**Article:**

Firth, N. orcid.org/0000-0003-1984-6869, Saxon, D. orcid.org/0000-0002-9753-8477, Stiles, W.B. et al. (1 more author) (2019) Therapist and clinic effects in psychotherapy: a three-level model of outcome variability. Journal of Consulting and Clinical Psychology, 87 (4). pp. 345-356. ISSN 0022-006X

https://doi.org/10.1037/ccp0000388

# Therapist and clinic effects in psychotherapy: a three-level model of outcome variability

**Nick Firth, David Saxon, William B. Stiles, & Michael Barkham**



10.1037/ccp0000388

## Abstract

Objective: The study aimed to 1) investigate the effect of treatment location on clinical outcomes for patients receiving psychological therapy (a clinic effect, akin to the concept of a therapist effect), and 2) assess the impact of explanatory individual and aggregate demographic and process variables on the clinic and therapist effects.

Method: The sample comprised 26,888 patients, seen by 462 therapists, across 30 clinics. Mean patient age was 38 years (69% female, 90% White, 92% planned ending). The dependent variable was patients' post-therapy score on the Clinical Outcomes in Routine Evaluation – Outcome Measure. An incremental three-level multilevel model was constructed. Markov Chain Monte Carlo estimation created 95% probability intervals for the clinic and therapist effects.

Results: A three-level model with no explanatory variables detected a clinic effect of 8.2%, significantly larger than the therapist effect of 3.2%. Adding explanatory variables significantly reduced the clinic effect to 1.9% but did not significantly alter the therapist effect (3.4%). Patient-level symptom severity and employment status, and clinic-level percentage of White patients and healthcare sector explained the most clinic outcome variance and overall outcome variance.

Conclusions: Substantial variability in clinical outcomes was found between clinics providing psychological therapy. Socioeconomic mix of patients explained significant proportions of variability at the clinic level but not the therapist level. Clinical implications include the need to go beyond the therapist-patient interaction in order to deliver effective psychological therapy. Future research is also needed to identify the mechanisms by which clinic and/or area-level factors impact on clinical outcomes.

Keywords: therapist effect, clinic effect, psychological therapy, outcome, deprivation

Public Health Significance Statements: This study demonstrates that typical outcomes for people receiving psychological therapy vary systematically across clinics in the UK. Levels of (un)employment and ethnic/racial composition may help to explain between-clinic differences in effectiveness. It is important to consider the broader socioeconomic and geographic context in which therapy is offered to improve the effectiveness of psychological interventions.

Substantial variability exists in the extent of symptom improvement and other benefits from psychological therapies (Baldwin & Imel, 2013; Barkham, Lutz, Lambert, & Saxon, 2017; Bohart & Greaves Wade, 2013). Such findings occur regardless of whether researchers adopt a trial design (e.g., Vittengl et al., 2016) or employ practice-based datasets (e.g., Pybis, Saxon, Hill, & Barkham, 2017). Much of this variability in outcomes is understood to depend on patient factors - particularly initial symptom severity and socio-economic deprivation, with more severe or deprived patients having poorer outcomes (Bohart & Greaves Wade, 2013; Hamilton & Dobson, 2002). Evidence also suggests that process factors such as the number of sessions attended by patients, as well as patient engagement, are related to clinical outcome (Barrett, Chua, Crits-Christoph, Gibbons, & Thompson, 2008; Bohart & Greaves Wade, 2013; Stulz, Lutz, Kopta, Minami, & Saunders, 2013).

Therapists also contribute to this variability, however. Some therapists consistently deliver better outcomes than other therapists, even after controlling for patient factors in their case mix (e.g., Baldwin & Imel, 2013; Barkham, Lutz, Lambert, & Saxon, 2017; Saxon & Barkham, 2012). This therapist effect typically accounts for between 5-10 per cent of variance in patient outcomes (Baldwin & Imel, 2013; Johns, Barkham, Kellett, & Saxon, 2018). Therapist characteristics associated with this effect include empathy, alliance, professional self-doubt, and deliberate practice (Goldberg et al., 2016; Nissen-Lie et al., 2015; Wampold, Baldwin, grosse Holtforth, & Imel, 2017).

Similarly, the clinic where a patient is seen may also have an effect. Studies of multiple healthcare organizations and clinics show considerable variability in outcomes (Delgadillo, Asaria, Ali, & Gilbody, 2016; Royal College of Psychiatrists, 2013). Controlling for patient factors, a clinic (or site) effect of 1.8 per cent was found by Pybis et al. (2017), indicating the amount of variance in patient outcomes attributable to differences between clinics. Clinic effects may reflect systematic differences in clinical population characteristics, therapist recruitment practices, resource allocation, accessibility, etc. Additionally, clinic effects may reflect geographic and socio-economic factors in their patient population, such as levels of social support, safety, adequate housing, and socioeconomic deprivation (Barkham, Delgadillo, Firth, & Saxon, 2018; Clark et al., 2018; Delgadillo et al., 2016). This latter kind of effect may be more aptly termed a neighborhood effect. Although there is little research into neighborhood effects on psychological therapy outcomes, there is growing evidence that the local neighborhood impacts individual physical health (e.g. Pickett & Pearl, 2001).

In summary, patient, therapist, and clinic factors can all contribute to the variability in patient outcomes. However, little is known about how these three sources relate and interact with each other to produce the variability reported in the literature. Just as therapist effects research has produced therapist-targeted interventions, research into clinic effects could lead to clinic level interventions to address contributing factors and improve outcomes.

Although studies of therapist and clinic effects have controlled for patient variables, therapist effect estimates have been derived from either a single clinic (e.g., Firth, Barkham, Kellett, & Saxon, 2014) or did not consider differences between clinics in the analysis (e.g., Green, Barkham, Kellett, & Saxon, 2014). Similarly, clinic effect estimates have not considered differences between therapists (Pybis et al., 2017). Despite these shortcomings due largely to sample limitations, such studies have identified important interactions between therapists and patient variables, and between clinics and patient variables. Most consistently, the effect of initial severity on patient outcomes has been found to vary between therapists and between clinics (Pybis et al., 2017; Schiefele et al., 2016). Also, therapists vary in how the number of sessions attended affects outcomes (Saxon, Firth, & Barkham, 2017).

The three sources of variability (patient, therapist and clinic) are not independent – they are levels with a hierarchical structure. Patients are nested within therapists, who are in turn nested within clinics. To assess relative influences in such cases, multilevel modeling (MLM) methods are recommended (Goldstein, 2010; Snijders & Bosker, 2012). MLM explicitly models variability as statistical variance at each level simultaneously, whilst appropriately modeling explanatory variables at each level (for example, accounting for patient case mix) and any interactions between levels (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Such methods require large samples of patients, therapists and clinics, which are more likely to come from naturalistic settings than from randomized controlled trials (Elkin, 1999).

The aims of the current study were two-fold. First, to estimate the size of therapist and clinic effects in a heterogeneous, naturalistic sample of patients receiving psychological therapies. Second, to assess the impact on outcome of the relationships between patient demographic and process variables and the variability between therapists and between clinics. We hypothesized that a significant clinic effect would be detected despite controlling for therapist variability and patient variables. We expected that patient variables (particularly symptom severity) might partly explain clinic effects, whether at the patient level or in aggregate, but had no clear hypotheses regarding the extent to which this would occur.

# Method

## Study dataset

The study sample was drawn from the CORE National Research Database 2011 (see Stiles, Barkham, & Wheeler, 2015). The initial database comprised 104,474 patients seen by 2,442 therapists at 52 psychological therapy clinics across the United Kingdom (UK). Ethical approval was covered by National Research Ethics Service application 05/Q1206/128 (amendment 3). The therapists were counselors, psychotherapists and clinical psychologists. Individual therapist characteristics data were unavailable in this database. The most common psychological intervention models delivered to patients included person-centered, psychodynamic, cognitive behavioral and supportive therapies. The mean unplanned ending rate per therapist was 33.0% (SD = 28.4), and per site was 33.5% (SD = 11.4).

Included clinics were from predominantly urban areas and ranged across five sectors of care provision: primary care, secondary care, university, voluntary, or workplace. All patients from two additional sectors (tertiary and private) were excluded in the process of applying the exclusion criteria below. Contributing factors included exceedingly high percentages of missing data in the tertiary sector (95% of patients had required data missing), and relatively low initial patient numbers in the private sector (n = 442).

In the UK mental health services delivery system, primary, secondary, and tertiary care are typically offered in separate National Health Service (NHS) settings within a region. The primary care sector, which includes community health centers and general practitioner clinics (family practices) offering predominantly short-term counseling, is usually the first point of contact. Secondary and tertiary clinics are more specialized, provide longer-term psychotherapy, and usually require referrals from a primary care clinic. Patients can also access therapy through voluntary organizations and charities, in university and workplace counseling centers, or in private practices. For patients, mental health services are typically free at the point of delivery except in the private sector.

In selecting patients for analysis, patient inclusion criteria were applied first, followed by therapist and clinic inclusion criteria. These criteria aimed to produce a sample that would provide adequate sample sizes of patients, therapists and clinics to produce robust estimates of effects at each level (Schiefele et al., 2016).

Patient inclusion criteria were as follows: 1) patient age was between 16 and 95 years old; 2) patients received an individual intervention (rather than a group, family, or couple intervention); 3) patient ethnic origin, employment status, and attendance data were recorded; and 4) valid pre- and post-treatment outcome measure scores were recorded. Therapist and clinic inclusion criteria were as follows: 1) each therapist worked with 10 or more eligible patients, and 2) each clinic included 5 or more eligible therapists. These criteria reduced the original sample of 104,474 patients to 26,888 as shown in Figure 1.

The final sample of 26,888 patients had a mean (SD) age of 38.4 (12.94) years and 69.3% were female. They were seen by 462 therapists, across 30 clinics with a mean (SD) number of patients per therapist of 58.2 (71.4) and mean (SD) number of therapists per clinic of 15.4 (12.4). In this study sample, the mean recorded rate of unplanned endings per therapist was 9.0% (SD = 8.8), and per site was 9.2% (SD = 6.9).

**Measures**

**Clinical Outcomes in Routine Evaluation Outcome Measure.** (CORE-OM). The CORE-OM (Barkham et al., 2001; Evans et al., 2002) is a 34-item measure of psychological distress. Items assess the following domains: symptoms (depression, anxiety, physical problems, and trauma), functioning (general functioning, and in close and social relationships), subjective wellbeing, and risk (to self and others). Each item is scored on a 5-point scale from 0-4. Items are anchored as follows: not at all, only occasionally, sometimes, often, and all or most of the time. Item scores are averaged and multiplied by 10 to produce a full measure clinical score of 0-40 with higher scores indicating greater distress. A clinical cut-off score of 10 has been found to optimally discriminate clinical and non-clinical samples (Connell et al., 2007). The CORE-OM demonstrates internal consistency of $\alpha$ = .93-.95 (Barkham, Gilbert, Connell, Marshall, & Twigg, 2005), test-retest reliability of .88 at one-month intervals (Barkham, Mullin, Leach, Stiles, & Lucock, 2007), and strong convergent validity with measures such as the Beck Depression Inventory (BDI-II) and Clinical Interview Scale – Revised (CIS-R) (Cahill et al., 2006; Connell et al., 2007). In this study the CORE-OM was administered prior to the first therapy session and following the last therapy session as part of routine practice at all clinics. As it was not collected every therapy session, patients who dropped out of treatment rarely had a measure for their last session attended.

Although the CORE-OM score at the last session was the study outcome, comparisons of effectiveness also considered statistically reliable, clinically significant pre-post change using the methods described by Jacobson and Truax (1991). A pre-post change of 5 points or more on the CORE-OM was taken as statistically reliable change, while change from above the clinical cut-off of 10 to below was clinically significant (Connell et al., 2007). Patients who met both of these criteria at outcome were considered statistically recovered.

**The CORE Assessment form and CORE End of Therapy form**. These forms are completed by therapists at intake and the end of therapy respectively. They record referral information, patient demographics (gender, age, employment status, ethnic origin), data on the nature, severity and duration of presenting problems, the number of sessions the patient attended, whether the ending was planned or unplanned, and which type(s) of therapy the patient received (Mellor-Clark, Barkham, Connell, & Evans, 1999). Ethnic origin, which became important in our analyses, was recorded in nine categories originally drawn from a UK government list (https://www.ons.gov.uk/methodology/classificationsandstandards/measuringequality/ethnicgroupnationalidentityandreligion#ethnic-group): 1 Asian (Bangladeshi); 2 Asian (Indian); 3 Asian (Pakistani); 4 Asian (E. African); 5 Asian (Chinese); 6 Black (African) ; 7 Black (Caribbean) ; 8 White (English/European) ; and 9 Other.

### Study sample characteristics

No formal diagnoses were available in the dataset. However, most patients were reported by therapists to be experiencing anxiety (71.8%) and/or depression (54.0%), with 14.6% and 7.3% respectively at a severe level. Incidence of depression was broadly similar across sectors, but the secondary care sector had the highest proportion at the severe level (12.0%). Secondary care was also characterized by a much larger proportion of patients with personality problems (37.2% with 7.9% at a severe level) compared to other sectors which ranged from 1.6% to 3.5%, with between 0.2% and 0.4% at a severe level.

Overall, the mean (SD) CORE-OM score pre-therapy was 17.8 (6.25); scores ranged from 17.2 (6.53) in the voluntary sector to 21.1 (7.05) in secondary care. Similarly, the proportion meeting criteria for clinical distress ranged from 85.9% in the voluntary sector to 94.1% in the secondary care sector with an overall rate of 89.4%.

Table 1 describes the study sample and shows comparisons between included and excluded patients. Effect sizes are also shown. Included patients were on average older, more likely to be female, White, and employed, had lower pre- and post-therapy outcome scores, attended more sessions and a higher proportion of offered sessions, and were more likely to have a planned ending (all p-values <0.001; all significant after Bonferroni correction). Patients of therapists excluded due to having fewer than 10 patients with eligible data (Figure 1) had significantly less pre-post improvement in CORE-OM scores ($M = 8.3$, $SD = 6.9$, $n = 3,326$) compared with patients of therapists with 10 or more eligible patients ($M = 9.0$, $SD = 6.7$, $n = 28,147$), $t(4100.0) = -5.7$, $p < .001$. Note that completion of treatment was not required for inclusion; however, completion of both pre- and post-treatment measures was required. Because relatively few patients who dropped out of treatment completed all measures, most (92%) of the included clients were treatment completers.

**Explanatory variables**

Patient level variables were: pre-treatment CORE-OM score, age, employment status, ethnic origin, sessions planned, sessions attended, and percentage of planned sessions attended. Two patient demographic variables, employment status and ethnic origin, were collapsed due to small numbers in some categories. Employment status was coded as: employed (comprising part-time employment, full-time employment), not employed (comprising receiving welfare benefits, unemployed, retired), or other role (comprising part-time student, full-time student, houseperson, other). Also, as 87% of the population of the UK is White (Office for National Statistics, 2011), with the remaining 13% representing a number of non-White race/ethnicity groups, non-White groups were collapsed into a single category. Using the UK ethnic origin categories on the CORE Assessment form (see list shown earlier; Evans et al., 2002), ethnic origin was recoded as White (i.e., category 8 White [English/European]), which accounted for 90.2% of the patients, versus Non-white (all other categories).

Therapist-level and clinic-level aggregates of patient level variables were derived from the original database of all patients (N = 104,474). This produced therapist and clinic level explanatory variables to represent the composition of each therapist's caseload and each clinic's clinical population. Therefore, in total, there were 22 potential explanatory variables as follows: Seven patient-level variables, an aggregate of each patient-level variable at the therapist and clinic level (14 total), and a variable specifying the sector of the clinic.

**Multilevel modeling analysis**

The primary analysis comprised multilevel modeling using the MLwiN software (Rasbash, Charlton, Browne, Healy, & Cameron, 2016). The post-therapy CORE-OM score was the dependent variable. This and the pre-therapy CORE-OM scores were log-transformed to correct issues of heteroskedasticity. Models were developed from a single level model to two- and then three-level models with each development tested for significance by comparing the reduction in the -2*loglikelihood value against the chi-square critical value for the additional degrees of freedom. Significant reductions indicate improvements in model fit. Significance of random effects additionally required reductions in deviance information criterion (DIC) values dervied from Markov Chain Monte Carlo (MCMC) simulation.

Due to the large number of explanatory variables, a Bonferroni correction was applied to z-score testing of model coefficients for all main effect, interaction, and random slope tests in order to correct the family-wise error rate. The Bonferroni correction was calculated a priori on the basis of the number of explanatory variables (22). Thus, the resulting (more conservative) z-score critical value of 3.06 was used (corresponding to a per-observation p-value of .0023, or 99.8% confidence).

Patient-level, therapist-level, and clinic-level variables were tested for significance, followed by interactions between significant variables, and finally random slopes at both the therapist and clinic level. Each significant variable was then tested in isolation within a three-level model and variables ranked according to the overall unexplained variance that each variable explained, and a final model was then reconstructed by adding variables in order from highest to lowest variance explained. The therapist effect and clinic effect were recalculated after each variable was added. As is standard, these effects were defined as the percentages of overall unexplained (or residual) variance associated with the therapist level or clinic level respectively, and are akin to intra-class correlation coefficients (Rasbash, Steele, Browne, & Goldstein, 2012). Clinic and therapist effects represent the degree to which the variability between clinics and the variability between therapists in a clinic are associated with patient outcomes.

Iterative Generalised Least Squares (IGLS) estimation was used in the construction of each model as variables were added and tested. Markov Chain Monte Carlo (MCMC) simulation procedures were then applied to a) the 1-level, 2-level, and 3-level model with no explanatory variables, b) the final 3-level model after inclusion of all significant variables, and c) two 2-level variants of the final model, in order to compare variance distributions and calculate DIC values. MCMC simulation used parameter estimates produced by IGLS as 'priors' to produce a chain of parameter estimates from which medians and means could be derived. In addition, MCMC allowed for the calculation of 95% probability intervals (PrIs) around estimates of effects. These are similar to confidence intervals and represent the 2.5

and 97.5 percentile values in the simulation chain (Browne, 2016).

A sensitivity analysis was carried out using a modified sample, including only clinics with 10 or more therapists per clinic. This reduced the potential bias due to therapist outliers, but also reduced the number of clinics in the sample and confidence in clinic-level estimates.

## Results

Initial analysis considered treatment effectiveness assessed by outcome scores and recovery rates. The results are presented overall, by therapists, and by clinics. These are followed by the development of the multilevel model, the identification of significant explanatory variables and the therapist and clinic effect estimations. Finally, the relationships between variables at the different levels in the model are considered in more detail.

### Outcomes

For patients, the mean (SD) CORE-OM outcome score was 8.8 (6.33) with a mean (SD) pre- to post-therapy change of 9.0 (SD = 6.69). Of patients scoring above the clinical cut-off (i.e., CORE-OM score ≥10 or more) at pre-therapy (n = 24,027), 58.4% met the criteria for reliable and clinically significant improvement (RCSI).

The mean (SD) RCSI rate for clinics was 57.5% (13.02) with a range across the 30 clinics of 23.4% - 75.2%. The mean (SD) rate for therapists was 57.3% (17.00), with a range across the 462 therapists of 6.7% - 100%, although the number of patients from which these were derived was small in many cases. Considering only therapists who saw 50 or more patients (n = 129), the range was 15.5% to 91.1%.

### Multilevel model development

A single, patient-level outcome model was compared with a 2-level model, with patients at level 1, a random intercept for therapists at level 2, and no explanatory variables. The 2-level model showed a significant improvement in model fit ($\chi^2(1) = 1801.51$; $p < 0.001$; reduction in DIC = 2283.34). A random intercept for clinics added at level 3 to produce a simple 3-level model was a further significant improvement ($\chi^2(1) = 216.99$; p < 0.001; reduction in DIC = 76.92).

Potential predictor variables, interactions, and random slopes were then tested for significance to produce a final model containing the statistically significant variables (full model specification available online as Supplemental Material). Following the inclusion of significant fixed effect variables, the multilevel model fit was re-tested. Recognizing the therapist level and clinic level again improved model fit, with significant reductions in -2*loglikelihood values ($\chi^2(1) = 870.66$; p < 0.001 and $\chi^2(1) = 52.93$; p < 0.001 respectively).

Of seven random slopes tested, slopes for intake severity and attendance at the therapist level also improved model fit ($\chi^2(2) = 24.64$; p < 0.001; reduction in DIC = 32.47, and $\chi^2(3) = 25.17$; p < 0.001; reduction in DIC = 37.24 respectively), indicating that the relations between intake severity and outcome, and attendance and outcome varied between therapists. The positive covariances with outcome for these two variables (0.004 and 0.003 respectively; see Supplementary Material) indicate that the variability between therapists increased as patient severity increased and attendance increased. However, for the latter, the standard error indicates uncertainty regarding the extent of the differences in the slopes. There were no significant random slopes for the clinic level. The final model, including significant explanatory variables, reduced the overall unexplained variance in the original 3-level empty model by 23.2%. Model assumptions of homoscedasticity and Normality of residuals at each level were tested and met. MCMC simulation indicated 40,000 iterations were sufficient for the convergence of model estimates.

**Explanatory variables**

Of the seven patient level variables, intake severity, employment status, ethnic origin, age, and attendance were significant predictors of outcome (all p<.001), while sessions planned and sessions attended were not significant. Greater intake severity, not being employed, being of non-White ethnic origin, being older and attending a lower percentage of planned sessions were associated with poorer outcome. Of the ten interactions between the five significant patient variables, only the interaction between severity and employment status was significant ($p<.001$). The effect of severity on outcome was greater for patients who were not employed, compared with patients who were employed.

For sector, a clinic level variable, there were no significant differences between the coefficients for primary care, university, voluntary, and workplace clinics, therefore they were collapsed into one category and compared to secondary care clinics. The model results indicated that secondary clinics were associated with poorer outcomes ($p<.001$).

No variables aggregated at the therapist level were significant in the model. However, patient ethnic origin aggregated at the clinic level, to represent the percentage of patients treated at the clinic who were White, was associated with outcome ($p<.001$). A larger proportion of White patients in a clinic population was associated with improved clinic outcomes for both White and non-White patients compared to clinics with a lower proportion of White patients. This was in addition to the effect of individual patient ethnic origin. All ten cross-level interactions and the one clinic-level interaction tested were all non-significant.

**Clinic and therapist effects**

The final 3-level model produced a clinic effect of 1.9% (PrI = 0.8% – 3.7%) and a therapist effect of 3.4% (PrI = 2.7% – 4.2%) after controlling for fixed effect explanatory variables. In order to assess how the recognition of a third level affected these effects, two 2-level models were constructed (patient/clinic and patient/therapist respectively). In these models, a clinic effect of 2.8% (PrI = 1.5% – 4.8%) and a therapist effect of 4.9% (PrI = 4.0% - 5.9%) were found, indicating that recognition of the third level reduced each effect by about 30%.

Figure 2 describes the variability between clinics and therapists in the final 3-level model by plotting their model residuals with 95% confidence intervals (CIs). These represent the unexplained outcome variability associated with individual clinics and individual therapists. More effective clinics and therapists are shown on the left of the charts, indicating a larger reduction in the log-transformed outcome scores.

In Figure 2, the dashed lines, where the residuals are zero, represent the average clinic or therapist, and only those clinics or therapists whose 95% CIs do not cross zero can be considered significantly different from average. Figure 2 indicates 4 (13.3%) clinics were more effective than average while 2 (6.7%) clinics were less effective than average and the difference between these two groups of clinics was significant as their 95% CIs did not overlap. The patient recovery rate for the more effective clinics was 69.7%, while the rate for patients seen at the less effective clinics was 48.5%. The recovery rate for patients seen at the majority of clinics, considered to be average, was 55.7%. Similarly, 18 (3.9%) therapists were more effective than average with an overall patient recovery rate of 77.2% compared with a rate of 41.4% for the 18 (3.9%) therapists who we less effective. The recovery rate for the 426 (92.2%) average therapists was 58.0%.

A sensitivity analysis was conducted on a sub-sample that required 10 or more therapists per clinic, compared with 5 or more therapists per clinic in the main analysis. This sub-sample comprised 22,535 patients, seen by 394 therapists, across 19 clinics. Results were generally comparable. All significant explanatory variables in the main analysis remained significant in the sensitivity analysis and the final clinic effect (1.7%, PrI = 0.6% - 3.8%) and therapist effect (3.6%, PrI = 2.8% - 4.6%) also approximated those in the main analysis.

**Explaining the clinic effect**

During model development, the 3-level model with no explanatory variables indicated a clinic effect of 8.2% (PrI= 4.8% – 13.6%), significantly larger than the therapist effect of 3.2% (PrI = 2.5% - 4.0%). However, as noted above, in the final model, the clinic effect was much reduced to 1.9%, indicating that the variables in the model explained 76.8% of the initial clinic effect. In contrast, the therapist effect in the final model had changed little, at 3.4% (Figure 3).

Figure 4 describes the changes in clinic and therapist effects as each variable, interaction, and random slope was added to the model. These were added in order of the percentage of the overall residual variance each explained when added in isolation to the empty 3-level model. The effects in Figure 4 were estimated using IGLS, which produced slightly smaller estimates than MCMC. Figure 4 also shows that patient severity, patient employment status, sector of the clinic, and the proportion of White patients in the clinic population explained most of the variability between clinics. Table 2 shows the percentages of residual variance from the empty 3-level model explained as each fixed effect variable was added.

Table 2 shows that patient severity explained 29.6% of the initial clinic level variance, as well as the largest proportion of variance at the therapist level (12.4%) and patient level (16.2%). Other variables explained little additional variance at the therapist and patient levels, but patient employment status (16.4%), clinic sector (16.1%), and the proportion of White patients in the clinic population (19.6%) explained considerable amounts of clinic level variance. The remaining variables explained around 1% of clinic level variance.

**Ethnic origin and outcomes**

Figure 4 and Table 2 indicate that in addition to the proportion of White patients in the clinic population being a significant predictor of outcome, it also explained a large amount of the variance at the clinic level. To describe how this variable and the patient level ethnic origin variable predict patient outcomes, Figure 5 plots the predicted CORE-OM outcome scores (not log-transformed) for White and non-White patients treated at clinics with different percentages of White patients in their populations. The lines represent predicted outcome scores for patients with mean or reference category values on other variables in the model.

Figure 5 shows that White patients on average had better outcomes than non-White patients regardless of the ethnic composition of the clinic population, and that outcomes improve for both White and non-White patients at clinics with higher proportions of White patients. In addition, Figure 5 also suggests that the difference between outcomes for White and non-White patients increased at clinics with higher proportions of White patients. However, this difference between around 0.25 of a point on CORE-OM for populations with the smallest proportions of White patients and around one point for the populations with the largest proportions of White patients (see Figure 5), was not significant in the final model. The removal from the model of either of the ethnic origin variables made little or no difference to the model coefficient of the other remaining ethnic origin variable.

# Discussion

To our knowledge, this study is the first to use a 3-level multilevel model comprising clinics, therapists, and patients to model patient outcomes in psychological therapy. Our results indicate that modeling a hierarchical, nested structure in the data produced the best fit, and was able to identify predictors of outcome and estimate the size of clinic and therapist effects in a single model. The model was also able to describe the relationships and interactions between different factors across the three levels, and how these contributed in different ways to the variability in patient outcomes.

## Patient variables

Supporting findings from previous MLM studies, the results show that most of the variability in patient outcomes was associated with differences between patients (e.g., Saxon & Barkham, 2012; Wampold & Brown, 2005). Also, the patient variables associated with outcome and explaining much of that variability (in particular intake severity, as well as employment status, ethnic origin, age and sessions attended) have been identified previously (e.g., Firth et al., 2015; Garfield, 1994). However, in the current study, the relations between outcomes and these variables have been estimated while also controlling for the relations with both clinics and therapists.

Cross-level interactions were also found. The significant effects of intake severity and attendance on outcome were found to be moderated by the therapist that the patient saw. In contrast, the effects of patient employment status, age and ethnic origin were the same regardless of the therapist. These findings are perhaps unsurprising as patient symptom severity is more directly linked to the therapeutic process and arguably by extension, the therapist. The relations between the patient variables and patient outcome were similar for all clinics; that is, there was no significant effect of random slopes. However, patient ethnic origin, aggregated to represent the ethnic composition of the clinic population, was found to explain a considerable amount of the variability between clinic outcomes.

**Therapist effects**

The therapist effect of 3.4% was smaller than the effects most commonly found of between 5% and 10% (Baldwin & Imel, 2013). However, the current study also included the clinic level in the model; excluding the clinic level produced a therapist effect of 4.9%, more similar to published effects (Johns et al., 2018). Thus, one potential reason for larger effects found elsewhere is that they may incorporate an unrecognized clinic effect.

The size of the therapist effect may appear small, but it was both statistically significant and clinically significant. Patients were almost twice as likely to recover if seen by above average therapists compared with below average therapists. In addition, as noted above, the effect of random therapist slopes for patient intake severity and attendance indicate that there was greater variability between therapists the more severe the patient’s condition. This finding is consistent with other large studies of therapist effects (e.g., Berglar et al., 2016; Schiefele et al., 2016) and indicates that more effective therapists are particularly more effective in the treatment of more severe patients.

Unlike the clinic effect, the therapist effect remained consistent after controlling for variables. We found therapist caseload mix, represented by patient variables aggregated at the therapist level, was not associated with outcome. However, therapists' personal qualities may have more impact, as indicated by recent research (e.g. Green et al., 2014; Goldberg et al., 2016; Nissen-Lie et al., 2015; Wampold et al., 2017).

**Clinic effects**

We know of only one previous study that has estimated the size of clinic effects using MLM, finding a smaller but broadly comparable effect of 1.8%, in a 2-level model using a different outcome measure and dataset (Pybis et al., 2017). Approximately half of the unadjusted clinic effect was explained by patient-level severity and employment status, suggesting a selection effect. This has important implications for healthcare providers using pay for performance (a.k.a. outcomes-based) payment models. Our study found that a considerable additional amount of variability between clinics was explained by two clinic level variables. These clinic level variables were the clinic sector and the percentage of a clinic's population who were White English/European. Compared to other sectors, treatment in a secondary care clinic was associated with poorer outcomes. Secondary care clinics tend to work with patients with more complex or treatment-resistant difficulties that may not have been fully captured in the available variables. The second clinic level variable, the percentage of a clinic's population that were White English/European, was a more surprising finding, particularly as it was in addition to an individual patient's ethnic origin and explained more of the outcome variance.

**Ethnic origin, deprivation, and location**

The finding that larger proportions of ethnic minority patients in the clinic population was associated with poorer outcomes for all patients accessing that clinic (after accounting for individual ethnic origin), implies that population ethnic composition reflects a distinct underlying factor in the population. Factors such as racism, implicit bias, and micro-aggressions may have contributed to this effect. However, the minimal association between individual ethnicity and outcomes, combined with a much stronger population effect impacting equally on White and minority patients, arguably challenges this hypothesis. Further research into these factors may be beneficial to clarify any such contribution.

Other possible explanations may be factors associated with deprivation. There is consistent evidence linking minority status and deprivation (Jivraj & Khan, 2013; Platt, 2007; United States Census Bureau, 2013), particularly in more urban areas, as in the current sample (Aldridge, Parekh, MacInnes, & Kenway, 2011). Analysis of 2011 UK Census data indicates that ethnic minorities are more likely to live in the most deprived areas. For example, 37% of the UK Bangladeshi population and 20% of the Caribbean population live in areas in the top decile of multiple deprivation, while around 7% of the White British population live in those areas (Jivraj & Khan, 2013).

It might be hypothesized, therefore, that the clinic ethnic composition was a good proxy measure of community deprivation. This would support a recent UK study of national primary care data that showed locality deprivation to be associated with patient outcomes (Clark et al., 2018). Therefore, clinics with relatively poorer outcomes may have served relatively more deprived communities, and the clinic effect detected in this study may in part reflect a neighborhood or locality effect. An alternate, but complementary, hypothesis is that deprivation may be impacting on provision of care due to reduced funding and resources. This would be an example of the inverse care law (Hart, 1971).

One implication of our findings is that comparisons of clinic effectiveness can only be fair if the characteristics of their patient populations are taken into consideration in the analysis. Failure to do so may result in clinics in more deprived areas or with more difficult to treat patients being penalized. Further studies are required that include therapist and clinic variables along with socioeconomic and geographic variables to tease apart their unique contributions to patient outcome. These should include variables such as income per patient, measures of neighborhood or area-level deprivation, and the racial and ethnic characteristics of clinicians and clinic staff or the presence of cultural competence training.

**Caveats**

The findings above come with a number of caveats, most of which concern the data sample. The disadvantages (and advantages) of using routinely collected data for research purposes have been well documented (e.g., Barkham, Stiles, Lambert, & Mellor-Clark, 2010). However, as it is routine data that is used administratively to monitor and compare clinic effectiveness, there is a strong argument for using this same data to study the variability in outcomes in clinics. In the current study, although the large sample allowed for multilevel analysis, the lack of therapist and clinic variables – a limitation of UK service datasets generally – limited our understanding of the possible reasons for the variability found.

Despite the large sample, wide CIs in the caterpillar plots and PrIs for the clinic effects in particular indicate a degree of uncertainty regarding some of the findings. Future studies with larger clusters might produce more robust estimates for model parameters.

The final caveats concern the generalizability of findings. They came from a heterogeneous sample of UK clinics and may not be generalizable to clinics in other countries with different configurations of provision and different clinic population characteristics. Also, as outcome measures came mainly from patients who completed therapy, the results may only be generalizable to therapy completer samples. Naturalistic data, including the present data are not well-suited to intention to treat analyses, as a “first-observation-carried-forward” approach would have considerably reduced variability due to large numbers of patients showing no apparent pre-post change (Barkham, Stiles, Connell, & Mellor-Clark, 2012). Large multi-clinic datasets containing sessional outcome measures are greatly needed in order that analyses can appropriately model non-completion.

## Conclusion

Our primary aim was to estimate the size of therapist and clinic effects using a three-level model. Confirming previous findings, patient outcomes varied systematically across both therapists and clinics, with patient severity being the variable most strongly associated with outcome variability. However, the overarching implication of our findings is that the effectiveness of therapy is not restricted to the therapist-patient interaction, and that the broader sociodemographic, socioeconomic, and geographic context in which the patient lives and in which the therapy is provided may substantially contribute to patient outcome.

**References**

Aldridge, H., Parekh, A., MacInnes, T., & Kenway, P. (2011). Monitoring poverty and social exclusion 2011: Joseph Rowntree Foundation.

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), Bergin and Garfield's handbook of psychotherapy and behavior change (6th ed., pp. 258-297). New York, NY: Wiley.

Barkham, M., Delgadillo, J., Firth, N., & Saxon, D. (2018). Practice-based evidence and the law of variability in psychological treatment. Revista Argentina de Clinica Psicologica, 27, 115-135.

Barkham, M., Gilbert, N., Connell, J., Marshall, C., & Twigg, E. (2005). Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings. British Journal of Psychiatry, 186, 239-246. https://doi.org/10.1192/bjp.186.3.239

Barkham, M., Lutz, W., Lambert, M. J., & Saxon, D. (2017). Therapist effects, effective therapists, and the law of variability. In L. G. Castonguay & C. E. Hill (Eds.), Therapist effects: Toward understanding how and why some therapists are better than others (pp. 13-36). Washington, DC: American Psychological Association. https://doi.org/10.1037/0000034-002

Barkham, M., Mullin, T., Leach, C., Stiles, W. B., & Lucock, M. (2007). Stability of the CORE-OM and the BDI-I prior to therapy: Evidence from routine practice. Psychology and Psychotherapy-Theory Research and Practice, 80, 269-278. https://doi.org/10.1348/147608306X148048

Barkham, M., Stiles, W.B., Connell, J., & Mellor-Clark, J. (2012). Psychological treatment outcomes in routine NHS services: What do we mean by treatment effectiveness? Psychology and Psychotherapy: Theory, Research and Practice, 85, 1-16.

Barkham, M., Stiles, W. B., Lambert, M. J., & Mellor-Clark, J. (2010). Building a rigorous and relevant knowledge base for the psychological therapies. In M. Barkham, G. E. Hardy & J. Mellor-Clark (Eds.), Developing and delivering practice-based evidence: A guide for the psychological therapies. Chichester, UK: Wiley-Blackwell. https://doi.org/10.1002/9780470687994.ch2

Berglar, J., Crameri, A., von Wyl, A., Koemeda-Lutz, M., Köhler, M., Staczan, P., … Tschuschke, V. (2016). Therapist effects on treatment outcome in psychotherapy: A multilevel modelling analysis. International Journal of Psychotherapy, 20, 61-80. Retrieved from http://www.ijp.org.uk/

Barrett, M. S., Chua, W. J., Crits-Christoph, P., Gibbons, M. B., & Thompson, D. (2008). Early withdrawal from mental health treatment: Implications for psychotherapy practice. Psychotherapy, 45, 247-267. https://doi.org/10.1037/0033-3204.45.2.247

Bohart, A. C., & Greaves Wade, A. (2013). The client in psychotherapy. In M. J. Lambert (Ed.), Bergin and Garfield's handbook of psychotherapy and behavior change (6th ed., pp. 219-257). New York, NY: Wiley.

Browne, W. J. (2016). MCMC Estimation in MLwiN: Centre for Multilevel Modelling, University of Bristol.

Cahill, J., Barkham, M., Stiles, W. B., Twigg, E., Hardy, G. E., Rees, A., & Evans, C.. (2006). Convergent validity of the CORE measures with measures of depression for clients in cognitive therapy for depression. Journal of Counseling Psychology, 53, 253-259. https://doi.org/10.1037/0022-0167.53.2.253

Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., & Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. Lancet, 391(10121), 679-686. https://doi.org/10.1016/S0140-6736(17)32133-5

Connell, J., Barkham, M., Stiles, W. B., Twigg, E., Singleton, N., Evans, O., & Miles, J.N.V. (2007). Distribution of CORE-OM scores in a general population, clinical cut-off points and comparison with the CIS-R. British Journal of Psychiatry, 190, 69-74. https://doi.org/10.1192/bjp.bp.105.017657

Delgadillo, J., Asaria, M., Ali, S., & Gilbody, S. (2016). On poverty, politics and psychology: the socioeconomic gradient of mental healthcare utilisation and outcomes. British Journal of Psychiatry, 209, 431-432. https://doi.org/10.1192/bjp.bp.115.171017

Elkin, I. (1999). A major dilemma in psychotherapy outcome research: Disentangling therapists from therapies. Clinical Psychology-Science and Practice, 6, 10-32. https://doi.org/10.1093/clipsy.6.1.10

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: psychometric properties and utility of the CORE-OM. British Journal of Psychiatry, 180, 51-60. https://doi.org/10.1192/bjp.180.1.51

Firth, N., Barkham, M., Kellett, S., & Saxon, D. (2015). Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis. Behaviour Research and Therapy, 69, 54-62. https://doi.org/10.1016/j.brat.2015.04.001

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), Handbook of psychotherapy and behavior change (4th ed.). New York: Wiley.

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., … Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. Journal of Counseling Psychology, 63, 1-11. https://doi.org/10.1037/cou0000131

Goldstein, H. (2010). Multilevel statistical models (4th Ed. ed.). Sussex, UK: Wiley. https://doi.org/10.1002/9780470973394

Green, H., Barkham, M., Kellett, S., & Saxon, D. (2014). Therapist effects and IAPT Psychological Wellbeing Practitioners (PWPs): A multilevel modelling and mixed methods analysis. Behaviour Research and Therapy, 63, 43-54. https://doi.org/10.1016/j.brat.2014.08.009

Hamilton, K. E., & Dobson, K. S. (2002). Cognitive therapy of depression: Pretreatment patient predictors of outcome. Clinical Psychology Review, 22, 875-893. https://doi.org/10.1016/S0272-7358(02)00106-X

Hart, J. T. (1971). The inverse care law. The Lancet, 297, 405-412. https://doi.org/10.1016/S0140-6736(71)92410-X

Jacobson, N. S., & Truax, P. (1991). Clinical significance - A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology, 59, 12-19. https://doi.org/10.1037/0022-006X.59.1.12

Jivraj, S., & Khan, O. (2013). Ethnicity and deprivation in England: How likely are ethnic minorities to live in deprived neighbourhoods? : Centre on Dynamics of Ethnicity, University of Manchester.

Johns, R., Barkham, M., Kellett, S., & Saxon, D. (2018). A systematic review of therapist effects: A critical narrative update and refinement to Baldwin and Imel's (2013) review. Clinical Psychology Review. Review. https://doi.org/10.1016/j.cpr.2018.08.004

Mellor-Clark, J., Barkham, M., Connell, J., & Evans, C. (1999). Practice-based evidence and standardized evaluation: Informing the design of the CORE system. European Journal of Psychotherapy & Counselling, 2, 357-374. https://doi.org/10.1080/13642539908400818

Nissen-Lie, H. A., Rønnestad, M. H., Høglend, P. A., Havik, O. E., Solbakken, O. A., Stiles, T. C., ... Monsen, J. T. (2015). Love yourself as a person, doubt yourself as a therapist? Clinical Psychology & Psychotherapy, 24, 48-60. https://doi.org/10.1002/cpp.1977

Office for National Statistics. (2011). 2011 Census: Population estimates for the United Kingdom.

Pickett, K. E., & Pearl, M. (2001). Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. Journal of Epidemiology and Community Health, 55, 111-122. https://doi.org/10.1136/jech.55.2.111

Platt, L. (2007). Poverty and ethnicity in the UK. Bristol, UK: Policy Press.

Pybis, J., Saxon, D., Hill, A., & Barkham, M. (2017). The comparative effectiveness and efficiency of cognitive behaviour therapy and generic counselling in the treatment of depression: evidence from the 2nd UK National Audit of psychological therapies. BMC Psychiatry, 17, 13. https://doi.org/10.1186/s12888-017-1370-7

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2016). MLwiN (Version 2.36): Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W., & Bryk, A. S. (Eds.). (2002). Hierarchical linear models: applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage Publications.

Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. Journal of Consulting and Clinical Psychology, 80, 535-546. https://doi.org/10.1037/a0028898

Saxon, D., Firth, N., & Barkham, M. (2017). The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage, and non-completion. Administration and Policy in Mental Health and Mental Health Services Research, 44, 705-715. https://doi.org/10.1007/s10488-016-0750-5

Schiefele, A.-K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J…Lambert, M. J. (2016). Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies. Administration and Policy in Mental Health and Mental Health Services Research, 44, 598-613. https://doi.org/10.1007/s10488-016-0736-3

Snijders, T. A. B., & Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.). London: Sage Publishers.

Stiles, W. B., Barkham, M., & Wheeler, S. (2015). Effect of duration of psychological therapy on recovery and improvement rates: evidence from UK routine practice. British Journal of Psychiatry, 207, 115-122. https://doi.org/10.1192/bjp.bp.114.145565

Stulz, N., Lutz, W., Kopta, S. M., Minami, T., & Saunders, S. M. (2013). Dose-effect relationship in routine outpatient psychotherapy: Does treatment duration matter? Journal of Counseling Psychology, 60, 593-600. https://doi.org/10.1037/a0033589

United States Census Bureau. (2013). Poverty rates for selected detailed race and hispanic groups by state and place: 2007–2011: U.S. Department of Commerce.

Vittengl, J. R., Jarrett, R. B., Weitz, E., Hollon, S. D., Twisk, J., Cristea, I… Cuijpers, P. (2016). Divergent outcomes in cognitive-behavioral therapy and pharmacotherapy for adult depression. American Journal of Psychiatry, 173, 481-490. https://doi.org/10.1176/appi.ajp.2015.15040492

Wampold, B. E., Baldwin, S. A., grosse Holtforth, M., & Imel, Z. E. (2017). What characterizes effective therapists? In L. G. Castonguay & C. E. Hill (Eds.), Therapist effects: Toward understanding how and why some therapists are better than others (pp. 37-53). Washington: American Psychological Association.

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. Journal of Consulting and Clinical Psychology, 73, 914-923. https://doi.org/10.1037/0022-006X.73.5.914

Table 1

Comparisons between included and excluded patients

| Variable | Included patients (n = 26888) | Excluded patients (n = 77586) | Included/excluded difference | Effect size |
|---|---|---|---|---|
| Mean patient age (SD) | 38.4 (12.9) | 34.8[b] (13.2) | t(49127.5) = -38.3* | Cohen's d = 0.28 |
| Female | 69.3% | 66.0%[c] | $\chi^2$(1, N = 103082) = 97.8* | OR = 1.16 |
| White | 90.2% | 84.3%[d] | $\chi^2$(1, N = 92877) = 541.8* | OR = 1.71 |
| Employment status[a] | | | $\chi^2$(2, N = 100084) = 3361.8* | |
| Employed | 60.4% | 40.1%[e] | | OR = 2.28 |
| Other role | 25.6% | 42.0%[e] | | OR = 0.48 |
| Not employed | 14.0% | 17.8%[e] | | OR = 0.75 |
| Mean pre-therapy CORE-OM (SD) | 17.8 (6.2) | 18.3 (6.8)[f] | t(56429.1) = 9.4* | Cohen's d = 0.08 |
| Mean post-therapy CORE-OM (SD) | 8.8 (6.3) | 9.8 (7.0)[g] | t(17454.0) = 12.7* | Cohen's d = 0.15 |
| Mean therapy sessions (SD) | 8.1 (9.0) | 7.8 (13.9)[h] | t(61566.7) = -3.1* | Cohen's d = 0.03 |
| Mean session attendance (SD) | 90.2% (15.5) | 76.3% (26.2)[h] | t(59958.3) = -82.9* | Cohen's d = 0.65 |
| Planned ending | 92.0%[i] | 49.9%[j] | $\chi^2$(1, N = 63543) = 12580.6* | OR = 11.55 |

CORE-OM = Clinical Outcomes in Routine Evaluation – Outcome Measure.

[a]employment status categories defined as follows: employed (part-time employment, full-time employment), not employed (receiving benefits, unemployed, retired), other role (part-time student, N/A, houseperson, full-time student, other), [b]n = 72423, [c]n = 76194, [d]n = 65989, [e]n = 73196, [f]n = 59027, [g]n =10479, [h]n =35896, [i]n= 26780, [j]n = 36763.

* p < .001.

Table 2

Proportion of residual variance from an empty 3 level model explained by each additional fixed effect variable, in order of input to an incremental 3 level model.

| | | Additional Percentage of Original Variance Explained | | | |
|---|---|---|---|---|---|
| Variable | Level of model | Overall (%) | Level 3 (%) | Level 2 (%) | Level 1 (%) |
| Patient Severity | Level 1 (Patient) | 17.1 | 29.6 | 12.4 | 16.2 |
| Patient Employment | Level 1 (Patient) | 1.1 | 16.4 | 1.9 | 0.8 |
| Sector | Level 3 (Clinic) | 0.5 | 16.1 | <0.1 | <0.1 |
| Percentage White | Level 3 (Clinic) | 0.4 | 19.6 | 0.4 | <0.1 |
| Patient Ethnic Origin | Level 1 (Patient) | <0.1 | 0.5 | <0.1 | 0.1 |
| Patient Attendance | Level 1 (Patient) | <0.1 | <0.1 | 1.6 | 0.1 |
| Patient Age | Level 1 (Patient) | <0.1 | 0.2 | 2.0 | 0.2 |

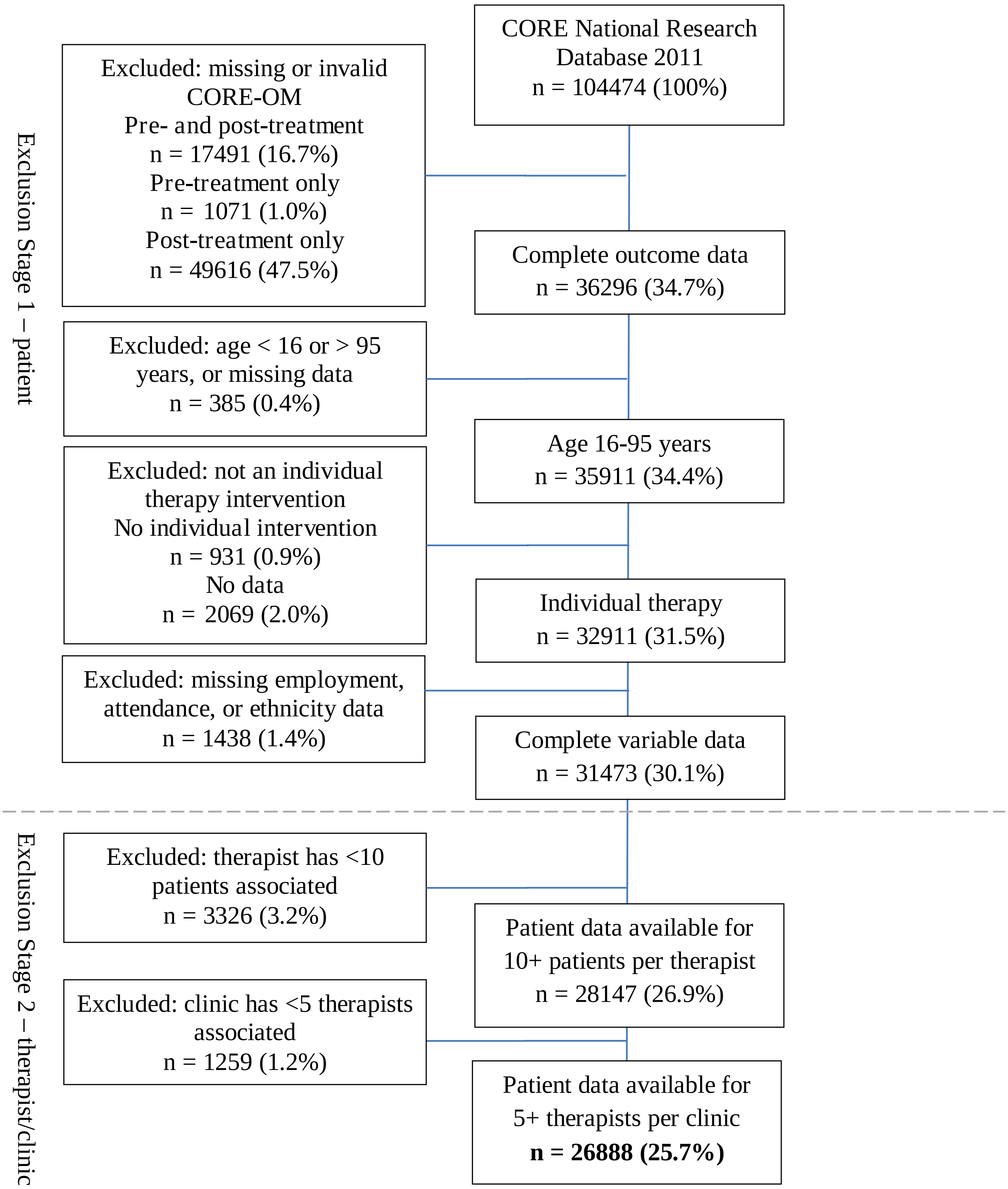


Figure 1. Inclusion and exclusion of patients from the Clinical Outcomes in Routine Evaluation (CORE) national database. CORE-OM = CORE Outcome Measure.

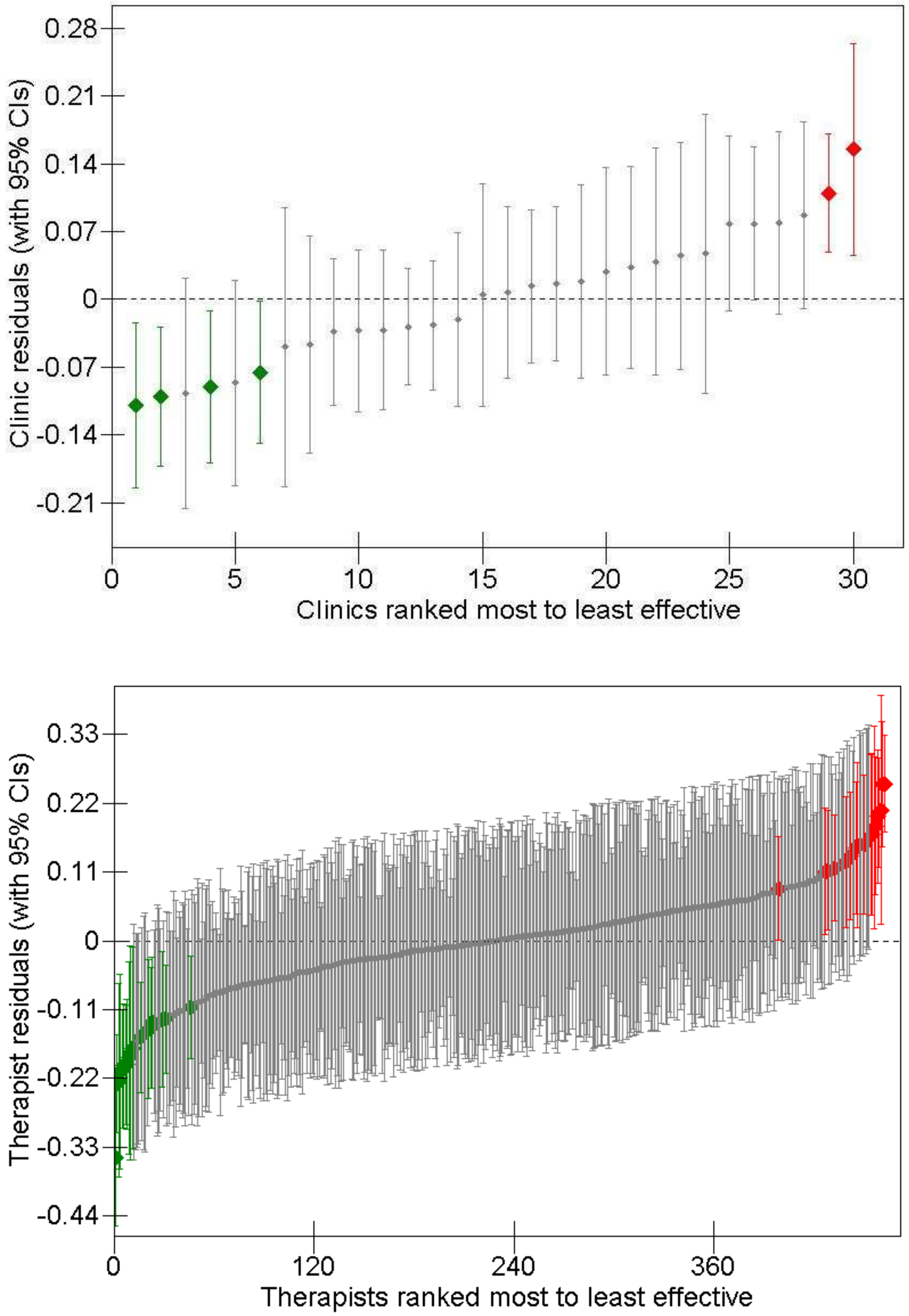


Figure 2. Markov Chain Monte Carlo estimated caterpillar plots of clinic variability (top) and therapist variability (bottom)

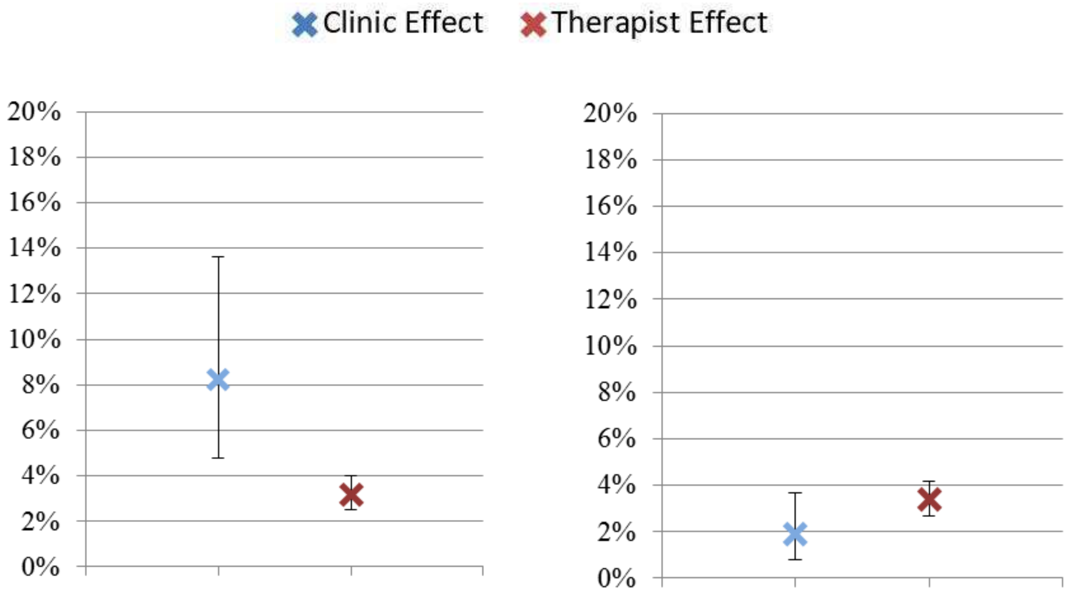


Figure 3. Markov Chain Monte Carlo estimated community effect and therapist effects with no explanatory variables (left), and after the inclusion of all significant explanatory variables (right). Vertical lines indicate 95% probability intervals for each effect.

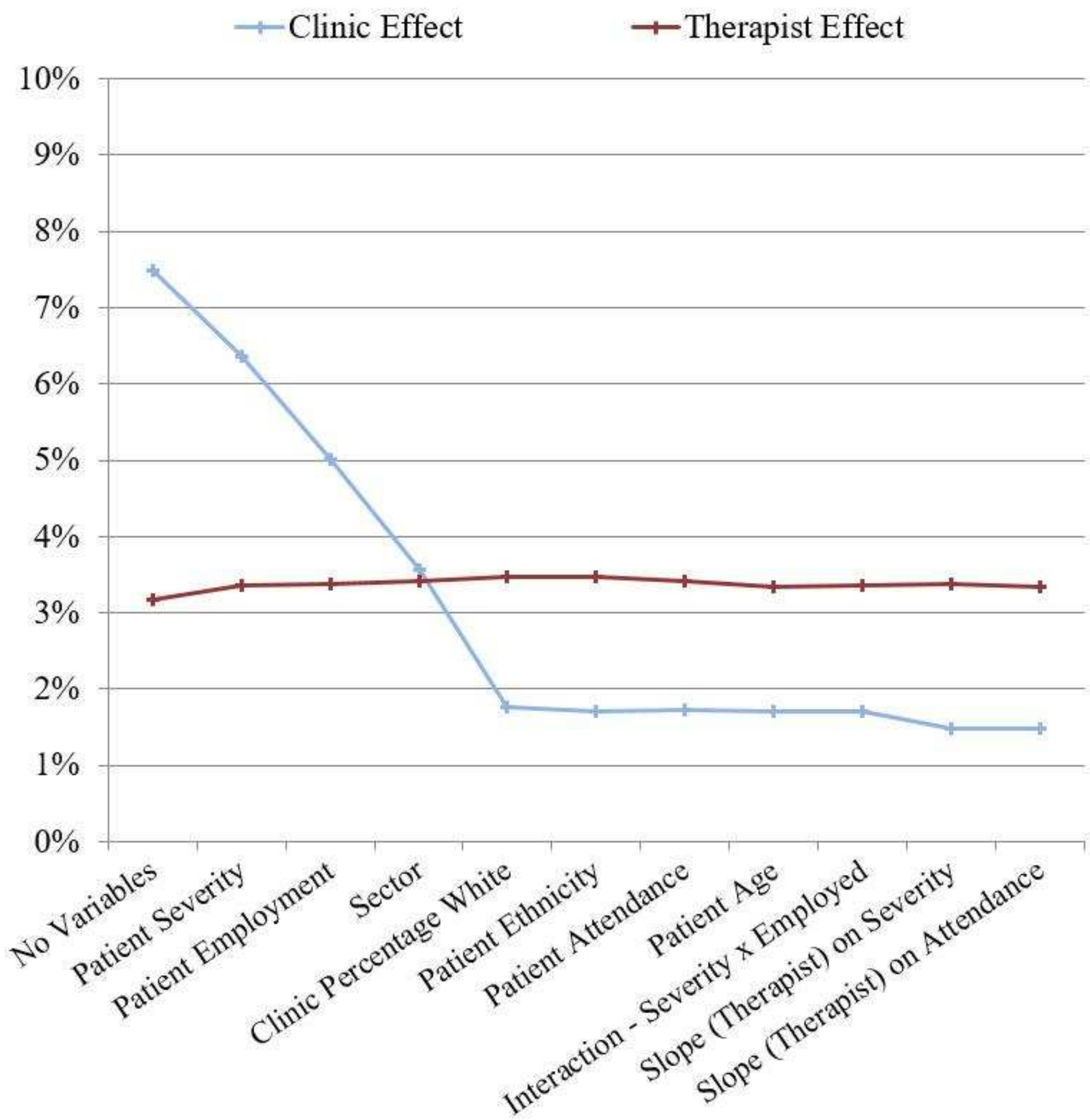


Figure 4. Iterative Generalised Least Squares (IGLS) estimated clinic effect and therapist effect, as each explanatory variable is incrementally added to the multilevel model (IGLS estimated values do not necessarily correspond exactly to Markov Chain Monte Carlo estimated values).

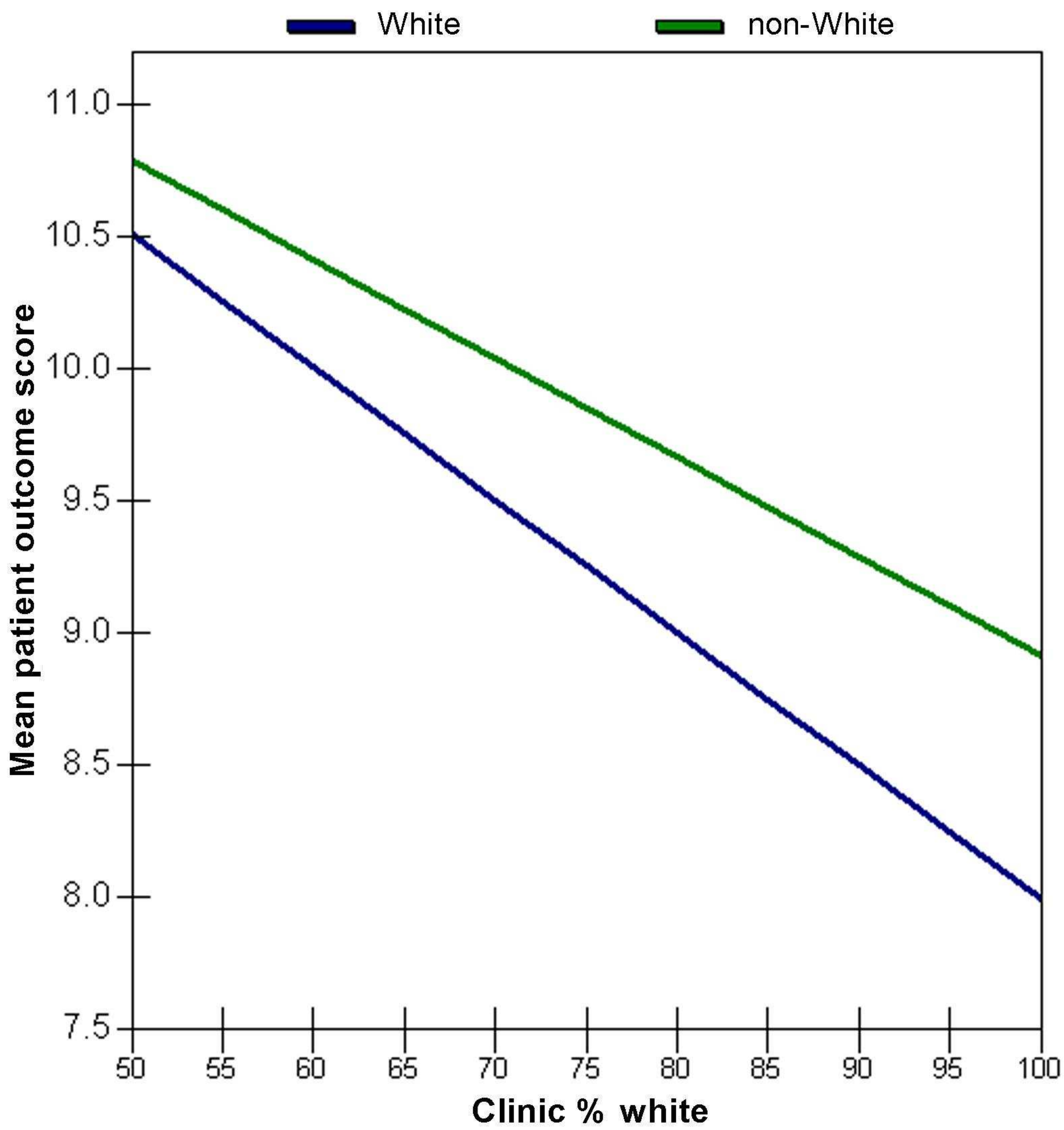


Figure 5. Predicted outcome scores for White and non-White patients treated at clinics with different percentages of White patients in their populations from 50% - 100% (Markov Chain Monte Carlo estimation).

Supplemental Material:

The final Markov Chain Monte Carlo multilevel model specification is included below for those with experience of multilevel models. CORE = Clinical Outcomes in Routine Evaluation – Outcome Measure score.

$$\begin{aligned}
\text{ln(PostCORE+1)}_{ijk} = {} & \beta_{0jk} + \beta_{1j}(\text{ln(PreCORE+1)-gm})_{ijk} + -0.169(0.013)\text{Employed}_{ijk} + \\
& -0.071(0.016)\text{Other Role}_{ijk} + 0.261(0.080)\text{SecondaryCare}_{k} + \\
& -0.007(0.001)(\text{Service\%white-gm})_{k} + -0.070(0.013)\text{White}_{ijk} + \\
& \beta_{7j}(\text{Attendance-gm})_{ijk} + 0.003(0.000)(\text{Age-gm})_{ijk} + \\
& -0.065(0.021)(\text{ln(PreCORE+1)-gm}).\text{employed}_{ijk} + e_{ijk}
\end{aligned}$$

$$\beta_{0jk} = 2.243(0.025) + v_{0k} + u_{0jk}$$
$$\beta_{1j} = 0.725(0.017) + u_{1jk}$$
$$\beta_{7j} = -0.169(0.031) + u_{7jk}$$

$$v_{0k} \sim \text{N}(0, \sigma^2_{v0}) \quad \sigma^2_{v0} = 0.007(0.003)$$

$$\begin{bmatrix} u_{0jk} \\ u_{1jk} \\ u_{7jk} \end{bmatrix} \sim \text{N}(0, \Omega_u) \ : \ \Omega_u = \begin{bmatrix} 0.013(0.002) & & \\ 0.004(0.002) & 0.010(0.003) & \\ 0.003(0.004) & -0.008(0.006) & 0.068(0.019) \end{bmatrix}$$

$$e_{ijk} \sim \text{N}(0, \sigma^2_e) \quad \sigma^2_e = 0.370(0.003)$$
$$\textit{Deviance(MCMC)} = 49603.284 (26888 \text{ of } 26888 \text{ cases in use})$$