# 1 *Missing Data*

***Geert Molenberghs and Emmanuel Lesaffre***

## 1.1 Introduction

Data from longitudinal studies in general, and from clinical trials in particular, are prone to incompleteness. As incompleteness usually occurs for reasons outside of the control of the investigators and may be related to the outcome measurement of interest, it is generally necessary to reflect on the process governing incompleteness. Only in special but important cases is it possible to ignore the missingness process.

When patients are examined repeatedly in a clinical trial, missing data can occur for various reasons and at various visits. When missing data result from patient dropout, the missing data have a *monotone* pattern. *Nonmonotone* missingness occurs when there are intermittent missing values as well. The focus here will be on dropout. Reasons typically encountered are adverse events, illness not related to study medication, uncooperative patient, protocol violation, ineffective study medication, loss to follow-up, and so on.

When referring to the missing-value, or nonresponse, process, we will use the terminology of Little and Rubin [1]. A nonresponse process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *nonrandom* (MNAR). In the context of likelihood and Bayesian [2] inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, MCAR and MAR are *ignorable*, whereas a nonrandom process is nonignorable. Thus, under ignorable dropout, one can literally ignore the missingness process and nevertheless obtain valid estimates of, say, the treatment. The above definitions are conditional on including the correct set of covariates into the model. An overview of the various mechanisms, and their (non-)ignorability under likelihood, Bayesian, or frequentist inference, is given in Table 1.

Consider the case in which only one follow-up measurement per patient is made. When dropout occurs in a patient, leaving the investigator without follow-up measures, one is usually forced to discard such a patient from analysis, thereby violating the intention to treat (ITT) principle, which stipulates that all randomized patients should be included in the primary analysis and according to the randomiza-

Table 1: Overview of Missing Data Mechanisms

| Acronym | Description | Likelihood/Bayesian | Frequentist |
|---|---|---|---|
| MCAR | missing completely at random | ignorable | ignorable |
| MAR | missing at random | ignorable | non-ignorable |
| MNAR | missing not at random | non-ignorable | non-ignorable |

tion scheme. Of course, the effect of treatment can be investigated under extreme assumptions, such as, for example, a worst-case and a best-case scenario, but such scenarios are most often not really helpful. The focus of this article will be on analysis techniques for repeated measurements studies.

Early work regarding missingness focused on the consequences of the induced lack of balance of deviations from the study design [3, 4]. Later, algorithmic developments took place, such as the expectation-maximization algorithm (EM) [5] and multiple imputation [6, 7]. These have brought likelihood-based ignorable analysis within reach of a large class of designs and models. However, they usually require extra programming in addition to available standard statistical software.

For a long time, clinical trial practice has put a strong emphasis on methods such as *complete case analysis* (CC) and *last observation carried forward* (LOCF) or other simple forms of imputation. Claimed advantages include computational simplicity, no need for a full longitudinal model analysis (e.g., when the scientific question is in terms of the last planned measurement occasion only), and for LOCF, compatibility with the ITT principle. However, a CC analysis assumes MCAR, and the LOCF analysis makes peculiar assumptions about the (unobserved) evolution of the response, underestimates the variability of the response, and ignores the fact that imputed values are no real data.

In recent times, concerted efforts have been done to bring clinical trial practice in line with contemporary scientific views regarding the prevention and handling of incomplete data. In 2010, a US National Academy of Sciences report was published, propagating a shift away from overly simple methods, in favor of using direct likelihood and Bayesian methods, inverse probability weighting, and multiple imputation [8, 9].

A likelihood-based or Bayesian [10] longitudinal analysis requires only MAR, uses all data (obviating the need for both deleting and filling in data), and is consistent with the ITT principle. Furthermore, it can also be shown that the incomplete sequences contribute to estimands of interest (treatment effect at the end of the study), even early dropouts. For continuous responses, the linear mixed model is popular and is a direct extension of analysis of variance (ANOVA) and MANOVA approaches, but more broadly valid in incomplete data settings. For categorical responses and count data, so-called marginal (e.g., generalized estimating equations, GEEs) and random-effects (e.g., generalized linear mixed-effects models, GLMMs) approaches are in use. Although GLMM parameters can be fitted using maximum likelihood, the same is not true for the frequentist GEE method, but modifications have been proposed to accommodate the MAR assumption [11–13].

Multiple imputation is another approach gaining clout for the analysis of incomplete clinical trial data [1, 14, 15].

Finally, MNAR missingness can never

be fully ruled out based on the observed data only. It is argued that, rather than going either for discarding MNAR models entirely or for placing full faith on them, a sensible compromise is to make them a component of a sensitivity analysis.

## 1.2 Methods in Common Use

We will focus on two relatively simple methods that have been and still are in extensive use. Detailed accounts of simple methods to handle missingness are given by various authors [8, 16–19].

### 1.2.1 Complete Case Analysis

A *complete case analysis* includes only those cases for analysis for which all measurements were recorded. This method has obvious advantages. It is very simple to describe, and because the data structure is as would have resulted from a complete experiment, standard statistical software can be used without additional work. Furthermore as the entire estimation is performed on the same subset of completers, there is a common basis for inference. Unfortunately, the method suffers from severe drawbacks. First, there is nearly always a substantial loss of information. The impact on precision and power is dramatic. Furthermore, such an analysis will only be representative for patients who remain on study. Of course a complete case analysis could have a role as an auxiliary analysis, especially if a scientific question relates to it. A final important issue about a complete case analysis is that it is only valid when the missingness mechanism is MCAR. However, severe bias can result when the missingness mechanism is MAR but not MCAR. This bias can go both ways, i.e., either overestimating or underestimating the true effect.

### 1.2.2 Last Observation Carried Forward

A method that has received a lot of attention [20–22] is the *last observation carried forward* (LOCF). As noted, in the LOCF method, whenever a value is missing, the last observed value is substituted. For the LOCF approach, the MCAR assumption is necessary but not sufficient for an unbiased estimate. Indeed, it further assumes that subjects' responses would have been constant from the last observed value to the endpoint of the trial. These conditions seldom hold [17]. In a clinical trial setting, one might believe that the response profile *changes* as soon as a patient goes off treatment and even that it would flatten. However, the constant profile assumption is even stronger. Therefore, carrying observations forward may bias estimates of treatment effects and underestimate the associated standard errors [17, 23–27]. Further more this method artificially increases the amount of information in the data, by treating imputed and actually observed values on equal footing.

Despite its shortcomings, LOCF has been the longstanding method of choice for the primary analysis in clinical trials because of its simplicity, ease of implementation, and the belief that the potential bias from carrying observations forward leads to a "conservative" analysis in comparative trials. An analysis is called conservative when it leads to no treatment difference, whereas in fact there is a treatment difference. However, reports of anticonservative or liberal behavior of LOCF are common [28–32], which means that a LOCF analysis can create a treatment effect when none exists. Thus, the statement that LOCF analysis has been used to provide a conservative estimate of treatment effect is unacceptable.

Baseline observation carried forward (BOCF) has been proposed as a method that partially overcomes LOCF's short-

comings, while being consistent with ITT. That said, it has been criticized severely as well [8, 33].

It is often quoted that LOCF or CC, although problematic for parameter estimation, produces randomization-valid hypothesis testing, but this is questionable. First, in a CC analysis, partially observed data are selected out, with probabilities that may depend on post-randomization outcomes, thereby undermining any randomization justification. Second, if the focus is on one particular time point, e.g., the last one scheduled, then LOCF plugs in data. Such imputations, apart from artificially inflating the information content, may deviate in complicated ways from the underlying data [8]. Third, although the size of a randomization-based LOCF test may reach its nominal size under the null hypothesis of no difference in treatment profiles, there will be other regions of the alternative space where the power the LOCF test procedure is equal to its size, which is completely unacceptable.

Historically, an important motivation behind the simpler methods was their ease of use. Indeed, the main advantage, shared with complete case analysis, is that complete data software can be used. However, with the availability of commercial software tools, such as, for example, the SAS procedures MIXED, NLMIXED, and GLIMMIX, and the R lme4 and nlme libraries, this motivation no longer applies. This is also the motivation of Little *et al* [8] to move away from them.

## 1.3 An Alternative Approach to Incomplete Data

A graphical illustration is first provided, using an artificial example, of the various simple methods that have been considered, and then so-called direct likelihood analysis is discussed.

### 1.3.1 Illustration of Simple Methods

Take a look at an artificial but insightful example, depicted in Figure 1, which displays the results of the traditional methods, CC and LOCF, next to the result of an MAR method. In this example, the mean response is supposed to be linear. For both groups (completers and dropouts), the slope is the same, but their intercepts differ. Patients with incomplete observations dropped out half way through the study; e.g., because they reached a certain level of the outcome. It is obviously an MAR missingness mechanism. Using a method, valid under the MAR assumption, yields the correct mean profile, being a straight line centered between the mean profiles of the completers and incompleters. If one would perform a CC analysis, the fitted profile would coincide with the mean profile of the complete cases (bold line). Next, under LOCF, data are imputed (dashed line). The resulting fitted profile will be the bold dashed line. Clearly, both traditional methods produce an incorrect result.

Furthermore, in a standard available case analysis (AC), one makes use of the information actually available. One such set of estimators could be the treatment-specific mean at several designed measurement occasions. With a decreasing sample size over time, means later in time would be calculated using less subjects than means earlier in time. Figure 1 shows a dramatic instance of this approach, evidently due to the extreme nature of this illustrative example. The key message is that such an approach cannot remove major sources of bias.

### 1.3.2 Direct Likelihood Analysis

For continuous outcomes, Verbeke and Molenberghs [17] describe likelihood-based

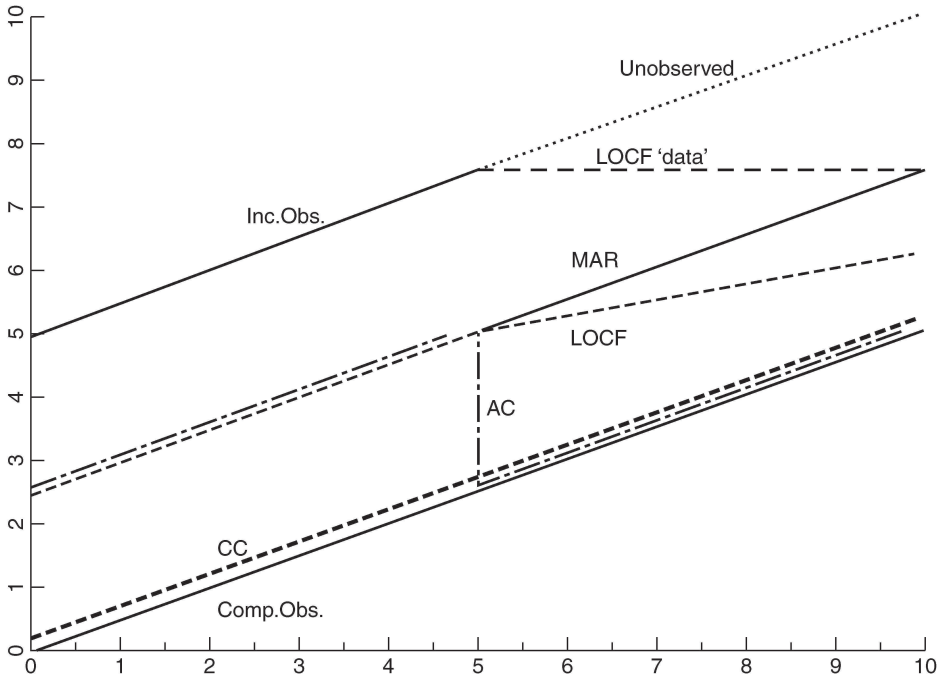


Figure 1: Artificial situation, illustrates the results of the traditional MCAR methods—CC and LOCF—next to the result of the direct likelihood method.

mixed-effects models, which are valid under the MAR assumption. Indeed, for longitudinal studies, where missing data are involved, a mixed model only requires that missing data are MAR. As opposed to the traditional techniques, mixed-effects models permit the inclusion of subjects with missing values at some time points (both dropout and intermittent missingness).

This likelihood-based MAR analysis is also termed *likelihood-based ignorable analysis* or, as used in the remainder of this article, a *direct likelihood analysis*. In such an analysis, the observed data are used without deletion nor imputation. In so doing, appropriate adjustments are made to parameters at times when data are incomplete, due to the within-patient correlation.

Thus, even when interest lies, for example, in a comparison between the two treatment groups at the last occasion, such a full longitudinal analysis is a good approach, because the fitted model can be used as the basis for inference at the last occasion.

In many clinical trials, the repeated measures are balanced in the sense that a common (and often limited) set of measurement times is considered for all subjects, which allows the *a priori* specification of a "saturated" model. For example, a full group-by-time interaction for the fixed effects combined with an unstructured covariance matrix. Such a model specification is sometimes termed mixed-effects model repeated-measures analysis (MMRM) [22]. Thus, MMRM is a particular form of a linear mixed model, relevant for acute phase confirmatory clinical trials, fitting within the direct likelihood paradigm. Moreover, this direct likelihood MMRM analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) approaches, but more generally valid when they are incomplete. This response is an unequivocal answer to the common criticism that a direct likelihood method is making strong assumptions. Indeed, its coincidence with MANOVA for data without missingness shows that the assumptions made are very mild. Therefore, it constitutes a very promising alternative for CC and LOCF. When a relatively large number of measurements is made within a single subject, the full power of random effects modeling can be used [17].

The practical implication is that a software module with likelihood estimation facilities and with the ability to handle incompletely observed subjects manipulates the correct likelihood, providing valid parameter estimates and likelihood ratio values. Note that similar arguments apply to the Bayesian case [2, 10, 13].

A few cautionary remarks are warranted. First, when at least part of the scientific interest is directed toward the nonresponse process, obviously both processes need to be considered. Under MAR, both questions can be answered separately, which implies that a conventional method can be used to study questions in terms of the outcomes of interest, such as treatment effect and time trend, whereafter a separate model can be considered to study missingness. Second, likelihood inference is often surrounded with references to the sampling distribution (e.g., to construct measures of precision for estimators and for statistical hypothesis tests [34]). However, the practical implication is that standard errors and associated tests, when based on the observed rather than the expected information matrix and given that the parametric assumptions are correct, are valid. Third, it may be hard to rule out the operation of an MNAR mechanism. This point was brought up in Section 1.1 and will be discussed further in Section 1.7.

## 1.4 Illustration: Orthodontic Growth Data

As an example, we use the orthodontic growth data, introduced by Potthoff and Roy [35] and used by Jennrich and Schluchter [36]. The data have the typical structure of a clinical trial and are simple yet illustrative. They contain growth measurements for 11 girls and 16 boys. For each subject, the distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14. Figure 2 presents the 27 individual profiles. Little and Rubin [1] deleted 9 of the $[(11 + 16) \times 4]$ measurements, rendering 9 incomplete subjects, which even though it is a somewhat unusual practice has the advantage of allowing a comparison between the incomplete data methods and the analysis of the original, complete data. Deletion is confined to the age 10 measurements, and rougly speaking, the complete observations at age 10 are those with a higher measurement at age 8. Some emphasis will be placed on ages 8 and 10, the typical dropout setting, with age 8 fully observed and age 10 partially missing.

The simple methods and direct likelihood method from Sections 1.2 and 1.3 are now compared using the growth data. For this purpose, a linear mixed model is used, assuming an unstructured mean, i.e., assuming a separate mean for each of the eight age $\times$ sex combinations, together with an unstructured covariance structure, and using maximum likelihood (ML) as well as restricted maximum likelihood (REML). The mean profiles of the linear mixed model using maximum likelihood for all four datasets, for boys, are given in Figure 3. The girls' profiles are similar and hence not shown.

Next to this longitudinal approach, a full MANOVA analysis and a univariate ANOVA analysis will be considered, i.e., one per time point. For all of these analyses, Table 2 shows the estimates and standard errors for boys at ages 8 and 10, for the original data and all available incomplete data, as well as for the CC and the LOCF data.

First, the group means for the boys in the original dataset in Figure 3 are considered; i.e., relatively a straight line is observed. Clearly, there seems to be a linear trend in the mean profile.

In a complete case analysis of the growth data, the 9 subjects that lack one measurement are deleted, resulting in a working dataset with 18 subjects. This result implies that 27 available measurements will not be used for analysis, a severe penalty on a relatively small dataset. Observing the profiles for the CC dataset in Figure 3, all group means increased relative to the original dataset but mostly so at age 8. The net effect is that the profiles overestimate the average length.

For the LOCF dataset, the 9 subjects that lack a measurement at age 10 are completed by imputing the age 8 value. It is clear that this procedure will affect the apparently increasing linear trend found for the original dataset. Indeed, the imputation procedure forces the means at ages 8 and 10 to be more similar, thereby destroying the linear relationship. Hence, a simple, intuitively appealing interpretation of the trends is made impossible.

In case of direct likelihood, two profiles can now be observed: one for the observed means and one for the fitted means. These two coincide at all ages except age 10. As mentioned, the complete observations at age 10 are those with a higher measurement at age 8. Due to the within-subject correlation, they are the ones with a higher measurement at age 10 as well, and therefore, the fitted model corrects in the appropriate direction. The consequences of this are very important. Although it is believed that the fitted means do not follow the observed means all that well, this neverthe-

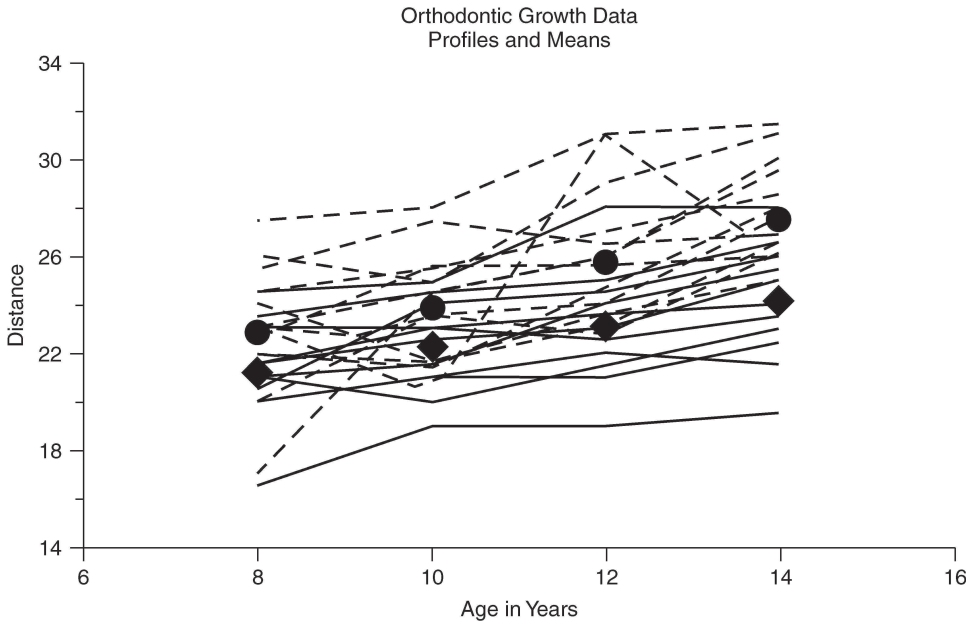


Figure 2: Orthodontic growth data. Raw and residual profiles. (Girls are indicated with solid lines. Boys are indicated with dashed lines.)

less is precisely what should be observed. Indeed, as the observed means are based on a nonrandom subset of the data, the fitted means take into account all observed data points, as well as information on the observed data at age 8, through the measurements that have been taken for such children, at different time points.

As an aside, note that, in case of direct likelihood, the observed average at age 10 coincides with the CC average, whereas the fitted average does not coincide with anything else. Indeed, if the model specification is correct, then a direct likelihood analysis produces a consistent estimator for the average profile, as if nobody had dropped out. Of course, this effect might be blurred in relatively small datasets due to small-sample variability. Irrespective of the small-sample behavior encountered here, the validity under MAR and the ease of implementation are good arguments that favor this direct likelihood analysis over other techniques.

Now compare the different methods by means of Table 2, which shows the estimates and standard errors for boys at age 8 and 10, for the original data and all available incomplete data, as well as for the CC data and the LOCF data.

Table 2 shows some interesting features. In all four cases, a CC analysis gives an upward biased estimate, for both age groups. This result is obvious, because the complete observations at age 10 are those with a higher measurement at age 8, as shown before. The LOCF analysis gives a correct estimate for the average outcome for boys at age 8. This result is not surprising because there were no missing observations at this age. As noted, the estimate for boys of age 10 is biased downward. When the incomplete data are analyzed, we see from Table 2 that direct likelihood

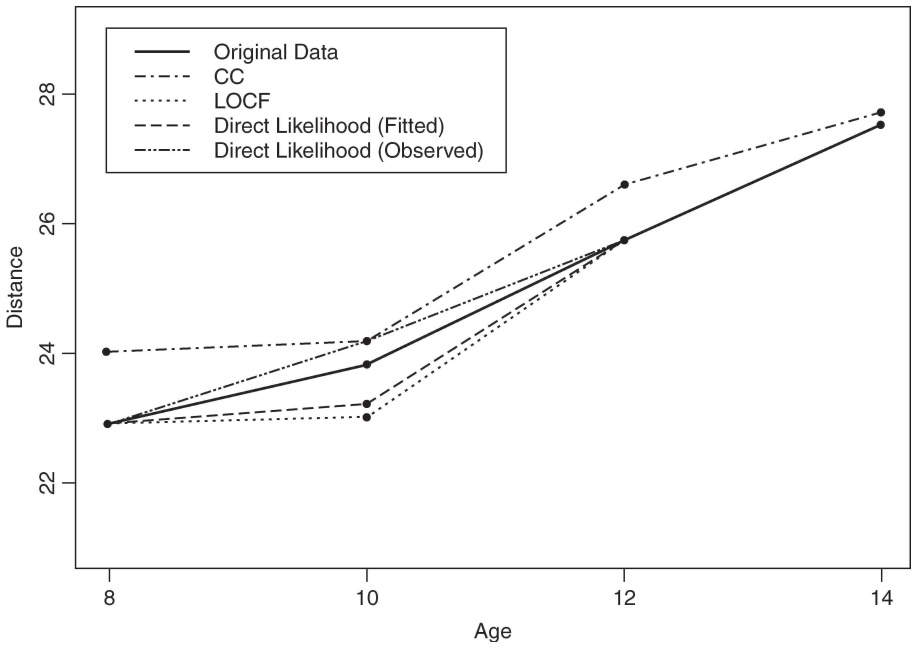


Figure 3: Orthodontic growth data. Profiles for the original data, CC, LOCF, and direct likelihood for boys.

Table 2: Orthodontic Growth Data. Comparison of analyses based on means at completely observed age 8 and incompletely observed age 10 measurement.

| Method | Boys at Age 8 | Boys at Age 10 |
|---|---|---|
| **Original Data** | | |
| Direct likelihood, ML | 22.88 (0.56) | 23.81 (0.49) |
| Direct likelihood, REML | 22.88 (0.58) | 23.81 (0.51) |
| MANOVA | 22.88 (0.58) | 23.81 (0.51) |
| ANOVA per time point | 22.88 (0.61) | 23.81 (0.53) |
| **All Available Incomplete Data** | | |
| Direct likelihood, ML | 22.88 (0.56) | 23.17 (0.68) |
| Direct likelihood, REML | 22.88 (0.58) | 23.17 (0.71) |
| MANOVA | 24.00 (0.48) | 24.14 (0.66) |
| ANOVA per time point | 22.88 (0.61) | 24.14 (0.74) |
| **Complete Case Analysis** | | |
| Direct likelihood, ML | 24.00 (0.45) | 24.14 (0.62) |
| Direct likelihood, REML | 24.00 (0.48) | 24.14 (0.66) |
| MANOVA | 24.00 (0.48) | 24.14 (0.66) |
| ANOVA per time point | 24.00 (0.51) | 24.14 (0.74) |
| **Last Observation Carried Forward Analysis** | | |
| Direct likelihood, ML | 22.88 (0.56) | 22.97 (0.65) |
| Direct likelihood, REML | 22.88 (0.58) | 22.97 (0.68) |
| MANOVA | 22.88 (0.58) | 22.97 (0.68) |
| ANOVA per time point | 22.88 (0.61) | 22.97 (0.72) |

produces good estimates. The MANOVA and ANOVA per time point analyses give an overestimation of the average of age 10, like in the CC analysis. Furthermore, the MANOVA analysis also yields an overestimation of the average at age 8, again the same as in the CC analysis.

Thus, direct likelihood shares the elegant and appealing features of ANOVA and MANOVA for fully observed data, but it is superior with incompletely observed profiles.

## 1.5 Inverse Probability Weighting

For non-Gaussian outcomes, apart from random-effects models, non-likelihood models have also been developed [18], the most popular one undoubtedly being *generalized estimating equations* (GEE) [18, 37]. This method essentially allows one to confine attention to the specification of the first moments of the outcome sequence, i.e., the mean structure. When data are incomplete, GEE is generally valid under MCAR only. Therefore, Robins, Rotnitzky, and Zhao [11, 19, 38] have developed so-called *weighted* generalized estimating equations (WGEE), as well as a number of refinements and extensions in subsequent papers, to allow use of GEE under not only MAR, but even under MNAR settings. The method rests on Horvitz-Thompson ideas [39], weighing contributions by the inverse probability of being observed. The method is elegant and enjoys good properties, but requires specification of a model for the weights. More recently, these WGEE have been extended toward so-called doubly robust estimating equations, where the weighting idea is supplemented with the use of a predictive model for the unobserved responses, given the observed ones. There are several excellent reviews [12, 38, 40]. The methodology has been extended to other non-likelihood-based methods, such as pseudo-likelihood [41].

## 1.6 Multiple Imputation

This method was introduced by Rubin in 1978 [6, 7] and has become an important approach for dealing with the statistical analysis of incomplete data. Many reviews and textbooks are available [1, 14, 15, 19]. Originally developed for sample surveys, the method has spread across a variety of statistical applications, including epidemiology, medical statistics, and in particular also clinical trials. Tools for multiple imputation have been incorporated into several standard statistical software packages.

The basic principle of multiple imputation (MI) is to replace each missing value with a set of $M$ plausible values. Each value can be considered a Bayesian draw from the conditional distribution of the missing observation given the observed data, in such a way that the set of imputations properly represents the information about the missing value that is contained in the observed data for the chosen model. The imputations produce $M$ "completed" datasets, each of which is analyzed using the method that would have been appropriate had the data been complete. The model for the latter analysis is called the *substantive* model, while that used to produce the imputations is called the *imputation* model. A key asset of the MI procedure is that, to a certain extent, these two models can be considered separately. MI is most straightforward to use under MAR, and most software implementations make this assumption. However, it is quite possible to apply it in MNAR settings [14]. Multiple imputation involves three distinct phases or, using Rubin's [6] terminology, tasks:

1. The missing values are filled in $M$ times to generate $M$ complete data sets.

2. The $M$ complete data sets are analyzed by using standard procedures.

3. The results from the $M$ analyses are combined into a single inference.

It is worth noting that the first and third tasks can be conducted by the SAS procedures MI and MIANALYZE, respectively, at least for particular imputation models. The second task is performed using one of the standard data analytic procedures. As stated earlier, implementations in R and other packages are also available.

One of the advantages of multiple imputation is that it can easily deal with situations where not only outcomes but also covariates are incomplete [15]. One then formulates a joint distribution over outcomes and covariates simultaneously, from which imputations are then drawn.

## 1.7 Sensitivity Analysis

While likelihood, Bayesian, and semiparametric methods, under the assumption of MAR, have been embraced as primary analyses for incomplete data [8], all models make assumptions about the so-called predictive distribution, i.e., the distribution governing the missing data, given the observed ones. Such assumptions are by default unverifiable from the data, while they may have an impact on the inferences drawn. It is therefore necessary to explore how sensitive the conclusions drawn are to the unverifiable assumptions. Many sensitivity analyses will take the form of exploring how deviations from MAR towards MNAR change the conclusions. There are several extensive reviews [13, 19].

We broadly define a sensitivity analysis as one in which several statistical models are considered simultaneously and/or where a statistical model is further scrutinized using specialized tools, such as diagnostic measures. This qualitative definition encompasses a wide variety of useful approaches. The simplest procedure is to fit a selected number of (MNAR) models which are all deemed plausible; alternatively, a preferred (primary) analysis can be supplemented with a number of modifications. The degree to which conclusions (inferences) are stable across such ranges provides an indication of the confidence that can be placed in them. Modifications to a basic model can be constructed in different ways.

Such analyses can be complemented with appropriate (global and/or local) influence analyses [42]. Another route is to construct pattern-mixture models, where the measurement model is considered, conditional upon the observed dropout pattern, and to compare the conclusions with those obtained from the selection model framework, where the reverse factorization is used [43, 44]. Alternative sensitivity analyses frameworks are provided by Robins, et al. [40], Forster and Smith [45] who present a Bayesian sensitivity analysis, and Raab and Donnelly [46]. A further paradigm, useful for sensitivity analysis, is so-called shared parameter models, where common latent or random effects drive both the measurement process as well as the process governing missingness [47, 48].

Nevertheless, ignorable analyses may provide reasonably stable results, even when the assumption of MAR is violated, in the sense that such analyses constrain the behavior of the unseen data to be similar to that of the observed data. A discussion of this phenomenon in the survey context has been given in Rubin, et al. [49]. These authors first argue that, in well-conducted experiments (some surveys and many confirmatory clinical trials), the assumption of MAR is often to be regarded as a realistic one. Second, and very important for confirmatory trials, an MAR analysis can be specified *a priori* without additional work relative to a situation with

complete data. Third, although MNAR models are more general and explicitly incorporate the dropout mechanism, the inferences they produce are typically highly dependent on the untestable and often implicit assumptions built in regarding the distribution of the unobserved measurements given the observed ones. The quality of the fit to the observed data need not reflect at all the appropriateness of the implied structure governing the unobserved data. Based on these considerations, it is recommended, for primary analysis purposes, the use of ignorable likelihood-based methods or appropriately modified frequentist methods. To explore the impact of deviations from the MAR assumption on the conclusions, one should ideally conduct a sensitivity analysis [17].

## 1.8 Conclusion

In conclusion, direct likelihood and Bayesian analyses, robust semi-parametric methods, or multiple imputation are preferable because they use all available information, without the need neither to delete nor to singly impute measurements or entire subjects. It is theoretically justified whenever the missing data mechanism is MAR, which is a more relaxed assumption than MCAR, necessary for simple analyses, such as CC; LOCF is even then not guaranteed to provide unbiased results. There is no distortion of statistical information, because observations are neither removed (such as in CC analysis) nor added (such as in LOCF analysis). As stated, MAR itself cannot be verified from the data, and hence a form of sensitivity analysis should ideally be conducted.

**Acknowledgments.** The authors gratefully acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy).

## References

[1] R. J. A. Little, and D. B. Rubin, *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, 2002.

[2] E. Lesaffre and A. Lawson, *Bayesian Biostatistics*. New York: John Wiley & Sons, 2012.

[3] A. Afifi, and R. Elashoff, Missing observations in multivariate statistics I: Review of the literature. *J. Am. Stat. Assoc.* 1996; **61**: 595–604.

[4] H. O. Hartley, and R. Hocking, The analysis of incomplete data. *Biometrics*, 1971; **27**: 7783–7808.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc. Series B*, 1977; **39**: 1–38.

[6] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, 1987.

[7] D. B. Rubin, Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. In: *Imputation and Editing of Faulty or Missing Survey Data*. Washington, DC: U.S. Department of Commerce, 1978, pp. 1–23.

[8] R. J. A. Little, R. D'Agostino, K. Dickerson, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. Shih, J. P. Siegel, and H. Stern, *The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials.* National Research Council. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press, 2010.

[9] C. H. Mallinckrodt, *Preventing and Treating Missing Data in Longitudinal Clinical Trials*. Cambridge: Cambridge University Press, 2013.

[10] M. J. Daniels and J. W. Hogan, *Missing Data in Longitudinal Studies: Strate-*

*gies for Bayesian Modeling and Sensitivity Analysis*. New York: Chapman & Hall/CRC, 2008.

[11] J. M. Robins, A. Rotnitzky, and L. P. Zhao, Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* 1995; **90**: 106–121.

[12] G. Fitzmaurice, M. Davidian, G. Molenberghs, and G. Verbeke, *Longitudinal Data Analysis*. Handbooks of Modern Statistical Methods. New York: Chapman & Hall/CRC, 2009.

[13] G. Molenberghs, G. Fitzmaurice, M. G. Kenward, G. Verbeke, and A. A. Tsiatis, *Handbook of Missing Data*. Handbooks of Modern Statistical Methods. New York: Chapman & Hall/CRC, 2013.

[14] J. R. Carpenter and M. G. Kenward, *Multiple Imputation and Its Applications*. Chichester: John Wiley & Sons, 2013.

[15] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall/CRC, 2010.

[16] G. Verbeke, and G. Molenberghs, *Linear Mixed Models in Practice: A SAS-Oriented Approach*. Lecture Notes in Statistics 126. New York: Springer-Verlag, 1997.

[17] G. Verbeke, and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag, 2000.

[18] G. Molenberghs and G. Verbeke, *Models for Discrete Longitudinal Data*. New York: Springer-Verlag, 2005.

[19] G. Molenberghs and M.G. Kenward, *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons, 2007.

[20] O. Siddiqui, and M. W. Ali, A comparison of the random-effects pattern mixture model with last observation carried forward (LOCF) analysis in longitudinal clinical trials with dropouts. *J. Biopharm. Stat.* 1998; **8**: 545–563.

[21] C. H. Mallinckrodt, W. S. Clark, R. J. Carroll, and G. Molenberghs, Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *J. Biopharm. Stat.* 2003; **13**: 179–190.

[22] C. H. Mallinckrodt, T. M. Sanger, S. Dube, D. J. Debrota, G. Molenberghs, R. J. Carroll, W. M. Zeigler Potter, and G. D. Tollefson, Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biol. Psychiatry*, 2003; **53**: 754–760.

[23] R. D. Gibbons, D. Hedeker, I. Elkin, D. Waternaux, H. C. Kraemer, J. B. Greenhouse, M. T. Shea, S. D. Imber, S. M. Sotsky, and J. T. Watkins. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Arch. Gen. Psychiatry*, 1993; **50**: 739–750.

[24] A. Heyting, J. Tolboom, and J. Essers. Statistical handling of dropouts in longitudinal clinical trials. *Stat. Med.* 1992; **11**: 2043–2061.

[25] P. W. Lavori, R. Dawson, and D. Shera. A multiple imputation strategy for clinical trials with truncation of patient data. *Stat. Med.* 1995; **14**: 1913–1925.

[26] C. H. Mallinckrodt, W. S. Clark, and R. D. Stacy. 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 Inform. J.* 2001; **35**(4): 1215–1225.

[27] C. H. Mallinckrodt, W. S. Clark, and R. D. Stacy. Accounting for dropout bias using mixed-effects models. *J. Biopharm. Stat.* 2001; **11**(1 & 2): 9–21.

[28] M. G. Kenward, S. Evans, J. Carpenter, and G. Molenberghs. *Handling missing responses: Time to leave Last Observation Carried Forward (LOCF) behind, Submitted for publication*.

[29] G. Molenberghs, H. Thijs, I. Jansen, C. Beunckens, M. G. Kenward, C. Mallinckrodt, and R. J. Carroll. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 2004; **5**: 445–464.

[30] C. H. Mallinckrodt, J. G. Watkin, G. Molenberghs, and R. J. Carroll. Choice of the primary analysis in longitudinal clinical trials. *Pharm. Stat.* 2004; **3**: 161–169.

[31] R. J. A. Little, and L. Yau. Intent-to-treat analysis in longitudinal studies with drop-outs. *Biometrics* 1996; **52**: 1324–1333.

[32] G. Liu and A. L. Gould. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *J. Biopharm. Stat.* 2002; **12**: 207–226.

[33] M. G. Kenward and G. Molenberghs, Last observation carried forward: a crystal ball? *J. Biopharm. Stat.* 2009; **19**: 872–888.

[34] M. G. Kenward, and G. Molenberghs, Likelihood based frequentist inference when data are missing at random. *Stat. Sci.* 1998; **12**: 236–247.

[35] R. F. Potthoff, and S. N. Roy, A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 1964; **51**: 313–326.

[36] R. I. Jennrich, and M. D. Schluchter, Unbalanced repeated measures models with structured covariance matrices. *Biometrics* 1986; **42**: 805–820.

[37] K.-Y. Liang and S. L. Zeger, Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.

[38] A. A. Tsiatis, *Semiparametric Theory and Missing Data*. New York: Springer-Verlag 2006.

[39] W. G. Cochran, *Sampling Techniques*. New York: John Wiley & Sons, 1977.

[40] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein, Semiparametric regression for repeated outcomes with non-ignorable non-response. *J. Am. Stat. Assoc.* 1998; **93**: 1321–1339.

[41] G. Molenberghs, M. G. Kenward, G. Verbeke, and B. Teshome Ayele, Pseudo-likelihood estimation for incomplete data. *Stat. Sinica* 2011; **21**: 187–206.

[42] G. Verbeke, G. Molenberghs, H. Thijs, E. Lesaffre, and M. G. Kenward, Sensitivity analysis for non-random dropout: A local influence approach. *Biometrics* 2001; **57**: 7–14.

[43] B. Michiels, G. Molenberghs, L. Bijnens, T. Vangeneugden, and H. Thijs. Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Stat. Med.* 2002; **21**: 1023–1041.

[44] H. Thijs, G. Molenberghs, B. Michiels, G. Verbeke, and D. Curran, Strategies to fit pattern-mixture models. *Biostatistics* 2002; **3**: 245–265.

[45] J. J. Forster, and P. W. Smith, Model-based inference for categorical survey data subject to non-ignorable non-response. *J. Roy. Stat. Soc. Series B* 1998; **60**: 57–70.

[46] G. M. Raab, and C. A. Donnelly, Information on sexual behaviour when some data are missing. *Appl. Stat.* 1999; **48**: 117–133.

[47] M. C. Wu, and K. R. Bailey, Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics* 1989; **45**: 939–955.

[48] M. C. Wu, and R. J. Carroll, Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988; **44**: 175–188.

[49] D. B. Rubin, H. S. Stern, and V. Vehovar, Handling "don't know" survey responses: The case of the Slovenian plebiscite. *J. Am. Stat. Assoc.* 1995; **90**: 822–828.

GEERT MOLENBERGHS
EMMANUEL LESAFFRE
KU Leuven Leuven, Belgium
Erasmus University Rotterdam, the Netherlands
Hasselt University, Belgium