Psychopharmacology (2014) 231:1987–1999
DOI 10.1007/s00213-013-3344-x

ORIGINAL INVESTIGATION

The impact of missing data on clinical trials: a re-analysis of a placebo controlled trial of *Hypericum perforatum* (St Johns wort) and sertraline in major depressive disorder

Anneke C. Grobler · Glenda Matthews · Geert Molenberghs

Received: 23 May 2013 / Accepted: 18 October 2013 / Published online: 15 November 2013

Abstract
Rationale and objective *Hypericum perforatum* (St John's wort) is used to treat depression, but the effectiveness has not been established. Recent guidelines described the analysis of clinical trials with missing data, inspiring the reanalysis of this trial using proper missing data methods. The objective was to determine whether hypericum was superior to placebo in treating major depression.
Methods A placebo-controlled, randomized clinical trial was conducted for 8 weeks to determine the effectiveness of hypericum or sertraline in reducing depression, measured using the Hamilton depression scale. We performed sensitivity analyses under different assumptions about the missing data process.
Results Three hundred forty participants were randomized, with 28 % lost to follow-up. The missing data mechanism was not missing completely at random. Under missing at random assumptions, some sensitivity analyses found no difference between either treatment arm and placebo, while some sensitivity analyses found a significant difference from baseline to week 8 between sertraline and placebo (−1.28, 95 % credible interval [−2.48; −0.08]), but not between hypericum and placebo (0.56, [−0.64;1.76]). The results were similar when the missing data process was assumed to be missing not at random.
Conclusions There is no difference between hypericum and placebo, regardless of the assumption about the missing data process. There is a significant difference between sertraline and placebo with some statistical methods used. It is important to conduct an analysis that takes account of missing data using valid statistically principled methods. The assumptions about the missing data process could influence the results.

Keywords St John's wort · *Hypericum perforatum* · Herbal medicine · Antidepressant · Sertraline · Hamilton depression scale · Bayesian · Multiple imputation · Missing at random · Missing not at random

Introduction

Hypericum perforatum

H. perforatum (St John's wort), is a herbal remedy used in the treatment of depression, especially in European countries (Fegert et al. 2006). It was shown to be more effective than placebo (Kalb et al. 2001) in treating depression and is believed to have fewer side effects than standard antidepressive therapies (Kasper et al. 2010; Linde et al. 1996). Some studies and meta-analyses have found hypericum to be as effective as standard antidepressive therapies (Linde et al. 2008; Rahimi et al. 2009), while other studies found no difference between hypericum and placebo (Shelton et al. 2001). Because different studies had contradictory results about the effectiveness of hypericum compared to placebo and standard antidepressive drugs, a trial was designed to compare both a standard antidepressive therapy (sertraline) and hypericum to placebo (Hypericum Depression Trial Study Group 2002). Sertraline

Trial Registration: Clintrials.gov, NCT00005013, http://www.clinicaltrials.gov/ct2/show/NCT00005013?term=Hypericum+perforatum+major+depression

A. C. Grobler (✉)
Centre for the AIDS Programme of Research in South Africa (CAPRISA), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Private Bag X7, Durban 4013, South Africa
e-mail: grobler@ukzn.ac.za

G. Matthews
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

G. Molenberghs
I-BioStat, Universiteit Hasselt and KU Leuven, Leuven, Belgium

(trademark Zoloft) is an antidepressant of the selective serotonin reuptake inhibitor (SSRI) class. The practicing clinician will be interested in these results, since herbal preparations should be used only when evidence exist that it is efficacious.

Missing data

Missing data are common in longitudinal clinical trials. Over the last 30 years, methods were designed for the proper analysis of clinical trial data when missing data are present. This is summarized in two guidance documents by the European Medicines agency, "Guideline on Missing Data in Confirmatory Clinical Trials"(CHMP 2010), and the National Research Council in the United States of America, "The Prevention and Treatment of Missing Data in Clinical Trials"(National Research Council 2010). The second report was recently summarized in the New England Journal of Medicine (NEJM) (Little et al. 2012). In the same issue, the journal announced new review policies regarding missing data (Ware et al. 2012). Reviewers will in future look at aspects of trial design that reduce the impact of missing data. Weighting or model-based methods will be preferred over complete case analysis or single imputation methods. The NEJM will in future also require sensitivity analysis when missing data are extensive.

Rubin (1976) described three different missing data mechanisms based on the level of dependence between the missing data process and the measurement process. These are missing completely at random, missing at random, and missing not at random. Data are missing completely at random when the probability of dropout is independent of both observed and unobserved data, for example, when a sample was lost in the laboratory or a patient did not attend a visit due to transport problems. Data are missing at random when the reason for dropout is known and associated with trial-related events (Carpenter et al. 2002). The reason for dropout can depend on observed data, but not on unobserved data, for example, when a participant who is doing poorly is subsequently discontinued from the trial by the clinician or as per participant's choice and a poor efficacy outcome is recorded in the study database. When neither missing completely at random nor missing at random is valid, data are missing not at random. In this instance, the missingness can be explained by unobserved outcomes, for example, when a participant whose condition worsens stops coming to the clinic, and this worsened condition is not observed. The missing data mechanism cannot be determined using the data observed, except possibly to confirm that the missing data mechanism is not missing completely at random.

Ten years ago, it was standard practice to analyze clinical trials using complete case analysis or single imputation techniques including last observation carried forward (LOCF). This is changing in favor of more appropriate methods and recommendations now advise against these (Carpenter et al. 2004; Carpenter et al. 2013; CHMP 2010; Mallinckrodt et al. 2001; Molenberghs and Kenward 2007; Molenberghs et al. 2004; National Research Council 2010; Ratitch et al. 2013). In 2002, the hypericum clinical trial was published using complete case analysis and last observation carried forward (Hypericum Depression Trial Study Group 2002). In this paper, we reanalyzed the data up to week 8, using principled methods suggested by current guidelines, which did not exist in 2002. We use this trial as an example of the changes in analyses that are required if the new guidelines are to be adopted.

Materials and methods

Trial design

The trial was a randomized, double-blind, parallel-arm, 8-week outpatient trial of hypericum, sertraline, or placebo treatment for major depressive disorder, followed by 18 weeks of double-blind continuation treatment in participants meeting response criteria at week 8. The study consisted of an acute phase (the first 8 weeks) and an optional continuation phase (from weeks 8 to 26). The focus of this paper is on the acute phase only. The specific inclusion and exclusion criteria are given in the trial publication (Hypericum Depression Trial Study Group 2002). The eligibility of participants was assessed, after which they gave written informed consent and participated in a 1-week placebo run-in. Participants meeting eligibility criteria after the run-in were randomized to one of the three treatment arms in a 1:1:1 ratio. Participants were assessed weekly from week 1 to week 8. The Hamilton depression scale (HAM-D) (Hamilton 1960), global assessment of functioning (GAF) scale, Clinical Global impressions scale for severity (CGI-S) and improvement (CGI-I), and the Beck Depression Inventory (BDI) were assessed at all visits. Other safety-related information was also collected, such as vital signs, adverse events, and blood chemistry and hematology (Hypericum Depression Trial Study Group 2002). We analyze the response over time on HAM-D, which is a measure of depression, with a higher score indicative of more severe depression.

The primary hypothesis was whether hypericum is superior to placebo after 8 weeks. The endpoint was defined as the change in the HAM-D score from baseline to week 8. The principal comparison was between the hypericum and placebo arms. The sertraline arm was included as an active control arm to validate the study, but no comparison between hypericum and sertraline was intended and the trial was not powered for such a comparison or for multiple comparisons with placebo (Hypericum Depression Trial Study Group 2002). Details about the proportion of patients discontinued, timing of discontinuation, and reason for discontinuation are provided.

Statistical methods

In the original analysis, treatment differences in the change in HAM-D total score from baseline to week 8 were evaluated through a random coefficient regression model. Available longitudinal scores through week 8 were modelled as a linear function of fixed effects for treatment, site, sex, week, and treatment by week, with random intercept and slope for each patient. A secondary analysis was restricted to participants who completed the acute phase (completer analysis), and analysis of covariance models used last observation carried forward; although these results were not given in the paper (Hypericum Depression Trial Study Group 2002). An analysis of the data in the continuation phase was also done in a separate paper. In this analysis, the HAM-D scores for completers at the final time point were compared and last observation carried forward was applied (Sarris et al. 2012). While the random coefficient regression model could be appropriate if the missing data mechanism is missing completely at random or missing at random, last observation carried forward is not an appropriate way of handling missing data, because it might inflate the Type I error and create bias in the estimation of mean change from baseline while producing standard errors that are too small (Mallinckrodt et al. 2001). The random coefficient regression model takes into account the expectation of the missing measurements, given the observed measurements and is valid and unbiased under missing at random (Molenberghs and Kenward 2007). It can be thought of as aiming to estimate the treatment effect that would have been observed if all participants had continued on treatment for the full study duration (CHMP 2010).

We reanalyzed the data using principled missing data techniques. Assumptions were made about the missing data mechanism and a series of sensitivity analyses were done under these assumptions. If the missing data mechanism is missing completely at random, an available case or complete case analysis would be valid, although this analysis would have reduced power because of the exclusion of some participants. If one assumes the missing data mechanism is missing at random, several methods would be valid, including maximum likelihood models, multiple imputation, and Bayesian analysis.

Likelihood-based approaches use a parametric model to formulate a statistical model for the missing data and base inferences on the likelihood function of the incomplete data. The objective is to draw inference about a parameter, θ, in a model $f\ (y|\theta)$ for the response data that is not fully observed. Under the missing at random assumption, θ and the missing data model are functionally independent and missing data can be treated as ignorable. In this case, inference is drawn about θ without having to specify a model that relates the missing data process to the observed data (National Research Council 2010).

In the absence of missing data, likelihood-based methods entail the maximization of the full data likelihood. With incompleteness, this likelihood is replaced by the observed data likelihood, where the individual likelihoods are integrated over the missing values, $\prod_{i=1}^{N} \int f\left(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{miss}}, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}\right) \mathrm{d}\boldsymbol{y}_i^{\text{miss}}$; where y indicate the outcome variable, r is the missing data indicator, θ and ψ are parameter vectors describing the measurement and missingness processes, respectively, and N is the number of participants. Under ignorability and missing completely at random or missing at random missingness, the integral can be rewritten as an integral over the missing values and the distribution of the missing data mechanism (under a selection model). Under missing completely at random, this becomes $\prod_{i=1}^{N} f\left(\boldsymbol{y}_i^{\text{obs}}|\boldsymbol{\theta}\right) f(\mathbf{r}_i|\boldsymbol{\psi})$ and under missing at random this becomes $\prod_{i=1}^{N} f\left(\boldsymbol{y}_i^{\text{obs}}\ |\boldsymbol{\theta}\right) f\left(\boldsymbol{r}_i|\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{\psi}\right)$. We fitted a model with fixed and random effects, including treatment, week, and the interaction between treatment and week, adjusted for repeated measures.

Single imputation has several limitations. When analyzing observed data, it is assumed that measurements are made with error. To assume that if data are missing, we can impute the missing value without error (a single value) is unrealistic. With conditional mean imputation, the imputed data are much less variable than the observed data would have been. Thus, analyzing imputed data as observed data leads to an underestimation of standard errors, p values, and confidence intervals (Carpenter and Kenward 2007; CHMP 2010).

Multiple imputation, first suggested by Rubin (1976), overcomes the limitations of single imputation. Multiple imputation is done in three steps. In step 1, plausible values for missing observations are imputed that reflect uncertainty about the missing data models, generally assuming the missing data process is missing at random. These values are used to fill in or impute the missing values. This process is repeated, resulting in the creation of several complete data sets, taking into account the uncertainty in estimating both the relationship between the variables and the residual variability. These provide a representation of the distribution of the missing data given the observed. In step 2, each of these data sets is analyzed using complete data methods that would have been appropriate had there been no missing data. In Step 3 the results are combined, taking the uncertainty regarding the imputations into account (Rubin 1976). This additional step is needed to correctly estimate the variability of quantities estimated from a completed data set. These results are unbiased and have approximately the correct standard error (Horton and Kleinman 2007; Molenberghs and Kenward 2007).

Multiple imputation is said to be proper if it leads to consistent asymptotically normal estimators, correct variance estimators, and valid tests. Generally, the imputation will be

proper if all sources of variability and uncertainty are included in the imputation model, including prediction errors of the individual values and errors of estimation in the fitted coefficients of the imputation model (White et al. 2011). Multiple imputation is done using Bayesian predictive distribution and Monte Carlo Markov Chain sampling to generate the imputation, assuming that the data follows a multivariate normal distribution. The model used for imputation should include all the variables included in the analysis model (to ensure proper imputation), and all variables that could improve the prediction.

Multiple imputation with 100 imputations was used to analyze the change from baseline to week 8. The imputation model included all observed values of HAM-D, age, sex, race, duration of depression, BDI, CGI-S, CGI-I, and GAF scale at baseline. The variables were selected because they were believed to be factors that could predict HAM-D scores or were included in the analysis model. The imputed datasets were used to get the estimates of the change from baseline to week 8, and the appropriate p values and summary statistics were calculated using Rubin's rule (Rubin 1976). The imputed datasets were also analyzed with a mixed model as described previously.

At its core, multiple imputation uses Bayesian techniques, since the imputations are sampled from a Bayesian posterior distribution. Fully Bayesian approaches, where multiple datasets are not imputed, are appropriate for the analysis of missing data by specifying priors on all the parameters and specifying distributions for the missing covariates (Daniels and Hogan 2008; Horton and Kleinman 2007). The missing data are then sampled from their conditional distribution via the Gibbs sampler, an algorithm that samples a Markov chain where the kernel is the product of the sequentially updated full conditional distributions of the parameters and the stationary distribution is the posterior distribution (Geman and Geman 1984).

In Bayesian analyses, parameters are treated as random variables. Probability statements are made about the model parameters and not about the data. Bayesian analysis has three components. The prior distribution, $p(\theta)$, reflects the distribution of the parameters before the data are seen. The likelihood, $L(\theta|D)$, gives the distribution of the observed data. The posterior distribution uses Bayes' theorem to combine information from the prior distribution and the likelihood and expresses uncertainty about the unknown parameters after seeing the data. In ignorable methods, the posterior distribution is $p(\theta|D) \propto p(\theta)L(\theta|D)$. The Bayesian inference is done by specifying a model, specifying prior distributions for the parameters of the model, and then updating the prior information on the parameters using the model specified and the data observed to obtain the posterior distribution of the parameters. In addition to specifying a missingness model, assumptions about the missing data and the uncertainty around the missing data can be made explicit through the prior distributions (Daniels and Hogan 2008). The priors can be used to encode information about the missing data process. Bayesian methods are a natural way of handling missing data because a probability distribution is estimated for each missing value, allowing for uncertainty to be captured. Missing data are treated as additional unknown quantities, thus, no distinction is made between missing data and unknown parameters. After specifying an appropriate joint model for the observed and missing data and the model parameters, posterior samples of the model parameters and missing values will be generated using Markov Chain Monte Carlo methods (Mason et al. 2012).

Letting Y_{ij} be the HAM-D score for participant i at occasion j, where $Y_{ij} \sim \text{Normal}(\mu_{ij}, \sigma^2)$, the following Bayesian model was fitted:

$$\mu_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \mathrm{S}_i + \beta_3 \mathrm{H}_i + (\beta_4 S_i + \beta_5 H_i) t_{ij}$$

where, t_{ij} indicates occasion j (week_j) for participant i. S_i is an indicator variable for the sertraline arm and equals 1 if participant i belongs to the sertraline arm and is 0 otherwise. Similarly, H_i is an indicator variable for the hypericum arm and equals 1 if participant i belongs to the hypericum arm and is 0 otherwise. A participant i belonging to the placebo arm will thus have $\mathrm{H}_i = \mathrm{S}_i = 0$. The placebo arm is therefore the reference group. We assume N independent participants.

Vague priors were specified for the unknown parameters and are given in Table 4. Sensitivity analyses were performed with various prior distributions, to assess the sensitivity of the Bayesian models to the choice of prior distribution. Priors could also have been used to include data about the missingness process. One could potentially use different prior distributions for each of the treatment arms, if the prior beliefs about the missingness mechanism warranted this.

Under missing not at random, an assumption we cannot test using the observed data, Bayesian analyses, pattern mixture models, selection models, and shared parameter models can be used, at the expense of increased complexity.

Under a missing not at random assumption, we fitted three Bayesian models under several different assumptions about the missing data mechanism. Bayesian analysis provides a flexible way to model missing not at random data, using a selection model factorization of a joint model, consisting of a model of interest and a model of missingness. The same model for the observed data was fitted as under missing at random and a model of missingness of the form $m_{ij} \sim \text{Bernoulli}(p_{ij})$, $\text{logit}(p_{ij}) = \theta_0 + \Delta y_{ij}$ was added, where $i = 1, \ldots, 340$ indicates the participant and $j = 1, \ldots 8$ indicates the visit, m_{ij} is a binary missing value indicator for y_{ij}. This model allows the missingness to depend on the value that would have been observed.

A second missing not at random model was fitted where the model of missingness had the following form: logit $p_{i,w}=\theta_0+\theta_1 y_{i,w-1}+\theta_2(y_{i,w}-y_{i,w-1})$. This model allows the missingness to depend on the value that would have been observed at the current occasion, where the value is possibly missing, as well as on the previous observed value. In so doing, in line with what is oftentimes done in a selection model specification, missingness can be seen to depend on both the level of the outcome (represented by the average of the current and previous values) and the increment between the previous and current values. It should be noted that the data do not carry information on θ_2 in the usual sense. While under a likelihood and Bayesian paradigm parameters may be identified, the usual asymptotics may not hold, in the sense that information accrual may be minimal with increasing sample sizes. This subtle issue is discussed and illustrated with data analyses and simulations in Jansen et al. (2006) and the references therein. Practically, it means that parameters distinguishing missing not at random from missing at random under the posited model may be identifiable from the observed data, but only barely so.

A third missing not at random model was fitted where the model of missingness has the following form: logit $p_{ij}=\theta_0+\theta_1 y_{ij}+\theta_2 y_{ij}S_i+\theta_3 y_{ij}H_i$. This model allows the missingness to depend on the unobserved HAM-D value, while allowing for a different mechanism in each treatment arm by including the HAM-D score by treatment interaction. The priors are given in Table 4. The priors were chosen to be flat and therefore uninformative. The parameters were varied from extremely flat priors, to less flat priors in order to investigate whether the models were sensitive to the choice of prior. Many other models for nonrandom missingness could be fitted, depending on the assumptions made regarding the missing data mechanism.

The assumptions made by the models fitted were tested. The assumption of linear regression was tested by plotting the studentized residuals against the predicted means. This plot showed no deviations from the assumption of linear regression. The normality assumption was tested by looking at the normal probability plot of the residuals. No deviation from normality was present. Different variance covariance structures were fitted and the most appropriate was chosen using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

All analyses were performed on the data collected from baseline to week 8, using SAS version 9.3 software (SAS Institute Inc, Cary, NC), with the exception of the Bayesian analyses, which were performed using OPENBUGS version 3.2.2. All assumptions of linear regression were satisfied.

Results

The trial enrolled 428 participants in the run-in phase and 340 were randomized to the tree treatment arms. The demographics of the sample and the CONSORT diagram is described elsewhere (Hypericum Depression Trial Study Group 2002). The mean pooled baseline HAM-D score was 22.8 (SD=2.7). A large number of participants was lost to follow-up before week 8; 28.8 % in the sertraline arm, 27.4 % in the hypericum arm, and 27.6 % in the placebo arm. At the end of week 8, the percentage participants who dropped out were similar across the three arms; however, participants discontinued sooner in the active arms, especially in the sertraline arm. The extent of and reason for missing data at each visit is given in Table 1.

A central question is whether the missing data are missing completely at random, missing at random, or missing not at random. Although the data cannot exclude missing not at random, it can hold some evidence of informative missingness. Figures 1 and 2 compare the HAM-D score of participants who dropped out at the next visit to those of participants who did not drop out at the next visit. Participants who dropped out are different from those who did not drop out. Prior to week 4, dropouts had lower HAM-D scores than those who did not drop out. From week 4 onwards, dropouts had higher HAM-D scores than participants who did not drop out; thus, dropout depends at least on observed HAM-D score. Among demographic and baseline variables, duration of depression and age were significantly associated with dropout in a logistic regression.

The fact that some withdrawals were due to insufficient response suggested that HAM-D score should not be analyzed without taking missing data into account. In addition, the fact that participants started dropping out sooner in the sertraline arm might reflect that tolerability issues were related to dropout, casting further doubt on the suitability of a missing at random model. Ignoring participants who drop out due to insufficient response or tolerability issues will introduce an important bias in the complete case analysis. In this context, the missing at random based model expresses the assumption that a participant's observed history is deemed adequate to derive his or her probability of dropping out. Here, history is to be understood as the combination of the patient's outcomes from the beginning of the study up to but not including the current one, and the covariate information, collected at baseline and during follow-up, up to the current time. Up to week 8, a missing completely at random analysis is not appropriate; a missing at random analysis might be appropriate, but missing not at random cannot be ruled out. Assumptions about the missing data, other than missing at random should be considered.

In the original paper, the mean HAM-D scores using available case analysis were given in a figure. According to this figure, the sertraline arm showed the largest improvement over time. It also included a random coefficient regression analysis on the longitudinal HAM-D total scores. This analysis detected a downward linear trend with week (p value<0.001). Linear trends with week did not

Table 1 Number of participants attending each visit

	Sertraline $N=111$		Hypericum $N=113$		Placebo $N=116$	
Number with Hamilton depression score at	*N*	Number missing (%)	*N*	Number missing (%)	*N*	Number missing (%)
Baseline	111	–	113	–	116	–
Week 1	101	10 (9.0)	101	12 (10.6)	111	5 (4.3)
Week 2	90	21 (18.9)	102	11 (9.7)	107	9 (7.8)
Week 3	90	21 (18.9)	100	13 (11.5)	94	22 (19.0)
Week 4	89	22 (19.8)	97	16 (14.2)	99	17 (14.7)
Week 6	82	29 (25.1)	91	22 (19.5)	93	23 (19.8)
Week 7	79	32 (28.8)	82	31 (27.4)	84	32 (27.6)
Week 8	79	32 (28.8)	82	31 (27.4)	84	32 (27.6)
Enter continuation phase	49		38		42	
Reasons not completing acute phase (week 8)		$N=32$		$N=31$		$N=32$
Loss to follow-up	10	31.3 %	8	25.8 %	7	21.9 %
Insufficient response	7	21.9 %	6	19.4 %	11	34.4 %
Withdrew consent	8	25.0 %	7	22.6 %	8	25.0 %
Adverse event	5	15.6 %	2	6.5 %	3	9.4 %
Protocol violation	2	6.3 %	8	25.8 %	3	9.4 %

differ significantly by treatment (hypericum versus placebo, p value=0.59; sertraline versus placebo, p value=0.18). If this analysis was done using all data points while fitting a mixed model using maximum likelihood estimates, this analysis is consistent with missing at random assumptions (Hypericum Depression Trial Study Group 2002). From the original paper, model estimates for the mean change from baseline to week 8 in HAM-D score were calculated for each of the treatment arms (Table 2). The conclusion was that neither hypericum nor sertraline was superior to placebo. The authors highlighted the high level of improvement in the placebo arms often seen in depression trials (Hypericum Depression Trial Study Group 2002).

We compared the change from baseline to week 8 using multiple imputation. The change was slightly larger in all arms using multiple imputation, but the p values were similar to the previous analysis. We conclude that there is no difference between either of the treatment arms and the placebo arm (Table 2).

The results for the change from baseline to week 8 using multiple imputation and a mixed model are presented in Tables 2 and 3, respectively. Multiple imputation and likelihood-based methods make similar assumptions about the missing data, namely that it is missing at random. The conclusions are expected to be similar. Neither analysis found either of the treatments to be different from placebo. Analyzing the data

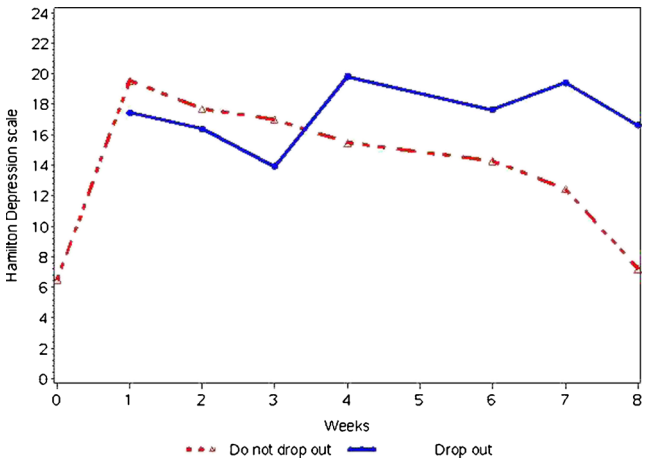

Fig. 1 Sample mean of Hamilton depression score at each week for all treatment groups combined. The *triangles* are the means for participants who did not drop out before the subsequent measurement. The *circles* are the means for participants who dropped out before the subsequent measurement

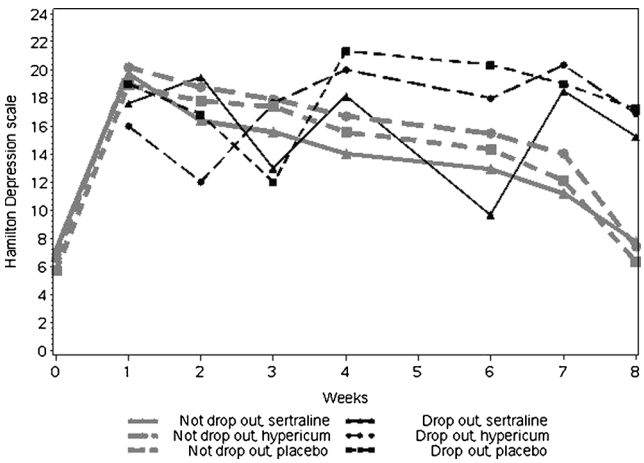

Fig. 2 Sample mean of Hamilton depression score at each week by treatment group. Not drop out gives the means for participants who did not drop out before the subsequent measurement. Drop out gives the means for participants who dropped out before the subsequent measurement

Table 2 Change in Hamilton depression score (week 8 to baseline)

	Mean change from baseline (standard error of the mean change from baseline) [95 % confidence interval]			*p* value	
	Hypericum ($n=113$)	Placebo ($n=116$)	Sertraline ($n=109$)	Hypericum vs placebo	Sertraline vs placebo
Last observation carried forward, change from baseline to week 8	−8.42 (0.68) [−9.76; −7.09]	−8.89 (0.73) [−10.33; −7.45]	−9.68 (0.64) [−10.94; −8.42]	0.64	0.42
From 2002 paper[a]	−8.68 (0.68) [−10.01; −7.35]	−9.20 (0.67) [−10.51; −7.89]	−10.53 (0.72) [−11.94; −9.12]	0.59	0.18
Multiple imputation: change calculated on the imputed values	−9.99 (0.74) [−11.45; −8.54]	−10.01 (0.80) [−11.58; −8.43]	−11.47 (0.67) [−12.78; −10.17]	0.99	0.16
Maximum likelihood models under missing at random assumptions, change based on model results[b]	−9.98 (0.76) [−11.47; −8.49]	−10.12 (0.74) [−11.58; −8.66]	−11.41 (0.78) [−12.94; −9.87]	0.19	0.23
				Significant according to posterior 95 % credible interval	
Missing at random Bayes model (Prior 1)	−8.80 (0.44) [−9.65; −7.93]	−9.36 (0.43) [−10.2; −8.52]	−10.64 (0.45) [−11.53; −9.76]	Not significant	Significant
Missing not at random Bayes model 1 (Prior 1)	−8.79 (0.44) [−9.65; −7.94]	−9.40 (0.43) [−10.20; −8.53]	−10.64 (0.45) [−11.52; −9.77]	Not significant	Significant
Missing not at random Bayes model 2 (Prior 1)	−8.79 (0.44) [−9.65; −7.94]	−9.35 (0.43) [−10.20; −8.51]	−10.65 (0.45) [−11.53; −9.77]	Not significant	Significant
Missing not at random Bayes model 3 (Prior 1)	−8.79 (0.44) [−9.65; −7.93]	−9.37 (0.43) [−10.21; −8.53]	−10.65 (0.45) [−11.53; −9.77]	Not significant	Significant

A lower score means a greater improvement

Vs versus

[a] Values are linear model estimates, and *p* values adjusted for site and sex based on modelling the longitudinal measures at week 0 and weeks 1 through 8 as a linear function of site, sex, treatment, week, and treatment by week, with a random intercept and slope over time for each participant including all available data. Results taken from the 2002 paper (Hypericum Depression Trial Study Group 2002).

[b] All participants included and all values modelled under missing at random assumptions. The model included site, treatment, week, treatment by week and a random intercept and slope over time for each participant

Table 3 Results fitting a mixed model

	With multiple imputation				Without multiple imputation			
	Estimate	Standard error	95 % CI	*p* value	Estimate	Standard error	95 % CI	*p* value
Intercept	22.70	0.25	22.22; 23.19		23.17	0.25	22.68; 23.66	
Week	−1.15	0.09	−1.33; −0.96	<0.001	−1.23	0.08	−1.40;−1.07	<0.001
Sertraline versus placebo	0.09	0.35	−0.60; 0.78	0.79	−0.08	0.36	−0.75; 0.66	0.81
Hypericum versus placebo	0.42	0.35	−0.26; 1.10	0.23	0.27	0.35	−0.43; 0.97	0.45
Week by sertraline	−0.23	0.13	−0.48; 0.02	0.08	−0.19	0.12	−0.43; 0.05	0.12
Week by hypericum	0.01	0.13	−0.24; 0.25	0.95	0.07	0.12	−0.16; 0.31	0.55

CI Confidence interval

using a likelihood-based model found the only significant effect to be week, indicating that depression decreased over time (Table 3).

The choice of a vague prior did not change the results appreciably in the Bayesian missing at random analysis (Table 4). The results were similar to the maximum likelihood model, with one notable exception. Under the maximum likelihood model, none of the treatment effects was significant; however, under this model, the interaction between the sertraline arm and week was significant, indicating that the decrease in HAM-D score over time was larger in the sertraline arm than in the placebo arm, regardless of the choice of prior.

Under the missing not at random assumption, we did several Bayesian analyses as a sensitivity analysis. The results under the first missing not at random Bayes model did not differ appreciably from the results under the missing at random Bayes model. Under prior sets 1 and 2, the results were almost similar. The posterior means were slightly different under the less flat and therefore more nonrandom priors used in prior set 3. Prior set 3 was the most informative prior used in that provided the strongest prior belief against the missing not at random assumption. However, the conclusions were the same regardless of prior set used. There was a significant interaction between sertraline and week, meaning that the sertraline arm had a larger decrease in HAM-D score than the placebo arm over time. Under missing not at random model 2, the posterior means for the β-coefficients were similar to the posterior means under missing not at random model 1, with the exception of sertraline and hypericum under prior sets 2 and 3. This model was more sensitive to the choice of prior than model 1. The conclusion drawn is the same, except under prior set 2, where the interaction between sertraline and week just did not meet statistical significance.

These results need to be interpreted with caution. First, as was discussed in Jansen et al. (2006), a test for missing not at random versus missing at random is valid only under the untestable assumption that the missing not at random alternative is correctly specified. Secondly, even then, this test has been shown not to have the usual power behavior, simply because there is information missing. Evidently, this problem cannot be avoided, hence the need for sensitivity analysis.

Missing not at random model 1 is sometimes called "protective" (Michiels and Molenberghs 1997) and is specific in the sense that dropout can depend on the current, possibly unobserved, measurement, but not on the previous one. Intuitively, it is a mirror image of a commonly used missing at random model, where missingness depends on the previous but not current value. Assuming that previous and current values are often relatively similar, these models are often not too different from each other, establishing that they retain some of the stability of missing at random models. Model 2, on the other hand, allows missingness to depend on previous and current measurements, and therefore also on the increment between them. This is a profound departure from missing at random, and about the increment there is often not a lot of information in the data, because it is by definition unknown for someone dropping out, at the time of dropout (Jansen et al. 2006). As a result, it is expected that the model is more sensitive to unverifiable assumptions or choices made, such as prior specification.

Discussion

The conclusion drawn in the original paper was that there was no difference between either the hypericum arm and placebo or the sertraline arm and placebo. This was taken to mean that the study results were inconclusive. Although it showed that hypericum was no better than placebo, it also did not find the expected difference between sertraline and placebo (Hypericum Depression Trial Study Group 2002). The same was found when the continuation data was analyzed, and the placebo effect was again noted and discussed (Sarris et al. 2012). The last observation carried forward analysis used previously was not a plausible assumption in this instance because of the week effect found. It penalized the arm with higher or earlier dropout; in this instance, the sertraline arm. This explains why the original analyses had inconclusive results. We reanalyzed the

Table 4 Hamilton depression score posterior means, standard deviations, and credible intervals according to Bayesian analysis

Missing at random model

	Prior set 1			Prior set 2			Prior set 3		
	β ~ Normal (0,10 000), precision~Gamma (0.001, 0.001).			β ~ Normal (0,1000), precision~Gamma (0.01, 0.01).			β ~ Normal (0,10), precision~Gamma (0.1,0.1).		
Parameter	Mean	SD	95 % credible interval	Mean	SD	95 % credible interval	Mean	SD	95 % credible interval
β_0 – Intercept	21.01	0.48	20.08; 21.95	21.0	0.47	20.06; 21.94	20.58	0.46	19.68; 21.49
β_1 – Week	−1.17*	0.05*	−1.28; −1.07*	−1.17*	0.05*	−1.28; −1.07*	−1.15*	0.05*	−1.26; −1.05*
β_2 – Sertraline	−0.20	0.67	−1.58; 1.05	−0.17	0.68	−1.54; 1.1	0.24	0.65	−1.07; 1.47
β_3 – Hypericum	0.74	0.66	−0.59; 1.99	0.76	0.66	−0.57; 2.01	1.15	0.66	−0.15; 2.40
β_4 – Interaction week and sertraline	−0.16*	0.08*	−0.31;−0.01*	−0.16*	0.08*	−0.31; −0.01*	−0.18*	0.08*	−0.33; −0.03*
β_5 – Interaction week and hypericum	0.07	0.08	−0.08; 0.22	0.07	0.08	−0.08; 0.22	0.05	0.08	−0.09; 0.21
Standard deviation from τ	3.94	0.06	3.82; 4.07	3.94	0.06	3.82; 4.07	3.94	0.06	3.82; 4.07
Standard deviation from τ2	4.36	0.19	4.01; 4.76	4.36	0.19	4.36; 4.76	4.37	0.19	4.01; 4.76

Missing not at random (MNAR) model 1

	1st MNAR model: prior set 1			1st MNAR model: prior set 2			1st MNAR model: prior set 3		
	β~ Normal(0,10 000), precision~Gamma (0.001, 0.001), θ_0 ~logistic (0,1), Δ~Normal (0, 10 000)			β~ Normal (0,1000), precision~Gamma (0.01, 0.01), θ_0 ~ logistic (0,1), Δ~Normal (0, 100)			β ~ Normal (0,10), precision~Gamma (0.1,0.1), θ_0 ~ logistic (0,1), Δ~Normal (0, 10).		
Parameter	Mean	SD	95 % credible interval	Mean	SD	95 % credible interval	Mean	SD	95 % credible interval
β_0 – Intercept	21.00	0.45	20.12; 21.93	21.0	0.48	20.07; 21.94	20.53	0.47	19.59; 21.46
β_1 – Week	−1.17*	0.05*	−1.28; −1.06*	−1.17*	0.06*	−1.28; −1.06*	−1.15*	0.05*	−1.26; −1.04*
β_2 – Sertraline	−0.21	0.66	−1.49; 1.08	−0.20	0.67	−1.49; 1.14	0.31	0.68	−1.02; 1.65
β_3 – Hypericum	0.76	0.65	−0.52; 2.04	0.74	0.68	−0.60; 2.09	1.19	0.65	−0.09; 2.48
β_4 – Interaction week and sertraline	−0.16*	0.08*	−0.31; −0.01*	−0.16*	0.08*	−0.32; −0.01*	−0.19*	0.08*	−0.34; −0.03*
β_5 – Interaction week and hypericum	0.07	0.08	−0.08; 0.22	0.07	0.08	−0.08; 0.23	0.05	0.08	−0.10; 0.20
Δ	0.01	0.05	−0.18; 0.07	0.02	0.03	−0.02; 0.07	0.02	0.02	−0.02; 0.07
θ_0	−4.12*	0.84*	−5.18; −0.89*	−4.28*	0.49*	−5.27; −3.37*	−4.30*	0.47*	−5.29; −3.43*
Standard deviation from τ	3.94	0.06	3.82; 4.07	3.94	0.06	3.82; 4.07	3.94	0.06	03.82; 4.07
Standard deviation from τ2	4.36	0.19	4.00; 4.75	4.36	0.19	4.00; 4.75	4.37	0.19	4.01; 4.76

Missing not at random model 2

	2nd MNAR model: prior set1	2nd MNAR model: prior set 2	2nd MNAR model: prior set 3
	β ~ Normal (0,10 000), precision~Gamma (0.001, 0.001), θ_0,θ_1 and θ_2 ~ logistic (0,1)	β ~Normal (0,1000), precision~Gamma (0.01, 0.01), θ_0,θ_1 and θ_2 ~ logistic (0,1),	β ~ Normal (0,10), precision~Gamma (0.1,0.1), θ_0,θ_1 and θ_2 ~ logistic (0,1)

Table 4 (continued)

Parameter	Mean	SD	95% credible interval	Mean	SD	95% credible interval	Mean	SD	95% credible interval
β_0 – Intercept	21.02	0.50	20.03; 21.98	21.12	0.48	20.18; 22.03	20.7	0.46	19.77; 21.61
β_1 – Week	−1.18*	0.05*	−1.28; −1.07*	−1.18*	0.06*	−1.3; −1.08*	−1.16*	0.05*	−1.27; −1.06*
β_2 – Sertraline	−0.10	0.73	−1.49; 1.35	−0.28	0.67	−1.56; 1.04	0.17	0.66	−1.10; 1.48
β_3 – Hypericum	0.79	0.71	−0.58; 2.21	0.67	0.69	−0.69; 2.00	1.06	0.66	−0.23; 2.34
β_4 – Interaction week and sertraline	−0.17*	0.08*	−0.33; −0.01*	−0.16	0.08	−0.31; <0.001	−0.18*	0.08*	−0.33; −0.03*
β_5 – Interaction week and hypericum	0.07	0.08	−008; 0.22	0.08	0.08	−0.07; 0.24	0.06	0.08	−0.10; 0.21
Delta	−0.22	99.4	−195; 191	0.11	10.05	−19.74; 19.85	−0.04	3.15	−6.19; 6.21
θ_0	−4.35*	0.59*	−5.59; −3.26*	−4.35*	0.55*	−5.51; −3.3*	−4.31*	0.55*	−5.43; −3.30*
θ_1	0.03	0.03	−0.02; 0.09	0.03	0.03	−0.02; 0.09	0.03	0.03	−0.02; 0.08
θ_2	0.20*	0.07*	0.04; 0.34*	0.20*	0.07*	0.05; 0.33*	0.19*	0.07*	0.03; 0.32*
Standard deviation from τ	3.97	0.06	3.84; 4.10	3.97	0.07	3.84; 4.10	3.96	0.07	3.84; 4.09
Standard deviation from τ2	0.05	0.19	0.04; 0.06	4.37	0.19	4.02; 4.76	4.37	0.19	4.01; 4.75
Missing not at random model 3									
	3rd MNAR model: prior set1 β ~ Normal(0, 10000), precision~Gamma (0.001, 0.001), θ_0 ~ logistic (0,1), θ_1,$\theta_2$andθ_3 ~ Normal (0, 10 000)			3rd MNAR model: prior set 2 β ~ Normal(0,1000), precision~Gamma (0.01, 0.01), θ_0 ~ logistic (0,1), θ_1,θ_2,θ_3 ~ Normal (0, 100)			3rd MNAR model: prior set 3 β ~ Normal(0,10), precision~Gamma (0.1,0.1), θ_0 ~ logistic (0,1), θ_1,θ_2,θ_3 ~ Normal (0, 10)		
β_0 Intercept	21.03	0.47	20.09; 21.97	20.98	0.48	20.03; 21.91	21.0	0.45	20.13; 21.89
β_1 Week	−1.17*	0.05*	−1.28; −1.07*	−1.17*	0.05*	−1.28; −1.06*	−1.17*	0.05*	−1.28; −1.06*
β_2 Sertraline	−0.23	0.68	−1.56; 1.11	−0.17	0.68	−1.51; 1.17	−0.19	0.64	−1.45; 1.04
β_3 Hypericum	0.72	0.67	−0.60; 2.04	0.78	0.67	−0.55; 2.08	0.76	0.65	−0.53; 2.04
β_4 Interaction week and sertraline	−0.16*	0.08*	−0.31; −0.01*	−0.16*	0.08*	−0.31; −0.01*	−0.16*	0.08*	−0.31; −0.01*
β_5 Interaction week and hypericum	0.07	0.08	−0.08; 0.22	0.07	0.08	−0.08; 0.22	0.07	0.08	−0.08; 0.22
θ_0	−4.29*	0.47*	−5.23; −3.42*	−4.31*	0.50*	−5.36; −3.37*	−4.29*	0.48*	−5.27; −3.36*
θ_1	0.04	0.03	−0.02; 0.09	0.04	0.03	−0.02; 0.09	0.04	0.03	−0.02; 0.09
θ_2	−0.01	0.02	−0.05; 0.03	−0.01	0.02	−0.04; 0.03	−0.01	0.02	−0.04; 0.03
θ_3	−0.04	0.02	−0.09; 0.002	−0.04	0.02	−0.09; 0.00	−0.04*	0.02*	−0.09; −0.001*
σ	3.94	0.06	3.82; 4.07	3.94	0.06	3.82; 4.07	3.94	0.06	3.82; 4.07
σ2	4.36	0.19	4.00;4.75	4.36	0.19	4.01; 4.75	4.36	0.19	4.01; 4.75

Letting Y_{ij} be the HAM-D score for participant i at occasion j, where Y_{ij}~Normal(μ_{ij},σ^2), the following Bayesian model was fitted: $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ where, t_{ij} indicates occasion j ($week_j$) for participant i. S_i is an indicator variable for the sertraline arm and equals 1 if participant i belongs to the sertraline arm and is 0 otherwise. Similarly, H_i is an indicator variable for the hypericum arm and equals 1 if participant i belongs to the hypericum arm and is 0 otherwise. A participant i belonging to the placebo arm will thus have $H_i=S_i=0$. The placebo group is therefore the reference group. We assume N independent participants

SD Standard deviation

*Statistically significant at 0.05 level

data using methods that are appropriate with missing data. We fitted various models using different assumptions about the missing data and various analysis methods. Under this extended sensitivity analysis, we draw a different conclusion.

The missing data in this study was not missing completely at random. We did a range of sensitivity analyses under missing at random and missing not at random assumptions. Some of our conclusions are similar to the original analysis, but both the missing at random and missing not at random analyses using Bayesian methods lead to the conclusion that sertraline is significantly better than placebo in reducing depression symptoms over 8 weeks. No difference was found between hypericum and placebo in any of the analyses. Our sensitivity analysis penalized the sertraline group less than the previous analysis and therefore showed that there was a difference between sertraline and placebo. The change from baseline to week 8 for sertraline was −10.64 (95 % CI −11.52, −9.77) and placebo was −9.36 (95 % CI −10.2, −8.52) according to the missing at random Bayesian model with prior set 1.

While there are strong similarities between likelihood and Bayesian missing at random analyses, a key difference is the absence versus presence of a prior specification. The impact of

Table 5 Summary of all the statistical methods used; their key features and main assumptions

Method	Key features	Assumptions
Missing at random		
Likelihood-based approaches	Parametric model; Draw inference about a parameter, θ, in a model f (y|θ) for the response data that is not fully observed. Model with fixed and random effects, including treatment, week and the interaction between treatment and week, adjusted for repeated measures	Missing at random Missing data mechanism is ignorable. No need to specify a model that relates the missing data process to the observed data.
Multiple imputation	Produce several different imputed data sets. The imputed values are random draws from the posterior predictive distribution of the missing data, given the observed data. Apply likelihood-based estimation methods to each data set. Parameter estimates are averaged across the several analyses. Standard errors are calculated using Rubin's (1987) formula that combines variability within and between data sets. The imputation model included: HAM-D, age, sex, race, duration of depression, BDI, CGI-S, CGI-I, and GAF scale at baseline.	Missing at random Data are from a multivariate normal distribution. Missing values can occur on any of the variables. Missing data mechanism is ignorable.
Bayesian missing at random model	Bayesian model: Y_{ij}~Normal(μ_{ij},σ^2) $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ The Bayesian inference is done by specifying a model and prior distributions for the parameters of the model, and then updating the prior information on the parameters using the model specified and the data observed to obtain the posterior distribution of the parameters. Vague priors were specified for the unknown parameters and are given in Table 4.	Missing at random Parameters are treated as random variables. Probability statements are made about the model parameters and not about the data
Missing not at random		
Bayesian missing not at random model 1	Bayesian model: Y_{ij}~Normal(μ_{ij},σ^2) $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ Plus a model of missingness: $logit\ p_{i,w}=\theta_0+\theta_1 y_{i,w-1}+\theta_2(y_{i,w}-y_{i,w-1})$. Key features as described for previous missing at random model. Vague priors were specified for the unknown parameters and are given in Table 4.	Missing not at random Parameters are treated as random variables. Probability statements are made about the model parameters and not about the data
Bayesian missing not at random model 2	Bayesian model: Y_{ij}~Normal(μ_{ij},σ^2) $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ Plus a model of missingness: $logit\ p_{i,w}=\theta_0+\theta_1 y_{i,w-1}+\theta_2(y_{i,w}-y_{i,w-1})$. Key features as described for previous missing not at random model. Vague priors were specified for the unknown parameters and are given in Table 4.	Missing not at random Parasmeters are treated as random variables. Probability statements are made about the model parameters and not about the data
Bayesian missing not at random model 3	Bayesian model: Y_{ij}~Normal(μ_{ij},σ^2) $\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 S_i+\beta_3 H_i+(\beta_4 S_i+\beta_5 H_i)t_{ij}$ Plus a model of missingness: $logit\ p_{ij}=\theta_0+\theta_1 y_{ij}+\theta_2 y_{ij} S_i+\theta_3 y_{ij} H_i$. Key features as described for previous missing not at random model. Vague priors were specified for the unknown parameters and are given in Table 4.	Missing not at random. This model allows the missingness to depend on the unobserved HAM-D value, while allowing for a different mechanism in each treatment arm by including the HAM-D score by treatment interaction.

the prior is perhaps one of the most studied topics in Bayesian analyses, already in the context of no missing data. It suggests that the missing at random based Bayesian analysis can be more sensitive to assumptions made than the missing at random based likelihood or multiple imputation analyses, simply because more assumptions have to be made.

This implies that the conclusion reached 10 years ago could be amended to state that hypericum does not seem to provide any benefit over placebo, in a trial where it could not be ruled out that the active comparator could provide a slight benefit over placebo. This illustrates the point that not taking account of missing data in the analysis could introduce bias and lead to incorrect results.

Adjusting the analysis to take missing data into account does not imply changing the proposed estimate of effectiveness. The measure of effectiveness reported in the original paper was change from baseline to week 8. We analyzed the same estimate under missing at random assumptions using either multiple imputation or likelihood-based methods (Table 5). In general, multiple imputation allows any measure of effectiveness, since the analysis of choice is done in the second step.

The trial design only continued participants with a full response at week 8 to the continuation phase. Thus only a fraction of the participants (37.9 %) will have data in the continuation phase. Because of nonrandom exclusion of participants, the comparability of the three treatment arms after week 8 is not equivalent to a randomized trial. At best, this provides an observational study about the longer term effects of the drugs. Any efficacy analysis in this continuation phase should be interpreted with caution. This design should be discouraged, unless the objective is to estimate sustained response in those who responded initially. Because of the small number of participants in the continuation phase, the correct handling of missing data in this phase is important. The missing data mechanism is probably missing at random by design, since missingness can be predicted by the response to treatment at week 8. Missing not at random missing data cannot be excluded either, since additional mechanisms could also contribute to the missing data in this phase. It becomes even more important to analyze the data from week 8 onwards using appropriate methods for missing data.

The analysis was done with standard statistical software, using resources that should be available to most researchers. The unavailability of software should no longer be a reason not to do the proper principled analyses in the presence of missing data.

Conclusion

There is no difference between hypericum and placebo, regardless of the assumption about the missing data process, but there is a significant difference between sertraline and placebo with some of the analyses assuming a missing at random missing data process and when a missing not at random missing data process is assumed. The assumptions about the missing data process could influence the results, as is shown by this example. This reanalysis of the original data, using proper missing data processes, changes the original conclusion of the trial. The original conclusion was that the trial was inconclusive, since the active control arm was not superior to placebo. The findings using these methods conclude that the sertraline arm could be superior to placebo under certain assumptions about the missing data process. This means that the original trial was not inconclusive, but found that hypericum was not superior to placebo. It is important to conduct an analysis that takes account of missing data using valid statistically principled methods.

Acknowledgments Data used in the preparation of this article were obtained from the limited access datasets (version 4.1) distributed from the NIH-supported "A Placebo-Controlled Clinical Trial of a Standardized Extract of H. perforatum in Major Depressive Disorder" (Hypericum). This is a multisite, clinical trial of persons with depression comparing the effectiveness of randomly assigned medication treatment. The purpose of this trial is to study the acute efficacy and safety of a standardized extract of the herb *H. perforatum* (St. John's wort) in the treatment of patients with major depression. The study was supported by NIMH contract # N01MH70007 to the Duke University Medical Center. The ClinicalTrials.gov identifier is NCT00005013. This manuscript reflects the views of the authors and may not reflect the opinions or views of the Hypericum Study Investigators or the NIH. Anneke Grobler had full access to the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Geert Molenberghs gratefully acknowledges financial support from the IAP research Network P7/06 of the Belgian Government (Belgian Science Policy).

Conflict of interest The authors declare that they have no conflicts of interest.

References

Carpenter J, Kenward MG (2007) Missing data in randomised controlled trials—a practical guide

Carpenter J, Kenward MG, Evans S, White I (2004) Last observation carried forward and last observation analysis. Letter to the editor. Statistics in medicine 23

Carpenter J, Pocock S, Lamm CJ (2002) Coping with missing data in clinical trials: a model-based approach applied to asthma trials. Statistics in medicine 21:1043–1066

Carpenter J, Roger J, Kenward M (2013) Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions and inference via multiple imputation. J Biopharm Stat 23:1352–1371

CHMP (2010) Guideline on missing data in confirmatory clinical trials. European Medicines Agency, London

Daniels MJ, Hogan JW (2008) Missing data in longitudinal studies. Strategies for Bayesian modeling and sensitivity analysis, 1st edn. Chapman & Hall/CRC, Boca Raton, FL

Fegert J, Kolch M, Zito JM, Glaeske G, Janhsen K (2006) Antidepressant use in children and adolescents in Germany. J Child Adolesc Psychopharmacol 16:197–206

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesion restoration of images. IEEE Trans on Pattern Anal and Mach Intell 6:721–741

Hamilton M (1960) A rating scale for depression. J Neurol, Neurosurg Psychiatry 23:56–62

Horton NJ, Kleinman KP (2007) Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 61:79–90

Hypericum Depression Trial Study Group (2002) Effect of *Hypericum perforatum* (St John's wort) in major depressive disorder. J Am Med Assoc 287:1807–1814

Jansen I, Hens N, Molenberghs G, Aerts M, Verbeke G, Kenward MG (2006) The nature of sensitivity in missing not at random models. Comput Stat and Data Anal 50:830–858

Kalb R, Trautmann-Sponsel RD, Kieser M (2001) Efficacy and tolerability of Hypericum extract WS 5572 versus placebo in mildly to moderately depressed patients. Pharmacopsychiatry 34:96,103

Kasper S, Gastpar M, Moller HJ, Muller WE, Volz HP, Dienel A, Kieser M (2010) Better tolerability of St. John's wort extract WS 5570 compared to treatment with SSRIs: a reanalysis of data from controlled clinical trials in acute major depression. Int Clin Psychopharmacol 25:204–213

Linde K, Berner MM, Kriston L (2008) St John's wort for major depression. Cochrane Database of Systematic Reviews 4.

Linde K, Ramirez G, Mulrow CD, Pauls A, Weidenhammer W, Melchart D (1996) St John's wort for depression—an overview and meta-analysis of randomised clinical trials. BMJ 313:253–258

Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD, Rotnitzky A, Scharfstein D, Shih WJ, Siegel JP, Stern H (2012) The prevention and treatment of missing data in clinical trials. N Engl J Med 367:1355–1360

Mallinckrodt CH, Clark WS, David SR (2001) Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward. Drug Inf J 35:1215–1225

Mason A, Richardson S, Plewis I, Best N (2012) Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. J Offician Stat 28:279–302

Michiels B, Molenberghs G (1997) Protective estimation of longitudinal categorical data with nonrandom dropout. Commun in Stat, Theory and Methods 26:65–94

Molenberghs G, Kenward MG (2007) Missing data in clinical studies. Wiley, Chichester

Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, Carroll RJ (2004) Analyzing incomplete longitudinal clinical trial data. Biostatistics 5:445–464

National Research Council (2010) The prevention and treatment of missing data in clinical trials. The National Academic Press, Washington DC

Rahimi R, Nikfar S, Abdollahi M (2009) Efficacy and tolerability of *Hypericum perforatum* in major depressive disorder in comparison with selective serotonin reuptake inhibitors: a meta-analysis. Prog Neuro-Psychopharmacol Biol Psychiatry 33:118–127

Ratitch B, O'Kelly M, Tosiello R (2013) Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. Pharmaceutical Statistics: n/a-n/a.

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Sarris J, Fava M, Schweitzer I, Mischoulon D (2012) St John's Wort (*Hypericum perforatum*) versus sertraline and placebo in major depressive disorder: continuation data from a 26-week RCT. Pharmacopsychiatry 45:275–278

Shelton RC, Keller MB, Gelenberg A, Dunner DL, Hirschfeld R, Thase ME, Russell J, Lydiard B, Crits-Christoph P, Gallop R, Todd L, Hellerstein D, Goodnick P, Keitner G, Stahl SM, Halbreich U (2001) Effectiveness of St John's wort in major depression: a randomized controlled trial. JAMA 285:1978–1986

Ware JH, Harrington D, Hunter DJ, D'Agostino RB (2012) Missing Data. N Engl J Med 367:1353–1354

White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. Stat Med 30: 377–399