SMGroup

SM Journal of Public Health and Epidemiology

Editorial

Should we Ban the use of 'Last Observation Carried forward' Analysis in Epidemiological Studies?

Shoop SJW*

Arthritis Research UK Centre for Epidemiology, University of Manchester, UK

Article Information

Received date: May 25, 2015
Accepted date: June 04, 2015
Published date: Jun 12, 2015

***Corresponding author**

Stephanie JW Shoop, Arthritis Research UK Centre for Epidemiology, University of Manchester, Manchester, M13 9PT, 01612751655, UK, Email: stephanie.shoop@postgrad.manchester.ac.uk

What is missing data?

Whenever patients are involved in research, the occurrence of missing information is inevitable. Examples of missing data include missing data points, as in incomplete forms, or loss of entire follow-ups due to patient attrition [1]. These missing data are a plague upon an Epidemiologist's research, affecting the quality of both the datasets and conclusions that can be drawn.

Why is missing data important?

Frequentist statistical models require information from all patients on all included variables. If missing data are ignored, the analyses become 'Complete Case' (CC) and all patients with missing data are dropped [1]. Here, valuable data, both in terms of information and the expense of recruitment, are lost. In many cases, using CC data will reduce sample sizes and therefore the power of the study [2]. To maximise statistical power of collected data and therefore be able to draw more accurate conclusions, information on all patients should be utilised.

The types of missing data: MCAR, MAR and MNAR

The first consideration is in what form the data are missing. Three types of 'missingness' have been described by Little and Rubin [3] (Table 1).

MCAR

In the first instance, data Missing Completely At Random (MCAR), missing data do not depend on observed or un-observed values (Table 1) [3]. For example, in epidemiological studies, patients may have moved house or a questionnaire was lost in the post, both of which are usually unrelated to the risk factors or outcomes of interest. In these cases, patients who have dropped-out should be very similar to patients who remain in the study.

MAR

Where data are Missing at Random (MAR), missing data depend only on variables which have been collected and therefore are observed in the dataset (Table 1) [3]. An example in epidemiological studies is where routing clinical data is used from multiple clinicians. Each clinician or hospital may collect patient data differently. In rheumatology, the total number of swollen or tender joints may be collected more frequently by certain clinicians [4], leading to missing data on 'joint count'. Where patients have missing 'joint counts', the values should be similar to patients where this information has been collected. It is likely only due to observed data, i.e. the 'clinician' that the data are missing.

MNAR

The final type of missingness, data Missing Not at Random (MNAR), is both the most difficult to assess and subsequently to deal with. If data is MNAR, the missing data depend on information which is not observed i.e. the missing values themselves (Table 1). In placebo-controlled clinical trials of drug interventions, patients in the placebo arm may drop out due to a lack of efficacy [5]. In this case, the outcome 'efficacy' is directly related to the values of the missing data. MNAR data may also partially relate to observed data (Table 1) [3], for example, patients with greater work hours may have a greater incentive to drop out if they feel their treatment is ineffective.

What is last observation carried forward?

Various methods have been suggested in order to use data on all patients recruited to an epidemiological study. These range from Last Observation Carried Forward (LOCF) to multiple imputation and Bayesian imputation methods. In this initial method, LOCF, the last captured value is carried forward when missing data arise, and is assumed not to change over time [6]. This

means that when a patient is lost to follow-up, their last values, for example of disease activity, are presumed to stay exactly same. LOCF is a highly popular tool for dealing with missing data [6], despite the strong assumptions it requires (Box 1).

- Patient information or outcomes do not change when data becomes missing
- A single data point can be used to estimate a distribution of potential values.

Box 1: Assumptions under LOCF [7].

Table 1: Types of data that missingness mechanisms relate to [3].

Type of data missingness	Missing relies on:	
	Observed data	Un-observed data
Missing Completely at Random (MCAR)		
Missing at Random (MAR)	✓	
Missing Not at Random (MNAR)	(✓)	✓

Is LOCF an appropriate method of missing data inference?

In certain circumstances, LOCF may be appropriate. For example, where variables are fixed and are guaranteed not to change over time, the assumptions outlined in Box 1 may hold. This is evident in cohort studies when patient demographics, for example date of birth, will not change. Another example is of genetic mutations such as in BRCA1/2 genes in patients at risk of breast and ovarian cancer [8], which do not change over time. In addition, the use of all patient data in LOCF will result in increased study power [2], although alternatives to LOCF also result in increased power [7].

However, where variables change over time, assuming that they remain constant will result in inaccuracies that may drastically affect any inferences drawn. This problem is not restricted to any one type of missingness mechanism (Table 1); for data MCAR, MAR and MNAR, LOCF assumes that the passing of time does have no effect on the measured variables [6].

For MCAR or MAR mechanisms, missing data are similar to present data and are therefore easily inferred from information on other patients using techniques such as multiple imputation [9]. However, LOCF ignores these sources. In both clinical trials and observational cohorts, patient condition is likely to change over time. For clinical trials, the intervention and placebo arms will likely both experience improvements in disease status, from a combination of an active substance or placebo effect [10]. For patients with missing data at follow-up, carrying forward previous disease state will result in a population with a poorer average disease status since a poorer disease status is carried forward. In observation studies, particularly those with a long follow-up period, increasing age of the participants may coincide with worsening conditions or co morbidities. Carrying forward previous disease state in these patients will result in wrongly inferring a healthier patient population.

For data MNAR, the state of the missing data depends on the data itself, and is therefore more difficult to infer. In placebo-controlled clinical trials of drugs, greater drop-out may be experienced in the intervention arm compared with the placebo-controlled arm due to increased side-effects associated with an active substance [11]. In this case, assuming a previous disease state will wrongly increase the safety profile of the active drug. Figure 1 highlights a basic example whereby in a trial of 200 patients, greater drop-out is experienced in the drug arm at the point of adverse events (Figure 1). If no data were missing, patients in the intervention arm would be at five times greater risk of suffering adverse events (Figure 1a). However, if half of the patients experiencing adverse events in the drug arm are lost to follow-up, missing data become MNAR. If the previous disease state, i.e. no adverse events, were inferred for missing patients using LOCF analysis, the relative risk is halved (Figure 1b). In fact, by using complete analysis and ignoring all lost patient data, a slightly more accurate relative risk is achieved (Figure 1c). However, by ignoring the reason that these patients dropped-out, both latter results may lead to unexpected adverse events if the drug is brought to market.

In longitudinal observational studies from hospital populations, patients may not return to the clinic if their condition has improved. If LOCF were used in this circumstance, the effect of an intervention may be wrongly exaggerated (Figure 2). Figure 2 describes a cohort in which patients have been prescribed one of two drugs: Drug A, which is given intravenously at the surgery, or Drug B, which is given orally and can be taken at home. The primary outcome is improvement. If the data were complete in this scenario, the drugs perform similarly with relative risk of 1 (Figure 2a). However, greater adherence may occur in drug A that has intra-venous rather than oral administration. Therefore, if more patients on drug B than drug A drop-out, the missing data again become MNAR. If LOCF analyses

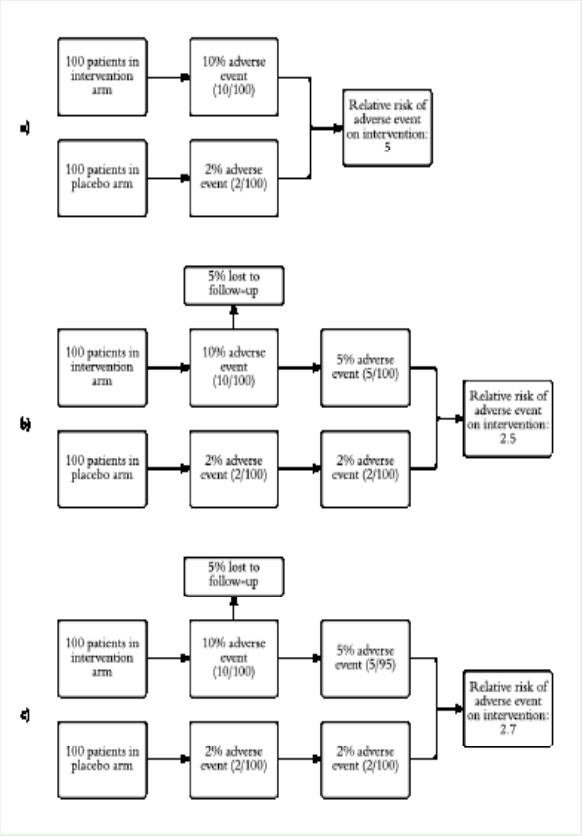

Figure 1: The relative risk of adverse event on an intervention compared to placebo in a clinical trial under the following scenarios: a) No missing data; b) Outcome data MNAR and LOCF analysis used; and c) Outcome data MNAR and CC analysis used.

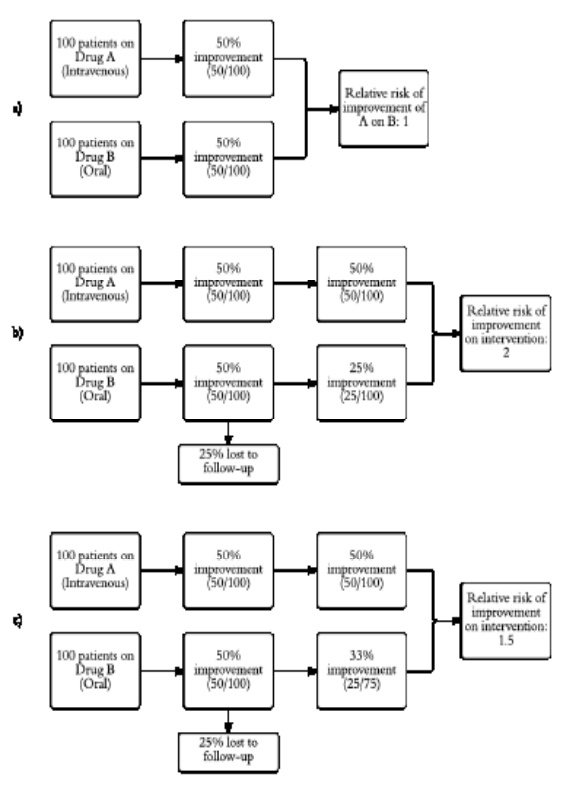

Figure 2: The relative risk of improvement of an intravenous drug compared with an orally administered drug in an observational epidemiological study. The relative risk is calculated under the following scenarios: a) No missing data; b) Outcome data MNAR and LOCF analysis used; and c) Outcome data MNAR and CC analysis used.'

were used assuming the previous disease state, i.e. not improved, for patients that dropped-out, the relative risk of improvement from Drug A doubles (Figure 2b). Again, if CC analyses were undertaken, the relative risk is an intermediate between the two analyses at 1.5 (Figure 2c).

Other than the potential inaccurate values from LOCF analysis, the precision of these estimates is, too, often flawed. Since data are missing, the true values fall within a spectrum, or distribution, of potential values. However, a single data point is used to estimate this distribution in LOCF (Box 1). Ignoring this distribution and only imputing a single point results in greater precision of the eventual point estimates [7], potentially misleading readers about the variability of the results gained.

Whilst using all patient data is a definite advantage of LOCF, this may not outweigh the disadvantages highlighted. This is particularly evident in light of statistically principled alternatives to LOCF. The use of multiple imputation or Bayesian methods also utilise all patient data and are widely available to epidemiologists including in free software packages such as in 'R'. However, by using not only missing patients' data but those from every other participant, and by taking variability in the imputation process into account, the inferences from these methods are far superior to that of LOCF. In addition, adjustments can be made for data MCAR, MAR and MNAR leading to more accurate reporting.

Conclusion

The use of LOCF is statistically un-principled, with assumptions that are only occasionally justifiable. Whilst this method may be applicable in rare circumstances, the alternatives should be promoted for all epidemiological researchers and may hopefully result in better quality inferences, and therefore more accurate results to translate into clinical practice.

References

1. Graham JW. Missing Data: Analysis and Design (Statistics for Social and Behavioural Sciences). 2012.
2. Streiner D, Geddes J. Intention to treat analysis in clinical trials when there are missing data. Evid. Based. Ment. Health. 2001; 4: 70-71.
3. Little RJA, Rubin DM. Statistical analysis with missing data. New York: John Wiley and Sons. 1987.
4. Pincus T, Segurado OG. Most visits of most patients with rheumatoid arthritis to most rheumatologists do not include a formal quantitative joint count. Ann Rheum Dis. 2006; 65: 820-822.
5. Boers M. Add-on or step-up trials for new drug development in rheumatoid arthritis: a new standard? Arthritis Rheum. 2003; 48: 1481-1483.
6. Chandan S, Jones MP. Bias in the last observation carried forward method under informative dropout. Journal of Statistical Planning and Inference. 2009; 139: 246-255.
7. Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G. Handbook of Missing Data Methodology. Florida: Chapman & Hall/CRC. 2014.
8. Narod SA. Modifiers of risk of hereditary breast and ovarian cancer. Nat.Rev. Cancer. 2002; 2: 113-123.
9. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009; 338: b2393.
10. Altman DG. Practical Statistics for Medical Research. Florida: Chapman & Hall/CRC. 1990.
11. Glasser SP. Essentials of Clinical Research. Springer.