Stop the abuse: A plea for a more principled approach to the analysis of time-to-event endpoints with competing risks, with a focus on analysis of AEs

Kaspar Rufibach
Methods, Collaboration, and Outreach Group, Roche Basel
ISCB Milan, 28th August 2023

Acknowledgments

- Thomas Künzel.
- SAVVY consortium, specifically Regina Stegherr, Jan Beyersmann, Claudia Schmoor, Tim Friede.
- X-industry working group on estimands for time-to-event endpoints.
- Competing risks + estimands: Jan Beyersmann, Marcel Wolbers.
- Comments on linkedin post.

Extended version of this talk, incl. recording (BBS talk from earlier this year):

www.kasparrufibach

Take home messages

Need accurate estimates of P(AE) + comparison between arms.

IP and (1 - KM) biased irrespective of what we use them for.

Bias "does not cancel out" when comparing P(AE) between arms in RCT.

Let me explain.

Estimation of P(AE)

What does the incidence proportion estimate?

Incidence proportion in interval from 0 to t:

$$\widehat{IP}_E(t) \quad = \quad \frac{\text{Number of patients with AE in } [0, t] \text{ and that this AE is observed}}{n_E}.$$

$\widehat{IP}_E(t)$ estimates:

P(AE happens in $[0, t]$ and that this AE is observed **before censoring**).

$\widehat{IP}_E(t) \leq \widehat{P}(\text{AE happens in } [0, t]) \Rightarrow \widehat{IP}_E(t)$ **underestimates** absolute AE risk.

With censoring it is unclear which quantity $\widehat{IP}_E$ is estimating.

Simple incidence proportion is biased if we have unequal follow-up or censoring.

Estimate P(AE) using time-to-AE

Consider time-to-first-AE

Redefine question: Consider **time-to-first-AE**.

- Estimate P(AE happens in $[0, t]$) using 1 - Kaplan-Meier.
- Correctly accounts for **censoring**.
- Consistently estimates AE risk at t, accounting for varying follow-up.

What does (1 - $\widehat{\text{KM}}$) with censoring of CEs estimate?

Administrative censoring: patients may still experience event at later time point.

Not for CEs!

What does (1 - $\widehat{\text{KM}}$) with censoring of CEs estimate?

- **Violates independent censoring assumption**:
  - Patient censored at death will NEVER experience AE.
  - Patients who will never experience AE treated as if they could still have one.
- Less than 100% of patients experience AE **before** death:
  - Some die before AE $\Rightarrow$ P(AE) $<$ 1.
  - But (1 - $\widehat{\text{KM}}$) approaches 1 $\Rightarrow$ naive (1 - $\widehat{\text{KM}}$) **overestimates** P(AE).

1 - Kaplan-Meier is biased if we have competing events.

Is this relevant at all?

How large can the bias be?

The SAVVY project

9 pharma

9 pharma + 3 universities

The SAVVY project

Data from **17 RCTs** in various indications.

200 - 7171 patients.

186 AEs.

SAVVY webpage

Goal: compare bias of estimators.

What is "gold standard"?

Gold standard: Aalen-Johansen estimator

What is "best" estimator to benchmark against?

Estimator	Accounts for censoring	Accounts for CEs
Incidence proportion	No	Yes
1 - Kaplan-Meier	Yes	No
Aalen-Johansen estimator	Yes	Yes

All **nonparametric**: no constant hazard assumption.

Aalen-Johansen:

- Generalizes Kaplan-Meier to competing risk and general multistate models.
- **No censoring**: Aalen-Johansen = incidence proportion.
- **No competing events**: Aalen-Johansen = (1 - Kaplan-Meier).

Bias of common estimators of AE risk

Estimation of AE risk

Incidence proportion:

- Accounts for CEs but not censoring.
- **Underestimation of P(AE) up to factor THREE!**

1 - Kaplan-Meier:

- Accounts for censoring but not CEs.
- **Overestimation of P(AE) up to factor FIVE!**

SmPC frequency categories

SmPC frequency categories:

- Very rare: $< 0.01\%$.
- Rare: $< 0.1\%$.
- Uncommon: $< 1\%$.
- Common: $< 10\%$.
- Very common: $\geq 10\%$.

		gold-standard Aalen-Johansen				
		very rare	rare	uncommon	common	very common
incidence proportion	very rare	**6**				
	rare		**0**			
	uncommon			**6**		
	common				**86**	2
	very common					**86**
1-Kaplan-Meier	very rare	**6**				
	rare		**0**			
	uncommon			**4**		
	common			2	**72**	
	very common				14	**88**

Potential impact on (labeling +) reimbursement!

Bias of common estimators of relative AE risk

Estimation of relative AE risk

Incidence proportion:

- Over- and underestimation observed.
- **Overestimation of RR up to factor of almost 3.**

1 - Kaplan-Meier:

- Over- and underestimation observed.
- **Underestimation of RR up to factor of $>$4.**

IQWiG categorization of evidence

IQWiG categorization of evidence applied to HR, IQWiG (2017):

- No effect: 1 included in CI,
- Minor: upper bound of CI in interval [0.9; 1) for HR < 1.
- Considerable: upper bound of CI in interval [0.75; 0.9).
- Major: upper bound < 0.75.

		HR Cox for AE			
		(0) no effect	(a) minor	(b) considerable	(c) major
RR gold-standard Aalen-Johansen	(0) no effect	**42**	3	3	1
	(a) minor	9	**2**	1	
	(b) considerable	4	1	**3**	2
	(c) major	2		4	**17**

Effect measure may have **large impact** on decision.

Potential impact on (labeling +) reimbursement!

Arm-wise bias does not cancel out in relative comparisons.

Comparison of ESTIMATORS.

Irrespective of what you choose as ESTIMAND.

Ultimately: not a question whether it matters!

Use appropriate statistical method from the start!

Now we have seen what does not work.

But what does work?

Aalen-Johansen: properly accounts for varying follow-up times and competing risks.

Take home messages

Need accurate estimates of P(AE) + comparison between arms.

IP and (1 - KM) biased irrespective of what we use them for.

Bias "does not cancel out" when comparing P(AE) between arms in RCT.

How would good look like in ten years?

Clear specification of goal:

- Determine and monitor **safety profile** of drug.
- Assess **causality** of (unexpected) safety signals.
- Balance **risk & benefit**.
- **Estimate risk** (probability) of an AE and enable safety differentiation.
- **Predict** patient-level drivers of AEs.
- Support characterisation of benefit in terms of **comorbidities**.

Derive **estimand**.

Inform **data collection**.

Chose appropriate **estimator / statistical analysis method**.

Call to action!

Estimate **disease-specific P(AE)'s**, properly discussing therapeutic area specific CEs.

Influence **updating of guidelines**.

Use Aalen-Johansen in a real clinical trial.

Resources

SAVVY webpage:

- Exemplary code for all methods.
- All papers and talks.
- Papers:
  - SAP: Stegherr et al. (2021a).
  - Methods: Stegherr et al. (2021c).
  - 1-sample: Stegherr et al. (2021b).
  - 2-sample: Rufibach et al. (2022).
- Effective statistician podcasts:
  - About SAVVY: https://theeffectivestatistician.com/the-analysis-of-adverse-events-done-right-savvy/.
  - 200th episode with 10% most downloaded podcasts: https://theeffectivestatistician.com/200th-episode/.

Extended version of this talk, incl. recording (BBS talk from earlier this year):

www.kasparrufibach

Thank you for your attention.

kaspar.rufibach@roche.com

http://www.kasparrufibach.ch

References I

- Aalen, O., Borgan, O. and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.
- Allignol, A., Schumacher, M., Wanner, C., Drechsler, C. and Beyersmann, J. (2011). Understanding competing risks: a simulation point of view. *BMC medical research methodology* **11** 86.
- Andersen, P. K., Borgan, O., Hjort, N. L., Arjas, E., Stene, J. and Aalen, O. (1985). Counting process models for life history data: A review [with discussion and reply]. *Scandinavian Journal of Statistics* **12** 97–158.
- Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine* **31** 1074–1088. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4385
- Beyersmann, J., Allignol, A. and Schumacher, M. (2012). *Competing Risks and Multistate Models with R*. Springer.
- Beyersmann, J., Friede, T. and Schmoor, C. (2020). Design aspects of covid-19 treatment trials: Improving probability and time of favourable events.
- Bühler, A., Cook, R. J. and Lawless, J. F. (2022). Multistate models as a framework for estimand specification in clinical trials of complex processes. *Statistics in Medicine* **n/a**. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9675
- Conner, S. C. and Trinquart, L. (2021). Estimation and modeling of the restricted mean time lost in the presence of competing risks. *Statistics in medicine* **40** 2177–2196.
- Gooley, T. A., Leisenring, W., Crowley, J. and Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med* **18** 695–706.
- ICH (2019). Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R1). https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf.
- IQWiG (2017). General Methods, Version 5.0. Institute of Quality and Efficiency in Health Care. https://www.iqwig.de/en/methods/methods-paper.3020.html

References II

- Latouche, A., Allignol, A., Beyersmann, J., Labopin, M. and Fine, J. P. (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *J Clin Epidemiol* **66** 648–653.
- Marcus, R., Davies, A., Ando, K., Klapper, W., Opat, S., Owen, C., Phillips, E., Sangha, R., Schlag, R., Seymour, J. F., Townsend, W., Trneny, M., Wenger, M., Fingerle-Rowson, G., Rufibach, K., Moore, T., Herold, M. and Hiddemann, W. (2017). Obinutuzumab for the First-Line Treatment of Follicular Lymphoma. *N. Engl. J. Med.* **377** 1331–1344.
- McCaw, Z. R., Tian, L., Vassy, J. L., Ritchie, C. S., Lee, C. C., Kim, D. H. and Wei, L. J. (2020). How to Quantify and Interpret Treatment Effects in Comparative Clinical Studies of COVID-19. *Ann Intern Med* **173** 632–637.
- Peters, S., Camidge, D. R., Shaw, A. T., Gadgeel, S., Ahn, J. S., Kim, D.-W., Ou, S.-H. I., Pérol, M., Dziadziuszko, R., Rosell, R., Zeaiter, A., Mitry, E., Golding, S., Balas, B., Noe, J., Morcos, P. N., Mok, T. and Investigators, A. T. (2017). Alectinib versus crizotinib in untreated alk-positive non-small-cell lung cancer. *The New England journal of medicine* **377** 829–838.
- Putter, H., Fiocco, M. and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* **26** 2389–2430.
- Rufibach, K., Stegherr, R., Schmoor, C., Jehl, V., Allignol, A., Boeckenhoff, A., Dunger-Baldauf, C., Eisele, L., Künzel, T., Kupas, K., Friedhelm, L., Trampisch, M., Zhao, Y., Friede, T. and Beyersmann, J. (2022). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY) – comparison of adverse event risks in randomized controlled trials. *Statistics in Biopharmaceutical Research, accepted* .
  `https://arxiv.org/abs/2008.07881`
- Schumacher, M., Ohneberg, K. and Beyersmann, J. (2016). Competing risk bias was common in a prominent medical journal. *Journal of clinical epidemiology* **80** 135–136.
- Stegherr, R., Beyersmann, J., Jehl, V., Rufibach, K., Leverkus, F., Schmoor, C. and Friede, T. (2021a). Survival analysis for adverse events with varying follow-up times (savvy): Rationale and statistical concept of a meta-analytic study. *Biometrical journal. Biometrische Zeitschrift* **63** 650–670.
- Stegherr, R., Schmoor, C., Beyersmann, J., Rufibach, K., Jehl, V., Brückner, A., Eisele, L., Künzel, T., Kupas, K., Langer, F., Leverkus, F., Loos, A., Norenberg, C., Voss, F. and Friede, T. (2021b). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY)-estimation of adverse event risks. *Trials* **22** 420.

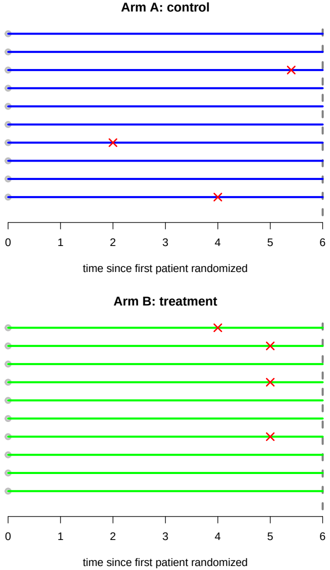
References III

- Stegherr, R., Schmoor, C., Lübbert, M., Friede, T. and Beyersmann, J. (2021c). Estimating and comparing adverse event probabilities in the presence of varying follow-up times and competing events. *Pharm Stat* **20** 1125–1146.
- Varadhan, R., Weiss, C. O., Segal, J. B., Wu, A. W., Scharfstein, D. and Boyd, C. (2010). Evaluating health outcomes in the presence of competing risks: a review of statistical methods and clinical applications. *Medical care* **48** S96–105.
- Putter, H., Stensrud, M. J., Tchetgen Tchetgen, E. J. and Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat Med* **39** 1199–1236.

Backup

Treatment works

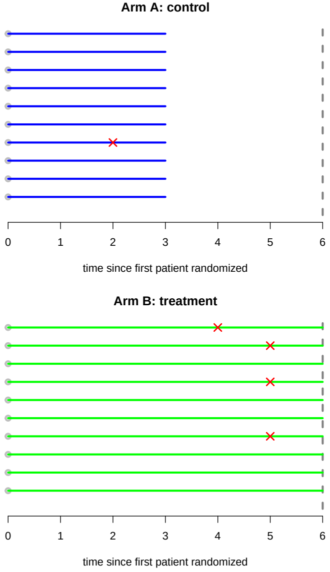
Estimation of P(AE)



- 2-arm RCT.
- 10 patients per arm.
- All patients randomized on same day.
- All patients observed for 6 months.

$P(\text{AE in A}) = 3 \;/\; 10 = 0.30,$
$P(\text{AE in B}) = 4 \;/\; 10 = 0.40.$

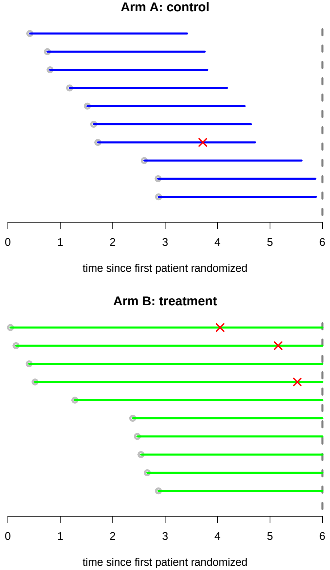
Estimation of P(AE): treatment works



- 2-arm RCT.
- 10 patients per arm.
- All patients randomized on same day.
- Hazard ratio for PFS = 0.5, stop AE recording after PFS event.

P(AE in A) = 1 / 10 = 0.10,
P(AE in B) = 4 / 10 = 0.40.

Estimation of P(AE): treatment works + staggered entry



- 2-arm RCT.
- 10 patients per arm.
- Patients enter trial over time.
- All patients observed until cutoff.
- Hazard ratio for PFS = 0.5, stop AE recording after PFS event.

P(AE in A) = 1 / 10 = 0.10,
P(AE in B) = 4 / 10 = 0.40.

Before you ask...

Before you ask...

Focus on bias - what about variability?

- Focus today with IP rarely on variability either!
- Simulation study for 2-arm comparisons: Stegherr et al. (2021c).

We do not collect data necessary to estimate P(AE) with AJE?

- ICH E9(R1) estimands addendum: **clinical trial objective** dictates data collection and analytical method!
- Clarify **clinical trial objective** also for analysis of safety!
- **Proper definition of CE** requires understanding and discussion of therapeutic area.

Before you ask...

Does normalization by exposure time not solve the problem?

- **Incidence density**. See backup for details.
- A priori estimates **AE hazard**, not P(AE). Can be turned into estimator of P(AE).
- Assumes **exponentiality** of AE hazard.
- Incidence density for each CE.

Can we use IP for "signal detection" or other purposes?

> Biases = statistical properties of IP, (1 - KM).
>
> Independent of what we use estimates of P(AE) for!

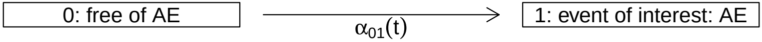
Causality

Aalen-Johansen:

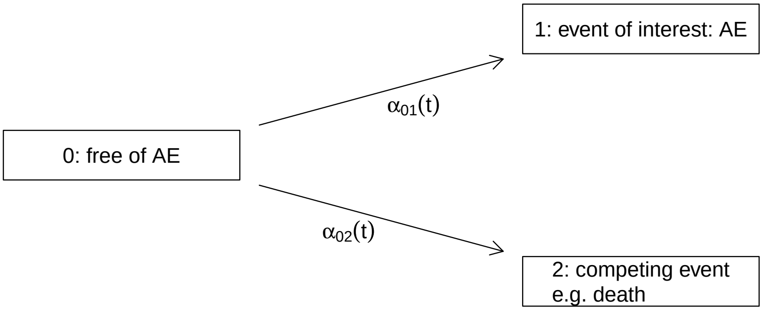
- Estimates cumulative incidence function.
- **Censoring**: if random, e.g. administrative censoring $\Rightarrow$ does not destroy causal interpretation.
- Competing events: intervention on observation process differs from intervention affecting the patient. Young et al. (2020), Rufibach et al. (2022).

Competing risks and the estimand addendum

One event – time to AE



Add competing event



Competing event vs. intercurrent event

Definition **competing event**, Gooley et al. (1999):

> *We shall define a* **competing risk** *as an event whose occurrence either precludes the occurrence of another event under examination or fundamentally alters the probability of occurrence of this other event.*

Definition **intercurrent event**, ICH (2019):

> *Events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest.*

Intercurrent event definition $\approx$ competing event definition.

ICH (2019) does not say anything about competing risks though.

Death: competing risk + intercurrent event (?).

Clinical questions of interest and their estimators

Extending Table 1 in Varadhan et al. (2010).

Clinical question	Target of inference	Estimator	Comment
What is hazard / probability of AE or death, whatever happens earlier?	Event-free survival ("composite")	Kaplan-Meier	1to1 correspondence between hazard and probability.
What is hazard / probability of AE, accounting for the possibility that patients may die before experiencing an AE?	Cause-specific hazards	Nelson-Aalen	- Key measure to compare groups in RCT. - Evaluate impact of risk factors.
	Cumulative incidence	Aalen-Johansen	- Interest in absolute risk ("probability"). - Benefit-risk of an intervention.
What is hazard / probability of AE in world where patients would not die?	Survival function ("hypothetical")	1 - KM with censoring deaths	- **Rarely (to say the least) of clinical interest.** **- Maybe for other CEs.** **- Estimation: assumption about "independence" of competing events - neither sensible nor needed!**

Did we get our clinical questions answered?

Yes!

Did we need ICH E9(R1) language or strategies?

No!

Conclusions:

Clearly formulate clinical question.

None of the five strategies in the addendum needed to model competing risk.

Random variable vs. stochastic process formulation

Endpoints like OS: model using **random variable** X with CDF F, hazard h, etc.

Competing risk, multistate models:

- Avoid random variables: temptation of latent failure time models (backup).
- Use **stochastic process** formulation, see e.g. Beyersmann et al. (2012):
  - $X(t) \in \{0, 1, 2\}$, $t \geq 0$: state occupied by individual at time $t \geq 0$.
  - $X(t) = j$ if event j has occurred in $[0, t]$.
  - $T := \inf\{t \ : \ X_t \neq 0\}$, X_T = state occupied at T.
  - Competing risk data: (T, X_T).

Andersen et al. (1985):

> *In life history analysis, time and random phenomena occurring in time play an essential role, and it seems therefore more natural to study life history analysis in terms of the theory of* **stochastic processes***. Thus, the formulation in terms of random variables may have contributed to hampering the researchers working in the field of survival analysis, or failure time analysis, from extending their otherwise fine methodology to more general life history models.*

Marry competing risk with ICH E9(R1) **if you must**

Definition of **variable** in ICH E9(R1) addendum:

> *The variable (or endpoint) to be obtained for each patient that is required to address the clinical question.*

No one says this must be **univariate**!

Marry competing risk with ICH E9(R1) **if you must**:

Attribute	**Definition**
Treatment	generic
Population	generic
Variable	(T, X_T)
Intercurrent event(s)	None left from competing risk, maybe others.
Summary measure	Depends on clinical question: hazard ratio, cumulative incidence.

Alternative proposal for general estimands for MSMs: Bühler et al. (2022).

Competing risk models: population quantities

"Cause-specific survival function":

$$S_k(t) \quad = \quad \exp[A_{0j}(t)].$$

- S_k is **NOT** marginal survival function!
- Only has this interpretation if competing event time distributions and censoring distribution are **independent**.
- Then marginal distribution describes event time distribution in world where competing events do not occur.

Competing risk models: hazard vs. probability

Transition probabilities in general multistate models:

$$P_{lj}(s,t) \quad := \quad P(X(t) = j | X(s) = l, \text{Past}).$$

Competing risk:

- $P_{0j}(0,t)$ referred to as **cumulative incidence**.
- Expected proportion of patients experiencing event of type j over course of time.

Cumulative incidence for $j = 1, 2$:

$$\begin{aligned}
P(T \leq t, X_T = j) &= P_{0j}(0,t) \\
&= P(X(t) = j | X(0) = 0) \\
&= \int_0^t P(T > v-)\alpha_{0j}(v)dv \\
&= \int_0^t \exp\Big(-A_{01}(v-) - A_{02}(v-)\Big)\alpha_{0j}(v)dv.
\end{aligned}$$

Competing risk models: population quantities

How is competing risk data generated? Two-step simulation process:

1. Determine time T at which event occurs via all-cause hazard $\alpha(t)$.
2. Event type X_T for given time T: determined via multinomial experiment that decides with probability $\alpha_{0j}(T)/\alpha(T)$ on $X_T = j$.

Beyersmann et al. (2012), Allignol et al. (2011).

Hazards completely determine stochastic behaviour of competing risks process.

Doing now what patients need next

R version and packages used to generate these slides:

R version: R version 4.2.3 (2023-03-15 ucrt)

Base packages: stats / graphics / grDevices / utils / datasets / methods / base

Other packages: ggplot2 / etm / cmprsk / mvna / prodlim / survival / reporttools / xtable

This document was generated on 2023-08-29 at 07:42:20.