebo washout period. Among the 154 subjects, 49 received fluoxetine, 53 received imipramine, and 52 received placebo. The fluoxetine and imipramine were found to be similar in the previous analyses [1] and are combined into a single drug group in the current analyses. The depression outcome mean for the 12 time points is plotted in Figure 1 for the placebo and drug groups. On average, the placebo group shows some improvement in depressive symptoms while the drug group improvement is larger, especially after week 5. The outcome mean at the endpoint is 13.5 (SD = 8.6) for the placebo group and 8.1 (SD = 7.1) for the drug group, a significant overall reduction in depressive symptoms for those who are on active medication. A common way to designate response is a drop of at least 50 per cent in the depression score between the baseline and the end point [2], in this case at week 10. Using this rule 38 per cent of the placebo group and 65 per cent of the drug group are responders (in case of missing data at week 10, the last available observation is used).[‡]

[‡][1] Did not use the 50 per cent approach but instead a Clinical Global Impression (CGI) rating of 'very much improved' or 'much improved' to estimate a 23 per cent response rate for the placebo group and a 52 per cent response rate for the drug group, but the additional CGI outcome will not be used here for simplicity.

DOI: 10.1002/sim

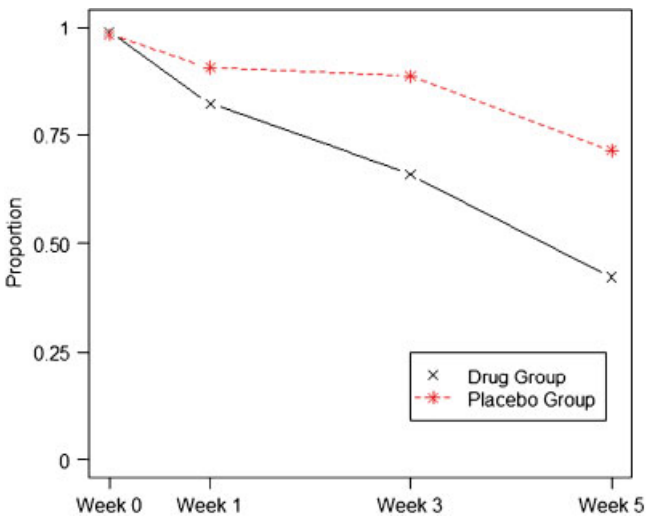

Figure 2. Schizophrenia trial.

1.2. Schizophrenia trial

Data for this example come from an NIMH schizophrenia collaborative study where severity of illness was measured for 437 schizophrenics randomized to one of four treatments: placebo, chlorpromazine, fluphenazine, or thioridazine. As in [3], the drug groups are combined into one group resulting in 329 subjects in the drug group and 108 in the placebo group. Subjects were measured once before treatment and three more times during five subsequent weeks. Following [3], the analysis considers the item 'Severity of Illness' (IMPS #79), originally scored as Normal, Borderline mentally ill, Mildly ill, Moderately ill, Markedly ill, Severely ill, and Among the most extremely ill, but dichotomized between the categories Mildly ill and Moderately ill. The proportion of subjects having more severe illness is shown in Figure 2 for the placebo and drug groups. The proportion of subjects in the more ill category at the end point is 0.71 for the placebo group and 0.42 for the drug group. The decline for the placebo group suggests a natural decline in severity or a placebo effect.

1.3. Alternative approaches

The conventional estimates of drug efficacy consist of the difference between the drug and placebo response rates and, in case of the antidepressant example, the mean difference between the drug and placebo groups of the depression score at the endpoint of the study. This approach has two shortcomings. First, the drug effect attains a causal interpretation as a treatment assignment effect, not an effect of the drug *per se*. Second, the assessment does not use all of the longitudinal data.

1.3.1. Causal modeling. As seen in the two examples, placebo-controlled randomized trials for antidepressants and other drugs often show a response for a sizeable percentage of the subjects in the placebo group. Due to randomization, potential placebo responders can be assumed to exist also in the drug group. This makes it difficult to assess the true drug effect in the sense that a

DOI: 10.1002/sim

large percentage of the responders in the drug group may have responded also under placebo. In statistical terms, the mean difference at the study endpoint is an 'intent-to-treat' estimate. This is a causal effect of the treatment assignment, but not necessarily a causal effect of the drug. That is, for some subjects their response to the drug is likely to be very similar to their response under placebo, whereas others may be presumed to do much differently on the two conditions. Here, a causal effect estimate is understood in the sense of the Rubin causal model [4–7]. A causal effect approach leads to the estimation of mean differences between drug and placebo groups within 'principal strata' consisting of homogeneous groups of individuals [8]. For example, to evaluate the causal effect of the drug, one important principal stratum consists of individuals, who would respond to the drug but would not respond to the placebo. The group membership is not observed but latent, which leads to finite mixture modeling with latent classes (see, e.g. [9]). This mixture modeling approach will be pursued here.

1.3.2. End points versus trajectories. End-point analysis has the disadvantage of drawing on information from a single time point. For the antidepressant trial, the outcome at week 10 may contain irrelevant time-specific sources of depression variation such as day-to-day fluctuations. For example, the average outcome at weeks 8, 9, and 10 may capture a longer-term level at the end of the trial that better represents the true end-point depression level. Also, in the schizophrenia trial the binary outcome allows a drop only from the category of 0 to the category of 1. As an alternative to end-point analysis, the trajectory shape over time for the continuous outcome or the modeling of the probability of illness for a dichotomous outcome can be studied and may be estimated by a random effects repeated measures growth model that draws on the information from all time points [3, 10]. The idea of considering trajectory shape in research on depression medication has also been proposed in the psychiatric literature by [11], although not using a formal statistical growth model. A generalized, finite mixture, version of such growth modeling [12, 13] will be applied in a causal modeling context.

1.4. Paper outline

The outline of the paper is as follows. In Section 2, the causal model of [4] will be briefly reviewed as a background for the modeling. Section 3 discusses several versions of causal models for drug trials. Section 4 describes growth mixture modeling (GMM) and applies it to the longitudinal data from the antidepressant and schizophrenia trials. Section 5 discusses a Monte Carlo simulation to study the quality of estimation. Section 6 concludes.

2. THE AIR MODEL

To define causal effects in a drug trial setting, it is instructive to first consider the Angrist, Imbens and Rubin ([4]; AIR from now on) discussion of causal inference in potential outcomes terms applied to randomized trials with non-compliance. In AIR notation, Z is a binary 0/1 treatment/control variable, D is a binary 0/1 variable indicating that the subject does not versus does take up the treatment, and Y is the outcome. AIR considered $Y_i(Z_i, D_i(Z_i))$ for individual i and defined the causal effect of Z on Y for individual i as the counterfactual expression $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$. Although this quantity cannot be observed, the average causal effect can be identified

DOI: 10.1002/sim

and estimated under suitable conditions. AIR considered four classes of subjects:

- Never takers ($D_i(1)=0, D_i(0)=0$): subjects who would not take up treatment if randomized to either treatment or control (causal effect $=0$ under the exclusion restriction)
- Compliers ($D_i(1)=1, D_i(0)=0$): subjects who would take up treatment if randomized to treatment and otherwise not (causal effect $=Y_i(1,1)-Y_i(0,0)$)
- Defiers ($D_i(1)=0, D_i(0)=1$): subjects who would do the opposite of their treatment assignment (causal effect $=-(Y_i(1,1)-Y_i(0,0))$)
- Always takers ($D_i(1)=1, D_i(0)=1$): subjects who would take up treatment if randomized to either treatment or control (causal effect $=0$ under exclusion restriction)

Of particular interest is the average causal effect in the Complier class, the complier-average causal effect ('CACE'). To extract this effect, AIR considered two key assumptions. Under the monotonicity assumption the probability of Defier class membership is taken as zero (this was an unlikely case in their application). Under the exclusion restriction the causal effect is taken as zero for Never takers as well as for Always takers (i.e. no effect of randomization). In some applications, the probability of the Always taker class membership is also zero due to the design of the study, where subjects do not have access to treatment unless invited.

The outcome mean difference for the drug and placebo groups, $\mu_1-\mu_0$, is referred to as the 'intention-to-treat' (ITT) effect. The ITT effect is thereby a weighted mean difference over the four classes, ignoring that subjects react differently to the treatment invitation.

3. A 4-CLASS DRUG TRIAL MODEL

This paper considers four latent classes analogous to those of the AIR model: Never Responders, Drug Only Responders, Placebo Only Responders, and Always Responders. These may be seen as principal strata in the sense of [8], providing homogeneous groups within which more meaningful comparison between the drug and placebo groups can be made. Principal stratum membership is not influenced by treatment and 'can be used as any pre-treatment covariate, such as age category' ([8], p. 21). For applications of principal stratification see, for example [14–16].

The four latent classes, their probabilities, and their outcome means for placebo (0) and drug (1) groups, are summarized in Table I. A number of different GMM expressed as in Table I will be considered by placing no or some restrictions on the probabilities and equalities between some of the means.

Table I. Four types of subjects, probabilities, and means for placebo (0) and drug (1) groups.

Placebo	Drug group		
Group	Non-responder	Responder	
Non-responder	Never responder π_n, μ_{n0}, μ_{n1}	Drug only responder π_d, μ_{d0}, μ_{d1}	Non-responder
Responder	Placebo only responder π_p, μ_{p0}, μ_{p1}	Always responder π_a, μ_{a0}, μ_{a1}	Responder
	Non-responder	Responder	

DOI: 10.1002/sim

The prevalences of the four types of subjects in Table I are of clinical interest. For Never Responder subjects, neither placebo nor drug is effective and the subjects may be switched to another drug. For Drug Only Responder subjects, only drug is effective. In the antidepressant trial one would expect $\mu_{d1} < \mu_{d0}$, for example. This case is of particular interest from a pharmaceutical point of view because this may be viewed as the subjects experiencing a real drug effect. For Placebo Only Responder subjects, only placebo is effective. This is presumably a smaller group, who may have adverse effects to drug. For Always Responder subjects, both placebo and drug are effective. These subjects may not need the drug, although they may benefit more from the drug than placebo.

The marginals of Table I give the subjects who respond to placebo and drug, respectively, where a higher marginal response rate for drug as compared with placebo is typically taken to indicate a beneficial drug effect. The drug responders considered in the marginal distribution, however, consist of a mixture of two different sub-populations: those who respond only to the drug and those who would have responded to the placebo as well. Similarly, the placebo responders consist of a mixture of those who respond only to placebo and those who respond to both placebo and drug. The sizes of the sub-populations represented by the cells of Table I are not observed, but have to be estimated using a mixture model.

It is assumed that the outcome at pre-randomization time points is drawn from the four different sub-populations, representing different subject types already before the start of the treatment, where the type-specific parameters for the outcome at the pre-randomization time points are unaffected by subsequent treatment. For each of the four types of subjects, post-randomization development is assumed to be different depending on whether the subject receives placebo or drug. In this way, there are potentially eight different average post-randomization trajectories. For example, for Drug Only Responder subjects, the response is not realized if the subject is randomized to placebo, only if randomized to drug. Similarly, for Placebo Only Responder subjects, the response is not realized if the subject is randomized to drug. For Always Responder subjects the degree of response may not be the same for subjects randomized to placebo as for subjects randomized to drug.

Finally, the actual degree of response or non-response need not be the same for the four types of subjects. For example, for subjects randomized to the drug group, the degree of response may be different for Drug Only Responder subjects and Always Responder subjects. For subjects randomized to placebo, the degree of response may be different for Placebo Only Responder subjects and Always Responder subjects.

The model of Table I will be referred to as the 4-class, 8-mean model. The placebo and drug group means are obtained as a finite mixture over the four latent classes,

$$\mu_0 = \pi_n \mu_{n0} + \pi_d \mu_{d0} + \pi_p \mu_{p0} + \pi_a \mu_{a0} \tag{1}$$

$$\mu_1 = \pi_n \mu_{n1} + \pi_d \mu_{d1} + \pi_p \mu_{p1} + \pi_a \mu_{a1} \tag{2}$$

The model of Table I can be seen as using a single latent class variable with four categories, or two cross-classified latent class variables each having two categories (Responders and Non-Responders). The latter approach is useful when considering covariates that may predict response, in that such covariates may be different for placebo response and drug response.

The antidepressant trial data can be used to illustrate the estimation of this model. Here, the endpoint at week 10 is used. A finite mixture model with four latent classes (four mixture components) is used under the assumption of normality for the outcome within each latent class (see, e.g. [17]). Only the means are varying across classes, with class-invariant variances. The

DOI: 10.1002/sim

Table II. Summary of antidepressant analyses.

Model	LL	#parameters	BIC	Never per cent	Drug per cent	Placebo per cent	Always per cent	Tx
Week 10 analysis								
1. ITT (1c), effect=5	−510	4	1041					
2. 4c, 8m	−481	13	1028	31	28	0	41	11
3. 50 per cent rule, 3c, 4m (AIR)				35	27	0	38	13
4. 3c, 4m (AIR)	−487	8	1014	40	8	0	18	2
5. 3c, 2m	−491	6	1012	27	28	0	45	15
GMM analysis								
6. ITT (1c), effect=6	−4708	33	9583					
7. 3c, 2m	−4688	37	9562	26	35	0	39	14
8. 3c, 4m (AIR)	−4676	43	9570	4	38	0	58	4
9. 4c, 8m	−4652	58	9597	28	26	4	42	18

N, never responder; D, drug only responder, P, placebo only responder; A, always responder; Tx, treatment effect for drug only responders.

maximum-likelihood loglikelihood (LL) value, # parameters, Bayesian information criterion (BIC) value, and class percentages are summarized in Table II under model 2. The ITT model is shown as a comparison as model 1, having two means, one for the placebo group and one for the drug group. The BIC is better for the hypothesized 4-class model than the 1-class ITT model. The entropy is 0.69. In these data, the 4-class analysis produces two latent classes corresponding to Never Responder (high mean for both the drug and the placebo groups) at the expense of the Placebo Only Responder class (low mean for placebo group and high mean for drug group) which is not represented. It is interesting to note that the 28 per cent Drug Only Responder percentage is considerably lower than the marginal 69 per cent drug response (the marginal placebo response percentage is the same as the Always Responder percentage, 41 per cent). In other words, according to the model most of the subjects responding to the drug would have also responded to placebo. The Drug Only Responder effect corresponds to approximately 1 1/2 SD. The ITT effect of 5 is less than half of that.

The above analysis illustrates that there is no guarantee that the four classes will have the anticipated interpretation of Never Responder, Drug Only Responder, Placebo Responder, and Always Responder. For example, a non-responder class may be replaced by an added responder class, resulting in two classes with different degrees of response. Although parameter constraints may be added to guarantee the anticipated interpretation, if the data do not support this, estimates on the boundaries of the constraints will be produced and the solution will fit the data less well. Alternatively, more classes may be added. As described in Section 4.1, however, the growth mixture analysis, using all available time points, recovers the four hypothesized classes.

3.1. The 3-class, 4-mean AIR model

The monotonicity and exclusion restriction assumptions of the AIR model do not have direct counterparts in a drug trial but the terms will be kept here for simplicity. The analogous assumptions suggest ways that the 4-class, 8-mean model can be simplified, producing models that are parsimonious, and are possibly more easily replicable in samples of limited sizes. In particular,

DOI: 10.1002/sim

AIR used a 3-class, 4-mean model which can be identified in terms of first-order moments. This model will now be discussed.

Under the monotonicity assumption there are no Placebo Only Responders ($\pi_p = 0$), resulting in a 3-class model. In the drug trial context, the monotonicity assumption implies that subjects who do not respond to the drug would also not respond to placebo. In this context a placebo effect may be viewed as a result of the attention and empathy of the nursing staff. The assumption may be reasonable from a substantive point of view if a subject is not a placebo responder and an ineffective drug can be seen as a placebo. A violation of this assumption would occur if a subject has an adverse effect to the drug and therefore does not benefit, whereas he/she would have benefitted from placebo. It is possible that this class has a relatively low prevalence.

Under the exclusion restriction $\mu_{n0} = \mu_{n1}$, $\mu_{a0} = \mu_{a1}$ in the notation of Table I. These two assumptions state that subjects in the Never Responder class have the same outcome mean irrespective of being randomized to placebo or drug, and subjects in the Always Responder class have the same outcome mean irrespective of being randomized to placebo or drug. A violation of this assumption would occur if, for example, among subjects in the Always Responder class the drug response is stronger than the placebo response.

Applying both the monotonicity and the exclusion restrictions results in a 3-class, 4-mean model. It follows from (1) and (2) that the difference between the drug and placebo means can be expressed as

$$\mu_1 - \mu_0 = \pi_d(\mu_{d1} - \mu_{d0}) \tag{3}$$

identifying the average causal effect of the drug as

$$\mu_{d1} - \mu_{d0} = (\mu_1 - \mu_0)/\pi_d \tag{4}$$

Due to the monotonicity assumption π_d can be estimated because the proportion of subjects who respond in the placebo group gives the proportion of Always Responders among those who respond in the drug group. These proportions are observable quantities if one defines response, for example, as a depression score drop of at least 50 per cent. This identifies the parameters of the 3-class, 4-mean model. Identification of the general 4-class, 8-mean model can be shown in line with [18, chapter 3]. The parameter recovery of the corresponding 4-class GMM is demonstrated in the Monte Carlo simulation study of Section 5.

A moment-estimator is suggested by (4) using sample proportions and means. As stated earlier, for the antidepressant trial data the proportion of responders in the placebo and drug groups are estimated as 38 per cent and 65 per cent, respectively, using the criterion of an end-point drop of at least 50 per cent. It should be emphasized that this is not an endorsement of using the 50 per cent rule, but the approach is simply used here as a contrast to finite mixture modeling. With the assumption of zero probability of Placebo Only Responders, this means that there are 38 per cent Always Responders. This gives a π_d estimate for the prevalence of the Drug Only Responder class of 27 per cent (65 per cent–38 per cent). This implies that for the drug group, among the 65 per cent responders ($n = 66$), only 42 per cent ($n = 28$) are Drug Only Responders, whereas the remaining 58 per cent ($n = 38$) of the drug responders would have responded also under placebo. It follows that the Never Responder prevalence is estimated as 35 per cent. With the endpoint outcome sample means of 9.3 and 12.9, (4) gives the average causal effect estimate of the Hamilton D28 improvement as $13((12.9 - 9.3)/0.27)$ for Drug Only Responders.

The above moment-based estimates of the prevalences and Drug Only Responder outcome mean difference are summarized in Table II as model 3, having 3 classes and 4 means. Model 3 can

DOI: 10.1002/sim

be compared with the maximum-likelihood estimated finite mixture modeling reported earlier for the 4-class, 8-mean model 2. The estimates from the two approaches are similar. In model 2, the 50 per cent rule is replaced by the finite mixture model and its assumptions of within-class normality and class-invariant variances. Table II also shows model 4 which is the counterpart to the moment-estimated 3-class, 4-mean AIR model 3, but using a maximum-likelihood-estimated finite mixture model. The entropy is 0.83. The Drug Only Responder percentage now obtains a low value of only 8 per cent with a Drug Only Responder treatment effect of only 2. Compared with the model 2 and model 3 results, the maximum-likelihood results for model 4 do not seem to give a plausible representation of the data.

3.2. A 3-class, 2-mean model

A more parsimonious version of the 3-class, 4-mean AIR model under monotonicity and exclusion restrictions can be formulated. It uses only one mean for responders and one mean for non responders,

$$\mu_{n0} = \mu_{d0} \tag{5}$$

$$\mu_{d1} = \mu_{a1} \tag{6}$$

Given that the exclusion restriction assumes $\mu_{n0} = \mu_{n1}$ and $\mu_{a0} = \mu_{a1}$, this results in only 2 means with the assumptions that:

1. A non-responder mean is the same if
   (a) the person is in the placebo group and in the Never Responder class,
   (b) the person is in the placebo group and in the Drug Only Responder class,
   (c) the person is in the drug group and in the Never Responder class.
2. A responder mean is the same if
   (a) the person is in the drug group and in the Drug Only Responder class,
   (b) the person is in the drug group and in the Always Responder class,
   (c) the person is in the placebo group and in the Always Responder class.

Because of variations in degree of non-response and degree of response in the above settings, there is no substantive reason to believe that the above assumptions are exactly true. For example, a non-responder mean may be higher in the placebo group for the Never Responder class than for the Drug Only Responder class because the Never Responder class subjects may be harder to cure. Or, a responder mean in the drug group may be lower in the Always Responder class than in the Drug Only Responder class because of an additive effect of the drug and the placebo. The assumptions may, however, approximate reality in a drug trial to a sufficient degree.

The 3-class, 2-mean model has the advantage of simplicity and can capture essential features in the data. Two of the latent classes assume the same post-randomization means for placebo and drug groups, whereas the third class allows these means to be different for placebo and drug. Latent classes characterized in this way are likely to be found in many trials because some subjects respond and do not respond in both placebo and drug groups, and some subjects respond only to drug.

The maximum-likelihood estimation results for the 3-class, 2-mean model are summarized in Table II as model 5. It can be seen that model 5 has the best BIC among the models for Week 10

DOI: 10.1002/sim

analysis. The entropy is 0.64. Compared with model 4, the more parsimonious model 5 is more in line with the results of model 2 and model 3.

4. GMM OF DRUG TRIALS

GMM has the potential of uncovering important information about classes of responders and non-responders in clinical trials extending the above models to longitudinal settings where not only the end point outcome is considered but also the trajectory throughout the trial. GMM combines random effects modeling in conventional repeated measures analysis with finite mixture modeling using latent class variables to represent qualitatively different classes of trajectories [12, 13, 19]. GMM is currently used in a wide variety of settings, see, for example [20] for an application to the joint study of PSA development and prostate cancer survival, [21] for an application to identifying trajectories of positive affect and negative events following myocardial infarction, and [22] for an application to growth modeling with non-ignorable dropout in a depression trial.

The 4-class model discussed above will now be presented in GMM terms. Let p_i and d_i be the binary latent class variables for individual i in the placebo and drug group, respectively. The probability of latent class membership is modeled by the logistic regressions

$$\log[P(p_i=1|z_p)/P(p_i=2|z_p)]=\alpha_p+\gamma_p z_{pi} \tag{7}$$

$$\log[P(d_i=1|z_d)/P(d_i=2|z_d)]=\alpha_d+\gamma_d z_{di} \tag{8}$$

where z_{pi} is a covariate influencing class membership of the placebo latent class variable p_i, and z_{di} is a covariate influencing class membership of the drug latent class variable d_i. Let 1 refer to the non-responder class, and 2 the responder class. The relationship between p_i and d_i is expressed via the log odds ratio

$$\log\left[\frac{\pi_{11i}/\pi_{12i}}{\pi_{21i}/\pi_{22i}}\right]=\omega_i \tag{9}$$

where

$$\pi_{kli}=P(p_i=k,d_i=l|z_{pi},z_{di}) \tag{10}$$

It may be noted that this model assumes that treatment status does not influence latent class membership. Class membership is conceptualized as a quality characterizing a subject before entering the trial. As an alternative, one may hypothesize that class membership arises as a function of treatment, with a single class during the pre-treatment period. This approach will not be explored here. If treatment influences only the class membership probabilities and not the random effect means directly, then the distinction between the four hypothesized latent classes of subjects cannot be made. Also, if the model allows treatment to influence class membership, the principal stratification interpretation of [8] referred to in Section 3 is not valid.

Consider the depression outcome y_{ti} observed at time point t for individual i. Let η denote random effects, let a_t denote time, and let ε_t denote residuals containing measurement error and time-specific variation. In line with the real-data analysis in Section 4.1, it is assumed that the outcome is observed at two pre-randomization time points. For the first, pre-randomization piece,

DOI: 10.1002/sim

the means of the random effects vary as a function of the combination of placebo latent class $k(k=1,2)$ and drug latent class $l(l=1,2)$,

$$y_{ti}^{\text{pre}}|_{p_i=k,d_i=l}=\eta_{0i}^{\text{pre}}+\eta_{1i}^{\text{pre}}a_t+\varepsilon_{ti}^{\text{pre}} \tag{11}$$

with $a_1=0$ to center at baseline, and random effects

$$\eta_{0i}^{\text{pre}}|_{p_i=k,d_i=l}=\alpha_{0kl}^{\text{pre}}+\zeta_{0i}^{\text{pre}} \tag{12}$$

$$\eta_{1i}^{\text{pre}}|_{p_i=k,d_i=l}=\alpha_{1kl}^{\text{pre}}+\zeta_{1i}^{\text{pre}} \tag{13}$$

With only two pre-randomization time points, the model is simplified by specifying a non-random slope, $V(\zeta_1^{\text{pre}})=0$, for identification purposes. All pre-randomization parameters are assumed to be equal across the placebo and drug groups.

Assume for simplicity a single drug and denote the treatment status for individual i by the dummy variable w_i ($w=0$ for the placebo group and $w=1$ for the drug group). For the second, post-randomization piece, a quadratic growth model is specified ($t=3,4,\ldots,T$),

$$y_{ti}|_{p_i=k,d_i=l}(w_i)=\eta_{0i}(w_i)+\eta_{1i}(w_i)(a_t-c)+\eta_{2i}(w_i)(a_t-c)^2+\varepsilon_{ti}(w_i) \tag{14}$$

where the a_t values are set according to the distance in timing of measurements and c is a constant such as the average of a_t. The random effects are allowed to be influenced by the group dummy covariate w, their distributions varying as a function of the combination of trajectory classes k and l,

$$\eta_{0i}|_{p_i=k,d_i=l}(w_i)=\alpha_{0kl}+\gamma_{0kl}w_i+\zeta_{0i}(w_i) \tag{15}$$

$$\eta_{1i}|_{p_i=k,d_i=l}(w_i)=\alpha_{1kl}+\gamma_{1kl}w_i+\zeta_{1i}(w_i) \tag{16}$$

$$\eta_{2i}|_{p_i=k,d_i=l}(w_i)=\alpha_{2kl}+\gamma_{2kl}w_i+\zeta_{2i}(w_i) \tag{17}$$

The residuals ζ_i in the first and second piece have a 4×4 covariance matrix $\Psi_{k,l}$, here taken to be constant across the k,l classes. The residuals ε_{ti} of the two pieces have a $T\times T$ covariance matrix $\Theta_{k,l}$, here taken to be constant across classes as well. All residuals are assumed i.i.d. and normally distributed. For simplicity, $\Psi_{k,l}$ and $\Theta_{k,l}$ are assumed to not vary across treatment groups, although this can be relaxed. In the actual analyses in Sections 4.1 and 4.2, the drug status covariate is represented by yet another latent class variable, where the latent status is known (this adds one extra class probability parameter, which could be ignored, but is included in the reporting of all the models). This creates a total of eight classes where the variances can be allowed to vary over subsets of those classes.

As seen in (15)–(17), the placebo group ($w_i=0$) consists of subjects that vary in the means of the growth factors, which are represented by α_{0kl}, α_{1kl}, and α_{2kl}. This gives the average development in the absence of medication for each of the four types of subjects of Table I. Because of randomization, the placebo and drug groups are assumed to be statistically equivalent at the first two time points. Drug effects are described in the second piece by γ_{0kl}, γ_{1kl}, and γ_{2kl} as a change in the development that can be different for the four types of subjects.

This model allows the assessment of drug response in the presence of placebo response both in terms of γ_{0kl}, γ_{1kl}, and γ_{2kl} and in terms of the probabilities of (7)–(9), giving the prevalence of each of the four types of subjects of Table I.

DOI: 10.1002/sim

The analysis may use the above model in an exploratory way or by restricting the parameters of the second growth piece to correspond to the hypothesized non-responder and responder classes. Used in an exploratory way, the resulting four types of subjects may not have the interpretation used for Table I. For example, instead of the Placebo Only Responder type two sets of Drug Only Responder types may be found, differing in their non-response/response characteristics. Restrictions on the parameters can be applied, for example, by forcing the estimated outcome mean at the last time point to be less than a certain value indicating response, and greater than a certain value indicating non-response.

An equivalent way to formulate the model is as in [12] using a single latent class variable that has four categories corresponding to the four types of subjects in Table I. This approach does not emphasize the hypothesis that the four types of Table I arise as a combination of an individual being prone to placebo and/or drug response. It also does not enable separate covariates for the latent class covariates as in (7), (8). Nevertheless, the single latent class variable approach is used in the present analyses for simplicity given that no covariates are included. As mentioned, the placebo-drug dummy variable is handled via an additional latent class variable with known class status.

The models discussed may be estimated by maximum-likelihood using the Mplus program [23]. Mplus was used for both the real-data and Monte Carlo analyses. Mplus scripts for the analyses are available from the first author. For a technical description, see [12, 13].

The choice of the number of latent classes in mixture modeling is often guided by the minimum of the BIC, penalizing models with many parameters [24, 25],

$$\mathrm{BIC} = -2\log L + r\log n \tag{18}$$

where $\log L$ is the loglikelihood, r is the number of free parameters in the model, and n is the sample size. The lower the BIC value, the better the model. BIC, however, is not always reliable for small sample sizes but may underestimate the number of classes for samples of size 200 and below [26]. Classification of subjects into the latent classes can be carried out based on the estimated posterior probabilities of class membership [17]. A summary measure of the classification quality is given by the entropy measure (see, e.g. [27]),

$$E_K = 1 - \frac{\sum_i \sum_k (-\hat{p}_{ik} ln\, \hat{p}_{ik})}{n \ln K} \tag{19}$$

where $\hat{p}_{ik}$ denotes the estimated posterior probability for individual i in class k. Entropy values range from 0 to 1, where entropy values close to 1 indicate clear classifications in that the entropy decreases for probability values that are not close to 0 or 1. Values of at least 0.8 typically represent good classification quality.

4.1. GMM applied to the antidepressant trial

The GMM will here be applied to the repeated measures data from the antidepressant trial, whereas the next section treats the schizophrenia trial. For the antidepressant trial the GMM approach uses a quadratic growth function for the second, post-randomization piece. As a first step, the drug and placebo groups are analyzed separately to show the trajectory features.

4.1.1. Separate GMM analysis of drug and placebo groups. Table III summarizes the results of GMM of the drug and placebo groups analyzed separately. For the drug group, BIC points to three

DOI: 10.1002/sim

Table III. Summary of separate analyses of drug and placebo groups.

# classes	LL	#parameters	BIC
Drug group ($n=102$)			
1	−3000	27	6125
2	−2980	33	6114
3	−2965	39	6110
4	−2959	45	6126
Placebo group ($n=52$)			
1	−1597	27	3300
2	−1587	33	3305
3	−1579	39	3313
4	−1572	45	3322

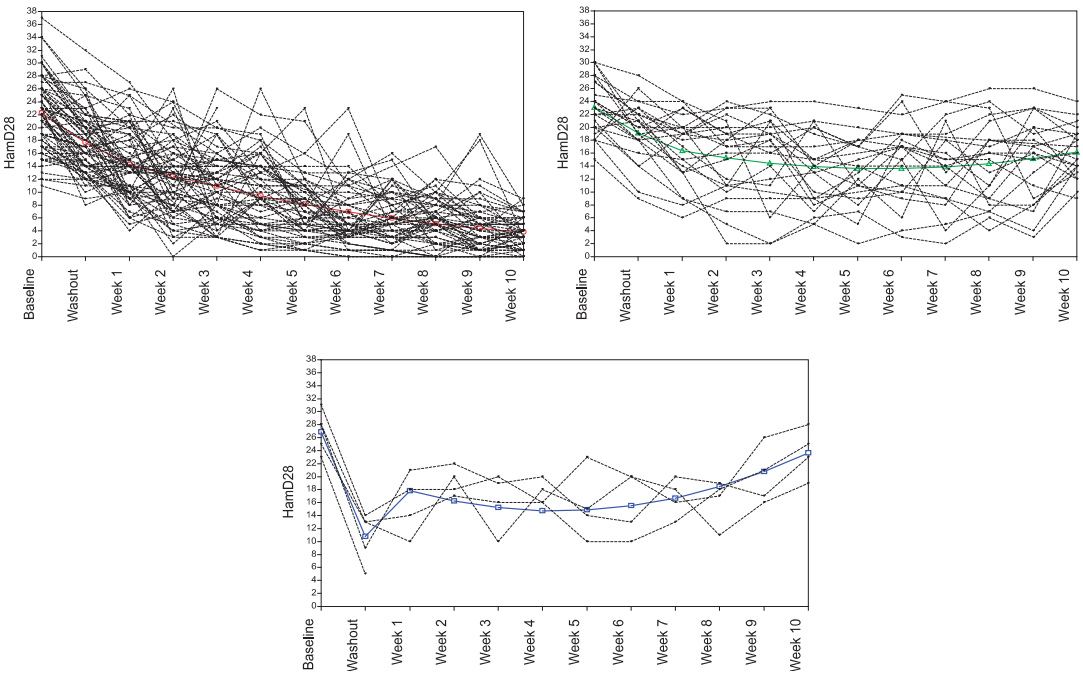

Figure 3. Observed trajectories divided into three classes for the drug group (class 1 is top left, class 2 is top right, and class 3 is at the bottom).

classes. As mentioned earlier, however, the low sample size may make BIC less trustworthy and suggest too few classes. The 3-class solution has an entropy of 0.87. The posterior probabilities are used to classify observed trajectories for subjects most likely to belong to each of the three classes as shown in Figure 3. The mean curves of the 3-class solution for the drug group are shown in the top part of Figure 4. Class 1 is drug responder class containing 67 per cent of the subjects. Class 2

DOI: 10.1002/sim

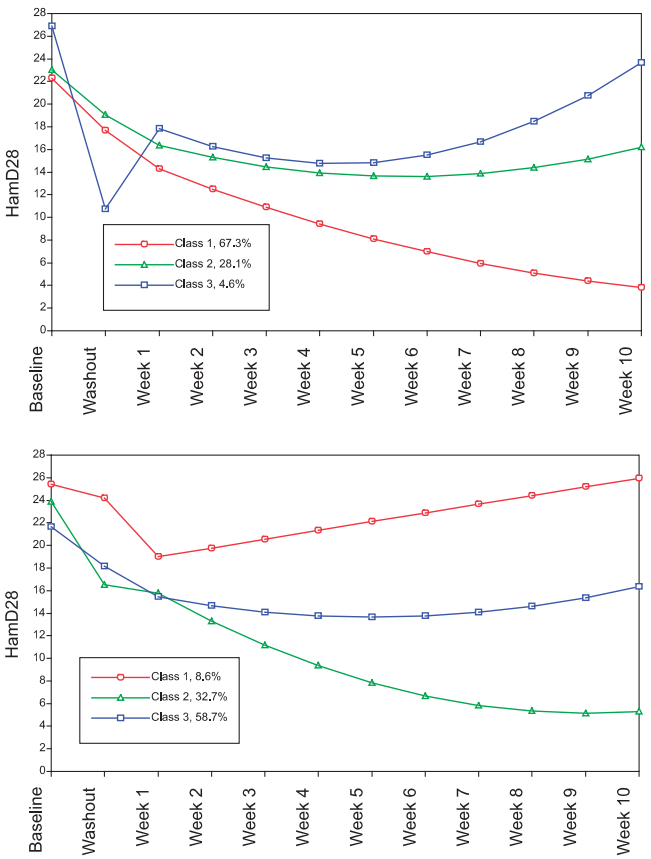

Figure 4. Estimated mean curves for 3-class model for drug group (top) and placebo group (bottom).

is a drug non-responder class containing 28 per cent. Class 3 is a drug non-responder class with volatile development, containing 5 per cent. The 4-class solution gives two classes very similar in shape and prevalence to class 1 and class 3, whereas class 2 is split into two non-responder classes.

Judged by BIC, the placebo group analyses suggest that a conventional, single-class growth model is sufficient. Again, the low sample size may cause BIC to underestimate the number of classes. The mean curves for the 3-class solution for the placebo group are shown in the bottom part of Figure 4. For the placebo group only one-thirds are in the responder class, which is in contrast with the two-thirds in the responder class for the drug group. What is not clear from these analyses, however, is what portion of the drug responders and drug non-responders would have been responders and non-responders under placebo. For this the joint analysis of both groups is needed.

4.1.2. Joint GMM analysis of drug and placebo groups. Four models are fitted as summarized in the bottom part of Table II, labeled Growth mixture analysis. Judging by BIC, the parsimonious model 7 with three classes and two sets of means is better than model 6, the ITT 1-class random effect repeated measures model. For model 7, the Drug Only Responder prevalence is

DOI: 10.1002/sim

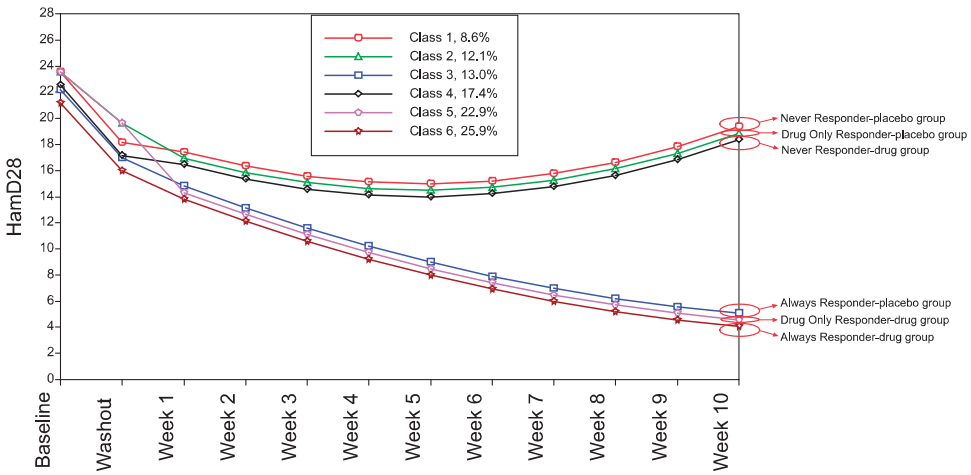

Figure 5. Estimated mean curves for model 7: 3-class model, two sets of means.

estimated as 35 per cent and with a week 10 treatment effect of 14 units on the Hamilton D28 scale. The entropy is 0.67. The results for model 7 are not far from those of models 2, 3, and 5.

Figure 5 shows the estimated mean trajectories for model 7, the 3-class model with two sets of means. The modeling uses the approach of letting the treatment dummy covariate be represented by a latent class variable with known classes as described in Section 4. This results in a total of six latent classes: classes 1–3 are for the placebo group and classes 4–6 are the corresponding classes for the drug group. The top three curves are identical and the bottom three curves are identical, but the curves have been jiggled here to show the class membership. It is seen that classes 1 and 4 represent Never Responders (non-response in both groups), classes 2 and 5 represent Drug Only Responders (non-response in placebo group and response in drug group), and classes 3 and 6 represent Always Responders (response in both groups).

Figure 6 shows the estimated mean trajectories for model 8, the 3-class model with four sets of means. The entropy is 0.91. As for the previous figure, classes 1–3 are for the placebo group and classes 4–6 are the corresponding classes for the drug group. Classes 1 and 4 represent Never Responders and classes 3 and 6 represent Always Responders. For classes 2 and 5, however, the outcome is unclear. Although the class 5 trajectory for the drug group ends at a lower Week 10 value than the corresponding class 2 trajectory for the placebo group, the class 5 mean trajectory ends with a high value at Week 10. It is therefore unclear if this can be characterized as a Drug Only Responder class. The class percentages for this solution are also quite different than for the other models, with the Never Responder class prevalence estimated as only 4 per cent, which does not seem plausible. As for the week 10 analysis using model 4, the 3-class, 4-mean model is therefore a questionable representation of the data.

Model 9, the 4-class model with eight sets of means, does not have a better BIC than the other models, but as discussed in Section 4, BIC tends to underestimate the number of latent classes in small samples. The entropy is 0.84. Table IV gives the estimated class prevalences and Figure 7 shows the estimated mean curves for the four types of subjects divided into the placebo and drug groups, resulting in eight classes of curves.

DOI: 10.1002/sim

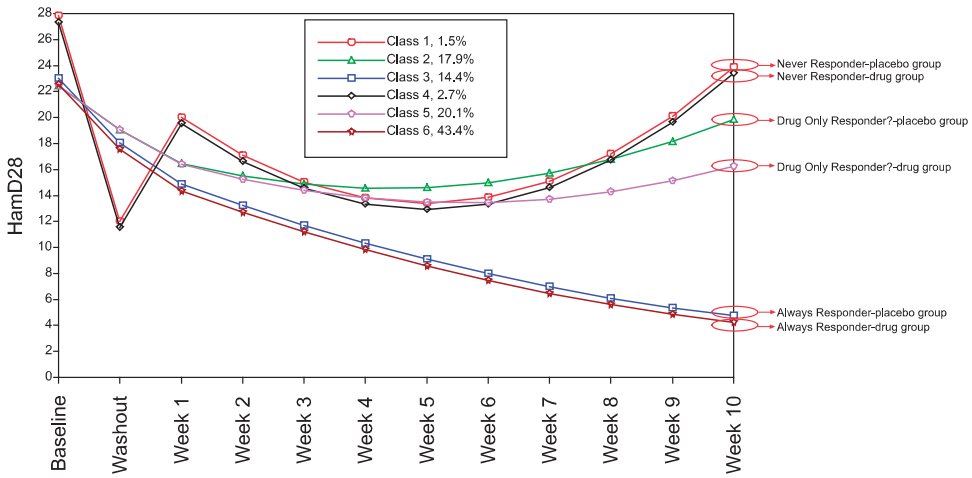

Figure 6. Estimated mean curves for model 8: 3-class model, four sets of means (AIR model).

Table IV. Antidepressant trial prevalence of four types of subjects under 4-class model with eight sets of means.

Placebo	Drug group		
Group	Non-responder	Responder	
Non-responder	Never responder 28 per cent	Drug only responder 26 per cent	54 per cent
Responder	Placebo only responder 4 per cent	Always responder 42 per cent	46 per cent
	32 per cent	68 per cent	

Figure 7 shows that the Never Responder subjects are found in class 2 for the placebo group and in class 6 for the drug group. Their week 10 means are around 15. As seen in Table IV, the prevalence of this type of subjects is estimated as 28 per cent.

The Drug Only Responder subjects are found in class 3 for the placebo group and class 7 for the drug group. The prevalence of this type of subjects is estimated as 26 per cent. The estimated week 10 treatment effect is 18 (corresponding to an estimated mean of 5 for the drug group and an estimated mean of 23 for the placebo group), which corresponds to a little over two SDs. The 95 per cent confidence interval for the treatment effect is $15.2-21.5$.

The Placebo Only Responder subjects are found in class 1 for the placebo group and class 5 for the drug group. It is seen that the placebo response is temporary, limited to weeks 4–7, with a later upswing in depression. This type of subjects has the highest baseline score of about 28 and a more volatile development with a sharp increase in depression around week 1. The prevalence of this type of subjects is estimated as only 4 per cent.

The Always Responder subjects are found in class 4 for the placebo group and class 8 for the drug group. Their estimated mean at week 10 is only around 4. The prevalence of this type of subjects is estimated as 42 per cent.

DOI: 10.1002/sim

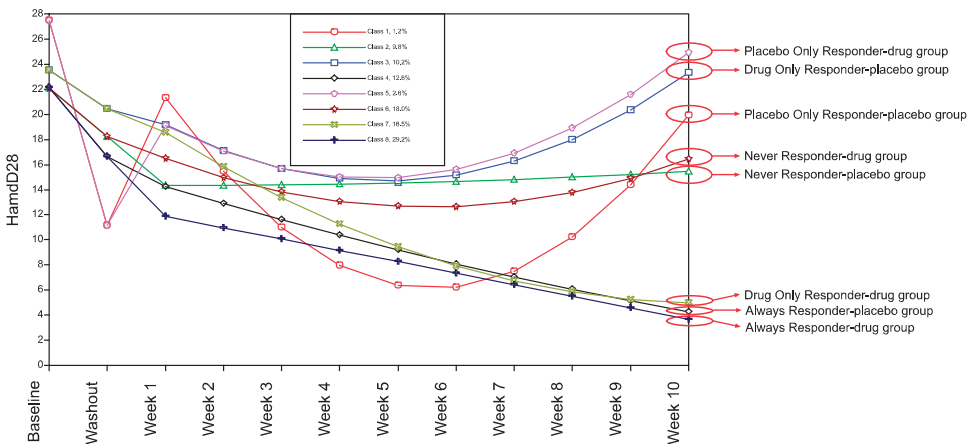

Figure 7. Estimated mean curves for model 9: 4-class model, eight sets of means.

Table IV also gives the estimated marginal response rates for placebo and drug. The placebo response rate is 46 per cent, whereas the drug response rate is 68 per cent.

The antidepressant analysis results are summarized in Table II, listing the models in order of appearance. All models are estimated using maximum likelihood except model 3, which uses the moment estimator of Section 3.1. With the exception of the two maximum-likelihood estimated 3-class, 4-mean AIR models, model 4 and model 8, the estimates are on the whole rather close. Only model 9, the 4-class GMM with eight sets of means, uncovers the hypothesized four classes of Table I. For this model the drug effect of 18 at week 10 for the Drug Only Responder class corresponds to a little over two SDs in terms of the total variation at week 10.

Previous attempts to isolate placebo response in antidepressants trials by statistical modeling include [28], where five trajectory categories were hypothesized for an individual treated with an active drug and where placebo subjects can fall into only one of the first three categories: (A) non-responders, (B) non-responders with initial placebo effect, (C) placebo responders, (D) true drug responders, and (E) mixture effects responders. This classification does not have the clarity of the potential outcomes—principal stratification approach used in the current paper. For example, the requirement that placebo subjects cannot occupy category (D), corresponding to Drug Only Response, is in contrast with the view of the current paper that the categories exist as principal strata before randomization. Also, it is not clear if, for example, treatment subjects in category (D) would fall in category (A) if they had been given placebo. Nevertheless, the trajectory types of the first four categories are found in the current paper, including the category (B) trajectory type seen for the Placebo Only Responder class. The fifth category (E) (‘subjects who have an initial improvement due to nonspecific effects and then experience a drug effect’) represents a more fine-grained distinction than used here. More recently, [29] used infinite mixtures in an attempt at isolating drug effects in the presence of placebo effects.

4.2. GMM applied to the schizophrenia trial

The schizophrenia trial example offers two new growth mixture features. First, the outcome is binary instead of continuous. Second, the fact that the outcome is categorical makes it possible to

DOI: 10.1002/sim

test the model against data using conventional likelihood-ratio χ^2 testing against the unrestricted model represented by a multinomial distribution for the corresponding frequency table. Also, in this trial the sample size is larger ($n=437$) so that the use of BIC to help decide on the number of classes is more reliable.

For the schizophrenia trial data the ITT model, the 3-class, 2-mean model, and the 4-class, 8-mean model will be discussed. Here, a linear logistic growth model with random intercept and random slope is applied (time was specified in weeks, not taking the square root as in previous growth analyses of these data). The conventional, ITT single-class random effects model gives LL$=-858$ with eight parameters, and BIC$=1765$. The week 10 estimated probabilities are 0.60 and 0.27 for the placebo and drug groups, respectively. The 3-class, 2-mean model gives LL$=-840$ with 11 parameters, and a better BIC of 1748. The entropy is 0.75. The 4-class, 8-mean model gives LL$=-836$ with 19 parameters, and a worse BIC of 1788. The entropy is 0.72. Given the binary outcomes, a likelihood-ratio χ^2 test for the frequency table of all response patterns is also available for evaluating the fit of the three models. The 1-class model gives $\chi^2=77$ with 23 degrees of freedom, whereas the 3-class model improves the model fit to $\chi^2=42$ with 20 degrees of freedom. The 4-class model obtains $\chi^2=33$ with 12 degrees of freedom.

The prevalences for the 3-class model are estimated as: Never Responder class 45 per cent, Drug Only Responder class 27 per cent, and Always Responder Class 28 per cent. The 4-class does not show a Placebo Only Responder class, but instead two non-responder classes. The prevalences are Never Responder class 60 per cent, Drug Only Responder class 27 per cent, and Always Responder Class 13 per cent.

The estimated mean probability curves for the 3-class model are shown in Figure 8. Curves are shown for six classes, where the first three are for the placebo group and the next three the corresponding classes for the drug group. Classes 1 and 4 have the same curves as do classes 3 and 6, but these classes are jiggled to be slightly different in order for the curves to show up. The Drug Only Responders appear as class 2, showing non-response in the placebo group and as class 5 showing response in the drug group. The figure shows that Drug Only Responders have a quicker improvement than Always Responders (class 3 and class 6 for placebo and drug groups, respectively). Although the model has only responder and non-responder mean parameters, the linear logistic growth model produces this differential improvement due to different starting points at week 0.

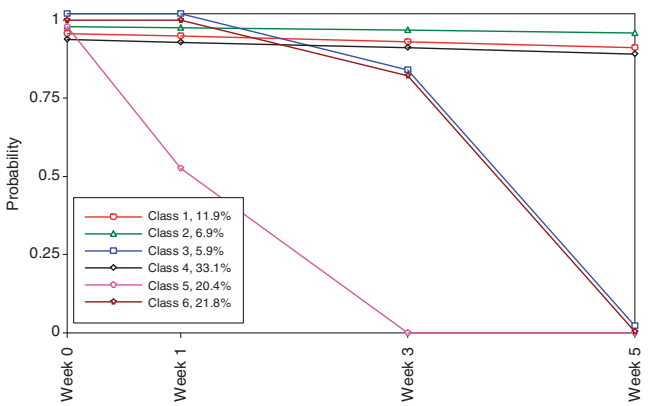

Figure 8. Schizophrenia trial: 3-class, 2-mean model.

DOI: 10.1002/sim

GMM of these data was also carried out in [30] (the outcome was kept in ordinal form and the square root of week was used). The model is similar in that it allows for responder and non-responders in both the treatment and the placebo groups. The model is different in that only two classes are used and that it does not allow for different drug effects in the classes. Investigating their model, it was found that the drug effect was significantly different across the classes. This model resulted in LL = −842 with 11 parameters, and BIC = 1751, which is a slightly worse BIC value than for the proposed 3-class, 2-mean model. More importantly, the modeling in [30] makes no attempt at causal inference regarding potential outcomes and therefore does not make a distinction between the classes of Never Responders, Drug Only Responders, Placebo Only Responders, and Always Responders. Because of this, it cannot make inference about the Drug Only Responder rate.

5. MONTE CARLO SIMULATIONS

The 4-class model of Section 4 will now be used in a Monte Carlo simulation with characteristics similar to that of the real-data analysis of the antidepressant trial. The aim is to study how well the parameter values can be recovered under different conditions.

Two pre-randomization time points are considered together with 10 post-randomization time points. The hypothetical prevalence for each of the four types of subjects is given in Table V. For the post-randomization piece of the GMM in (14)–(17), a linear function is used for simplicity. The choice of class-varying parameters results in the mean trajectories as shown in Figure 9. Note that a lower score indicates a lower level of depression. The first four classes of the figure correspond to the four types of subjects in the placebo group and the last four classes correspond to the four types of subjects in the drug group. For each group, the four classes are numbered in the order Never Responder, Drug Only Responder, Placebo Only Responder, and Always Responder. At week 12 the SD for the outcome is approximately eight. The week 12 means for non-responders and responders are about two SDs apart, except for Drug Only Responders in the placebo group (‘class 2’) for whom the mean difference is a little over one SD. In terms of within-class SDs the week 12 mean differences are about three to four SDs. Monte Carlo simulation results are presented for $n = 500$, $n = 200$, and $n = 100$ using 1000 replications. To limit the results shown, the focus is on the estimated probabilities for the four cells of Table V and the week 12 treatment effects expressed as drug mean–placebo mean for those four cells. A negative value indicates that the drug is beneficial. The placebo and drug group marginal responder probabilities are labeled π. Tables VI–VIII show the results for $n = 500$, $n = 200$, and $n = 100$.

Table VI shows that all quantities are well estimated with little bias, good standard error estimates, and good 95 per cent coverage for $n = 500$. The estimated power to reject the favorable

Table V. Hypothesized prevalence of four types of subjects for Monte Carlo simulation study.

Placebo	Drug group		
Group	Non-responder (per cent)	Responder (per cent)	
Non-responder	25	35	60
Responder	5	35	40
	30	70	

DOI: 10.1002/sim

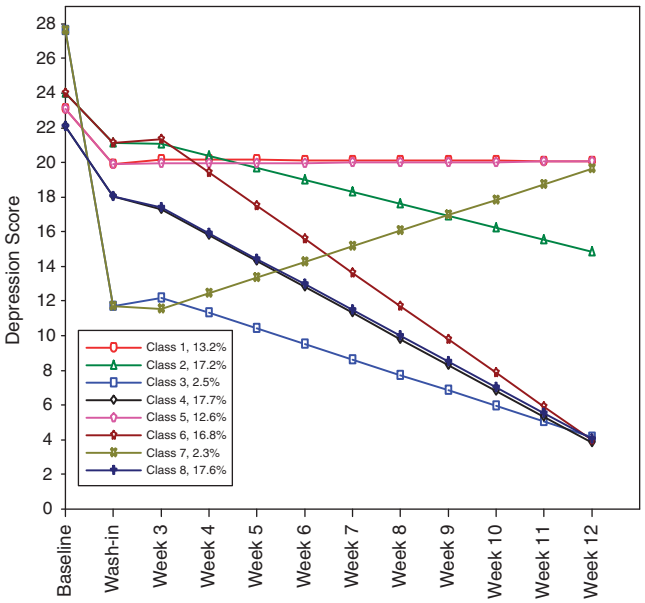

Figure 9. Growth mixture model for placebo and drug responders.

Table VI. Monte Carlo results for $n=500$.

Quantity	True value	Mean estimate	SD	Mean SE	MSE	95 per cent coverage	Power
π_n	0.250	0.2485	0.0271	0.0295	0.0007	0.962	0.999
π_d	0.350	0.3505	0.0399	0.0435	0.0016	0.958	1.000
π_p	0.050	0.0521	0.0134	0.0145	0.0002	0.953	0.970
π_a	0.350	0.3489	0.0350	0.0372	0.0012	0.951	1.000
Placebo π	0.400	0.4010	0.0345	0.0362	0.0012	0.959	1.000
Drug π	0.700	0.6994	0.0275	0.0288	0.0008	0.951	1.000
$\mu_{n1}-\mu_{n0}$	0.000	−0.0141	0.8313	0.8334	0.6906	0.954	0.046
$\mu_{d1}-\mu_{d0}$	−11.000	−10.9306	1.1641	1.1049	1.3586	0.877	0.997
$\mu_{p1}-\mu_{p0}$	16.000	15.9790	1.7084	1.6751	2.9161	0.919	0.998
$\mu_{a1}-\mu_{a0}$	0.000	−0.0494	1.1060	1.0094	1.2245	0.844	0.156

drug effect among Drug Only Responders ($\mu_{d1}-\mu_{d0}$) is high, 0.997. For the smaller sample of $n=200$ in Table VII, the results are still good and only slightly worse than for $n=500$. Table VIII for $n=100$ still shows acceptable results but the coverage for the Drug Only Responder effect is now problematic due to a bias in the estimated standard error. The mean estimates are, however, still quite good at this sample size, which is reassuring given that few trials involve more than this many subjects. For $n=50$ (data not shown), also the point estimates show large biases.

DOI: 10.1002/sim

Table VII. Monte Carlo results for $n = 200$.

Quantity	True value	Mean estimate	SD	Mean SE	MSE	95 per cent coverage	Power
π_n	0.250	0.2474	0.0483	0.0513	0.0023	0.919	0.984
π_d	0.350	0.3513	0.0699	0.0771	0.0049	0.936	0.972
π_p	0.050	0.0580	0.0264	0.0289	0.0008	0.921	0.771
π_a	0.350	0.3434	0.0604	0.0694	0.0037	0.939	0.986
Placebo π	0.400	0.4014	0.0593	0.0636	0.0035	0.937	0.989
Drug π	0.700	0.6946	0.0464	0.0507	0.0022	0.935	0.999
$\mu_{n1} - \mu_{n0}$	0.000	0.0561	1.5384	1.4625	2.3673	0.922	0.078
$\mu_{d1} - \mu_{d0}$	−11.000	−10.9769	1.9278	1.7088	3.7132	0.847	0.976
$\mu_{p1} - \mu_{p0}$	16.000	15.8315	3.2447	2.4249	10.5460	0.831	0.983
$\mu_{a1} - \mu_{a0}$	0.000	0.0507	1.6597	1.5074	2.7543	0.832	0.168

Table VIII. Monte Carlo results for $n = 100$.

Quantity	True value	Mean estimate	SD	Mean SE	MSE	95 per cent coverage	Power
π_n	0.250	0.2435	0.0671	0.0642	0.0045	0.872	0.947
π_d	0.350	0.3514	0.0961	0.0939	0.0092	0.872	0.905
π_p	0.050	0.0674	0.0391	0.0332	0.0018	0.893	0.544
π_a	0.350	0.3377	0.0858	0.0799	0.0075	0.871	0.959
Placebo π	0.400	0.4051	0.0839	0.0802	0.0071	0.888	0.981
Drug π	0.700	0.6890	0.0662	0.0658	0.0045	0.901	0.997
$\mu_{n1} - \mu_{n0}$	0.000	0.1296	2.3661	1.8794	5.6098	0.838	0.162
$\mu_{d1} - \mu_{d0}$	−11.000	−11.0642	2.7479	1.9818	7.5475	0.772	0.960
$\mu_{p1} - \mu_{p0}$	16.000	15.5789	4.4785	2.1952	20.2146	0.650	0.968
$\mu_{a1} - \mu_{a0}$	0.000	0.0404	2.3167	1.6840	5.3632	0.799	0.201

6. DISCUSSION

The Monte Carlo simulation study suggests that the parameters of the proposed GMM can be well recovered for sample sizes of at least 200 and settings similar to those used in the simulations. Point estimates, but not SEs, are reasonably well estimated even at $n = 100$, but not much below this sample size. The antidepressant trial data analysis uses $n = 154$, which therefore approaches the lower limit of what is an acceptable sample size. Future studies could investigate whether the more parsimonious 3-class model with two sets of means may do better at low sample sizes and/or with binary outcomes.

The clinical trial data analyses using GMM provide an interesting view of the drug response. For the antidepressant trial data the 4-class GMM-estimated marginal rate of drug response was 68 per cent versus 46 per cent for placebo. The Drug Only Responder rate was estimated as 26 per cent. For the schizophrenia trial data the estimated marginal rate of drug response was 70 per cent versus 40 per cent for placebo, whereas the Drug Only Responder rate was estimated as 35 per cent. The GMM approach emphasizes that the marginal drug response rate is obtained as a mixture of Drug Only Responders and Always Responders. For both data sets, at least half of those who respond to the drug are subjects who would also respond to placebo. This finding challenges the convention

DOI: 10.1002/sim

of assessing a drug effect by using the marginal response rates. The ability to uncover this finding illustrates the strength of the principal stratification, mixture modeling approach. As summarized in Table II for the antidepressant trial data it is also noteworthy that the effect size estimated by GMM for the Drug Only Responder class is much larger than that of the ITT approach.

It is also noteworthy that antidepressant trials typically truncate the study sample by eliminating subjects who show a strong placebo response during the washout period, that is, before randomization. Owing to this, the placebo response rate is underestimated. Using the proposed approach such sample truncation is not necessary.

In summary, the paper points to the usefulness of augmenting the conventional approach of comparing marginal response rates and using ITT effect estimates with the approach of GMM-estimated prevalences for the four classes of subjects. From a clinical research point of view, it is of special interest to consider the prevalence of the Drug Only Responder class and the Drug Only Responder drug effect estimate.

The mixture approach can be expanded in several ways. Further investigations can attempt to include predictors of the latent class membership for placebo and drug response. Also, it would be possible to examine the variation in the percentage of Drug Only Responders across multiple trials involving different populations to understand which subgroups are more likely to benefit from a particular medication. GMM can potentially also be used in a dynamic fashion as a basis for switching subjects onto different treatment regimes.

ACKNOWLEDGEMENTS

The authors thank Patrick McGrath for providing the antidepressant trial data. They acknowledge helpful comments from Michael Sobel, Patrick McGrath, Booil Jo, and Elizabeth Stuart. The research was supported by grant 1 R21 AA10948-01A1 from NIAAA, by NIMH and NIDA under grant no. MH40859, and by grant P30 MH066247.

REFERENCES

1. McGrath P, Stewart JW, Janal MN, Petkova E, Quitkin FM, Klein DF. A placebo-controlled study of fluoxetine versus imipramine in the acute treatment of atypical depression. *American Journal of Psychiatry* 2000; **157**: 344–350.
2. Montgomery SS. Clinically relevant effect sizes in depression. *European Neuropsychopharmacology* 1994; **4**: 283–284.
3. Hedeker D, Gibbons RD. *Longitudinal Data Analysis*. Wiley: New York, 2006.
4. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**:444–445.
5. Holland P. Causal inference, path analysis, and recursive structural equations models. In *Sociological Methodology*, Clogg C (ed.). American Sociological Association: Washington, DC, 1988; 449–493.
6. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.
7. Rubin DB. Bayesian inference for causal effects. *The Annuals of Statistics* 1978; **6**:34–58.
8. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; **58**:21–29.
9. Little RJ, Yau LHY. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychological Methods* 1998; **3**:147–159.
10. Gibbons RD, Hedeker D, Waternaux C, Davis JM. Random regression models: a comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin* 1988; **24**:438–443.
11. Quitkin FM, Rabkin JG, Ross D, Stewart JW. Identification of true drug response to antidepressants. Use of pattern analysis. *Archives of General Psychiatry* 1984; **41**:782–786.

DOI: 10.1002/sim

12. Muthén B, Brown CH, Masyn K, Jo B, Khoo ST, Yang CC, Wang CP, Kellam S, Carlin J, Liao J. General growth mixture modeling for randomized preventive interventions. *Biostatistics* 2002; **3**:459–475.
13. Muthén B, Asparouhov T. Growth mixture modeling: analysis with non-Gaussian random effects. In *Longitudinal Data Analysis*, Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G (eds). Chapman & Hall/CRC Press: Boca Raton, FL, 2008; 143–165.
14. Rubin DB. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* 2004; **31**:161–170.
15. Mealli F, Imbens GW, Ferro S, Biggeri A. Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics* 2004; **5**:207–222.
16. Gallop R, Small DS, Lin JY, Elliott MR, Joffe M, Ten Have TR. Mediation analysis with principal stratification. *Statistics in Medicine* 2009; **28**:1108–1130.
17. McLachlan GJ, Peel D. *Finite Mixture Models*. Wiley: New York, 2000.
18. Titterington DM, Smith AFM, Makov UE. *Statistical Analysis of Finite Mixture Distributions*. Wiley: New York, 1985.
19. Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 1999; **55**:463–469.
20. Lin H, Turnbull BW, McCulloch CE, Slate E. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* 2002; **97**:53–65.
21. Elliott MR, Gallo JJ, Ten Have TR, Bogner HR, Katz IR. Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics* 2005; **6**:119–143.
22. Beunckens C, Molenberghs G, Verbeke G, Mallinckrodt C. A latent-class mixture model for incomplete longitudinal Gaussian data. *Biometrics* 2008; **64**:96–105.
23. Muthén B, Muthén L. *Mplus User's Guide*. Muthén and Muthén: Los Angeles, CA, 2008.
24. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978; **6**:461–464.
25. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1993; **90**:773–795.
26. Nylund KL, Asparouhov T, Muthén B. Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Structural Equation Modeling* 2007; **14**:535–569.
27. Ramaswamy V, DeSarbo W, Reibstein D, Robinson W. An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science* 1993; **12**:103–124.
28. Tarpey T, Petkova E, Ogden T. Profiling placebo responders by self-consistent partitioning of functional data. *Journal of the American Statistical Association* 2003; **98**:850–858.
29. Tarpey T, Petkova E, Govindarajulu U. Predicting potential placebo effect in drug treated subjects. *The International Journal of Biostatistics* 2009; **5**. DOI: 10.2202/1557-4679.1152.
30. Xu W, Hedeker D. A random-effects mixture model for classifying treatment response in longitudinal clinical trials. *Journal of Biopharmaceutical Statistics* 2001; **11**:253–273.