Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): summary of findings and assessment of existing guidelines

Regina Stegherr

Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin

22.05.2025

Acknowledgements

- SAVVY consortium: specifically Kaspar Rufibach (also for part of the slides), Jan Beyersmann, Claudia Schmoor, Tim Friede.
- My part of this work was conducted at Ulm University as part of my PhD.

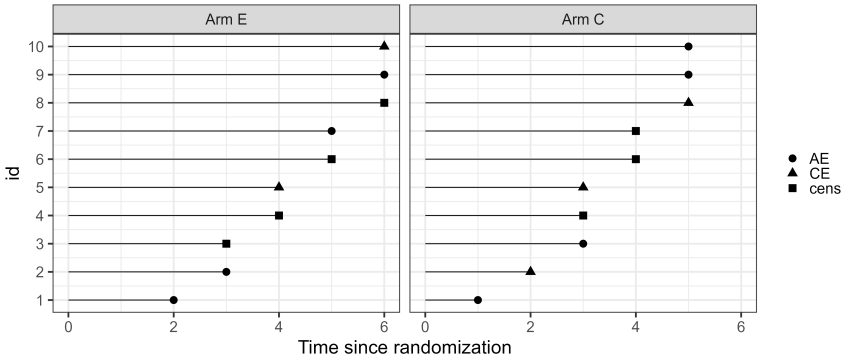
Take home messages

1. Need accurate estimates of Adverse Event (AE) probability and comparison between arms.
2. Incidence Proportion and (1-Kaplan-Meier) **biased** regardless of what we use them for.
3. Bias "does not cancel out" when comparing AE probabilities between arms in randomized clinical trials.

Agenda

1. Take home messages
2. Estimation of the risk of an AE
   - Example
   - The SAVVY project
   - Estimators of the AE probability
   - Example
   - Bias of common estimators of AE probability
   - Bias of common estimators of relative AE risk
3. Assessment of existing guidelines
4. Take home messages
5. Resources and future plans

Example



Challenges

- Varying follow-up times
- Censoring
- Competing Events

The SAVVY project

Survival analysis for **A**d**V**erse events with **V**ar**Y**inf follow-up times:

Goal: improve analyses and reporting of AE data in clinical trial through use of **survival techniques** appropriately dealing with

- varying follow-up times between arms
- censoring
- competing events

The SAVVY consortium

9 pharma + 3 universities


The Empirical Study of the SAVVY project

Federated learning: central analysis team:

- Developed macors (`R` + SAS). Now included in `savvyr` `R`-package
- Analyses on individual patient data ran by the sponsor.
- Only aggregated data was shared.
- Central team performed meta-analysis

The Empirical Study of the SAVVY project

Federated learning: central analysis team:

- Developed macors (R + SAS). Now included in `savvyr` R-package
- Analyses on individual patient data ran by the sponsor.
- Only aggregated data was shared.
- Central team performed meta-analysis

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

RCTs with 200 - 7171 patients included.

186 AEs: selected by sponsor.

The SAVVY project - scientific questions

1. For estimation of **AE probability**, how biased are commonly used estimators (especially incidence proportion and 1-Kaplan-Meier) in presence of **censoring, varying follow-up times, competing events** and (in case of incidence densities) a **restrictive parametric model**?

The SAVVY project - scientific questions

1. For estimation of **AE probability**, how biased are commonly used estimators (especially incidence proportion and 1-Kaplan-Meier) in presence of **censoring, varying follow-up times, competing events** and (in case of incidence densities) a **restrictive parametric model**?
2. What is the **bias of common estimators that quantify the relative risk of experiencing an AE** between two treatment arms in a RCT?

The SAVVY project - scientific questions

1. For estimation of **AE probability**, how biased are commonly used estimators (especially incidence proportion and 1-Kaplan-Meier) in presence of **censoring, varying follow-up times, competing events** and (in case of incidence densities) a **restrictive parametric model**?
2. What is the **bias of common estimators that quantify the relative risk of experiencing an AE** between two treatment arms in a RCT?
3. Can trial characteristics be identified that help **explain the bias** in estimators?

The SAVVY project - scientific questions

1. For estimation of **AE probability**, how biased are commonly used estimators (especially incidence proportion and 1-Kaplan-Meier) in presence of **censoring, varying follow-up times, competing events** and (in case of incidence densities) a **restrictive parametric model**?
2. What is the **bias of common estimators that quantify the relative risk of experiencing an AE** between two treatment arms in a RCT?
3. Can trial characteristics be identified that help **explain the bias** in estimators?
4. How does the use of potentially biased estimators **impact quantification of AE probabilities** and relative effects in regulatory settings?

The Empirical Study of the SAVVY project

Competing Events (CEs) in SAVVY:

- **Death**: AE after death impossible.
- **All-cause**: Death, loss to follow-up, withdrawal of consent, treatment discontinuation, and progression: CEs after which AE is in principle still possible but not observed due to premature end of follow-up

The Empirical Study of the SAVVY project

Competing Events (CEs) in SAVVY:

- **Death**: AE after death impossible.
- **All-cause**: Death, loss to follow-up, withdrawal of consent, treatment discontinuation, and progression: CEs after which AE is in principle still possible but not observed due to premature end of follow-up

Censoring: designated end of follow-up reached without having an AE or CE; administrative not triggered by course of disease

Incidence proportion

- Incidence proportion in experimental arm in interval $[0, t]$

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

- $\widehat{\text{IP}}_E(t)$ estimates P(AE happens in $[0, t]$ and AE is observed **before censoring**)
- Identical follow-up times in all patients are assumed.
- $\widehat{\text{IP}}_E(t) \leq \widehat{P}(\text{AE happens in } [0, t]) \Rightarrow \widehat{\text{IP}}_E(t)$ **underestimates** AE probability

1-Kaplan-Meier

- **Time-to-first-AE** considered.
- Estimates P(AE happens in $[0, t]$).
- Correctly accounts for **censoring**.
- **Competing events censored** at their event times $\Rightarrow$ **Overestimation** of the AE probability

Incidence density

- Incidence density in experimental arm in interval $[0, t]$

$$\text{ID}_E(t) = \frac{\text{Number of patients with observed AE in } [0, t]}{\text{patient-time at risk restricted by } t}$$

- Estimator of hazard rate.
- Parametric assumption of constant hazards.
- Probability scale transformation: $1 - \exp(-\text{ID}_E(t) \cdot t)$ (**parametric version of 1-Kaplan-Meier**).

Incidence density

- Incidence density in experimental arm in interval $[0, t]$

$$\text{ID}_E(t) = \frac{\text{Number of patients with observed AE in } [0, t]}{\text{patient-time at risk restricted by } t}$$

- Estimator of hazard rate.
- Parametric assumption of constant hazards.
- Probability scale transformation: $1 - \exp(-\text{ID}_E(t) \cdot t)$ (**parametric version of 1-Kaplan-Meier**).
- Probability transform accounting for competing events:

$$\frac{\text{ID}_E(t)}{\text{ID}_E(t) + \overline{\text{ID}}_E(t)} \left(1 - \exp\left(-t \cdot \left(\text{ID}_E(t) + \overline{\text{ID}}_E(t)\right)\right)\right)$$

with $\overline{\text{ID}}_E(t) = \frac{\text{Number of patients with observed CE in } [0, t]}{\text{patient-time at risk restricted by } t}$

Aalen-Johansen estimator (AJE)

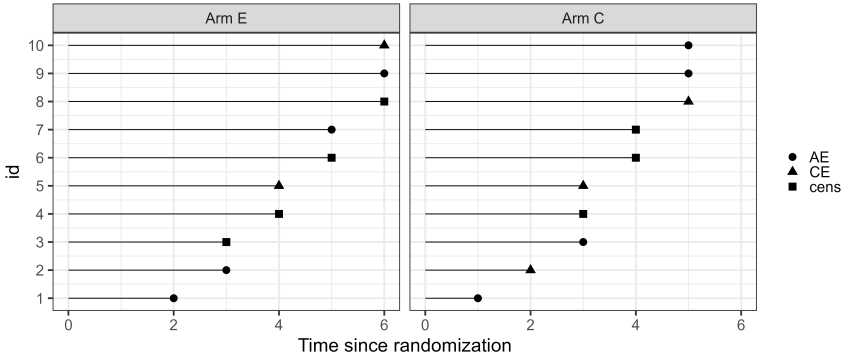
$$CIF(\tau) = \sum_{u \in (0,\tau]} \prod_{v \in (0,u)} \left(1 - \Delta\hat{\Lambda}(v) - \Delta\hat{\tilde{\Lambda}}(v)\right) \Delta\hat{\Lambda}(u)$$

- Generalizes Kaplan-Meier to multiple event types.
- Accounts for **censoring** and **competing events**.
- Non-parametric estimator.
- **Gold-standard** in our comparisons.

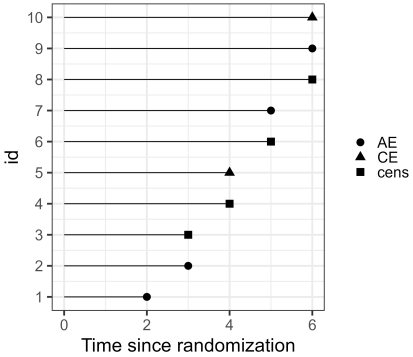
Possible sources of bias

	Accounts for censoring	Makes no constant hazard assumption	Accounts for CE
Incidence proportion	No	Yes	Yes
1-Kaplan-Meier	Yes	Yes	No
Probability transform incidence density ignoring CEs	Yes	No (AE hazard)	No
Probability transform incidence density accounting for CEs	Yes	No (AE and CE hazard)	Yes
death only Aalen-Johansen	Yes	Yes	Yes (death only)
Aalen-Johansen	**Yes**	**Yes**	**Yes**

Example: AE probability estimation



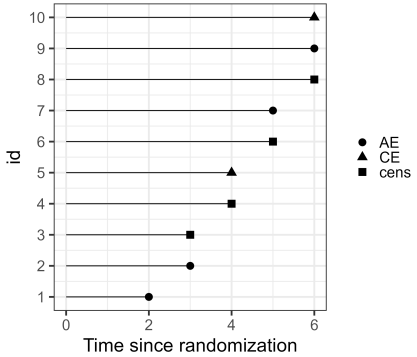
Example: AE probability estimation in Arm E



Estimation at maximal follow-up (time 6):

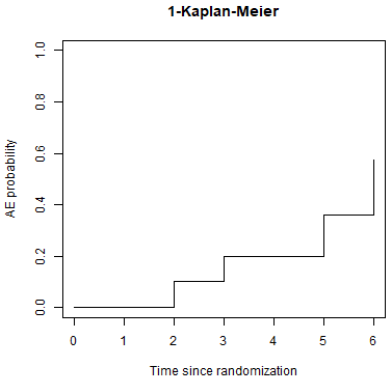
- Incidence proportion: $4/10 = 0.4$

Example: AE probability estimation in Arm E

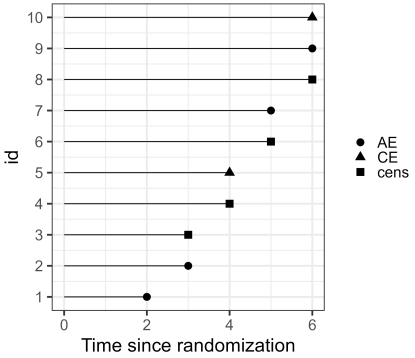


Estimation at maximal follow-up (time 6):

- Incidence proportion: $4/10 = 0.4$
- 1-Kaplan-Meier: 0.57



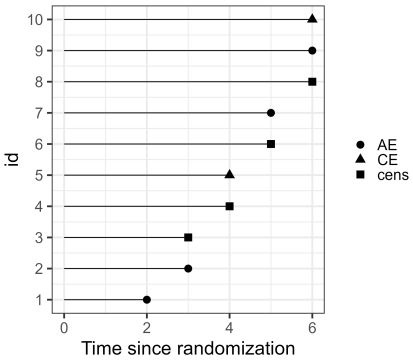
Example: AE probability estimation in Arm E



Estimation at maximal follow-up (time 6):

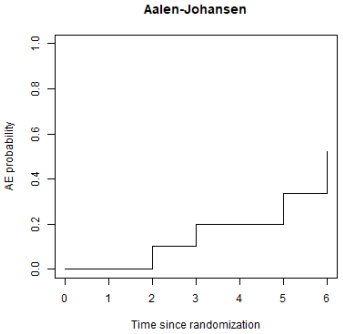
- Incidence proportion: $4/10 = 0.4$
- 1-Kaplan-Meier: 0.57
- Probability transform incidence density
  - ignoring CE: 0.42
  - accounting for CE: 0.37

Example: AE probability estimation in Arm E

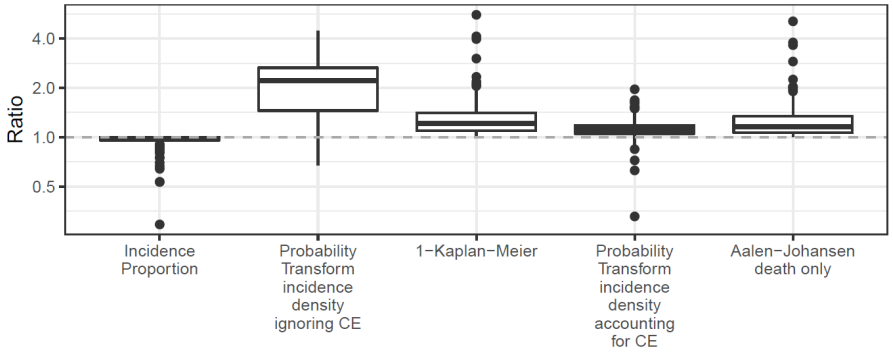


Estimation at maximal follow-up (time 6):

- Incidence proportion: $4/10 = 0.4$
- 1-Kaplan-Meier: 0.57
- Aalen-Johansen estimator: 0.52



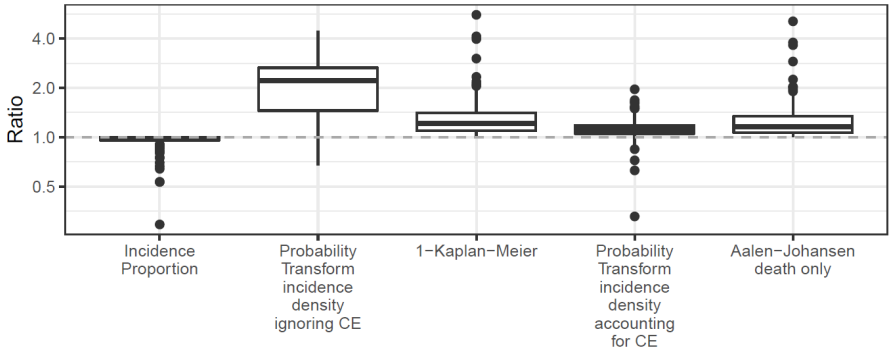
Ratio of estimators to AJE at maximal observed time



Incidence proportion:

- **Underestimation** of AE probability **up to the factor THREE!**
- Datasets have many CEs & few censoring ⇒ overall performance not bad

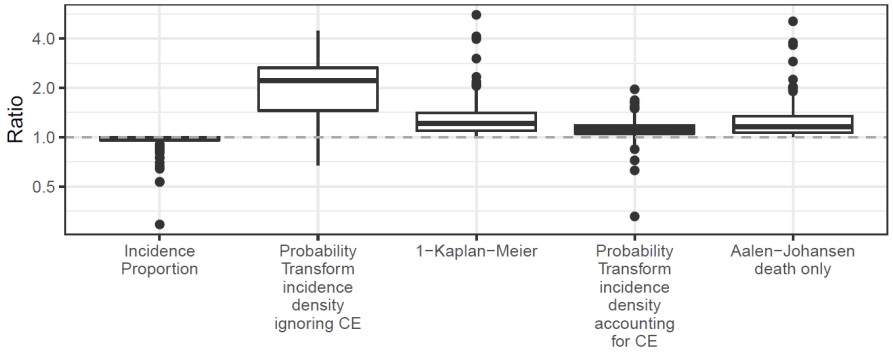
Ratio of estimators to AJE at maximal observed time



1-Kaplan-Meier:

- **Overestimation** of AE probability **up to the factor FIVE!**
- Accounts for censoring but not CEs.

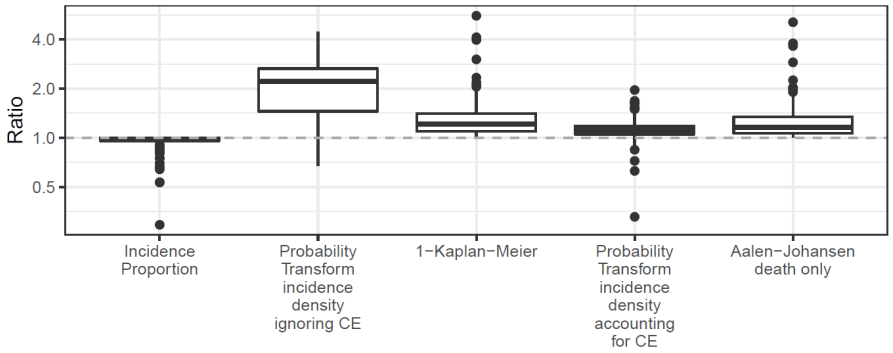
Ratio of estimators to AJE at maximal observed time



Probability transforms incidence density:

- Accounting for CE important.
- Parametric Kaplan-Meier on average worse than non-parametric one.

Ratio of estimators to AJE at maximal observed time



Death only Aalen-Johansen:

- Definition of CEs important but disease specific $\Rightarrow$ impacts amount of censoring and competing events

Comparison of frequency categories

- Incidence proportion vs gold-standard Aalen-Johansen estimator

		gold-standard Aalen-Johansen				
		very rare (<0.01%)	rare (<0.1%)	uncommon (<1%)	common (<10%)	very common (>=10%)
IP	very rare	**6**	0	0	0	0
	rare	0	**0**	0	0	0
	uncommon	0	0	**6**	0	0
	common	0	0	0	**86**	2
	very common	0	0	0	0	**86**

Comparison of frequency categories

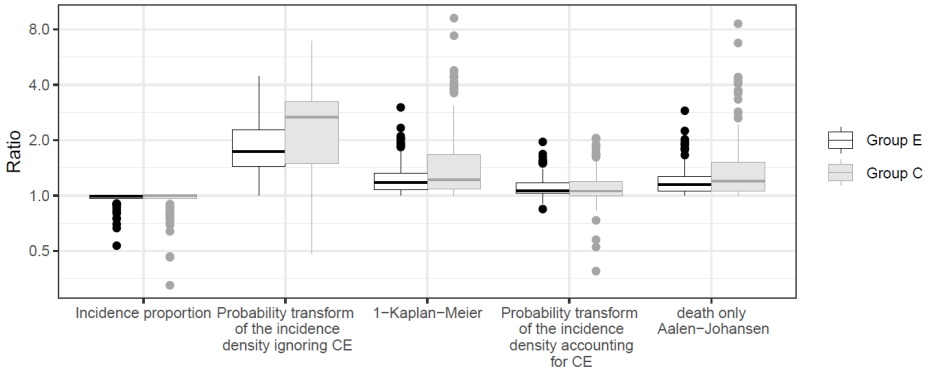
- Incidence proportion vs gold-standard Aalen-Johansen estimator

		gold-standard Aalen-Johansen				
		very rare (<0.01%)	rare (<0.1%)	uncommon (<1%)	common (<10%)	very common (>=10%)
IP	very rare	**6**	0	0	0	0
	rare	0	**0**	0	0	0
	uncommon	0	0	**6**	0	0
	common	0	0	0	**86**	2
	very common	0	0	0	0	**86**

- 1-Kaplan-Meier vs gold-standard Aalen-Johansen estimator

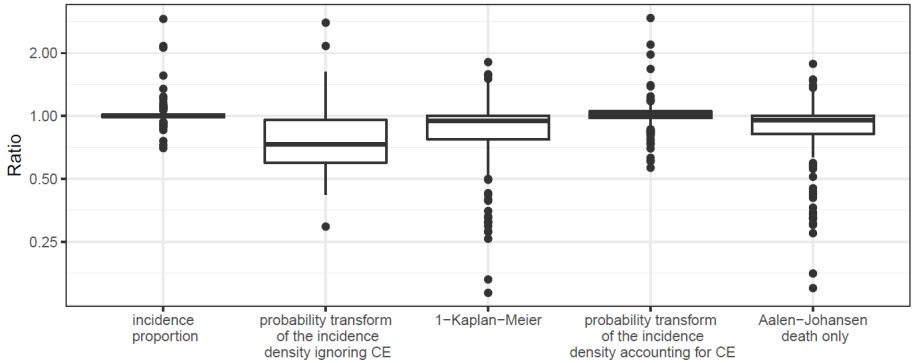
		gold-standard Aalen-Johansen				
		very rare (<0.01%)	rare (<0.1%)	uncommon (<1%)	common (<10%)	very common (>=10%)
1-KM	very rare	**6**	0	0	0	0
	rare	0	**0**	0	0	0
	uncommon	0	0	**4**	0	0
	common	0	0	2	**72**	0
	very common	0	0	0	14	**88**

Ratio of estimators to AJE per arm at common maximal follow-up



- As percentage of AEs, CEs and censoring in both arms differ, over-/underestimation different in both arms

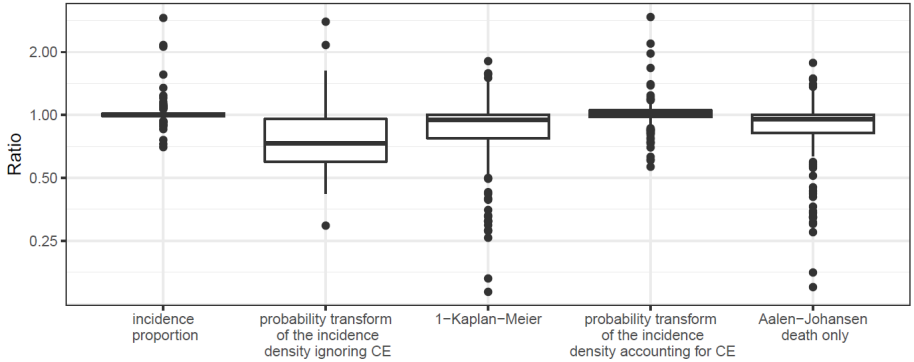
Ratio of RR of estimators to RR of AJE at common maximal follow-up



Incidence proportion:

- Over- and underestimation observed.
- **Overestimation** of relative risk **up to the factor of almost 3.**

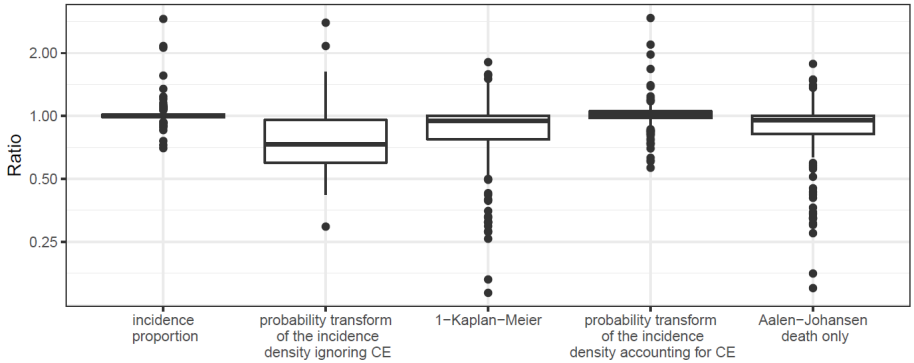
Ratio of RR of estimators to RR of AJE at common maximal follow-up



1-Kaplan-Meier:

- Over- and underestimation observed.
- **Underestimation** of relative risk **up to the factor of >4**

Ratio of RR of estimators to RR of AJE at common maximal follow-up



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

Impact on quantification of relative effects

- Categorization motivated by IQWiG General Methods Version 5.0 (2017)
  (0) no effect: $1 \in$ CI
  (a) minor: RR<1 & $\text{CI}_{\text{upper}} \in [0.9, 1)$ or RR>1 & $\text{CI}_{\text{lower}} \in (1, 1.11]$
  (b) considerable: RR<1 & $\text{CI}_{\text{upper}} \in [0.75, 0.9)$ or RR>1 & $\text{CI}_{\text{lower}} \in (1.11, 1.33]$
  (c) major: RR<1 & $\text{CI}_{\text{upper}} < 0.75$ or RR>1 & $\text{CI}_{\text{lower}} > 1.33$

Impact on quantification of relative effects

- Categorization motivated by IQWiG General Methods Version 5.0 (2017)
  - (0) no effect: $1 \in$ CI
  - (a) minor: RR<1 & $\text{CI}_{\text{upper}} \in [0.9, 1)$ or RR>1 & $\text{CI}_{\text{lower}} \in (1, 1.11]$
  - (b) considerable: RR<1 & $\text{CI}_{\text{upper}} \in [0.75, 0.9)$ or RR>1 & $\text{CI}_{\text{lower}} \in (1.11, 1.33]$
  - (c) major: RR<1 & $\text{CI}_{\text{upper}} < 0.75$ or RR>1 & $\text{CI}_{\text{lower}} > 1.33$

		gold-standard Aalen-Johansen			
		(0) no effect	(a) minor	(b) considerable	(c) major
incidence proportion	(0) no effect	**84**	5		
	(a) minor	3	**10**	2	
	(b) considerable	1	2	**12**	2
	(c) major	1		1	**33**
1-Kaplan-Meier	(0) no effect	**84**	9	4	8
	(a) minor	3	**6**	3	3
	(b) considerable	2	1	**7**	5
	(c) major		1	1	**18**

Comparison of two gold-standards

- Compare conclusions of the RR calculated with the Aalen-Johansen estimator (all events) to conclusions of the hazard ratio calculated with the Cox model.

		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**

- Different estimands: Cox HR - relative effect based on AE hazard, RR Aalen-Johansen - based on probabilities
- Hazard of CE also with impact on Aalen-Johansen estimator
- **For complete picuture: Report both relative risk based on Aalen-Johansen and hazard ratio of Cox model**

Can trial characteristics be identified that help explain the bias in estimators?

- Meta-regression identified
  - Amount of censoring
  - Amount of competing events

  as leading factors of the bias
- However, mathematical considerations also show that the timing of the events is important (e.g., censoring after last AE does not impact the bias of the incidence proportion).

Guidelines

Guideline	Acknowledges varying FU	Proposes incidence density	Acknowledges constant hazard assumption	Proposes life-table/ 1-KM	Acknowledges CEs
ICH E3	x			x	
ICH E9	x	x		x	
SmPC	x	x			
CIOMS	x	x	x	x	
FDA premarketing	x	x	x	x	
Consort HARM 2022 update	x				

Take home messages

1. Need accurate estimates of AE probability and comparison between arms.
2. Incidence Proportion and (1-Kaplan-Meier) **biased** regardless of what we use them for.
3. Bias "does not cancel out" when comparing AE probabilities between arms in randomized clinical trials.

Resources

SAVVY webpage: `https://numbersman77.github.io/savvy/`

- Exemplary code for all methods.
- All papers and talks.
- Papers:
  - SAP: Stegherr et al. (2021, Biometrical Journal)
  - Methods: Stegherr et al. (2021, Pharmaceutical Statistics)
  - 1-sample: Stegherr et al. (2021, Trials)
  - 2-sample: Rufibach et al. (2023, Statistics in Biopharmaceutical Research)
  - Summary: Rufibach et al. (2024, Trials)

Future plan

- Estimate disease-specific AE probabilities, properly discussing therapeutic area specific competing events
- Influence updating of guidelines

Thank you for your attention.

regina.stegherr@charite.de

Estimand

- Time-to-1st-event and type-of-1st-event, i.e., no AEs after treatment discontinuation are considered
- All six estimators target the **same estimand** (understood as population quantity): the probability $P(\text{AE in } [0, t])$
- In simple situations without censoring or varying follow-up times, i.e., when all patients are observed the same amount of time, $P(\text{AE in } [0, t])$ can easily be estimated by the incidence proportion
- But as soon as varying follow-up times and/or censoring are present, the **incidence proportion will be biased**
- We did not attempt to define what a fit-for-purpose estimand to quantify the safety risk could be
- Focus is on the **statistical properties of commonly used estimators** in presence of varying follow-up and CEs

Estimand

Not entirely clear how to combine ICH E9(R1) addendum and competing events

- Varadhan et al. (2010) considered estimands in presence of competing events in sense of population quantities
- Five attributes of SAVVY target of estimation within the ICH E9(R1) estimand framework

Estimand

Not entirely clear how to combine ICH E9(R1) addendum and competing events

- Varadhan et al. (2010) considered estimands in presence of competing events in sense of population quantities
- Five attributes of SAVVY target of estimation within the ICH E9(R1) estimand framework
  - *Treatment*: generic
  - *Population*: generic
  - *Variable/endpoint*: Time-to-1st-event (composite of AE and CE), with indication of type of event (stochastic process formulation)
  - *Summary measure*: arm-wise probabilities $P(\text{AE in } [0, t])$ (one sample), respectively the relative risk of arm-wise probabilities $P(\text{AE in } [0, t])$
  - *Intercurrent events*: CEs do not affect the existence of the measurements because different CEs are simply different values of precisely one random variable; One could argue that CEs are thus simply made part of the variable attribute of the estimand

Competing event vs. intercurrent

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 (?).