When it is better to estimate a slope with only one point

Rodolphe Thiebaut [1 2 *] **, Walker Sarah** [3]

[1] *Centre épidémiologie et biostatistique INSERM : U897, Université Victor Segalen - Bordeaux II, FR*

[2] *Department of Infectious Diseases and Microbiology Institute of Child Health, University College London, GB*

[3] *Medical Research Council Clinical Trial Unit, London,GB*

* Correspondence should be adressed to: Thiebaut Rodolphe <rodolphe.thiebaut@isped.u-bordeaux2.fr>

MESH Keywords Bias (Epidemiology) ; Biological Markers ; Computer Simulation ; standards ; Data Interpretation, Statistical ; Humans ; Models, Statistical

One of the five basic postulates of Euclidean geometry is that one can draw a straight line between any two points 1. Therefore, intuitively two (or more) measurements seem essential in order to investigate the change in a biomarker. This note illustrates how this simple intuition fails, leading to imprecise and biased results in common situations. We will consider the biomarker CD4+ T Lymphocytes count (CD4) in a trial or observational study of Human Immunodeficiency Virus (HIV) infection as an example. The primary objective of a trial could be to evaluate the change in CD4 after the initiation of a randomised intervention, or analysis of a change in the biomarker could be a retrospective exploratory analysis in both types of study. In either case, some participants may not have two marker measurements available, and this is likely to occur more frequently in the latter because this statistical analysis was not planned. The question is: should these participants be excluded or not? In other words, should one use a complete data set or the whole sample?

Today, most statistical packages include methods which work with different numbers of measurements for each participant, technically termed "unbalanced data" 2, 3. Such approaches allow all participants to be included, those with only one measurement as well as those fortunate enough to have more. One such method is a linear mixed or random effects model 2, 4. The general idea is to model the absolute levels of biomarkers from all participants - participants with only one value can still contribute population level information about the average biomarker levels at the time of their measurement. However, this kind of model does not just estimate overall levels in the study population – it assesses how different individual participants are from this population average. By making the reasonable assumption that individual level baselines (intercepts) and subsequent changes (slopes) in the participant population come from a statistical distribution (usually Normal), one can predict the slope of a patient with a single measurement only, for example, just at baseline.

How does it work intuitively?

When measurements from an individual are not available during follow-up, the method behaves as if that participant had contributed the follow-up values of patients whose baseline level and other measured characteristics are most similar.

How valid is the estimation of a slope when some patients only have one marker measurement?

In statistics, it is usual to distinguish three kinds of missing data 5–7: 1) 'missing completely at random' (MCAR), 2) 'missing at random' (MAR) or 3) 'informatively missing'. The first situation occurs when a marker value is missing at a given time independently of other values of the marker or explanatory variables (measured or not). For example, it happens if the test tube is broken by accident. A missing observation is 'missing at random' when the probability of not observing this value may depend on previously observed values. For instance, values could be missing after a participant reaches a threshold level (e.g. 200 cells/μL) that defines the end of the follow-up in a given study. Finally, a missing observation is 'informatively missing' when the probability that data are missing depends on some unobserved values. This situation could happen when a participant misses a visit because they are too sick (related to very low CD4). A linear mixed or random effects model as described above will provide unbiased estimates when missing measurements are due to either of the first two mechanisms. This is not the case for the third mechanism.

What about the intuitive alternative of excluding patients with only one measurement available?

Excluding some participants without a second measurement for particular reasons is analogous to a 'complete case analysis', although this is more often recognised in the context of missing explanatory variables. Such 'complete case analysis' is known to be valid **only** when the data are missing completely at random, meaning in our example **only** if the fact that a patient has no follow-up measurement is not linked to the baseline value or any explanatory variable. In this situation, excluding patients with only one measurement from analyses will produce estimates which are unbiased but with a poorer precision (higher variance, larger confidence intervals) compared to estimates from random effect models, that is, will give basically the same results but with reduced statistical power to detect genuine differences. However, a more likely situation is that data are 'missing at random', that is whether or not a biomarker value is observed depends on at least one of the prior values or other explanatory variables; then excluding participants with only one measurement leads to selection bias. For instance, this situation occurs when participants with lower baseline values are more likely lost to follow-up and also have the poorest treatment response (positive correlation between baseline value and slope value). In this case, MCAR assumption is violated; and excluding participants with only one baseline measurement overestimates the increase in marker values.

An example

We illustrate this with a simulation study based on the increase in CD4+ cell count after initiation of antiretroviral treatment in HIV infected patients 8. Lets take a study population of 100 patients with three measurements, i.e. at time 0 (treatment initiation), 1 and 2 years. Patients started a new treatment with an average of 200 cells/μL and have a mean increase of 100 cells/μL per year, strongly and positively associated with the baseline value (r=0.88): the CD4+ increase is more rapid in those with higher baseline CD4+. From this complete simulated dataset, we generated two incomplete datasets by simulating loss to follow-up of different groups of participants. In the first incomplete dataset, we deleted values at year 1 and 2 for a random selection of half of the participants. In the second, observations at year 1 and year 2 were deleted for all patients with CD4+ less than 200 cells/μL at baseline. With each incomplete dataset (i.e. deleting randomly or if baseline CD4<200), changes in CD4+ cell count were estimated by two different analyses using a random effects model including an intercept (for baseline) and a slope (for rate of change), and compared with analysis of the complete simulated dataset (N=100 patients with 300 measurements) which would not have been available had this been a real study. The first analysis included **all** the data that would have been available if this had been a real study (N=50 patients with 3 measurements each, 50 patients with a baseline measurement only, total 200 measurements) and the second excluded subjects with one measurement only (N=50 with 3 measurements each, total 150 measurements) as intuition based on Euclid would suggest.

Case 1: random loss to follow-up

As expected, when participants were randomly lost to follow-up, the three statistical analyses gave similar estimates of mean CD4 at baseline and during follow-up (Table 1). However, excluding those with only one measurement led to an increase of almost 40% in the standard error of the estimates compared to the analyses performed either on the complete dataset or including all patients. In the present example, this led to slightly wider confidence intervals (Table 1).

Case 2: loss to follow-up when baseline CD4 below 200 cells/μL

All data points and estimated trajectories are shown in Figure 1. Estimates from the complete dataset and from all available data also including participants with only one measurement were unbiased (Table 1), that is the baseline CD4 and rate of change in CD4 were both estimated correctly compared to the underlying model used to generate the data. In contrast, estimates restricted to participants with two or more measurements were biased upward: with a difference of +60 cells/μL for the baseline value and +31 cells/year for the rate of increase. This result is expected because we generated 'missing at random' (but not 'missing completely at random') data by deleting subsequent observations in those with baseline CD4 < 200 cells/μL); the complete case analysis that excludes those with only one available measurement excludes patients who are likely to have had poorer responses, leading to biased estimates. The random effects model provides unbiased estimates from the dataset including all patients even if half of them had only one measurement. The model could predict the value of CD4+ for the missing values during follow-up (in red in Figure 1) thanks to the information provided by the other patients and the available value at baseline.

Conclusion

Of course, in this artificial example, we knew that analysis excluding participants with only one measurement was likely to overestimate the increase in CD4+ because of the way we generated the data. Unfortunately, when faced with real-life data it is not generally possible to grasp the impact of excluding such participants. Here, we would like to endorse the use of the most efficient tools available for analysing clinical data. In particular, focussing on the analysis of change in biomarkers, we highlight that it would be far more effective to include all patients, even those with only one measurement, providing that a relevant statistical approach is used. Moreover, this result, whilst well known to statisticians and even used very widely in population pharmacokinetics studies 9, can be generalised to even more unbalanced data such as patients having very different numbers of available measurements 6. Also, other alternative approaches, such as multiple imputation, are also relevant in the context of the present note 10.

However, it should be recognised that all models make assumptions, whether or not these are immediately obvious to clinical readers. Excluding participants with only one measurement assumes that these participants are representative of the study population. Other simple methods such as the commonly used single imputation based on the last available measure (last observation carried forward, LOCF) make a major assumption of stability of marker values 11 and also underestimate variability by treating the single imputation as real observed data. The linear mixed or random effects methods described above make other less stringent assumptions that nevertheless might not be true. For instance, if the marker is informatively missing then other approaches will be needed 12. Because it is very difficult, if possible at all, to identify the type of missingness mechanism, it is generally recommended to check the robustness of the results through sensitivity analyses 6. However, the best way for handling missing information is trivially to have no missing observations in the first place 7. The idiom "Garbage in, garbage out" still holds.

Ackowledgements:

The authors declare no conflict of interest. RT was supported by a grant from the French National Agency for Research on AIDS and viral hepatitis (ANRS). We would like to thank Di Gibb, Nigel Klein and Daniel Commenges for their helpful comments on the manuscript.

References:

- 1. Billingsley HM The Elements of Geometrie of the Most Auncient Philosopher Euclide of Megara. London John Day; 1570-
- 2. Laird NM , Ware JH Random-effects models for longitudinal data. Biometrics. 1982; 38: 963- 74
- 3. Jennrich RI , Schluchter MD Unbalanced repeated-measures models with structured covariance matrices. Biometrics. 1986; 42: 805- 20
- 4. Laird NM Missing data in longitudinal studies. Stat Med. 1988; 7: 305- 15
- 5. Little R , Rubin D Statistical analysis with missing data. New York John Wiley & Sons; 1987;
- 6. Molenberghs G , Thijs H , Jansen I , Beunckens C , Kenward MG , Mallinckrodt C , Carroll RJ Analyzing incomplete longitudinal clinical trial data. Biostatistics. 2004; 5: 445- 64
- 7. Altman DG , Bland JM Missing data. Brit Med J. 2007; 334: 424-
- 8. Thiébaut R , Jacqmin-Gadda H , Walker S , Sabin C , Del Amo J , Porter K , Dabis F , Chêne G CASCADE Collaboration Determinants of response to first HAART regimen in naive patients with an estimated time since HIV seroconversion. HIV Med. 2006; 7: 1- 9
- 9. Whiting B , Kelman AW , Grevel J Population pharmacokinetics. Theory and clinical application. Clinical Pharmacokinetics. 1986; 11: 387- 401
- 10. Kenward MG , Carpenter J Multiple imputation: current perspectives. Stat Methods Med Res. 2007; 16: 199- 218
- 11. Cozzi-Lepri A , Smith GD , Mocroft A , Sabin CA , Morris RW , Phillips AN A practical approach to adjusting for attrition bias in HIV clinical trials with serial marker responses. AIDS. 1998; 12: 1155- 61
- 12. Touloumi G , Pocock SJ , Babiker AG , Darbyshire JH Impact of missing data due to selective dropouts in cohort studies and clinical trials. Epidemiology. 2002; 13: 347- 55

Figure 1

Data and estimation of mean CD4 cells/µL according to data used: complete dataset (grey), excluding patients without follow-up measurements due to low baseline CD4 cell count (black), including all patients (red). Note: estimates from the complete dataset (in grey) are very similar to with estimates from model fitted with patients having one measurement available or more (in red) (see Table 1).

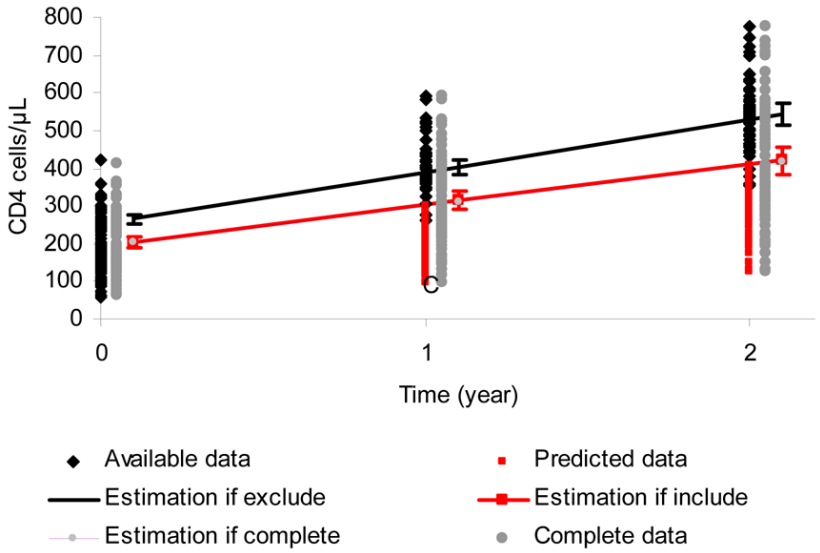

Table 1

Estimated baseline CD4+ and slope per year (mean and 95% confidence intervals) according to data used in analysis and missing data mechanism.

Methods/Time	Estimate of Baseline CD4 (cells/μL)	Estimate of rate of change (slope) (cells/μL/year)
Complete data (100 pts, 300 measures)	**205 (191;219)**	**+105 (96;114)**
Case 1: lost to follow-up completely at random		
(i) Including all patients (100 pts, 200 measures)	204 (189;220)	+110 (99;121)
(ii) Excluding those with only one measurement (50 pts, 150 measures)	204 (183;224)	+110 (97;123)
Case 2: lost to follow-up if baseline CD4 <200 cells/μL		
(i) Including all patients (100 pts, 200 measures)	205 (191;220)	+107 (97;117)
(ii) Excluding those with only one measurement (50 pts, 150 measures)	265 (253;278)	+138 (128;148)