Practical Application with implementation details: Estimating the causal treatment effect in subgroups defined by a post-baseline biomarker

Dr. Dominik Heinzmann, Global Head PHC Data Science Oncology, Roche

BBS Training Series 2021, Feb 2, "A gentle Introduction to Causal Thinking"

Disclosures

- Dominik Heinzmann is a full time employee of F. Hoffmann-La Roche Ltd.
- Any opinions, findings, and conclusions expressed in this work are those of the presenters and do not necessarily reflect those of F. Hoffmann-La Roche Ltd.

- Practical application presented here is partially based on presentation given by Heinzmann and Kong at the BBS seminar “RCTs meeting causal inference: principal stratum strategy and beyond” (Sept 7, 2020).

Acknowledgments

- Sabine Lauer, Lauer Research, Germany
- Shengchun Kong, PD Data Sciences, Genentech, US
- Marcel Wolbers, PD Data Sciences, Roche, Basel
- Kaspar Rufibach, PD Data Sciences, Roche, Basel

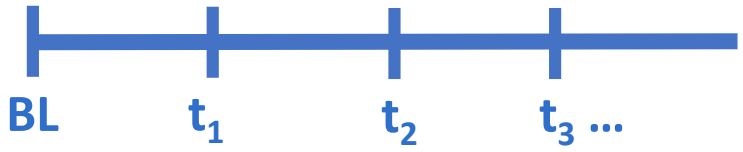
(i) Anti-drug antibodies and principal stratum approach

(ii) Principal ignorability score-weighting

(iii) Application

Practical example: Anti-drug antibodies (ADA)

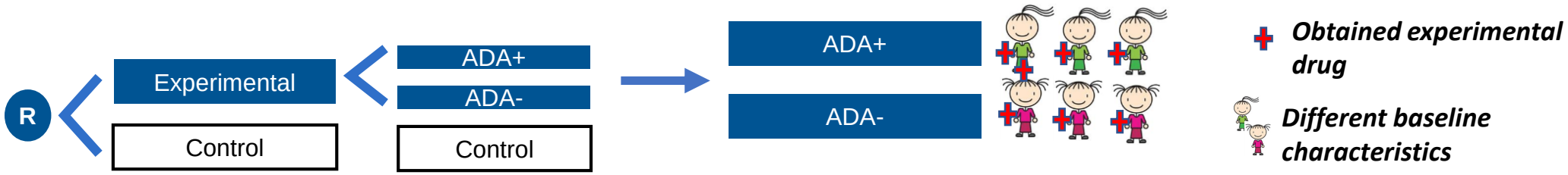
- Novel biologic treatments may provoke an unwanted immune response and form ADAs
- ADA can negatively impact safety, PK, PD and/or efficacy of such a biologic treatment
- ADA measured with defined schedule for a patient, e.g.



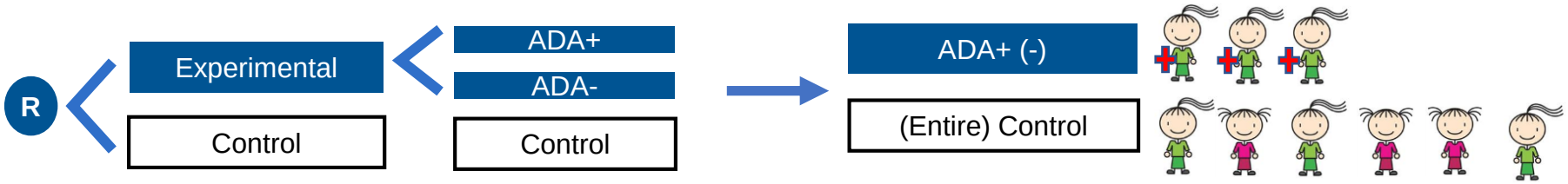
- **ADA-positive** if either new ADA formed, or ADA titer meaningfully increased after initiation of treatment
- **ADA** = **Intercurrent event** in the language of ICH E9 addendum
  - ADA is a post-randomization variable induced / influenced by treatment
  - ADA has potential impact on the interpretation of the clinical outcome

Practical example: Scientific questions of interest

- **Compare ADA+ and ADA- on treatment arm only:** No, we are interested in the treatment effect

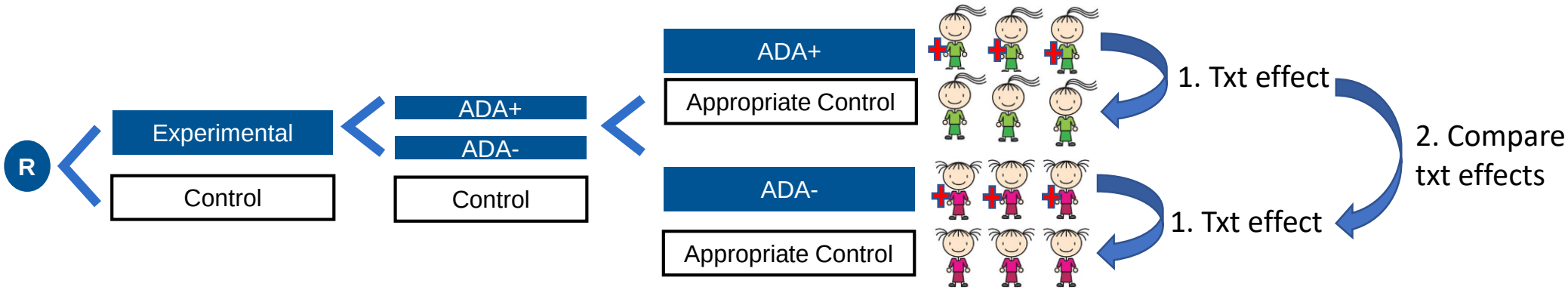


- **Compare ADA+ (ADA-) to entire control:** Naïve analysis -> Not estimating causal treatment effect as effect may be influenced by differences in baseline features



Practical example: Scientific questions of interest

- **Compare ADA+ (ADA–) to the appropriate control** *(principal stratum approach)*
  1. An assessment of whether the ADA+ (ADA–) subgroup derives benefit from experimental treatment
  2. A comparison of treatment effects between each ADA subgroup compared to corresponding control



Obtained experimental drug

Different baseline characteristics

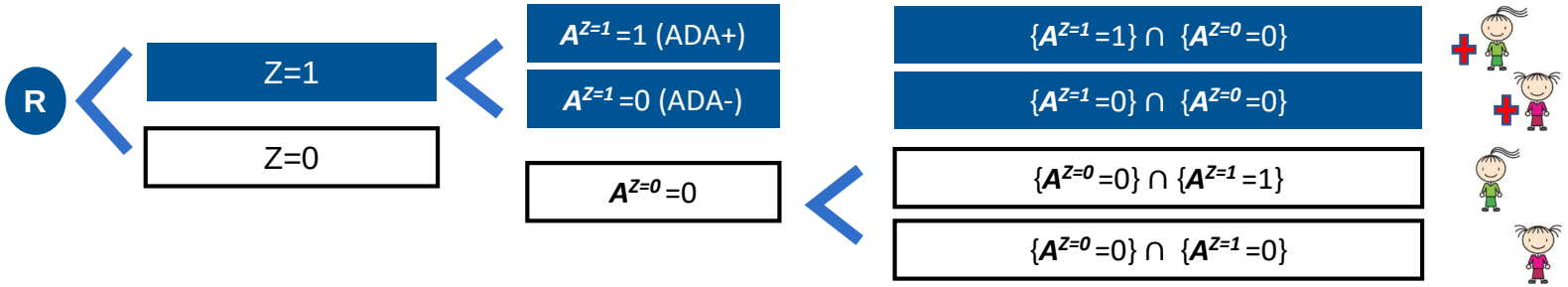
(i) Anti-drug antibodies and principal stratum approach

(ii) Principal ignorability score-weighting

(iii) Application

Potential outcomes framework

- **Practical example:** RCT with ADA measured only in experimental arm, not in control arm
- ***Z:*** treatment assignment (1=treated with experimental txt, 0=treated with control)
- ***A:*** intercurrent event
  - $A^{z=1}$ = Potential outcome of ADA under treatment **z=1**
  - $A^{z=0}$ = Potential outcome of ADA under treatment **z=0**, always =0 for practical example of ADA



- Each patient has two potential outcomes $A^{z=1}$, $A^{z=0}$ and before treatment happens both have the potential to become the *actual* outcome

- ***Y:*** outcome: Time to event variable (e.g. overall survival)
  - $Y^{z=1}$ = Potential outcome under treatment **z=1**
  - $Y^{z=0}$ = Potential outcome under treatment **z=0**

- Finally, let ***X*** be a set of baseline covariates

Potential outcomes framework

- **Causal treatment effect**: Comparison of potential outcomes
  - $\{Y_i^{z=1}; i \text{ in } \boldsymbol{S}\}$ vs. $\{Y_i^{z=0}; i \text{ in } \boldsymbol{S}\}$ on a common patient population $\boldsymbol{S}$

- For estimation of the **causal treatment effect** for ADA+ (analogous for ADA-), we need to estimate
  - distributions of $Y^{z=1} \mid A^{z=1}=1$ and $Y^{z=0} \mid A^{z=1}=1$

- Distribution of $Y^{z=1} \mid A^{z=1}=1$ can be estimated from observed data

$$P(Y^{z=1} > t | A^{z=1} = 1) = P(Y^{z=1} > t | Z = 1, A^{z=1} = 1) = P(Y > t | Z = 1, A = 1)$$

  - 1. equality: Treatment assignment independent of A and $\boldsymbol{X}$, which is satisfied in RCT*
  - 2. equality: C*onsistency assumption*, i.e. having a well-defined treatment such that a patient's potential outcome equals that observed in the trial,

- Distribution of $Y^{z=0} \mid A^{z=1}=1$ cannot be estimated from trial data without further assumptions
  - $Y^{z=0}$ and $A^{z=1}$ are not jointly observed in the same patient in a RCT

- Literature provides a variety of assumptions to enable drawing causal conclusions on the effect of treatment in the principal stratum *(Ding and Lu, 2017)*

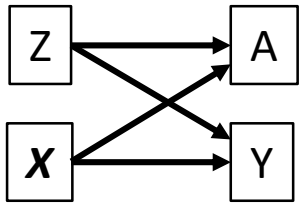
*In case of a landmark approach as outlined on slide 12, landmark population needs to be close to overall population

Principal Ignorability *(conditional independence)*

- Aim: Allow estimation of distribution of $Y^{z=0}$ | $A^{z=1}$=1 based on observed data

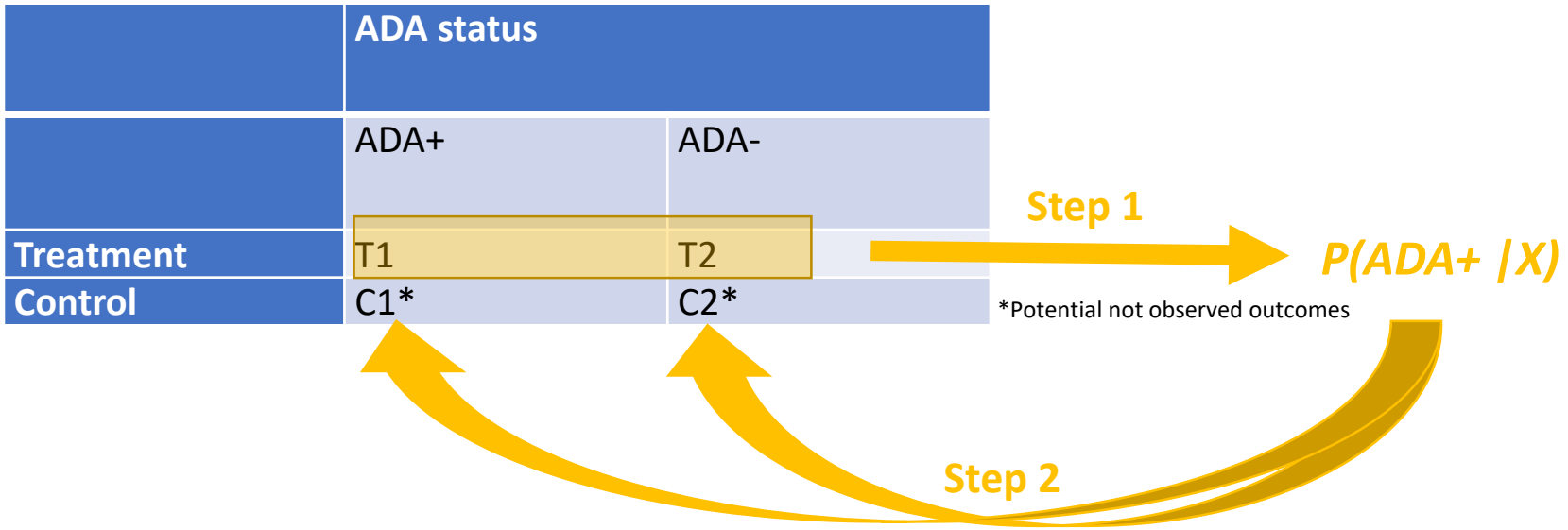
- **Principal ignorability assumption:** $Y^{z=0} \perp\!\!\!\perp A^{z=1} \mid \mathbf{X}$

- Whether a patient would be $A^{z=1}$=1 (i.e. ADA+) under treatment is independent of their outcome where they assigned to the control group, conditional on ***X***
- Simple **causal diagram (DAG)** that is compatible with this assumption for the control arm in our setting *(Dukes et al. 2021)*



- Similar to "no unmeasured confounding" assumptions often used in propensity score approaches in observational studies, but principal ignorability is an assumption *across-worlds*
  - *Across-worlds:*
    - Potential outcomes across treatment and control are never observed jointly
    - For a patient on the control arm we can observe $Y^{z=0}$ (i.e outcome on control) but not $A^{z=1}$ (i.e. ADA status on treatment)
    - For a patient on the experimental treatment arm, we can observe $A^{z=1}$ but not $Y^{z=0}$

- Weighted Placebo ("Control") Patients Approach (Bornkamp & Bermann 2020)
- Principal Ignorability Score-weighting (Stuart & Jo 2015)



Step 1. Use groups T1 and T2 to fit a logistic regression model of ADA status on baseline covariate X.

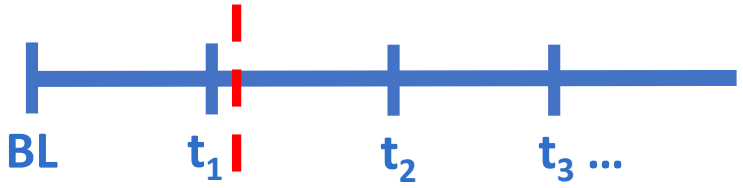
Step 2. Use prediction model in step 1 to get the estimated probability of being ADA+ if taking experimental treatment for subjects in control group (C1*+C2*).

Step 3. Use weight 1 for patients in group A, and the weight calculated in step 2 as the weight for control group (C1*+C2*) to fit a weighted Cox regression model to estimate the hazard ratio and associated CI.

Landmark analysis

- Our example ***Y*** is a time to event variable (overall survival; OS)
- OS may be a competing risk to observe the intercurrent event *A*
  - *Risk of **immortal time bias**:*
    - *Patients in experimental arm would need to live long enough to observe A*
    - *Immortal time: Time to study entry to first post-randomization ADA assessment*
    - *Patients in control arm can die during this "immortal time"*
    - *Not accounting for leads to bias favoring the experimental arm subgroups*

- **Landmark analysis** approach can be applied, such that *A*:=occurrence/absence of ADA *at a fixed landmark time point*



- **Selection** of an appropriate landmark time point is discussed in application section
  - **Remark:** Need to be early enough such that landmark population is close to the overall clinical trial population so "randomization" property holds such that treatment assignment is independent of ***X*** and *A*

(i) Anti-drug antibodies and principal stratum approach

(ii) Principal ignorability score-weighting

(iii) Application

Data set: General structure

1. Landmark*

2. BL covariates

A	B	C	D	E	F	G	H	I	J	K	L	M	N
SUBJID	ACTARM	SAFFL	ADALM4	CNSR	AVAL	AVALU	OS1MNTH	AGE65	BECOG	SEX	BNLR	TOBHX	RACEGRP
1	CONTROL	Y	CONTROL	0	12.51140454	MONTHS	Y	>= 65	1	F	4.916666667	PREVIOUS	WHITE
2	EXPERIMENTAL	Y	ADA- 4 weeks	0	11.78987509	MONTHS	Y	< 65	0	M	3.333333333	NEVER	WHITE
3	EXPERIMENTAL	Y	ADA- 4 weeks	1	30.82499026	MONTHS	Y	< 65	0	F	1.994764398	CURRENT	WHITE
4	CONTROL	Y	CONTROL	0	8.285732628	MONTHS	Y	>= 65	0	F	4.454545455	PREVIOUS	WHITE
5	EXPERIMENTAL	Y	ADA- 4 weeks	0	4.651730563	MONTHS	Y	>= 65	0	M	4.700680272	CURRENT	WHITE
6	CONTROL	Y	CONTROL	1	21.69861406	MONTHS	Y	< 65	0	M	3.369975787	PREVIOUS	WHITE
7	CONTROL	Y	CONTROL	0	8.146336513	MONTHS	Y	< 65	1	M	2.111111111	PREVIOUS	WHITE
8	CONTROL	Y	CONTROL	0	9.533927037	MONTHS	Y	>= 65	1	F	2.172413793	PREVIOUS	WHITE
9	CONTROL	Y	CONTROL	1	25.25452277	MONTHS	Y	< 65	0	F	2.645833333	CURRENT	WHITE

*Patient with OS $Y^{a=i}$ < landmark time point (i=0,1) are excluded

- ADA incidence in landmark population: ~22%

Implementation details: Selection of baseline covariates $\boldsymbol{X}$

- **Reminder**
  - For standard RCT analyses, randomization is sufficient to make causal inference on treatment effect
  - For the treatment effect in a principal stratum in a RCT, principal ignorability assumptions needs to be satisfied on top of randomization
    - $Y^{z=0} \perp\!\!\!\perp A^{z=1} \mid \boldsymbol{X}$
  - Hence one needs to adjust for all baseline covariates that make the potential outcomes of ADA and the final outcome of overall survival for patients on control independent.
  - Principal ignorability by itself is a causal assumption. To justify, you need scientific and medical reasoning. Data can help, but on its own is not sufficient.
    - E.g. it may be that you have not included all baseline covariates $\boldsymbol{X}$
    - or if even the right set of covariates exist because it could be that the counterfactual ADA status in control patients could have a direct impact on their outcome Y

Implementation details: Selection of baseline covariates *X*

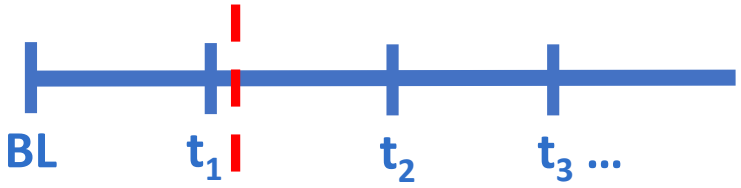
- **Formally,** only baseline covariates that influence both ADA and overall survival shall be included

- **General**
  - Should be selected prospectively, based on literature review, clinical and statistical expert input,
  - Directed Acyclic Graphs (DAG) are useful to encode assumptions regarding both, the ADA and OS mechanisms.
  - Statistical modelling can be used to select candidates reasonably likely to influence OS or ADA separately.

- **Particular**
  - Often challenging to assess if baseline covariate influences ADA, easier to assess if that covariate influences OS.
  - If a baseline covariate is related to OS but its influence on ADA is uncertain, include that covariate as it does not introduce bias and can improve precision of your estimate *(Brookhart et al. 2006)*
  - If a baseline covariate is only influencing ADA but not OS, the covariate should be omitted
    - Can introduce bias and decrease precision since such a covariate has the same structure as instrumental variables as it is related to ADA but unrelated to OS (except through ADA) *(Myers et al 2011)*

Implementation details: Landmark analysis

- **Landmark analysis** to address immortal time bias



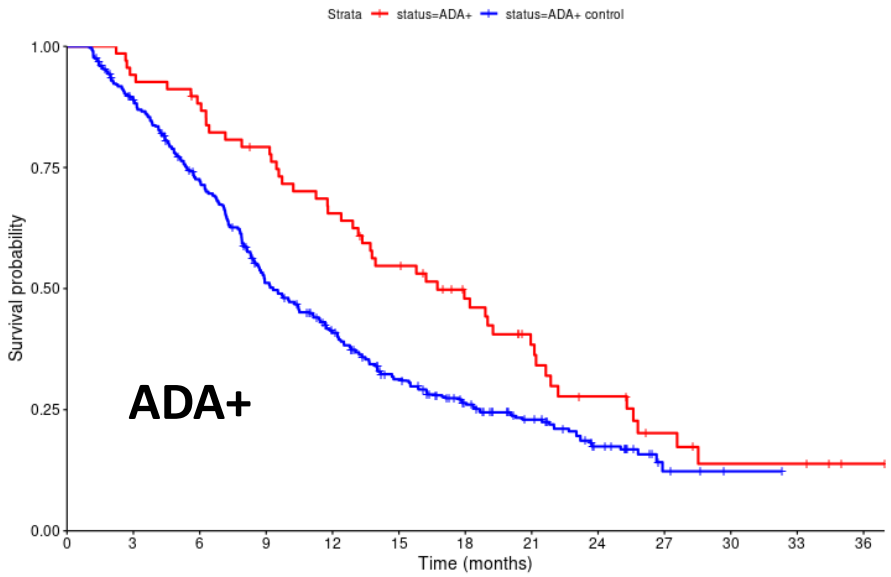
- **Reminder**: Landmark population needs to be very close to the overall population such that “randomization” holds

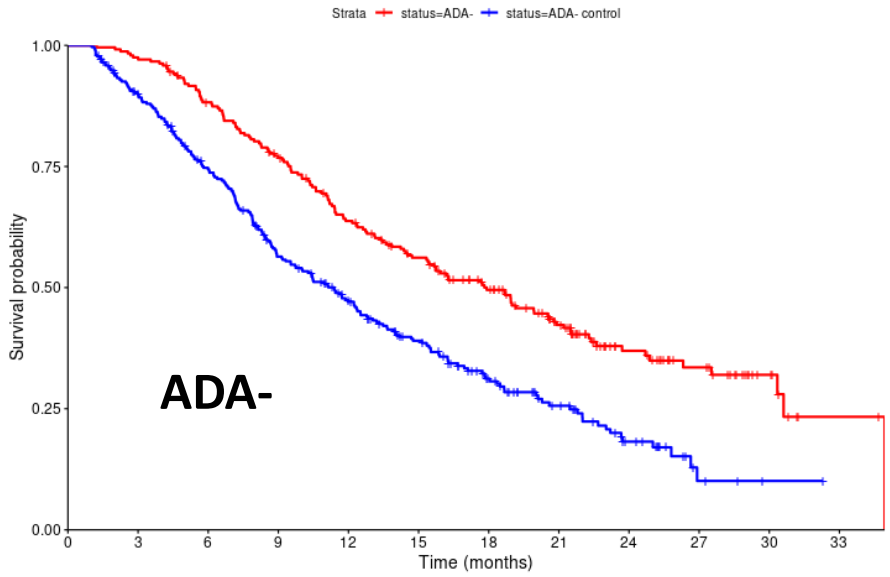
- **Practical consideration choice of landmark time point for ADAs:**
  - Guided by the scheduled ADA assessments
    - E.g. in oncology often at BL + every 3 weeks afterwards just prior to next dose
  - Aim for an early time point to have the population with observable *A* close to the overall trial population
    - Later landmark time point causes more patients dying or being censored prior to landmark
  - Aim to catch at the landmark time point a large proportion of the patients in the trial who are ADA-positive at any time point in the trial.
    - In many practical applications for ADAs, this will be the case.

- **In our example:** We selected a landmark timepoint of **4 weeks**
- **Remark:** For the more general case when you have missing ADA status at landmark for some patients, more advanced methodology can be applied *(Kong et al. 2020)*

Results

Groups	Hazard ratio (95% Confidence Interval)
ADA+ vs appropriate control	0.59 (0.43, 0.81)
ADA- vs appropriate control	0.57 (0.46, 0.71)

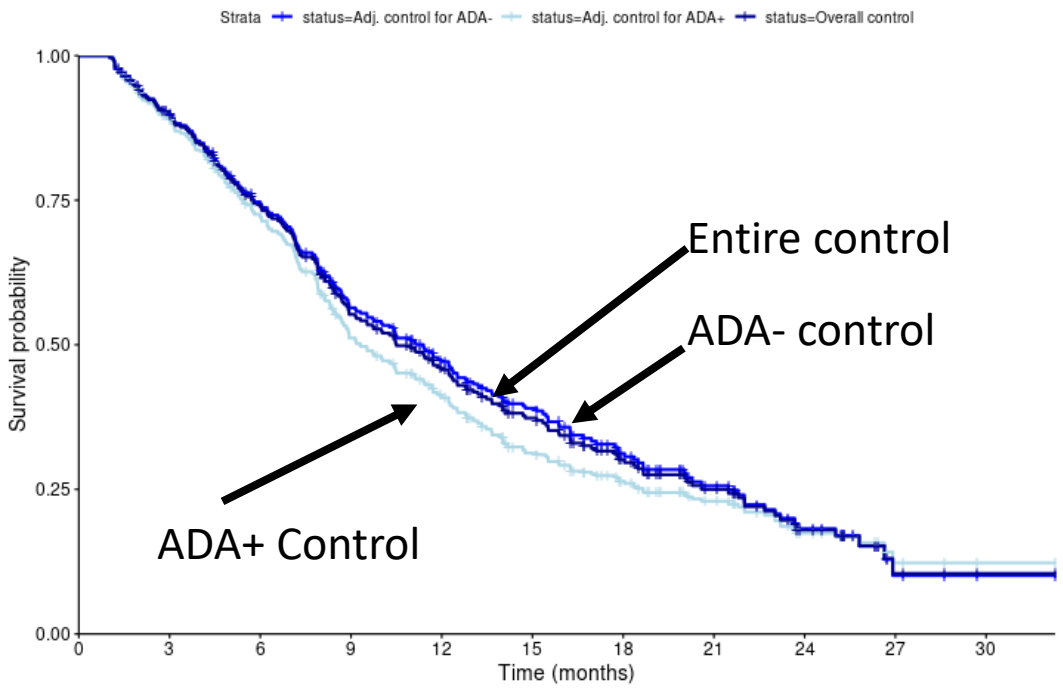


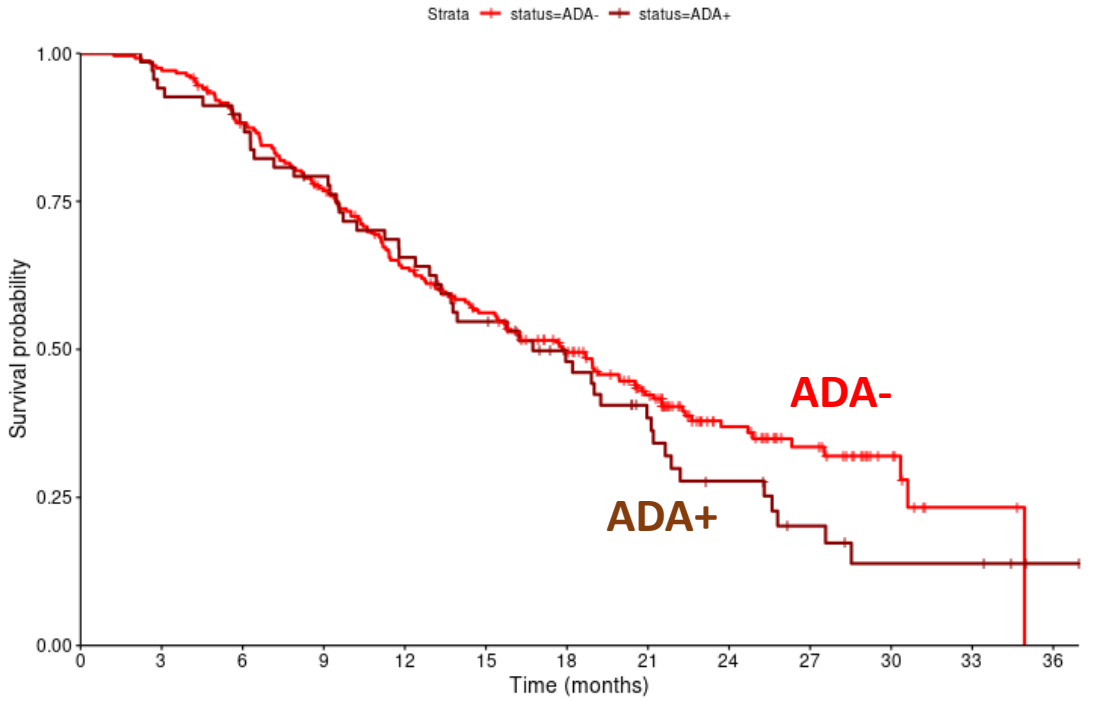


Observation: Similar and strong treatment effect in both strata, suggesting no clinically meaningful impact of appearance of ADA on outcome

Results

Results principal stratum approach *(previous slides)*	HR (95% CI)
ADA+ vs appropriate control	0.59 (0.43, 0.81)
ADA- vs appropriate control	0.57 (0.46, 0.71)





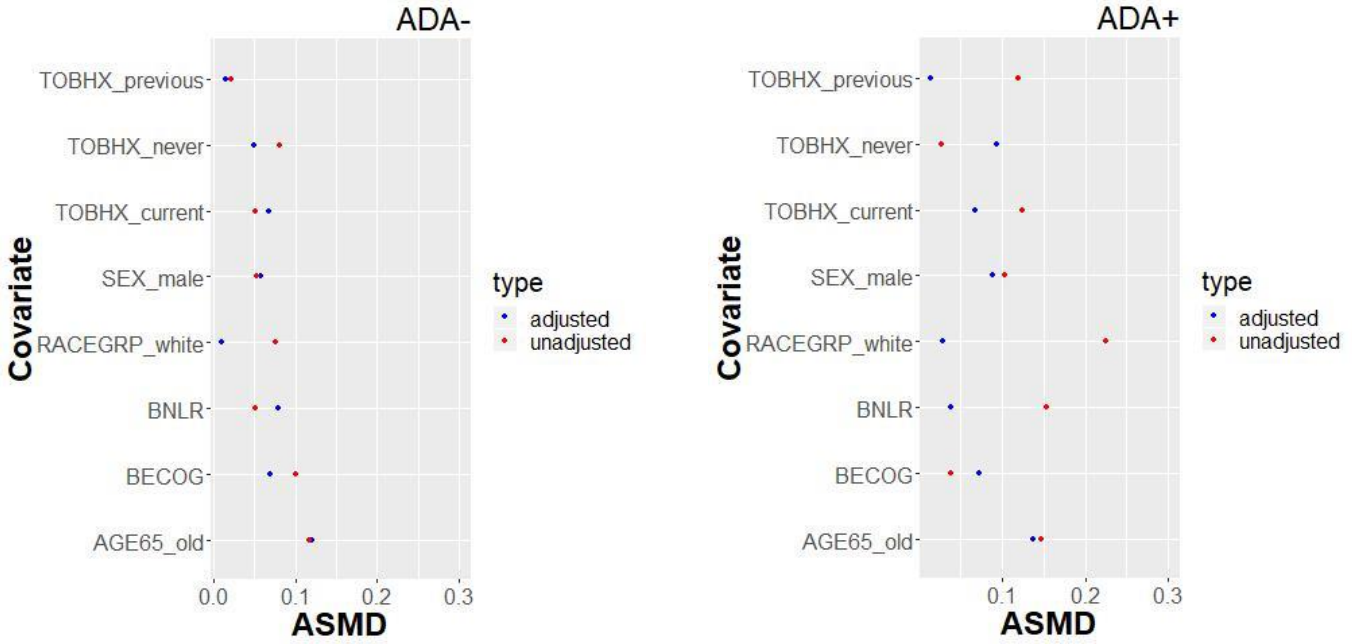
vs entire control (not estimating a causal effect)	HR (95% CI)
ADA+ vs entire control	0.67 (0.49, 0.92)
ADA- vs entire control	0.56 (0.45, 0.69)

Only treatment arm (not estimating a causal effect)	HR (95% CI)
ADA+ vs ADA- (unadjusted)	1.19 (0.86, 1.64)
ADA+ vs ADA- (adjusted)	1.16 (0.84, 1.62)

Model diagnostics

- **Balancing of covariates:** Assess if weighting using estimated propensity score induces a balance in measured covariates between ADA stratum and appropriate control

$$\text{ASMD} := \frac{|\bar{X}_{treatment} - \bar{X}_{control}|}{\sqrt{(s^2_{treatment} + s^2_{control})/2}}$$



- Not influenced by sample size, hence can be used to compare balancing of measured covariates when different weights are assigned to the same patient in control
- **Challenges:**
  - Literature suggests compare to constants like 0.1 or 0.25 but no concrete guidance in applied situations
  - Principal stratum analysis
    - usually is in setting with "small" N, ie either stratum $\{A^{z=1}=1\}$ or $\{A^{z=1}=1\}$ will have limited sample size
    - Additionally, in trials with n:1 randomization, one may have a limited pool of control arm patients not allowing to strongly balance covariate distributions among treatment arms

Model diagnostics

Supplemental model diagnostic framework

- **"Balance similarity" with a RCT**
  - Compare expected number of covariates with ASMD > constant (eg 0.1 and 0.25) for a RCT with same sample size *(details see Appendix 1)*

- **"Inversions"**
  - Assess proportion of ASMDs which get larger after adjustment

Stratum	N	Randomization ratio	No. of BL covariates	#Expected features >0.1	#Obs. Features* >0.1	#Rev. features> 0.1	#Expected features >0.25	#Obs. Features* >0.25	#Rev. features> 0.25
ADA+	68	1:1	6	3.4	1	0	0.9	0	0
ADA-	240	1:1	6	1.6	1	1	0	0	0

Balance similarity Inversion

*For covariates with multiple levels, if one level is above threshold (eg TOBHX_previous) then the covariate TOBHX is accounted as >threshold

Discussion & Links

- Causal thinking provides a transparent way to discuss causal effects
- Key benefit it that assumptions made are explicit and thus can appropriately be considered in the interpretation of the results from the analyses
- The application discussed shows that causal thinking (potential outcome framework) can help to implement the ICH E9 addendum, in particular the principal stratum estimand approach

Code available

- **Example presented:** R code available at https://github.com/openpharma/BBS-causality-training
- **Alternative example** with code based on Bornkamp, Kaspar et al (2020): https://oncoestimand.github.io/princ_strat_drug_dev/princ_strat_example.html

REFERENCES

- **Bornkamp B, Bermann G.** (2020). “Estimating the treatment effect in a subgroup defined by an early post-baseline biomarker measurement in randomized clinical trials with time-to-event endpoint”, *Statistics in Biopharmaceutical Research, 12, 1, 19-28.*
- **Bornkamp, B., Rufibach, K., Lin, J., Liu, Y., Mehrotra D., Roychoudhury S, Schmidli S., Shentu Y., and Wolbers M.** (2020). “Principal Stratum Strategy: Potential Role in Drug Development.” Industry Working Group "Estimands in Oncology". https://arxiv.org/abs/2008.05406.
- **Brookhart MA., Schneeweiss S., Rothman KJ., et al.** (2006), “Variable selection for propensity score models“, *Am J Epidemiol,*163, 1149−56.
- **Ding P., and Lu J.** (2017), “Principal Stratification Using Principal Scores,” *Journal of the Royal Statistical Society*, Series B, 79, 757–777.
- **Dukes O., Lancker K.V., Bornkamp B., Heinzmann D., Rufibach K., and Wolbers M.** (2021), ”Letter to the editor: On identification of the principal stratum effect in patients who would comply if treated”, *Statistics in Biopharmaceutical Research*, DOI: 10.1080/19466315.2021.1872697
- **Heinzmann D., and Kong S.** (2020), “Principal stratum strategy to investigate anti-drug antibody impact on outcome in randomized controlled trials”, *BBS Seminar Principal Stratification and beyond* http://bbs.ceb-institute.org/wp-content/uploads/2020/09/06-Heinzmann.pdf
- **Kong S., Heinzmann D., Lauer S. and Tian L.,** (2020), “Weighted Approach for Estimating Effects in Principal Strata with Missing Data for a Categorical Post-Baseline Variable in Randomized Controlled Trials”, https://arxiv.org/abs/2101.04263
- **Myers J.A., Rassen A.R., Gagne J.J., et al. (2011),** “Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates”, *Am J Epidemiol*, 174, 1213-1222.
- **Stuart E. A., and Jo, B.**, (2015), “Assessing the sensitivity of methods for estimating principal causal effects”, *Statistical methods in medical research,* 24 (6), 657-674.

Q & A

Appendix 1

"Balance Similarity" with a RCT

- In a RCT, for feature *m*, we have
- $D_m = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}}} = \frac{\bar{X}_T - \bar{X}_C}{s\sqrt{\frac{1}{n_T} + \frac{1}{n_C}}} \sim t_{n_T + n_C - 2}$ since $s_T^2 = s_C^2 = s^2$
- $D_m = SDM \cdot \frac{1}{\sqrt{\frac{1}{n_T} + \frac{1}{n_C}}}$ with SDM being the standardized mean difference
- Hence $P(|SDM| > z) = P\left(|SDM| \frac{1}{\sqrt{\frac{1}{n_T} + \frac{1}{n_C}}} > z \frac{1}{\sqrt{\frac{1}{n_T} + \frac{1}{n_C}}}\right)$ which can be computed from the t-distribution
- Finally, for M features, the expected number of features with */SDM/>z*, follows from Binomial(M, $P(|SDM| > z)$)
- **Application:**
  - RCT with 1:1 randomization and n=38, on average, 3.1 covariates would have ASMD values > 0.25, and on average, 7.3 covariates would have ASMD values > 0.1
  - This can now be compared with the observed counts to make an assessment of the model balancing

- **Remark:** Covariates with multiple levels (e.g. *race*), the covariate is considered above the threshold if the ASMD of any of the levels exceed the threshold