Bone Marrow Transplantation (2020) 55:1–3
https://doi.org/10.1038/s41409-019-0431-6

EDITORIAL

Statistical analyses of clinical trials in haematopoietic cell transplantation or why there is a strong correlation between people drowning after falling out of a fishing boat and marriage rate in Kentucky

Robert Peter Gale[1] · Mei-Jie Zhang[2]

Received: 6 November 2018 / Revised: 13 December 2018 / Accepted: 20 December 2018 / Published online: 26 March 2019

We are pleased to introduce a 12-part series on topics in statistics and epidemiology which we hope will be interesting and useful to readers of *BONE MARROW TRANSPLANTATION*. Our objective is to raise awareness of on some of the complexities in design and interpretation of data from clinical trials. We hope that these typescripts will generate a lively discussion. More on this below.

Analyses of data from transplant studies are especially complex because of several issues including: (1) *time-to-treatment* bias. For example, subjects who are potential transplant recipients need to live long enough to receive a transplant. Moreover, this bias is not random with the worst subjects dying before a transplant can be done. This bias also confounds donor type. Finding a HLA-matched unrelated donor transplant takes longer than finding a HLA-haplotype-matched donor, such that subjects awaiting a transplant from an HLA-matched unrelated donor may fall by the wayside; (2) subject selection bias. Subjects with comorbidities or frailties maybe excluded even if they survive sufficiently long to potentially receive a transplant. Transplant studies often fail to state: These conclusions apply only to subjects receiving a transplant and should not be assumed to apply to all persons with the disease and disease state being studied. A good example is results of transplants in persons >65 years. This is a highly selected cohort for which the true denominator of potential recipients is unknown. A good outcome in persons >65 years receiving a transplant cannot be applied to most persons >65 years; (3) competing causes of therapy– failure, e.g. transplant-related mortality (TRM), graft-failure and graft-versus-host disease (G*v*HD). If you die of G*v*HD, you are not at-risk to relapse. This is another confounded relationship because persons with G*v*HD are less likely to relapse than those without G*v*HD. (4) Imprecision with false-negative and e-positive results such as who really has G*v*HD *versus* a drug- or virus infection-related rash (or both). And more such as the universal problems of measurement error and the role of chance in clinical outcomes Fig. 1, issues no clinical scientist likes to consider.

Another example is an over–emphasis on *point-estimate*s rather than confidence intervals. Some people think that the *point-estimate*, say median survival of 2 years, is the *true* value. This is almost never so. What we can say is that the *true* value has a 95% probability of being within the confidence intervals (credibility limits if we are using a Bayesian approach). Sometimes it is higher, sometimes lower. As such, you should look at the confidence interval, not the point-estimate when comparing outcomes and effect sizes. Also, there is widespread misunderstanding of what a *P*-value is and what it means [1]. We could list >20 mis–interpretations of the *P*-value, such as whether the results obtained occurred by chance. Wrong. A *P*-value is the likelihoodthat the statistical model being tested (usually the *null* hypothesis) is a good explanation of the results (or more extreme results). A *P*-value of 0.06 does not mean that the results are not significant. Rather, it means a statistical model other than the *null* hypothesis better explains the data (or more extreme data). Other important issues are to analyze non-proportional hazards common in transplant studies, proper use of sensitivity analyses, whether to use fixed- or random-effect models, differences between association and correlation and between these and *cause-and-effect* and others. Few of us appreciate the magnitude of undetermined (latent) variables in predicting outcomes. In

✉ Robert Peter Gale
robertpetergale@gmail.com

[1] Imperial College London, London, UK

[2] Medical College of Wisconsin and CIBMTR, Milwaukee, WI, USA

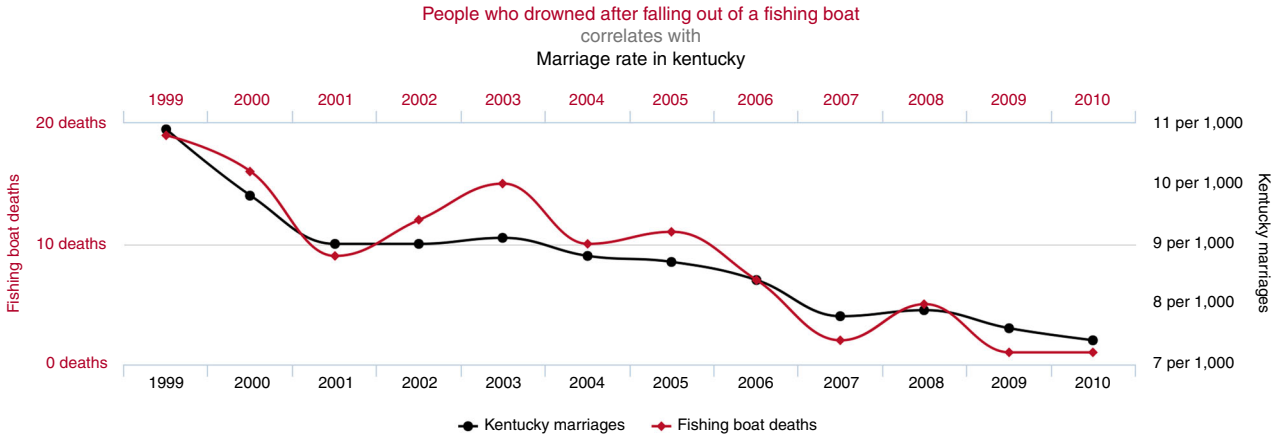

Fig. 1 People who drowned after falling out of a fishing boat correlates with marriage rate in Kentucky

observational studies, such as those from the CIBMTR and EBMT, we can adjust for known but not latent variables, which account for about one-half of the observed variance. These issues will be covered in typescripts to the following in this series.

Clinical trials in our field are rarely randomized and many lack appropriate controls. Even when there is a randomized trial, blinding is rarely possible, forcing us to consider observer biases. Our randomized trials are lucky to have 300 subjects; most trials in cardio–vascular disease, diabetes and breast cancer have 10,000 subjects or more. This means that we must often rely on answers from observational databases. This is not necessarily bad, but we discussed the advantages and limitations of this approach elsewhere [2]. Other approaches are propensity score analyses and Monte Carlo simulations. But what exactly are these and what are their advantages and limitations? Often, therapy recommendations are based on the so-called expert consensus panels, the limitations of which are well-known [3].

Take the seemingly simple setting where we do a randomized trial for one variable, say two different therapies, and get different outcomes between the arms. Intuitively, one might assume that the intervention was responsible for the different outcomes, namely, that the intervention variable is *causal*. However, this assumption is sometimes wrong. Causality is a tricky business in statistics and epidemiology and the subject of the first typescript in this series appearing in this issue: *Causal Inference in Randomized Clinical Trials*. Later in the series, we will discuss data from experimental psychology exploring how the human mind works (or does not work), which results in thin slicing, fuzzy logic and incorrect conclusions.

Several sources estimate that about ½ of published peer-reviewed studies are wrong [4]. This is also estimated to be so for about ½ of physician practices considered *standard-of-care* (Which is better: hot or cold compresses for a sprained ankle? One must be wrong but both are widely recommended) [5]. Setting the statistical significance barrier at $P < 0.05$ guarantees many incorrect conclusions. But what to do? Lower the *P*-value to 0.005 [6] and get lots of under–powered studies with potential false-negatives OR keep the *P*-value at 0.05 and require confirmation [7]. We also cover this question later in the series.

We hope that the readers will enjoy these typescripts and find them useful, maybe even fun. We welcome suggestions for topics which can be e-mailed to us or Profs. Lazarus and Mohty. We also hope to start a lively discussion on these topics and answer questions. We can be reached on Twitter #BMTStats. Our operators are standing by. ☺

Acknowledgements R.P.G. acknowledges support from the National Institute of Health Research (NIHR) Biomedical Research Centre funding scheme.

Compliance with ethical standards

Conflict of interest R.P.G. is a part-time employee of Celgene Corporation. M.-J.Z. declares no conflict of interest.

References

1. Gale RP, Zhang M-J. What is the *P*-value anyway? Bone Marrow Transplant. 2016;51:1439–40.

2. Gale RP, Eapen M, Logan B, Zhang MJ, Lazarus H. Comparing therapy-options: are observational database studies and expert opinion as good (or better) than randomized trials? Bone Marrow Transplant. 2009;43:435–46.
3. Gale RP, Eapen M, Logan B, Zhang MJ, Lazarus HM. Are there roles for observational database studies and structured quantification of expert opinion to answer therapy controversies in transplants? Bone Marrow Transplant. 2009;43:743.
4. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2:e124
5. Cifu A, Prasad V. Ending Medical Reversal: Improving Outcomes, Saving Lives. Baltimore: Johns Hopkins University Press; 2015.
6. Ioannidis JPA. The proposal to lower *P* value thresholds to .005. JAMA. 2018;319:1429–30.
7. Barach P, Thomas RL, Lipshultz SE. Lowering the *P* value threshold. JAMA. 2018;320:936.