Leveraging baseline covariate adjustment methods in oncology time-to-event studies

Daniel Backenroth
November 3, 2025
Savannah, Georgia
BASS
Draft

Daniel Backenroth

Fellow, Biostatistics at Johnson & Johnson

Based in Raritan NJ

6 years at J&J

Previous work experience at Flatiron Health

Agenda

1. Why covariate adjustment?
2. FDA/EMA guidance
3. Conditional vs. unconditional treatment effects
4. Covariate-adjusted log-rank test
5. Covariate-adjusted hazard ratio estimator
6. How to decide what to adjust for
7. Cross-fitting when adjusting for many variables
8. Adaptive designs
9. Group sequential design considerations

Why covariate adjustment?

Covariate adjustment can increase power of a study to detect a treatment effect often at **little or no cost**

Additional power can be used to:

- Increase chance of study success
- Reduce time to analysis
- Reduce size of trial

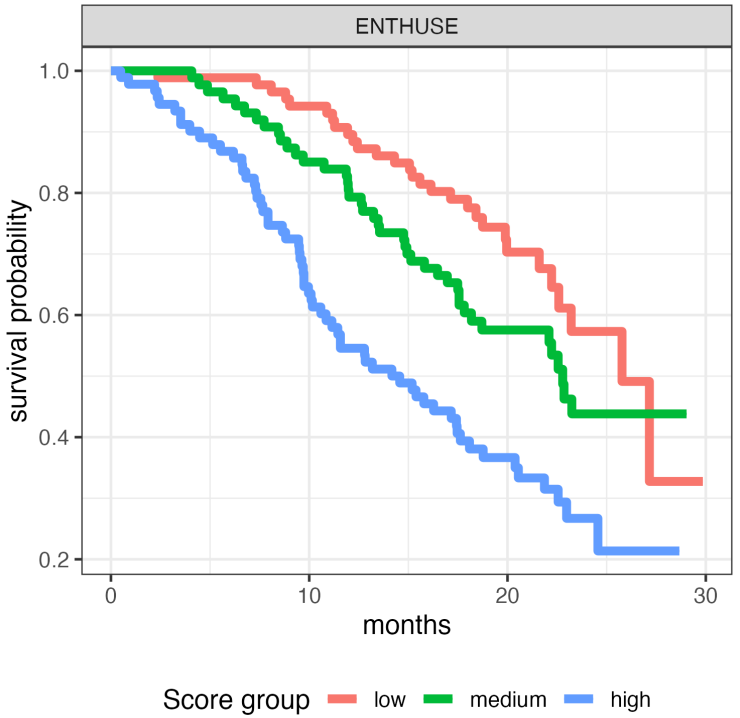
Why covariate adjustment?

Covariate adjustment can increase power of a study to detect a treatment effect, often at **little or no cost**

Additional power can be used to:

- Increase chance of study success
- Reduce time to stopping
- Reduce size of trial

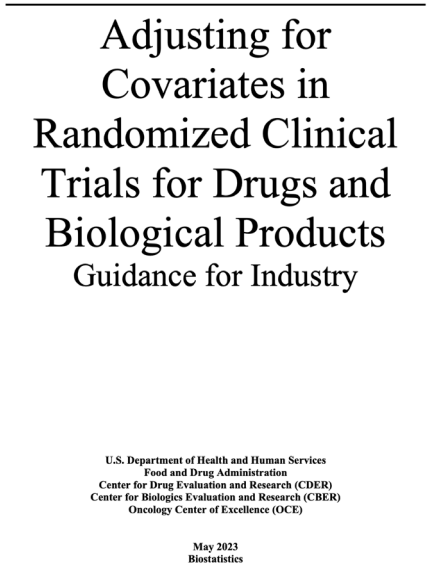
Adjustment for Halabi score in OS analysis in mCRPC



Assuming variance reduction of 13%:

- *Keeping n (# patients) fixed:*
  - study powered at 90% for HR=0.7 can have ~93% power
- *Keeping n and power fixed:*
  - average time to stopping can be reduced by 4 months (~10%)
- *Reducing n by 5%:*
  - can maintain power at 90% **and** reduce average time by 2 months

Regulatory guidance FDA/EMA


Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products

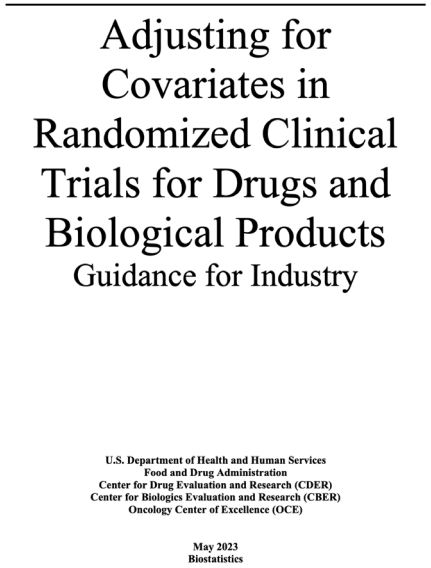
Guidance for Industry

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Oncology Center of Excellence (OCE)

May 2023
Biostatistics

Incorporating prognostic baseline covariates in the design and analysis of clinical trial data can result in a more efficient use of data to demonstrate and quantify the effects of treatment. Moreover, this can be done with minimal impact on bias or the Type I error rate —from FDA guidance

Regulatory guidance FDA/EMA



Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products

Guidance for Industry

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Oncology Center of Excellence (OCE)

May 2023
Biostatistics

Incorporating prognostic baseline covariates in the design and analysis of clinical trial data can result in a more efficient use of data to demonstrate and quantify the effects of treatment. Moreover, this can be done with minimal impact on bias or the Type I error rate.

	FDA	EMA
When pre-specification required:	Before unblinding of comparative data	Before first patient in
Which covariates to adjust for:	Anything you want	Only a few most clinically/prognostically important variables
When to consult with agency:	Use in adaptive design/ *estimating a conditional effect* /using novel methodology/ *adjusting for many covariates*/data-adaptive covariate selection/ *complex covariate-adaptive randomization*	Unknown

Conditional vs. unconditional treatment effects

- When adjusting only for treatment, the Cox PH estimates a **marginal hazard ratio** , which is the average hazard ratio for all participants in the trial

When adjusting for covariates or stratifying, the Cox PH model

- Estimates a **conditional hazard ratio** , a treatment effect that is conditional on the values of the stratification variables
- Even if treatment effect is the same for all individuals, this conditional treatment effect will be different (further from 1) than the marginal treatment effect
- This is called non-collapsibility
- Conditional HR only equals marginal HR when both equal 1

More on conditional vs. unconditional treatment effects

Assume two strata, high-risk and low-risk

Within each stratum treatment indicator $I_i \sim Bern(0.5)$

In high-risk stratum survival times $T_i \sim Exp(\mathbf{0.03} * 0.5^{I_i})$

In low-risk stratum survival times $T_i \sim Exp(\mathbf{0.01} * 0.5^{I_i})$

Conditional hazard ratio is 0.5 for all patients

Conditional hazard ratio is *further from 1* than marginal hazard ratio

For this data-generating process, **marginal HR is approximately 0.55**

More on conditional vs. unconditional treatment effects

Assume two strata, high-risk and low-risk

Within each stratum treatment indicator $I_i \sim Bern(0.5)$

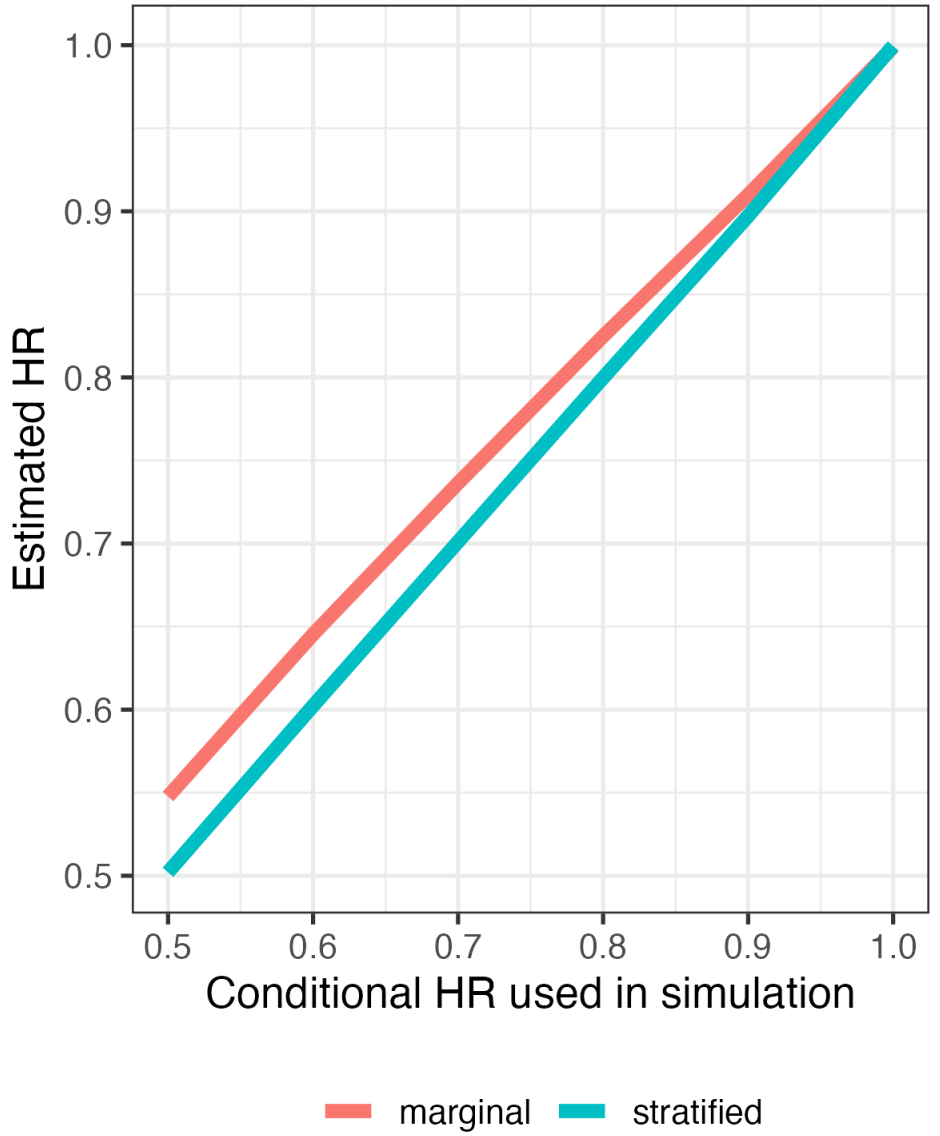
In high-risk stratum survival times $T_i \sim rexp(0.03 * 0.5^{I_i})$

In low-risk stratum survival times $T_i \sim rexp(0.01 * 0.5^{I_i})$

Conditional hazard ratio is 0.5 for all patients

Conditional hazard ratio is further from 1 than marginal hazard ratio

For this data-generating process, **marginal HR is approximately 0.55**



FDA guidance is more positive about unconditional treatment effect estimation

As part of the pre-specification of the estimand of interest, **sponsors should specify whether the treatment effect of interest in an analysis is a conditional or unconditional treatment effect** .

FDA guidance is more positive about unconditional treatment effect estimation

As part of the pre-specification of the estimand of interest, sponsors should specify whether the treatment effect of interest in an analysis is a conditional or unconditional treatment effect .

Conditional treatment effects

Sponsors **should discuss** with the relevant review divisions specific proposals …containing nonlinear regression to estimate **conditional treatment effects** for the primary analysis…results can be difficult to interpret if the model is misspecified and treatment effects substantially differ across subgroups.

Unconditional treatment effects

Sponsors **can** perform covariate-adjusted estimation and inference for an **unconditional treatment effect** …in the primary analysis of data from a randomized trial. The method used should provide valid inference under approximately the same minimal statistical assumptions that would be needed for unadjusted estimation in a randomized trial.

What are some reasons to prefer unconditional rather than conditional models?

- Resulting HR doesn't depend on what variables are adjusted for
  - Although in a stratified model HR does depend on stratification variables
- There are simulation studies that show potential for Type I error rate inflation when using a Cox PH model if adjustment covariates don't satisfy PH (Jiang et al., Stat Med 2008)

Covariate-adjusted log-rank test

- Proposed by Ye et al. (Biometrika 2024)
- Like test of Lu and Tsiatis (Biometrika 2008) that is specifically cited in FDA guidance, but provides guaranteed asymptotic efficiency gains under covariate - adaptive randomization (e.g., stratified permuted block randomization)
- Ye et al. also provide **a marginal** covariate-adjusted HR estimator under the Cox PH assumption
- Stratified versions of the test and estimator are also available

Biometrika (2024), **111**, 2, *pp.* 691–705 https://doi.org/10.1093/biomet/asad045
Advance Access publication 27 July 2023

Covariate-adjusted log-rank test: guaranteed efficiency gain and universal applicability

BY TING YE

Department of Biostatistics, University of Washington, Box 351617, Seattle, Washington 98195, U.S.A.
tingye1@uw.edu

JUN SHAO

Department of Statistics, University of Wisconsin, 1300 University Avenue, Madison, Wisconsin 53706, U.S.A.
shao@stat.wisc.edu

AND YANYAO YI

Global Statistical Sciences, Eli Lilly and Company, 839 S Delaware St., Indianapolis, Indiana 46285, U.S.A.
yi_yanyao@lilly.com

SUMMARY

Nonparametric covariate adjustment is considered for log-rank-type tests of the treatment effect with right-censored time-to-event data from clinical trials applying covariate-adaptive randomization. Our proposed covariate-adjusted log-rank test has a simple explicit formula and a guaranteed efficiency gain over the unadjusted test. We also show that our proposed test achieves universal applicability in the sense that the same formula of test can be universally applied to simple randomization and all commonly used covariate-adaptive randomization schemes such as the stratified permuted block and the Pocock–Simon minimization, which is not a property enjoyed by the unadjusted log-rank test. Our method is supported by novel asymptotic theory and empirical results for Type-I error and power of tests.

Some key words: Covariate calibration; Minimization; Permuted block; Pitman's relative efficiency; Stratification; Time-to-event data; Validity and power of tests.

1. INTRODUCTION

In clinical trials, adjusting for baseline covariates has been widely advocated as a way to improve efficiency for demonstrating treatment effects 'under approximately the same minimal statistical assumptions that would be needed for unadjusted estimation' (ICH E9, 1998; EMA, 2015; FDA, 2023). In testing for an effect between two treatments with right-censored time-to-event outcomes, adjusting for covariates using the Cox proportional hazards model

Construction of covariate-adjusted log-rank test

1. Write log-rank test numerator as sum of patient level contributions O_{ij}: $U_L = \frac{1}{n}\sum_{i=1}^{n}[I_i O_{i1} - (1 - I_i)O_{i0}]$
   - I_i is the treatment indicator for patient i
   - O_{ij} is derived outcome, contribution of patient i on treatment j to score function

Construction of covariate-adjusted log-rank test

1. Write log-rank test numerator as sum of patient level contributions O_{ij}: $U_L = \frac{1}{n}\sum_{i=1}^{n}[I_i O_{i1} - (1 - I_i)O_{i0}]$
   - I_i is the treatment indicator for patient i
   - O_{ij} is derived outcome, contribution of patient i on treatment j to score function
2. Regress the O_{ij} on baseline covariates X_i separately in each treatment arm to get coefficients β_j

Construction of covariate-adjusted log-rank test

1. Write log-rank test numerator as sum of patient level contributions O_{ij}: $U_L = \frac{1}{n}\sum_{i=1}^{n}[I_i O_{i1} - (1 - I_i)O_{i0}]$
   - I_i is the treatment indicator for patient i
   - O_{ij} is derived outcome, contribution of patient i on treatment j to score function
2. Regress the O_{ij} on baseline covariates X_i, separately in each treatment arm, to get coefficients β_j
3. Add contrast across treatment arms of predictions from these regression models to get adjusted numerator:
   $U_{CL} = U_L - \frac{1}{n}\sum_{i=1}^{n}[I_i(X_i - \bar{X})^T\beta_1 - (1 - I_i)(X_i - \bar{X})^T\beta_0]$

Why does this work?

$$U_{CL} = U_L - \frac{1}{n}\sum_{i=1}^{n}[\overbrace{I_i(X_i - \bar{X})^T\beta_1}^{\text{Augmentation tx arm}} - \overbrace{(1 - I_i)(X_i - \bar{X})^T\beta_0}^{\text{Augmentation ctrl arm}}]$$

- Because of randomization, each of the "augmentation terms" has mean 0

Why does this work?

$$U_{CL} = U_L - \frac{1}{n}\sum_{i=1}^{n}[\overbrace{I_i(X_i - \bar{X})^T\beta_1}^{\text{Augmentation tx arm}} - \overbrace{(1-I_i)(X_i - \bar{X})^T\beta_0}^{\text{Augmentation ctrl arm}}]$$

- Because of randomization, each of the "augmentation terms" has mean 0
- The β_j minimize the variance of U_{CL}. As long as the O_{ij} are correlated with the baseline covariates X_i, variance σ^2_{CL} will be reduced relative to σ^2_L:
  - $\sigma^2_{CL} = \sigma^2_L - \pi(1-\pi)(\beta_1 + \beta_0)^T\Sigma_X(\beta_1 + \beta_0)$
  - π: randomization probability to treatment 1

Covariate-adjusted stratified log-rank test

For guaranteed efficiency gain, we need to include in the set of covariates one indicator for each randomization stratum, and we need to fit the regression models separately by arm

Covariate-adjusted stratified log-rank test

For guaranteed efficiency gain, we need to include in the set of covariates one indicator for each randomization stratum, and we need to fit the regression models separately by arm

Example:

A6181120 trial in mCRPC (Michaelson et al., JCO 2014)

4 stratification factors

- ECOG PS
- docetaxel-resistant vs. intolerant
- type of disease progression at entry
- previous therapy with a VEGF pathway inhibitor

In each arm we have 15 stratum indicators (2^4-1), so if we adjust for one variable we have 16 covariates in each arm and 32 total!

Lesson: adjustment requires not too many strata (or drop some stratification variables from adjustment)

Covariate-adjusted hazard ratio estimator

Ye et al. also provide an estimator of marginal HR under a proportional hazards assumption, either stratified or unstratified

$$\lambda_1(t) = \lambda_0(t)e^{\theta}$$

$$\lambda_{1z}(t) = \lambda_{0z}(t)e^{\theta} \quad \text{for every } z$$

In regulatory submissions may be asked to show this estimator is unbiased and that CIs constructed using standard error estimator have the right coverage

How to decide what to adjust for

- Easiest: use existing ”prognostic score”—weighted sum of baseline characteristics

- Numerous prognostic scores have been developed in various oncology indications
  - Halabi score for mCRPC
  - Bachet score for mCRC
  - ISS stage for multiple myeloma
  - IPI and R/R IPI scores for DLBCL
- And cardiovascular indications
  - The PREVENT calculator

How to decide what to adjust for

- Easiest: use existing ”prognostic score”-weighted sum of baseline characteristics

- Numerous works have developed prognostic scores in various oncology indications
  - Halabi score for mCRPC
  - Bachet score for mCRC
  - ISS stage for multiple myeloma
  - IPI and R/R IPI scores for DLBCL
- And cardiovascular indications
  - The PREVENT calculator

- Common variables in the oncology scores:
  - **Lab variables**
    - e.g., PSA, ALB, HGB
  - Locations, numbers of metastatic sites
  - Age
  - **ECOG PS**
  - Chromosomal abnormalities
  - Disease-specific biomarkers (PSA)

Prognostic score examples

Bachet score in mCRC

Table S.5. Common multivariate model and weighted prognostic factors with Hazard Ratios (HR) and Confidence Intervals (CI) in first-line of construction dataset

		Common multivariate model			The weighted prognostic factors
Variable		**HR**	**95%CI**	***P*-value**	
ECOG PS	0	Reference			1
	1	1.33	[1.28-1.37]	< .001	1.33
	≥ 2	1.77	[1.63-1.92]	< .001	1.77
HGB	< 12	Reference			1
	≥ 12	0.85	[0.82-0.88]	< .001	0.85
PLT	< 400	Reference			1
	≥ 400	1.15	[1.11-1.20]	< .001	1.15
WBC/ANC	< 1.45	Reference			1
	≥ 1.45	0.77	[0.75-0.80]	< .001	0.77
LDH	< 1 UNL	Reference			1
	≥ 1 UNL	1.29	[1.24-1.34]	< .001	1.29
ALP	< 1 UNL	Reference			1
	1 ≤ < 3 UNL	1.19	[1.15-1.24]	< .001	1.19
	≥ 3 UNL	1.59	[1.50-1.69]	< .001	1.59
N of metastasic site	0-1	Reference			1
	≥ 2	1.36	[1.32-1.41]	< .001	1.36

Halabi score inmCRPC

Table 2. Multivariable Model Predicting Overall Survival Using Cancer and Leukemia Group B–90401 Training Set

Factor	Hazard Ratio	95% CI
Opioid analgesic use (yes *v* no)	1.09	1.00 to 1.30
LDH > 1 ULN (yes *v* no)	1.40	1.16 to 1.65
Disease site		
Bone/bone + LN *v* LN	1.06	1.00 to 1.36
Visceral *v* bone/bone + LN	1.27	0.96 to 1.51
Visceral *v* LN	1.34	1.00 to 1.76
ECOG PS		
1 *v* 0 (or 2 *v* 1)	1.36	1.15 to 1.58
2 *v* 0	1.84	1.33 to 2.49
Albumin	0.89	0.77 to 1.00
Hemoglobin	0.94	0.88 to 1.00
PSA	1.02	1.00 to 1.06
Alkaline phosphatase	1.16	1.00 to 1.30

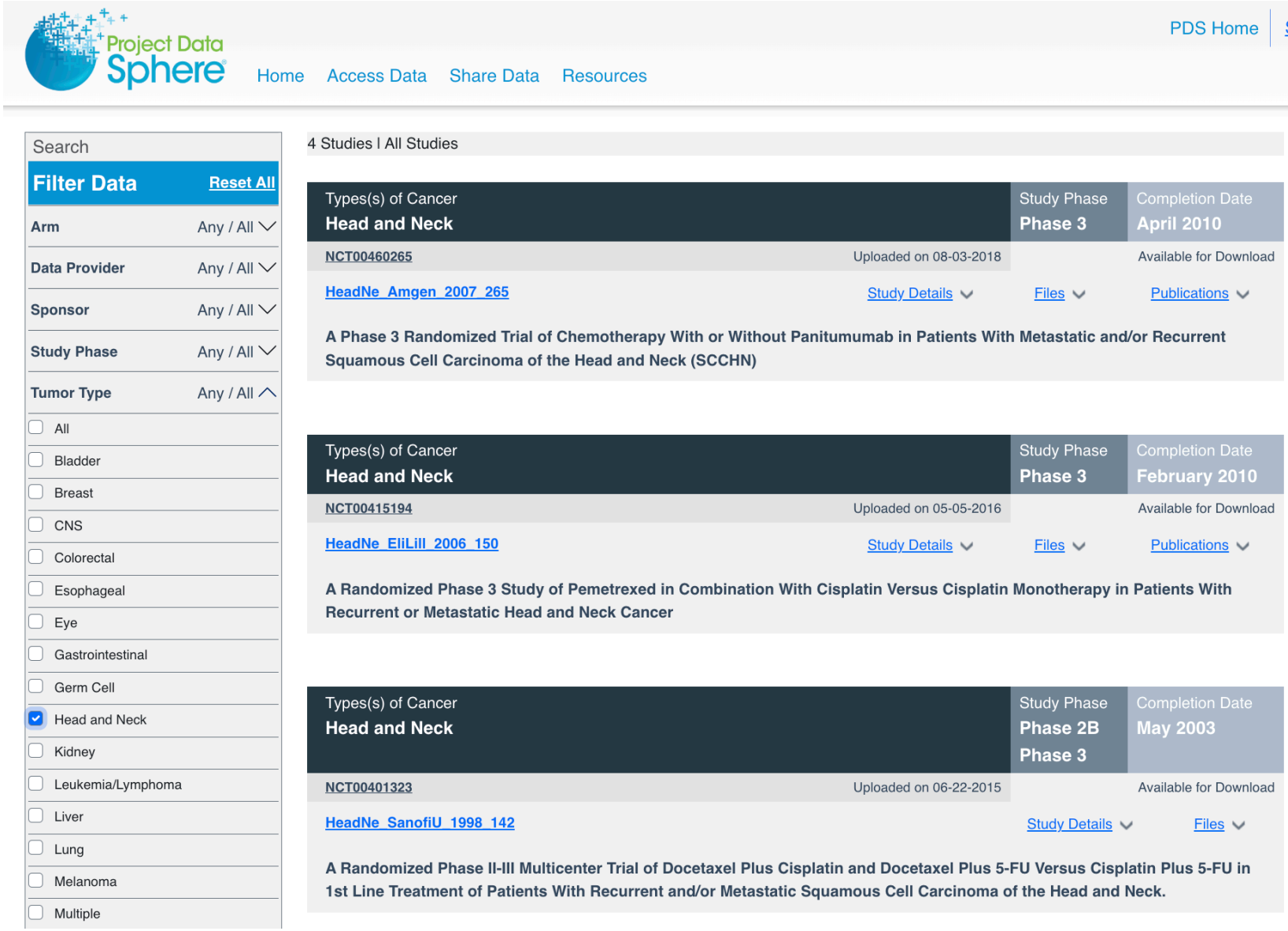
Abbreviations: ECOG PS, Eastern Cooperative Oncology Group performance status; LDH, lactate dehydrogenase; LN, lymph node; PSA, prostate-specific antigen; ULN, upper limit of normal.

What if a score doesn’t exist

- If you have data, it’s not difficult to make your own score

What if a score doesn't exist

- If you have data, it's not difficult to make your own score
- If you don't have data, look for some on Project Data Sphere



Making a score

1. Extract relevant baseline data (labs, ECOG PS, etc…) and outcomes from trial datasets
2. Choose some data for model building and some (ideally, a separate trial dataset) for validation
3. Fit a Cox PH model with a LASSO penalty to select variables and determine their coefficients in the linear combination

Making a score example (head & neck cancer): Dataset selection

Table 1: Studies Used to Estimate a Prognostic Score for First Line Treatment of R/M HNSCC

Study; Year of protocol	Treatments	Year of protocol	# of events/# of patients per arm available in Project Data Sphere	Use in modeling
NCT00460265; 2006	Cisplatin and 5-FU with or without panitumumab	2006	Control: 204/260 Treatment: 221/260	Prognostic score creation and assessment
NCT00401323; 1997	Docetaxel+cisplatin vs. docetaxel+5-FU vs. cisplatin+5-FU	1997	Control (Cisplatin+5-FU): 243/282	Prognostic score creation and assessment
NCT00415194;	Pemetrexed+cisplatin vs. placebo+cisplatin	2006	Control: 320/397	Assessment only

Table 3: Completeness Rates of Each Variable in Table 2

	NCT00460265	NCT00401323	NCT00415194
n	514	279	343
Age = not missing (%)	100.0	100.0	100.0
Sex = not missing (%)	100.0	100.0	100.0
Site = not missing (%)	100.0	100.0	98.5
ECOGBL = not missing (%)	100.0	100.0	100.0
Surgery = not missing (%)	100.0	100.0	0.0
Radiotherapy = not missing (%)	100.0	100.0	0.0
TNMstage = not missing (%)	99.4	88.5	0.0
BMI = not missing (%)	100.0	100.0	99.4
ALB = not missing (%)	96.7	76.3	0.0
ALP = not missing (%)	99.4	97.8	95.9
HGB = not missing (%)	100.0	99.3	99.4
LDH = not missing (%)	98.1	75.3	0.0
WBC = not missing (%)	100.0	97.1	99.4
PLT = not missing (%)	100.0	97.8	99.4

Making a score example (head & neck cancer): score construction and validation

Table 4: Prognostic score model coefficients

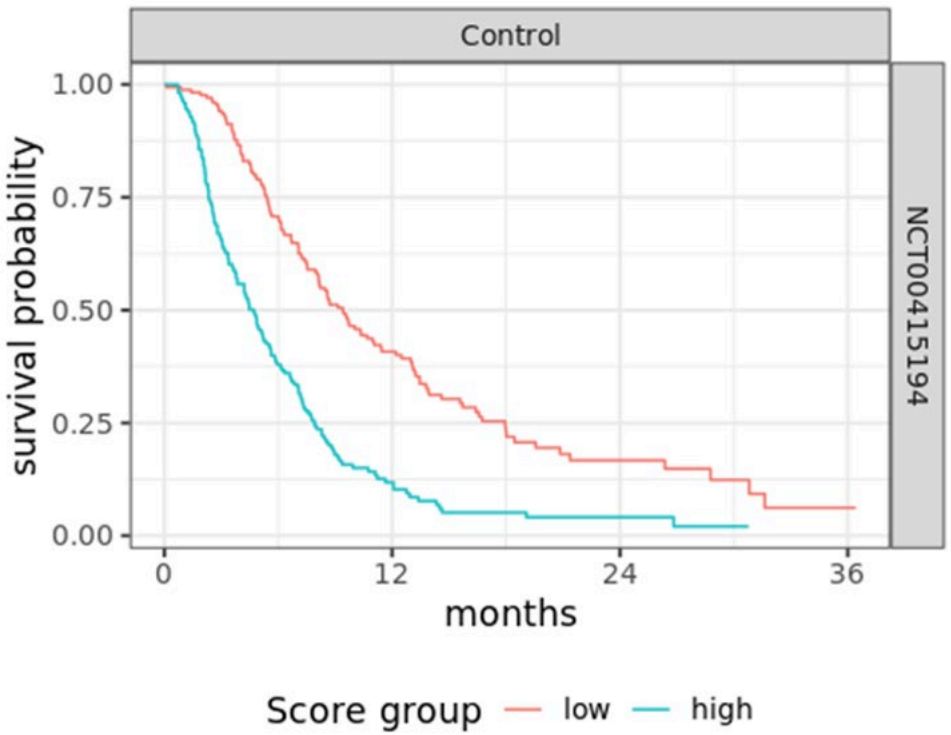
covariate	coefficient (log hazard ratio)
ECOG PS	0.2699
BMI	-0.03513
log(White blood cell count, 10^9/L)	0.5640
Hemoglobin, g/dL	-0.1151
Platelet count, 10^9/L	0.001207
Prior radiotherapy	0.5283

Making a score example (head & neck cancer): score construction and validation

Table 4: Prognostic score model coefficients

covariate	coefficient (log hazard ratio)
ECOG PS	0.2699
BMI	-0.03513
log(White blood cell count, 10^9/L)	0.5640
Hemoglobin, g/dL	-0.1151
Platelet count, 10^9/L	0.001207
Prior radiotherapy	0.5283

Figure 2: Overall Survival in NCT00415194 Stratified by High/Low R/M HNSCC score.



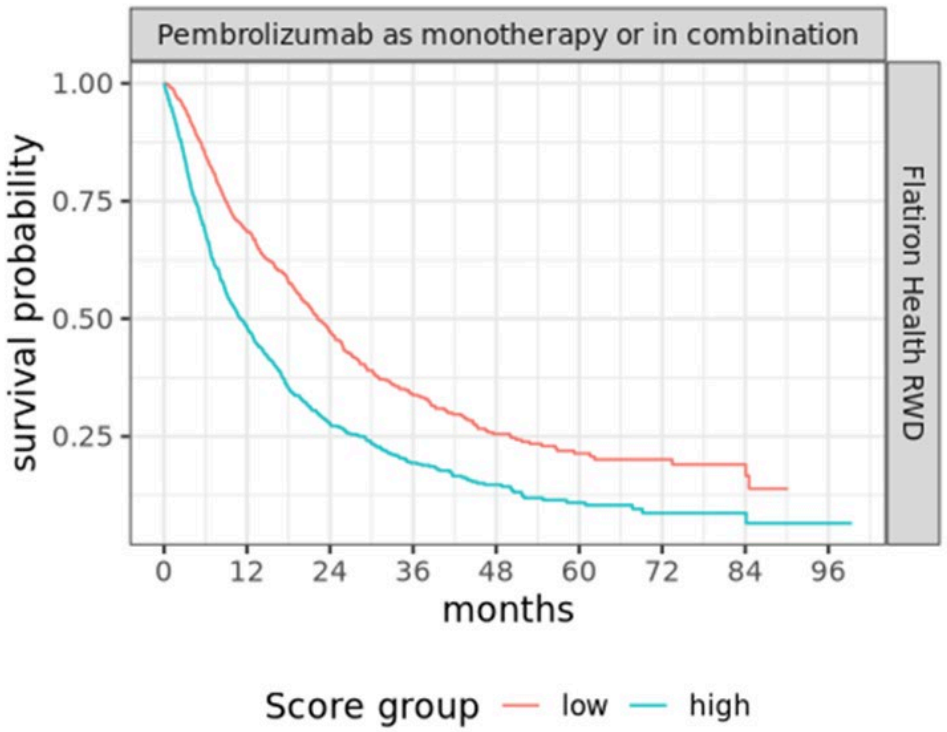
Additional validation in RWE

The Project Data Sphere datasets are very old

- None have data on patients treated on current SoC

Can use RWE to confirm that score still has prognostic validity

Figure 3 Overall Survival in Flatiron Health RWD Stratified by High/Low R/M HNSCC score



This study used the US-based, electronic health record-derived deidentified Flatiron Health Research Database[1].

Flatiron Health. Database Characterization Guide. Flatiron.com. Published March 18, 2025. Accessed September 5, 2025. https://flatiron.com/database -characterization

Assessing variance reductions using bootstrap (Li et al., JRSSSB 2023)

Repeat 1,000 times:

1. Bootstrap dataset
2. Randomly assign treatment, as closely as possible following actual randomization method from trial, e.g., stratified permuted block randomization
3. Calculate standard error for unadjusted and adjusted analysis (se_{unadj} and se_{adj})
4. Variance reduction is defined as $VR = 1 - \frac{se_{adj}^2}{se_{unadj}^2}$
5. We can report the median VR and its CI (using percentile bootstrap)

Assessing variance reductions using bootstrap

Repeat 1,000 times:

1. Bootstrap dataset
2. Randomly assign treatment, as closely as possible following actual randomization method from trial, e.g., stratified permuted block randomization
3. Calculate standard error for unadjusted and adjusted analysis (se_{unadj} and se_{adj})
4. Variance reduction is defined as $VR = 1 - \frac{se_{adj}^2}{se_{unadj}^2}$
5. We can report the median VR and its CI (using percentile bootstrap)

Results in test dataset:

VR=13% (95% CI: 7.4%19.4%)

Assessing variance reductions using bootstrap

Repeat 1,000 times:

1. Bootstrap dataset
2. Randomly assign treatment, as closely as possible following actual randomization method from trial, e.g., stratified permuted block randomization
3. Calculate standard error for unadjusted and adjusted analysis (se_{unadj} and se_{adj})
4. Variance reduction is defined as $VR = 1 - \frac{se_{adj}^2}{se_{unadj}^2}$
5. We can report the median VR and its CI (using percentile bootstrap)

Results in test dataset:

VR=13% (95% CI: 7.4%-19.4%)

If variance reduction is X%:

- Study with (100-X)% of the patients and adjustment has same power as
- Study with 100% of the patients and no adjustment

So here sample size could be reduced by up to 13% (with caveats to be discussed later)

Assessing variance reductions using bootstrap

Repeat 1,000 times:

1. Bootstrap dataset
2. Randomly assign treatment, as closely as possible following actual randomization method from trial, e.g., stratified permuted block randomization
3. Calculate standard error for unadjusted and adjusted analysis (se_{unadj} and se_{adj})
4. Variance reduction is defined as $VR = 1 - \frac{se_{adj}^2}{se_{unadj}^2}$
5. We can report the median VR and its CI (using percentile bootstrap)

Results in test dataset:

VR=13% (95% CI: 7.4%-19.4%)

If variance reduction is X%:

- Study with (100-X)% of the patients and adjustment has same power as
- Study with 100% of the patients and no adjustment

So here sample size could be reduced by up to 13% (with caveats to be discussed later)

Different methods can be used to assess variance reductions also in the training data, without being too "optimistic", but those check only "internal" validity

Some more regulatory considerations + potential concerns

- Regulators may be concerned about potential mismatch between standard analysis and covariate-adjusted analysis
  - Should be a topic for investigation: why was there mismatch?
  - **But requiring significance with both standard and covariate -adjusted analysis takes away all the efficiency benefits**
- Regulators may be concerned about potential misalignment between test and estimator
  - May not be a bigger problem with adjusted than with unadjusted test—can use simulation to investigate, depends on presence of NPH
- Regulators may be interested in interaction of treatment with prognostic variables/score
  - Consider using Cox PH models with splines to model HR as smooth function of prognostic score

What if we adjust for many variables?

- When adjusting for many variables:
  - Greater power gains are possible

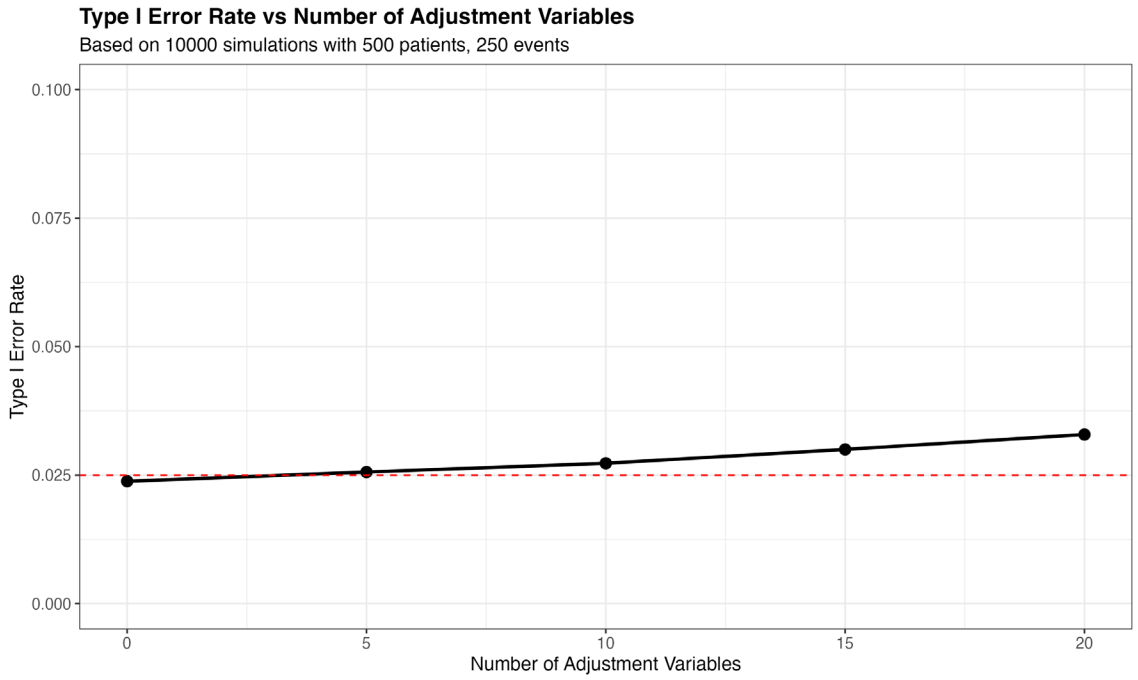
Project Data Sphere study	Variance reduction from adjustment with score	Variance reduction from adjustment with variables in score
A6181120	28.8%	34.9%
ENTHUSE	13.5%	23.5%
ENTHUSE M1C	21.3%	26.0%
MAINSAIL	7.0%	12.3%

What if we adjust for many variables?

- When adjusting for many variables:
  - Greater power gains are possible

Project Data Sphere study	Variance reduction from adjustment with score	Variance reduction from adjustment with variables in score
A6181120	28.8%	34.9%
ENTHUSE	13.5%	23.5%
ENTHUSE M1C	21.3%	26.0%
MAINSAIL	7.0%	12.3%

- But Type I error rate goes up



How cross-fitting works

With too many variables

- Coefficients aren't good estimates of the true coefficients anymore, may be unstable, numerator and denominator of adjusted log-rank test statistic won't be estimated properly

Change to diagram

To prevent this, we can use crossfitting

- Divide up the data into 10 pieces
- For each piece, construct the log-rank test statistic using regression coefficients estimated using the other 9 pieces
- Then construct the overall log-rank test statistic using the estimates from the 10 pieces
  - Add up numerators
  - Add up denominators
  - Divide

Cross-fitting

Used simulation study based on real Project Data Sphere dataset (A6181120) to illustrate cross-fitting performance

We simulate 4 strata, 2 binary adjustment variables and 4 continuous adjustment variables

Using the stratified log -rank test, this means **in each arm** we have $3+2+4=9$ covariates

Type I error rates and power for different test strategies

scenario	# events	Stratified log-rank test	Covariate-adjusted stratified log-rank test without cross-fitting	Covariate-adjusted stratified log-rank test with cross-fitting
Null	150	0.0262 (0.0248-0.0276)	0.0305 (0.0290-0.0321)	0.0256 (0.0243-0.0271)
	440	0.0257 (0.0244-0.0272)	0.0263 (0.0250-0.0278)	0.0248 (0.0234-0.0262)
Alternative	440	0.7420 (0.7333-0.7505)	0.8765 (0.8699-0.8829)	0.8704 (0.8636-0.8769)

Type I error rate inflation is effectively mitigated by cross-fitting, at both sample sizes

Lesson: Simulation required if many variables are adjusted for to make sure operating characteristics of test are good

Need to account for prevalence of strata

How to incorporate covariate adjustment into study design

- We can keep sample size and target number events the same, use covariate adjustment to increase power
  - To quantify power, use "effective" number of events: divide by 1– variance reduction in Schoenfeld's formula

How to incorporate covariate adjustment into study design

- We can keep sample size and target number events the same, use covariate adjustment to increase power
  - To quantify power, use "effective" number of events - divide by 1– variance reduction -in Schoenfeld's formula

Example

We assume HR=0.7 and want 90% power, need 331 events by Schoenfeld's formula

We assume variance reduction of 13%

288 events would give 90% power with covariate adjustment.

If we stick with 331 events, effective number of events is 331/(1– 0.13)~380, which gives us 93.5% power

Adaptive designs

- Alternatively, we can use information monitoring to trigger analyses

- We approximate information accumulated with expression $I \approx \frac{1}{\mathrm{se}(\hat{\delta})^2}$.

- We calculate required information using expression $I = \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta_a}\right]^2$.

  - δ_a is assumed log HR under alternative hypothesis (assumes proportional hazards)
  - $\mathrm{se}(\hat{\delta})$ is standard error of estimate of log HR
    - Without covariate adjustment, can be approximated with $\sqrt{\frac{d}{4}}$, d is # events

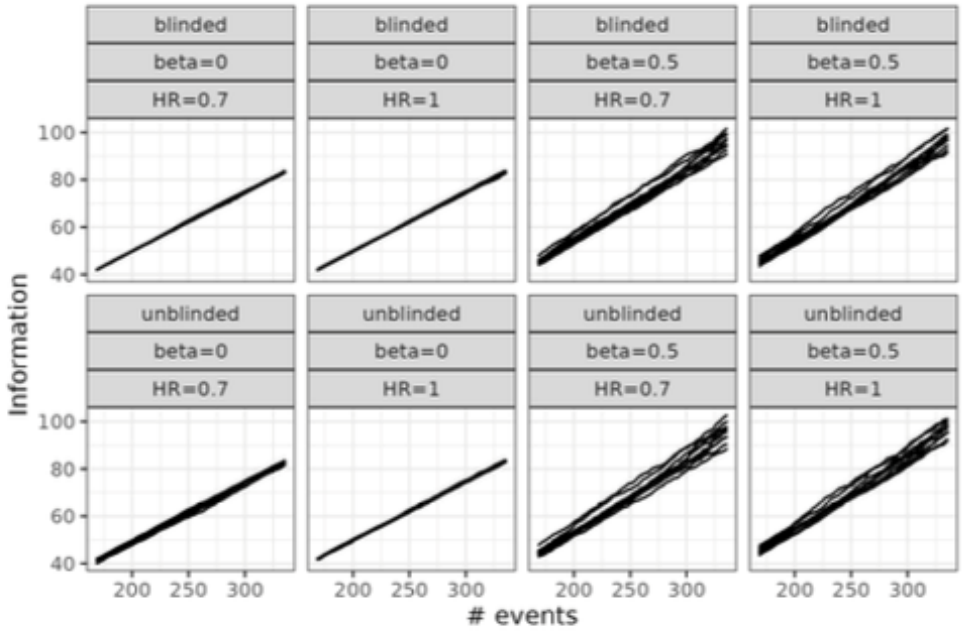
$\mathrm{se}(\hat{\delta})$ can be estimated using blinded data

- Randomly assign treatment to each patient, calculate $\mathrm{se}(\hat{\delta})$
  - Follow as closely as possible randomization method applied in trial, e.g., permuted block randomization
  - Repeat 10 times and take median

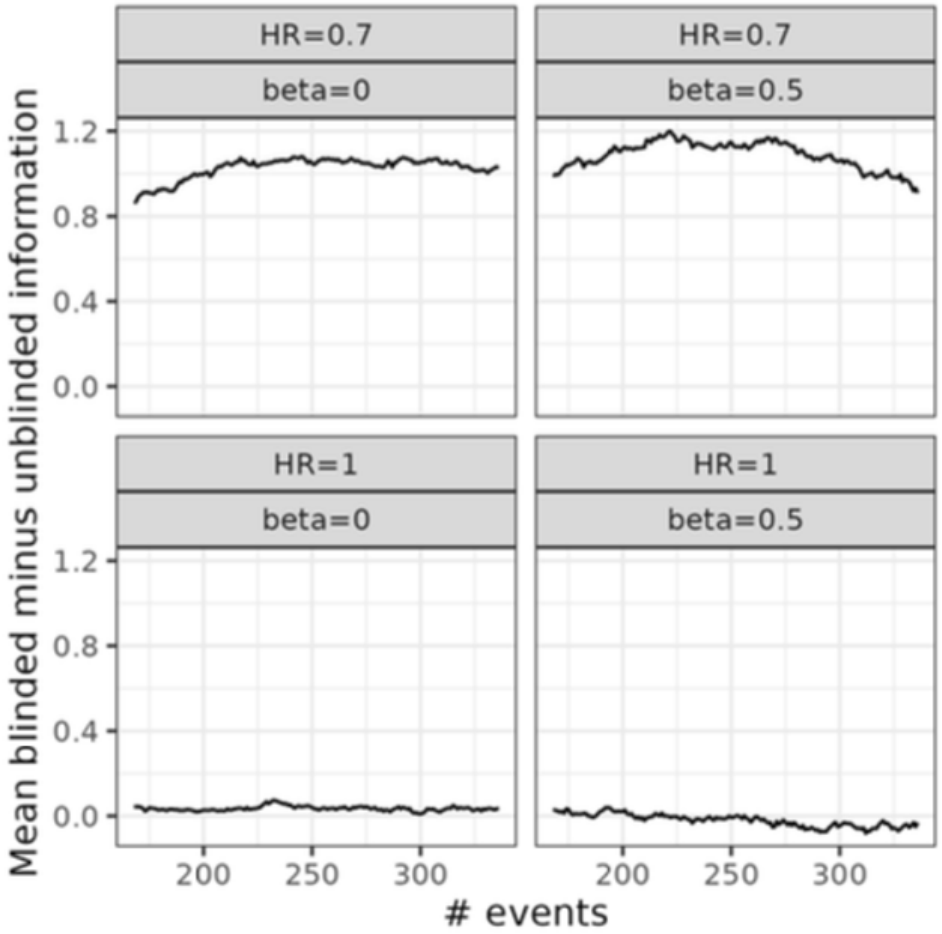
$\mathrm{se}(\hat{\delta})$ can be estimated using blinded data

- Randomly assign treatment to each patient, calculate $\mathrm{se}(\hat{\delta})$
  - Follow as closely as possible randomization method applied in trial, e.g., permuted block randomization
  - Repeat 10/25 times and take median

Information accumulation over time with and without covariate



Slight bias from blinded information estimation



Blinded information estimate slightly biased upwards when HR is not equal to 1

- Less information accumulates with each event when risk set is imbalanced
- Can be corrected given HR design assumption

Adaptive designs

Using an information adaptive design, we trigger analysis exactly when we think power is adequate (under PH assumption)

- We can either **reduce sample size** and assume that efficiency gains will mean smaller number of events will be enough (so analysis will not be delayed)
- Or we can **keep sample size the same**, so that there is **only upside potential**

Adaptive designs

We ran simulation:

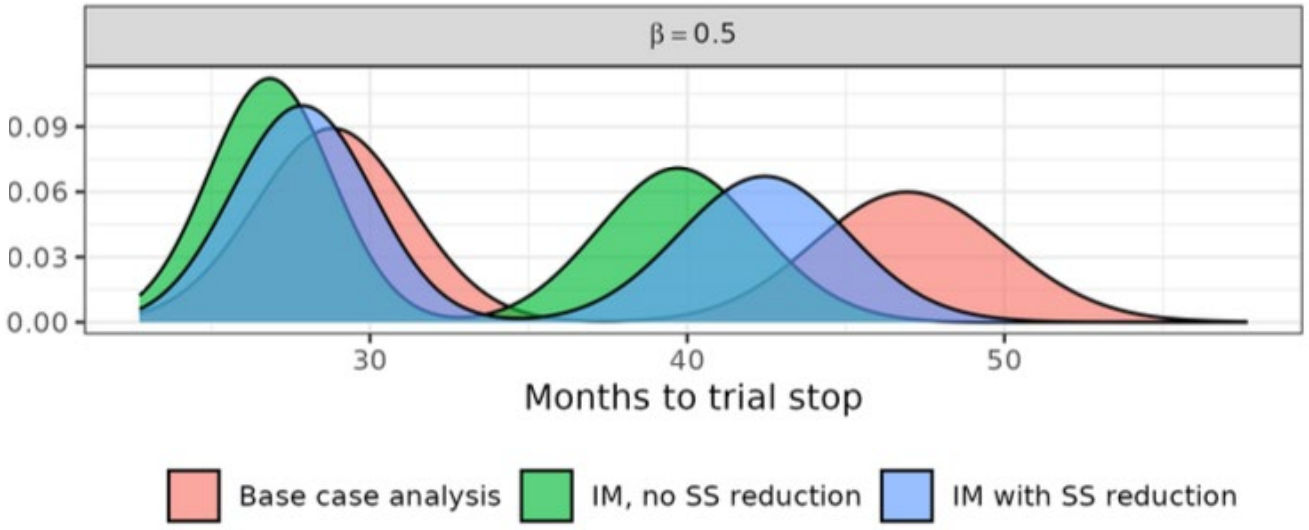
- One interim analysis
- True variance reduction is 13%
- Three designs considered, all with 90% power
  - Base case (unadjusted test, event-triggered)
  - **Information monitoring, no sample size reduction**
  - Information monitoring, 5% fewer patients enrolled

Adaptive designs

We ran simulation:

- One interim analysis
- True variance reduction is 13%
- Three designs considered, all with 90% power
  - Base case (unadjusted test, event-triggered)
  - Information monitoring, no sample size reduction
  - Information monitoring, 5% sample size reduction

Scenario where covariate is prognostic



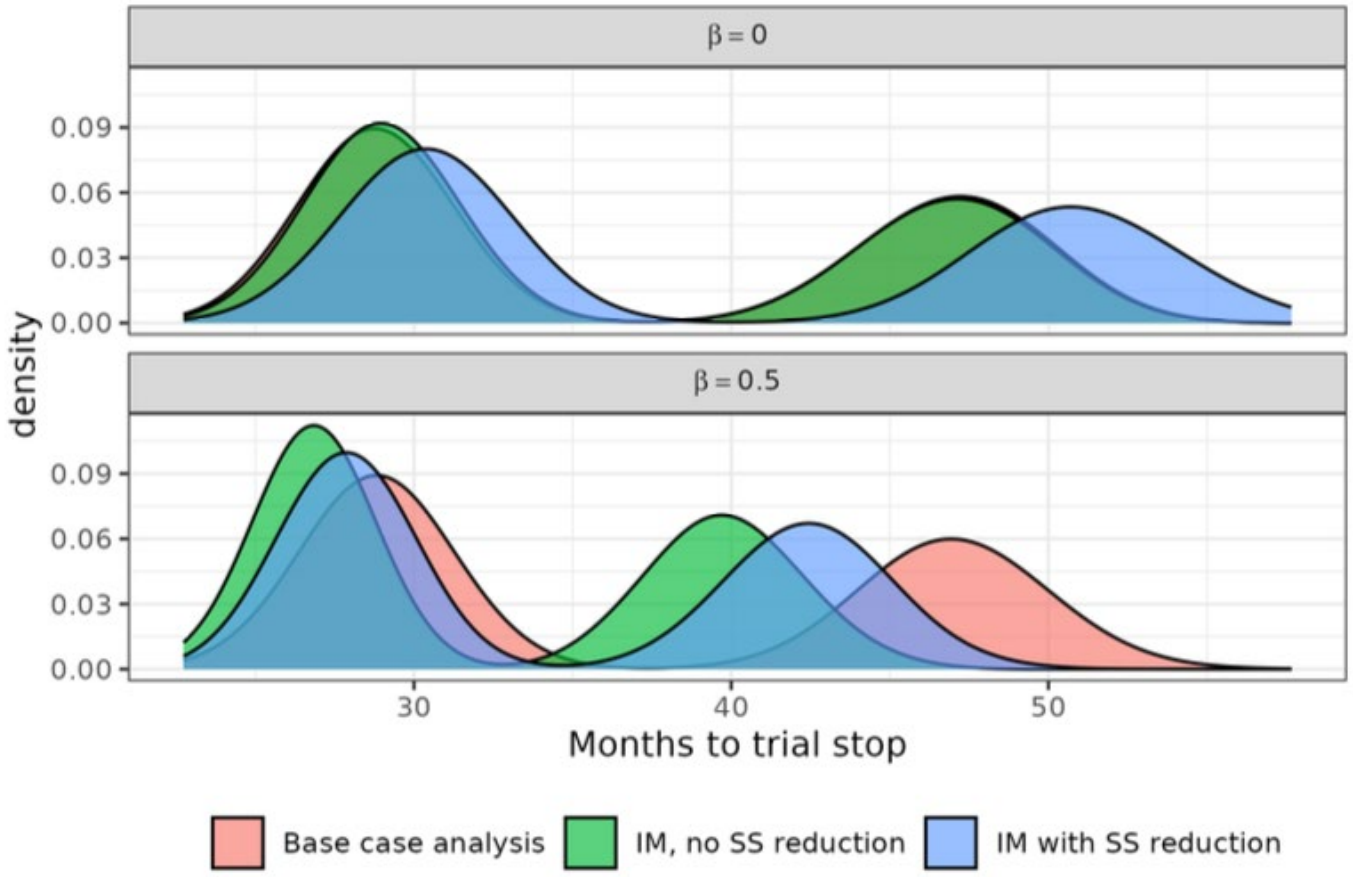
Distribution is bimodal because of interim analysis

Adaptive designs

We ran simulation:

- One interim analysis
- True variance reduction is 13%or 0%
- Three designs considered, all with 90% power
  - Base case (unadjusted test, event-triggered)
  - Information monitoring, no sample size reduction
  - Information monitoring, 5% sample size reduction

Scenario where covariate is not prognostic



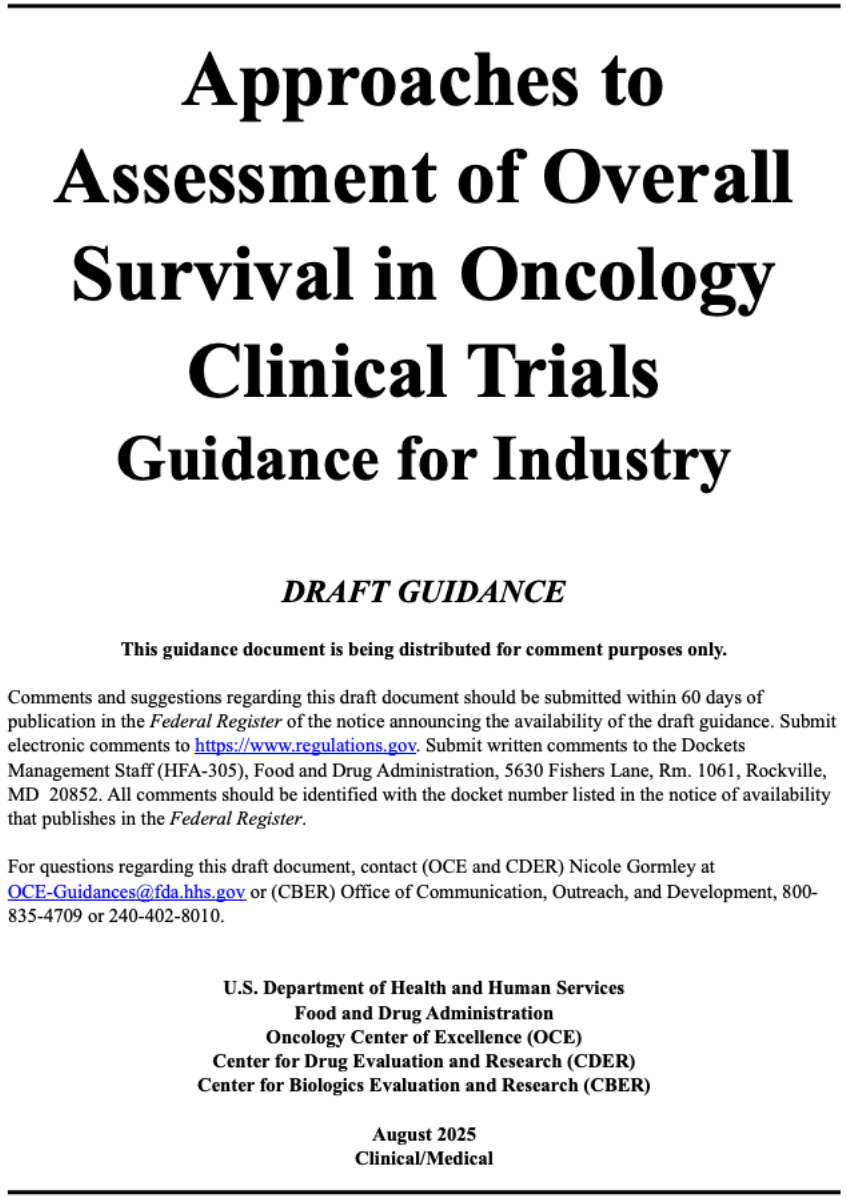
Regulatory considerations

In recent oncology-specific draft guidance FDA said:

> The timing of the interim and final overall survival analyses should be event-driven rather than based on a pre-specified time period.

Unclear if information monitoring will satisfy FDA

- Information is monotonic (with unusual exceptions) with event number, so results will not be that different
- Specifying clear rules for when information targets are met should helpsatisfy regulators



Approaches to Assessment of Overall Survival in Oncology Clinical Trials

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

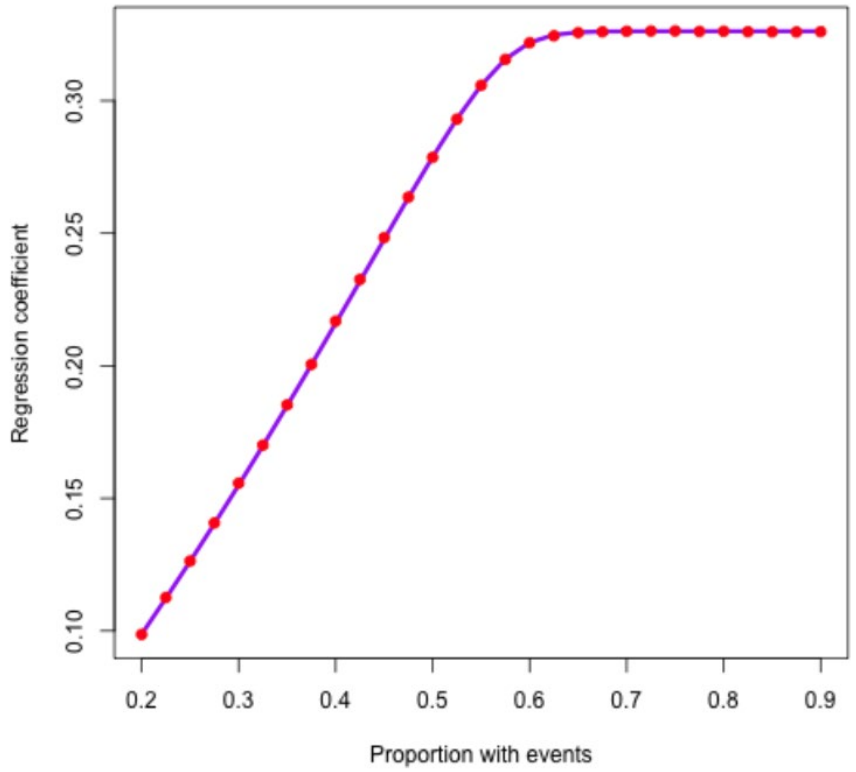
For questions regarding this draft document, contact (OCE and CDER) Nicole Gormley at OCE-Guidances@fda.hhs.gov or (CBER) Office of Communication, Outreach, and Development, 800-835-4709 or 240-402-8010.

U.S. Department of Health and Human Services
Food and Drug Administration
Oncology Center of Excellence (OCE)
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

August 2025
Clinical/Medical

Group sequential design considerations

- Important to note that efficiency gains will likely change over course of the trial
- Results from toy example with one binary covariate below
- Regression coefficients change



Purple lines comes from theory, red dots from simulation

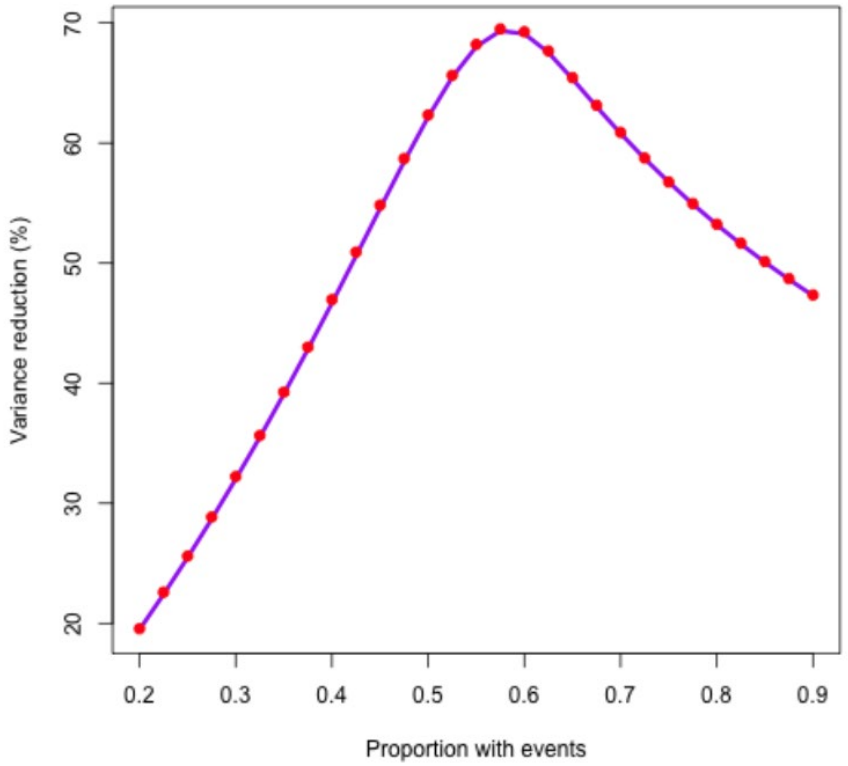
Group sequential design considerations

- Important to note that efficiency gains will likely change over course of the trial
- Results from toy example with one binary covariate below

- So do variance reductions $1 - \sigma^2_{\mathrm{CL}}/\sigma^2_{\mathrm{L}}$
  - Remember $\sigma^2_{CL} = \sigma^2_L - \pi(1-\pi)(\beta_1 + \beta_0)^T \Sigma_X (\beta_1 + \beta_0)$
  - Remember approximation $\sigma^2_L = \mathrm{d}/4$ (d is # events)

- Important to note that efficiency gains will likely change over course of the trial
- Results from toy example with one binary covariate below

- So do variance reductions $1-\sigma^2_{CL}/\sigma^2_L$
  - Remember $\sigma^2_{CL} = \sigma^2_L - \pi(1-\pi)(\beta_1 + \beta_0)^T \Sigma_X (\beta_1 + \beta_0)$
  - Remember approximation $\sigma^2_L = \mathrm{d}/4$ (d is # events)



Purple lines comes from theory, red dots from simulation

Two consequences of regression coefficients changing

Two principal consequences:

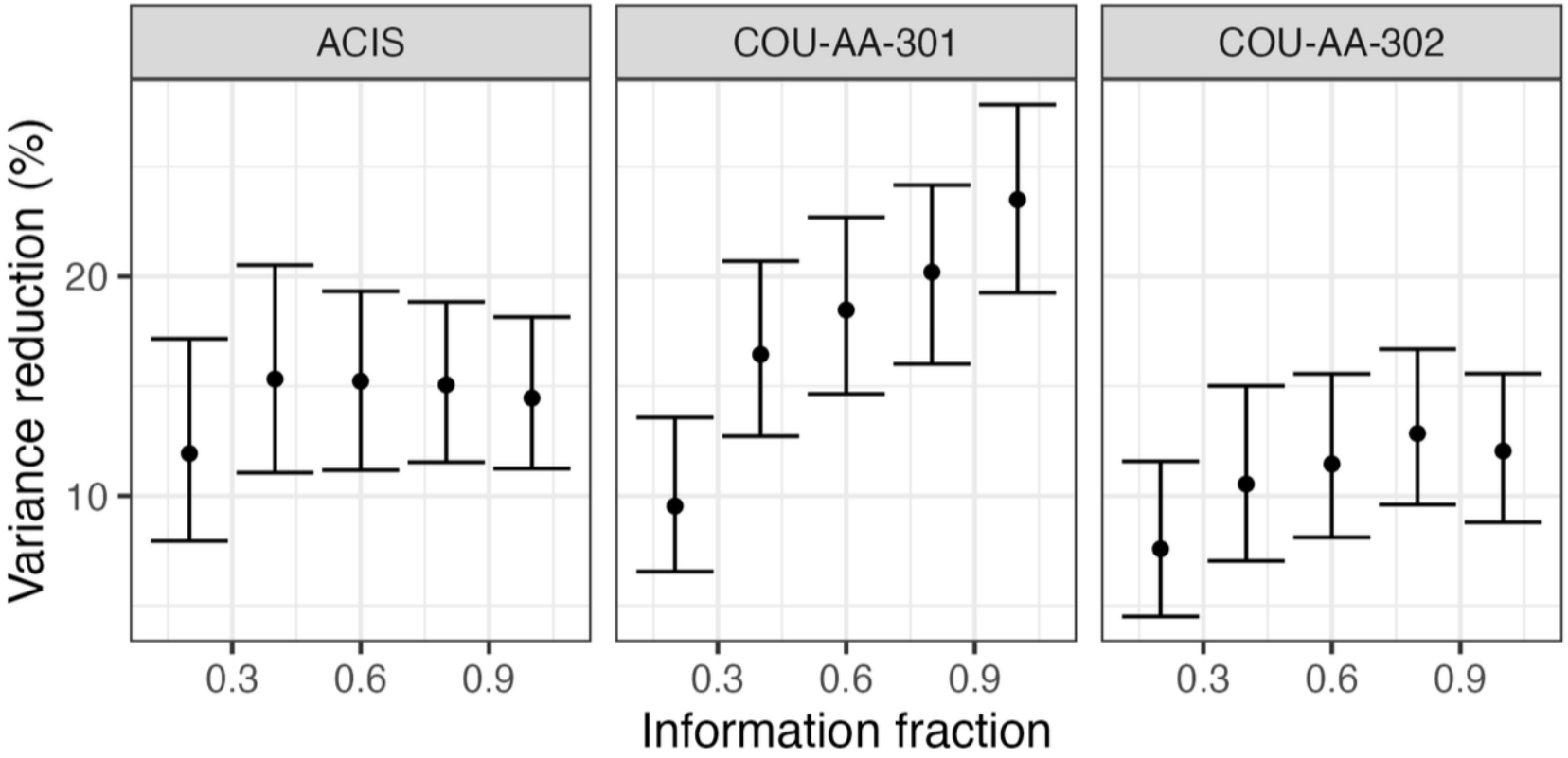
- Can’t assume same efficiency gain at interim and final analyses, in a group sequential design
  - Makes blinded event number re-targeting complicated

Two consequences of regression coefficients changing

- Can’t assume same efficiency gain at interim and final analyses, in a group sequential design
  - Makes blinded event number re-targeting complicated

- No longer have independent increments
  - Usually not much of an effect on Type I error rate
  - See Tsiatis & Davidian (2025), Van Lancker (2025)

Illustration of variance reductions changing over time in real data



ACIS: Saad et al., LancetOnc 2021
COU-AA-301: de Bono et al., NEJM 2011
COU-AA-302: Ryan et al., NEJM 2013

Conclusion

- Covariate adjustment in oncology studies can bring substantial benefits
  - More powerful, smaller or faster studies
- Finding covariates to adjust for is straightforward
  - Focus on labs (hemoglobin, albumin, etc…)
  - Search for publicly available data on Project Data Sphere
    - Note very little heme data available
- Adaptive designs can be used to maximize the benefits, information monitoring can be useful

Acknowledgments

J&J colleagues

- Shiva Dibaj
- Jozefien Buyze
- Fredrik Öhrn
- SanneRoels

Ting Ye (University of Washington)

Kelly Van Lancker (Ghent University)

References

Bachet et al., Characteristics of Patients and Prognostic Factors Across Treatment Lines in Metastatic Colorectal Cancer: An Analysis From the Aide et Recherche en Cancérologie Digestive Database, JCO 2025 (mCRC prognostic score)

Efron, An introduction to the bootstrap, CRC Press 1993 (optimism-corrected efficiency gain estimation)

EMA, Adjustment for baseline covariates in clinical trials- Scientific guideline, 2015

FDA, Adjusting for covariates in randomized clinical trials for drugs and biological products, 2023

Halabi et al., Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer, JCO 2014 (mCRPC prognostic score)

Jiang et al., The type I error and power of non-parametric logrank and Wilcoxon tests with adjustment for covariates—A simulation study, Stat Med 2008

Li et al., Estimating the Efficiency Gain of Covariate-Adjusted Analyses in Future Clinical Trials Using External Data, JRSSSB 2023 (estimating efficiency gains using bootstrap)

Tsiatis and Davidian, Independent increments and group sequential tests, arxiv 2025 (influence functions for restoring independent increments)

Van Lancker et al., Combining covariate adjustment with group sequential, information-adaptive designs to improve randomized trial efficiency, Biometrics 2025 (group sequential design with covariate adjustment; information monitoring)

Ye et al., Covariate-adjusted log-rank test: guaranteed efficiency gain and universal applicability, Biometrika 2024 (covariate-adjusted log-rank test)

Thank you

If you have more questions, please contact:
Daniel Backenroth
dbackenr@its.jnj.com