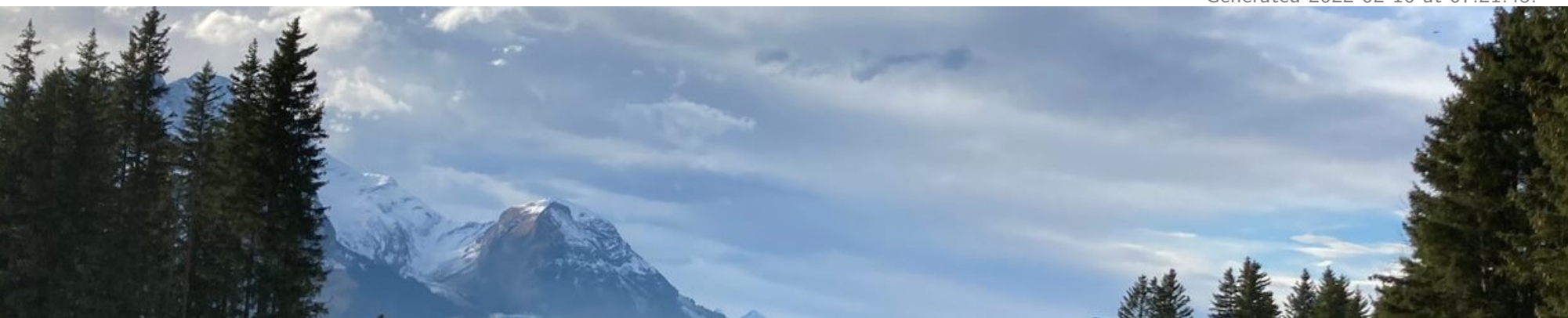
# Efficient effect estimation in pre-specified subgroups in forest plots for a time-to-event endpoint

*Kaspar Rufibach*
*Methods, Collaboration & Outreach Group, Department of Biostatistics, Roche Basel*
*Joint work with Marcel Wolbers (Roche) and Ke Li (then MSc student UZH)*
*University of Ulm, 10 February 2022*

Generated 2022-02-10 at 07:21:45.



# Motivating example

# (De)Motivating example

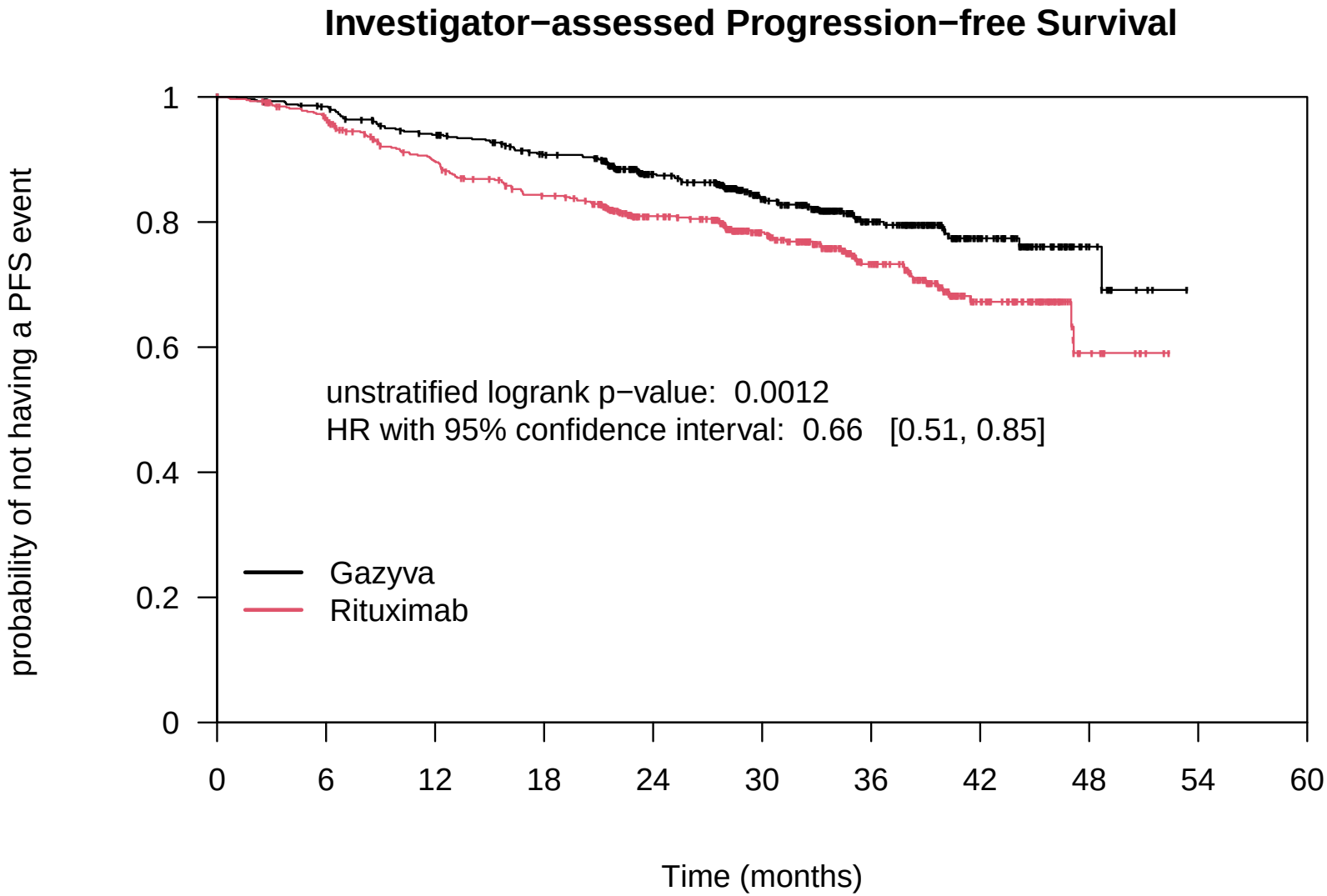
# Gallium

- **Population**: Treatment-naive follicular lymphoma (FL) patients.
- **Comparison**: Rituximab + chemotherapy vs. **Obinutuzumab** + chemotherapy.
- **Phase III, 1:1 randomized, open-label** clinical trial.
- Primary endpoint: investigator-assessed **progression-free survival**.
- 1202 patients.

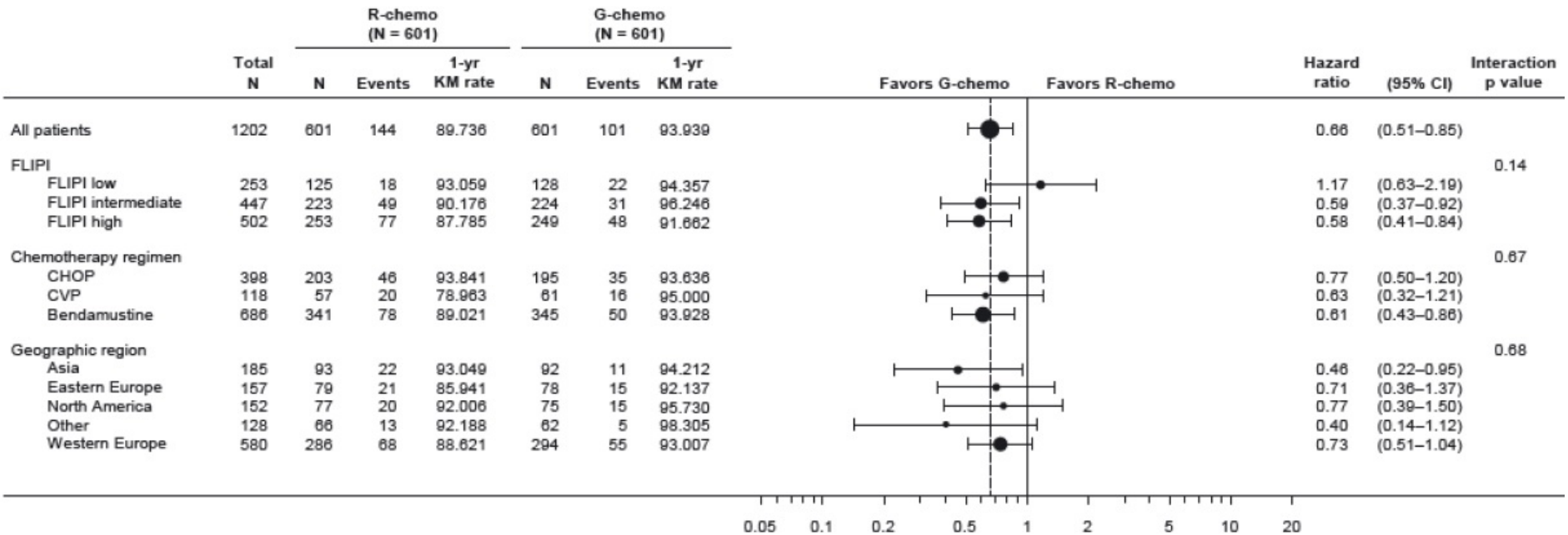
Marcus et al. (2017), NEJM.

Pre-defined subgroups: 15 variables $\Rightarrow$ 49 subgroups.

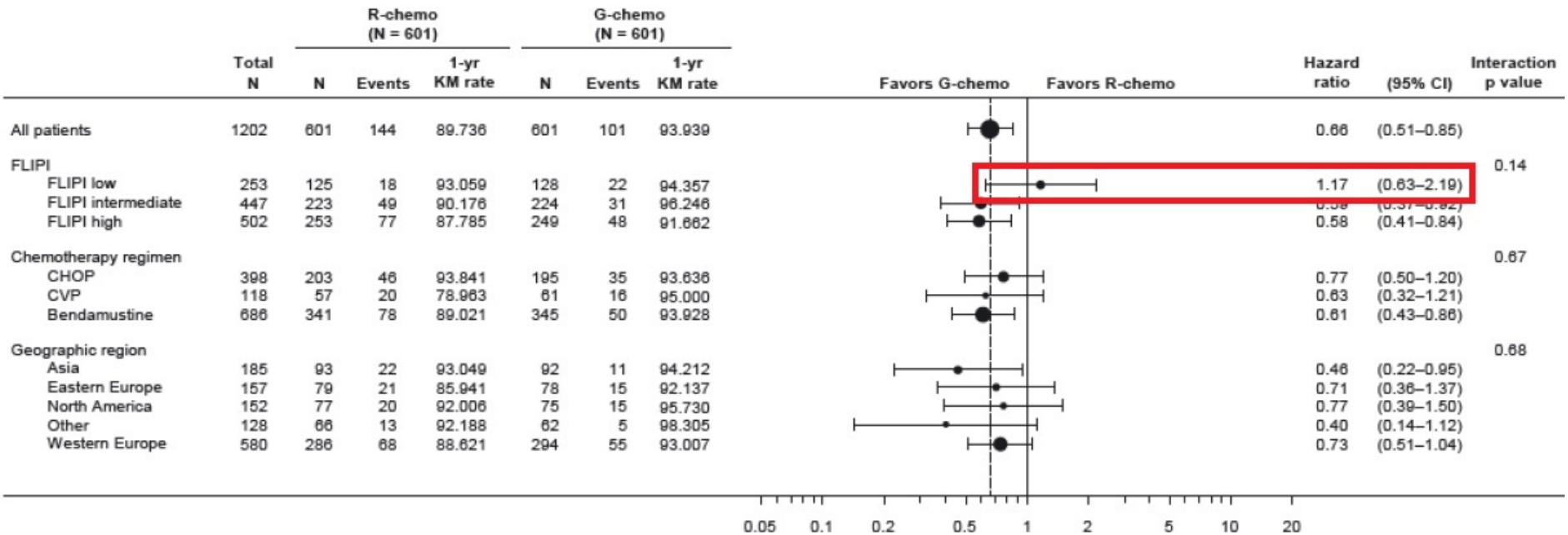
# Gallium: PFS


Investigator−assessed Progression−free Survival
probability of not having a PFS event
1
0.8
0.6
0.4
0.2
0
unstratified logrank p−value: 0.0012
HR with 95% confidence interval: 0.66 [0.51, 0.85]
Gazyva
Rituximab
0
6
12
18
24
30
36
42
48
54
60
Time (months)
601 570 536 502 405 278 168 75 13
601 562 505 463 378 266 160 68 10

# Gallium: forest plot of stratification factors



# Gallium: forest plot of stratification factors



# FLIPI

**FLIPI**: prognostic score in follicular lymphoma.

Effects:

- Overall: 0.66.
- FLIPI low: 1.17, based on 40 events.
- FLIPI intermediate: 0.59.
- FLIPI high: 0.58.

# EMA feedback

Major objection!

> **In the FL FLIPI low...subgroup, there seems to be no difference between the two treatment arms. Therefore, the MAH is asked to discuss the results in the FL FLIPI low patients...**

Company feedback: usual discussion about why one should not overinterpret this effect.

# Statistician get back to work!


"Data don't make any sense,
we will have to resort to statistics."

# Problem statement

# Setup

Confirmatory RCT, **powered** to detect effect in full population.

Pre-defined subgroups:

- 10 or more categorical baseline variables defining $\geq 20$ subgroups.
- Subgroups generated by different variables overlapping.

**Goals** of subgroup analyses:

- Assess **magnitude** of treatment effect in major subgroups.
- Investigate **homogeneity** of treatment effects across subgroups.
- Determine appropriate patient population for treatment use.

# Two standard subgroup estimates

Naive subgroup-specific effect estimate:

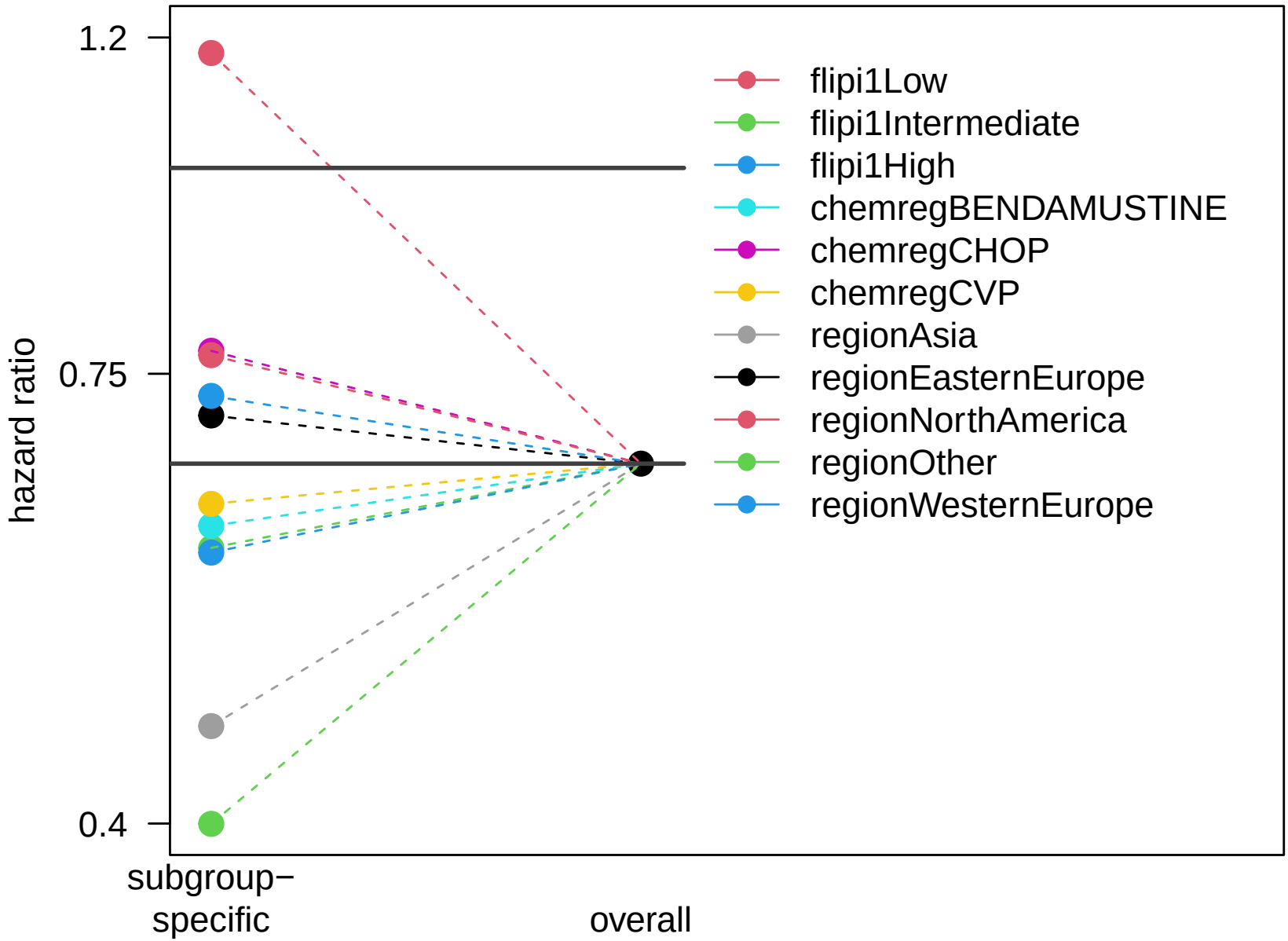
- **Unbiased**,
- **large variability**.

Naive overall effect estimate:

- **Biased** for a given subgroup,
- **low variability**.
- "In the absence of a critically interpreted subgroup analysis for which a high degree of confidence is warranted, the **best estimate of effect for a subgroup is the overall effect**." Oxman (2012).

Question: Can we interpolate between these two extremes?

# Naive estimates



# Idea

# Idea

Goal: Get a **shrunken** estimate.

Idea: bias naive subgroup-specific estimates to reduce variability $\Rightarrow$ reduce mean-squared error (MSE).

Two approaches:

1. Penalized **composite likelihood**.
2. **Marginalization** of penalized model for entire data.

**No penalization of overall treatment effect**.

# Method 1: penalized composite likelihood

# Penalized composite likelihood

**Model** for subgroup $\mathcal{S}_k$ with treatment indicator $z_i$:

$$h_i(t) \quad = \quad h_{0,k}(t) \cdot \exp(\beta_{\text{trt}} \cdot z_i + \beta_k \cdot z_i), i \in \mathcal{S}_k,$$

with

- $\beta_{\text{trt}}$: "overall" treatment effect,
- $h_{0,k}$: subgroup-specific baseline hazard,
- $\beta_k$: subgroup-specific deviation from "overall" effect.

Effect specific to subgroup $\mathcal{S}_k$: $\widehat{\beta}_{\text{trt}} + \widehat{\beta}_k$.

Estimation? Ignore overlap between subgroups, maximize **penalized composite log-likelihood**

$$\sum_{k=1}^{K} l_k(\beta_{\text{trt}}, \beta_k) - \lambda \sum_{k=1}^{K} |\beta_k|^q.$$

Penalties: $q = 1$ for Lasso, $q = 2$ for ridge.

# Penalized composite likelihood

Composite likelihood:

- Composite score function is linear combination of "valid" score functions.
- Provides **unbiased** estimate under usual regularity conditions on each likelihood component.
- Asymptotics: based on standard estimating equations theory.

Varin et al. (2011).

# Method 2: marginalization of a frequentist penalized Cox model

# Marginalization of penalized Cox model

**Global** model for full dataset with main effects and subgroup-treatment interactions for all subgroups $k$:

$$h_i(t) \quad = \quad h_0(t) \cdot \exp(\beta_{\text{trt}} z_i + \alpha_1 s_{1i} + \ldots + \alpha_K s_{Ki} + \beta_1 s_{1i} z_i + \ldots + \beta_K s_{Ki} z_i).$$

Fit global model using penalized likelihood.

**Non-collapsibility** of hazard ratio:

- **Estimand** $\beta_{\text{trt}}$ depends on whether we add covariates or not, even if covariates are not associated with response.
- Even if (conditional) hazard functions proportional between treatments and within every subgroup $\Rightarrow$ marginal hazard functions are **not proportional**, Ford et al. (1995).

# Marginalization of penalized Cox model

How to compute subgroup-specific treatment effect from global model?

1. Predict **covariate-dependent survival function** for each subject, for both control and treatment.
2. Predict "marginal" survival function in subgroup $\mathcal{S}_k$ for control and treatment through **averaging across subjects**.
3. Approximate hazard ratio in subgroup $\mathcal{S}_k$ through "discrete" **average hazard ratio**:

$$\text{AHR}_{OC} = \frac{\int S_{k,\text{ctrl}}(t) f_{k,\text{trt}}(t) \mathrm{d}t}{\int S_{k,\text{trt}}(t) f_{k,\text{ctrl}}(t) \mathrm{d}t}.$$

   Reduces to hazard ratio if hazards are proportional.

4. $\text{AHR}_{OC}$ corresponds to **odds of concordance**:

$$\text{AHR}_{OC} = \frac{P(T_{\text{ctrl}} > T_{\text{trt}})}{1 - P(T_{\text{ctrl}} > T_{\text{trt}})},$$

   see Schemper et al. (2009).

# Method 3: marginalization of a Bayesian penalized Cox model

# Marginalization of Bayesian penalized Cox model

Same as Method 2, but fit global model using Bayes.

**Horseshoe prior**, Carvalho et al. (2009):

- Outcome $(\mathbf{y}|\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ **sparse** parameter vector.
- Each regression coefficient $\beta_i$ (conditionally) independent with prior a scale-mixture of Normals:

$$\begin{aligned} (\beta_i|\lambda_i) &\sim N(0, \lambda_i^2), \\ (\lambda_i|\tau) &\sim C^+(0, \tau), \\ \tau &\sim C^+(0, 1). \end{aligned}$$

- Each $\beta_i$ mixed over own **local shrinkage** parameter $\lambda_i$, and each $\lambda_i$ has independent half-Cauchy prior with common, **global shrinkage** parameter $\tau$.
- Best predictive performance among various priors, shrinks "strong" less than "weak" effects
- Why "horseshoe"? $\mathbb{E}(\beta_i|\mathbf{y}) \overset{\tau^2=1}{=} (1 - \mathbb{E}(\kappa_i|\mathbf{y}))y_i$, and shrinkage coefficient $\kappa_i$ has prior $\text{Be}(1/2, 1/2)$.
- Implementation: via STAN and R package **brms** Bürkner (2018). Poisson regression for proportional hazards model with piecewise constant baseline hazard, Laird and Olivier (1981).

# Simulation

# Simulation setup

Simulation assumptions:

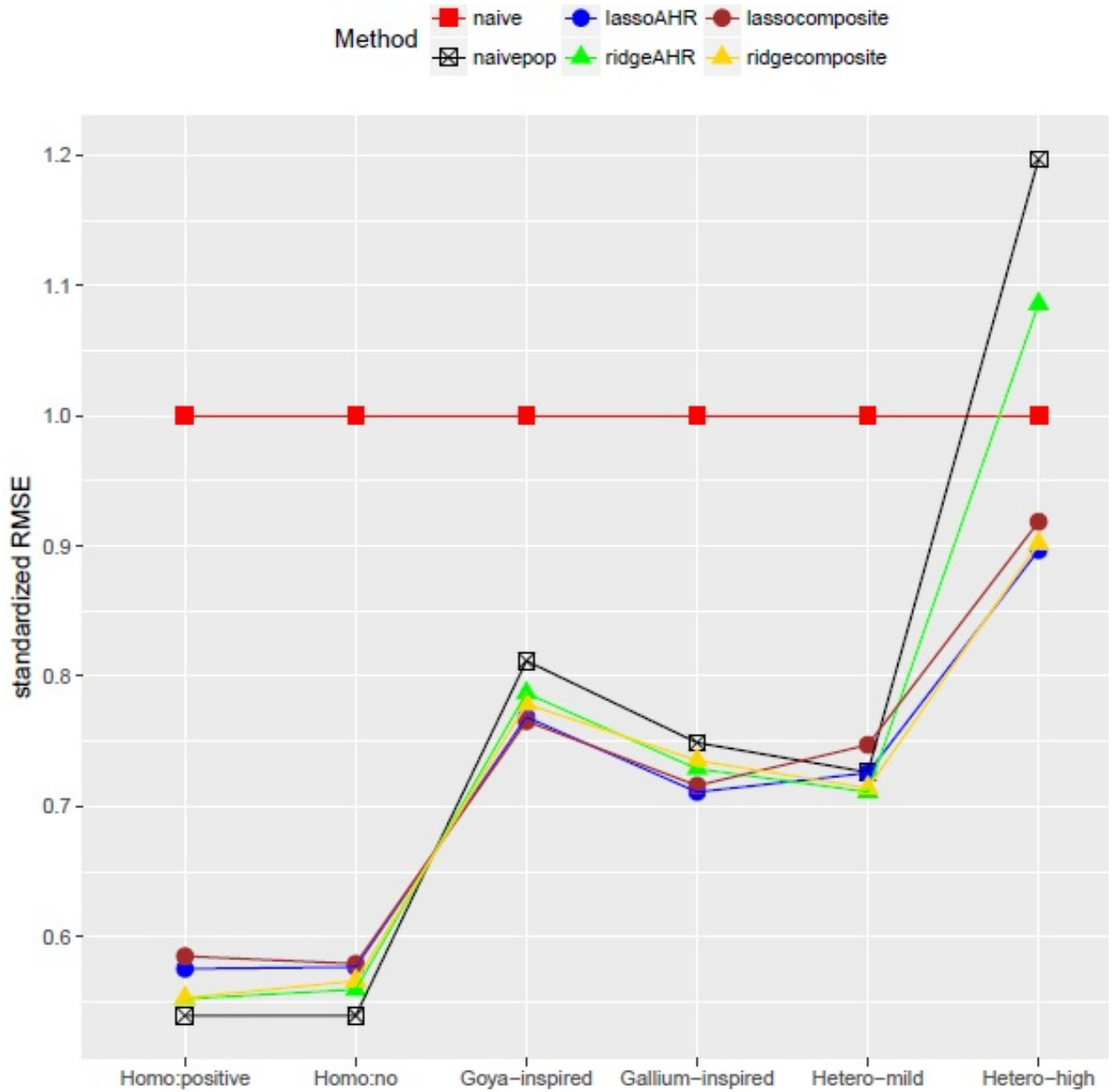
- Inspired by GALLIUM data: $n$=1202, 245 events.
- 10 subgrouping variables, defining 25 subgroups.
- 1000 simulation runs for each scenario.

# Simulation scenarios

| Scenario | Overall treatment effect (HR-scale) | No.subgroup with predictive effect | Differential subgroup treatment effect |
|---|---|---|---|
| Homo:positive | 0.67 | 0 | 0 |
| Homo:no | 1 | 0 | 0 |
| GOYA-inspired | 1 | 3 | 0.5 |
| GALLIUM-inspired | 0.67 | 3 | 1.2 |
| Hetero-mild | 1 | 15 | $\mathcal{N}(0, 0.2)$ |
| Hetero-high | 1 | 15 | $\mathcal{N}(0, 0.5)$ |

# Results

# Simulation results



# Simulation results

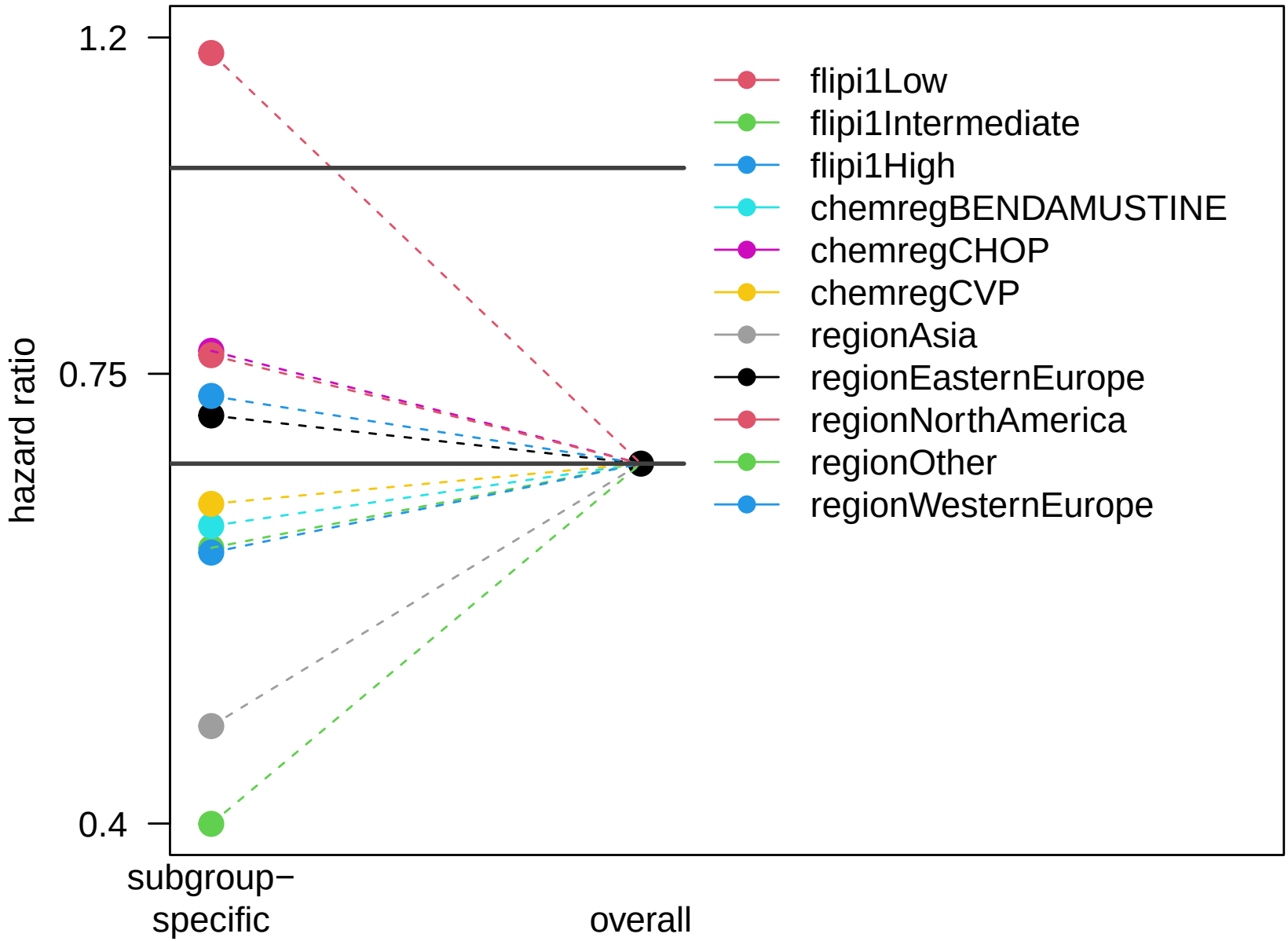
Main findings:

- Naive overall estimate: MSE as shrinkage estimators and uniformly better than naive subgroup-specific, **except** when highly heterogeneous effects.
- Shrinkage estimates: uniformly better MSE than naive subgroup-specific.
- No relevant difference in MSE for composite/marginalization.
- Lasso slightly better than ridge.
- Lasso (= Laplace prior with Bayes) may overshrink in sparse situations. "Fixed" with horseshoe prior (next slide).

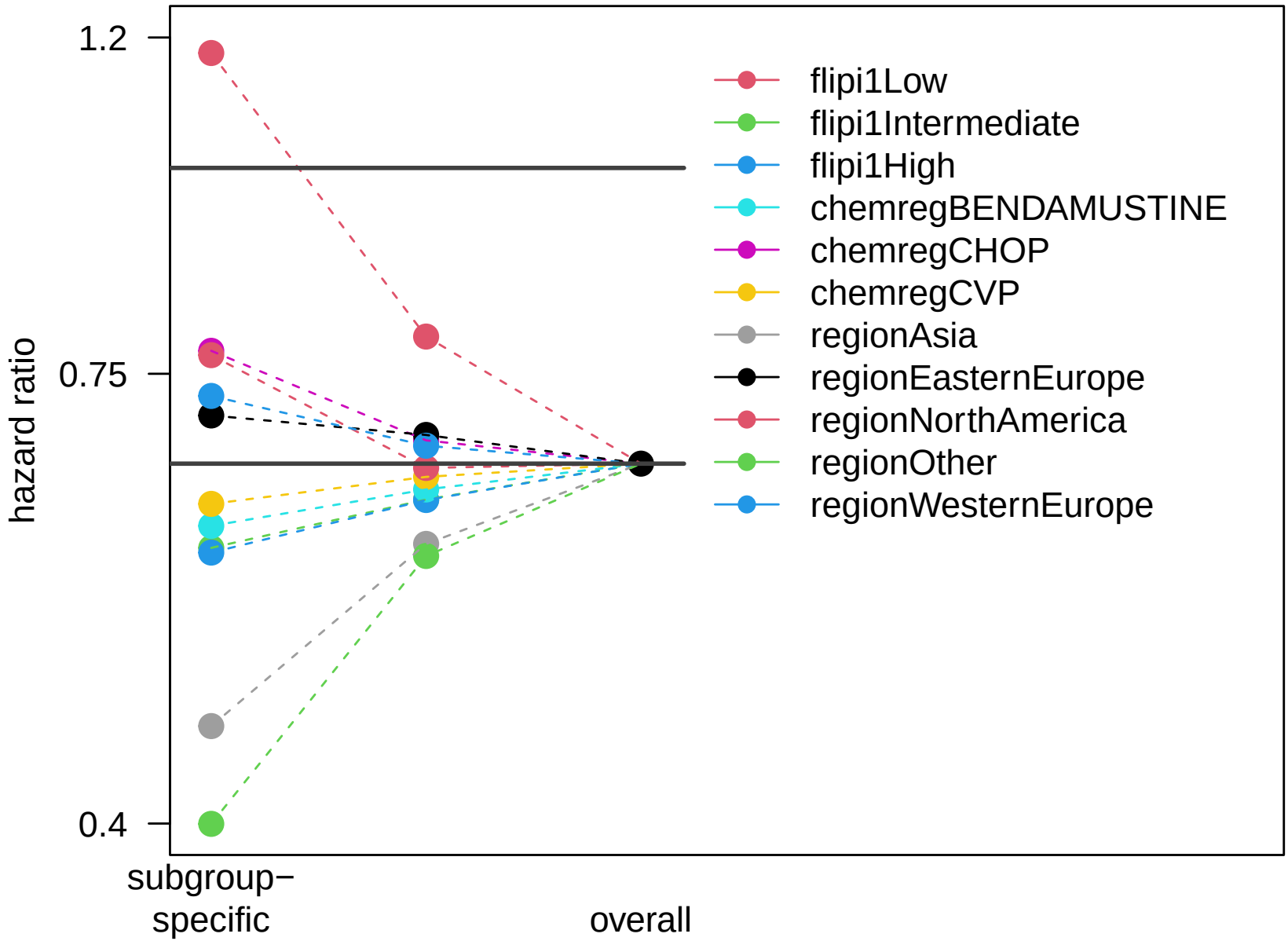
Gallium: **not** highly heterogeneous treatment effect $\Rightarrow$ naive overall good estimate $\Rightarrow$ use in FLIPI low subgroup?

# Gallium data

# Naive estimates



# Naive estimates – Bayesian interpolation



# Conclusions

# Conclusions

- New approaches for shrinkage estimation for pre-specified subgroups.
- Interpolate between naive subgroup-specific and naive overall. Degree of interpolation governed through **penalty parameter** or **prior choice**.
- Approaches can be made **fully automatic** $\Rightarrow$ relevant in regulatory environment.
- Fit using standard penalized regression models.
- Naive overall estimate has good MSE!

**Thank you for your attention.**

# References I


- Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* **10** 395–411. https://doi.org/10.32614/RJ-2018-017
- Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics* (D. van Dyk and M. Welling, eds.), vol. 5 of *Proceedings of Machine Learning Research*. PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. http://proceedings.mlr.press/v5/carvalho09a.html
- Ford, I., Norrie, J. and Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Stat Med* **14** 735–746.
- Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* **76** 231–240. https://www.tandfonline.com/doi/abs/10.1080/01621459.1981.10477634
- Marcus, R., Davies, A., Ando, K., Klapper, W., Opat, S., Owen, C., Phillips, E., Sangha, R., Schlag, R., Seymour, J. F., Townsend, W., Trneny, M., Wenger, M., Fingerle-Rowson, G., Rufibach, K., Moore, T., Herold, M. and Hiddemann, W. (2017). Obinutuzumab for the First-Line Treatment of Follicular Lymphoma. *N. Engl. J. Med.* **377** 1331–1344.
- Oxman, A. D. (2012). Subgroup analyses. *BMJ* **344** e2022.

# References II


- Schemper, M., Wakounig, S. and Heinze, G. (2009). The estimation of average hazard ratios by weighted cox regression. *Stat. Med.* **28** 2473–2489.
- Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21** 5–42.
  `http://www.jstor.org/stable/24309261`

# Backup slides.

# Company feedback to EMA

For major objection: 22p document, just short of a scientific paper.

Key arguments that FLIPI low patients should **still be treated**:

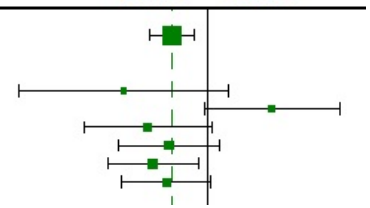
- **Interaction test** $p = 0.14$.
- 95% confidence interval from **0.63 to 2.19** $\Rightarrow$ covers overall effect of 0.66.
- No difference in patient characteristics FLIPI low vs. FLIPI non-low.
- **No biological rationale** why FLIPI low patients should profit less.
- Simulation study: given #subgroups analyzed probability to have observed #HRs $> 1$ is $\approx$ **70%**.
- Updated snapshot: HR decreased to **1.11**.

Can we get a pre-specified shrunken estimate?

# Statistician dig further

Subgroups by gender and FLIPI:

| Baseline Factors | Total n | R-chemo (N=601) n | R-chemo (N=601) Events | R-chemo (N=601) 1 Year KM rate | G-chemo (N=601) n | G-chemo (N=601) Events | G-chemo (N=601) 1 Year KM rate | Hazard Ratio | 95% Wald CI |
|---|---|---|---|---|---|---|---|---|---|
| All Patients | 1202 | 601 | 144 | 89.736 | 601 | 101 | 93.939 | 0.66 | (0.51, 0.85) |
| Flipi and Gender Interactions | | | | | | | | | |
| FL Low Women | 116 | 52 | 8 | 91.697 | 64 | 4 | 100.000 | 0.38 | (0.11, 1.27) |
| FL Low Men | 137 | 73 | 10 | 94.030 | 64 | 18 | 88.807 | 2.08 | (0.96, 4.52) |
| FL Intermediate Women | 222 | 115 | 22 | 93.588 | 107 | 11 | 96.970 | 0.50 | (0.24, 1.04) |
| FL Intermediate Men | 225 | 108 | 27 | 86.600 | 117 | 20 | 95.606 | 0.64 | (0.36, 1.14) |
| FL High Women | 301 | 154 | 41 | 92.596 | 147 | 22 | 94.239 | 0.53 | (0.32, 0.90) |
| FL High Men | 201 | 99 | 36 | 80.420 | 102 | 26 | 88.104 | 0.62 | (0.37, 1.03) |

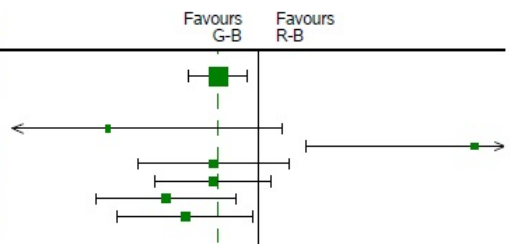



Even **more improbable** that **men with FLIPI low** profit differently from Gazyva!

# Statistician keep digging

Subgroups by gender and FLIPI within Bendamustine patients only:

| Baseline Factors | Total n | R-B (N=341) n | R-B Events | R-B 1 Year KM rate | G-B (N=345) n | G-B Events | G-B 1 Year KM rate | Hazard Ratio | 95% Wald CI |
|---|---|---|---|---|---|---|---|---|---|
| All Patients | 686 | 341 | 78 | 89.021 | 345 | 50 | 93.928 | 0.61 | (0.43, 0.86) |
| Flipi and Gender Interactions | | | | | | | | | |
| FL Low Women | 65 | 33 | 6 | 89.729 | 32 | 1 | 100.000 | 0.16 | (0.02, 1.32) |
| FL Low Men | 84 | 43 | 1 | 97.436 | 41 | 12 | 87.361 | 13.74 | (1.78, 105.81) |
| FL Intermediate Women | 127 | 66 | 13 | 89.244 | 61 | 7 | 96.429 | 0.58 | (0.23, 1.45) |
| FL Intermediate Men | 136 | 59 | 17 | 80.960 | 77 | 14 | 94.667 | 0.57 | (0.28, 1.17) |
| FL High Women | 162 | 82 | 21 | 94.895 | 80 | 7 | 94.595 | 0.32 | (0.14, 0.76) |
| FL High Men | 112 | 58 | 20 | 82.655 | 54 | 9 | 90.598 | 0.41 | (0.18, 0.93) |




So it is only **men with FLIPI low treated with Bendamustine**???

# Penalized composite likelihood – implementation

Implementation:

1. Transform data to long.
2. Fit penalized model to stacked data.
3. Optimize tuning parameter $\lambda$ using cross-validation (of patients!).

| ID | sex | flipi | time | status |
|---|---|---|---|---|
| 1 | f | low | 13 | 1 |
| 2 | f | high | 7 | 0 |
| 3 | m | NA | 23 | 0 |
| ... | ... | ... | ... | ... |

| ID | subgroup | time | status |
|---|---|---|---|
| 1 | sex.f | 13 | 1 |
| 2 | sex.f | 7 | 0 |
| 3 | sex.m | 23 | 0 |
| ... | ... | ... | ... |
| 1 | flipi.low | 13 | 1 |
| 2 | flipi.high | 7 | 0 |
| 3 | NA | 23 | 0 |
| ... | ... | ... | ... |

# SmPC

Summary of Product characteristics:

**4.4 Special warnings and precautions for use**

In order to improve the traceability of biological medicinal products, the trade name and batch number of the administered product should be clearly recorded (or stated) in the patient file.

Based on a subgroup analysis in previously untreated follicular lymphoma, the efficacy in FLIPI low risk (0-1) patients is currently inconclusive (see section 5.1). A therapy choice for these patients should carefully consider the overall safety profile of Gazyvaro plus chemotherapy and the patient-specific situation.

`http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_Information/human/002799/WC500171594.pdf`

*Doing now what patients need next*

**R version and packages used to generate these slides:**

R version: R version 4.1.1 (2021-08-10)

Base packages: grid / stats / graphics / grDevices / utils / datasets / methods / base

Other packages: brms / Rcpp / gbm / broom / glmnet / Matrix / filesstrings / googlesheets / readxl / forestplot / checkmate / magrittr / forcats / stringr / purrr / readr / tidyr / tibble / ggplot2 / tidyverse / reporttools / xtable / dplyr / biostatKR / survival

This document was generated on 2022-02-10 at 07:21:45.