Journal for ImmunoTherapy of Cancer

Moving beyond conventional stratified analysis to assess the treatment effect in a comparative oncology study

Ryan Sun ,[1] Zachary McCaw,[2] Lu Tian,[3] Hajime Uno,[4] Fangxin Hong,[4] Dae Hyun Kim,[5] Lee-Jen Wei[6]

To cite: Sun R, McCaw Z, Tian L, *et al*. Moving beyond conventional stratified analysis to assess the treatment effect in a comparative oncology study. *Journal for ImmunoTherapy of Cancer* 2021;**9**:e003323. doi:10.1136/jitc-2021-003323

RS and ZM contributed equally.

Accepted 22 October 2021

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[2]Insitro, South San Francisco, California, USA
[3]Department of Biomedical Data Science, Stanford University, Stanford, California, USA
[4]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, Massachusetts, USA
[5]Hinda and Arthur Marcus Institute for Aging Research, Harvard Medical School, Boston, Massachusetts, USA
[6]Department of Biostatistics, Harvard University T H Chan School of Public Health, Boston, Massachusetts, USA

Correspondence to
Dr Lee-Jen Wei;
wei@hsph.harvard.edu

ABSTRACT

In a comparative oncology study with progression-free or overall survival as the endpoint, the primary or key secondary analysis is routinely stratified by patients' baseline characteristics when evaluating the treatment difference. The validity of a conventional strategy such as a stratified HR analysis depends on stringent model assumptions that are unlikely to be met in practice, especially in immunotherapy studies. Thus, the resulting summary is generally neither valid nor interpretable. This article discusses issues with conventional stratified analyses and presents alternatives using data from KEYNOTE-189, a recent immunotherapy trial for treating patients with metastatic, non-squamous, non-small-cell lung cancer.

To increase precision or reduce bias in estimating the overall treatment effect in comparative oncology studies, analysis of progression-free or overall survival data is routinely stratified by baseline factors associated with patients' survival.[1–5] The treatment effect is typically summarized via a stratified hazard ratio (HR). The validity of this approach depends on two assumptions: first, that the proportional hazards (PH) assumption holds within each stratum, and second, that the HRs are the same across all strata. In practice, these stringent constraints are seldom met. Consequently, the estimated HR is invalid and difficult to interpret clinically.[6–9]

In this article we use data from KEYNOTE-189, a recent study for treating patients with metastatic, non-squamous, non-small-cell lung cancer,[10] to illustrate issues with conventional stratified analysis. We then discuss a simple, alternative stratified inference approach for assessing the overall treatment effect that has been discussed extensively in the statistical—but not medical—literature.[8 9] In contrast to conventional stratified analysis, this alternative approach appropriately estimates the overall treatment effect without requiring strong modeling assumptions. Moreover, it provides the flexibility to estimate the treatment effect for patient populations that may differ from the study population. Lastly, the proposed alternative remains valid in the presence of treatment effect heterogeneity across strata. We illustrate the general approach using two summary measures, the event rate at a specific time point t and the mean survival time up to t. Unlike the conventional stratified HR, these summaries of the between-group differences are assumption-free.

It is increasingly important to understand the fundamental ideas and assumptions underlying stratified analysis. In particular, stratified studies now commonly appear in immunotherapy research across various cancers, and contemporary oncology studies commonly show violations of the modeling assumptions needed for conventional stratified HR analysis. For instance, it is well-known that certain immunotherapies demonstrate delayed treatment effects and thus violate the PH assumptions needed for HR calculations. Although the statistical literature has discouraged the use of stratified HR methods,[9] clinical studies still almost exclusively apply this approach. The goals of this article are to reiterate the issues with stratified HR analysis and to bring attention to a robust alternative procedure, enabling improved scientific communication and ensuring the validity of conclusions drawn from clinical oncology studies.

CONVENTIONAL STRATIFIED SURVIVAL ANALYSIS

The KEYNOTE-189 study randomized patients with metastatic, non-squamous, non-small-cell lung cancer to receive a combination of pemetrexed/platinum chemotherapy plus either pembrolizumab or placebo (in a 2:1 treatment allocation ratio).[10] There were 387 and 191 patients in the pembrolizumab and placebo arms, respectively. The primary

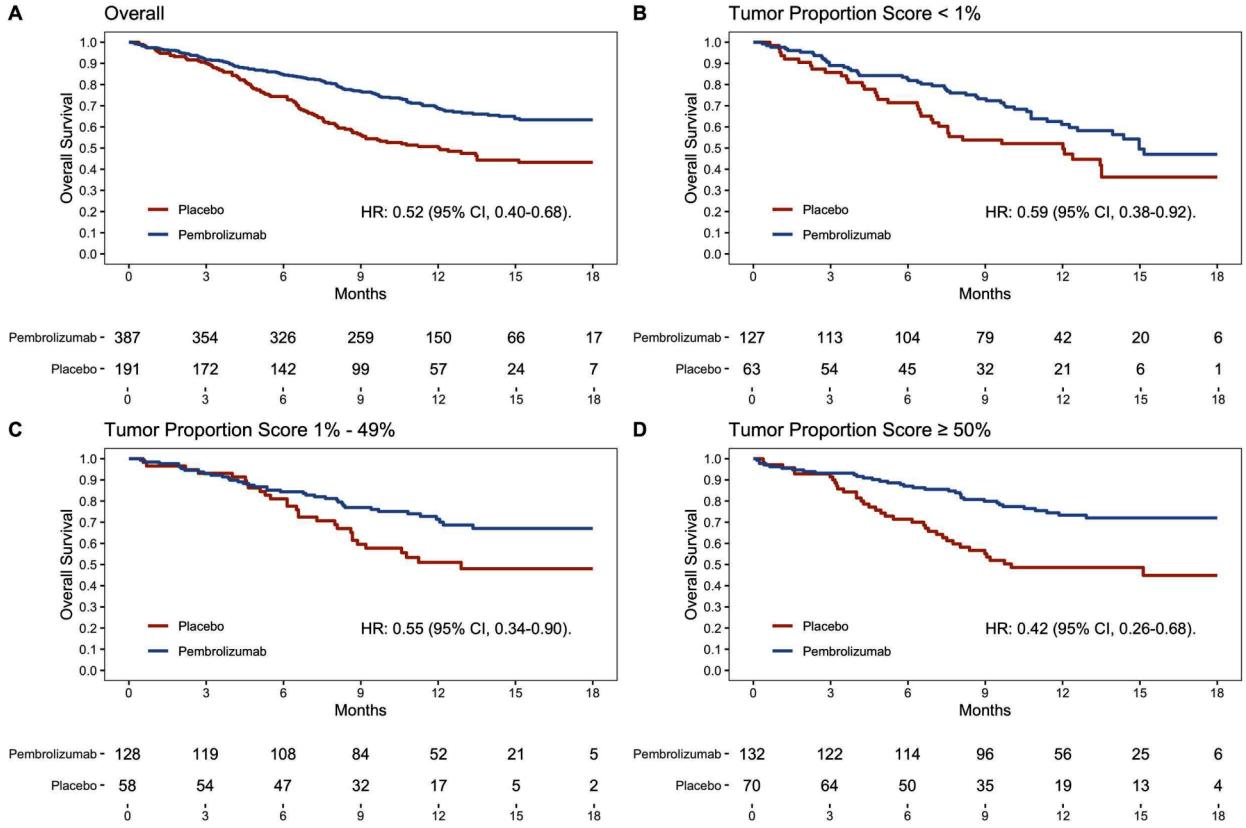

Figure 1 Kaplan-Meier curves based on reconstructed survival data among the overall population (A) and stratified by the baseline programmed death ligand 1 Tumor Proportion Score (B–D).

analysis for overall survival was stratified based on the patient's programmed death ligand 1 (PD-L1) Tumor Proportion Score (TPS), the choice of platinum-based therapies, and smoking history. For ease of illustration, we consider TPS as the only stratification factor in this article. Figure 1 presents Kaplan-Meier curves among all patients and for three strata defined by TPS <1%, 1%–49%, and ≥50%, obtained by reconstructing the survival data from figure 2 of the KEYNOTE-189 publication.[10 11]

In figure 1A, the overall unstratified HR is 0.52 (95% CI 0.40 to 0.68). Within strata, the HRs are 0.59, 0.55, and 0.42 (figure 1B–D). The Kaplan-Meier curves in figure 1A are not separated for approximately the first 3 months and are parallel after around 6 months. These patterns indicate that the PH assumption is not met in the overall study population when comparing overall survival between treatment and placebo. The profiles of the stratum-specific curves also suggest deviations from the PH assumption. For example, in figure 1C, the two survival curves are not distinguishable until month 6. Again, lacking PH, the clinical interpretation of the stratum-specific HR becomes unclear. The HR from the stratified Cox model is 0.53 (95% CI 0.41 to 0.70), which does not suggest that patients receiving pembrolizumab are 47% less likely to die than those receiving control because the hazard is not a probability measure like risk. More specifically, the hazard lacks basic properties that all probabilities possess. For instance, the hazard can be greater than 1, and the average hazard across strata is not equal to the overall hazard. Thus, it is inappropriate to interpret hazards as risks. Rather, the hazard quantifies the intensity or force of mortality, and the estimate of 0.53 ostensibly means that within each TPS stratum (as opposed to in the overall population), the ratio of hazards between the pembrolizumab and placebo groups is always 0.53. Even if the PH assumptions were valid for each TPS stratum, the stratum-specific HRs vary from 0.59 to 0.42, suggesting non-constant underlying HRs across strata. Using the aforementioned stratified HR of 0.53 to summarize the survival benefit is therefore problematic, and the results are not interpretable as providing the HR in the overall population.

A SIMPLE, ASSUMPTION-FREE ALTERNATIVE

For the KEYNOTE-189 example, the study population is a mixture of three subpopulations defined by the PD-L1 levels. Here we present a simple and robust stratified analysis procedure for estimating the overall treatment effect via two complementary summary measures. First, consider the 12-month survival rate as the summary measure of interest. The stratum-specific rates are listed in table 1. The basic idea is to obtain an overall survival rate for each arm separately by taking a weighted average of the stratum-specific survival rates. A stratum's weight is

Table 1 12-month survival rates for strata defined by the programmed death ligand 1 Tumor Proportion Score

Tumor Proportion Score (%)	12-month survival		Patients in stratum	Proportion of patients in stratum (%)	Rate difference (%)	OR of survival rates
	Pembrolizumab (%)	Placebo (%)				
<1	61.0	49.6	190	32.9	11.3	1.58
1–49	70.7	49.7	186	32.2	21.0	2.44
≥50	73.2	47.2	202	34.9	26.0	3.06

the proportion of all patients belonging to that stratum. The resulting overall survival rates of the two treatment arms are then compared using a difference or ratio. This approach is simple, intuitive, and has certain optimality properties demonstrated in statistical literature.[12 13] For KEYNOTE-189, the overall survival rate at 12 months for pembrolizumab is (see table 1 for numbers used in this calculation) (0.329×61.0%)+(0.322×70.7%)+(0.349×73.2 %)=68.3%.

The corresponding survival rate for placebo is 48.8%. From these two marginal rates, the odds ratio (OR; pembrolizumab vs placebo) is 2.27 (95% CI 1.56 to 3.30, p<0.001), favoring pembrolizumab. Specifically, the overall OR of 2.27 comes from {0.683×(1–0.488)}/{0.488×(1–0.683)} (numbers reflect rounding). Unlike conventional stratified methods which combine the three unequal stratum-specific ORs of 1.58 (where 1.58={0.61×(1–0.496)}/{0.496×(1–0.61)} is the standard form of the OR), 2.44, and 3.06, the alternative procedure provides a genuine OR, together with background survival rates for each arm. These background survival rates are essential for assessing the clinical utility of pembrolizumab over placebo. Moreover, we can readily calculate other summaries of the treatment effect from the overall rates of 68.3% and 48.8% for the two arms. For instance, the corresponding survival rate difference is 19.6% (95% CI 10.6% to 28.5%).

An alternative to the survival rate that captures both the short-term and long-term survival profile is the mean survival time across the study period. This approach has been discussed extensively in the unstratified setting.[14 15] Here, we present the corresponding stratified case, which has not been discussed in the medical literature. For the survival curves in figure 1, the higher the curve, the better the therapy. Thus, the larger the area under the curve, the better. In fact, the area under the survival curve across, for instance, 18 months of follow-up is the 18-month mean survival time. For strata from low to high TPS, the 18-month mean survival times are 12.9, 14.3, and 14.7 months for pembrolizumab and 10.8, 12.2, and 11.4 months for placebo. Taking the stratum-size weighted average, as illustrated for the aforementioned survival rates, gives 14.0 and 11.5 months for pembrolizumab and placebo, respectively; that is, a randomly selected patient followed for 18 months is expected to survive 14.0 months if treated with pembrolizumab. The difference of 2.5 months (95% CI 1.4 to 3.6 months, p<0.001) favors immunotherapy. Unlike the HR, this procedure does not require any modeling assumptions and has a straightforward, clinically meaningful interpretation.

In addition, the proposed method provides flexibility for exploring the effect of treatment in other patient populations that are composed of different mixtures of the three strata. For KEYNOTE-189, there were relatively equal proportions of patients in each stratum. Had the sample been predominantly composed of patients with high TPS, for instance, with stratum proportions (0.05, 0.15, 0.80), the mean survival time difference would have been 3.1 months rather than 2.5 months. In contrast, the conventional stratified inference procedure can only provide an estimate for the study population.

CONCLUSION

Stratification can improve precision and accuracy when reporting results, especially when the proportions of patients assigned to one arm vary markedly across strata. For randomized trials, this may occur for relatively small-sized or moderately-sized trials, or in subgroup analyses of larger studies. Moreover, non-trivial treatment imbalance can also occur with respect to other baseline variables that are highly associated with the survival outcome but not included in the randomization/stratification procedure. For observational studies, stratified analysis can substantially reduce bias owing to a lack of control over treatment allocations. When assessing heterogenous stratum-specific treatment effects, the proposed stratified procedure automatically provides appropriate estimates for the overall event rate or the mean survival time in each treatment arm.

The conventional stratified analysis procedure has undesirable constraints; its results are difficult to interpret and are often invalid. Appropriate alternatives are readily available for practical usage via publicly available computer packages (including at https://github.com/zrmacc/StratSurv). These alternatives also offer the flexibility to consider different target populations. We recommend that such procedures be used for stratified analysis in practice.

Acknowledgements We acknowledge clinicians at our respective cancer centers for valuable practitioner perspective on the text.

Contributors RS, ZM, and L-JW originated the idea. LT, HU, FH, DHK, and L-JW provided review of statistical methodology. RS, ZM, and L-JW wrote the manuscript. All authors read, revised, and approved the final paper.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests There are no competing interests.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement We used reconstructed data from the KEYNOTE-189 study. The data were reconstructed using the open-source reconstructKM R package. The software implementing our methodology is publicly available at https://githubcom/zrmacc/StratSurv.

ORCID iD

Ryan Sun http://orcid.org/0000-0003-1176-1561

REFERENCES

1 Rini BI, Plimack ER, Stus V, *et al*. Pembrolizumab plus axitinib versus sunitinib for advanced renal-cell carcinoma. *N Engl J Med* 2019;380:1116–27.
2 Johnston SRD, Harbeck N, Hegg R, *et al*. Abemaciclib combined with endocrine therapy for the adjuvant treatment of HR+, HER2-, node-positive, high-risk, early breast cancer (monarchE). *J Clin Oncol* 2020;38:3987–98.
3 Gebre-Medhin M, Brun E, Engström P, *et al*. Artscan III: a randomized phase III study comparing chemoradiotherapy with cisplatin versus cetuximab in patients with locoregionally advanced head and neck squamous cell cancer. *J Clin Oncol* 2021;39:38–47.
4 Shitara K, Van Cutsem E, Bang Y-J, *et al*. Efficacy and safety of pembrolizumab or pembrolizumab plus chemotherapy vs chemotherapy alone for patients with first-line, advanced gastric cancer: the KEYNOTE-062 phase 3 randomized clinical trial. *JAMA Oncol* 2020;6:1571–80.
5 Okamoto I, Nokihara H, Nomura S, *et al*. Comparison of carboplatin plus pemetrexed followed by maintenance pemetrexed with docetaxel monotherapy in elderly patients with advanced Nonsquamous non-small cell lung cancer: a phase 3 randomized clinical trial. *JAMA Oncol* 2020;6:e196828.
6 Struthers CA, Kalbfleisch JD. Misspecified proportional hazard models. *Biometrika* 1986;73:363–9.
7 Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984;71:431–44.
8 Jiang F, Tian L, Fu H, *et al*. Robust alternatives to ancova for estimating the treatment effect via a randomized comparative study. *J Am Stat Assoc* 2019;114:1854–64.
9 Tian L, Jiang F, Hasegawa T, *et al*. Moving beyond the conventional stratified analysis to estimate an overall treatment efficacy with the data from a comparative randomized clinical study. *Stat Med* 2019;38:917–32.
10 Gandhi L, Rodríguez-Abreu D, Gadgeel S, *et al*. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *N Engl J Med Overseas Ed* 2018;378:2078–92.
11 Guyot P, Ades AE, Ouwens MJNM, *et al*. Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC Med Res Methodol* 2012;12:9.
12 Tsiatis AA, Davidian M, Zhang M, *et al*. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Stat Med* 2008;27:4658–77.
13 Tian L, Cai T, Zhao L, *et al*. On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. *Biostatistics* 2012;13:256–73.
14 Uno H, Claggett B, Tian L, *et al*. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* 2014;32:2380–5.
15 Pak K, Uno H, Kim DH, *et al*. Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. *JAMA Oncol* 2017;3:1692–6.