Momal *et al. Trials (2023) 24:380*
https://doi.org/10.1186/s13063-023-07375-0

Trials

METHODOLOGY Open Access

More efficient and inclusive time-to-event trials with covariate adjustment: a simulation study

Raphaëlle Momal[1†], Honghao Li[1†], Paul Trichelair[1], Michael G. B. Blum[1] and Félix Balazard[1*]

Abstract

Adjustment for prognostic covariates increases the statistical power of randomized trials. The factors influencing the increase of power are well-known for trials with continuous outcomes. Here, we study which factors influence power and sample size requirements in time-to-event trials. We consider both parametric simulations and simulations derived from the Cancer Genome Atlas (TCGA) cohort of hepatocellular carcinoma (HCC) patients to assess how sample size requirements are reduced with covariate adjustment. Simulations demonstrate that the benefit of covariate adjustment increases with the prognostic performance of the adjustment covariate (C-index) and with the cumulative incidence of the event in the trial. For a covariate that has an intermediate prognostic performance (C-index=0.65), the reduction of sample size varies from 3.1% when cumulative incidence is of 10% to 29.1% when the cumulative incidence is of 90%. Broadening eligibility criteria usually reduces statistical power while our simulations show that it can be maintained with adequate covariate adjustment. In a simulation of adjuvant trials in HCC, we find that the number of patients screened for eligibility can be divided by 2.4 when broadening eligibility criteria. Last, we find that the Cox-Snell R^2_{CS} is a conservative estimation of the reduction in sample size requirements provided by covariate adjustment. Overall, more systematic adjustment for prognostic covariates leads to more efficient and inclusive clinical trials especially when cumulative incidence is large as in metastatic and advanced cancers. Code and results are available at https://github.com/owkin/CovadjustSim.

Keywords Covariate adjustment, Trial design, Cox regression, Eligibility criteria

Background

Adjustment for prognostic covariates improves precision and increases statistical power for treatment effect estimation in randomized clinical trials [1-4]. Randomization guarantees the validity of statistical analysis of randomized trials whether they are adjusted or unadjusted [5]. However, unadjusted analysis can be imprecise because of a large variability between patient outcomes that could be explained by several baseline covariates. Covariate adjustment for prognostic covariates accounts for outcome variation between patients, leading to a more precise estimation of the treatment effect while adjusting for random noise will lead to small decrements of power [1]. While adjustment on important covariates can correct for chance imbalance in important baseline covariates, adjustment covariates should be selected and prespecified at the trial design stage based on their prognostic value and not on any imbalance criterion assessed after randomization [6]. This methodological consensus is currently being translated into regulatory guidance: the European Medicines Agency (EMA) published a guideline in 2015 and the Food and Drug Administration

[†] Raphaëlle Momal and Honghao Li contributed equally.

*Correspondence:
Félix Balazard
felix.balazard@owkin.com
[1] Owkin Inc, New York, USA

(FDA) has issued a draft guidance in 2021 [7, 8]. Increase of precision when using covariate adjustment translates to a reduced sample size for reaching a target of statistical power, typically at least 80% in clinical trials.

For time-to-event trials that are frequent in oncology, we investigate to what extent trial and indication characteristics determine the impact of covariate adjustment on statistical power and on sample size requirements. These characteristics include the cumulative incidence of the event of interest at the end of follow-up, the prognostic performance of covariates, and the censoring rate. Understanding the relationship between cumulative incidence and reduction in sample size helps prioritize the disease indications where covariate adjustment is the most impactful.

We also evaluate whether covariate adjustment can help to broaden trial eligibility criteria. Eligibility criteria in clinical trials can be too restrictive which leads to limited generalizability as well as difficulty in enrollment [9, 10]. Beyond ensuring patient safety, restrictive eligibility might be used to ensure homogeneity in the trial population [11, 12]. In non-small cell lung cancer, it was shown using observational cohorts that many inclusion criteria are superfluous as they restrict the potential enrollment of trials even though the treatment is as efficacious for the excluded patients as for the included patients [13]. As covariate adjustment allows to analytically compensate for the heterogeneity in the patient population, we investigate whether adequate covariate adjustment could allow to broaden eligibility criteria while maintaining statistical power.

To answer both of those questions, we use parametric simulations as well as semi-synthetic simulations based on data from patients with resected HCC. In parametric simulations, event times are simulated based on an extensive exploration of the parameter space. The semi-synthetic simulations are based on TCGA data [14]. The covariate of interest, which is used for adjustment, is named HCCnet and it captures a prognostic signal on overall survival for HCC after resection [15]. More specifically, it is a continuous measure of the risk at per-patient level, with higher value indicating higher risk. For each patient, the HCCnet value is determined by applying the deep-learning model to the patient's histological slide. In both cases, the simulations rely on the proportional hazards assumption.

Last, we determine how sample size could be determined if the prognostic signal carried by the covariate is known a priori based on external data. For a continuous outcome, the Fleiss formula relates the sample size of the adjusted analysis (denoted N_{adj}), which is required for a given statistical power, to the sample size of the unadjusted analysis (denoted N_0). Denoting by r^2 the proportion of variance of the outcome explained by the covariate, the Fleiss formula states that the sample size needed for the adjusted analyses is reduced by r^2 compared to the unadjusted one $N_{\text{adj}} = N_0\left(1 - r^2\right)$ [16]. For instance, a correlation r of 0.5 between a baseline covariate and the outcome translates to sample size requirements for the adjusted analysis reduced by 25% compared to the unadjusted analysis. For a time-to-event outcome, there are several alternative definitions for the proportion of variation explained by a covariate. Different measures to compute the proportion of explained variance have been proposed for time-to-event analysis [17, 18]. Using the parametric simulations, we assess whether the Fleiss formula can be extended to the time-to-event setting.

Methods

Parametric simulations based on a time-to-event model

Parametric simulations are performed to estimate the observed reduction of the sample size requirement and assess its relationship with a single adjustment covariate's C-index and the cumulative incidence of the event at the end of the trial. Other parameters of interest are the size of the treatment effect, the Weibull shape of the baseline hazard function, and the drop-out rate. The simulations rely on the proportional hazard assumption.

Event times are generated following the Weibull distribution with shape w and scale depending on the treatment hazard ratio θ, and on a standard Gaussian covariate x. Censoring times T^{drop} are drawn from an exponential distribution with a specified drop-out rate d. Denoting z the treatment allocation variable, κ the intercept, and β the coefficient of x, this generative model can be formally summarized as follows for patient i:

$$\begin{cases} h_i(z) = \theta \exp\left(\kappa + \beta x_i\right), \\ T_i(z) \sim W\left(h_i(z)^{-\frac{1}{w}}, \omega\right), \\ \quad T_i^{drop} \sim \varepsilon(d). \end{cases}$$

All patients remaining at risk at 5 years are censored at that time. The treatment allocation is independent of the covariate and there is the same number of patients in both arms. For each set of input parameters, the auxiliary parameters κ and β are numerically optimized to reach prespecified values of the cumulative incidence at the end of the trial Λ and the C-index C evaluated in the control arm.

Once event times are simulated, the presence of a treatment effect is tested in an unadjusted analysis and an analysis adjusted for the covariate using the Wald test for the treatment coefficient in a Cox regression. The statistical power for the unadjusted analysis and the adjusted analysis is estimated on a grid of sample sizes based on 10,000 numerical replications per sample size [19]. The resulting power curves give the sample sizes N_{adj} and N_0 required to reach a power of 80% for both analyses, from which the reduction of sample size achieved with adjustment R^2_{obs} is deduced (Fig. 1). These simulations explore a wide range of parameter values (Table S1), allowing for an extensive study of R^2_{obs} behavior as a function of the cumulative incidence Λ and c-index C in different settings of proportional hazards.

To indicate what are the most relevant indications for covariate adjustment, we provide estimates of the cumulative incidence Λ in the control arm for several oncology trials. Cumulative incidence is estimated by reading the value of the Kaplan-Meier curves published in the manuscript describing the trial results. More details on parametric simulations can be found at https://github.com/owkin/CovadjustSim.

Semi-synthetic simulations based on HCC data from TCGA

To consider simulations that mimic distributions of covariates found in clinical data, we also perform semi-synthetic simulations of resected HCC patients. The covariate used for adjustment is a prognostic score based on hematoxylin and eosin stained (H&E) images processed with the HCCnet deep learning algorithm [15]. The deep learning model was trained on another dataset than TCGA. We consider the prognostic scores of HCCnet applied on 328 patients with early stage HCC from the TCGA HCC dataset [14, 15]. In the TCGA dataset, we have access to outcome measures including overall survival and 34 clinical variables with less than 50% of missing data in addition to the HCCnet prognostic covariate.

We impute all missing values among the 34 clinical variables. For imputation, we use factorial analysis for mixed data (FAMD), a principal component method for data involving both continuous and categorical variables [20]. The imputed variables used as adjustment are tumor staging (1% missing values) and Eastern Cooperative Oncology Group (ECOG) score which have 20% missing values. The imputed variables used as eligibility

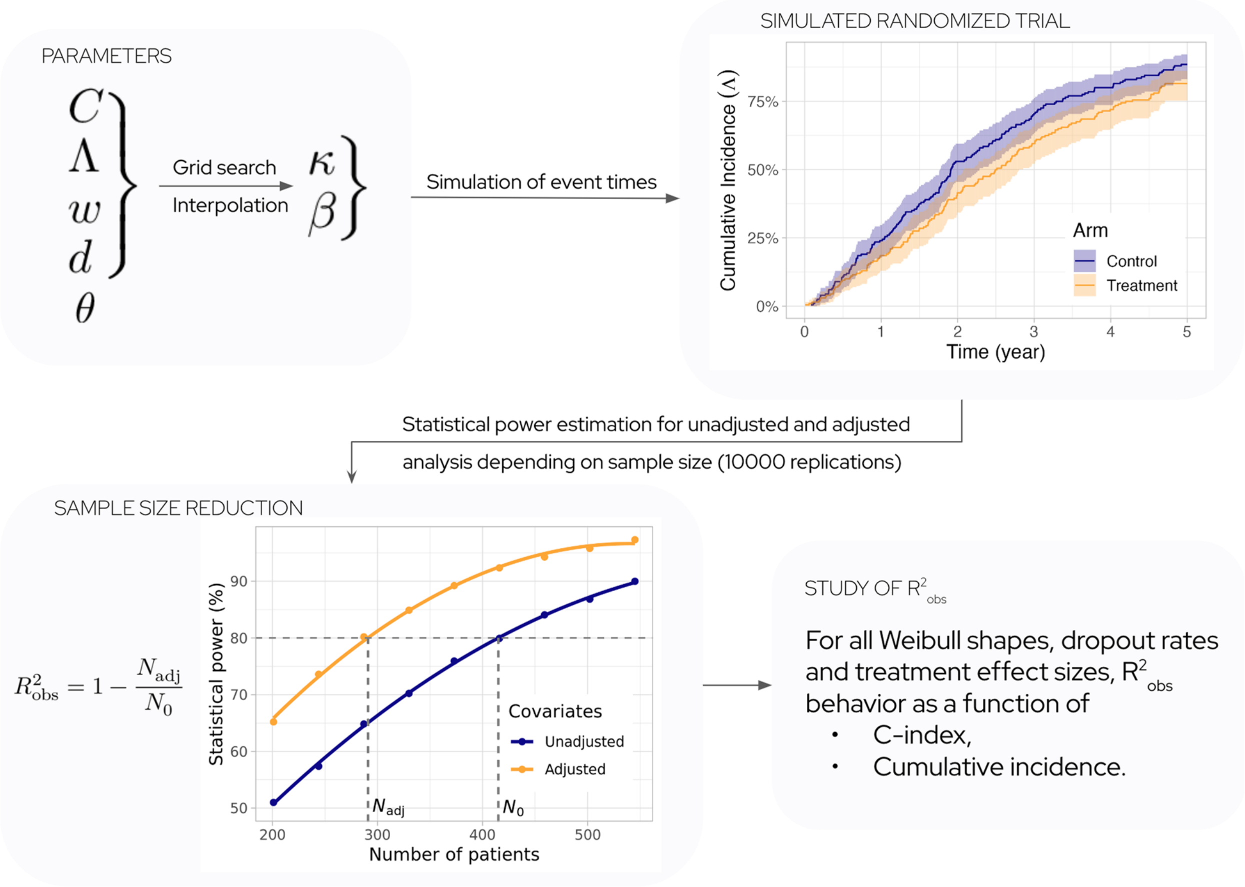
PARAMETERS
C
Λ
w
d
θ
Grid search
Interpolation
κ
β
Simulation of event times
SIMULATED RANDOMIZED TRIAL
Cumulative Incidence (Λ)
0%
25%
50%
75%
0
1
2
3
4
5
Time (year)
Arm
Control
Treatment
Statistical power estimation for unadjusted and adjusted
analysis depending on sample size (10000 replications)
SAMPLE SIZE REDUCTION
$R^2_{\text{obs}} = 1 - \frac{N_{\text{adj}}}{N_0}$
Statistical power (%)
50
60
70
80
90
200
300
400
500
Number of patients
N_{adj}
N_0
Covariates
Unadjusted
Adjusted
STUDY OF R^2_{obs}
For all Weibull shapes, dropout rates and treatment effect sizes, R^2_{obs} behavior as a function of
C-index,
Cumulative incidence.

Fig. 1 Workflow of parametric simulations. For a set of parameters corresponding to a clinical trial scenario, 10,000 instances of clinical trials are simulated to estimate statistical power. The parameters used to obtain the illustrative Kaplan-Meier curves and power curves are $C = 0.65, \Lambda = 0.9, w = 1.5, d = 0, \theta = 0.7$. The number of patients in the power curve is the sum of both arms

criteria in our simulation study are the ECOG score, the Child-Pugh classification (33% missing), the macrovascular invasion (15% missing), and B or C hepatitis infection status (15% and 5% missing values respectively).

The simulations follow the same assumptions as the parametric ones while preserving the observed survival curve and dependence of survival on covariates. To do so, a Cox model of overall survival is fitted on the available prognostic variables (tumor staging, ECOG score, and the HCCnet variable). For each simulated patient, we sample the clinical covariates from TCGA. The hazard rate is defined as for parametric simulations except that there is a matrix X of covariates instead of a single covariate, and β is replaced by $\widehat{\beta}$ the vector of coefficients obtained from the fitted Cox model. The Weibull distribution is replaced by the empirical survival function that depends on the hazard rate and on the baseline survival function $\widehat{S}_0$ fitted with the same Cox model on a null data point (baseline hazard):

$$\begin{cases} h_i(z) = \theta^z \exp\left(\kappa + \widehat{\beta}^T X_i\right), \\ \widehat{S}(t|z, X_i) = \widehat{S}_0(t)^{h_i(z)}, \\ T_i(z) \sim \widehat{S}(\bullet|z, X_i). \end{cases}$$

As before, all patients with events after 5 years are censored at that time.

We choose a sample size of 760 individuals as it is the average sample size of 4 ongoing trials for adjuvant treatment in early stage HCC [21-24]. The treatment effect size is set to $\theta = 0.72$ so that the estimated statistical power with adjustment for the clinical variables (tumor staging and ECOG score) is 80% for a sample size of 760 individuals. Randomization of the treatment assignment is stratified on tumor staging. To estimate the reduction of sample size obtained when adding HCCNet as adjustment covariate, we consider varying values of the sample size, find the minimal values where power reaches 80%, and compute the relative reduction of sample size compared to the sample size of 760 individuals. Statistical power is estimated based on 10,000 replications.

Effect of covariate adjustment when broadening eligibility criteria

Using the parametric simulations and the semi-synthetic simulations, we evaluate if the effect of covariate adjustment is changed when considering less restrictive inclusion criteria. These simulations assume that the treatment hazard ratio is constant across the entire population. For parametric simulations, the restricted inclusion criteria is based on the values of the prognostic covariate X. Only patients with values of X below the 80% quantile, i.e., patients at lower risk, are included in the simulated trial with the more restrictive eligibility criteria; this cohort is therefore expected to have a lower cumulative incidence than the less restrictive cohort. There are then 4 scenarios when combining the two possible eligibility criteria (all patients or restricted inclusion) and the two choices of adjustments (no adjustment or adjustment for X). Parameters of the simulations include the log hazard ratio of the covariate β, the intercept of the Cox model κ, the Weibull shape w, and the treatment hazard ratio r. We set $w = 1.5$, and $\theta = 0.7$. The remaining parameters β and κ are fixed so as to reach 0.65 of c-index and 0.9 of cumulative incidence in the control arm of the study with less restrictive criteria.

In the case of the HCC semi-synthetic simulations, we consider that including all TCGA patients selected for HCCnet validation is the less restrictive inclusion criteria and we define two additional levels of restricted eligibility criteria (Table 1). The mildly restrictive eligibility level has two inclusion criteria present in all 4 ongoing large trials for adjuvant treatment in early stage HCC [21-24]: only patients with a Child-Pugh score of A and with an ECOG status of 0 or 1 are included. The most restrictive eligibility criteria further restrict the ECOG status to 0 as in the STORM trial [25], exclude patients with a dual infection of hepatitis B and hepatitis C as in the KEYNOTE-937 trial [23] and exclude patients with macrovascular invasion as in the IMBRAVE050 trial [22]. We consider only the eligibility criteria that were available in the TCGA HCC dataset. In summary, more restrictive

Table 1 Definition of eligibility criteria used for the semi-synthetic simulations based on the TCGA dataset. The more restrictive eligibility criteria exclude patients with comorbidities who can be expected to have worse outcomes. *N* denotes the number of TCGA patients who meet the eligibility criteria

Eligibility level	Nested inclusion criteria	*N* (%)
Less restrictive	All TCGA patients selected for HCCnet validation [15]	328 (100%)
Mildly restrictive	Child Pugh classification is A ECOG ≤ 1	270 (82%)
Most restrictive	ECOG score of 0 No macrovascular invasion No cumulated hepatitis B and C infection	169 (52%)

eligibility criteria exclude patients with increased disease severity. The group with the most restrictive eligibility criteria is expected to have the lowest cumulative incidence, and the lowest HCCnet score on average. There are therefore 6 different scenarios when combining the three levels of eligibility criteria and the two choices of adjustment: whether or not HCCnet is considered as an adjustment variable in addition to tumor staging and ECOG. In the scenario with the most restrictive eligibility levels, every patient has an ECOG of 0 and therefore the analyses are not adjusted for ECOG.

For both types of simulations, changing the inclusion criteria changes the number of events which affects statistical power directly. To provide a fair comparison between the methods with or without adjustment, we present the statistical power of the different scenarios as a function of the number of events. In both cases, no drop-out was added and 10,000 replications were generated to evaluate statistical power. In both cases, patients at lower risk of the event are selected when we consider the more restrictive criteria. We also evaluate how broadening the eligibility criteria would impact the number of patients that need to be screened for enrollment to succeed.

Proposed R2 measures for time-to-event analysis

Several categories of measures have been proposed to extend the R^2 measure to time-to-event data [17, 18]. We consider explained variation (EV) and explained randomness (ER) measures. Explained variation measures are extensions of the proportion of explained variance that is used in linear regression. Explained randomness measures are based on entropy measures and compare the quantity of information contained in models with and without the covariates of interest. In the simulations, we study the behavior of four EV measures: R^2_{D}, R^2_{I}, R^2_{PM}, and R^2_{R} [26-28], and four ER measures: ρ^2_{k}, ρ^2_{WA}, ρ^2_{XOQ}, and R^2_{CS} [26, 27, 3129-]. The proposed R^2 measures are estimated over the grid of simulation parameters in Table S1 and are compared to the observed reduction in sample size. Each estimation of R^2 for a set of parameters is an average of 1000 R^2, each evaluated with a simulated dataset of 1000 control patients.

Results

Evaluation of the parameters impacting sample size reduction with parametric simulations

The parametric simulations show that the sample size reduction obtained with covariate adjustment varies between 0 and 86%. It increases as a function of the covariate prognostic performance measured with the C-index, and of cumulative incidence, which corresponds to the probability of an event (death, progression...) before the end of the follow-up period. When we consider a cumulative incidence of $\Lambda = 10\%$, covariate adjustment reduces the sample size by 3.1% for a covariate with a C-index of 0.65, by 9.5% for a C-index of 0.75, and by 32.7% for a C-index is 0.85. For an intermediate value of $\Lambda = 50\%$, the reduction is 16.8%, 42.7%, and 73.0% for the three C-index values of 0.65, 0.75, and 0.85. For a high cumulative incidence value of $\Lambda = 90\%$, the reduction is 29.1%, 61.3%, and 85.7% for the same values of C-index (Fig. 2).

Cumulative incidence values depend on indication, on the nature of the event (progression, death...), and on the duration of follow-up (Table 2). We find a wide range of values for cumulative incidence in several oncology trials. It ranges from 18.6% at 5 years for disease recurrence in early breast cancer to 98% at 3 years for death in metastatic pancreatic cancer (Table 2).

We find that other parameters of the simulations do not impact the reduction of sample size obtained with covariate adjustment. These additional parameters are the size of the treatment effect (hazard ratio), the Weibull shape parameter, and the drop-out rate (Figure S1).

The drop-out rates of $d = 0.01$ or $d = 0.1$ result in different average censoring rates depending on the values taken by other parameters. The median censoring rate (computed over the set of other parameters' value) before the end of follow-up was 7.6% when $d = 0.01$ (min-max: 0.8–47.3%) and 46.3% when $d = 0.1$ (min-max: 6.7–89.8%).

Comparing semi-synthetic HCC simulations and parametric simulations

We consider semi-synthetic simulations based on the TGCA HCC cohort to evaluate power gain obtained with a deep learning variable. We find that adjusting on the deep learning covariate HCCnet, in addition to tumor staging and ECOG, reduces the required sample size to reach 80% statistical power by $R^2_{\mathrm{obs}} = 1 - N_0/N_{\mathrm{adj}} \simeq 1 - 671/759 = 11.6\%$ (figure S2). For the sample size that provides a power of 80% when adjusting on ECOG and tumor staging only, the statistical power increases by 5% in absolute value when adjusting also on the deep learning covariate.

We evaluate the compatibility of this result with the results of the parametric simulations. The cumulative incidence of death in the HCC-TCGA population is 49% at 5 years. The Cox model with tumor staging and ECOG score as covariates has a C-index of 0.65 in the simulated population, while adding the HCCnet covariate results in a C-index of 0.70. We label by 1 the quantities associated with adjustment for the clinical variables (tumor staging and ECOG) and by 2 the quantities associated with the additional adjustment of the HCCnet covariate (tumor staging, ECOG, and HCCnet). Applying Fleiss equation

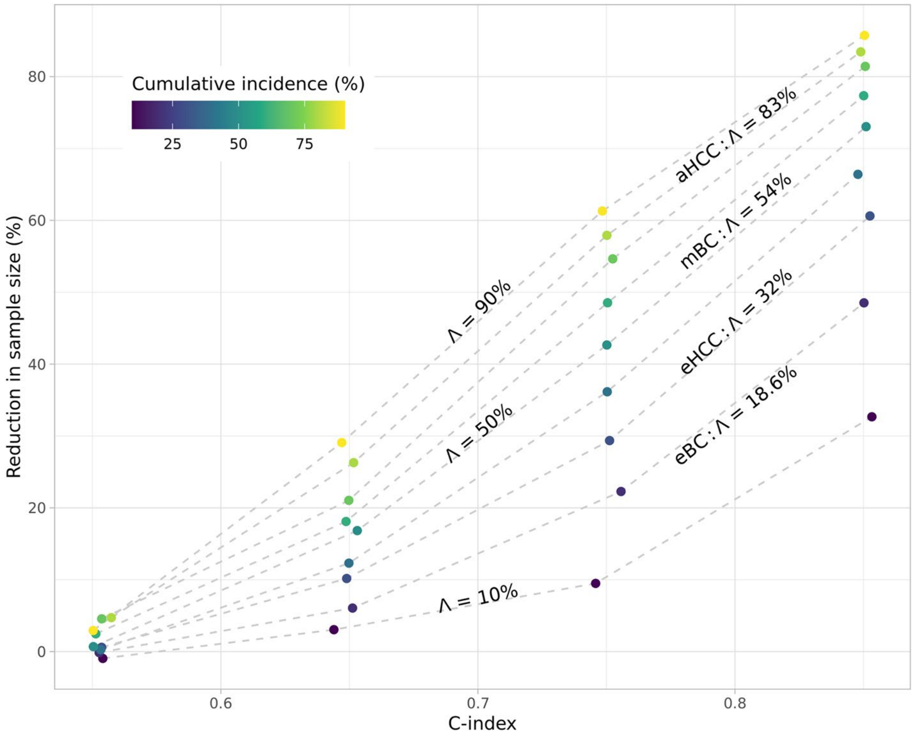

Fig. 2 Reduction in sample size R^2_{obs} as a function of the prognostic performance (C-index) of the covariate for a range of cumulative incidence values. Cumulative incidence Λ is measured at the end of the follow-up period. In the simulations, the hazard ratio is set at $\theta = 0.7$, the drop-out rate at $d = 0.01$, and the shape parameter of the Weibull distribution at $w = 1.5$. The cumulative incidence values that are provided for the breast cancer and HCC indications come from clinical trials selected in Table 2. eBC, early breast cancer; eHCC, early resectable hepatocellular carcinoma; mBC, metastatic breast cancer; aHCC, advanced hepatocellular carcinoma

Table 2 Cumulative incidence of events of interest in the control arms of a selection of trials. For a given C-index of a prognostic covariate, the impact of covariate adjustment will be larger for indications with large cumulative incidence of events. *HR+*, hormone receptor positive; *PD-L1+*, programmed death ligand 1 positive; *NSCLC*, non-small cell lung cancer

Indication	Trial	Cumulative incidence Λ in control arm
HR+ early breast cancer (eBC)	BIG 1-98 [32] Letrozole vs tamoxifen	Probability of disease recurrence at 5 years: 18.6%
HCC after resection of local ablation (eHCC)	STORM [25] Sorafenib vs placebo	Probability of death at 5 years: 32%
Metastatic hormone-sensitive prostate cancer	ENZAMET [33] Enzalutamide vs standard nonsteroidal antiandrogen therapy in addition to testosterone suppression	Probability of death at 4 years: 36%
PD-L1+ advanced NSCLC	KEYNOTE-024 [34] Pembrolizumab vs chemotherapy	Probability of death at 1.5 years: 50%
HR+ metastatic breast cancer in premenopausal patients (mBC)	MONALEESA-7 [35] Ribociclib vs placebo in addition to endocrine therapy	Probability of death at 3.5 years: 54%
Resected pancreatic cancer	PRODIGE 24 [36] Modified FOLFIRINOX vs gemcitabine	Probability of death at 5 years: 70%
Advanced HCC (aHCC)	CheckMate 459 [37] Nivolumab vs sorafenib	Probability of death at 3 years: 83%
Malignant pleural mesothelioma	CheckMate 743 [38] Nivolumab+Ipilimumab vs chemotherapy	Probability of death at 3 years: 85%
Metastatic pancreatic cancer	OXIPAN [39] FOLFIRINOX vs gemcitabine	Probability of death at 3 years: 98%

for the two adjustments using the $R^2_{\text{obs},i}$ obtained with parametric simulations (Fig. 2), we obtain

$$N_{adj,2}/N_{adj,1} = \left(1 - R^2_{obs,2}\right)/\left(1 - R^2_{obs,1}\right) \simeq 0.73/0.84 = 0.869 = 100\% - 13.1\%$$

Therefore, results obtained with semi-synthetic simulations are coherent with the findings of the parametric simulations.

It should be noted that the impact depends on the added prognostic performance of a covariate and is not linked to the specific nature of the covariate.

Covariate adjustment when broadening eligibility criteria

For the unadjusted analysis, statistical power for a fixed number of events is increased when restricting the eligibility criteria. By contrast, for the adjusted analysis, the broader inclusion criteria have the same statistical power as the narrower one (Fig. 3).

While the adjusted analyses with different eligibility criteria have the same statistical power, they imply a very different screened population size. Screened individuals are patients for which eligibility criteria is evaluated to test if they can be enrolled in the clinical trial. In the HCC example, the required size of the screened population is 667 for the less restrictive inclusion while it is 1629 for the most restrictive population. Therefore, the size of the screened population is divided by 2.4 when broadening eligibility criteria while attaining the same statistical power. This difference is explained by the smaller proportion of patients included as well as the smaller proportion of events with the restrictive eligibility criteria (34.8% at 5 years versus 44.2% in the entire population).

Fit with R^2 measures from the literature

We compare various R^2 measures for time-to-event endpoints to the reduction of sample size R^2_{obs} provided by covariate adjustment for the grid of parameters considered in parametric simulations (Figure S3). Most measures do not depend on the cumulative incidence of the event at the end of follow-up (Figure S3), which is not compatible with the results found for the reduction of sample size provided by covariate adjustment (Fig. 2). Most measures increase only as a function of the C-index (Figure S3). The Cox-Snell R^2_{CS} best captures the observed sample size reduction in all our simulations. The median absolute error is minimal for the R^2_{CS} and is 3.2% (first and third quartiles are 0.9% and 8.4% respectively). For large values of R^2_{CS}, R^2_{obs} is underestimated by R^2_{CS}. Median absolute error for other R^2 measures

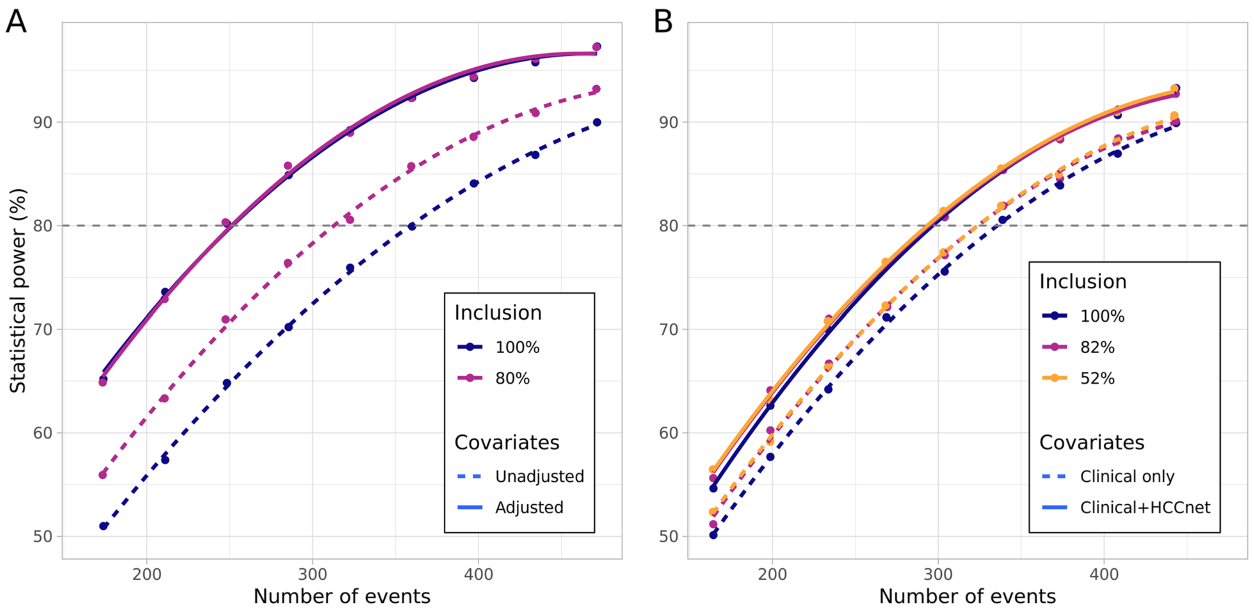

Fig. 3 Effect of broader eligibility criteria and of covariate adjustment on statistical power. Different inclusion statuses are shown by color and adjustment statuses by type of line, irrespective of color. (A) Results of the parametric simulations where the covariate adjusted for is a standard Gaussian. (B) Results of the semi-synthetic simulations based on the HCC-TCGA cohort. The clinical covariates adjusted for are ECOG score and tumor staging. The three levels of inclusion are based on eligibility criteria of past and ongoing trials outlined in Table 1. For all simulations, a constant treatment effect size is assumed across the population. More restrictive eligibility criteria exclude patients with higher disease severity

are 5.2% for R^2_{PM} (1.9–10.9%), 5.2% for R^2_{D} (1.9–11.0%), 6.6% for R^2_{R} (2.0–13.5%), 6.3% for R^2_{I} (2.5–17.6%), 6.4% for ρ^2_{WA} (2.6–17.7%), 7.2% forρ^2_{XOQ} (2.6–21.2%), and 7.5% forρ^2_{k} (2.6–21.2%).

Using the Fleiss formula and the Cox-Snell R^2_{CS} measure, we find that further adjusting on HCCnet—in addition to clinical covariates—in an adjuvant HCC trial would decrease the sample size by 9.2%, which is a slight underestimation of the 11.6% reduction in sample size found with the semi-synthetic simulations, and is coherent with the results of the parametric study presented above.

Discussion

The impact of covariate adjustment depends on several characteristics related to indications and clinical trials. Our simulations confirm the expected result that the power gains increase with the prognostic performance, measured by C-index, of the covariates used in covariate adjustment. Other parameters that were considered such as Weibull shape, drop-out rate, or effect size do not play an important role in determining power gain. Previous work on the topic already identified that the drop-out rate and effect size do not impact the precision gains obtained with covariate adjustment [2].

Cumulative incidence at the end of the follow-up period is another major determinant of the impact of covariate adjustment. Compared to earlier work [2], we considered a finite time horizon (i.e., follow-up of 5 years) which allowed us to identify the strong dependence on cumulative incidence. Dependence on cumulative incidence is related to the dependence on the prevalence of events that occur for binary outcomes [3]. We investigated cumulative incidence for several published trials in oncology. Covariate adjustment will have limited impact for trials of new endocrine therapies for early breast cancer. For indications with low cumulative incidence, prognostic information can be more useful to perform prognostic enrichment than for covariate adjustment [40]. For aggressive cancers such as mesothelioma, metastatic breast cancer, or metastatic pancreatic cancer, covariate adjustment provides notable gains in precision.

Another advantage of covariate adjustment is that it removes incentive to homogenize the population with restrictive eligibility criteria if we assume a constant treatment effect across the population. In both simulation scenarios, the adjusted analyses are just as powerful whether there are strict eligibility criteria or not. However, the size of the population that needs to be screened for inclusion can be reduced substantially with the least restrictive eligibility criteria. More importantly, broader eligibility criteria imply a broader potential target population. Adequate covariate adjustment can therefore go hand in hand with broader eligibility criteria that would allow easier enrollment as well as better generalizability of trial results. This would be in line with recent calls for less restrictive eligibility criteria [10, 11].

When comparing two designs of a clinical trial, one without covariate adjustment and the other with covariate adjustment, the adjusted trial will have the additional practical burden of data collection of the predefined covariates, e.g., digitizing histology slides to apply HCCnet. However, the associated cost can provide a large return on investment by improving the statistical power. This can lead to a reduction of the size of the population included in the trial and therefore a reduction in the time and effort spent. When comparing a trial with restrictive eligibility criteria and without adjustment with a trial with a broader eligibility criteria and with adjustment, the former will not have the advantage of less data collection given that the screening will require collecting a large amount of information. Further, the more inclusive adjusted trial will have the added advantage of reducing the size of the population considered during the recruitment and screening phase and the associated costs.

We evaluated the sample size reduction brought by covariate adjustment by investigating whether several R^2 measures could approximate the observed sample size reduction. We found that the Cox-Snell R^2_{CS} was the best approximation of our quantity of interest. The sample size with adjustment is then $N_{\mathrm{adj}} = N_0\left(1 - R^2_{\mathrm{CS}}\right)$ and this generalizes the Fleiss formula to a time-to-event outcome. When denoting n the number of patients, and l_0 and l_1 the log-likelihoods of a base model and a model adjusting for additional covariates, we have $R^2_{\mathrm{CS}} = 1 - \exp\left[-\frac{2}{n}(l_1 - l_0)\right]$ [31]. Other R^2 measures we consider were developed such as they do not depend on cumulative incidence explaining why they cannot approximate the reduction of covariate adjustment provided by covariate adjustment [17, 18].

Approximate sample size is of practical importance in the design of clinical trials. It could also be useful in the case of a blinded sample size reestimation when there is uncertainty on the prognostic performance of adjustment covariates and where the required number of events should be reevaluated at an interim stage. Blinded sample size reestimation procedures have been proposed for a continuous outcome and could be generalized for time-to-event outcomes [41].

As noted in the draft FDA guidance, covariate adjustment changes the target of estimation, a phenomenon called non-collapsibility [8]. When adjusting for a prognostic covariate and when there is a true treatment effect (e.g., hazard ratio not equal to 1), it is expected that the conditional estimand (e.g., hazard ratio) drifts further away from 1 compared to the marginal estimand and the variance is increased. Because the amount of drift is superior to the inflation of variance, statistical power resulting from covariate adjustment is increased [42] as confirmed in our simulations. If a marginal estimand is preferred, one can consider adjusted marginal estimators that target the estimand of the unadjusted analysis while leveraging the gain in precision offered by covariate adjustment [42, 43].

Our simulations study the effect of covariate adjustment on a relative measure of treatment effect, which is the hazard ratio. Absolute measures of efficacy such as restricted mean survival time or absolute risk reduction are also of interest and do not rely on the proportional hazards assumption. Estimation of those measures can also be improved by using the prognostic signal of covariates [44-46]. The extent to which our findings, for instance the dependence on cumulative incidence, generalize to this setting should be studied in further work.

Overall, we have shown that covariate adjustment reduces the sample size that is needed to reach a targeted statistical power. Reduction is particularly pronounced for indications where cumulative incidence is large. Furthermore, adequate covariate adjustment allows to maintain statistical power while relaxing eligibility criteria. New sources of prognostic covariates such as deep-learning models based on images can lead to more efficient trials.

Abbreviations

eBC	Early breast cancer
mBC	Metastasis breast cancer
ECOG	Eastern Cooperative Oncology Group
EMA	European Medicines Agency
ER	Explained randomness
EV	Explained variation
FAMD	Factorial analysis for mixed data
FDA	Food and Drug Administration
HCC	Hepatocellular carcinoma
aHCC	Advanced HCC
H&E	Hematoxylin and eosin
HR+	Hormone receptor positive
NSCLC	Non-small cell lung cancer
PD-L1+	Programmed deathligand 1 positive
TCGA	The Cancer Genome Atlas

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13063-023-07375-0.

Additional file 1: Table S1. Description of simulation parameters used for parametric simulations of the time-to-event model. **Figure S1.** Evolution of R^2_{obs} as a function of C-index, cumulative incidence, treatment effect, Weibull shape *w* and drop-out rate *d*. A: θ = 0.7, B: θ = 0.4. **Figure S2.** Power curves resulting from adjustment with clinical variables only (tumor staging and ECOG score) or with the additional deep learning HCCnet covariate. Covariates are sampled from the HCC patients of the TCGA dataset. **Figure S3.** Relationships between proposed R^2 measures and the reduction of sample size provided by covariate adjustment R^2_{obs} over the grid of parameters described in Table S1. Each point on each panel corresponds to a unique combination of values for the five parameters in Table S1. Each unique combination of parameters has therefore eight corresponding R^2 measures. The notations for different measures of R^2 follow mainly [28]. R^2_{CS} is the Cox-Snell R^2_{CS} which is related to the likelihood ratio between the model of interest and a null model. ρ^2_κ is a variation of R^2_{CS}. R^2_D is a transformation of the D measure. R^2_I is a variation of R^2_D. R^2_{PM} is a measure related to the squared Pearson correlation between the logarithm of transformed survival time and the term βX. $R^2_R$is the measure proposed by Royston in the same paper of reference [28]. $R^2_{WA}$is an approximated version of a more complex measure related to the Weibull model. $R^2_{X0Q}$is a measure named after the authors of the paper: Xu, O'Quigley [30].

Acknowledgements
Not applicable.

Authors' contributions
The author(s) read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The results analyzed and steps to reproduce the datasets generated during the current study are available at https://github.com/owkin/CovadjustSim. The prognostic scores from the HCCnet deep learning algorithm applied on the TCGA HCC dataset are not publicly available.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
All authors of the article are employees of Owkin Inc.

Received: 7 February 2023 Accepted: 12 May 2023
Published online: 06 June 2023

References

1. Kahan BC, Jairath V, Doré CJ, et al. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials. 2014;15:139. https://doi.org/10.1186/1745-6215-15-139.
2. Hernández AV, Eijkemans MJC, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power? Ann Epidemiol. 2006;16:41–8. https://doi.org/10.1016/j.annepidem.2005.09.007.
3. Hernández AV, Steyerberg EW, Habbema JDF. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol. 2004;57:454–60. https://doi.org/10.1016/j.jclinepi.2003.09.014.
4. Pocock SJ, Assmann SE, Enos LE, et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med. 2002;21:2917–30. https://doi.org/10.1002/sim.1296.
5. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869. https://doi.org/10.1136/bmj.c869.
6. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Control Clin Trials. 2000;21:330–42. https://doi.org/10.1016/s0197-2456(00)00061-1.

7. CHMP. Guideline on adjustment for baseline covariates in clinical trials. EMA 2015.
8. CDER. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products. US Food Drug Adm. 2021.https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products. Accessed 15 Jun 2021.
9. Kim ES, Bruinooge SS, Roberts S, et al. Broadening eligibility criteria to make clinical trials more representative: American Society of Clinical Oncology and Friends of Cancer Research Joint Research Statement. J Clin Oncol Off J Am Soc Clin Oncol. 2017;35:3737–44. https://doi.org/10.1200/JCO.2017.73.7916.
10. Research C for DE and. Enhancing the Diversity of Clinical Trial Populations — Eligibility Criteria, Enrollment Practices, and Trial Designs Guidance for Industry. US Food Drug Adm. 2020.https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enhancing-diversity-clinical-trial-populations-eligibility-criteria-enrollment-practices-and-trial. Accessed 24 Nov 2021.
11. Food and Drug Administration. Public Workshop: EVALUATING INCLUSION AND EXCLUSION CRITERIA IN CLINICAL TRIALS. Workshop Rep 2018;:12.
12. Averitt AJ, Weng C, Ryan P, et al. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. NPJ Digit Med. 2020;3:1–10. https://doi.org/10.1038/s41746-020-0277-8.
13. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature. 2021;592:629–33. https://doi.org/10.1038/s41586-021-03430-5.
14. Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell. 2017;169:1327-1341.e23. https://doi.org/10.1016/j.cell.2017.05.046.
15. Saillard C, Schmauch B, Laifa O, et al. Predicting survival after hepatocellular carcinoma resection using deep-learning on histological slides. Hepatology Published Online First: 2020.https://aasldpubs.onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1002/hep.31207. Accessed 18 May 2020).
16. Joseph L. Fleiss. Appendix: Sample-Size Determination. In: The Design and Analysis of Clinical Experiments. John Wiley & Sons, Ltd 1999. 369–417. https://doi.org/10.1002/9781118032923.app1
17. Choodari-Oskooei B, Royston P, Parmar MKB. A simulation study of predictive ability measures in a survival model II: explained randomness and predictive accuracy. Stat Med. 2012;31:2644–59. https://doi.org/10.1002/sim.5460.
18. Choodari-Oskooei B, Royston P, Parmar MKB. A simulation study of predictive ability measures in a survival model I: Explained variation measures. Stat Med. 2012;31:2627–43. https://doi.org/10.1002/sim.4242.
19. Schoenfeld DA. Sample-Size Formula for the Proportional-Hazards Regression Model. Biometrics. 1983;39:499. https://doi.org/10.2307/2531021.
20. Josse J, Husson F. missMDA: a package for handling missing values in multivariate data analysis. J Stat Softw 2016;70. https://doi.org/10.18637/jss.v070.i01
21. Bristol-Myers Squibb. A phase 3, randomized, double-blind study of adjuvant nivolumab versus placebo for participants with hepatocellular carcinoma who are at high risk of recurrence after curative hepatic resection or ablation. clinicaltrials.gov 2021. https://clinicaltrials.gov/ct2/show/NCT03383458. Accessed 23 Nov 2021.
22. Hoffmann-La Roche. A phase III, multicenter, randomized, open-label study of atezolizumab (Anti-PD-L1 Antibody) plus bevacizumab versus active surveillance as adjuvant therapy in patients with hepatocellular carcinoma at high risk of recurrence after surgical resection or ablation. clinicaltrials.gov 2021. https://clinicaltrials.gov/ct2/show/NCT04102098. (Accessed 23 Nov 2021).
23. Merck Sharp & Dohme Corp. A phase 3 double-blinded, two-arm study to evaluate the safety and efficacy of pembrolizumab (MK-3475) versus placebo as adjuvant therapy in participants with hepatocellular carcinoma and complete radiological response after surgical resection or local ablation (KEYNOTE-937). clinicaltrials.gov 2019. https://clinicaltrials.gov/ct2/show/NCT03867084. (Accessed 17 Aug 2020).
24. AstraZeneca. A phase III, randomized, double-blind, placebo-controlled, multi center study of durvalumab monotherapy or in combination with bevacizumab as adjuvant therapy in patients with hepatocellular carcinoma who are at high risk of recurrence after curative hepatic resection or ablation. clinicaltrials.gov 2021. https://clinicaltrials.gov/ct2/show/NCT03847428. Accessed 23 Nov 2021.
25. Bruix J, Takayama T, Mazzaferro V, et al. Adjuvant sorafenib for hepatocellular carcinoma after resection or ablation (STORM): a phase 3, randomised, double-blind, placebo-controlled trial. Lancet Oncol. 2015;16:1344–54. https://doi.org/10.1016/S1470-2045(15)00198-9.
26. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med. 2004;23:723–48. https://doi.org/10.1002/sim.1621.
27. Kent JT, O'Quigley J. Measures of dependence for censored survival data. Biometrika. 1988;75:525–34. https://doi.org/10.1093/biomet/75.3.525.
28. Royston P. Explained Variation for Survival Models. Stata J Promot Commun Stat Stata. 2006;6:83–96. https://doi.org/10.1177/1536867X0600600105.
29. O'Quigley J, Xu R, Stare J. Explained randomness in proportional hazards models. Stat Med. 2005;24:479–89. https://doi.org/10.1002/sim.1946.
30. Ronghui Xu, O'Quigley J. A. R. ^{2} type measure of dependence for proportional hazards models. J Nonparametric Stat 1999;12:83–107. https://doi.org/10.1080/10485259908832799
31. Cox DR, Snell EJ, Cox DR, et al. Analysis of binary data. 2. ed., 1. CRC Press reprint. Boca Raton, Fla.: : Chapman & Hall [u.a.] 1989.
32. Breast International Group (BIG) 1-98 Collaborative Group, Thürlimann B, Keshaviah A, et al. A comparison of letrozole and tamoxifen in postmenopausal women with early breast cancer. N Engl J Med. 2005;353:2747–57. https://doi.org/10.1056/NEJMoa052258.
33. Davis ID, Martin AJ, Stockler MR, et al. Enzalutamide with standard first-line therapy in metastatic prostate cancer. N Engl J Med. 2019;381:121–31. https://doi.org/10.1056/NEJMoa1903835.
34. Reck M, Rodríguez-Abreu D, Robinson AG, et al. Pembrolizumab versus chemotherapy for PD-L1–positive non–small-cell lung cancer. N Engl J Med. 2016;375:1823–33. https://doi.org/10.1056/NEJMoa1606774.
35. Im S-A, Lu Y-S, Bardia A, et al. Overall survival with ribociclib plus endocrine therapy in breast cancer. N Engl J Med. 2019;381:307–16. https://doi.org/10.1056/NEJMoa1903765.
36. Conroy T, Hammel P, Hebbar M, et al. FOLFIRINOX or gemcitabine as adjuvant therapy for pancreatic cancer. N Engl J Med. 2018;379:2395–406. https://doi.org/10.1056/NEJMoa1809775.
37. Yau T, Park J-W, Finn RS, et al. Nivolumab versus sorafenib in advanced hepatocellular carcinoma (CheckMate 459): a randomised, multicentre, open-label, phase 3 trial. Lancet Oncol. 2022;23:77–90. https://doi.org/10.1016/S1470-2045(21)00604-5.
38. Baas, Scherpereel A, Nowak AK, et al. First-line nivolumab plus ipilimumab in unresectable malignant pleural mesothelioma (CheckMate 743): a multicentre, randomised, open-label, phase 3 trial. Lancet. 2021;397:375–86. https://doi.org/10.1016/S0140-6736(20)32714-8.
39. Conroy T, Desseigne F, Ychou M, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. N Engl J Med. 2011;364:1817–25. https://doi.org/10.1056/NEJMoa1011923.
40. Kerr KF, Roth J, Zhu K, et al. Evaluating biomarkers for prognostic enrichment of clinical trials. Clin Trials. 2017;14:629–38. https://doi.org/10.1177/1740774517723588.
41. Friede T, Kieser M. Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. Pharm Stat. 2011;10:8–13. https://doi.org/10.1002/pst.398.
42. Daniel R, Zhang J, Farewell D. Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biom J. 2020;63:528–57. https://doi.org/10.1002/bimj.201900297.
43. Permutt T. Do covariates change the estimand? Stat Biopharm Res. 2020;12:45–53. https://doi.org/10.1080/19466315.2019.1647874.
44. Díaz I, Colantuoni E, Hanley DF, et al. Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. Lifetime Data Anal. 2019;25:439–68. https://doi.org/10.1007/s10985-018-9428-5.
45. Zhang M. Robust methods to improve efficiency and reduce bias in estimating survival curves in randomized clinical trials. Lifetime Data Anal. 2015;21:119–37. https://doi.org/10.1007/s10985-014-9291-y.
46. Parast L, Tian L, Cai T. Landmark estimation of survival and treatment effect in a randomized clinical trial. J Am Stat Assoc. 2014;109:384–94. https://doi.org/10.1080/01621459.2013.842488.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.