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

Regina Stegherr[1], Jan Beyersmann[1], Valentine Jehl[2], Kaspar Rufibach[3], Friedhelm Leverkus[4], Claudia Schmoor[5], Tim Friede[6] for the SAVVY project group

[1] Institute of Statistics, Ulm University, Ulm, Germany
[2] Novartis Pharma AG, Basel, Switzerland
[3] F. Hoffmann-La Roche, Basel, Switzerland
[4] Pfizer, Berlin, Germany
[5] Clinical Trials Unit, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany
[6] Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

25 August 2020

regina.stegherr@uni-ulm.de

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 assesent of therapies (Unkel et al., 2019).
- For **quantifying AE risks** in a time-to-first-event analysis several estimators have been suggested so far.
- **Compare commonly** used (but possibly biased) estimators to estimators **accounting for competing events** in time-to-event studies and also compare safety comparisons between treatment groups.
- 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).
- **Aim:** Investigate in an **empirical study** of several randomized controlled trials whether the use of different estimators for analyses of AEs leads to different conclusions about therapies' safety

Estimating AE probabilities: Commonly used but biased methods

- **Incidence proportion**: $\dfrac{\text{\# AEs in } [0, \tau]}{\text{\# patients}}$
  - 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
- **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 events (van Walraven et al. 2016, Schumacher et al. 2016)
  - But health technology agencies, e.g., IQWiG, still ask for Kaplan-Meier estimates
- **Incidence density**: $\mathrm{ID}(\tau) = \dfrac{\text{\# AE in } [0, \tau]}{\text{patient-time at risk restricted by } \tau}$
  - 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

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
- **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 events (van Walraven et al. 2016, Schumacher et al. 2016)
  - But health technology agencies, e.g., IQWiG, still ask for Kaplan-Meier estimates
- **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

Estimating AE probabilities: Commonly used but biased methods

- **Incidence proportion**: $\dfrac{\text{\# AEs in } [0,\tau]}{\text{\# patients}}$
  - 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
- **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 events (van Walraven et al. 2016, Schumacher et al. 2016)
  - But health technology agencies, e.g., IQWiG, still ask for Kaplan-Meier estimates
- **Incidence density**: $\mathrm{ID}(\tau) = \dfrac{\text{\# AE in } [0,\tau]}{\text{patient-time at risk restricted by } \tau}$
  - 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

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): $\frac{\mathrm{ID}(\tau)}{\mathrm{ID}(\tau) + \overline{\mathrm{ID}}(\tau)} \left(1 - \exp(-\tau \cdot [\mathrm{ID}(\tau) + \overline{\mathrm{ID}}(\tau)])\right)$

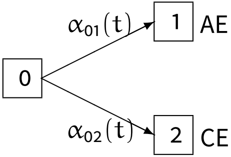
  with $\overline{\mathrm{ID}}(\tau) = \frac{\text{\# competing event in } [0,\tau]}{\text{patient-time at risk restricted by } \tau}$
  - Assumption of **constant hazards** for both (AE and competing event) hazards
  - Literature about incidence densities often neglects CEs (e.g. book ‘Analysis of incidence 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): $\frac{\mathrm{ID}(\tau)}{\mathrm{ID}(\tau) + \overline{\mathrm{ID}}(\tau)} \left(1 - \exp(-\tau \cdot [\mathrm{ID}(\tau) + \overline{\mathrm{ID}}(\tau)])\right)$

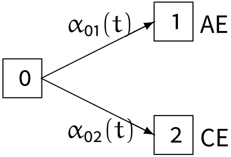
  with $\overline{\mathrm{ID}}(\tau) = \frac{\text{\# competing event in } [0,\tau]}{\text{patient-time at risk restricted by } \tau}$
  - Assumption of **constant hazards** for both (AE and competing event) hazards
  - Literature about incidence densities often neglects CEs (e.g. book ‘Analysis of incidence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

Definition of the competing event



- Time-to-1st-event and type-of-1st-event
- What are the possibilities of the 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 and any event of patients course of disease or treatment that stops the recording of the interesting type of AE (e.g. disease- or safety-related loss to follow-up, withdrawal of consent and discontinuation)

Definition of the competing event



- Time-to-1st-event and type-of-1st-event
- What are the possibilities of the 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 and any event of patients course of disease or treatment that stops the recording of the interesting type of AE (e.g. disease- or safety-related loss to follow-up, withdrawal of consent and discontinuation)

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 treatment group τ_E may be greater/smaller than largest observed event time in comparison group 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 $\tau = \min(\tau_E, \tau_C)$ (Considered for group comparisons)
- 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 treatment group τ_E may be greater/smaller than largest observed event time in comparison group 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 $\tau = \min(\tau_E, \tau_C)$ (Considered for group comparisons)
- 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)

Empirical Study

- The Statistical Analysis Plan can be found in Stegherr et al. (2020)
- Only **aggregated** data shared: Trial level analyses ran within the sponsor company / organization using SAS (and R) code provided ⇒ **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 studies and a total of 62 type of AEs (range 3 - 51 per study))
- **Main study**: 10 participating organizations contributing 17 studies including 186 types of adverse events

Comparison of frequency categories

AE probability in group E at maximum follow-up time

- Incidence proportion vs Aalen-Johansen estimator (all events) at maximum follow-up time

		Aalen-Johansen (all events)				
		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 Aalen-Johansen estimator (all events) at maximum follow-up time

		Aalen-Johansen (all events)				
		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 Aalen-Johansen estimator (all events) at maximum follow-up time

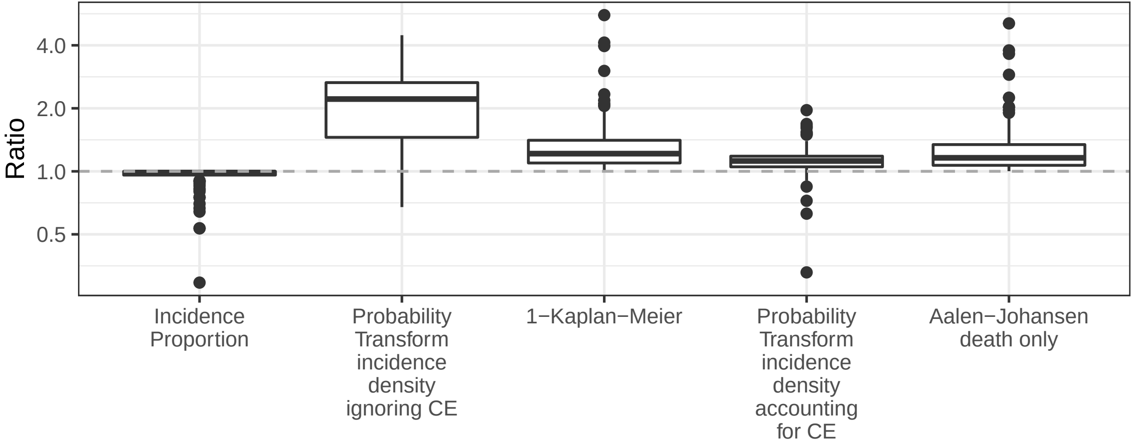
		Aalen-Johansen (all events)				
		very rare	rare	uncommon	common	very common
		(<0.01%)	(<0.1%)	(<1%)	(<10%)	(>=10%)
IP	very rare	6	0	0	0	0
	rare	0	0	0	0	0
	uncommon	0	0	6	0	0
	common	0	0	0	86	**2**
	very common	0	0	0	0	86

- 1-Kaplan-Meier vs Aalen-Johansen estimator (all events) at maximum follow-up time

		Aalen-Johansen (all events)				
		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/Aalen-Johansen estimator (all events)

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/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/Aalen-Johansen)) $\hat{\theta}_k$ with bootstrapped variance $\hat{\sigma}_k^2$, $k = 1, ..., 186$ types of AEs

- Normal-normal hierarchical model (NNHM): $\hat{\theta}_j | \theta_j \sim N(\theta_j, \sigma_j^2)$, $\theta_j | \theta, \rho \sim N(\theta, \rho^2), j = 1, ..., K$
- 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/Aalen-Johansen)) $\hat{\theta}_k$ with bootstrapped variance $\hat{\sigma}_k^2$, $k = 1, \dots, 186$ types of AEs

- Normal-normal hierarchical model (NNHM): $\hat{\theta}_j | \theta_j \sim N(\theta_j, \sigma_j^2)$, $\theta_j | \theta, \rho \sim N(\theta, \rho^2), j = 1, \dots, K$
- 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

Summary

- Time-to-event methods account for censoring, but Kaplan-Meier must not be used (on average 1.21-fold overestimation compared to Aalen-Johansen); Kaplan-Meier censors competing events and hence overestimates AE probabilities
- Ditto: Using one AE incidence density only (on average 2.1-fold overestimation). This bias is worse than simply using incidence proportions (on average 0.97-fold underestimation but minimum of 0.294 observed) if there are many competing events
- Ignoring competing events worse than assuming simple constant hazards model

Summary of comparison of AE risks between treatment groups

- Choice of estimator of AE probability crucial for group comparisons in terms of the relative risk (RR)
- Meta-analysis at maximum follow-up time (accounted for same length of follow-up in both groups): average $\mathrm{RR}_{estimator}/\mathrm{RR}_{gold-standard AJE}$

FU time	IP	Prob Trans ID	1-KM	Prob Trans ID CE	AJE (death)
$\min(\tau_E, \tau_C)$	0.997	0.732	0.838	0.977	0.860

- Incidence proportion on average comparable but there are also types of AEs for which the RR based on the incidence proportion is up to the 3-fold of RR based on the gold-standard AJE
- Different lengths of confidence intervals and different RR estimates may result in different conclusions of group comparisons

Summary of comparison of AE risks between treatment groups

- Choice of estimator of AE probability crucial for group comparisons in terms of the relative risk (RR)
- Meta-analysis at maximum follow-up time (accounted for same length of follow-up in both groups): average $\mathrm{RR}_{estimator}/\mathrm{RR}_{gold-standardAJE}$

FU time	IP	Prob Trans ID	1-KM	Prob Trans ID CE	AJE (death)
$\min(\tau_E, \tau_C)$	0.997	0.732	0.838	0.977	0.860

- Incidence proportion on average comparable but there are also types of AEs for which the RR based on the incidence proportion is up to the 3-fold of RR based on the gold-standard AJE
- Different lengths of confidence intervals and different RR estimates may result in different conclusions of group comparisons

Discussion

- The frequency categories do include more "common" and "very common" types of AEs than "very rare" or "rare" types of AE
- For the single 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 type of AE level (indication, MedDRA SOC) were considered in the meta-analysis
- Most studies are from oncology (12 of 17) which typically have few censorings and many competing events
- Recommendation: Always use Aalen-Johansen estimator with all events definition of CEs

References

- Stegherr, R., Beyersmann, J., Jehl, V., et al. (2020). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): Rationale and statistical concept of a meta-analytic study, Biometrical Journal (in press). Preprint available at arXiv:1912.00263
- Stegherr, R., Schmoor, C., Beyersmann, J., et al. (2020). Survival analysis for AdVerse events with VarYing follow-up time (SAVVY) — estimation of adverse event risks. (submitted)
- Rufibach, K., Stegherr, R., Schmoor, C., et al. (2020). Survival analysis for AdVerse events with VarYing follow-up time (SAVVY) — comparison of adverse event risks in randomized controlled trials. (submitted)
- 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., et al. (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.