Contemporary Clinical Trials 87 (2019) 105875

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Short Communication

Biostatistics pitfalls: Lessons learned from analysis of medical data

Guosheng Yin[a,b,*], Chenyang Zhang[a], Zhao Yang[a]

[a] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong
[b] Department of Biostatistics, MD Anderson Cancer Center, Houston, TX, USA

ARTICLE INFO

Keywords:
Data analysis
Model assumption
RMST

ABSTRACT

In several recent issues of *The Lancet*, we identified a few common pitfalls in the analysis of clinical trial and medical data in the published articles (Mok et al., 2019; Herrlinger et al., 2019; Reindl-Schwaighofer et al., 2019; He et al., 2019). Without careful validation of model assumptions, even the primary endpoint of the trial might be analyzed using improper statistical methods. We carried out an in-depth analysis of the statistical issues in four real clinical trials, which highlights the importance of statistics in the medical field. With every effort, biostatisticians need to work with clinicians closely to take the most appropriate statistical approaches to data analysis; otherwise the conclusions drawn from the data might be problematic or misleading.

1. Introduction

In several recent issues of *The Lancet*, we identified a few common pitfalls in the analysis of medical data in the published articles [1–4]. The inappropriate use of statistical methods may damage the soundness of the analysis, and thus the conclusions drawn from the statistical findings may be misleading or sometimes even false due to the fact that either the model assumptions are violated or the statistical methods are not applied in a proper way. These biostatistics pitfalls suggest that clinicians should work closely with well-trained biostatisticians to carry out sound statistical analysis in order to draw correct conclusions from these studies. It also highlights the crucial value of biostatistics in medicine and further corroborates the importance of collaborations between the two fields. The four real examples below demonstrate the necessity for such close collaborations from different perspectives. For example, in the clinical trial design stage, how to make the design more efficient as well as properly controlling the type I error to reduce the possibility of false positive; stratified analysis should be used cautiously, otherwise it may cause adverse effects if the data are overly stratified; the Cox proportional hazards model might not provide a suitable fit to the data when the proportional hazards assumption is violated and thus the reported hazard ratio would be meaningless.

Example 1. Comparison of pembrolizumab and chemotherapy using the restricted mean survival time.

Mok et al. [1] conducted an important trial (KEYNOTE-042) comparing pembrolizumab and chemotherapy for previously untreated, PD-L1-expressing non-small-cell lung cancer (NSCLC). The reported findings were based on the second interim analysis originally planned for the trial. The overall survival (OS) of the three subpopulations based on the PD-L1 tumor proportion scores (TPS $\geq$ 50%, $\geq$20%, $\geq$1%) were assessed under significance thresholds of 0.0122, 0.0120, 0.0124 sequentially to control the overall one-sided type I error rate at $\alpha = 0.025$. However, the first interim analysis had already spent the type I error rate of $\alpha = 0.01576$, and thus the remaining type I error rate would be 0.00924 which is the total type I error rate for the second interim analysis and the final analysis together. It is not clear how the three significance thresholds (0.0122, 0.0120, 0.0124) were derived. Moreover, Fig. 2 in Mok et al. [1] shows that the Kaplan-Meier OS curves cross somewhere between 6 months to 1 year for all PD-L1 TPS subpopulations, indicating potential violation of the proportional hazards assumption (e.g., $p = .006$ for the TPS $\geq$ 50% subpopulation using a residual-based test [5]). When the proportional hazards assumption is violated, the hazard ratio (HR) from the Cox proportional hazards model is not a meaningful quantity which cannot be interpreted as an average HR over time. As an alternative, we suggest a more robust (assumption-free) approach to quantifying treatment effect by using the restricted mean survival time (RMST). The RMST is defined as the area under the Kaplan-Meier curve up to the specific follow-up time [6–8]. Based on the OS data reconstructed from Fig. 2A in Mok et al. [1], the 36-month RMST using pembrolizumab was 20.3 months and that of chemotherapy was 16.6 months, i.e., NSCLC patients under pembrolizumab on average gained 3.7 months more lifetime during a 3-year period of follow-up (95% CI [1.4, 5.9]; $p = .001$) compared with those under chemotherapy.

* Corresponding author at: Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong.
E-mail address: gyin@hku.hk (G. Yin).

https://doi.org/10.1016/j.cct.2019.105875
Received 12 September 2019; Received in revised form 18 October 2019; Accepted 24 October 2019
Available online 30 October 2019

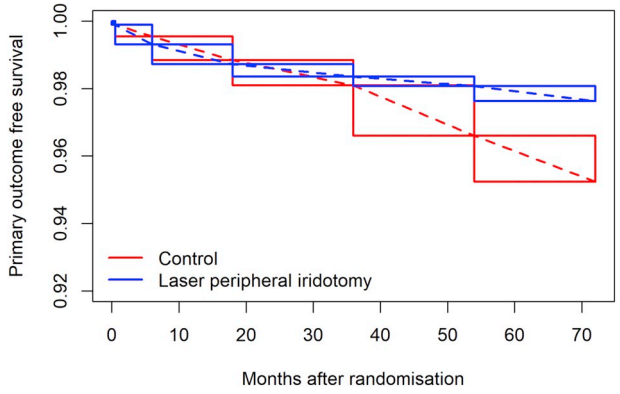

Fig. 1. Primary outcome free survival curves (rectangles) using interval-censored data with dotted lines connecting all diagonal lines of the rectangles.

Example 2. Stratified log-rank test for highly stratified data.

Herrlinger et al. [2] reported that lomustine-temozolomide combination therapy might improve survival compared with temozolomide standard therapy (p = .0492 from a stratified log-rank test). The modified intention-to-treat dataset with a total of 109 patients was used to perform the stratified log-rank test with up to 33 strata grouped by eleven center-based categories and three classes of recursive partitioning analysis. However, it is well-known that the stratified log-rank test may suffer from lack of efficiency [9] and power loss [10] when the stratum size is small (3 to 4 patients per stratum in this study), i.e., over-stratification. The test statistic would remain unchanged after the time point when the stratum only involves patients from one treatment arm, such that the remaining patients in that arm would make no contribution to the test statistic [9]. Moreover, a large number of small-size strata may further deteriorate power due to extreme imbalance in the number of patients at risk within each stratum [10]. The authors also conducted supplementary analyses under the unstratified and small strata cases. Fig. 1a in the Appendix of Herrlinger et al. [2] shows the Kaplan-Meier curve for overall survival and reports no statistically significant difference by the unstratified log-rank test (p = .6579) and a univariate Cox model (HR = 0.90, 95% CI [0.58, 1.41]); Fig. 1b displays the Kaplan-Meier curve under the small strata case using inverse probability weights with an HR of 0.90 (95% CI [0.47, 1.17]).

We reanalyzed the reconstructed data (from Fig. 3A in Herrlinger et al. [2]) using the number needed to treat (NNT) and RMST at the five-year follow-up. The estimated absolute risk reduction at five years was 3.5% (95% CI, [−17.6%, 24.6%]), corresponding to an estimated NNT of 29 (95% CI, NNTB 4 to ∞ to NNTH 6), where NNTB (NNTH) represents NNT to benefit (harm) one patient. Furthermore, there was no statistically significant difference in overall survival between the two groups using the RMST-based test (p = .161) and the unstratified log-rank test (p = .200), which are consistent with the insignificant survival difference in their supplementary analyses.

Example 3. Dummy coding for ordinal categorical variables in the Cox proportional hazards model.

Reindl-Schwaighofer et al. [3] reported that genetic mismatch of non-synonymous single nucleotide polymorphisms (nsSNPs) was associated with an increased risk of graft loss with an HR of 1.68 (95% CI [1.17, 2.41], p = .005) using the Cox proportional hazards model. The nsSNP mismatch was discretized by quartiles into four groups S1–S4, which was treated as a single ordinal variable in the reported analysis. Using such an ordinal-coding variable, the Cox model assumes a same HR across all adjacent interquartile range (IQR) groups. However, the Kaplan-Meier curves of four groups crossed as shown in Fig. 3 of Reindl-Schwaighofer et al. [3] indicating violation of the PH assumption. Therefore, the interpretation that each increase by a unit of one IQR had an HR of 1.68 is misleading. A more appropriate approach is to create three dummy-coding or indicator variables by taking one group as reference. As in Ovadia et al. [11], two HRs for three bile acid concentration groups were computed by taking one group as reference. We reanalyzed the survival data reconstructed from Fig. 3 in Reindl-Schwaighofer et al. [3] using the log-rank test, Cox model, and RMST, and the results are summarized in Table 1. A four-sample log-rank test indicated a significant overall survival difference among the four groups S1–S4 with p = .003. By taking S1 as reference, step-down tests identified significant differences between S1 and each of the other three groups. The HRs between adjacent quartile groups varied from 1.14 to 3.37 which violated the same HR assumption for the ordinal-coding variable. Dummy coding is recommended for ordinal variables unless the proportional hazards assumption is satisfied across all groups.

Example 4. Re-analysis of the data of laser peripheral iridotomy for the prevention of angle closure.

He et al. [4] reported that laser peripheral iridotomy had a significant prophylactic effect using the Cox proportional hazards model with a hazard ratio (HR) of 0.53 (95% CI [0.30, 0.92]; p = .024) and McNemar's test (p = .0041). For each patient, one eye was randomly selected for treatment and the other was left untreated. Both eyes were examined at six pre-specified time points, leading to a pair-matched, interval-censored study. The trial was extended from 36 months to 72 months with additional 155 participants due to a low event rate and, as a result, the significance threshold was adjusted to 0.025. Treating interval-censored data as right-censored may introduce bias and underestimate the variance, which might result in false positive findings. We reanalyzed the interval-censored data reconstructed from Fig. 3 of He et al. [4] using the interval-censored log-rank test [12], RMST and the Cox model [13]. Using interval-censored data, the log-rank test (p = .021) still supported a significant difference, while the Cox model yielded an HR of 0.52 (95% CI [0.29, 0.95]; p = .031), leading to an

Table 1
Re-analysis of functional renal allograft survival data[a] using the log-rank test, Cox proportional hazards model and restricted mean survival time (RMST)[b] by taking S1 as the reference group.

Comparison		Log-rank test	Hazard ratio	Difference in RMST
S2 vs S1	Estimate 95% CI		3.37 [1.09, 10.45]	−0.611 [−1.146, −0.075]
	p-Value	0.026	0·036	0.025
S3 vs S1	Estimate 95% CI		5.15 [1.73, 15.33]	−1.050 [−1.672, −0.428]
	p-Value	0.001	0.003	0.001
S4 vs S1	Estimate 95% CI		5.7 [1.94, 16.80]	−1.180 [−1.822, −0.537]
	p-Value	0.0003	0.002	0.0003

S1–S4 denote four categories grouped by the quartiles of nsSNP mismatch coding for transmembrane and secreted proteins.

[a] Survival data were reconstructed from Fig. 3 in the original paper [3] using the "digitize" package in R software, version 3.5.1 (R Project for Statistical Computing).

[b] The restricted mean survival time was estimated by calculating the area under the Kaplan-Meier curve at the ten-year follow-up using the "survRM2" package.

insignificant result at the significance threshold 0.025. The survival curves for interval-censored data display rectangles as show in Fig. 1 and the RMST can be calculated as the area under the curve by connecting the diagonal lines of all rectangles. Both differences in 72-month RMSTs using right-censored data (0.263, 95% CI [−0.434, 0.960]; $p = .459$) and interval-censored data (0.489, 95% CI [−0.308, 1.287]; $p = .229$) indicate no treatment benefit. To account for inter-eye correlations, the multivariate Cox model for pair-matched data should have been used [14].

2. Discussion

Overall, from the above four recently published studies in *The Lancet*, we have identified inappropriate statistical analysis of the data, while all the misuses of biostatistics can be avoided under close collaborations between clinicians and well-trained biostatisticians. Every clinical trial is equipped with a protocol. The protocol of a clinical trial provides a comprehensive description of the entire study, which typically includes a statistical analysis plan (SAP) that details the trial design (e.g., sample size and power, type I error rate, group sequential methods and alpha-spending function, Bayesian decision rules), statistical analysis (e.g., hypothesis tests, models and assumptions, treatment of missing data), analysis of population (e.g., intent-to-treat or per protocol population, subgroup analysis). Despite the detailed specifications of statistical methods in the SAP, it may happen that the trial design is implemented without much care, unexpected events occur during the course of the trial, and inappropriate statistical analyses are carried out for the collected data. Simply using convenient statistical software without careful validation of model assumptions and in-depth analysis of the statistical issues may not guarantee the soundness of the data analysis, which may result in false conclusions.

Declaration of Competing Interest

The authors declare no potential conflicts of interest

The authors have no funding to disclose

Acknowledgments

The research was supported by grant 17307318 from the Research Grants Council of Hong Kong.

References

[1] T.S. Mok, Y.L. Wu, I. Kudaba, et al., Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KRYNOTE-042): a randomized, open-label, controlled, phase 3 trial, Lancet 393 (2019) 1819–1830, https://doi.org/10.1016/S0140-6736(18)32409-7.

[2] U. Herrlinger, T. Tzaridis, F. Mack, et al., Lomustine-temozolomide combination therapy versus standard temozolomide therapy in patients with newly diagnosed glioblastoma with methylated MGMT promoter (CeTeG/NOA–09): a randomised, open-label, phase 3 trial, Lancet 393 (2019) 678–688, https://doi.org/10.1016/S0140-6736(18)31791-4.

[3] R. Reindl-Schwaighofer, A. Heinzel, A. Kainz, et al., Contribution of non-HLA incompatibility between donor and recipient to kidney allograft survival: genome-wide analysis in a prospective cohort, Lancet 393 (2019) 910–917, https://doi.org/10.1016/S0140-6736(18)32473-5.

[4] M. He, Y. Jiang, S. Huang, et al., Laser peripheral iridotomy for the prevention of angle closure: a single-centre, randomised controlled trial, Lancet 393 (2019) 1609–1618, https://doi.org/10.1016/S0140-6736(18)32607-2.

[5] P.M. Grambsch, T.M. Therneau, Proportional hazards tests and diagnostics based on weighted residuals, Biometrika 81 (1994) 515–526, https://doi.org/10.1093/biomet/81.3.515.

[6] I.R. Weir, G.D. Marshall, J.I. Schneider, et al., Interpretation of time-to-event outcomes in randomized trials: an online randomized experiment, Ann. Oncol. 30 (2019) 96–102, https://doi.org/10.1093/annonc/mdy462.

[7] G. Yin, Clinical Trial Design: Bayesian and Frequentist Adaptive Methods, 151–5 John Wiley & Sons, 2012, pp. 246–248.

[8] H. Uno, B. Claggett, L. Tian, et al., Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis, J. Clin. Oncol. 32 (2014) 2380–2385, https://doi.org/10.1200/JCO.2014.55.2208.

[9] D.A. Schoenfeld, A.A. Tsiatis, A modified log rank test for highly stratified data, Biometrika 74 (1987) 167–175, https://doi.org/10.1093/biomet/74.1.167.

[10] K. Akazawa, T. Nakamura, Y. Palesch, Power of logrank test and cox regression model in clinical trials with heterogeneous samples, Stat. Med. 16 (1997) 583–597, https://doi.org/10.1002/(SICI)1097-0258(19970315)16:5<583::AID-SIM433>3.0.CO;2-Z.

[11] C. Ovadia, P.T. Seed, A. Sklavounos, et al., Association of adverse perinatal outcomes of intrahepatic cholestasis of pregnancy with biochemical markers: results of aggregate and individual patient data meta-analyses, Lancet 393 (2019) 899–909, https://doi.org/10.1016/S0140-6736(18)31877-4.

[12] J. Sun, A non-parametric test for interval-censored failure time data with applications to AIDS studies, Stat. Med. 15 (1996) 1387–1395, https://doi.org/10.1016/S0140-6736(18)32607-2.

[13] C. Anderson-Bergman, Icenreg: regression models for interval censored data in R, J. Stat. Softw. 81 (2017) 1–23, https://doi.org/10.18637/jss.v081.i12.

[14] E.W. Lee, L.J. Wei, D.A. Amato, Cox-type regression analysis for large numbers of small groups of correlated failure time observations, in: J.P. Klein, P.K. Goel (Eds.), Survival Analysis: State of the Art, Springer Netherlands, Dordrecht, 1992, pp. 237–247.