Evaluation of event rate differences using stratified Kaplan-Meier estimates with Mantel-Haenszel weights & adjusted hybrid variance estimators

Bischofberger S[1], Buchner H[1], Göldner RG[2], Bleckert G[1], Zhu L[1], Haller B[3]
[1]*Staburo GmbH, Munich, Germany*
[2]*Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany*
[3]*Technical University of Munich, School of Medicine, Institute of AI & Informatics in Medicine*

BACKGROUND

- Assessment of differences in event rates is a common endeavor in the evaluation of efficacy and safety of new treatments in clinical trials (in particular in oncology).
- We investigate the performance of different hypothesis tests for an overall survival endpoint.
- Stratified analyses are desired and sometimes even required by regulators.
- We illustrate the necessity of non-zero variance estimates – especially in the presence of strong prognostic stratification effects.
- Focus: comparison of event rates via Kaplan-Meier estimates for a pre-specified time t_0
- One-sided test for superiority at significance level α=0.05.

GOALS:
- Choose best **stratum weights**
- Choose best Kaplan-Meier **variance estimator**
- Compare performance with **Cox model** (stratified)

SIMULATION STUDY

- 1 experimental arm and 1 control arm
- Per treatment group: 65 subjects divided into 3 strata
- Time of interest: $t_0 = 100$ (e.g., days)
- Survival time and censoring time from exponential distributions
- Constant censoring intensity λ_{cens}=0.005 (leads to 22%-38% of patients censored in scenario 2 below)
- Hazard rates of **active treatment group** (G=1) and **control group** (G=2) in stratum k: $\lambda_{k,G}$ for G=1,2
- Two underlying models:
  - **proportional hazard rates** that satisfy the **Cox model (COX)**
    $$\lambda_{k,2}(t) = c_P \lambda_{k,1}(t)$$
    for all t and with same hazard ratio $c_P \geq 1$ for all k.
  - **additive survival difference (ASD)** at time t_0:
    $$S_{k,2}(t_0) = S_{k,1}(t_0) - c_A$$
    for same difference in survival $c_A \geq 0$ for all k.
- 10,000 simulation runs
- Simulate proportional effects (hazard ratios): c_P = 2.0, 2.5, **3.0**, 3.5
- Simulate additive effects: c_A = 0.10, 0.15, **0.20**, 0.25
- 3 different allocations in table below but results only illustrated for scenario 2

Scenario	$S_{k,1}(t_0)$ for k=1,2,3	n per treatment group for k=1,2,3	Note
1	0.80, 0.50, 0.30	20, 30, 15	Base case
2	**0.95**, **0.70, 0.50**	**45**, **10, 10**	**Largest stratum with greatest $S_{k,1}(t_0)$**
3	**0.95**, 0.70, 0.50	**10**, 10, 45	Smallest stratum with greatest $S_{k,1}(t_0)$

Z-TESTS FOR KAPLAN-MEIER RATES

Z-test for 3 strata at time t_0:

$$Z = \frac{\sum_{k=1}^{3} w_k (\hat{S}_{k,1} - \hat{S}_{k,2})}{\left(\sum_{k=1}^{3} {w_k}^2 \left(\widehat{\sigma_{k,1}^2} + \widehat{\sigma_{k,2}^2}\right)\right)^{0.5}}$$

- $\hat{S}_{k,G}$: Kaplan-Meier estimator in stratum k and treatment group G=1,2 at time t_0
- $\widehat{\sigma_{k,G}^2}$: variance estimate of $\hat{S}_{k,G}$, w_k: stratum weights

VARIANCE ESTIMATORS

Common choice:
Greenwood formula:

$$\hat{S}_{k,G}^2 \sum_{t_i \leq t_0} \frac{d_i}{n_i(n_i - d_i)}$$

Problem: can become 0 if no events observed or if all subjects have an event → Z is not well-defined. In the simulations, Z is evaluated if at least one stratum with non-zero variance is present. Strata with 0 variance are excluded from the analysis.

Suggestion:
Borkowf's adjusted hybrid variance estimator [1]:

$$w_*(1 - w_*)/(n - m_c)$$

for $w_* = (1 - n^{-1})w + (2n)^{-1}$,
m_c: # censored subjects at t_0,
w: truncated version of $\hat{S}_{k,G}$

Assures non-zero variances for all t

STRATUM WEIGHTS

Inverse variance (IV) weight:

$$w_k = \frac{\left(\widehat{\sigma_{k,1}^2} + \widehat{\sigma_{k,2}^2}\right)^{-1}}{\sum_{j=1}^{4} \left(\widehat{\sigma_{j,1}^2} + \widehat{\sigma_{j,2}^2}\right)^{-1}}$$

Motivation:
"least-squares", i.e., minimal variance in Z [2]

Problem:
not defined for strata with 0 variance.

Mantel-Haenszel (MH) weight:

$$w_k = \frac{n_k m_k/(n_k + m_k)}{\sum_j \left(n_j m_j/(n_j + m_j)\right)}$$

for n_k and m_k number of treatment and control subjects in stratum k

Motivation: robust for sparse events [3]
if $n_k = m_k$:
w_k = subjects in stratum k / total number of subjects

2 underlying models with 2 different types of tests each:

	Test for ASD (H_0: $c_A = 0$)	Test for COX (H_0: $c_P = 1$)
Simulate: COX $c_P \geq 1$	Z-tests assumptions violated ✗	Cox regression correct model ✓
Simulate: ASD $c_A \geq 0$	Z-tests correct model ✓	Cox regression assumptions violated ✗

RESULTS

- The **violation of assumptions** for the respective mis-specified stratified tests increases with increasing effect size in the underlying model: The effects c_P and c_A are not the same across all strata if the data are generated from the respective other model.

c_A	c_P per stratum		
	stratum 1	stratum 2	stratum 3
0.10	3.168	1.432	1.322
0.15	4.350	1.676	1.515
0.20	**5.609**	**1.943**	**1.737**
0.25	6.954	2.239	2.000

Table 1: *Proportional hazard ratios c_P in data from ASD models with actual effect c_A.*

c_P	c_A per stratum		
	stratum 1	stratum 2	stratum 3
2.0	0.048	0.210	0.250
2.5	0.070	0.290	0.323
3.0	**0.093**	**0.357**	**0.375**
3.5	0.114	0.413	0.412

Table 2: *Additive survival effects c_A in data from COX models with actual effect c_P.*

- At least one stratum with Greenwood variance equal to 0 occurred only in <0.01% of all ASD simulation runs with c_A=0.2 and in 0.1% of all COX simulations with c_P=3.0 (in 2.9% of all simulations under H_0).
  The case with 0 variance in all three strata did not occur in this scenario.
- In scenario 3 (not discussed here), up to 45% of all simulation runs had a stratum with 0 variance under H_0.

Type I error:
- Cox regression controls type I error as expected.
- Stratified Z-tests with Greenwood variance inflate type I error and are excluded from the analysis of power below (dotted lines in Fig. 1).
- Unstratified Greenwood Z-test controls type I error.
- Z-tests with Borkowf's variance can be too conservative.

Test	Type I Error
Cox regression	0.050
Z: unstratified, Greenwood	0.040
Z: IV weight, Greenwood	**0.059**
Z: MH weight, Greenwood	**0.070**
Z: unstratified Borkowf	0.026
Z: IV weight, Borkowf	0.025
Z: MH weight, Borkowf	0.041

Table 3: *Type I errors based on simulations.*

Power of test:
- Data from COX model: The Cox regression performs best as expected, the Z-test with MH-weights and Borkowf is a reliable alternative.
- Data from ASD model: Z-test with IV-weights and Borkowf performs very well. Cox regression, unstratified Z-test with Greenwood and Z-test with MH-weights and Borkowf have almost identical power.

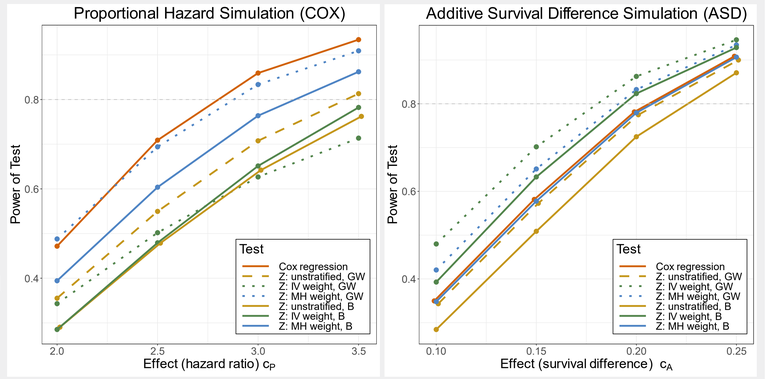

Figure 1: *Power vs effect size for underlying COX model (left) and ASD model (right).*

CONCLUSIONS

- **Z-tests for difference in survival at time t_0** are a **valuable alternative to Cox regression**, especially if the proportionality assumption does not hold.
  However, for small violations the Cox regression is still the model of choice.
- **Greenwood variance can easily become zero** in small or extreme strata (Kaplan-Meier = 0 or 1)
- Z-tests with **Borkowf's variance control type I error** – stratified Z-test with Greenwood does not
- **Mantel-Haenszel type weightings** seem promising
  (still assign a weight to strata with 0 variance)

more information, including more scenarios:

REFERENCES

[1] Borkowf, C.B., 2005. A simple hybrid variance estimator for the Kaplan-Meier survival function. Statist. Med., 24, pp. 827-851.
[2] Lachin, J. M., Biostatistical Methods. The Assessment of Relative Risks, 2nd ed. Wiley-Blackwell, 2011.
[3] Greenland, S. and Robins, J.M., 1985. Estimation of a common effect parameter from sparse follow-up data. Biometrics, 41, pp. 55-68.