Title

Limitation of site-stratified cox regression analysis in survival data: a cautionary tale of the PANAMO phase III randomized, controlled study in critically ill COVID-19 patients.

Journal

Current Controlled Trials in Cardiovascular Medicine, 25(1)

Authors

Sandrock, Christian

Song, Peter

Publication Date

2024-12-18

DOI

10.1186/s13063-024-08679-5

Peer reviewed

Trials

COMMENTARY

Open Access

Limitation of site-stratified cox regression analysis in survival data: a cautionary tale of the PANAMO phase III randomized, controlled study in critically ill COVID-19 patients

Christian E. Sandrock[1] and Peter X. K. Song[2*]

Abstract

Current guidelines tend to focus on a *p*-value threshold of a pre-specified primary endpoint tested in randomized controlled clinical trials to determine a treatment effect for a specific drug. However, a *p*-value does not always provide evidence on the treatment effect of a drug, especially when stratification of the data does not account for unforeseen variables introduced into the analysis. We report and discuss a rare case in which investigational site stratification in the pre-specified analysis method of a primary endpoint results in a loss of statistical power in the evaluation of the treatment effect due to data attrition of almost 17% of outcome data in the phase III randomized, controlled PANAMO study in critically ill COVID-19 patients. Other analyses utilizing no or different stratification (e.g., stratifying by country, region, pooling low enrollment clinical sites) evaluates 100% of patient data resulting in *p*-values suggesting a positive treatment effect ($p < 0.05$). We demonstrate how this technical artifact occurs by adjustment for site stratification within the Cox regression analysis for survival outcomes and how alternative stratification corrects this discrepancy.

Keywords Cox proportional hazards regression, Survival analysis, Vilobelimab, PANAMO, COVID-19

Introduction

Physicians and scientists often use the *p*-value to report statistical significance in the pre-specified analysis method for the defined primary and, if applicable, secondary outcome(s) when evaluating data of randomized controlled clinical trials. This approach is reflected in the development and adoption of international harmonized guidelines for Structure and Content of Clinical Study Reports (ICH E3) to regulatory authorities, for Good Clinical Practice (ICH E6) in the conduct, monitoring, and reporting of clinical trials, and Statistical Principles for Clinical Trials (ICH E9) which outlines the principles of statistical methodology applied to clinical trials for marketing applications submitted to regulatory authorities [1–3]. Rigorous data handling, analyses, and reporting ensures the validity of a statistical hypothesis test used in well-performed and controlled experiments such as a high-quality clinical trial. Reviewers of manuscripts are regularly confronted with claims by investigators and sponsors that post-hoc analyses of clinical trial subgroups using non-prespecified endpoints or when applying a different, non-prespecified test suggest treatment effects. To control for type I error, which means incorrectly rejecting the null hypothesis (i.e., claiming the drug has an effect

*Correspondence:
Peter X. K. Song
pxsong@umich.edu
[1] Division of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, UC Davis Medical Center, University of California, Davis, 2315 Stockton Blvd, Sacramento, CA 95817, USA
[2] School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI, USA

when it does not), researchers typically aim to control the overall type I error rate at level of 5%, which may present a technical challenge in some clinical studies that involve multi-stage decision making [4]. This is also referred to as a significance level (alpha) of 0.05 in a statistical test. Testing a hypothesis multiple times is problematic with no multiplicity adjustment as it suffers from an increased probability of type I error [5, 6].

Statistical methods can have limitations when data are missing from the analysis. Therefore, expert statisticians and trained reviewers of regulatory authorities will usually not rely on one *p*-value or one test alone to judge the quality and the robustness of a clinical data set. An important principle for the evaluation of randomized controlled clinical trial data is the "intention to treat" (ITT) principle, meaning that the data of all enrolled, randomized patients are assessed for a primary outcome analysis [7]. Analyzing all available data is important because any exclusion of data, be it arbitrary or be it intentional, can lead to a significantly biased assessment of the effectiveness of a drug or intervention. This ITT principle is also reflected in regulatory guidance documents such as FDA or ICH guidelines for the analysis methods applied for primary outcome analysis in clinical trials [8].

Background for the PANAMO study statistical analysis

The recently published PANAMO phase III study was a 1:1 randomized placebo-controlled, global clinical trial in adult patients with severe COVID-19 [9]. The study which randomized 369 patients was designed for a 90% power to demonstrate an improvement in 28-day all-cause mortality (the primary endpoint) for a new monoclonal antibody treatment binding and blocking the complement factor C5a (vilobelimab, tradename Gohibic™). C5a is a well-researched component of host immune response which can lead to tissue damage after infections [10]. The sponsor of the study suggested in its statistical analysis plan (SAP) protocol, under which the trial was conducted, to analyze this primary endpoint using an age-adjusted Cox regression without site stratification to assess the risk of all-cause mortality. The PANAMO trial was conducted between October 2020 with the last patient randomized in October 2021. The US Food and Drug Administration (FDA) then recommended to incorporate investigational site stratification as an adjustment in this Cox regression to account for potential cross-site heterogeneity in the latter stage of full enrollment for the study [11]. This approach was then adopted by the sponsor in the SAP shortly before the database was locked. Thus, the site-stratified Cox regression became the pre-specified method for the analysis of the clinical trial data, and the originally proposed analysis using a non-site stratified approach became a post-hoc analysis. It is important to note that changes to the SAP are frequently seen and accepted prior to the unblinding of many studies.

When the data from PANAMO was analyzed using site stratification within Cox regression, however, approximately 17% of all enrolled patients were excluded from contributing to the analyzed output. Consequently, the resulting *p*-value (0.094) reflected this exclusion due to reduced power of the study. However, the FDA within its published review acknowledged this technical limitation and adopted the originally proposed method (Cox regression without site stratification adjustment) as the "more reliable" method ($p=0.026$) and stated that "there were no data integrity issues" ultimately concluding on a significant mortality benefit being demonstrated [11]. This conclusion was the basis for an emergency use authorization of Gohibic in the US for the treatment of COVID-19 in hospitalized adults when initiated within 48 h of receiving IMV or ECMO [12]. The originally published article in Lancet Respiratory Medicine [9] also came to a similar conclusion within its abstract, "In addition to standard of care, vilobelimab improves survival of invasive mechanically ventilated patients with COVID-19 and leads to significant decrease in mortality," despite the reported negative *p*-value from the pre-specified analysis method, recognizing its limitation.

What happened statistically?

Within Cox regression analysis, adjustments can be made for confounders which are known or expected to have an impact on the outcome parameter survival [13]. Importantly, in the PANAMO phase III study, all Cox regression analyses were adjusted for age, as age has been demonstrated to impact survival in COVID-19. Adjusting for site stratification within Cox regression could be justified if site-specific heterogeneity was assumed to be a confounder for the outcome mortality. However, when adjusting for site stratification in Cox regression analysis, the Cox partial likelihood estimation method (i.e., the mathematical procedure) requires calculation of site-specific risk sets separately for each site to reflect heterogeneous baseline hazards across sites. Technically, for a site that has no events (e.g., no deaths) or a site that contains only one enrolled patient regardless of survival status, the corresponding risk set has no variability and thus does not contribute to the formation of the partial likelihood. Consequently, this excludes data from all such sites in the data analysis and creates data attrition, something not anticipated by the FDA's request.

In the PANAMO phase III study, this was the case for 61 patients (16.6% of the total enrollment): 55 patients

from sites with no events (i.e., deaths) plus 6 patients from single patient sites who died. By chance, these 6 patients from six singletons were all placebo deaths and not a single death from the vilobelimab group; thus, excluding them from the data analysis caused underestimation of the treatment effect. Because a factual exclusion of all patient outcome data from these sites is involved in the analysis, the resulting *p*-value was compromised. Therefore, this hidden bias due to a reduced effective sample size and unbalanced treatment allocation tipped the *p*-value above the significance level.

Empirically, we can verify the above insight by removing these 61 patients in a fit of the remaining data to the site-stratified Cox regression, which generated the identical output (*p*-value, hazard ratio and confidence intervals). When analyzing the data set with the originally proposed protocol method using Cox regression without site stratification, the analysis reported a positive finding with a hazard ratio (HR) of 0.67 and a *p*-value of 0.026 (Fig. 1), which was adopted by the FDA in its published review as the more reliable method.

In order to reflect the original motivation of site-stratified analysis to account for geographic diversity and population heterogeneity while addressing the technical challenge caused by local risk sets (e.g., confounding of race and health disparities), a country-level or region-level stratification may be deemed more appropriate. Also, one might argue that the healthcare system (country) may have more impact on mortality as it crucially impacts intensive care treatment modalities (i.e., which drugs are approved and paid for within the healthcare system) as well as unit staffing with qualified personnel and other factors. Fitting the country-stratified or the region-stratified Cox model, as well as the multilevel frailty Cox model with random effects to account region-specific heterogeneity, the resulting *p*-values for the treatment effect all suggested positive findings with the estimated hazard ratios varying in similar ranges (Fig. 1). The same phenomenon was repeated for the pre-specified sensitivity analysis using logistic regression as well as for a post-hoc simple group comparison via log rank test. When these same analyses were applied, the key secondary endpoint, 60-day all-cause mortality, comparable patterns of HRs, confidence intervals, and *p*-values were observed.

Conclusions

The PANAMO phase III analysis is a rare case in which the stratification in the pre-specified analysis method introduced an unintended bias such that the *p*-value reflected inaccurate or incomplete information about the entire available data set, the ITT population. The introduced unintended bias could be adequately explained and proven by analyzing the data set by Cox regression without site stratification or use of proper stratification by geographical region or small site pooling. Such additional evaluation should therefore not be dismissed as an "invalid" post-hoc analysis when the stratification in the pre-specified analysis leads to the described data attrition and sampling bias. If study population stratification is planned in a trial, the power analysis should take multiple subgroup analyses into account in the SAP a

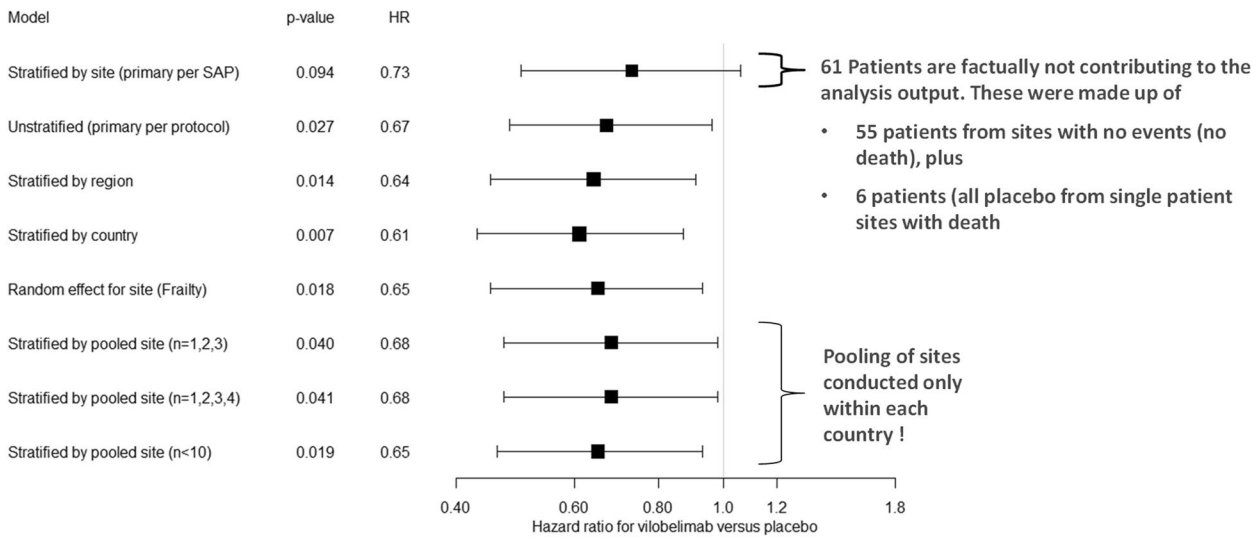

Fig. 1 Cox regression analyses performed on the phase III PANAMO study population. *p*-values, hazard ratios (HR), and confidence intervals for various age-adjusted and stratified Cox regression analyses (Model) within the PANAMO phase III primary outcome data for 28-day all-cause mortality

priori. It is also not a problem of alpha-inflation when applying an adequately stratified analysis to avoid data attrition that impacts randomization and exerts sampling bias. Journals, reviewers, and guideline committees have no standard for assessing data for such a case and reporting guidelines mandate, for good reasons, only to use the pre-specified analysis method. Unfortunately, falling short of the *p*-value cutoff of 0.05 for statistical significance impairs the validation of a study to report a positive finding of treatment benefits (i.e., meeting the primary endpoint) in the eyes of guideline committees, hospital formulary committees, health systems, and healthcare providers. Regulatory guidelines, on the other hand, mandate the application of the ITT principle for outcome analyses. It seems reasonable that experts charged with assessing and/or recommending on such a case should be encouraged to ask for or have the sponsors construct a more in-depth data evaluation aiming at judging the entire data set. Otherwise, the choice of the statistical method would overrule the actual information contained in the data set which contradicts the purpose of conducting well-controlled clinical trials. It is important to consider the described shortcoming of adjusting for site stratification within Cox regression when it leads to hidden data attrition and unbalanced treatment allocation. This may influence the analysis resulting in clinical conclusions based on a statistical decision of rejecting or accepting the null hypothesis of no treatment effect under a prefixed type I error, say, 0.05.

In conclusion, by considering the entirety of the ITT data set, vilobelimab demonstrated a positive survival benefit for all applied methods which included the 369 randomized patients, the ITT population in PANAMO [9], thus supporting the FDA's issuing of an EUA [11, 12]. Adjusting for site stratification within Cox regression can introduce a hidden data exclusion from smaller sites or sites with no events and a potential bias in treatment allocation, which should be carefully examined before being chosen as a pre-specified analysis method. Investigators in all controlled clinical trials should be aware of this limitation when considering applying site-stratified Cox regression analysis, and in such situation, the use of the random-effects modeling approach to address site-specific heterogeneity, as showcased in our frailty Cox model, may be considered.

Acknowledgements
Bruce P. Burnett and Niels Riedemann, both employees of InflaRx, contributed to ideation and editing of the commentary.

Authors' contributions
CES and PHKS co-wrote the commentary.

Funding
No funding was provided for the writing of this article. The PANAMO study was funded by InflaRx N.V. and supported by a grant from German Federal Government, grant number 16LW0113; VILO-COVID. The funder of the study, InflaRx N.V., had a role in the study design, data collection, data analysis, data interpretation, and writing of the report.

Data availability
As part of the site agreement signed before trial participation, investigators agreed to keep all aspects of the trial (including the resulting data) confidential. Data for the completed PANAMO trial will be shared according to applicable regulatory requirements. Supplementary data for Vlaar et al. [7] can be found at mmc1 (4).pdf.

Declarations

Ethics approval and consent to participate
The study was approved by the institutional review boards or ethics committees for each site and conducted in accordance with Good Clinical Practice guidelines and the Declaration of Helsinki of 1975. The study protocol (Reference 9; appendix pp 39–250) was approved by an institutional review board (Amsterdam UMC, Amsterdam, Netherlands; IRB 2020_067#B2020179, Study Title: A Pragmatic Adaptive Randomized Controlled Phase II/III Multicenter Study of IFX-1 in Patients with Severe COVID-19 Pneumonia—"PANAMO"). All patients or their legally authorized representatives provided written informed consent. In the Netherlands, Germany, and Russia, deferred consent procedures were allowed. Deferred consent involved randomization at investigators discretion according to pre-set criteria agreed on during ethical review of the protocol, followed by the request for patient's (deferred patient consent) or representative's (deferred proxy consent) informed consent during the study [Jansen TC, Kompanje EJO, Druml C, et al.: Deferred consent in emergency intensive care research: what if the patient dies early? Use the data or not? Intensive Care Med 2007; 33:894–900].

Consent for publication
Not applicable.

Competing interests
PHKS has acted as a statistical consultant for InflaRx GmbH.

Received: 4 March 2024 Accepted: 3 December 2024
Published online: 18 December 2024

References

1. ICH Harmonized Tripartite Guideline: Structure and Content of Clinical Study Reports E3, https://database.ich.org/sites/default/files/E3_Guideline.pdf. Accessed 14 Aug, 2024.
2. ICH Harmonized Guideline Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2), https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf. Accessed 14 Aug, 2024.
3. ICH Harmonized Tripartite Guideline: Statistical Principles for Clinical Trials E9, https://database.ich.org/sites/default/files/E9_Guideline.pdf. Accessed 14 Aug, 2024.
4. Shreffler J, Huecker MR. Type I and type II errors and statistical power. 2023 Mar 13. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023.
5. Proschan MA, Waclawiw MA. Practical guidelines for multiplicity adjustment in clinical trials. Control Clin Trials. 2000;21:527–39.
6. Parker RA, Weir CJ. Non-adjustment in multiple testing in multi-arm trials of distinct treatments: rationale and justification. Clin Trials. 2020;17(5):562–6.
7. McCoy CE. Understanding the intention-to-treat principle in randomized controlled trials. West J Emerg Med. 2017;18(6):1075–8.
8. Phillips A, Haudiquet V. ICH E9 guideline 'Statistical principles for clinical trials': a case study. Stat Med. 2003;22(1):1–11; discussion 13-17.
9. Vlaar APJ, Witzenrath M, van Paassen P, for the PANAMO study group, et al. Anti-C5a antibody (vilobelimab) therapy for critically ill, invasively mechanically ventilated patients with COVID-19 (PANAMO): a multicentre, double-blind, randomised, placebo-controlled, phase 3 trial. Lancet Respir Med. 2022;10(12):1137–46.

10. Guo RF, Ward PA. Role of C5a in inflammatory responses. Annu Rev Immunol. 2005;23:821–52.
11. Emergency Use Authorization (EUA) for Vilobelimab (IFX-1) Center for Drug Evaluation and Research (CDER) Review, https://www.fda.gov/media/167044/download?attachment. Accessed 14 Aug, 2024.
12. Emergency Use Authorization of Gohibic, Letter of Authorization, https://www.fda.gov/media/166823/download?attachment. Accessed 14 Aug. 2024.
13. George B, Seals S, Aban I. Survival analysis and regression models. J Nucl Cardiol. 2014;21(4):686–94.

Publisher's Note