EVALUATION OF Z-TESTS TO COMPARE FIXED TIME SURVIVAL PROBABILITIES USING STRATIFIED KAPLAN-MEIER ESTIMATES WITH DIFFERENT VARIANCE ESTIMATORS AND WEIGHTS

MARIA BLANCO[1], STEPHAN BISCHOFBERGER[1], BERNHARD HALLER[2], GABRIELE BLECKERT[1], EVA HOSTER[3], HANNES BUCHNER[1]

[1]Staburo GmbH, Munich, Germany [2]Technical University of Munich, Institute of AI and Informatics in Medicine

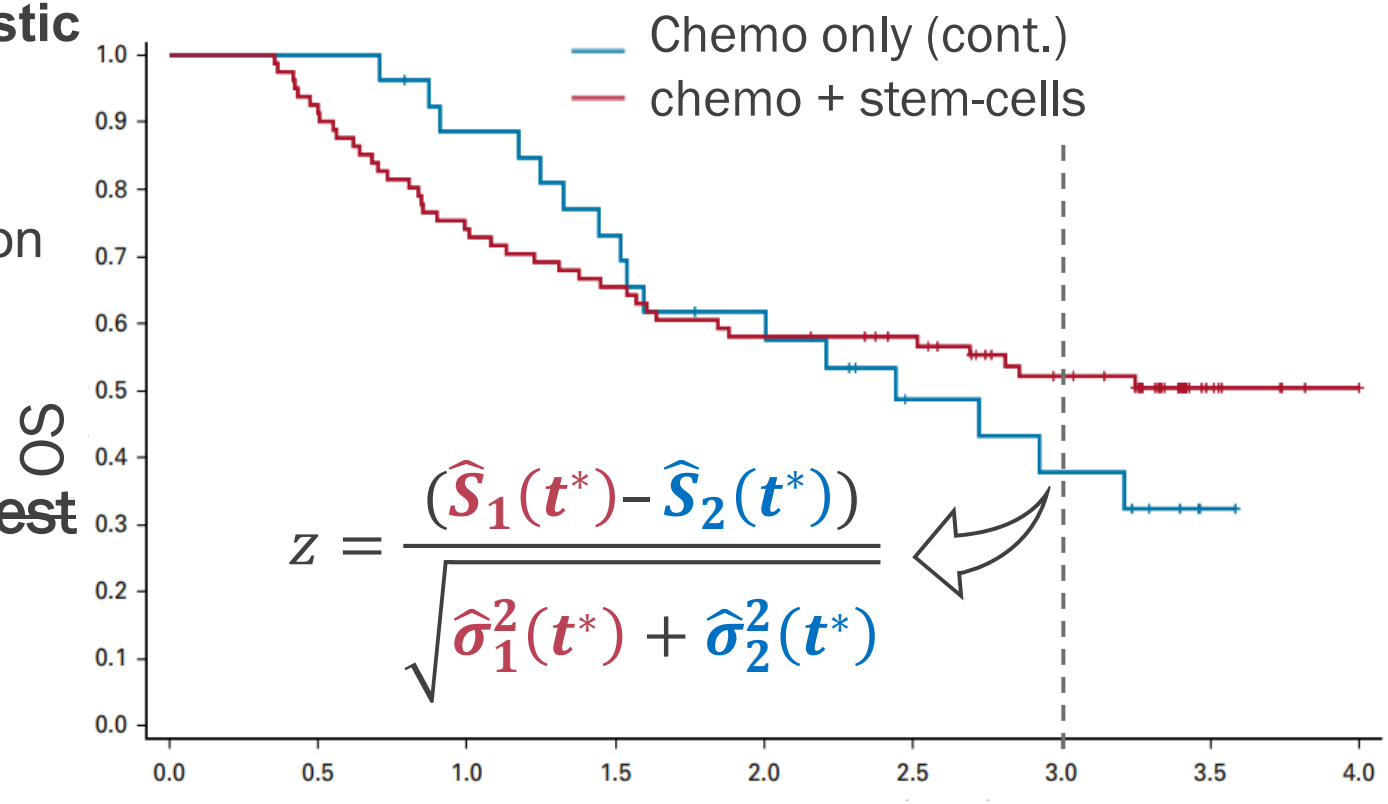
[3]Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians-Universität München

PSI 2026 CONFERENCE

15 - 17 JUNE 2026

The VidazaAllo study: Overall Survival

- **108** patients with **advanced myelodysplastic syndromes**
- Chemo induction **+** stem-cell transplantation
  vs
  Chemotherapy only (continuously)
  (by donor availability)
- **Non-proportional hazards →** ~~**Log-rank test**~~
- **Primary endpoint**: OS at 3 years (t^*)
- **Analysis:** KM-based Z-test



→ **Interpretability!**

Kröger N, Sockel K, Wolschke C, et al. Comparison Between 5-Azacytidine Treatment and Allogeneic Stem-Cell Transplantation in Elderly Patients With Advanced MDS According to Donor Availability (VidazaAllo Study). *Journal of Clinical Oncology*. 2021;39:3318-3327. doi:10.1200/JCO.20.02724

The VidazaAllo study: Event Free-Survival

- **Major secondary endpoint**:

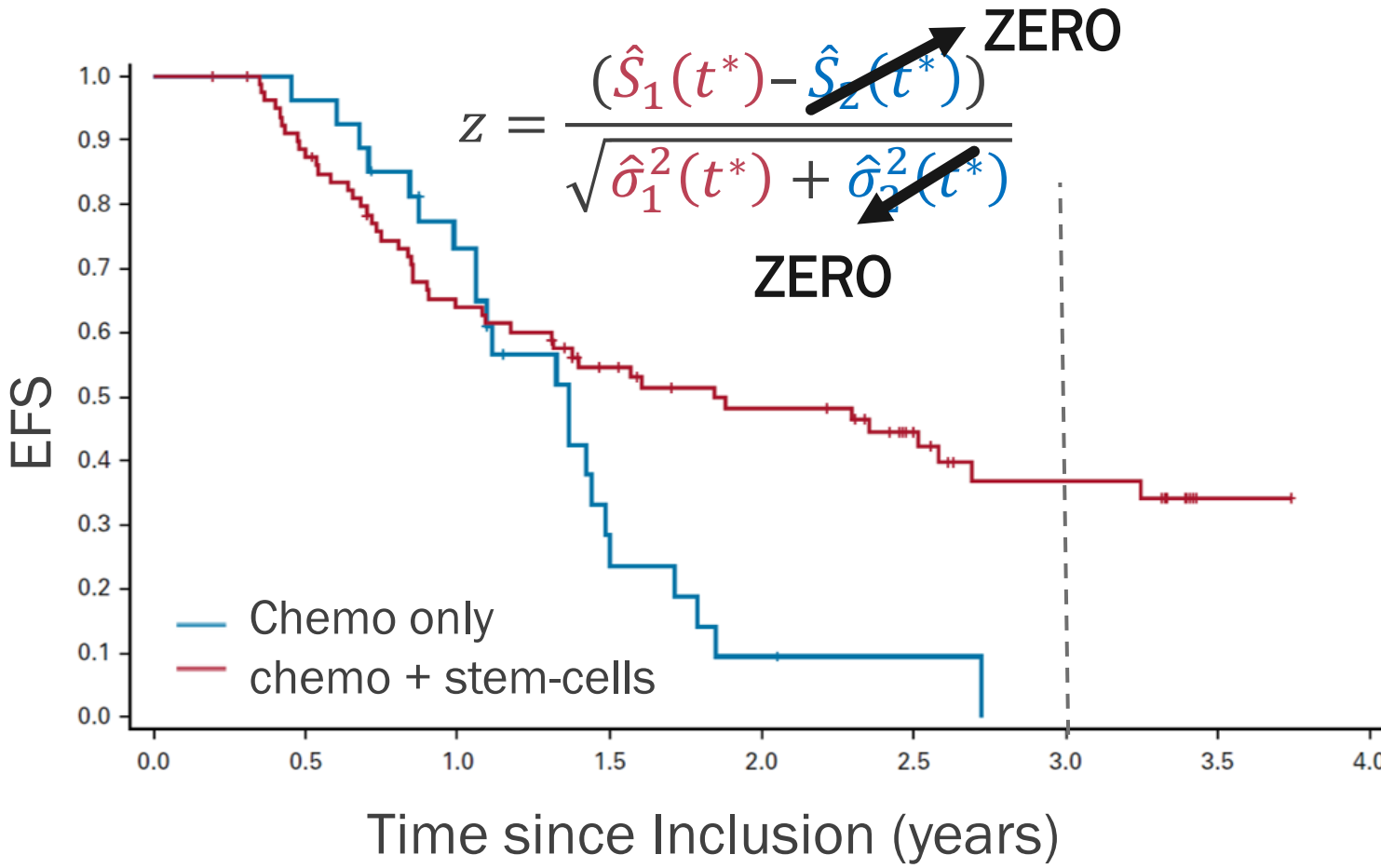
  EFS at 3 years

- **Non-PH**
- **Analysis:** KM-based Z-test

...motivated this work:

is there a better variance estimator?

And how can we incorporate stratification?



Kröger N, Sockel K, Wolschke C, et al. Comparison Between 5-Azacytidine Treatment and Allogeneic Stem-Cell Transplantation in Elderly Patients With Advanced MDS According to Donor Availability (VidazaAllo Study). *Journal of Clinical Oncology*. 2021;39:3318-3327. doi:10.1200/JCO.20.02724

AGENDA

1. ~~Motivation~~ ✓
2. The stratified Z-test
3. Research questions
4. Methods
5. Results
6. Conclusions & Recommendation

THE STRATIFIED Z-TEST

For K strata and 2 arms:

$$Z_{Strat} = \frac{\sum_{k=1}^{K} w_k (\hat{S}_{k,1}(t^*) - \hat{S}_{k,2}(t^*))}{\sqrt{\sum_{k=1}^{K} w_k^2 \left(\hat{\sigma}_{k,1}^2(t^*) + \hat{\sigma}_{k,2}^2(t^*)\right)}}$$

- $\hat{S}_{k,g}(t^*)$: Kaplan–Meier estimate in stratum k, arm g, at t^*
- w_k: weight assigned to stratum k
- $\hat{\sigma}_{k,g}^2(t^*)$: estimated variance of $\hat{S}_{k,g}(t^*)$

⚠️ WHEN DO ZERO VARIANCE ESTIMATES OCCUR?

For Greenwood's variance estimator:

$$\hat{\sigma}^2_{Greenw:\,k,g}\left(\hat{S}_{k,g}(t)\right) = \boldsymbol{\hat{S}_{k,g}(t)^2} \sum_{t_{k,g,i} \le t} \frac{\boldsymbol{d_{k,g,i}}}{n_{k,g,i}\left(n_{k,g,i} - d_{k,g,i}\right)}$$

1. All patients had the event before t^*
2. No events are observed before t^*

$d_{k,g,i}$ = events at t_i
$n_{k,g,i}$ = subjects at risk before t_i
Stratum k and arm g

More common in stratified analysis due to:

- Small strata
- **Strong prognostic factors**

RESEARCH QUESTIONS

$$Z_{Strat} = \frac{\sum_{k=1}^{K} \boldsymbol{w_k} (\hat{S}_{k,1}(t^*) - \hat{S}_{k,2}(t^*))}{\sqrt{\sum_{k=1}^{K} \boldsymbol{w_k^2} \left(\boldsymbol{\hat{\sigma}^2_{k,1}(t^*)} + \boldsymbol{\hat{\sigma}^2_{k,2}(t^*)} \right)}}$$

1 Variance estimator $\hat{\sigma}^2_{k,(.)}$:

Which variance estimator avoids zero variance problems and should be used?

2 Weighting method $\boldsymbol{w_k}$:

Which weighting method is optimal for stratified Z-test?

Q1:
WHICH VARIANCE ESTIMATOR SHOULD BE USED?

Candidates:

1. Greenwood*

Classic asymptotic estimator.

2. Peto*

Binomial-inspired approximation.

3. Non-parametric bootstrap*

Flexible, but can be computationally heavy.

4. Borkowf's "simple hybrid"*

Binomial-inspired variance with piecewise *"probability-weight"* using KM information.

5. Borkowf's "adjusted hybrid"

Simple + shrinkage → **Variance estimates** > 0
hybrid function

*Can produce zero variance estimates

Q2:
HOW SHOULD STRATA BE WEIGHTED?

Candidates:

1. Inverse-variance (IV) weight

- Precision-based
- More weight to more precise strata
- Undefined if both arm-specific variances in a stratum are zero ⚠️

2. Mantel-Haenszel (MH) weight

- Design-based
- Uses stratum size and allocation ratio
- With 1:1 allocation, reflects stratum proportion

COMPARED TESTS

Test type	Variants
Stratified KM-based Z-tests	All combinations of: 2 weights and 5 variance estimators
Unstratified KM-based Z-tests	All 5 variance estimators' versions
Complementary log-log Z-test	Unstratified (recommended by Klein et al. (2007))
Log-rank tests	Both unstratified and stratified

Methods

SIMULATION DESIGN & EVALUATION CRITERIA

Fixed settings:

- 2 arms, 3 strata + 1:1 allocation ratio
- t^* = 3 years
- One-sized $\boldsymbol{\alpha = 2.5\ \%}$
- 10000 replications per scenario

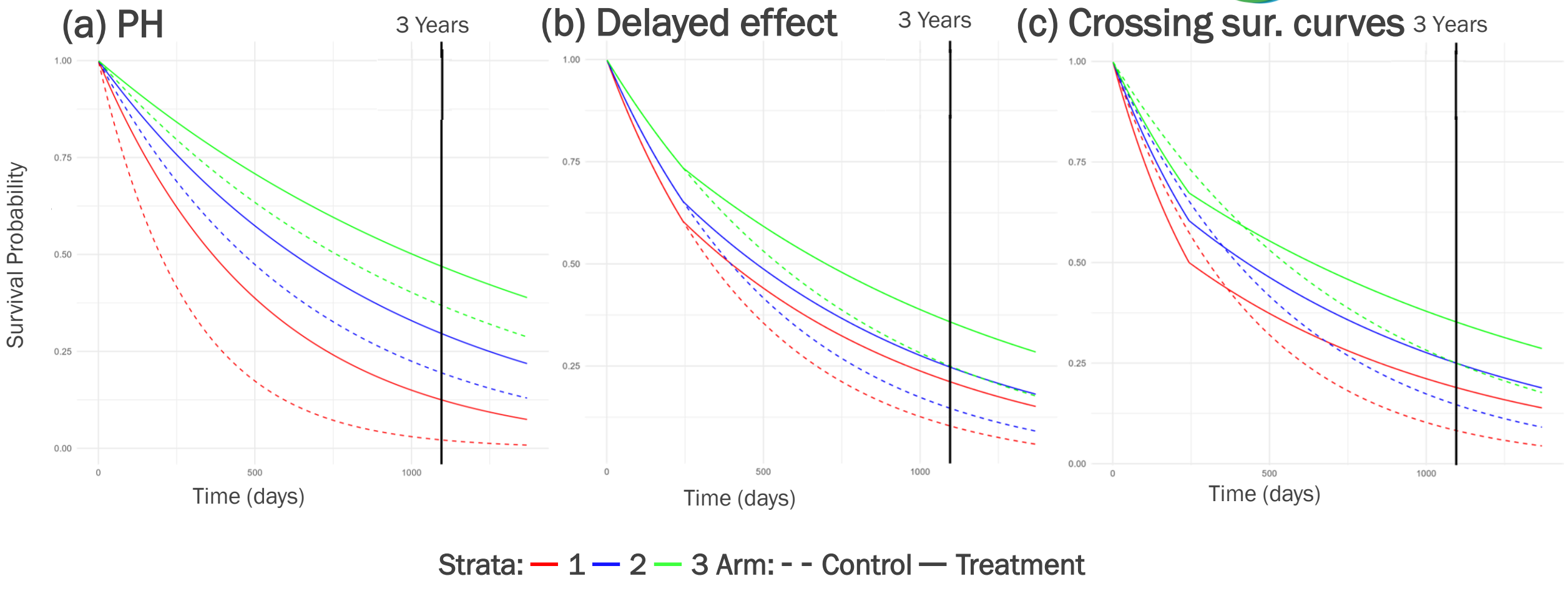
Evaluation:

- Power - punish zero variance cases
- Type I error – exclude zero variance cases
- Frequency of zero variance cases

Varied parameters:

1. Sample size
2. Survival difference at t^*
3. Others…

SIMULATED SCENARIOS: HAZARDS DYNAMICS



RESULTS

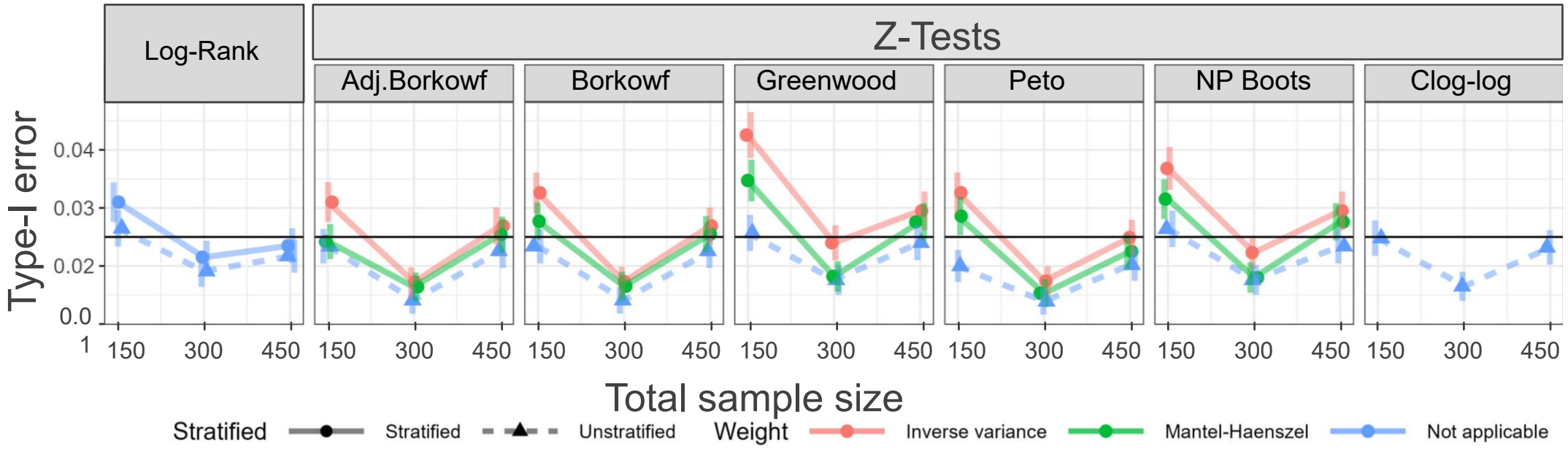
ZERO VARIANCE PROBLEM

Unstratified tests: not affected

Stratified Z-tests:

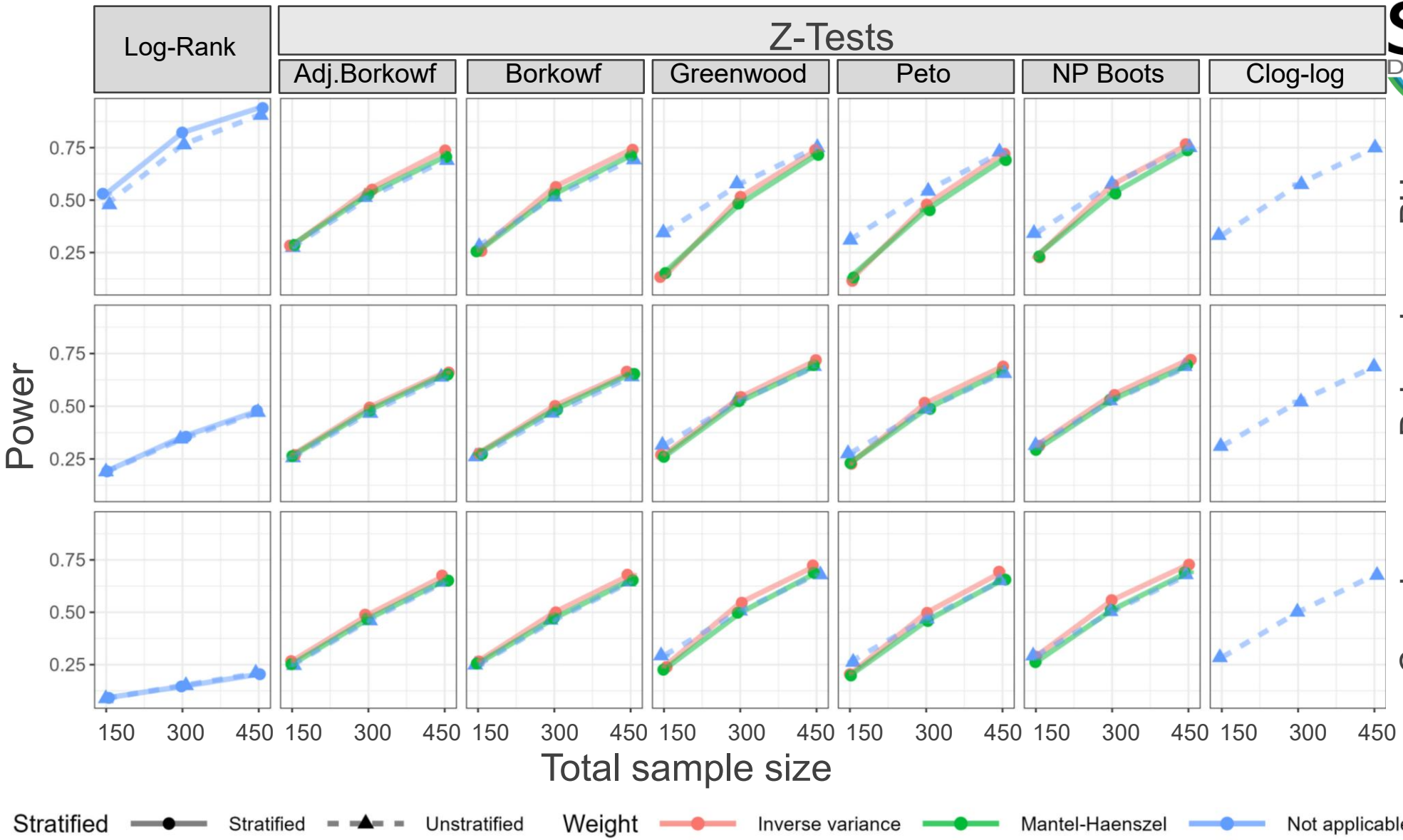
- **Greenwood / Peto:** 5.84% - 11.4%
- **Non-parametric bootstrap**: 2.7% - 6.3%
- **Borkowf's simple hybrid:** 0.17% - 1.5%
- **Borkowf's adjusted hybrid:** 0% → var estimate is always > 0 by construction

IMPACT OF TOTAL SAMPLE SIZE: TYPE-I ERROR



- **Inflation mainly with IV weights**, specially with for Greenwood and bootstrap for small N
- **MH weights** were **more stable** across sample sizes

IMPACT OF TOTAL SAMPLE SIZE: POWER



- **PH:**

  Best = strat. LR

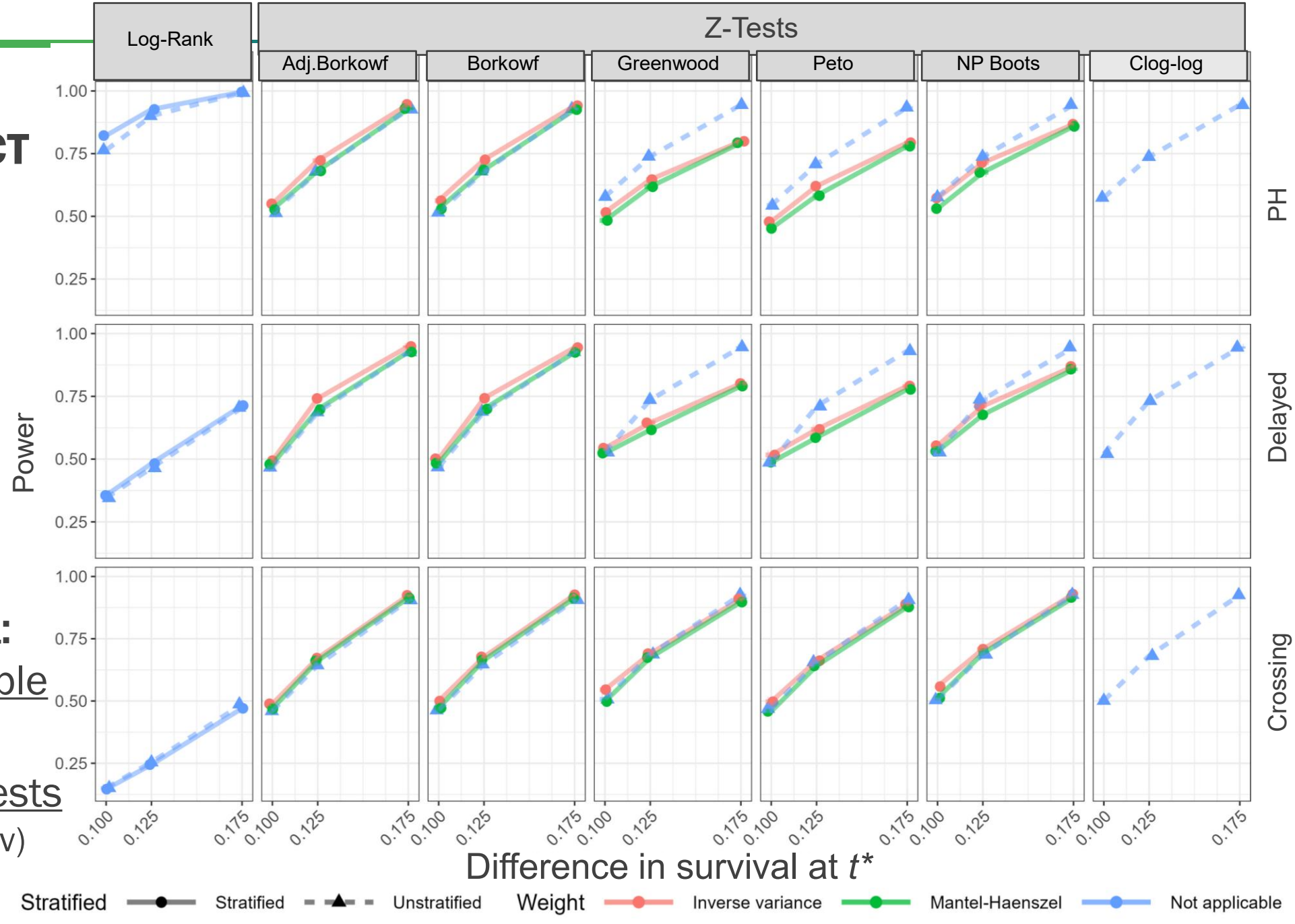
- **Non-PH:**

  LR < KM Z-tests

  Z-tests comparable

TREATMENT EFFECT (DIFFERENCE IN SURVIVAL AT T*)

- LR ≈100%
  for diff=0.175
- Z-tests = similar
  as previous slide
- **PH + delayed effect:**
  Borkowf's both simple and adj. hybrid **>** other strat. KM Z-tests
  (similar for crossing surv)



CONCLUSIONS & RECOMMENDATION

Type I error

- Mostly controlled near nominal 2.5%
- Except for some inflation with IV weights
- MH weights were more stable

Power

- **PH:** log-rank test
- **Non-PH:**
  - **Unstratified:** all comparable + no observed zero variance problem
  - **Stratified:** Borkowf adjusted hybrid best most of the times + no zero var problem by def.

CONCLUSIONS & RECOMMENDATION

Type I error

- Mostly controlled near nominal 2.5%
- Except for some inflation with IV weights
- MH weights were more stable

Power

- **PH:** log-rank test
- **Non-PH:**
  - **Unstratified:** all comparable + no observed zero variance problem
  - **Stratified:** Borkowf adjusted hybrid best most of the times + no zero var problem by def.

For stratified fixed-time comparisons under non-PH:

1. **Borkowf adjusted hybrid variance**

+

2. **Mantel-Haenszel weights**

→ Avoids zero variance cases
→ Preserves type I error
→ Competitive power across non-PH scenarios

→ Z-test benefit: interpretability!

STABURO GMBH

Aschauer Straße 26a
81549 Munich, Germany
+49.89.55271520
www.staburo.de

Contact us:
info@staburo.de

THANK YOU

FOR YOUR

TIME & ATTENTION!

REFERENCES (1/2)

- Borkowf, C. (2005). A simple hybrid variance estimator for the Kaplan–Meier survival function. *[No journal/publisher provided]*
- Brownie, C., Anderson, D. R., Burnham, K. P., & Robson, D. S. (1985). *Statistical inference from band recovery data—a handbook* (2nd ed., Resour. Publ. 156). U.S. Fish Wildl. Serv.
- Chauvel, C., & O'Quigley, J. (2014). Tests for comparing estimated survival functions. *Biometrika, 101*(3), 535–552.
- Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*, 187–202.
- Duke-Robert, J., & Margolis, M. (2018). Public workshop: Oncology clinical trials in the presence of non-proportional hazards. *[No further publication info provided]*
- Fleming, T. R., & Harrington, D. P. (1981). A class of hypothesis tests for one and two sample censored survival data. *Communications in Statistics – Theory and Methods, 10*(8), 763–794.
- Greenland, S., & Robins, J. (1985). Estimation of a common effect parameter from sparse follow-up data. *Biometrics, 41*, 55–68.
- Greenwood, M. (1926). A report on the natural duration of cancer. *[No further publication info provided]*
- Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika, 69*(3), 553–566.
- International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). (1998). *E9 statistical principles for clinical trials.*
- Klinglmüller, F., Fellinger, T., König, F., Friede, T., Hooker, A. C., Heinzl, H., Mittlböck, M., Brugger, J., Bardo, M., Huber, C., Benda, N., Posch, M., & Ristl, R. (2023). A neutral comparison of statistical methods for time-to-event analyses under non-proportional hazards. *[No journal/publisher provided]*
- Kröger, N., et al. (2021). Comparison between 5-azacytidine treatment and allogeneic stem-cell transplantation in elderly patients with advanced MDS according to donor availability (VIDAZA-ALLO study). *Journal of Clinical Oncology, 39*(30), 3318–3327.

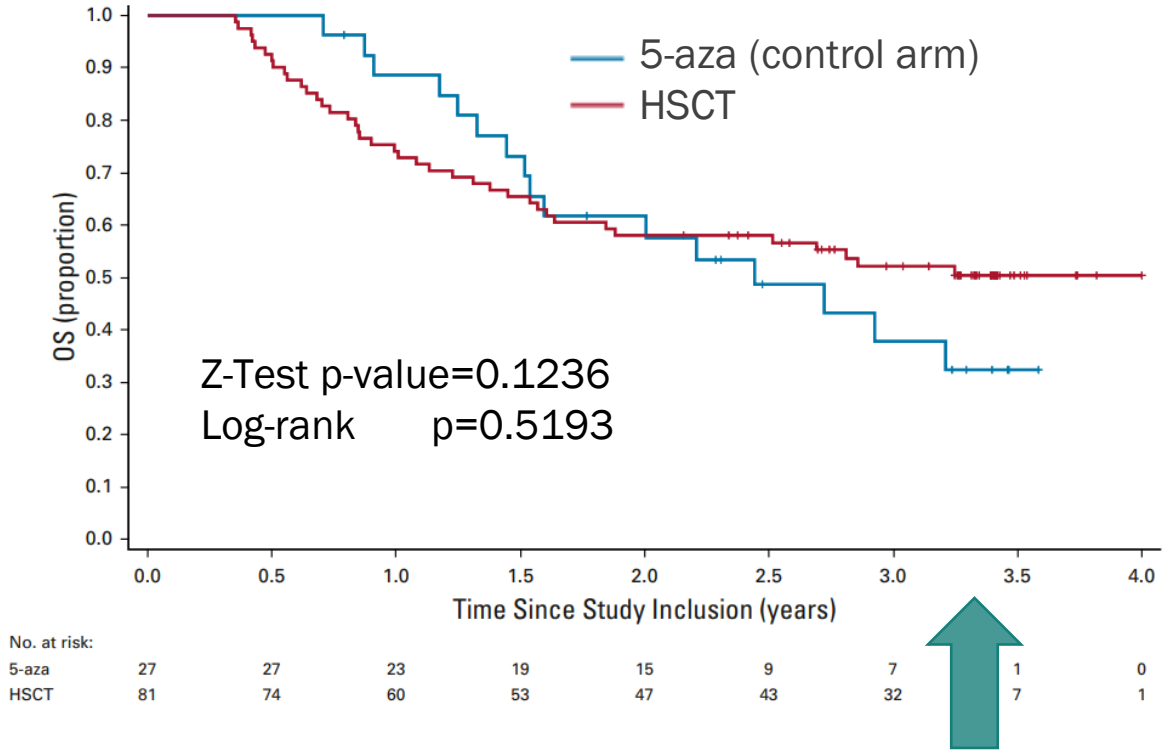
REFERENCES (2/2)

- Lachin, J. M. (2011). *Biostatistical methods: The assessment of relative risks* (2nd ed.). Wiley-Blackwell.
- Lin, R. S., Lin, J., Roychoudhury, S., Anderson, K. M., Hu, T., Huang, B., Leon, L. F., Liao, J. J., Liu, R., Luo, X., Mukhopadhyay, P., Qin, R., Tatsuoka, K., Wang, X., Wang, Y., Zhu, J., Chen, T.-T., Iacona, R., & Proportional Hazards Working Group, C.-P. N. (2020). Alternative analysis methods for time to event endpoints under nonproportional hazards: A comparative analysis. *Statistics in Biopharmaceutical Research, 12*(2), 187–198.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemo. Rep., 50*, 163–170.
- Peto, R., Pike, M., Armitage, P., et al. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer, 35*, 1–39.
- Royston, P., & Parmar, M. (2013). Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology, 13*, 152.
- Shen, Y.-L., Wang, X., Sirisha, M., Mulkey, F., Zhou, J., Gao, X., Zhang, L., Gwise, T., Tang, S., Theoret, M., Pazdur, R., & Sridhara, R. (2023). Nonproportional hazards—an evaluation of the MaxCombo test in cancer clinical trials. *Statistics in Biopharmaceutical Research, 15*(2), 300–309.
- Stummer, W., et al. (2006). Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: A randomised controlled multicentre phase III trial. *The Lancet Oncology, 7*(5), 392–401.
- Yang, S., & Prentice, R. (2010). Improved logrank-type tests for survival data using adaptive weights. *Biometrics, 66*, 30–38. https://doi.org/10.1111/j.1541-0420.2009.01243.x

BACKUP AND EXTENSIVE SLIDES

MOTIVATION: VidazaAllo Study

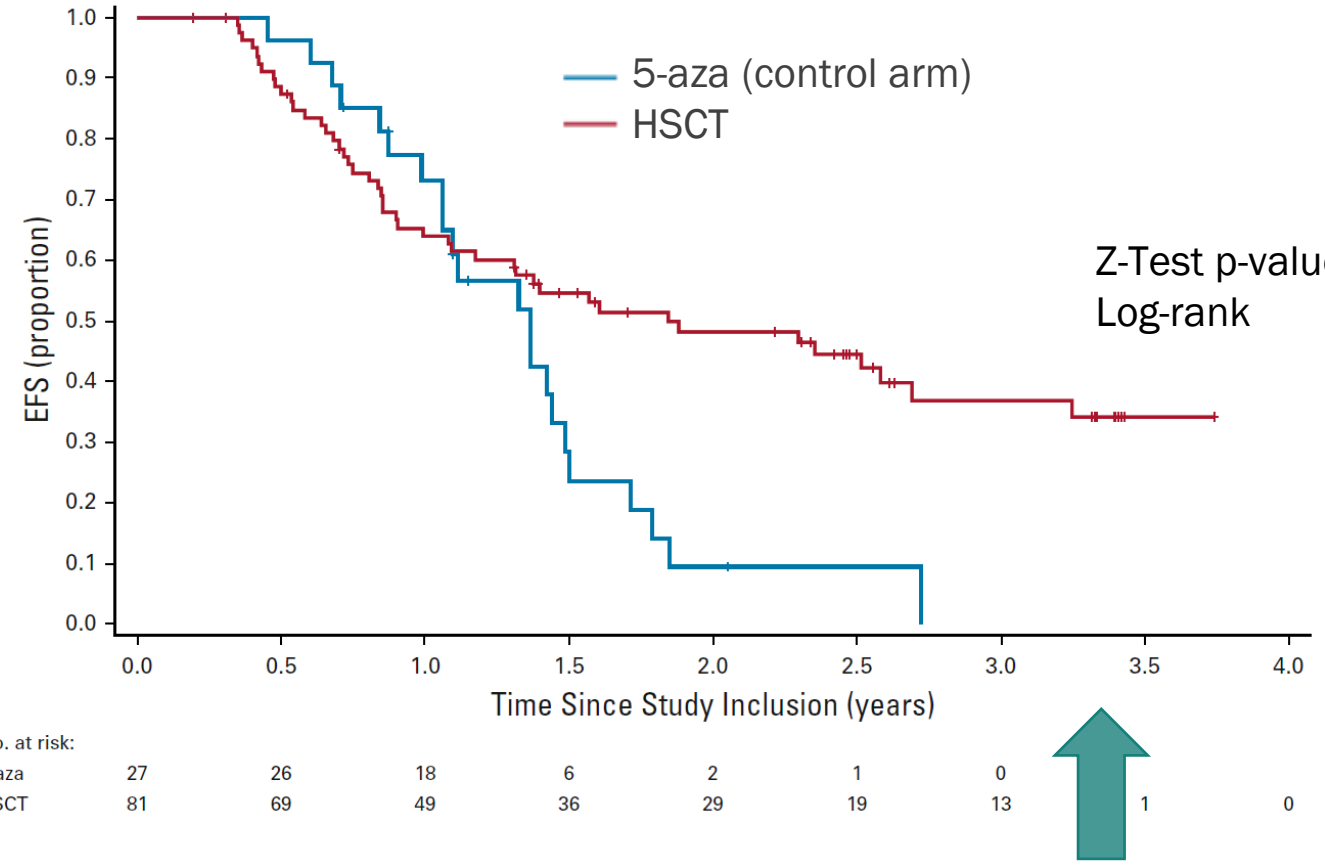
- 108 elderly patients (55–70 years) with advanced myelodysplastic syndromes (MDS)
- Compare allogeneic stem-cell transplantation (HSCT) after 5-azacytidine (5-aza) induction versus continuous 5-aza, dictated by donor availability (biological-assignment clinical trial)
- **Primary endpoint:** OS at 3 years (after assignment)
- **Challenge:** therapy-related mortality: HSCT is curative in terms of long-term survival but has higher mortality right after procedure



Non-Proportional hazards were expected – but no stratified analysis for primary analysis!

Kröger N, Sockel K, Wolschke C, et al. Comparison Between 5-Azacytidine Treatment and Allogeneic Stem-Cell Transplantation in Elderly Patients With Advanced MDS According to Donor Availability (VidazaAllo Study). *Journal of Clinical Oncology*. 2021;39:3318-3327. doi:10.1200/JCO.20.02724

VIDAZA-ALLO STUDY: EVENT-FREE SURVIVAL (EFS)



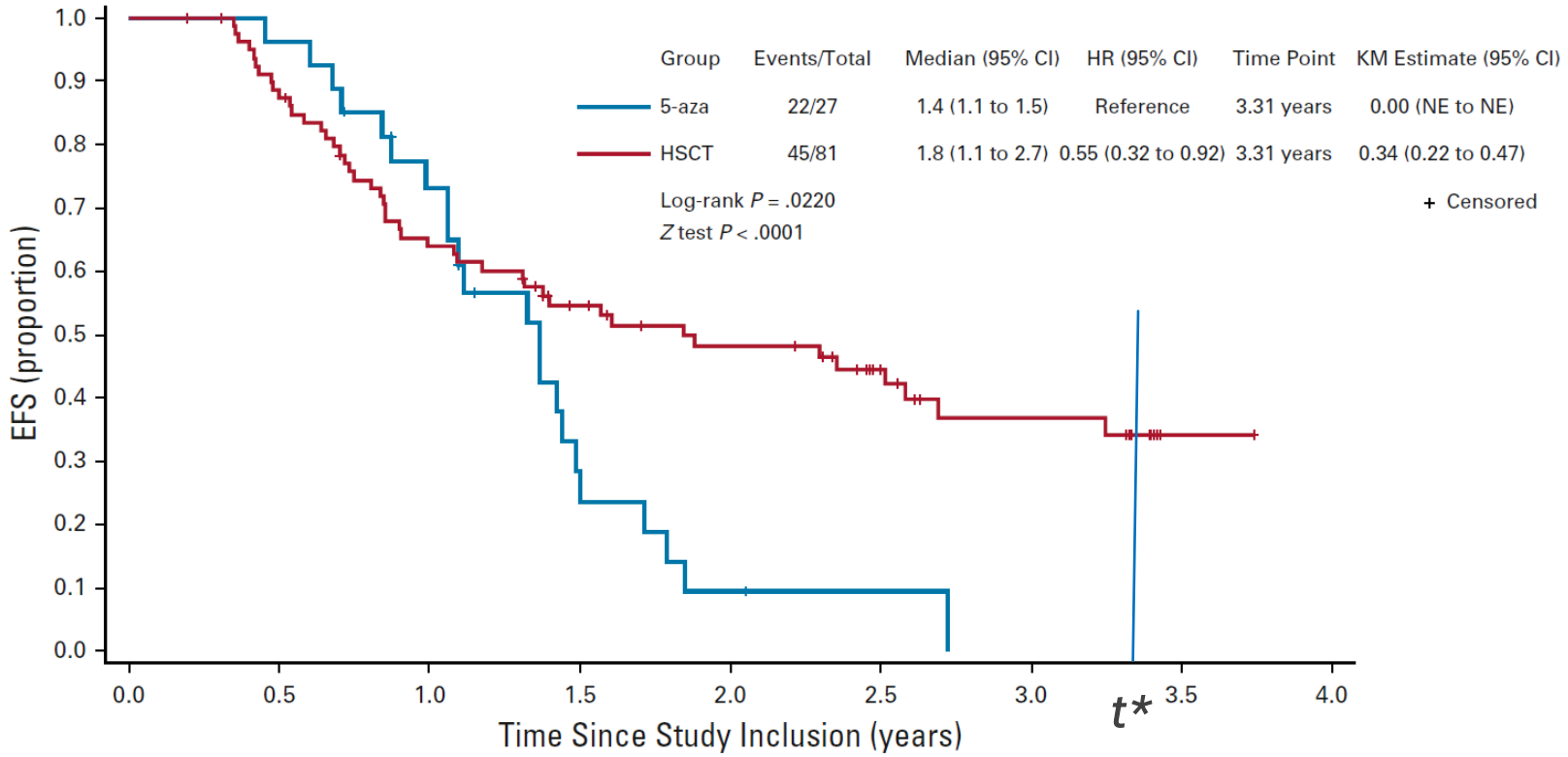
Z-Test p-value<0.0001
Log-rank p=0.0220

- Non-proportional hazards as expected also for EFS not only for OS
- Pronounced effect of HSCT vs 5-aza at 3 years after assignment
- Standard Z-test still computable since Greenwood variance is only zero in one arm but it …

…motivates this work:
is there a better variance alternative?
And how can we incorporate covariates/stratification?

Kröger N, Sockel K, Wolschke C, et al. Comparison Between 5-Azacytidine Treatment and Allogeneic Stem-Cell Transplantation in Elderly Patients With Advanced MDS According to Donor Availability (VidazaAllo Study). *Journal of Clinical Oncology*. 2021;39:3318-3327. doi:10.1200/JCO.20.02724

EVENT-FREE SURVIVAL



No. at risk:

5-aza	27	26	18	6	2	1	0		
HSCT	81	69	49	36	29	19	13	1	0

INTRODUCTION

- In clinical trials, the evaluation of a new drug's efficacy is often based on time-to-event endpoints.

  For oncology trials, examples include:

  - Overall Survival (OS)
  - Progression-Free Survival (PFS)

- Recommendations from regulators:
  - ICH E9: Stratification of prognostic factors during randomization ⟶ incorporation of stratification in the analysis
  - Pre-specification of the analysis ⟶ prevention of post-hoc adjustments ("cherry-picking")

COMPARISON OF TWO SURVIVAL CURVES

Comparison of survival endpoints when proportional hazards (PH) assumption is valid:

- The log-rank (LR) test (Mantel, 1966)
- Cox model (Cox, 1972)

But what if **PH** assumption **is expected to be severely violated**, i.e., in a non-proportional hazards (non-PH) scenario?

- Standard Cox model is miss-specified and LR loses power (Shen et al., 2023; Lin et al., 2020)
- No gold-standard alternative, but many proposed methods including:
  - Fleming-Harrington family and other variations of weighted LR tests (Fleming and Harrington, 1981; Harrington and Fleming, 1982; Yang and Prentice, 2010; Chauvel and O'Quigley, 2014)
  - Restricted Mean Survival Time (RMST) comparison (Royston and Parmar, 2013)
  - MaxCombo test (Duke-Robert and Margolis, 2018), discussion in Shen et al., 2023
  - **Fixed time-point survival rate comparison →** Z-test

Focus of the following Simulations

Z-TEST IN CLINICAL STUDIES

Z-test for **differences in Kaplan-Meier (KM) estimates at fixed time *t**:

- Proposed by Brownie et al. (1985)
- Already employed in phase II (Kröger et al., 2021) and phase III (Stummer et al., 2006) trials.
- Choice of t*: can be relatively straight-forward from a clinical perspective?

VARIANCE ESTIMATORS

- Greenwood

$$\widehat{\sigma}^2{}_{\text{Greenwood: }k,g}\left(\hat{S}_{k,g}(t)\right) = \hat{S}_{k,g}(t)^2 \sum_{t_{k,g,i} \le t} \frac{d_{k,g,i}}{n_{k,g,i}\left(n_{k,g,i} - d_{k,g,i}\right)}$$

with $d_{k,g,i}$ and $n_{k,g,i}$ denote the number of events at t_i and subjects at risk before t_i for stratum k and arm g, respectively.

- Peto

$$\widehat{\sigma}^2{}_{\text{Peto: }k,g}\left(\hat{S}_{k,g}(t)\right) = \frac{\hat{S}_{k,g}(t)^2\left(1 - \hat{S}_{k,g}(t)\right)}{n_{k,g,t}}$$

where $n_{k,g,t}$ is the number of subjects at risk before t in for stratum k and arm g.

- Non-parametric bootstrap

Iterative algorithm.

- Borkowf's "simple hybrid" estimator (Borkowf, 2005)

$$\widehat{\sigma}^2{}_{\text{Borkowf:k,g}}(\hat{S}_{k,g}(t)) = \frac{w_{k,g}(t)^2\left(1 - w_{k,g}(t)\right)}{n_{k,g} - m_{c,k,g}(t)}$$

where

$$w_{k,g}(t) = \hat{S}_{k,g}(t) I_{[0.5 \le \hat{S}_{k,g}(t) \le 1]} I_{[0.5 \le \hat{S}_{k,g}(t) \le 1]} + 0.5 I_{[\hat{S}_{k,g}(t) < 0.5 \le b_{max,k,g}(t)]} + b_{max,k,g}(t) I_{[\hat{S}_{k,g}(t) \le b_{max,k,g}(t) < 0.5]}$$

and

$$b_{max,k,g}(t) = 1 - \frac{m_{e,k,g}(t)}{n_{k,g}}$$

with $m_{e,k,g}(t)$ being the cumulative number of events in stratum k and arm g at time t and $n_{k,g}$ the respective population size.

- Borkowf's "adjusted hybrid" estimator (Borkowf, 2005)

Replace $w_{k,g}(t)$ by

$$w^*_{k,g}(t) = \left(1 - \frac{1}{n_{k,g}}\right) w_{k,g}(t) + \frac{1}{2n_{k,g}}$$

STRATA WEIGHTS

- **Inverse variance weight**

$$w_{IV,k} = \frac{\left(\hat{\sigma}_{k,1}^2 + \hat{\sigma}_{k,2}^2\right)^{-1}}{\sum_{j=1}^{K}\left(\hat{\sigma}_{j,1}^2 + \hat{\sigma}_{j,2}^2\right)^{-1}}$$

$\hat{\sigma}_{k,g}^2$: variance estimative for stratum k=1,…K, and arm g=1,2; undefined if all strata with zero variance in both treatment arms.

- **Mantel-Haenszel weight (Greenland and Robins, 1985)**

$$w_{MH,k} = \frac{n_k m_k/(n_k + m_k)}{\sum_{j=1}^{K} n_j m_j/(n_j + m_j)}$$

n_k, m_k: number of subjects in active treatment arm and in control arm, respectively, in stratum k.

For allocation ratio 1:1 in each stratum, $w_{MH,k}$ equals proportion of subjects in stratum k relative to total number of subjects.

SIMULATION: CLINICAL TRAILS WITH 2 TREATMENT ARMS AND 3 STRATA

- **Survival times:** exponential for control, piecewise exponential for treatment
- **Analysis time t^*:** 3 years
- **Allocation ratio:** 1:1 (stratified)
- **Recruitment time:** 6 months
- **Follow-up time:** at least 3 years
- **Significance level:** alpha=2.5% one-sided
- **Scenarios:** 42 unique scenarios
- **Replicates per scenario:** 10 000

Tested parameters (Reference value is underlined):

- **Total sample sizes:** 150, 300, 450
- **Patient distribution across 3 strata:**

  (200, 50, 50), (100, 100, 100), (50, 50, 200)
- **Effect sizes:** 0.1, 0.125, 0.175
- **Hazard dynamics:**

  PH, crossing survival curves, delayed effects, no treatment effect
- **Timing of hazard changes:** 12, months (more values in appendix)
- **Median censoring time in months:** 120 (10%) (more values in appendix)

Simulation framework was influenced by Klinglmüller et al. (2023) and R package SimNPH was used to generate the data.

POWER AND TYPE-I ERROR

- Power – punish zero variance cases:

$$Power = \frac{1}{n_{replicates}} \sum_{i=1}^{n_{replicates}} I(p\_value(i) \leq \alpha \; AND \; problem\_var(i) = FALSE)$$

where α = 2.5% and *problem_var(i)* indicates that simulation run *i* had at least one estimated variance as 0 or NaN in an entire treatment group, if unstratified, or in a stratum if stratified.

- Type I error – exclude zero variance cases:

$$Type\_I\_error = \frac{1}{n_{replicates} - n_{problem\_var}} \sum_{i=1}^{n_{replicates}} I(p\_value(i) \leq \alpha \; AND \; problem\_var(i) = FALSE),$$

- Confidence interval with binomial formula

STATISTICAL TESTS

- **The complementary log-log (cloglog) transformation of the Z-test:** unstratified

$$Z_{cloglog} = -\frac{log\left(-log\left(\hat{S}_1(t^*)\right)\right) - \log\left(-\log\left(\hat{S}_2(t^*)\right)\right)}{\sqrt{\frac{\tilde{\sigma}_1(t^*)^2}{log\left(\hat{S}_1(t^*)\right)^2} + \frac{\tilde{\sigma}_2(t^*)^2}{log\left(\hat{S}_2(t^*)\right)^2}}}$$

where $\tilde{\sigma}_g(t^*)^2$ is the sum part of the Greenwood estimator: $\tilde{\sigma}_g(t^*)^2 = \sum_{i:t_{g,i} \leq t^*} \frac{d_{g,i}}{n_{g,i}(n_{g,i} - d_{g,i})}$

- **Log-rank:** unstratified and stratified

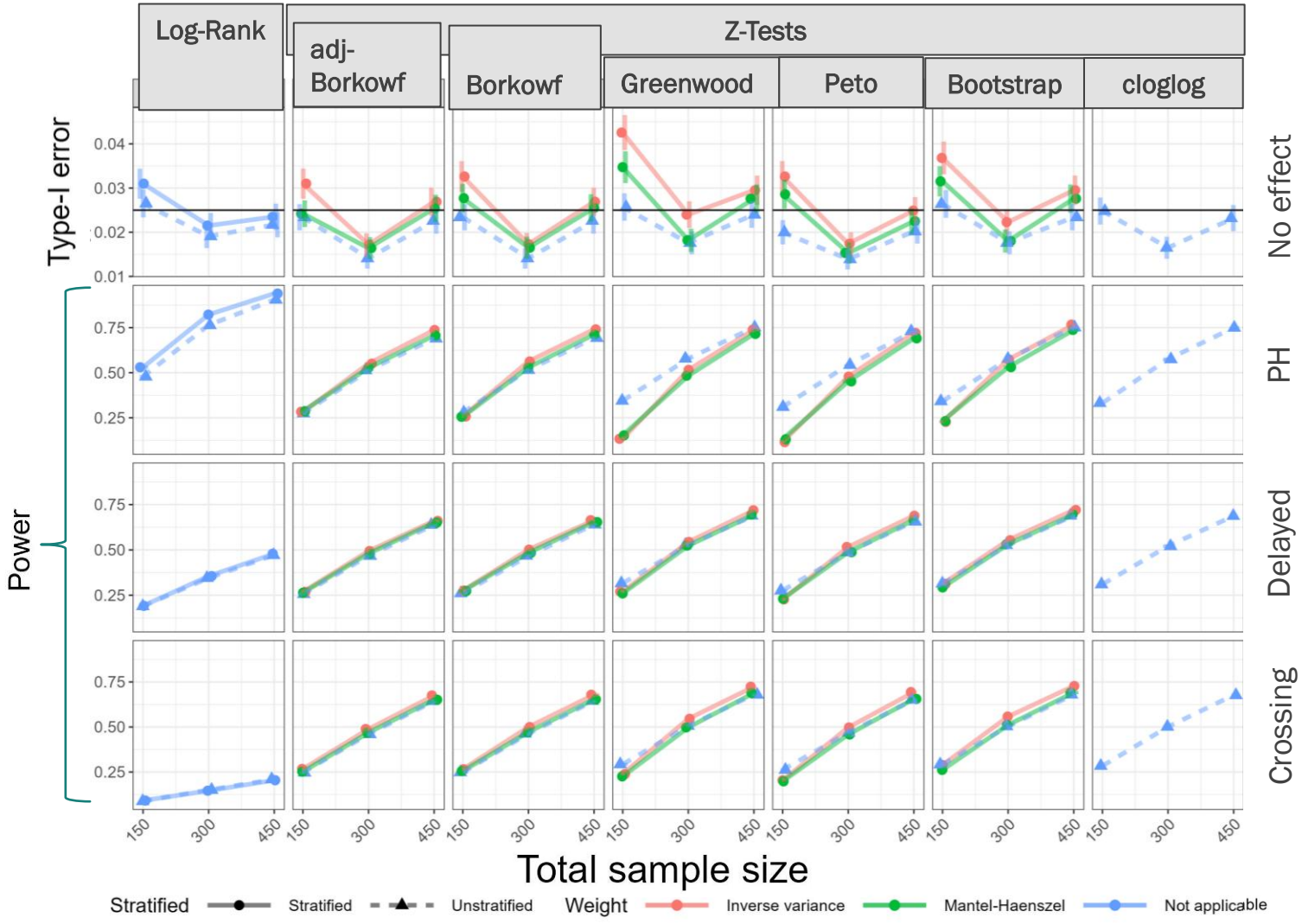
$$LR = \frac{U_L}{\sqrt{V_L}} \sim N(0,1) \text{ and } LR_{Strat} = \frac{U_{L,strat}}{\sqrt{V_{L,strat}}} \sim N(0,1)$$

where U_L = cumulative difference between number of observed events in arm 1 and

V_L = variance of the total number of observed events in arm 1

Klein, J. P., Logan, B., Harhoff, M., & Andersen, P. K. (2007). Analyzing survival curves at a fixed point in time. *Statistics in Medicine, 26*(24), 4505–4519.

IMPACT OF TOTAL SAMPLE SIZE



staburo DATA SCIENCE

Type-I Error

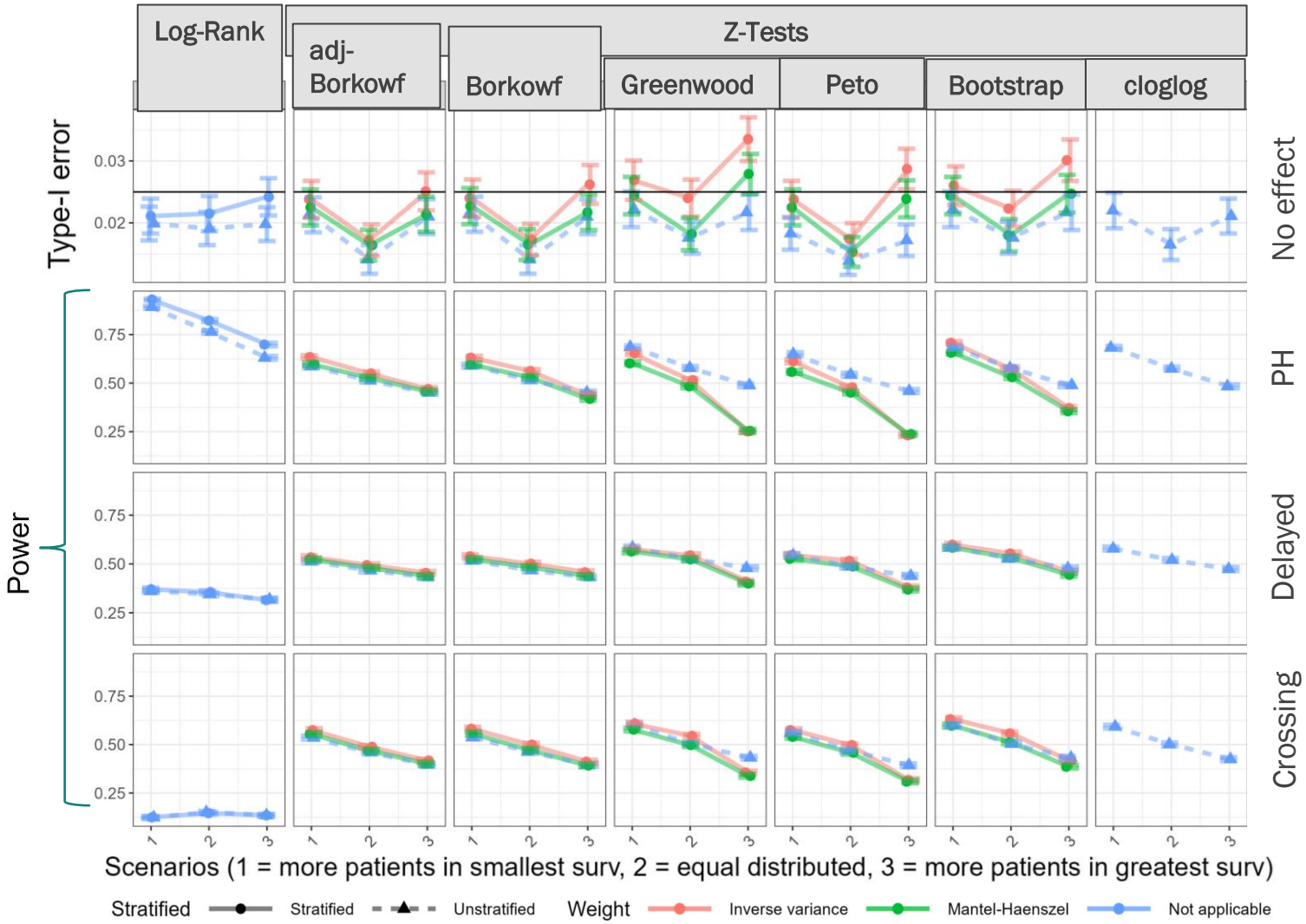
- Slightly inflated for **inverse variance weighting** especially when combined with **Greenwood** and **bootstrap** for **small N**
- **Therefore, use MH weights and not IV!**

Power

- **Log-rank** tests **best for PH** with marginal advantage for stratification
- **Log-rank loses** power below KM Z-tests **for both selected non-PH scenarios**
- KM **Z-tests comparable but unstratified better for Greenwood, Peto and Bootstrap for small N**

Therefore, under non-ph, (adjusted) Borkowf with MH is recommended for stratified tests!

PATIENT DISTRIBUTION ACROSS STRATA



Type-I Error

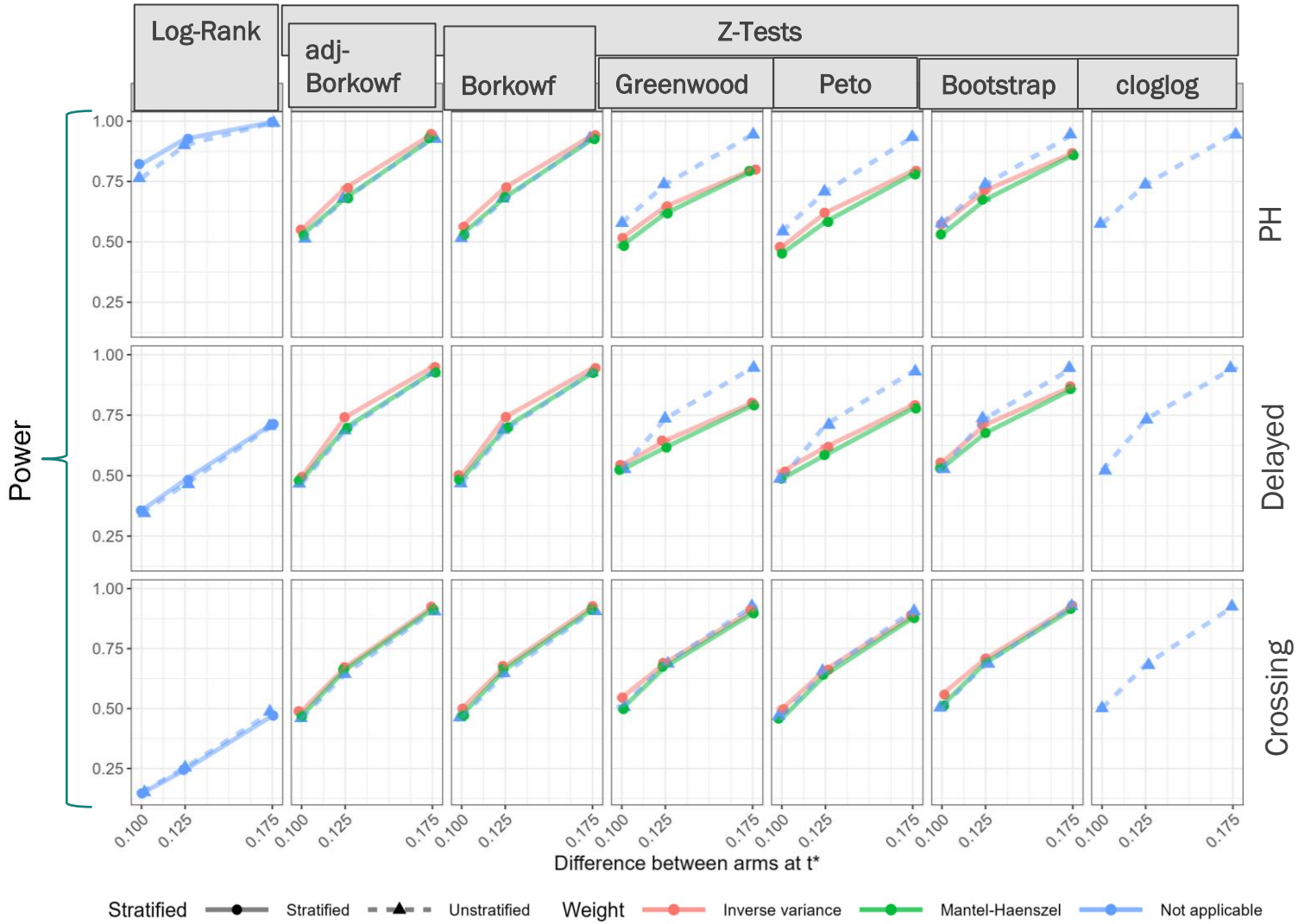
- Slightly inflated for **inverse variance weighting** for **Greenwood** and **bootstrap** in scenario 3 (2/3 of patients is strata with longest median)

Power

- Similar results as previous slide
- **Scenario 3: Borkowf's simple/adjusted hybrid** estimators **outperforms** other stratified KM Z-tests

Therefore, under non-ph, (adjusted) Borkowf with MH is recommended for stratified tests!

TREATMENT EFFECT (DIFFERENCE IN SURVIVAL AT T*)



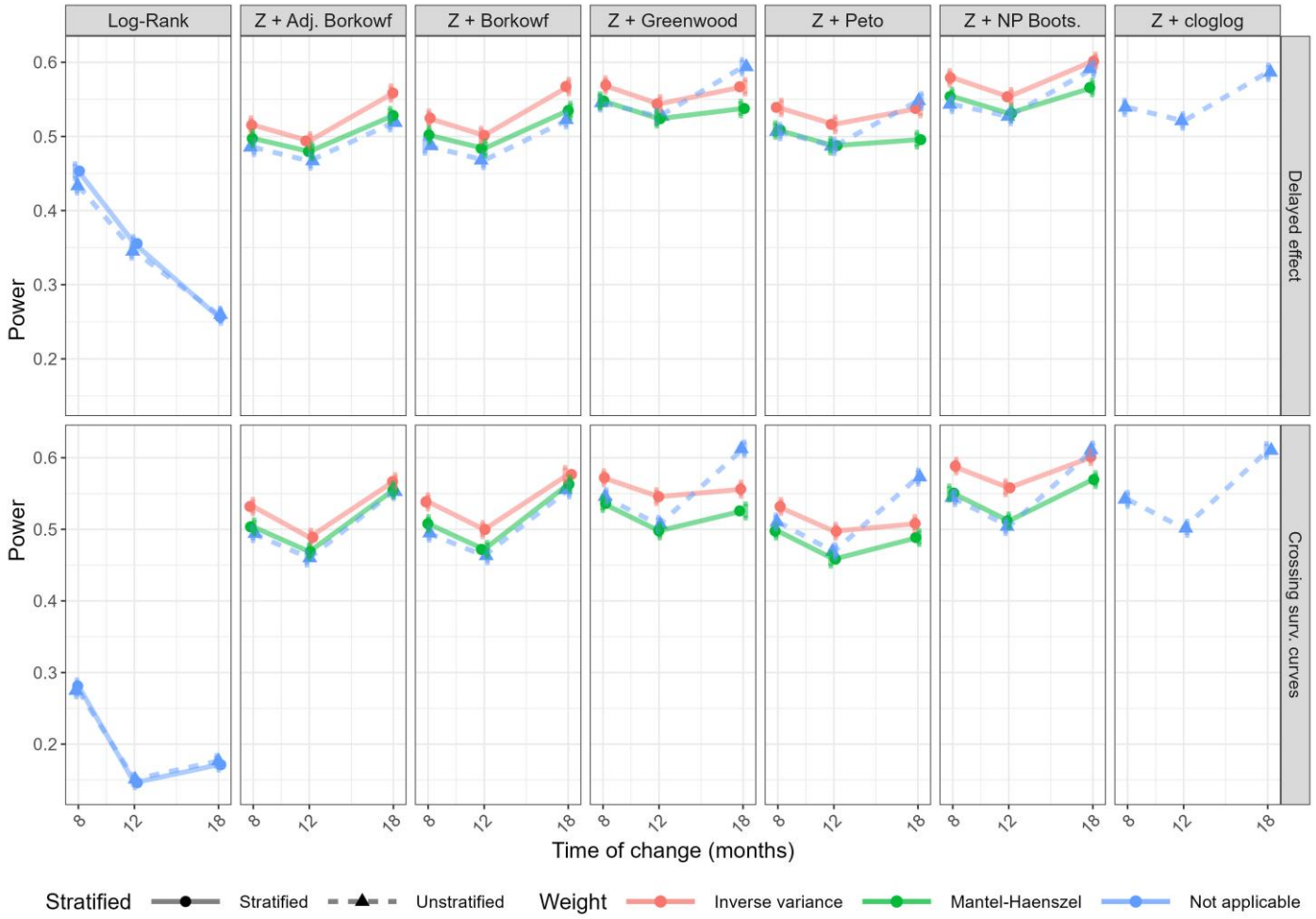
Note: **Same** survival difference **in each stratum** → **different** hazard ratios **across strata**

Power

- Log-rank ≈100% power for a difference of 0.175
- KM rate tests similar results as previous slides
- **PH + delayed** effect: **Borkowf's simple/adjusted hybrid** estimators **outperform** other stratified KM Z-tests (while all are the same for crossing)
- Unstratified tests all with consistently good and similar performance (Cloglog, Greenwood, Peto, bootstrap)

Therefore, under non-ph, (adjusted) Borkowf with MH is recommended for stratified tests!

TIME OF CHANGE IN TREATMENT HAZARD

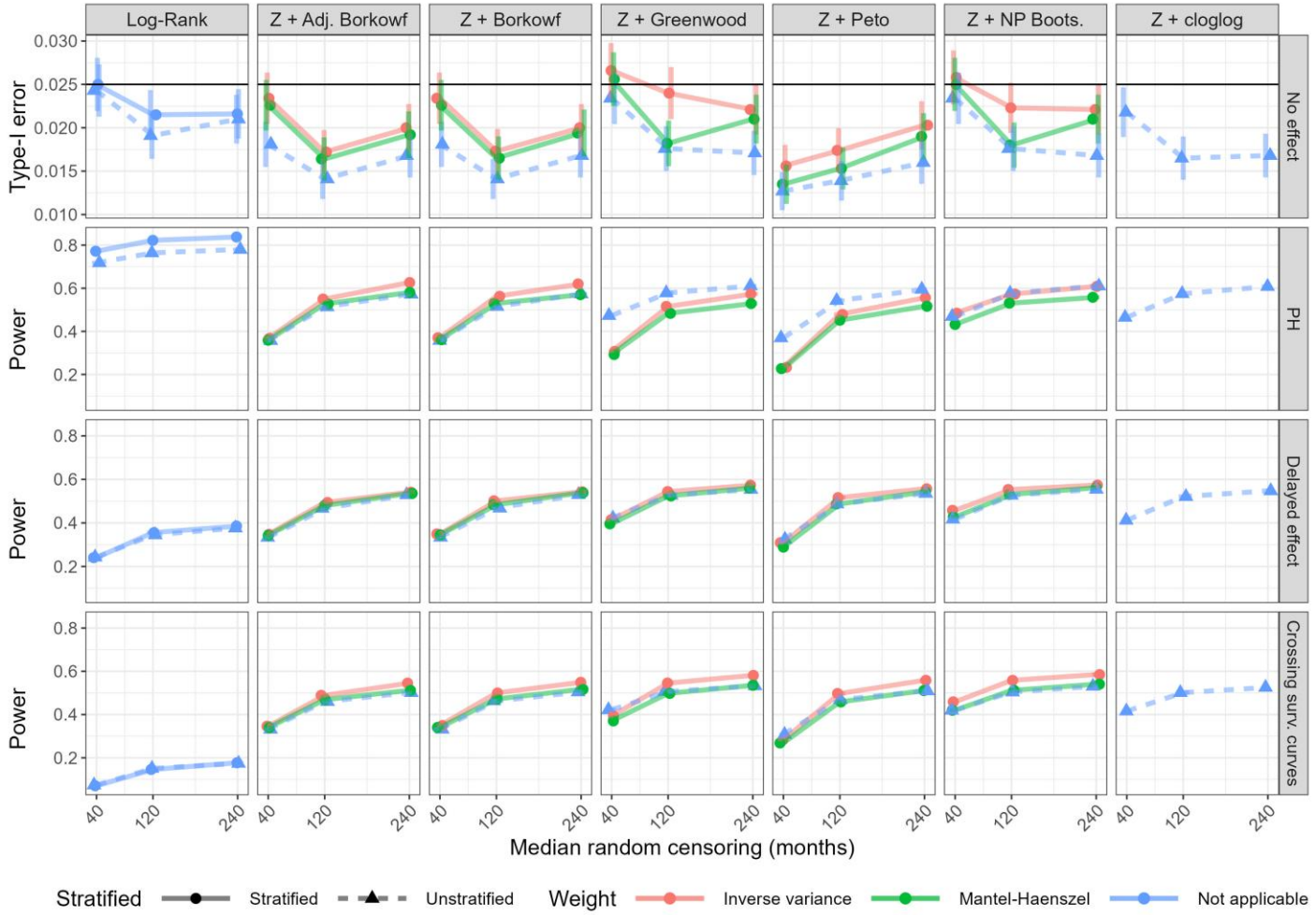


Note: Fixed treatment effect = 0.1 (difference in survival at 3 years) across all strata

Power

- Peto often lowest power among stratified estimators

AMOUNT OF CENSORING



Type-I Error

- Heavy censoring ⟶ Type I error slightly inflated for Greenwood & nonparametric bootstrap (both IV & MH weights)

Power

- Log-rank tests (LR) ⟶ Highest power in PH scenarios, loses power as PH assumption is violated
- KM rate tests ⟶ Similar performance across estimators
- Greenwood ⟶ Power loss in stratified PH cases
- Peto ⟶ Power loss in high censoring cases

Censoring times median in months: 40 (20%), 120 (10%), 240 (5%)

4. CONCLUSIONS

- **Type-I errors acceptable:** sometimes unexpectedly small, sometimes too high for stratified Z statistics with Greenwood, Peto, Bootstrap
- Log-rank: best test under PH, worst test under non-PH
- Sometimes power highest for unstratified compared with stratified Z tests, despite lower type-I-error.
- Under non-PH, unstratified Z-tests are comparable in terms of Type-I-error and power. Zero variance not a problem. Simplest unstratified tests (Z-test + Greenwood or c-loglog) are OK.
- Under non-PH and stratified tests, (adjusted) Bowkowf is best in terms of power (up to exceptions).
  We recommend MH weights because of potential alpha-inflation with IV weights.
  - Other practical idea for stratified analysis:
    1. Z-test with MH weights and Greenwood (“gold standard” variance)
    2. If Greenwood equal 0 in one stratum (both treatment arms), use adjusted Borkowf for all strata.

STUDY LIMITATIONS

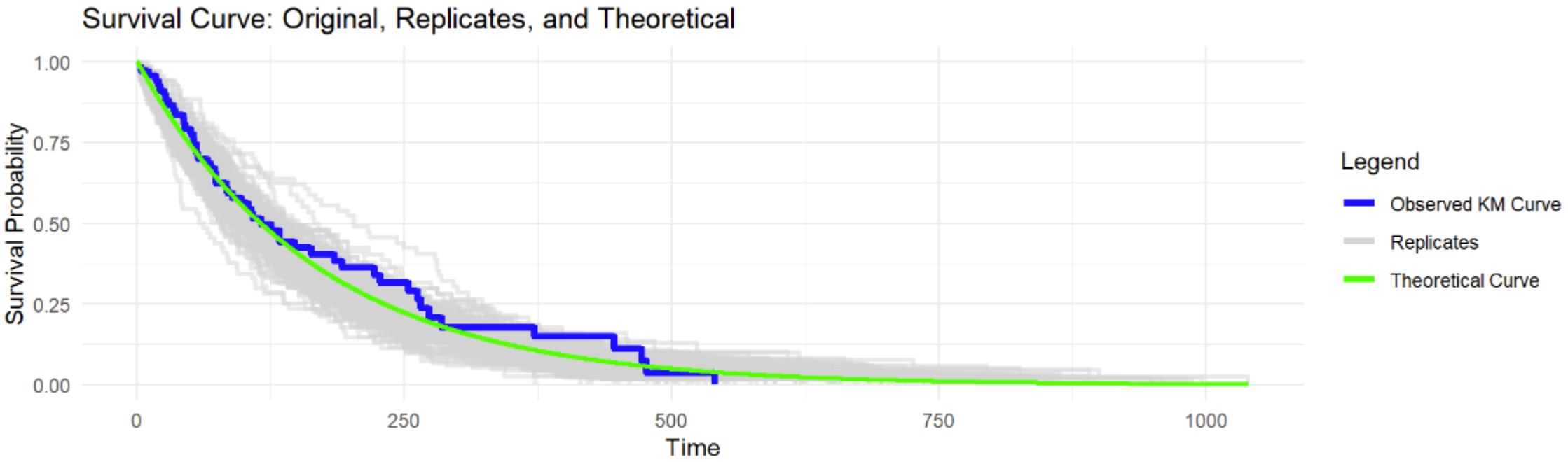
- Piecewise exponential distributions used for simulation ⟶ restrictive but broad hazard dynamics
- Sample size analysis used instead of number of observations & Log-rank test power depends on event count ⟶ May create unfair comparisons
- Zero variance estimates penalized in power calculation
- Most scenarios are in power regions far below the usual values (>=80%)

- Whole survival curves + additional endpoints can still be considered to assess the risk-benefit profile

PRELIMINARY SIMULATION: EVALUATING VARIANCE ESTIMATORS

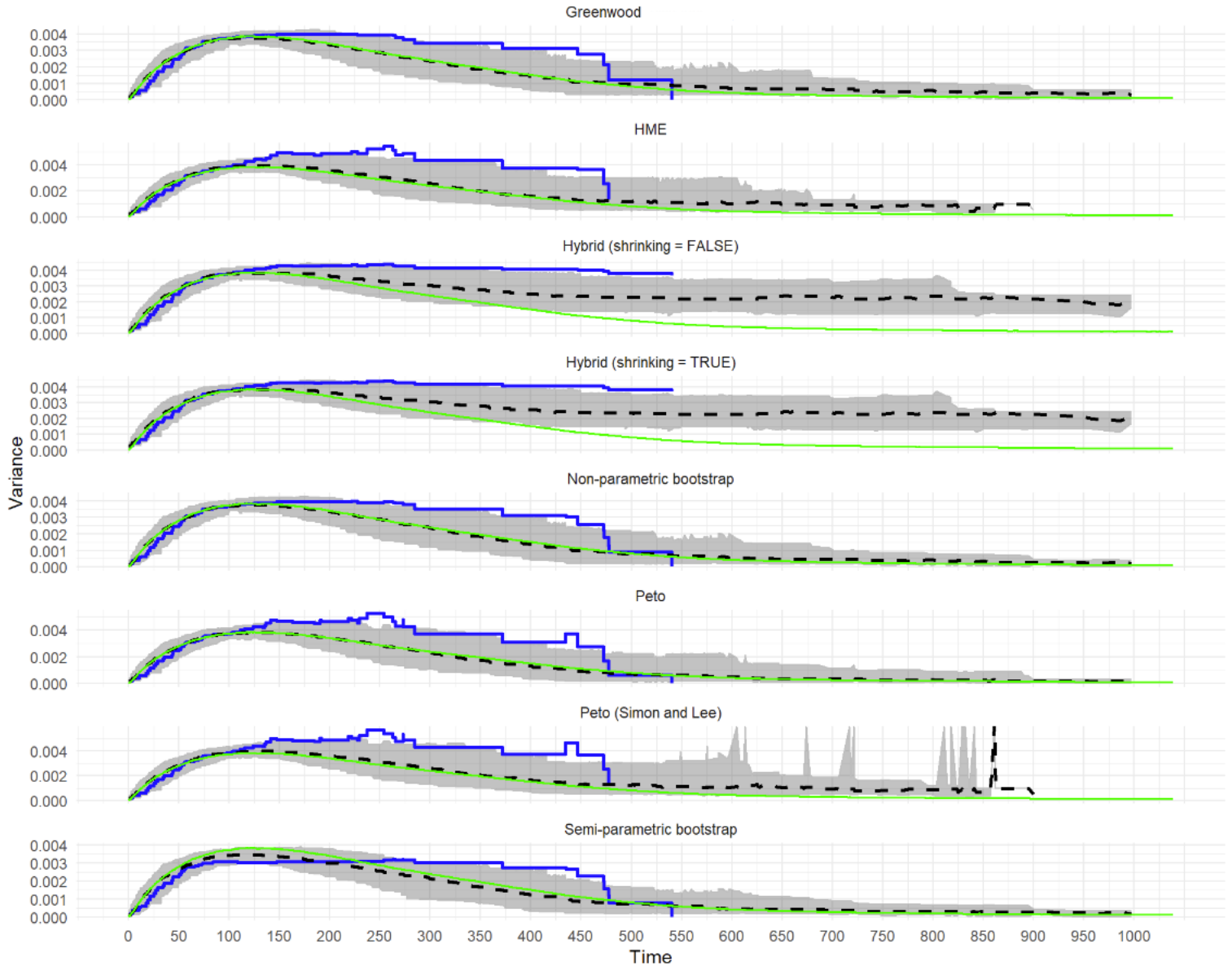
- **Objective:** assess performance and efficiency of variance estimators.
- **Motivation:** for each variance estimator that is included in the main simulation, 3 tests are added to the comparison (unstratified, MH and IV)
- **Design:**
  - Single survival curve simulation with random censoring.
  - Exponential distributions for both censoring and event times.
  - Two sample sizes: medium (70) and small (10).
- **Simulation parameters:**
  - Event hazard rate: 0.006 (Median time to event: ~3.8 months)
  - Censoring hazard rate: 0.001 (Median time to censoring: ~22.8 months)
  - 20,000 bootstrap samples
  - 10,000 replicates for true variance estimation
  - 100 standard replicates for performance validation.

PRELIMINARY SIMULATION – 70 PATIENTS



RESULTS FOR PRELIMINARY SIMULATION – 70 PATIENTS

- **Greenwood:** Observed variance follows true variance closely, with fluctuations towards the end.
- **HME:** Similar to Greenwood with minor deviations.
- **Hybrid (shrinking = FALSE):** Shows moderate accuracy but diverges towards the end.
- **Hybrid (shrinking = TRUE):** Moderate accuracy, slight overestimation towards the end.
- **Non-parametric bootstrap:** Good fit, some variability towards later times.
- **Peto:** More fluctuations, especially towards the end.
- **Peto (Simon and Lee):** Significant instability and high variability in later times.
- **Semi-parametric bootstrap:** Generally follows true variance with minor deviations.



PRELIMINARY SIMULATION – SELECTED ESTIMATORS

Selected estimators for main simulation:

1. **Greenwood:** Follows true variance closely, reliable performance.
2. **Hybrid (shrinking = FALSE):** Good overall performance with slight overestimation.
3. **Hybrid (shrinking = TRUE):** Consistent performance, ensures variance estimate is greater than zero.
4. **Non-parametric bootstrap:** Reliable, good fit, widely used.
5. **Peto:** Stable, especially in smaller sample sizes, despite some end fluctuations.

Excluded estimators:

- **HME:** Similar to Greenwood, but not included due to redundancy.
- **Peto (Simon and Lee):** High instability and variability, less reliable.
- **Semi-parametric bootstrap:** Minor deviations, no distinct advantage over non-parametric bootstrap.