Survival analysis for AdVerse events with VarYing follow-up times - Results of the empirical study of the SAVVY project

Regina Stegherr[1], on behalf of the SAVVY Study Group

[1] This work was conducted at the Institute of Statistics, Ulm University, Germany, and is not in any way related to my current work at the “Institut für Klinische Pharmakologie und Toxikologie, Pharmakovigilanz, und Beratungszentrum für Embryonaltoxikologie, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin”, Germany

30.08.2022

regina.stegherr@charite.de

Thanks to...

The SAVVY project group

- Academic leads: Jan Beyersmann (Ulm), Tim Friede (Göttingen) and Claudia Schmoor (Freiburg)
- Steering Committee: Valentine Jehl (Novartis), Friedhelm Leverkus (Pfizer), Kaspar Rufibach (Roche) and the academic leads
- Participating companies: Bayer, Boehringer Ingelheim, BMS, Janssen, Lilly, Merck, Novartis, Pfizer, Roche

Survival analysis for AdVerse events with VarYing follow-up times - SAVVY

- **Safety in terms of adverse events (AEs)** is a relevant aspect of risk-benefit assessent of therapies (Unkel et al., 2019).
- In analyses of AEs (of a certain kind), observation may be precluded by **death, progression** or some other competing event. Moreover, recording of AEs is limited to a restricted period of time (**censoring**) and varying follow-up times (Allignol et al., 2016).
- Overall aim: Improve reporting of AEs through the use of survival techniques appropriately dealing with varying follow-up times and competing events.
- **Empirical study**: Investigate in several randomized controlled trials whether the use of different estimators for analyses of AEs leads to different conclusions about therapies' safety.
- Comparison of **commonly** used (but biased) **estimators quantifying the AE probability** to estimators **accounting for competing events** in time-to-event studies and also compare safety comparisons between treatment groups.

Estimating AE probabilities: Commonly used but biased methods

- **Incidence proportion**: $\frac{\#\text{ AEs in } [0, \tau]}{\#\text{ patients}}$
  - Usually only calculated at the end of follow-up ⇒ Assumes identical follow-up times in all patients
  - **Underestimation** of AE probability in presence of censoring
- **Incidence density**: $\mathrm{ID}(\tau) = \frac{\#\text{ AE in } [0, \tau]}{\text{patient-time at risk restricted by } \tau}$
  - Assumption of constant hazards
  - Estimator of hazard rate ⇒ probability scale requires transformation: $1 - \exp\left(-\mathrm{ID}(\tau) \cdot \tau\right)$
  - Parametric version of 1-Kaplan-Meier
- **1-Kaplan-Meier**: competing events censored at their event time
  - **Overestimation** of AE probability in presence of competing events
  - About 50% of all Kaplan-Meier curves ignore competing risks (van Walraven et al. 2016, Schumacher et al. 2016)
  - But health technology assessment agencies, e.g., IQWiG, still demand Kaplan-Meier estimates

Estimating AE probabilities: Commonly used but biased methods

- **Incidence proportion**: $\frac{\text{\# AEs in } [0, \tau]}{\text{\# patients}}$
  - Usually only calculated at the end of follow-up $\Rightarrow$ Assumes identical follow-up times in all patients
  - **Underestimation** of AE probability in presence of censoring
- **Incidence density**: $\mathrm{ID}(\tau) = \frac{\text{\# AE in } [0, \tau]}{\text{patient-time at risk restricted by } \tau}$
  - Assumption of constant hazards
  - Estimator of hazard rate $\Rightarrow$ probability scale requires transformation: $1 - \exp\left(-\mathrm{ID}(\tau) \cdot \tau\right)$
  - Parametric version of 1-Kaplan-Meier
- **1-Kaplan-Meier**: competing events censored at their event time
  - **Overestimation** of AE probability in presence of competing events
  - About 50% of all Kaplan-Meier curves ignore competing risks (van Walraven et al. 2016, Schumacher et al. 2016)
  - But health technology assessment agencies, e.g., IQWiG, still demand Kaplan-Meier estimates

Estimating AE probabilities: Commonly used but biased methods

- **Incidence proportion**: $\frac{\text{\# AEs in } [0, \tau]}{\text{\# patients}}$
  - Usually only calculated at the end of follow-up ⇒ Assumes identical follow-up times in all patients
  - **Underestimation** of AE probability in presence of censoring
- **Incidence density**: $\mathrm{ID}(\tau) = \frac{\text{\# AE in } [0, \tau]}{\text{patient-time at risk restricted by } \tau}$
  - Assumption of constant hazards
  - Estimator of hazard rate ⇒ probability scale requires transformation: $1 - \exp\left(-\mathrm{ID}(\tau) \cdot \tau\right)$
  - Parametric version of 1-Kaplan-Meier
- **1-Kaplan-Meier**: competing events censored at their event time
  - **Overestimation** of AE probability in presence of competing events
  - About 50% of all Kaplan-Meier curves ignore competing risks (van Walraven et al. 2016, Schumacher et al. 2016)
  - But health technology assessment agencies, e.g., IQWiG, still demand Kaplan-Meier estimates

Estimating AE probabilities: Alternative, underused approaches

- **Aalen-Johansen estimator**: $\mathrm{CIF}(\tau) = \sum_{u \in (0,\tau]} \prod_{v \in (0,u)} \left(1 - \Delta\hat{\Lambda}(v) - \Delta\hat{\bar{\Lambda}}(v)\right) \Delta\hat{\Lambda}(u)$
  - **Gold-standard**: accounts for censoring and competing events and is not restricted to constant hazards (non-parametric)
  - Generalizes the Kaplan-Meier estimator to multiple event types
- **Probability transform of the incidence density accounting for competing events** (parametric version of Aalen-Johansen): $\dfrac{\mathrm{ID}(\tau)}{\mathrm{ID}(\tau) + \overline{\mathrm{ID}}(\tau)} \left(1 - \exp(-\tau \cdot [\mathrm{ID}(\tau) + \overline{\mathrm{ID}}(\tau)])\right)$

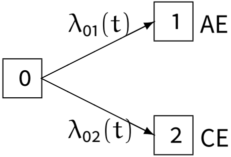
  with $\overline{\mathrm{ID}}(\tau) = \dfrac{\text{\# competing event in } [0,\tau]}{\text{patient-time at risk restricted by } \tau}$
  - Assumption of **constant hazards** for both (AE and competing event) hazards
  - In literature about incidence densities often neglected (e.g. book ‘analysis of incidence rates’ by Cummings, 2019)

Estimating AE probabilities: Alternative, underused approaches

- **Aalen-Johansen estimator**: $\mathrm{CIF}(\tau) = \sum_{u \in (0,\tau]} \prod_{v \in (0,u)} \left(1 - \Delta\hat{\Lambda}(v) - \Delta\hat{\bar{\Lambda}}(v)\right) \Delta\hat{\Lambda}(u)$
  - **Gold-standard**: accounts for censoring and competing events and is not restricted to constant hazards (non-parametric)
  - Generalizes the Kaplan-Meier estimator to multiple event types
- **Probability transform of the incidence density accounting for competing events** (parametric version of Aalen-Johansen): $\frac{\mathrm{ID}(\tau)}{\mathrm{ID}(\tau) + \overline{\mathrm{ID}}(\tau)} \left(1 - \exp(-\tau \cdot [\mathrm{ID}(\tau) + \overline{\mathrm{ID}}(\tau)])\right)$

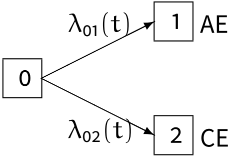
  with $\overline{\mathrm{ID}}(\tau) = \frac{\text{\# competing event in } [0,\tau]}{\text{patient-time at risk restricted by } \tau}$
  - Assumption of **constant hazards** for both (AE and competing event) hazards
  - In literature about incidence densities often neglected (e.g. book ‘analysis of incidence rates’ by Cummings, 2019)

Definition of the competing event



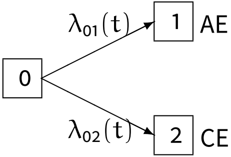
- Time-to-1st-event and type-of-1st-event
- **Adverse event (AE):** Event of interest
- Two possible definitions of a competing event (CE):
  - **Death only**: death without prior AE, i.e., events after which an AE can definitely not occur any more
  - **All events**: death, loss to follow up, withdrawal of consent, treatment discontinuation, and progression, i.e., competing events after which an AE in principle still could occur, but is not observed due to premature end of follow-up
- **Censoring:** designated end of follow-up reached without having an AE or a competing event as defined above; administrative not triggered by course of disease

Definition of the competing event



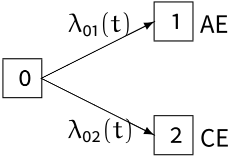
- Time-to-1st-event and type-of-1st-event
- **Adverse event (AE):** Event of interest
- Two possible definitions of a competing event (CE):
  - **Death only**: death without prior AE, i.e., events after which an AE can definitely not occur any more
  - **All events**: death, loss to follow up, withdrawal of consent, treatment discontinuation, and progression, i.e., competing events after which an AE in principle still could occur, but is not observed due to premature end of follow-up
- **Censoring:** designated end of follow-up reached without having an AE or a competing event as defined above; administrative not triggered by course of disease

Definition of the competing event



- Time-to-1st-event and type-of-1st-event
- **Adverse event (AE):** Event of interest
- Two possible definitions of a competing event (CE):
  - **Death only**: death without prior AE, i.e., events after which an AE can definitely not occur any more
  - **All events**: death, loss to follow up, withdrawal of consent, treatment discontinuation, and progression, i.e., competing events after which an AE in principle still could occur, but is not observed due to premature end of follow-up
- **Censoring:** designated end of follow-up reached without having an AE or a competing event as defined above; administrative not triggered by course of disease

Definition of the competing event



- Time-to-1st-event and type-of-1st-event
- **Adverse event (AE):** Event of interest
- Two possible definitions of a competing event (CE):
  - **Death only**: death without prior AE, i.e., events after which an AE can definitely not occur any more
  - **All events**: death, loss to follow up, withdrawal of consent, treatment discontinuation, and progression, i.e., competing events after which an AE in principle still could occur, but is not observed due to premature end of follow-up
- **Censoring:** designated end of follow-up reached without having an AE or a competing event as defined above; administrative not triggered by course of disease

Possible sources of bias

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

Group comparisons and follow-up times

- **Risk difference** or **relative risk** of incidence proportions may be misleading
  - Comparing two quantities that both underestimate the AE probability
  - Comparing two quantities **evaluated at different follow-up times**, i.e., largest observed event time in experimental treatment group E τ_E may be greater/smaller than largest observed event time in comparison group C τ_C (Incidence proportion only calculated at the end of follow-up, Bender et al., 2016) (referred to as maximum follow-up time)
- Evaluate estimators at $\tau = \min(\tau_E, \tau_C)$ (referred to common maximum follow-up time)
- As estimators (e.g. Kaplan-Meier) at the end of follow-up may have **larger variability** due to small numbers still at risk (Pocock et al. 2002):
  - Evaluate estimators at **earlier time point when more patients are still at risk**
  - Evaluate estimators at $\tilde{\tau} = \min(\tilde{\tau}_E, \tilde{\tau}_C)$, with $\tilde{\tau}_E(p)$ and $\tilde{\tau}_C(p)$ defined as event time when $p \cdot 100\%$ of all patients in group E and group C, respectively, are still at risk , e.g., $p = 0.9$ (P90), $p = 0.6$ (P60) and $p = 0.3$ (P30)

Group comparisons and follow-up times

- **Risk difference** or **relative risk** of incidence proportions may be misleading
  - Comparing two quantities that both underestimate the AE probability
  - Comparing two quantities **evaluated at different follow-up times**, i.e., largest observed event time in experimental treatment group E τ_E may be greater/smaller than largest observed event time in comparison group C τ_C (Incidence proportion only calculated at the end of follow-up, Bender et al., 2016) (referred to as maximum follow-up time)
- Evaluate estimators at $\tau = \min(\tau_E, \tau_C)$ (referred to common maximum follow-up time)
- As estimators (e.g. Kaplan-Meier) at the end of follow-up may have **larger variability** due to small numbers still at risk (Pocock et al. 2002):
  - Evaluate estimators at **earlier time point when more patients are still at risk**
  - Evaluate estimators at $\tilde{\tau} = \min(\tilde{\tau}_E, \tilde{\tau}_C)$, with $\tilde{\tau}_E(p)$ and $\tilde{\tau}_C(p)$ defined as event time when $p \cdot 100\%$ of all patients in group E and group C, respectively, are still at risk , e.g., $p = 0.9$ (P90), $p = 0.6$ (P60) and $p = 0.3$ (P30)

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
  - *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

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 $\mathrm{P}(\text{AE in } [0, t])$ (one sample), respectively the relative risk of arm-wise probabilities $\mathrm{P}(\text{AE in } [0, t])$
  - *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

Empirical Study

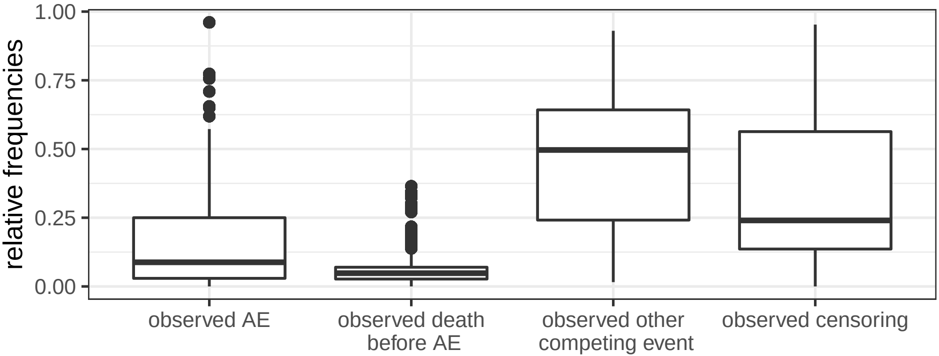
- Statistical Analysis Plan can be found in Stegherr et al. (2021, BiomJ)
- Only **aggregated** data shared with the project collaborators: Trial level analyses ran within the sponsor company / organization using SAS (and R) code provided by the project collaborators ⇒ **no release of individual patient data was required**.
- **Pilot study** to **develop SAS macros**, to **assess feasibility** of macros and output data structure, to **check output dataset** whether they contain all necessary information and to **train meta-analysis** and obtain early results (3 partners providing 5 trials and a total of 62 types AEs (range 2 - 51 per trial))
- **Main study**: 10 participating organizations contributing 17 trials including 186 types of AEs

Empirical study - Collected data

- Study characteristics: indication, severity of AE, comparison type, ...
- AE probability estimates in both groups with variances
- Group comparisons: relative risk (RR) and risk difference, hazard ratio
- Probability estimates of a composite endpoint
- Number of AEs, CEs (death and all events), censoring
- Minimum, median, mean and maximum follow-up time and time of AE (also separate for each group)

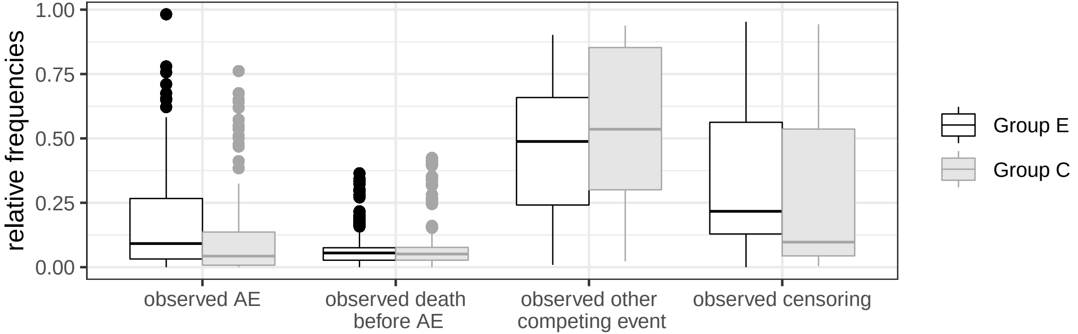
Empirical study - Description

- Twelve (71.6%) of the 17 trials were from oncology
- Nine (52.9%) were actively controlled and eight (47.1%) placebo controlled
- Between 200 and 7171 patients (median 443, IQR: [411, 1134]) were included in the trials
- Median follow-up time in treatment group 927 days (IQR: [449, 1380])
- Relative event frequencies in the experimental treatment group E at the maximum follow-up time



Empirical study - Description

- Twelve (71.6%) of the 17 trials were from oncology
- Nine (52.9%) were actively controlled and eight (47.1%) placebo controlled
- Between 200 and 7171 patients (median 443, IQR: [411, 1134]) were included in the trials
- Median follow-up time in treatment group 927 days (IQR: [449, 1380])
- Relative event frequencies per treatment group at the common maximum follow-up time



Comparison of frequency categories

AE probability in group E at maximum follow-up time

- 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**

Comparison of frequency categories

AE probability in group E at maximum follow-up time

- Incidence proportion vs gold-standard Aalen-Johansen estimator

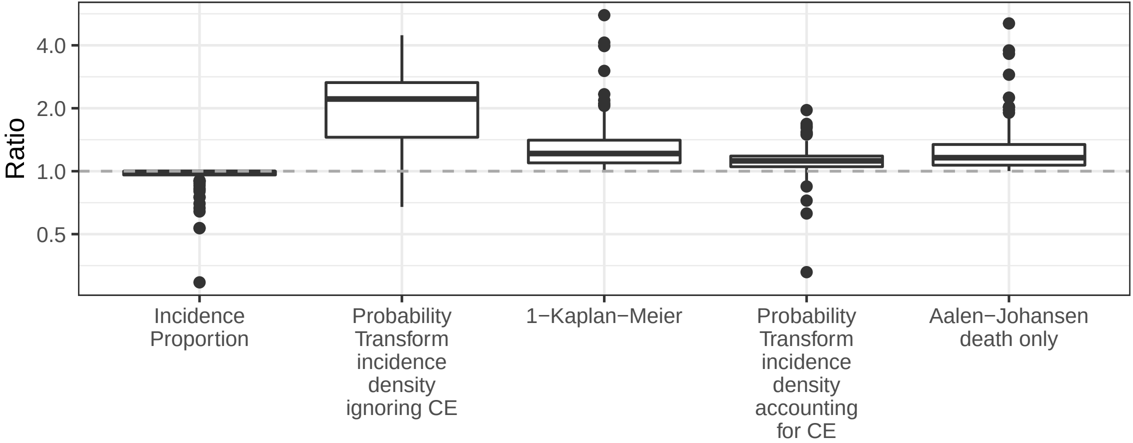
		gold-standard Aalen-Johansen				
		very rare	rare	uncommon	common	very common
		(<0.01%)	(<0.1%)	(<1%)	(<10%)	(>=10%)
IP	very rare	**6**	0	0	0	0
	rare	0	**0**	0	0	0
	uncommon	0	0	**6**	0	0
	common	0	0	0	**86**	2
	very common	0	0	0	0	**86**

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

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

Boxplots of the ratio estimator of interest/gold-standard Aalen-Johansen estimator

AE probability in group E at maximum follow-up time



Meta-analysis

AE probability in group E

Observed data: estimator of log-ratio (log(estimator/gold-standard Aalen-Johansen)) $\hat{\theta}_k$ with bootstrapped variance $\hat{\sigma}_k^2$, $k = 1, \ldots, 186$ types of AEs

- Normal-normal hierarchical model (NNHM): $\hat{\theta}_j | \theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$, $\theta_j | \theta, \rho \sim \mathrm{N}(\theta, \rho^2)$, $j = 1, \ldots, K$
- Interpretation of estimate $\hat{\theta}$ (intercept): $\exp(\hat{\theta})$ corresponds to the estimated average ratio

FU time	IP	Prob Trans ID	1-KM	Prob Trans ID CE	AJE (death)
maximum	0.972	2.097	1.214	1.130	1.170
P90	0.983	1.361	1.128	1.026	1.100
P60	1.000	1.138	1.062	1.006	1.050
P30	0.993	1.057	1.031	1.001	1.025

- Univariable and multivariable meta-regression to see what drives the size of the bias; Input variables: value of gold-standard estimator, proportion of censoring, proportion of competing events, maximal follow-up time in experimental group

Meta-analysis

AE probability in group E

Observed data: estimator of log-ratio (log(estimator/gold-standard Aalen-Johansen)) $\hat{\theta}_k$ with bootstrapped variance $\hat{\sigma}_k^2$, $k = 1, \ldots, 186$ types of AEs

- Normal-normal hierarchical model (NNHM): $\hat{\theta}_j | \theta_j \sim N(\theta_j, \sigma_j^2)$, $\theta_j | \theta, \rho \sim N(\theta, \rho^2), j = 1, \ldots, K$
- Interpretation of estimate $\hat{\theta}$ (intercept): $\exp(\hat{\theta})$ corresponds to the estimated average ratio

FU time	IP	Prob Trans ID	1-KM	Prob Trans ID CE	AJE (death)
maximum	0.972	2.097	1.214	1.130	1.170
P90	0.983	1.361	1.128	1.026	1.100
P60	1.000	1.138	1.062	1.006	1.050
P30	0.993	1.057	1.031	1.001	1.025

- Univariable and multivariable meta-regression to see what drives the size of the bias; Input variables: value of gold-standard estimator, proportion of censoring, proportion of competing events, maximal follow-up time in experimental group

Meta-analysis

AE probability in group E

Observed data: estimator of log-ratio (log(estimator/gold-standard Aalen-Johansen)) $\hat{\theta}_k$ with bootstrapped variance $\hat{\sigma}_k^2$, $k = 1, \ldots, 186$ types of AEs

- Normal-normal hierarchical model (NNHM): $\hat{\theta}_j | \theta_j \sim N(\theta_j, \sigma_j^2)$, $\theta_j | \theta, \rho \sim N(\theta, \rho^2)$, $j = 1, \ldots, K$
- Interpretation of estimate $\hat{\theta}$ (intercept): $\exp(\hat{\theta})$ corresponds to the estimated average ratio

FU time	IP	Prob Trans ID	1-KM	Prob Trans ID CE	AJE (death)
maximum	0.972	2.097	1.214	1.130	1.170
P90	0.983	1.361	1.128	1.026	1.100
P60	1.000	1.138	1.062	1.006	1.050
P30	0.993	1.057	1.031	1.001	1.025

- Univariable and multivariable meta-regression to see what drives the size of the bias; Input variables: value of gold-standard estimator, proportion of censoring, proportion of competing events, maximal follow-up time in experimental group

Impact of the choice of relative effect estimator for AE probabilities on qualitative conclusions

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

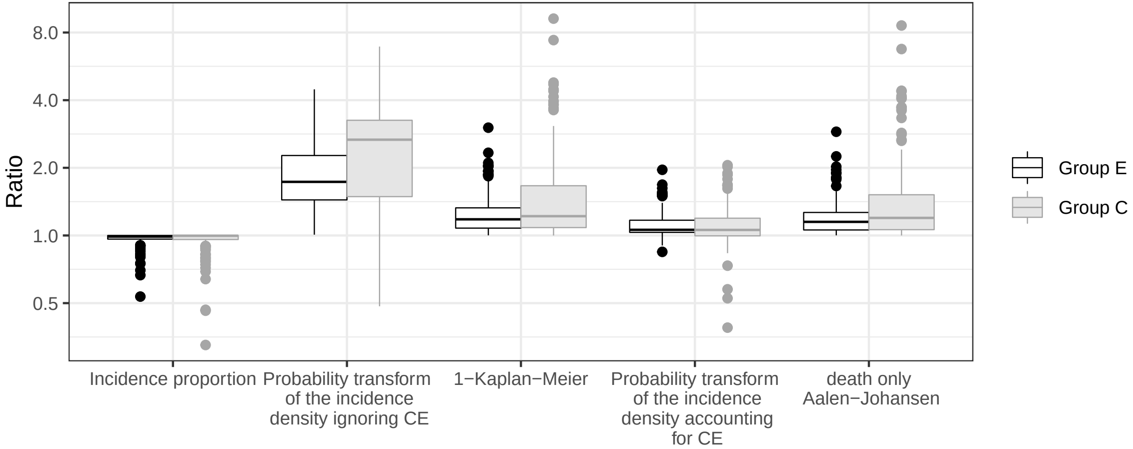
		gold-standard Aalen-Johansen			
		(0) no effect	(a) minor	(b) considerable	(c) major
incidence proportion	(0) no effect	**84**	5		
	(a) minor	3	**10**	2	
	(b) considerable	1	2	**12**	2
	(c) major	1		1	**33**
1-Kaplan-Meier	(0) no effect	**84**	9	4	8
	(a) minor	3	**6**	3	3
	(b) considerable	2	1	**7**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

Impact of the choice of relative effect estimator for AE probabilities on qualitative conclusions

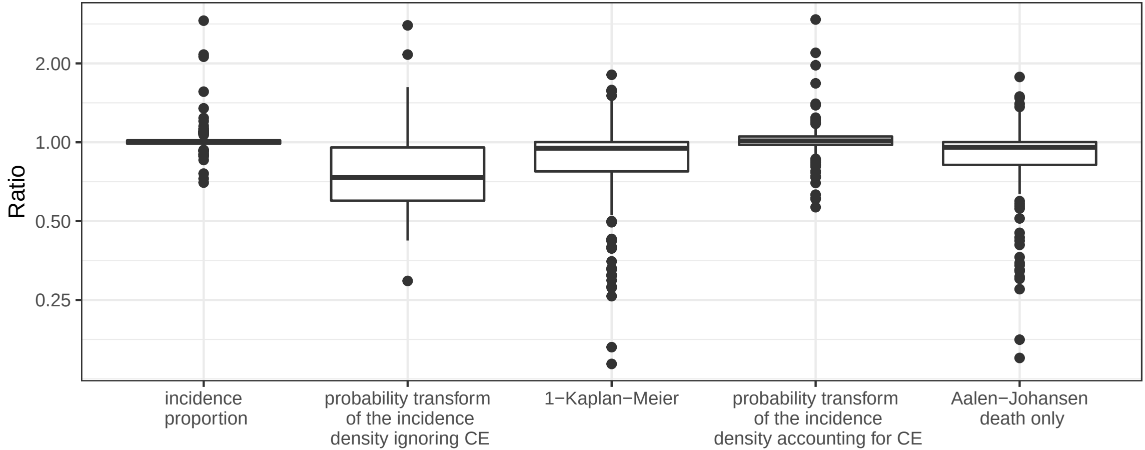
- Categorization motivated by IQWiG General Methods Version 5.0 (2017)
  - (0) no effect: $1 \in \mathrm{CI}$
  - (a) minor: $\mathrm{RR} < 1$ & $\mathrm{CI}_{\mathrm{upper}} \in [0.9, 1)$ or $\mathrm{RR} > 1$ & $\mathrm{CI}_{\mathrm{lower}} \in (1, 1.11]$
  - (b) considerable: $\mathrm{RR} < 1$ & $\mathrm{CI}_{\mathrm{upper}}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

		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**

AE probability in E and C at common maximum follow-up time



Ratio of RRs


Ratio
2.00
1.00
0.50
0.25
incidence proportion
probability transform of the incidence density ignoring CE
1−Kaplan−Meier
probability transform of the incidence density accounting for CE
Aalen−Johansen death only

Meta-analysis

Ratio of RRs

Observed data: estimator of log-ratio of RRs (log(RR estimator/RR gold-standard Aalen-Johansen)) $\hat{\theta}_k$ with bootstrapped variance $\hat{\sigma}_k^2$, $k = 1, \ldots, 186$ types of AEs

- Normal-normal hierarchical model (NNHM): $\hat{\theta}_j | \theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$, $\theta_j | \theta, \rho \sim \mathrm{N}(\theta, \rho^2)$, $j = 1, \ldots, K$
- Interpretation of estimate $\hat{\theta}$ (intercept): exp$(\hat{\theta})$ corresponds to the estimated average ratio

FU time	IP	Prob Trans ID	1-KM	Prob Trans ID CE	AJE (death)
common maximum	0.997	0.732	0.838	0.977	0.860
P90	0.999	0.803	0.883	0.994	0.901
P60	1.000	0.925	0.956	1.020	0.961
P30	1.001	0.977	0.991	1.025	0.992

Meta-analysis

Ratio of RRs

Observed data: estimator of log-ratio of RRs (log(RR estimator/RR gold-standard Aalen-Johansen)) $\hat{\theta}_k$ with bootstrapped variance $\hat{\sigma}_k^2$, $k = 1, \ldots, 186$ types of AEs

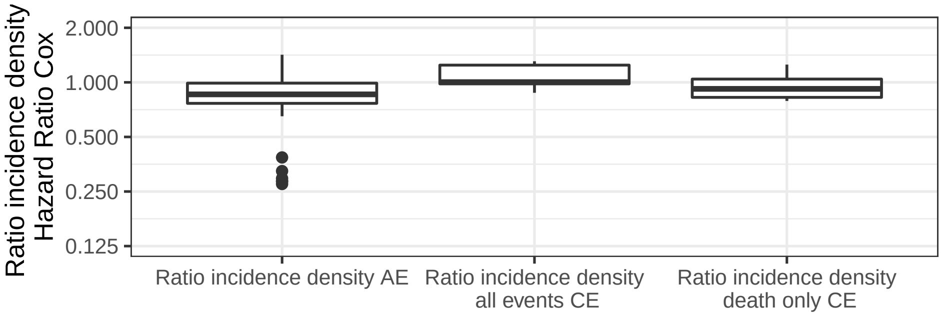
- Normal-normal hierarchical model (NNHM): $\hat{\theta}_j | \theta_j \sim N(\theta_j, \sigma_j^2)$, $\theta_j | \theta, \rho \sim N(\theta, \rho^2)$, $j = 1, \ldots, K$
- Interpretation of estimate $\hat{\theta}$ (intercept): $\exp(\hat{\theta})$ corresponds to the estimated average ratio

FU time	IP	Prob Trans ID	1-KM	Prob Trans ID CE	AJE (death)
common maximum	0.997	0.732	0.838	0.977	0.860
P90	0.999	0.803	0.883	0.994	0.901
P60	1.000	0.925	0.956	1.020	0.961
P30	1.001	0.977	0.991	1.025	0.992

Estimators of relative AE risk based on hazards

- Compare ratio of incidence densities to hazard ratio (HR) calculated by Cox model (gold-standard)
- Always consider all cause-specific hazards in a competing risks analysis (Latouche et al., 2013)
- Meta-analysis (Only consider maximum follow-up time)

FU time	Ratio incidence density AE	Ratio incidence density all events CE	Ratio incidence density death only CE
maximum	0.803	0.908	0.958



Comparison of two gold-standards

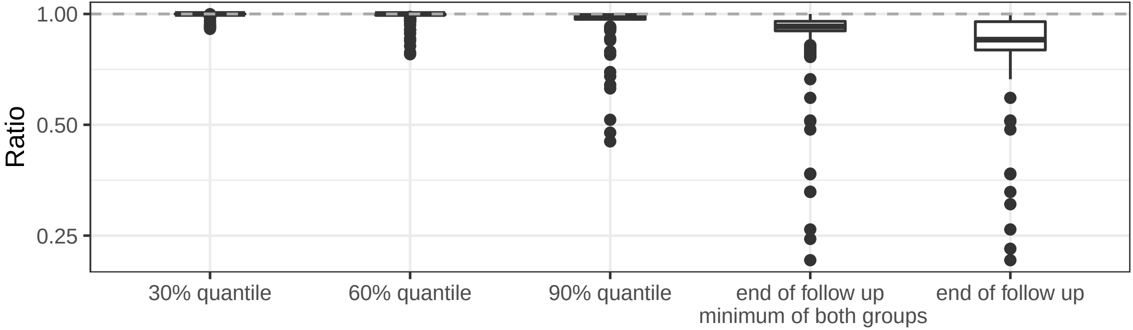
- Compare conclusions of the RR calculated with the Aalen-Johansen estimator (all events) to conclusions of the HR calculated with the Cox model.

		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

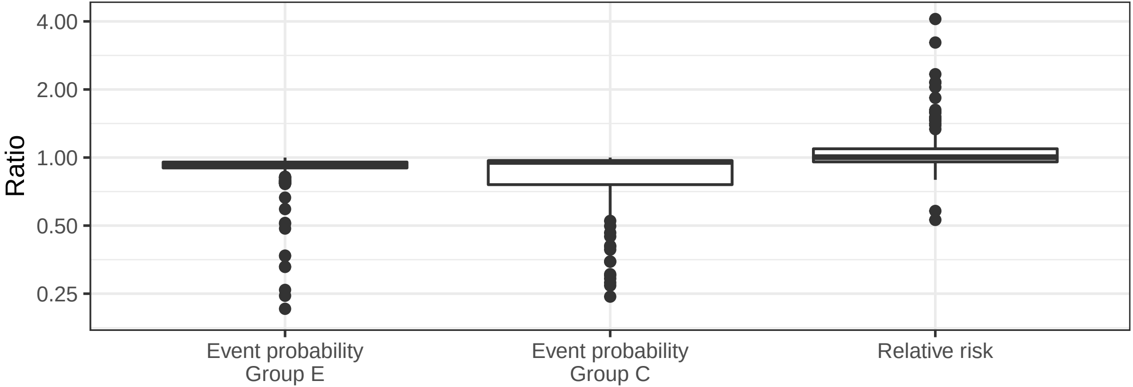
Composite Endpoint - Role of Censoring

- Role of censoring without complication of competing events
- Comparison of incidence proportion to 1-Kaplan-Meier using the all events definition of a competing event (here gold-standard)
- Experimental group composite event probability



Composite Endpoint - Role of Censoring

- Comparison of treatment groups
- Comparison of RR calculated with incidence proportion to RR calculated with 1-Kaplan-Meier using the all events definition of a competing event (here gold-standard)



Discussion

Competing Event Definition

- Definition of a competing event the main reason why the incidence proportion performed so well

		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**

		Aalen-Johansen (death only)				
		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	**5**	1	0
	common	0	0	0	**73**	15
	very common	0	0	0	0	**86**

- Definition of censoring should not differ between treatment groups

Discussion

Competing Event Definition

- Definition of a competing event the main reason why the incidence proportion performed so well

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

		Aalen-Johansen (death only)				
		very rare	rare	uncommon	common	very common
		(<0.01%)	(<0.1%)	(<1%)	(<10%)	(>=10%)
IP	very rare	**6**	0	0	0	0
	rare	0	**0**	0	0	0
	uncommon	0	0	**5**	1	0
	common	0	0	0	**73**	15
	very common	0	0	0	0	**86**

- Definition of censoring should not differ between treatment groups

Discussion

Estimand

Estimand in safety context (Stegherr et al., 2021 BiomJ, Unkel et al., 2019)

- *Treatment policy estimand*
  - Comparison of treatment groups with regard to AEs on entire follow-up period irrespective of intercurrent events as treatment discontinuation
  - Of interest for HTA bodies
- *While on treatment estimand*
  - Documentation of AEs often ended after treatment discontinuation
- In empirical study secondary data analysis of preexisting trials; Participating companies defined competing events trialspecific; Have to take data as collected

Our interest was **not** precisely quantifying the risk of a certain type of AE for a specific drug or therapeutic field, but a **methodological comparison** of estimators of absolute AE risk in presence of competing events

Discussion

Estimand

Estimand in safety context (Stegherr et al., 2021 BiomJ, Unkel et al., 2019)

- *Treatment policy estimand*
  - Comparison of treatment groups with regard to AEs on entire follow-up period irrespective of intercurrent events as treatment discontinuation
  - Of interest for HTA bodies
- *While on treatment estimand*
  - Documentation of AEs often ended after treatment discontinuation
- In empirical study secondary data analysis of preexisting trials; Participating companies defined competing events trialspecific; Have to take data as collected

Our interest was **not** precisely quantifying the risk of a certain type of AE for a specific drug or therapeutic field, but a **methodological comparison** of estimators of absolute AE risk in presence of competing events

Discussion

Guidelines need updates!

ICH

- Methods to analyze safety, e.g., E2, E3, or E9
- Describing analysis methods, primarily incidence proportions and incidence density
- E9 explicitly asks for "... appropriate use of **survival methods** to exploit the potential relationship of the incidence of adverse events to duration of exposure and/or follow-up"

EMA's anticancer guideline

- "... Kaplan-Meier analysis of selected adverse events which consider censoring of events, may be useful."
- Not being specific which events to censor ⇒ What about competing event?

Extension of the CONSORT statement on reporting of harms

- Recommends "Kaplan-Meier curves showing cumulative incidence of important adverse events can be helpful"
- Neither discusses censoring nor competing events

Discussion

Guidelines need updates!

ICH

- Methods to analyze safety, e.g., E2, E3, or E9
- Describing analysis methods, primarily incidence proportions and incidence density
- E9 explicitly asks for "… appropriate use of **survival methods** to exploit the potential relationship of the incidence of adverse events to duration of exposure and/or follow-up"

EMA's anticancer guideline

- "… Kaplan-Meier analysis of selected adverse events which consider censoring of events, may be useful."
- Not being specific which events to censor $\Rightarrow$ What about competing event?

Extension of the CONSORT statement on reporting of harms

- Recommends "Kaplan-Meier curves showing cumulative incidence of important adverse events can be helpful"
- Neither discusses censoring nor competing events

Discussion

Guidelines need updates!

ICH

- Methods to analyze safety, e.g., E2, E3, or E9
- Describing analysis methods, primarily incidence proportions and incidence density
- E9 explicitly asks for "... appropriate use of **survival methods** to exploit the potential relationship of the incidence of adverse events to duration of exposure and/or follow-up"

EMA's anticancer guideline

- "... Kaplan-Meier analysis of selected adverse events which consider censoring of events, may be useful."
- Not being specific which events to censor ⇒ What about competing event?

Extension of the CONSORT statement on reporting of harms

- Recommends "Kaplan-Meier curves showing cumulative incidence of important adverse events can be helpful"
- Neither discusses censoring nor competing events

Discussion

Limitations

- The frequency categories do include more "common" and "very common" types of AEs than "very rare" or "rare" AE
- For the single types of AEs using the gold-standard (Aalen-Johansen estimator all events) the RR is more often greater than 1 than smaller (or equal 1), i.e., AE probability more often greater in the experimental group E
- No hierachy levels beyond AE level (indication, MedDRA SOC) were considered in the meta-analysis
- Most studies are from oncology (12 of 17)

The empirical study illustrates possible biases but the results can not be generalized.

Discussion

Limitations

- The frequency categories do include more "common" and "very common" types of AEs than "very rare" or "rare" AE
- For the single types of AEs using the gold-standard (Aalen-Johansen estimator all events) the RR is more often greater than 1 than smaller (or equal 1), i.e., AE probability more often greater in the experimental group E
- No hierachy levels beyond AE level (indication, MedDRA SOC) were considered in the meta-analysis
- Most studies are from oncology (12 of 17)

The empirical study illustrates possible biases but the results can not be generalized.

Summary

- Choice of estimator crucial for estimation of AE probability and for group comparisons
- Results of empirical study inline with results of methodological considerations and simulations (Stegherr et al., 2021 Pharm Stat)
- Time-to-event methods account for censoring, but **Kaplan-Meier must not be used**
- If time-to-event methods are considered to account for censoring instead of simply using the incidence proportion, it is important to correctly analyze competing events
- If competing events are falsely analyzed as censored observations, bias will be induced. This bias is worse than completely disregarding the time-to-event structure by using the incidence proportion if there are many competing events
- Ignoring competing events worse than assuming simple constant hazards model
- Best choice: Always use Aalen-Johansen estimator to estimate AE probability and use RR calculated with Aalen-Johansen and HR from Cox model for **all** types of events that are considered for group comparisons

References

- Stegherr, R., Schmoor, C., Beyersmann, J., Rufibach, K., Jehl, V., Brückner, A., Eisele, L., Künzel, K., Kupas, K., Langer, F., Leverkus, F., Loos, A., Norenberg, C., Voss, F. and Friede, T. (2020). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY) — estimation of adverse event risks. *Trials*, 22(1), 1-13.
- Rufibach, K., Stegherr, R., Schmoor, C., Jehl, V., Allignol, A., Boeckenhoff, A., Dunger-Baldauf, C., Eisele, L., Künzel, T., Kupas, K., Leverkus, F., Trampisch, M., Zhao, Y., Friede, T. and Beyersmann, J. (2020) Survival analysis for AdVerse events with VarYing follow-up times (SAVVY) – comparison of adverse event risks in randomized controlled trials. arXiv:2008.07881 (submitted).
- Stegherr, R., Beyersmann, J., Jehl, V., Rufibach, K., Leverkus, F., Schmoor, C. and Friede, T. (2020). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): Rationale and statistical concept of a meta-analytic study. *Biometrical Journal*, *Biometrical Journal*, 63(3), 650-670.
- Stegherr, R., Schmoor, C., Lübbert, M., Friede, T. and Beyersmann, J. (2020). Estimating and comparing adverse event probabilities in the presence of varying follow-up times and competing events. *Pharmaceutical Statistics*, 20(6), 1125-1146.
- Allignol, A., Beyersmann, J. and Schmoor, C. (2016). Statistical issues in the analysis of adverse events in time-to-event data. *Pharmaceutical Statistics* **15**, 297–305.
- Bender, R., and Beckmann, L. (2019). Limitations of the incidence density ratio as approximation of the hazard ratio. *Trials*, **20**(1), 485.
- Pocock, S. J., Clayton, T. C. and Altman, D. G. (2002). Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *The Lancet* **359**, 1686–1689.
- 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.
- Unkel, S., Amiri, M., Benda, N., Beyersmann, J., Knoerzer, D., Kupas, K., Langer, F., Leverkus, F., Loos, A., Ose, C., Proctor, T., Schmoor, C., Schwenke, C., Skipka, G., Unnebrink, K., Voss, F. and Friede, T. (2019). On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. *Pharmaceutical Statistics* **18**, 165–183.
- 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**, 96–105.