Innovative Statistical Methods for Meta-Analyses With Between-Study Heterogeneity

Mengli Xiao, PhD
Department of Biostatistics and Informatics, Colorado School of Public Health
University of Colorado

May 24, 2023

No relevant disclosures.

Systematic reviews and meta-analysis: a lens through which evidence is viewed

Image from Murad et al. (2016).

Meta-analysis

Systematic reviews and meta-analysis

- Totality of evidence in medical and clinical research.
  One summary effect ← intervention effects from multiple *related* but *independent* studies.

Why meta-analysis?

- Improved statistical power & precision.
- Investigate reasons for heterogeneous intervention effects.

Sources of heterogeneity

Studies may be:

1. different but *related*.

Solution:

1. random-effects meta-analysis.

Sources of heterogeneity

Studies may be:

1. different but *related*.
2. different and *unrelated*, and should not be included in the same meta-analysis.

Solution:

1. random-effects meta-analysis.
2. **determining unrelated studies.**

Sources of heterogeneity

Studies may be:

1. different but *related*.
2. different and *unrelated*, and should not be included in the same meta-analysis.
3. modeled using covariates.

Solution:

1. random-effects meta-analysis.
2. **determining unrelated studies.**
3. known covariates $\rightarrow$ meta-regression; unknown covariates $\rightarrow$ **modeling the dependence of the effect on unknown covariates.**

Sources of heterogeneity

Studies may be:

1. different but *related*.
2. different and *unrelated*, and should not be included in the same meta-analysis.
3. modeled using covariates.

Solution:

1. random-effects meta-analysis.

3. known covariates $\rightarrow$ meta-regression;

Sources of heterogeneity

Studies may be:

1. different but *related*.
2. different and *unrelated*, and should not be included in the same meta-analysis.
3. modeled using covariates.

Solution:

1. random-effects meta-analysis.
2. **determining unrelated studies.**
3. known covariates $\rightarrow$ meta-regression; unknown covariates $\rightarrow$ **modeling the dependence of the effect on unknown covariates.**

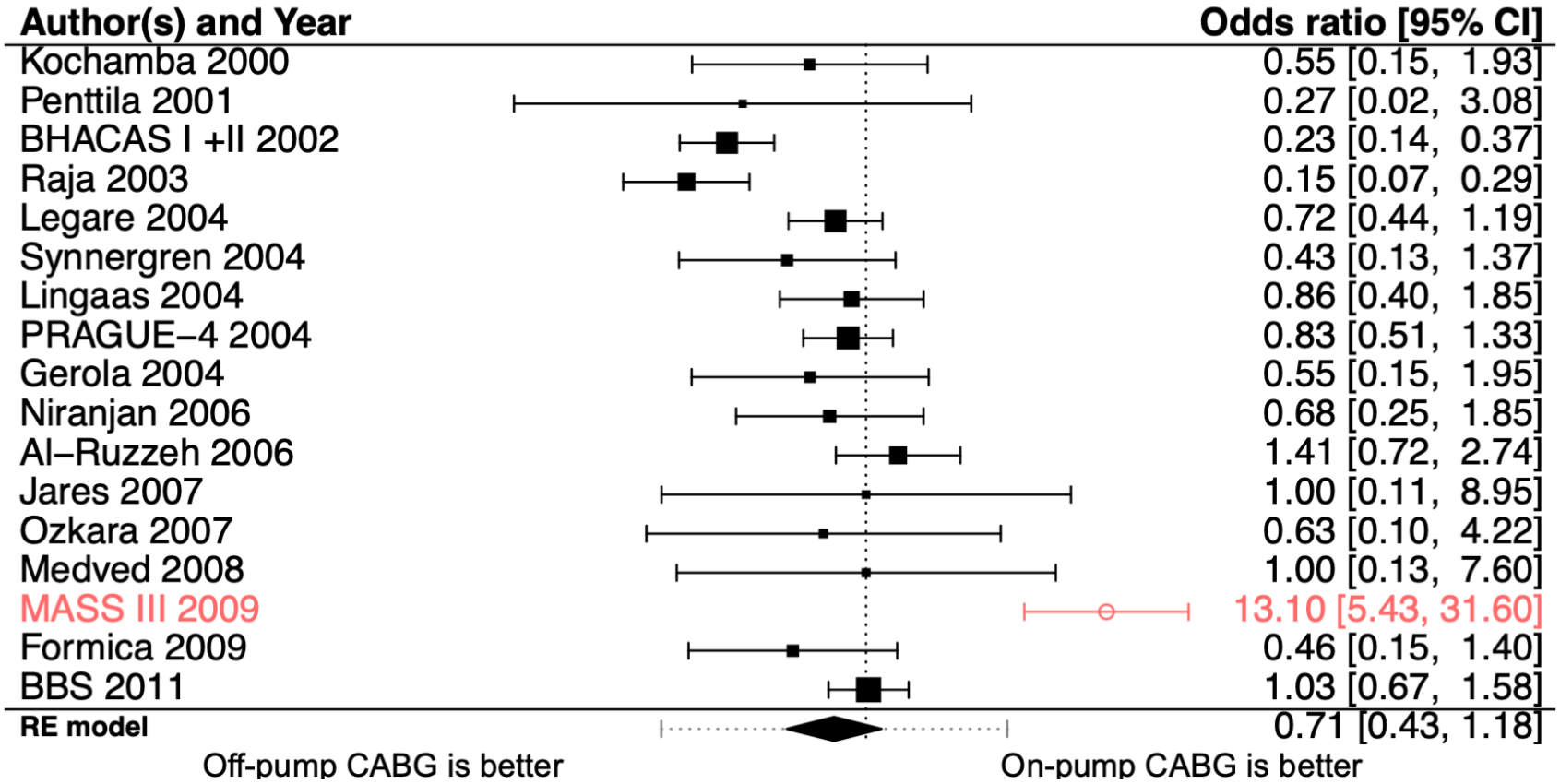
1. Determining unrelated studies

Quantifying replicability of multiple studies in a meta-analysis

How are unrelated studies defined and measured in meta-analyses with aggregated data?

Motivating example

- Off-pump (left) vs. on-pump (right) coronary artery bypass grafting (CABG) for reducing postoperative atrial fibrillation (POAF) occurrence, where CABG is a common treatment for ischemic heart disease.



Are studies replicable?

Definition of replicability

"obtaining ***consistent*** results across studies aimed at answering the same scientific question, each of which has obtained its own data." (National Academy of Sciences, 2019)

Why should we assess replicability in meta-analyses?

- Replication crisis of various scientific studies.
- The replication issues are rooted in systematic reviews and meta-analysis, e.g., subjective inclusion and exclusion of studies with different designs.
- No solution to a persistent dilemma: whether a meta-analysis should exclude studies inflating/deflating the meta-analysis summary.
- Practical guidelines exist, but no consensus on statistical tools.

Current methods for assessing replicability

- Few metrics for replicability in meta-analysis.
- No widely-used definition of replicability.
  - The p-value driven definition is problematic (Jaljuli et al., 2019).

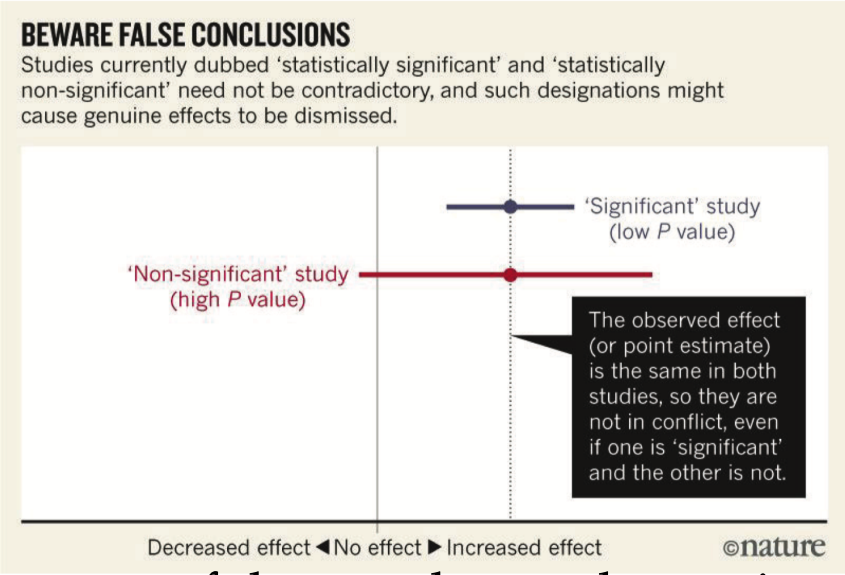


Figure: Using agreement of the p-value to determine replicability in National Academy of Sciences (2019)

Current methods for assessing replicability

- No established definition of replicability.
  - Concept of non-replicability is entangled with heterogeneity (Schauer and Hedges, 2020).
    - Common heterogeneity statistic: τ^2, Q and I^2.
    - Question: are τ^2, Q or I^2 sufficient to reflect non-replicability?

Why are heterogeneity statistics insufficient?

- **High between-study heterogeneity** still assume studies come from the same distribution, i.e., studies are ***replicable***.

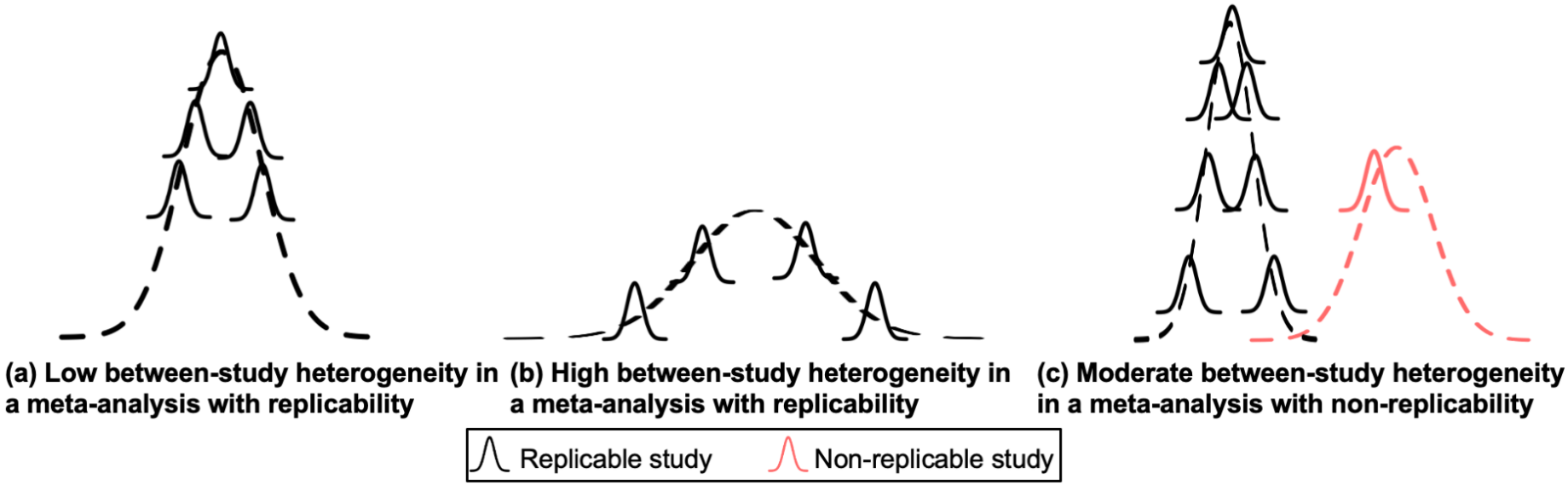


Figure: Random-effects meta-analyses with different levels of heterogeneity and replicability.

Statistics behind meta-analysis

In a meta-analysis with n studies, if μ is the overall/summary mean effect size, then for each study i (where $i = 1, \ldots, n$):

- y_i: observed effect size; s_i^2: within-study sample variance.
- Goal: obtain an estimate of μ from observed study effects, i.e., $\hat{\mu} = T(y_1, \ldots, y_n, s_1^2, \ldots, s_n^2)$.
- A random-effects model for the observed effect size in each study i:

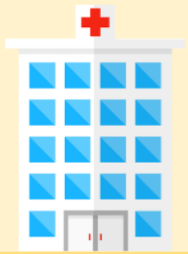
$$y_i = \mu + \epsilon_i, \qquad \epsilon_i \sim N(0, s_i^2 + \tau^2),$$

where τ^2 is the between-study variance and is estimated from the data.

Intuition

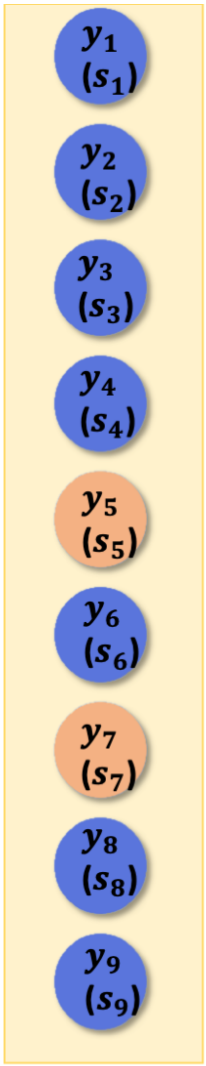
Study 1

Study 2



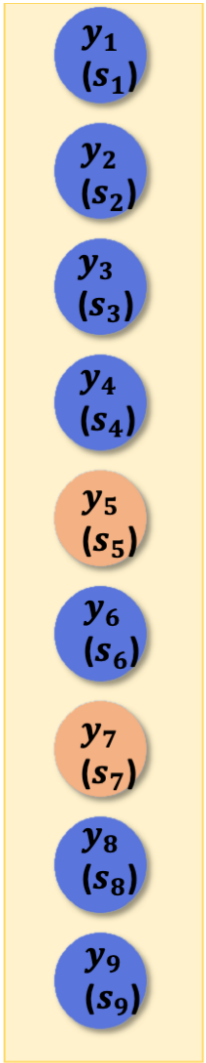
Study 9

Intuition

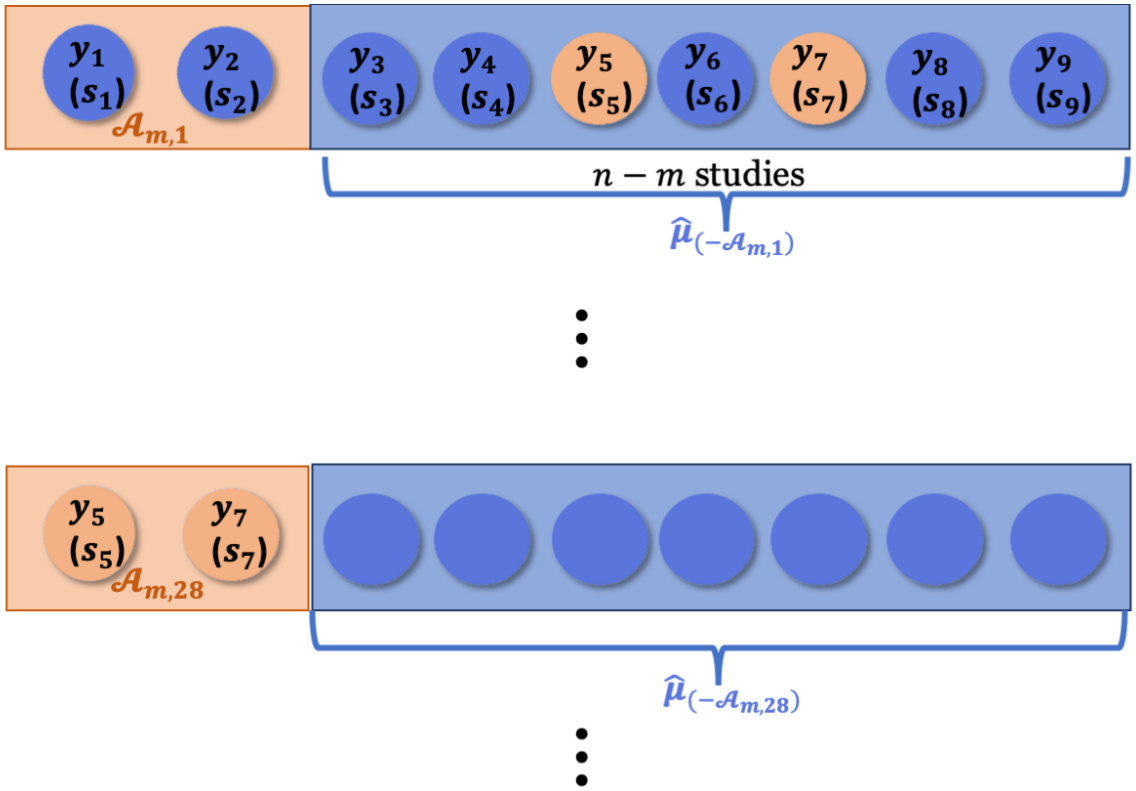


n studies,
m being non-replicable.

Intuition



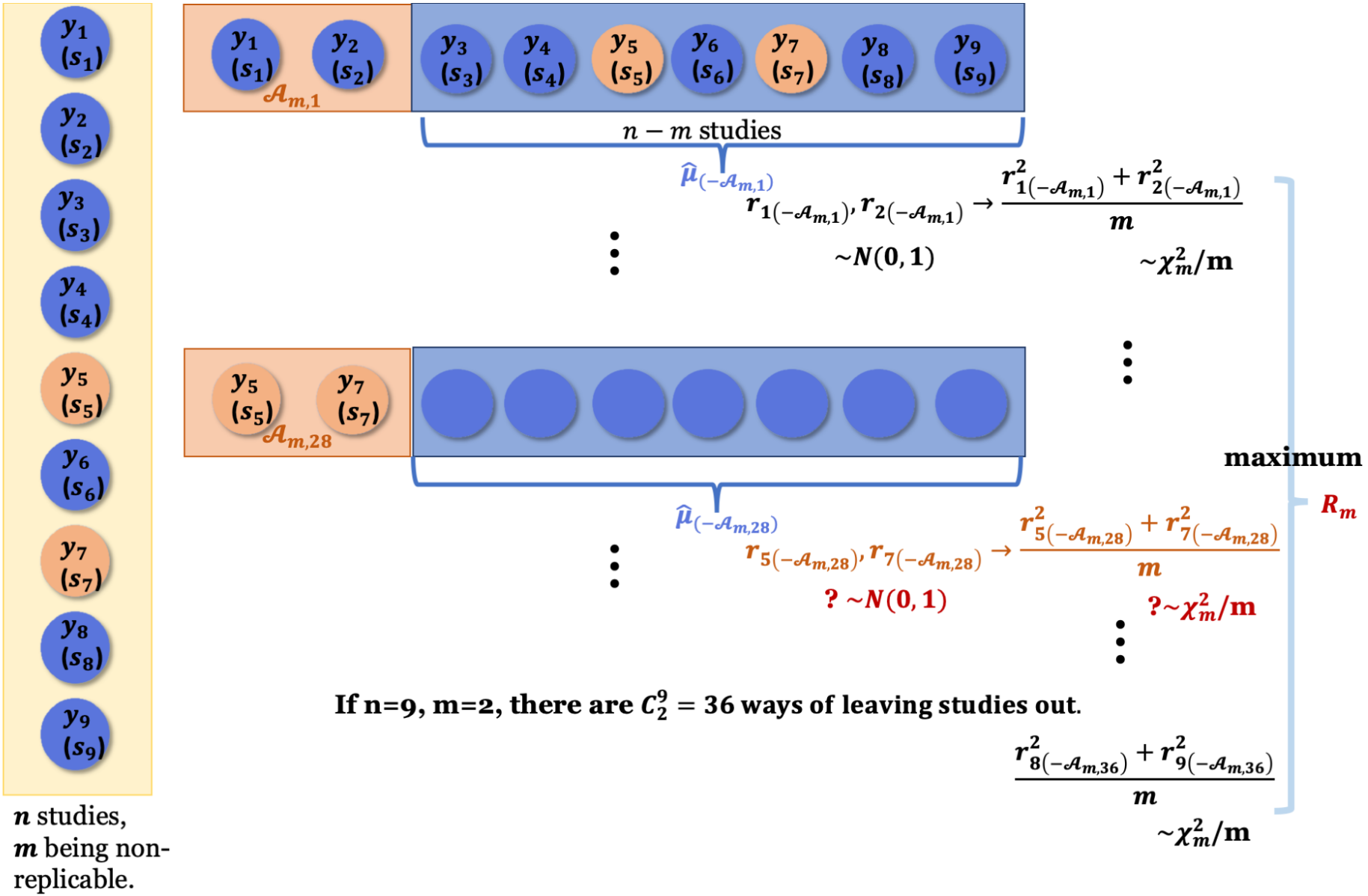
n studies, ***m*** being non-replicable.



If n=9, m=2, there are $C_2^9 = 36$ ways of leaving studies out.

Denote $\mathcal{A}_{m,k}$ as the set of m studies to be removed from the meta-analysis ($k = 1, \ldots, M_m$). Let $\hat{\mu}_{(-\mathcal{A}_{m,k})}$ and $\hat{\tau}^2_{(-\mathcal{A}_{m,k})}$ be the estimates of μ and τ^2 using the **remaining $n - m$ studies**, after omitting studies in $\mathcal{A}_{m,k}$.

Intuition



n studies, m being non-replicable.

Proposed measure: R_m

- How large is R_m?
  - We used the bootstrap method and an asymptotic approximation to describe the distribution of R_m $- >$ Hypothesis testing.

Simulation summary

- Simulation design

Parameter	Design
Number of studies (n)	$\{5,\ 10,\ 20,\ 50\}$
Between-study standard deviation (τ)	$\{0,\ 0.32,\ 0.55\}$
Within-study standard deviation (s)	$s \sim U\,(0.10,\ 0.22)$
	$s \sim U\,(0.22,\ 0.54)$
	$s \sim U\,(0.54,\ 1.41)$
Effect size difference ($\lvert\Delta\rvert$) or effect sizes (μ and μ')	$\lvert\Delta\rvert = 3$, μ=0, $\mu' = 3$

Note: Without loss of generality, $\mu = 0$ and $\mu' = 3$ but they can be any other values as long as $|\Delta| = |\mu' - \mu| = 3$.

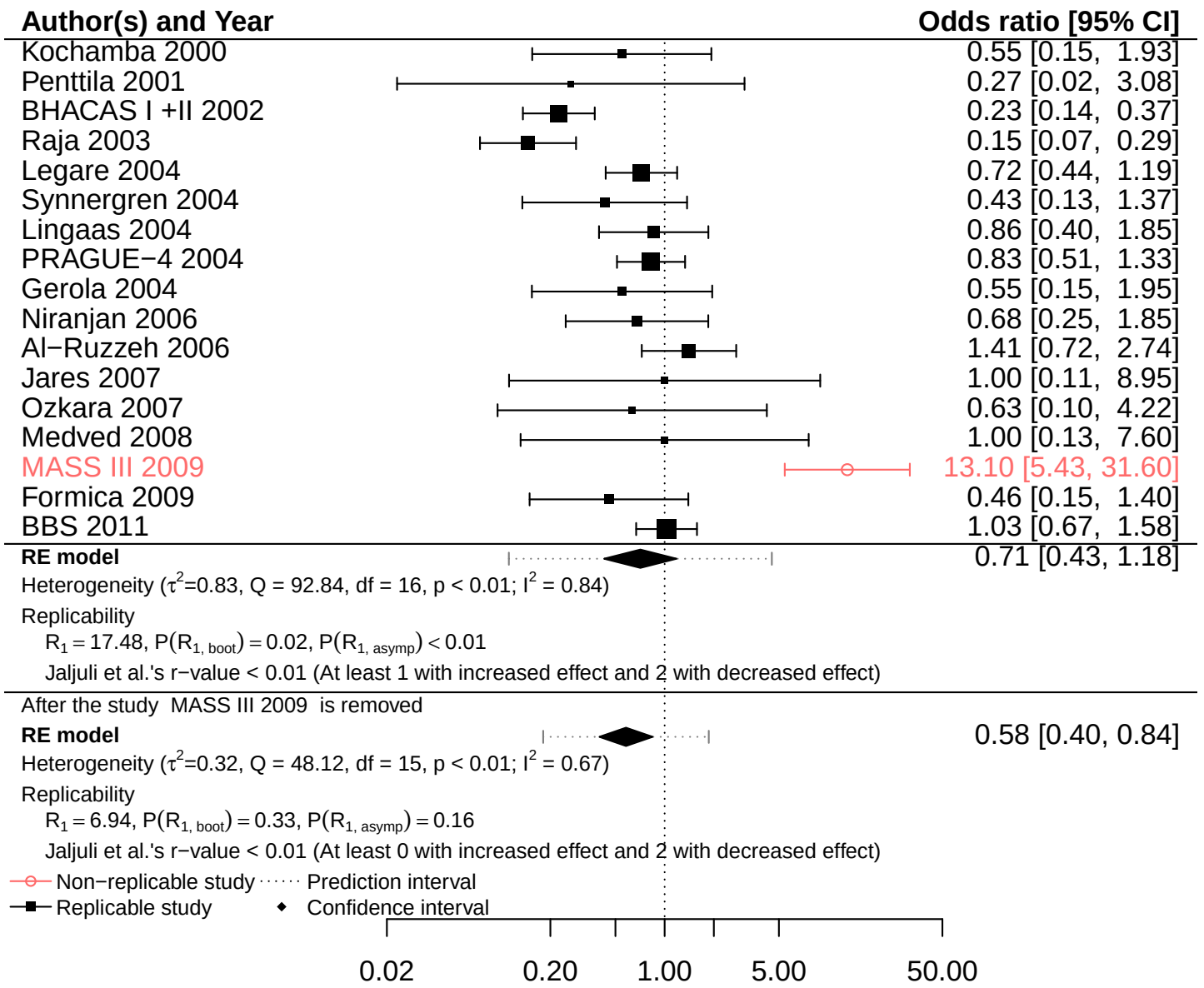
Simulation summary

- Type I error rate
  - Generally controlled (< 0.07) under the bootstrap method but inflated under the asymptotic approximation.
  - Controlled type I error rate in the setting close to our motivating study.
- Power
  - $>$ 80% power in most settings (including the case study).

Is off-pump CABG better than on-pump CABG in reducing postoperative complications?

- Meta-analysis in Møller et al. (2012) with 17 trials: heparin dose less than 300 mg/k vs. on-pump coronary artery bypass grafting (CABG) for reducing postoperative atrial fibrillation (POAF).
- Heparin prevents the blood from clotting during CABG.
- Odds of having POAF is the outcome measure.

Is off-pump CABG better?



Is off-pump CABG better?

- The population from study MASS III 2009 had different ethnicity and severe forms of atherosclerosis, comparing to other studies (Møller et al., 2012).
- Researchers should cautiously interpret the results of an MA that mixes studies with different population baseline characteristics.

R package

- To facilitate assessing replicability before any meta-analysis is undertaken, we developed an R package "repMeta".
- Replicability test and identifying non-replicable studies.

2. Modeling the dependence of the effect on unknown covariates

If studies are replicable (from the same distribution), is the meta-analysis summary measure sufficient to represent heterogeneous effects?

3 A bivariate generalized mixed-effects model (BGLMM)

- Study-level treatment effect heterogeneity
- Unknown study characteristics & baseline risks
- Methods
  - Exisiting methods
  - BGLMM for conditional treatment effects
- Results
  - A case study

Study-level treatment effect heterogeneity

Among studies from the same distribution, i.e., replicable studies, substantial treatment effects heterogeneity still exists. How?

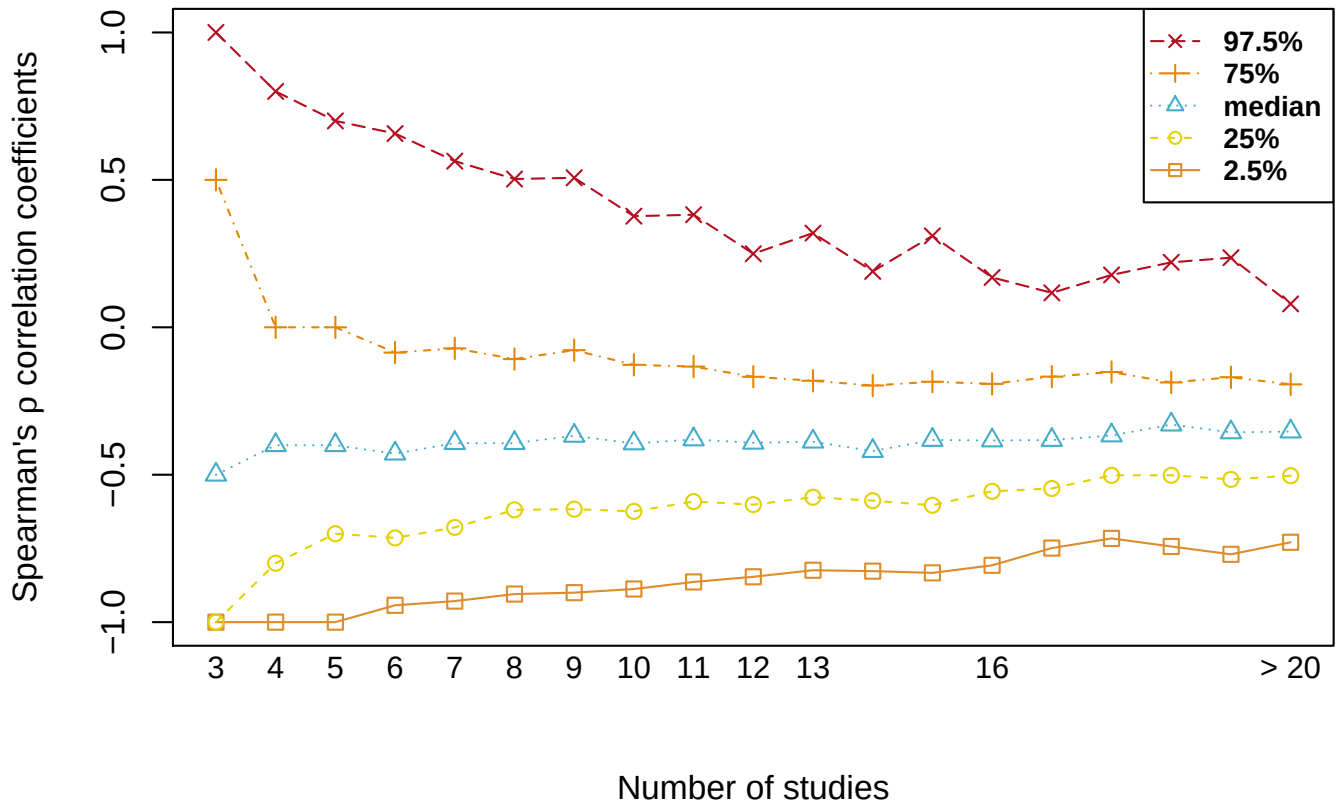
- The disease severity varies across study populations $\rightarrow$ the heterogeneity of 1) risk without treatment, and 2) responsiveness to treatment. (Brand and Kragt, 1992; Kravitz et al., 2004).
- Methodological diversity in outcome assessment (Higgins et al., 2019).

Baseline risk

- Control group risk or baseline risk is a proxy for study conditions.
- Readily available for binary outcomes.

OR is correlated with the baseline risk

Results from over 40,000 meta-analyses



- Both ORs and RRs are negatively correlated with baseline risks in most cases.
- The proportion of meta-analyses with negligible/minor correlations (|Spearman's ρ | <0.3 is the cut-off for negligible correlation (Mukaka, 2012))) is slightly higher for RR and baseline risks that for OR and baseline risks, and the difference is notable when the number of studies > 20.
  - The proportions of meta-analyses with negligible ρ_{OR} but non-negligible ρ_{RR} are 6.1% ($5 \leq n < 20$) and 5.4% ($n \geq 20$), and the proportions of meta-analyses with negligible ρ_{RR} but non-negligible ρ_{OR} are 6.3% ($5 \leq n < 20$) and 11.7% ($n \geq 20$).

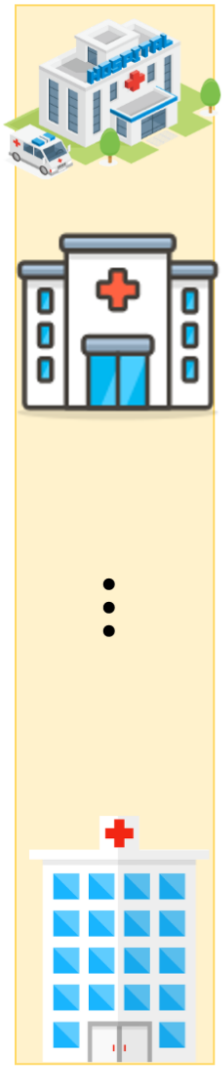
Existing approaches from the control risk regression

Several methods can address the dependence of treatment effect on baseline risks.

- Naïve weighted least square approach: regress treatment effect on the baseline risk (Brand and Kragt, 1992).
  - A negative correlation between estimation errors of treatment effect and baseline risk (Senn and Brand, 1994).
  - More importantly, the measurement error of the baseline risk needs be modeled (Sharp et al., 1996; Schmid et al., 2020).
- **Exact measurement error model**, or one-stage model (Schmid et al., 2004; Chu et al., 2012).
- Approximate measurement error model (Schmid et al., 2020).

⋮

Conditional effect estimates



Study 1

Study 2

Study 9

Treatment group

$$g(p_{i1}) = \mu_1 + \nu_{i1}$$

Control group

$$g(p_{i0}) = \mu_0 + \nu_{i0}$$

$$(\nu_{i0}, \nu_{i1})^T \sim N\left((0,0)^T, \Sigma\right) \qquad \Sigma = \begin{pmatrix} \sigma_0^2 & r\sigma_0\sigma_1 \\ r\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}$$

Conditional effect estimates

Study i, with any baseline risk p_0

Treatment effect? → $p_{i1}|p_0$?

Treatment group: $g(p_{i1}) = \mu_1 + \nu_{i1}$

Control group: $g(p_{i0}) = \mu_0 + \nu_{i0}$

$$(\nu_{i0}, \nu_{i1})^T \sim N\left((0,0)^T, \Sigma\right) \qquad \Sigma = \begin{pmatrix} \sigma_0^2 & r\sigma_0\sigma_1 \\ r\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}$$

$$E(g(p_{i1}) \mid p_{i0} = p_0) = \mu_1 - \frac{r\mu_0\sigma_1}{\sigma_0} + \frac{r\sigma_1}{\sigma_0} g(p_0)$$

$M(\cdot)$ denotes the median

$$\hat{\boldsymbol{p}}_{\boldsymbol{i1}}|\boldsymbol{p_0} = M(p_{i1} \mid p_{i0} = p_0) = g^{-1}\left(M(g(p_{i1}) \mid p_{i0} = p_0)\right) = \boldsymbol{g^{-1}}\left(\boldsymbol{E}(\boldsymbol{g}(\boldsymbol{p_{i1}}) \mid \boldsymbol{p_{i0}} = \boldsymbol{p_0})\right)$$

Treatment effect

$$\widehat{OR}_i|p_0 = \frac{\hat{p}_{i1}|p_0}{1 - \hat{p}_{i1}|p_0} \Big/ \frac{p_0}{1 - p_0}$$

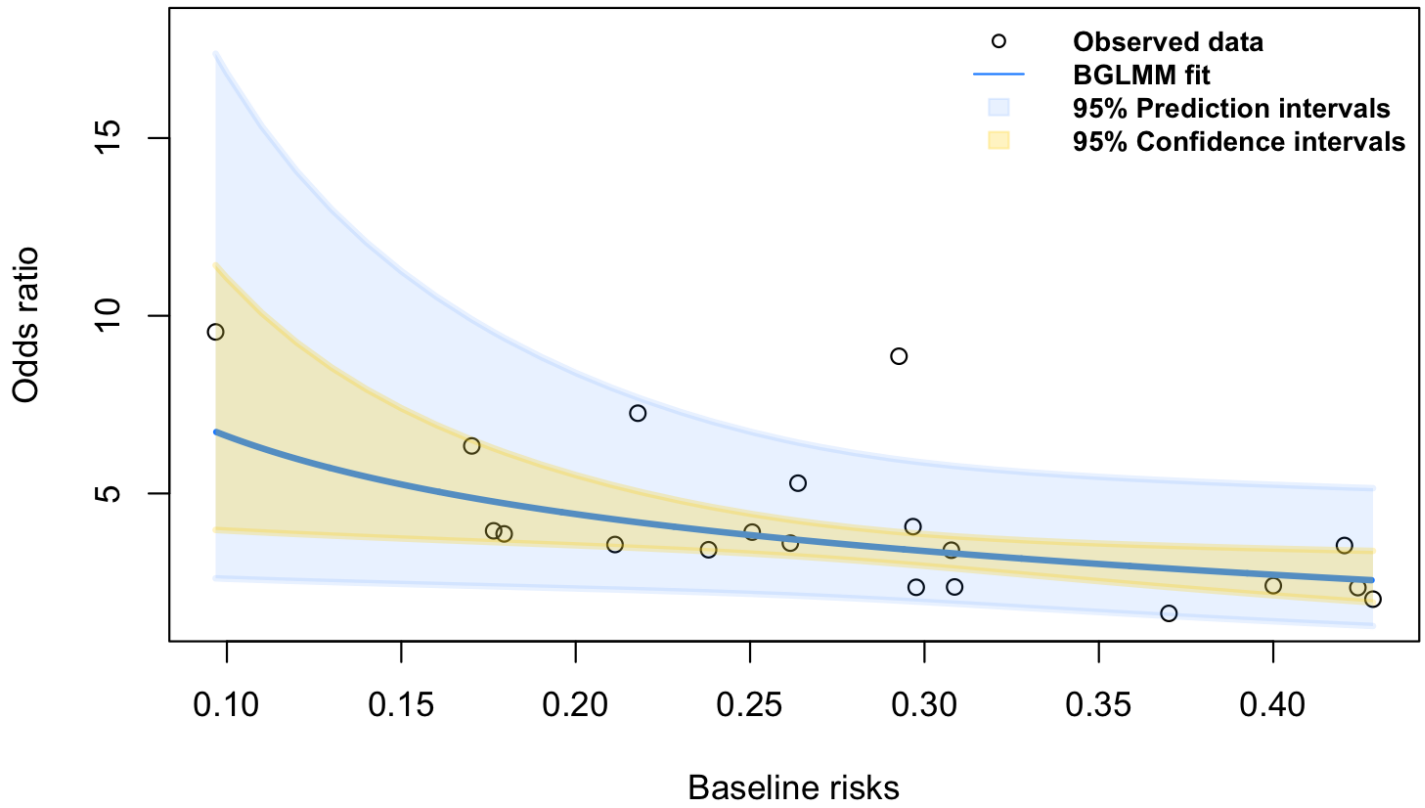
$$\widehat{RR}_i|p_0 = \frac{\hat{p}_{i1}|p_0}{p_0}$$

$$\widehat{RD}_i|p_0 = \hat{p}_{i1}|p_0 - p_0$$

Bivariate generalized linear mixed effects model (BGLMM)

- Baseline risk: p_{i0}; treatment group risk: p_{i1}.
- $\Sigma = \begin{pmatrix} \sigma_0^2 & r\sigma_0\sigma_1 \\ r\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}$ is the covariance matrix for the baseline and treatment risk in the logit scale.
- We fit the model in SAS using PROC NLMIXED.

A meta-analysis with 20 studies comparing the effect of oral sumatriptan (100 mg) versus placebo on mitigating the acute headache at 2 hours after drug administration (Derry et al., 2012).



- BGLMM 1) describes the non-linear dependence of ORs on baseline risks, 2) estimates the correlation r_0 as -0.68 (95% CI, -0.91 to -0.14), and 3) derives conditional ORs for any given baseline risk between 0 and 1.
- A single summary OR of 3.48 with 95% CI (2.93 to 4.14) versus conditioning OR ranging from 6.73 to 2.56 when the baseline risk increases from 0.10 to 0.43.

Baseline risks	A single summary OR	BGLMM-derived conditional OR
0.10		6.73
⋮	3.48	⋮
0.43		2.56

Summary

A meta-analysis with between-study heterogeneity may suffer from:

1. non-replicable study findings.

Solutions:

1. a replicability test and an R package "repMeta".

Contributions:

1. a new metric to suggest sources of effect discrepancy in a meta-analysis: confounding, selection bias, measurement error, etc. $\rightarrow$ improves future systematic reviews.

Summary

A meta-analysis with between-study heterogeneity may suffer from:

1. non-replicable study findings.
2. high heterogeneity with unknown covariates.

Solutions:

1. a replicability test and an R package "repMeta".
2. a bivariate model that accounts for baseline risks (Xiao et al., 2022a,b).

Contributions:

1. a new metric to suggest sources of effect discrepancy in a meta-analysis: confounding, selection bias, measurement error, etc. $\rightarrow$ improves future systematic reviews.
2. reporting effects conditioning on the baseline risk, such as using BGLMM, rather than relying on the overall summary alone.

Future work

- Improving replicability in meta-analysis for clinical decision making:
  - replicability analysis in a meta-analysis with individual participant data.
  - clinical guidance about replicability measures.
- Conditioning effect for the continuous outcomes in a meta-analysis.

Thank you!

Questions? (mengli.xiao@cuanschutz.edu)

References: I

Brand, R. and Kragt, H. (1992). Importance of trends in the interpretation of an overall odds ratio in the meta–analysis of clinical trials. *Statistics in Medicine*, 11(16):2077–2082.

Chu, H., Nie, L., Chen, Y., Huang, Y., and Sun, W. (2012). Bivariate random effects models for meta-analysis of comparative studies with binary outcomes: methods for the absolute risk difference and relative risk. *Statistical Methods in Medical Research*, 21(6):621–633.

Derry, C. J., Derry, S., and Moore, R. A. (2012). Sumatriptan (oral route of administration) for acute migraine attacks in adults. *Cochrane Database of Systematic Reviews*, (2).

Embrechts, P., Mikosch, T., and Klüppelberg, C. (1997). *Modelling Extremal Events: For Insurance and Finance*. Springer-Verlag, Berlin, Germany.

References: II

Gumbel, E. J. (1958). *Statistics of Extremes.* Columbia University Press, New York, NY.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A. (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.

Jaljuli, I., Benjamini, Y., Shenhav, L., Panagiotou, O., and Heller, R. (2019). Quantifying replicability and consistency in systematic reviews. *ArXiv.* Available at `https://arxiv.org/abs/1907.06856`.

Kravitz, R. L., Duan, N., and Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82(4):661–687.

References: III

Møller, C. H., Penninga, L., Wetterslev, J., Steinbrüchel, D. A., and Gluud, C. (2012). Off-pump versus on-pump coronary artery bypass grafting for ischaemic heart disease. *Cochrane Database of Systematic Reviews*, 3:Art. No.: CD007224.

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71.

Murad, M. H., Asi, N., Alsawas, M., and Alahdab, F. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127.

National Academy of Sciences (2019). *Reproducibility and Replicability in Science*. National Academies Press, Washington, DC.

Schauer, J. M. and Hedges, L. V. (2020). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin*, 146:701–719.

References: IV

Schmid, C. H., Stark, P. C., Berlin, J. A., Landais, P., and Lau, J. (2004). Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *Journal of Clinical Epidemiology*, 57(7):683–697.

Schmid, C. H., Stijnen, T., and White, I. (2020). *Handbook of Meta-analysis*. CRC Press, Boca Raton, FL.

Senn, S. and Brand, R. (1994). Importance of trends in the interpretation of an overall odds ratio in the meta–analysis of clinical trials. *Statistics in Medicine*, 13(3):293–296.

Sharp, S. J., Thompson, S. G., and Altman, D. G. (1996). The relation between treatment benefit and underlying risk in meta-analysis. *BMJ*, 313(7059):735–738.

References: V

Xiao, M., Chen, Y., Cole, S., MacLehose, R., Richardson, D., and Chu, H. (2022a). Is or "portable" in meta-analysis? time to consider bivariate generalized linear mixed model. *Journal of Clinical Epidemiology*, 142:280.

Xiao, M., Chu, H., Cole, S. R., Chen, Y., MacLehose, R. F., Richardson, D. B., and Greenland, S. (2022b). Controversy and debate: Questionable utility of the relative risk in clinical research: Paper 4: Odds ratios are far from "portable"—a call to use realistic models for effect variation in meta-analysis. *Journal of Clinical Epidemiology*.

SCT
SOCIETY FOR CLINICAL TRIALS

44th Annual Meeting
BALTIMORE
May 21 – 24, 2023

Efficient Evidence Synthesis for Evaluating Treatment Effects by Network Meta-Analysis and Combining Grouped Data and Individual Patient Data

J. Jack Lee, Ph.D., D.D.S.
Professor of Biostatistics

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Outline

- Example of Network Meta-Analysis on Aggregate Data (AD)
  - Contrast-based analysis
  - Arm-based analysis
- Extracting Individual Patient Data (IPD) from KM Curves
  - IPDfromKM
- Usage of IPD in Evidence Synthesis
  - Evaluate treatment effect in subgroups
  - Assessing surrogate endpoints
- Combining AD and IPD
- Concluding Remarks

Network Meta-analysis on Aggregate Data (AD)

Comparing Treatment Effect of Surgical Procedures In Treating Peri-Implantitis

J. Jack Lee [1], PhD, DDS, Feng Zhang [1], MS,
Kyuchul Oh/Keitetsu Kure [2], DDS, PhD, Chun-Teh Lee [2], DDS, MS, DMSc

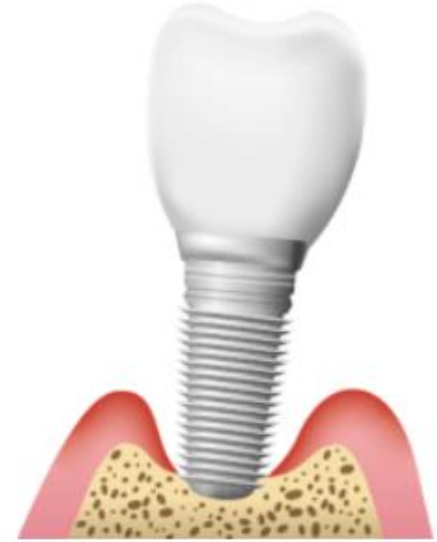
[1] The University of Texas MD Anderson Cancer Center
[2] The University of Texas School of Dentistry

Healthy

Mucositis

Peri-implantitis

Implant loss



Measure Probing Pocket Depth (PPD) in mm

Objectives

- Evaluate the treatment effects of surgical interventions in treating peri-implantitis.
- Use “debridement only” as a “control” arm.
  - Compare the effect of debridement only and additional interventions
  - Rank the effect of various treatments
- Apply network meta-analysis to synthesize information
  - Contrast-based analysis
  - Arm-based analysis

Trials Selection Criteria and Outcome Measures

- **Inclusion Criteria**
  - Only Randomized Clinical Trials
  - At least 12 months for followed-up
  - Published in the English language
  - Applied surgeries for treating peri-implantitis
  - Reported on at least one clinical or radiographic parameter

- **Outcome Measures Include:**
  (1) **Probing Pocket Depth (PPD) in mm**
  (2) Clinical Attachment Level (CAL) in mm
  (3) Bleeding On Probing (BOP) in %
  (4) Radiographic Marginal Bone Level (MBL) in mm
  (5) Gingival Recession (REC) in mm

Alternative Study ID by Contrasts and by Arms

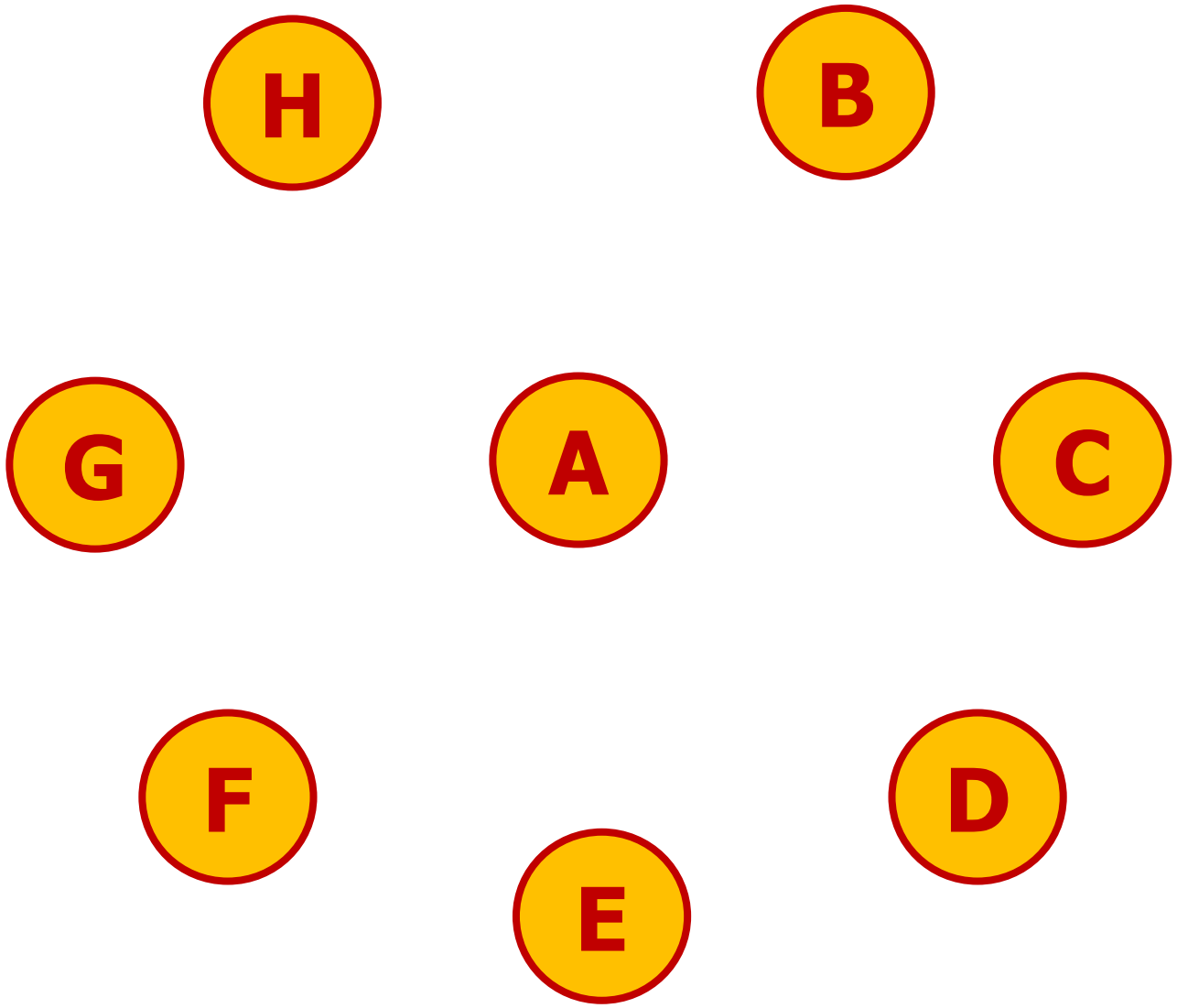
Alt ID	Author	Year	Arm	N
1	Emanuel	2020	A	14
			B	18
2	Renvert	2021	A	32
			C	34
3	Renvert	2018	A	20
			B	21
4	Isehed	2016	A	14
			E	15
5	Carcuac	2016	A	24
			A	24
			G	27
			G	25
6	Wohlfahrt	2012	A	16
			F	16
7	Hallström	2017	A	16
			G	15
8	Albaker	2018	A	13
			H	11
9	Jepsen	2016	A	30
			F	33
10	Schwarz	2009	B	9
			C	10

Alt ID	Author	Year	Arm	N
11	Wagener	2021	A	24
21	Tapia	2018	B	15
22		2018	B	15
23	Polymeri	2020	B	13
24		2020	B	11
31	Schwarz	2012	C	14
32		2012	C	10
33	Isler	2018	C	26
34		2018	C	26
35	Aghazadeh	2012	C	23
36		2012	C	22
41	de Waal	2012	D	15
42		2012	D	15
43	Romeo	2005, 2007	D	10
44		2005, 2007	D	7
45	de Waal	2014	D	22
46		2014	D	22

Arm Description

Arm A	**DBT (Surgical Debridement) alone**
Arm B	**DBT+Bone Graft**
Arm C	**DBT+Bone Graft+Membrane**
Arm D	**DBT+Resective surgery**
Arm E	**DBT+EMD**
Arm F	**DBT+Porous Titanium Granule**
Arm G	**DBT+Adjunct Antibiotics**
Arm H	**DBT+aPDT**

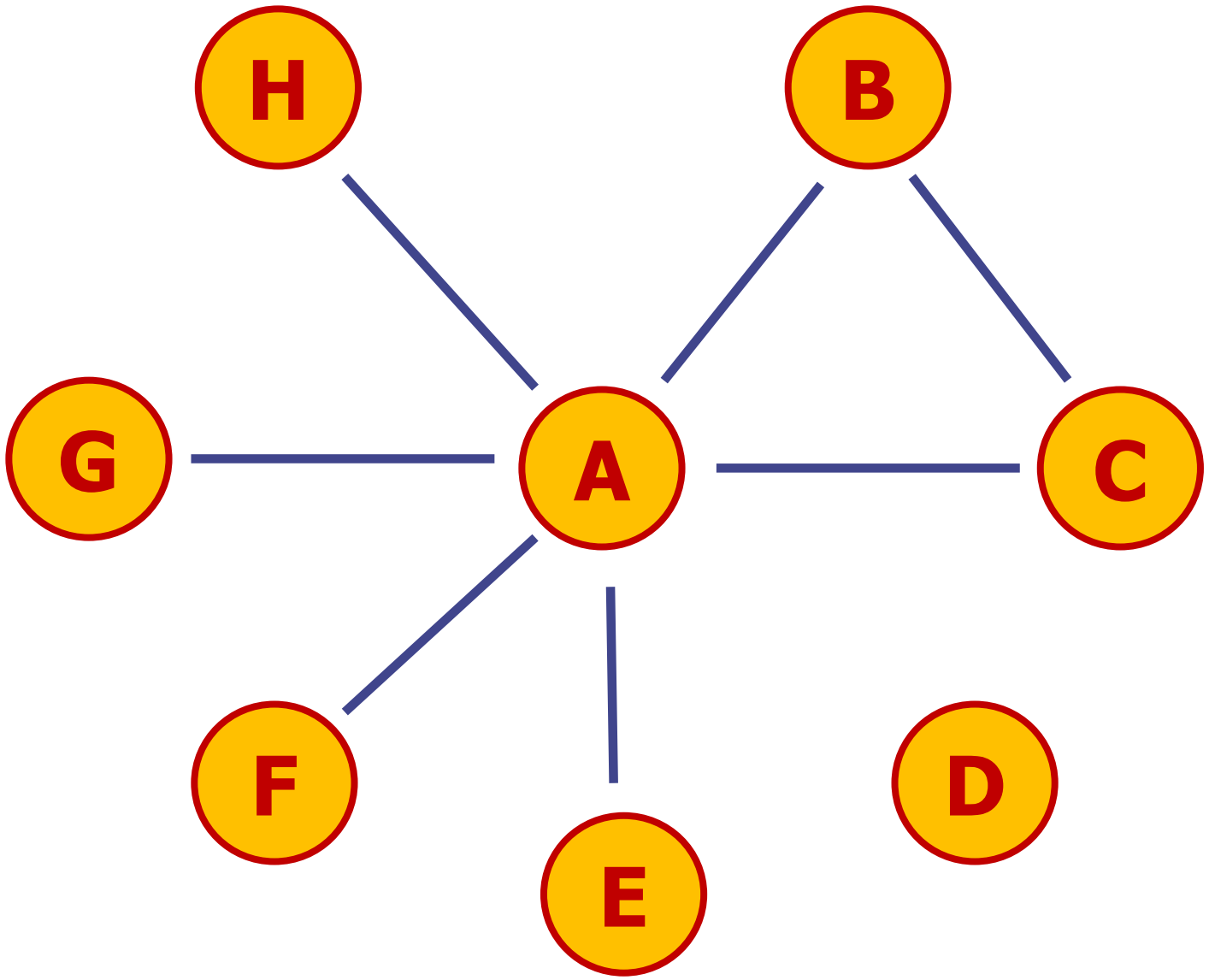
Network Diagram (Treatment Arms)



Arm Description

Arm A	DBT (Surgical Debridement) alone
Arm B	DBT+Bone Graft
Arm C	DBT+Bone Graft+Membrane
Arm D	DBT+Resective surgery
Arm E	DBT+EMD
Arm F	DBT+Porous Titanium Granule
Arm G	DBT+Adjunct Antibiotics
Arm H	DBT+aPDT

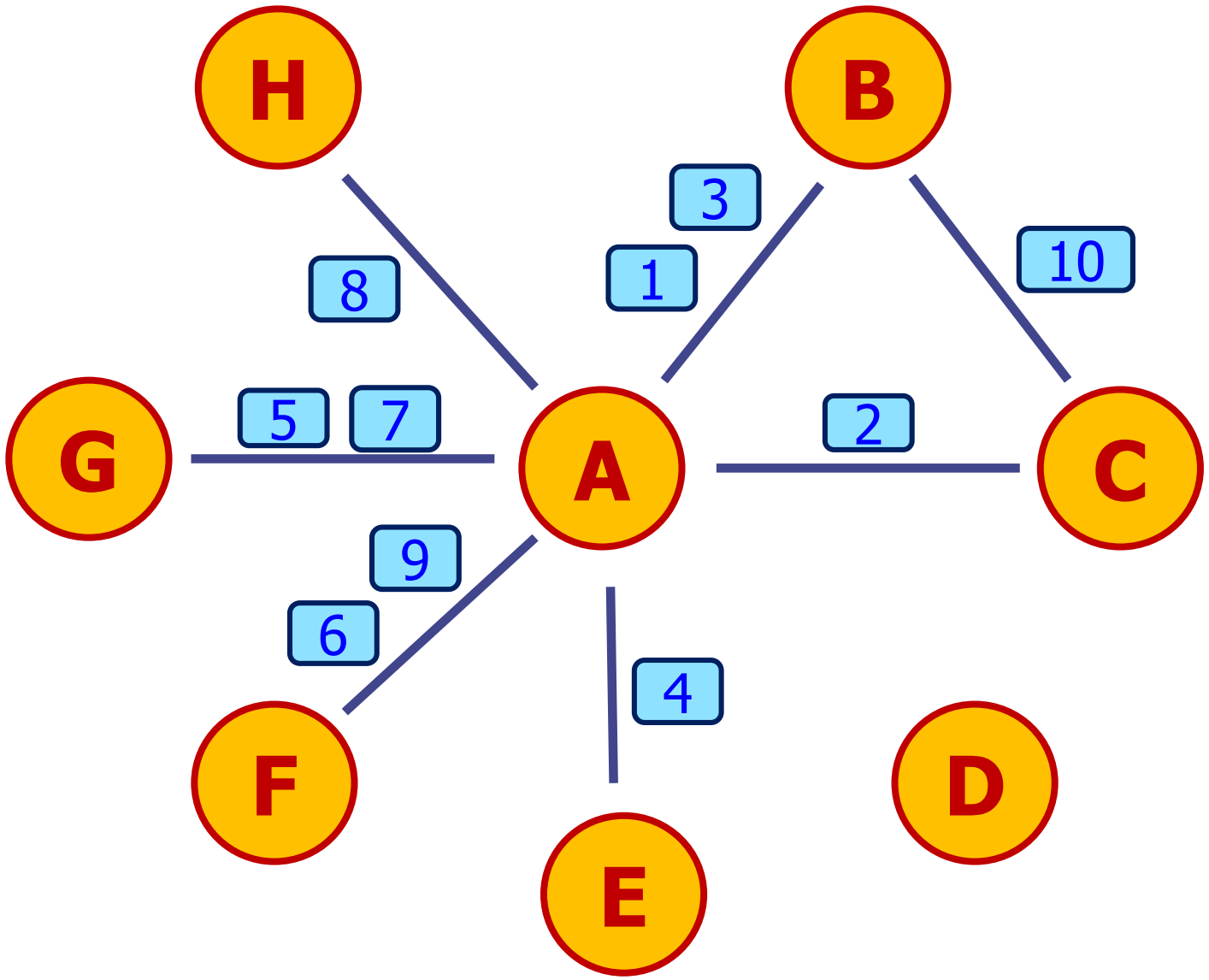
Network Diagram (Trials and Treatment Arms)



Arm Description

Arm A	DBT (Surgical Debridement) alone
Arm B	DBT+Bone Graft
Arm C	DBT+Bone Graft+Membrane
Arm D	DBT+Resective surgery
Arm E	DBT+EMD
Arm F	DBT+Porous Titanium Granule
Arm G	DBT+Adjunct Antibiotics
Arm H	DBT+aPDT

Network Diagram (Trials and Treatment Arms)



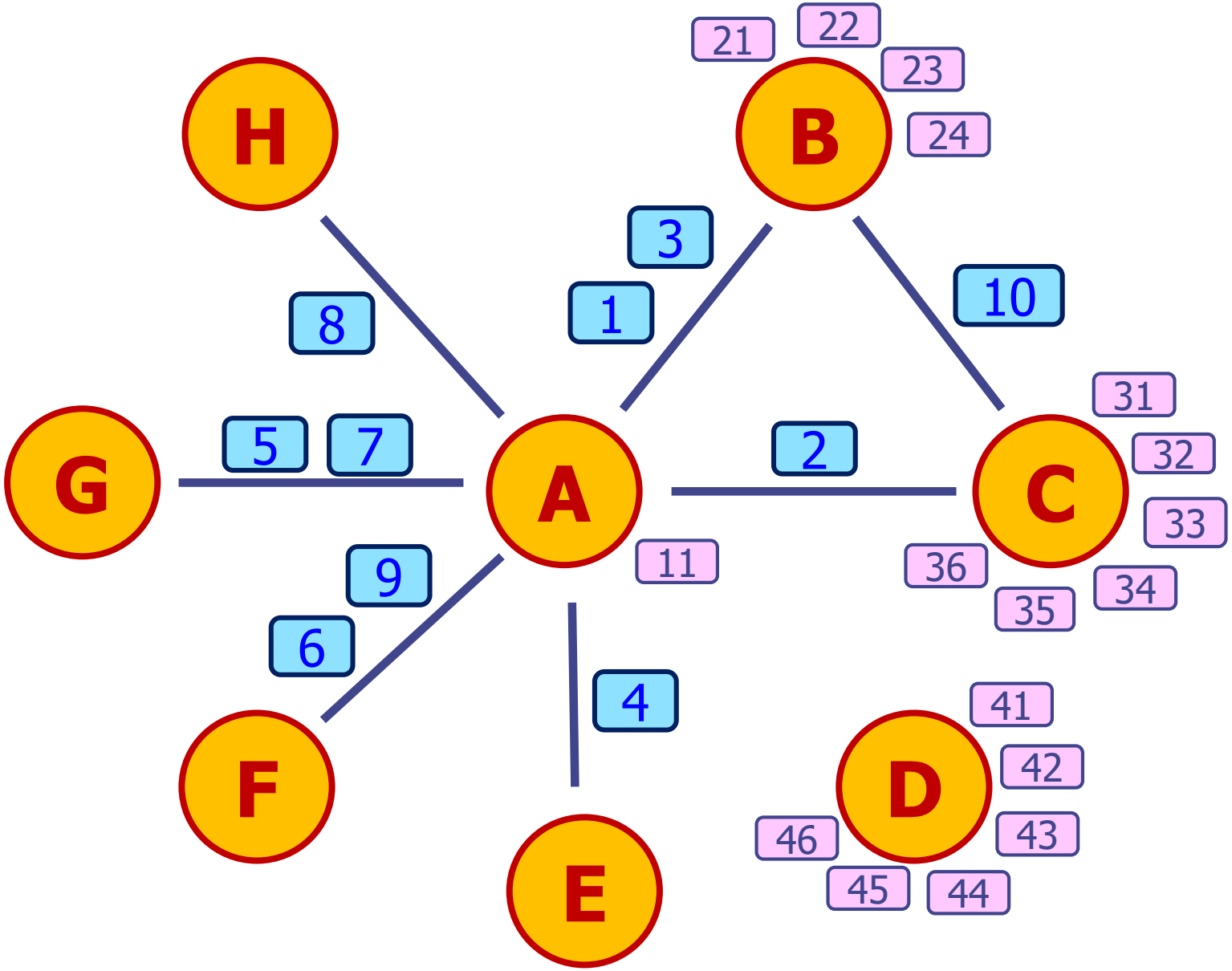
Arm Description

Arm A	DBT (Surgical Debridement) alone
Arm B	DBT+Bone Graft
Arm C	DBT+Bone Graft+Membrane
Arm D	DBT+Resective surgery
Arm E	DBT+EMD
Arm F	DBT+Porous Titanium Granule
Arm G	DBT+Adjunct Antibiotics
Arm H	DBT+aPDT

Arm

Trial

Network Diagram (Trials and Treatment Arms)

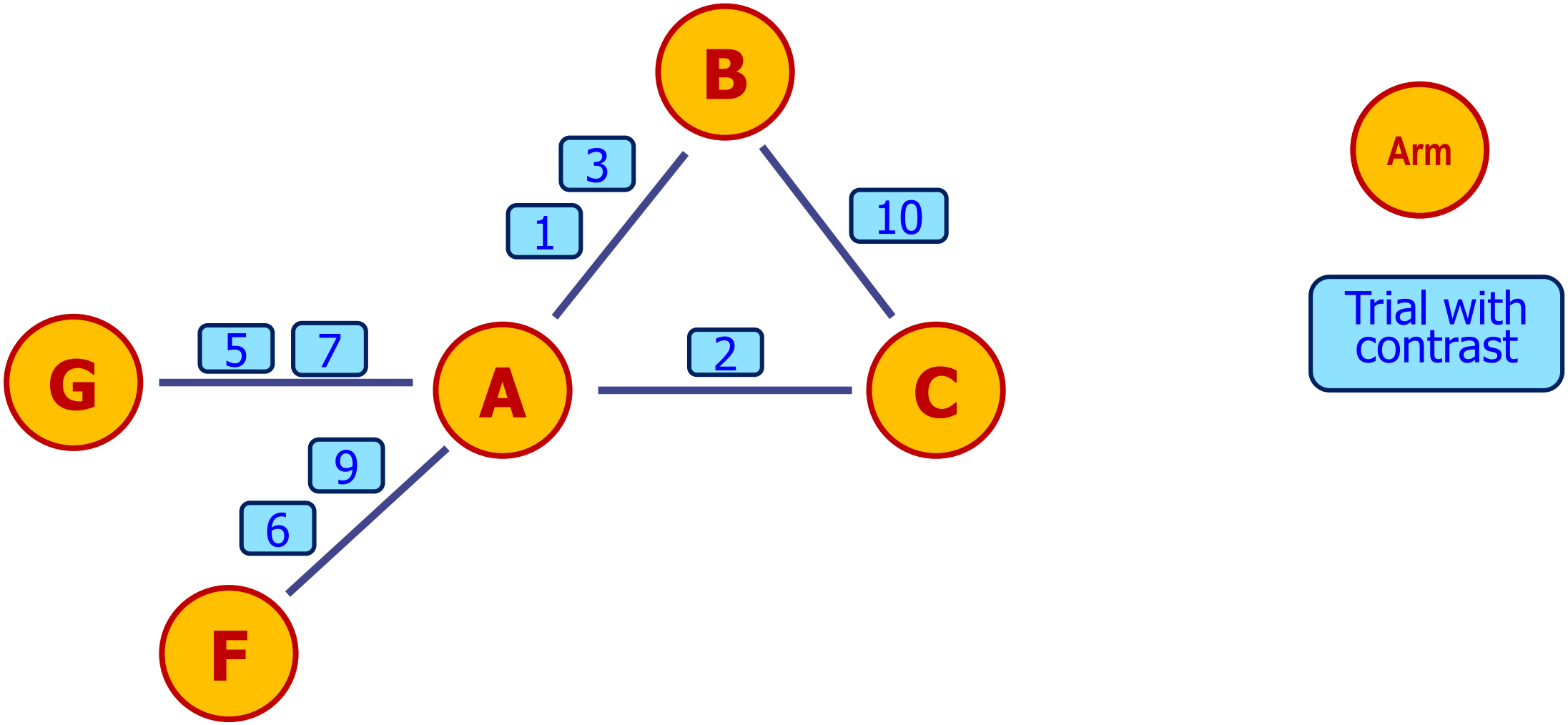


Arm

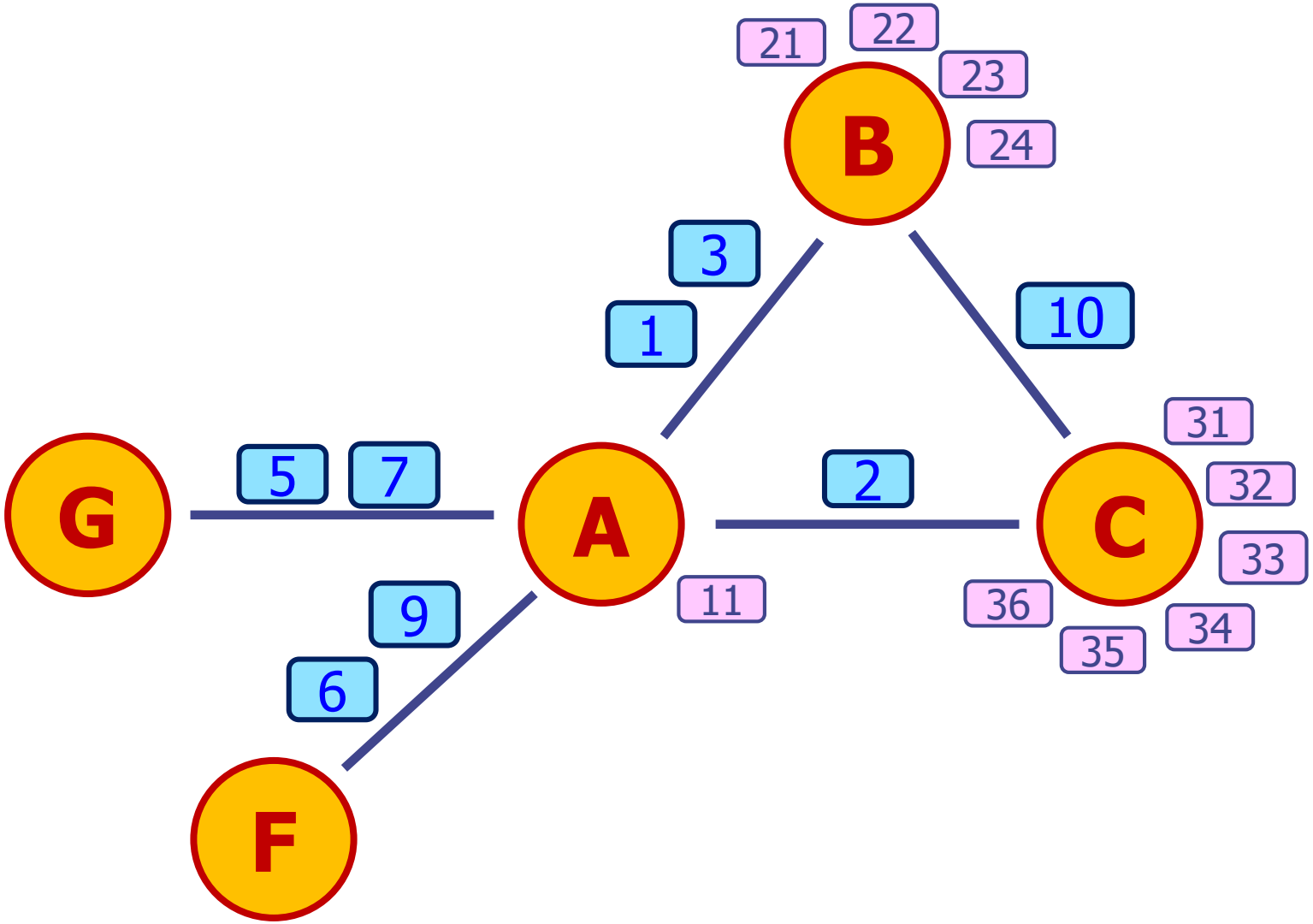
Trial with contrast

Trial with arm only

Analysis on PPD (Treatment Contrasts Only)



Analysis on PPD (Treatment Contrasts & Arms)



Arm

Trial with contrast

Trial with arm only

Statistical Methods for Network Meta-Analysis

- Continuous variable
  - Contrast-based analysis for Study i, arm k vs b

$$Y_{ibk} \sim N(\delta_{ibk},\ s_i{}^2)$$

$$\delta_{ibk} \sim N(\delta_{bk}, \tau_{bk}{}^2), \text{ where } \delta_{bk} = d_{1k} - d_{1b},$$

d_{1k} and d_{1b} are the treatment effect of arm k and b versus the reference arm 1.

prior for $d_{1k} \sim N(0, 10^3)$, $d_{1b} \sim N(0, 10^3)$, $d_{11} = 0, \tau_{bk} = \tau$

  - Arm-based analysis for Study i, arm k

$$Y_{ik} \sim N(\theta_{ik},\ s_{ik}{}^2),$$

where $\theta_{ik} = \mu_{ib}$ $\quad if\ k = b;$

$\theta_{ik} = \mu_{ib} + \delta_{ibk}$ $\quad if\ k > b$

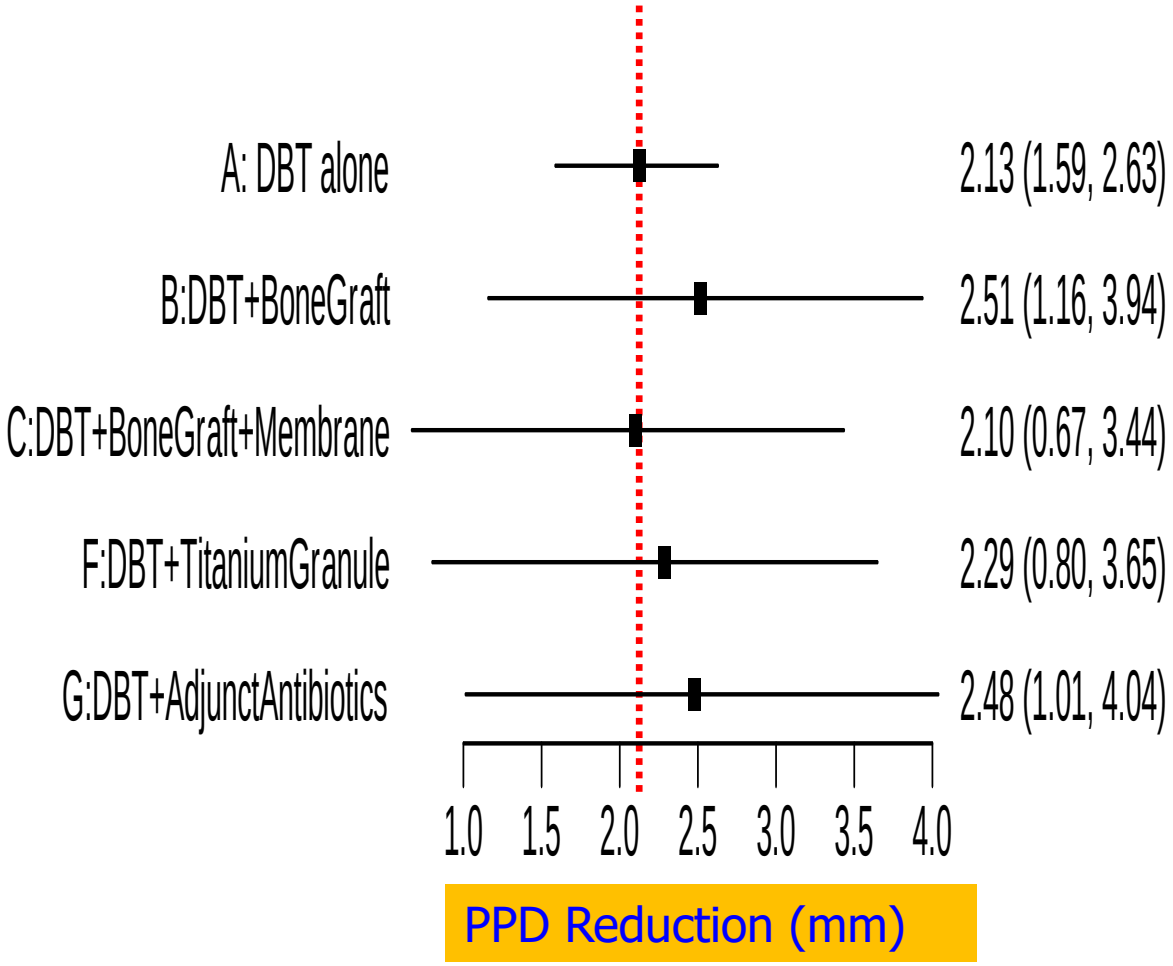
$$\delta_{ibk} \sim N(\delta_{bk}, \tau_{bk}{}^2)$$

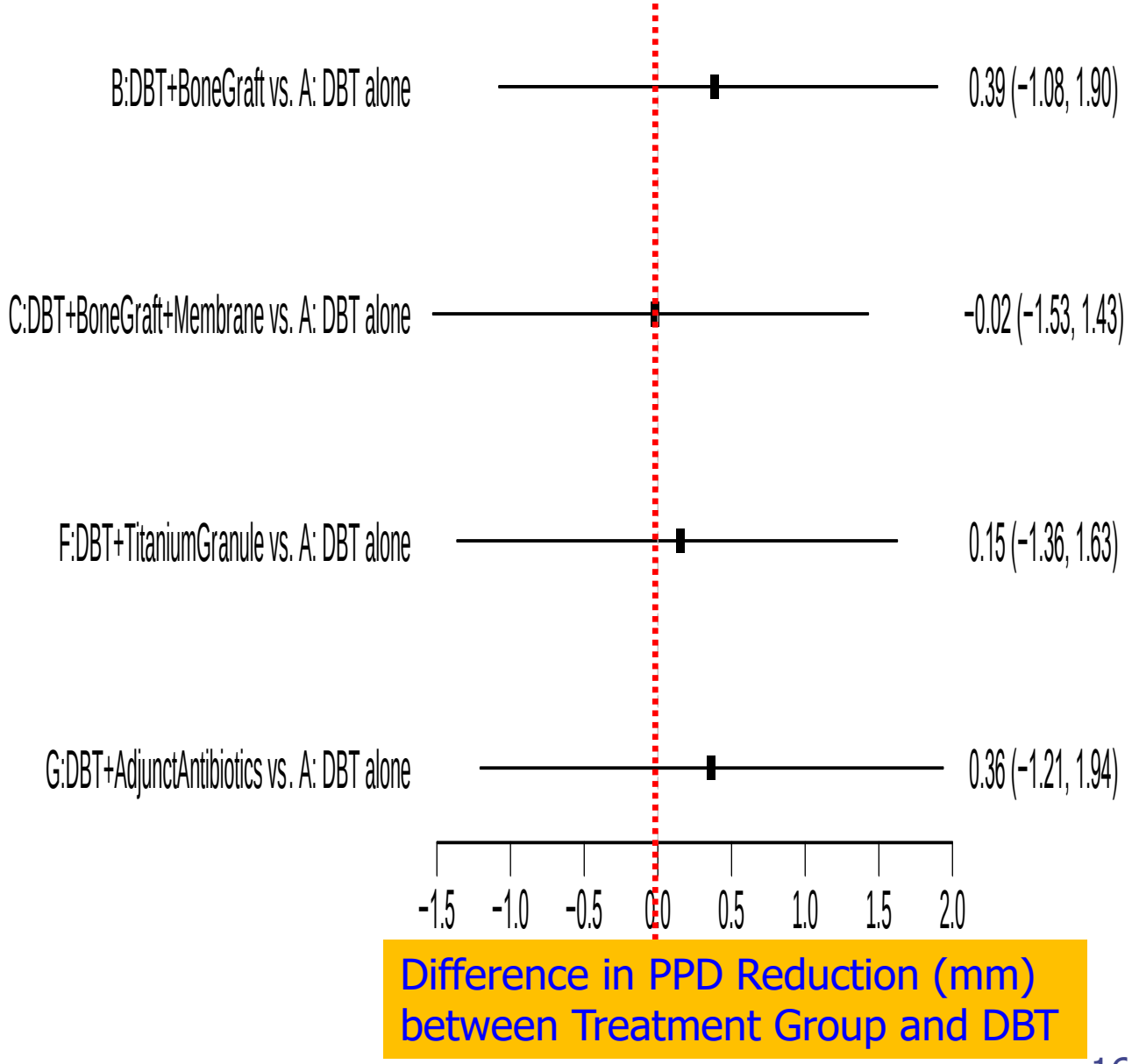
prior for $d_{1k} \sim N(0, 10^3)$, $d_{1b} \sim N(0, 10^3)$, $d_{11} = 0, \tau_{bk} = \tau$

Statistical Methods

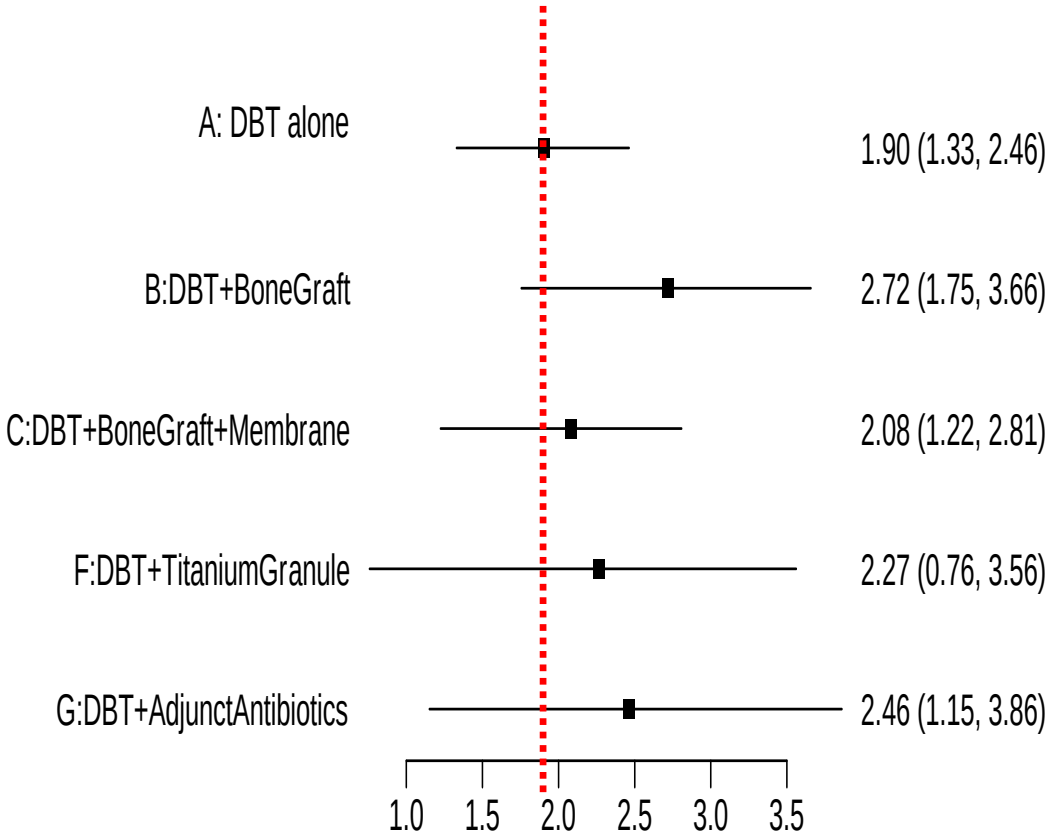
- R Packages for arm-based data
  - pcnetmeta
  - Gemtc
    - Both work under Bayesian framework. Prior can be specified.
- Assumptions
  - Endpoint follows normal distribution (test performed)
  - Take probing pocket depth (mm) as the primary endpoint.
    - Take the difference of baseline and 12-month measures
    - Positive number means pocket depth is reduced, hence, treatment works
  - Allow the entire covariance matrix to vary between trials, but to be sampled from a common *Wishart* distribution, $\sum_i \sim \tau^2\, Wishart(C; 2)$.
- Sensitivity analysis
  - Default, homogeneous equal correlation: inverse gamma, a = 0.001, b = 0.001
  - heterogeneous equal correlation: uniform (0,3)

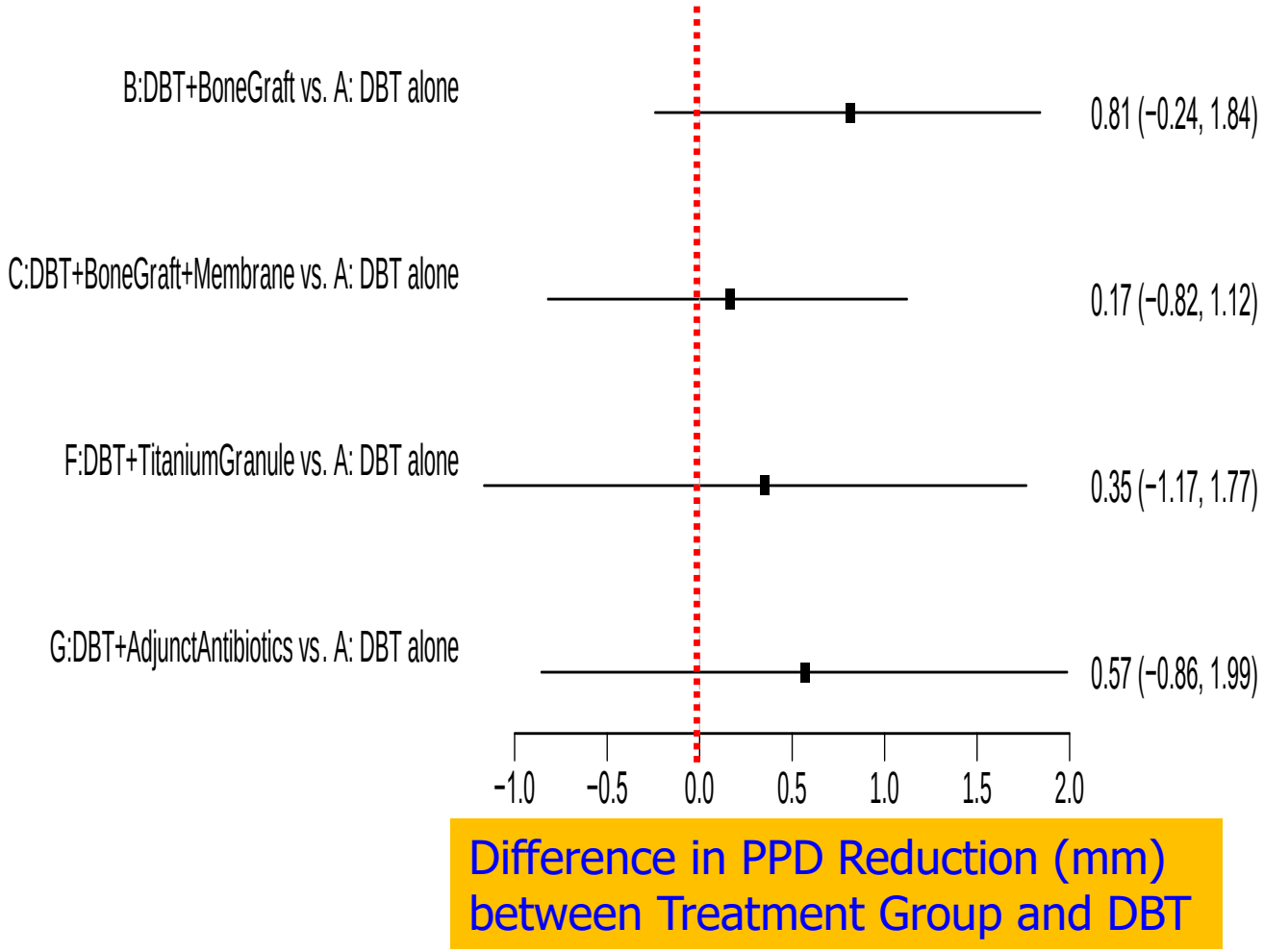
Results: Contrasts-Based Analysis



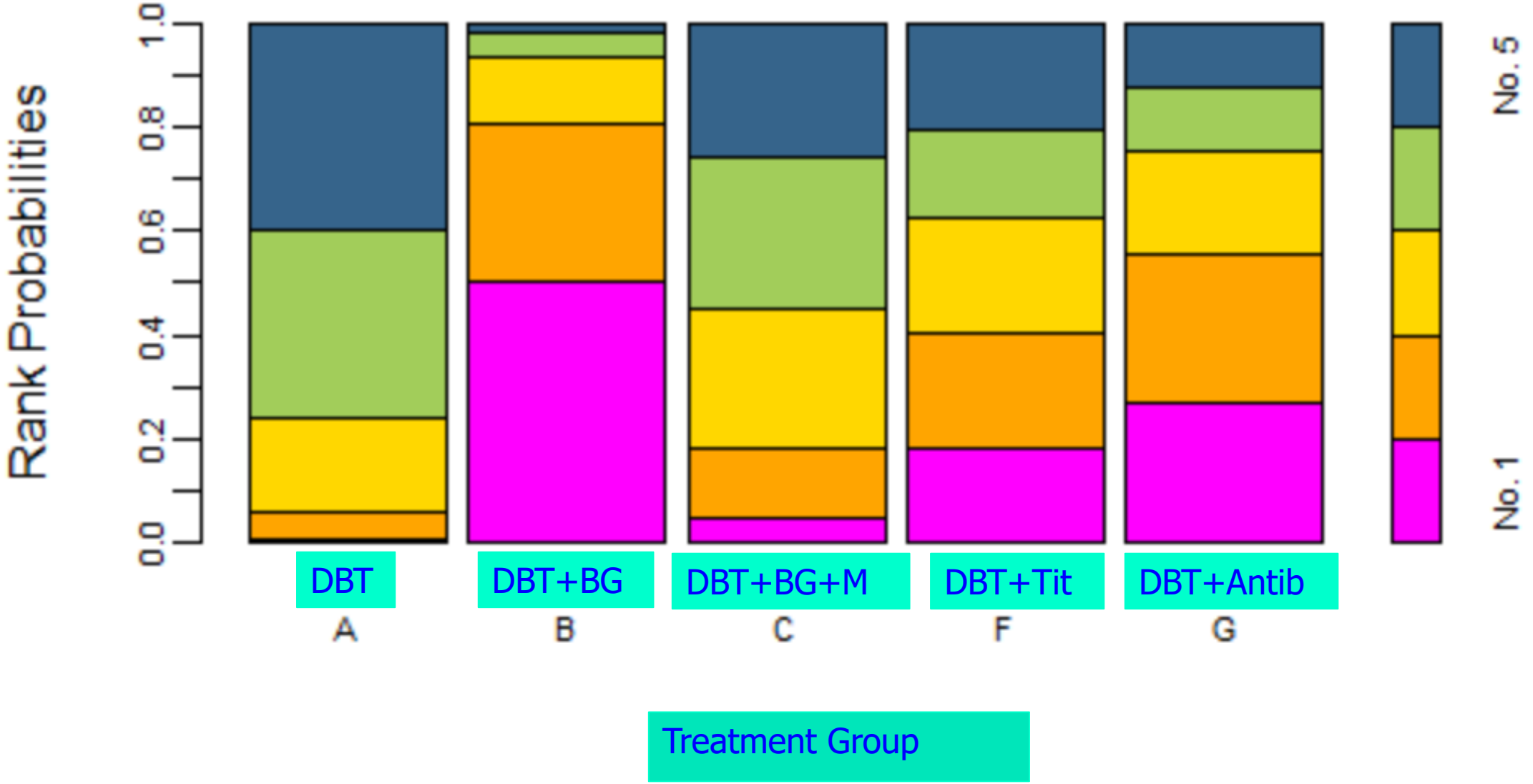


Results: Arm-based Analysis

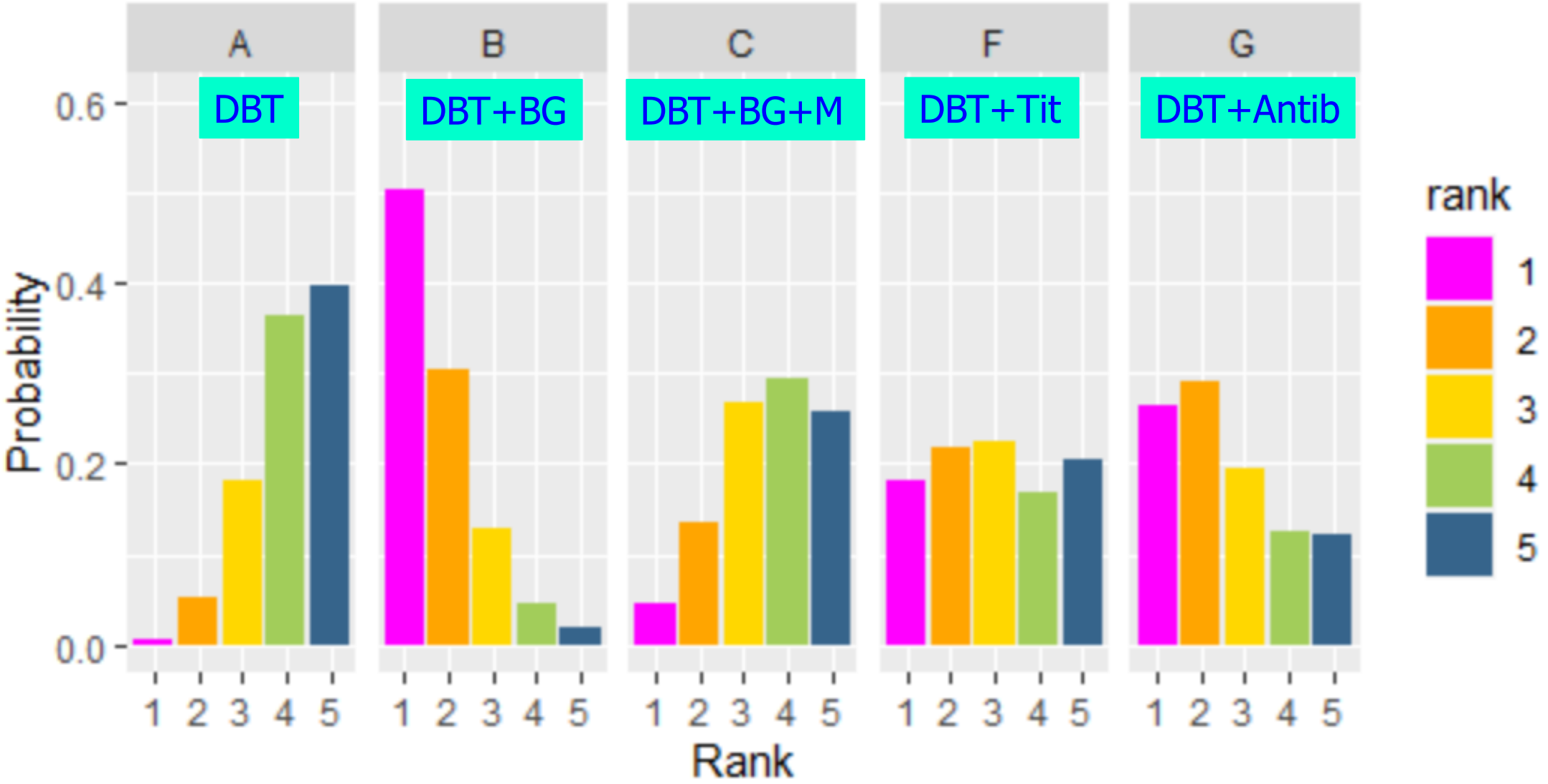




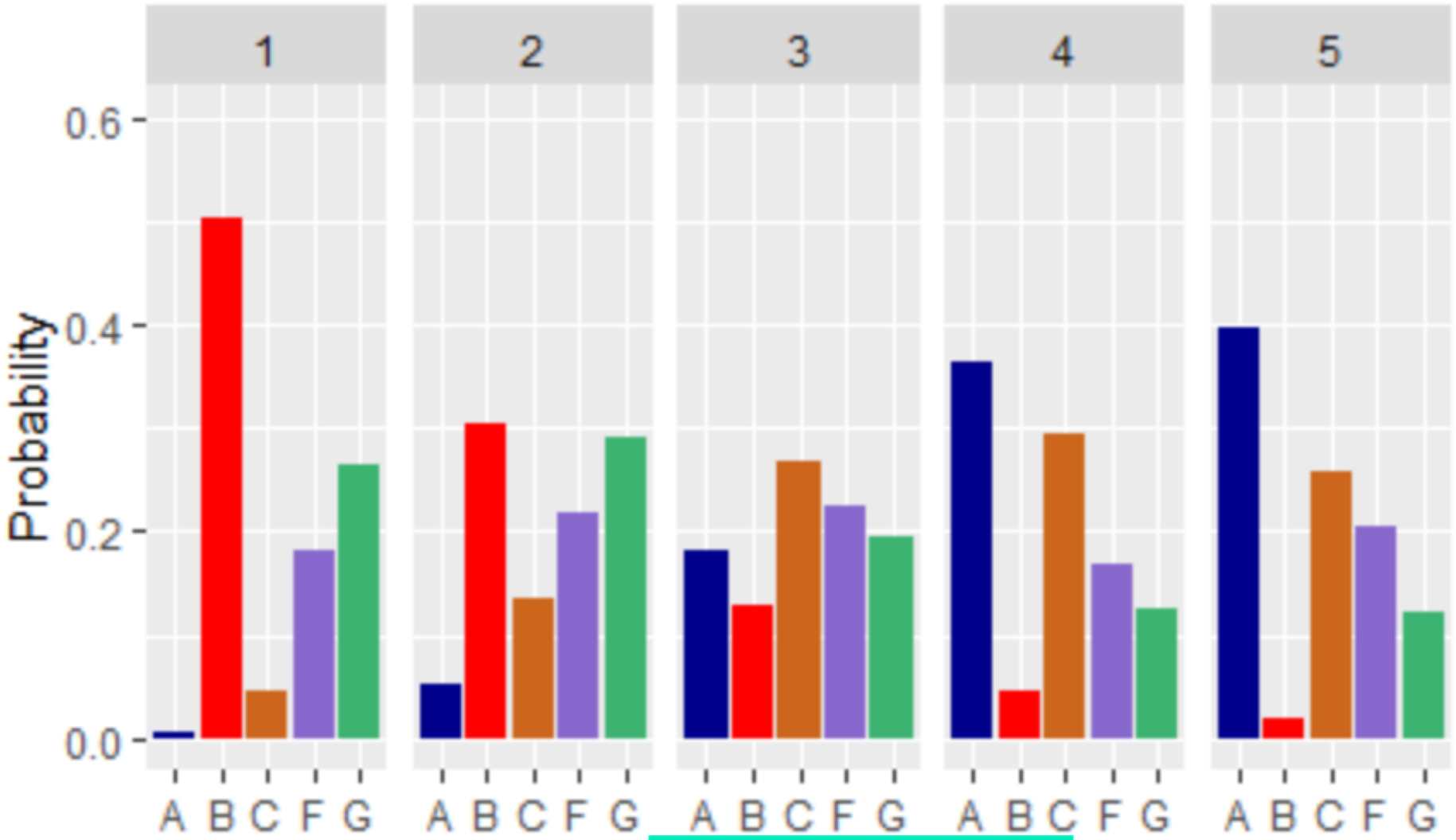
Treatment Ranking – Stacking Barchart (Arm-based)

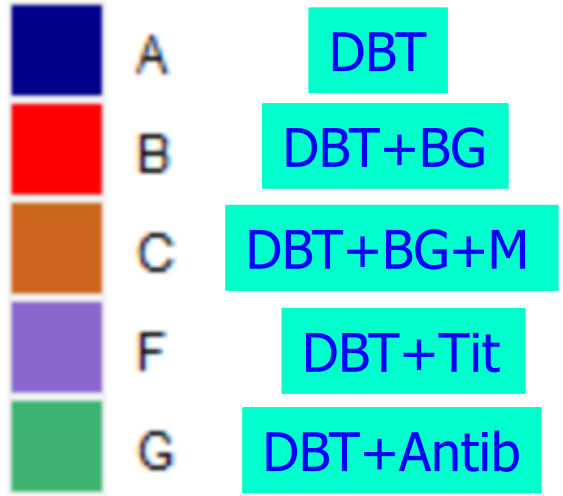


Treatment Ranking – Side-by-Side Barchart (Arm-based) [by Arm]



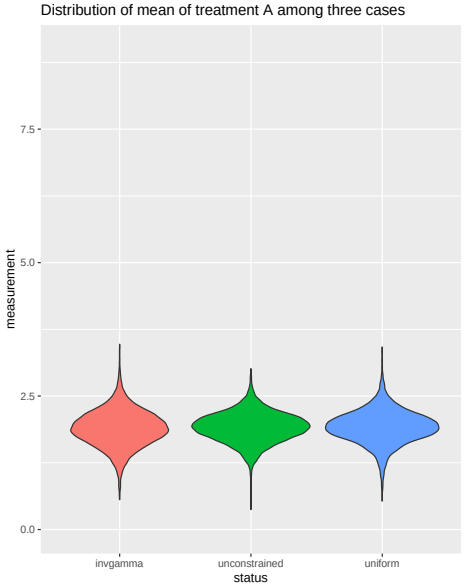
Treatment Ranking – Side-by-Side Barchart (Arm-based) [by Ranking]





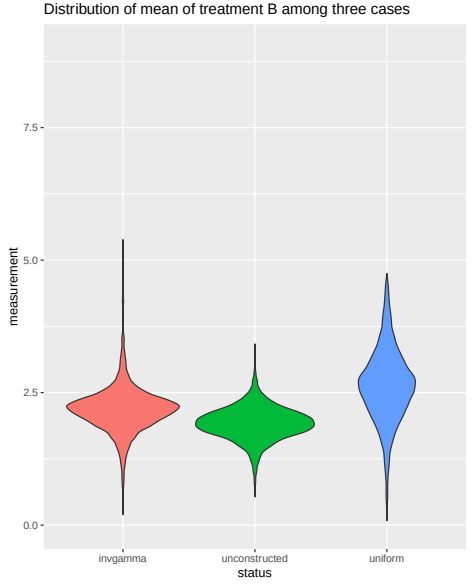
Sensitivity Analysis of Mean Treatment Effect by Varying Priors of the Covariance Matrix

Arm A



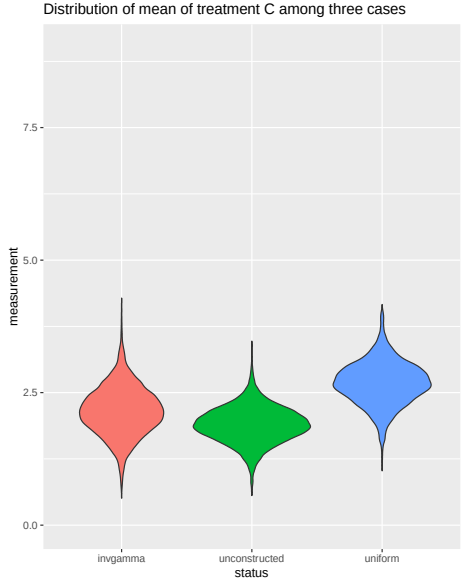
DBT

Arm B



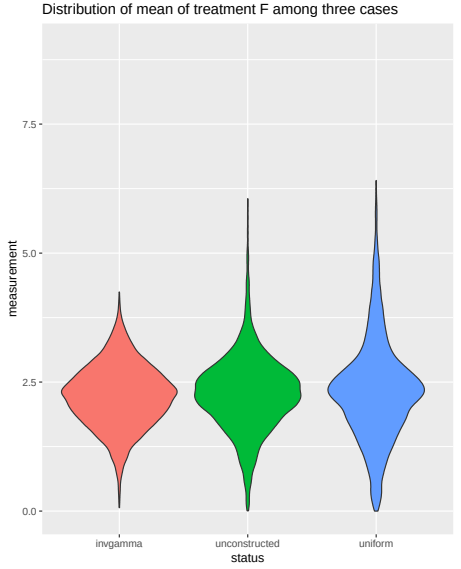
DBT+BG

Arm C



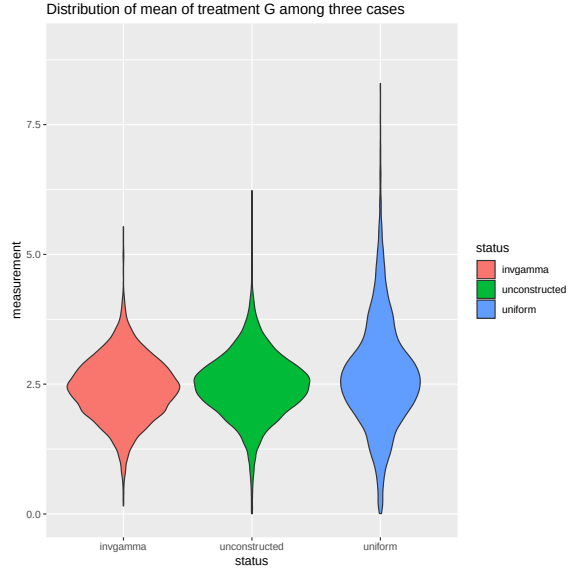
DBT+BG+M

Arm F



DBT+Tit

Arm G



DBT+Antib

Red: inverse gamma
Green: Unstructured
Blue: Uniform

Summary

- Network meta-analysis is a power tool for synthesizing evidence across trials with multiple treatment arms.
- Arm-based approach can incorporate more information than the contrast-based approach.
- Ranking of the treatment effect:
  1. DBT + graft
  2. DBT + antibiotics
  3. DBT + titanium granule
  4. DBT + graft + membrane
  5. DBT alone
- Common limitations and caveats of meta-analysis such as patient heterogeneity, sampling bias, reporting bias, etc. also apply.

Bayesian Network Meta-Analysis for Binary Endpoint or Survival Rate at a Landmark Time

- Compare 1-yr event-free survival among treatments
- Model: Let r_{ik}, p_{ik}, n_{ik} be the # of no-event, prob of no-event, sample size for trial i, treatment k. , $\delta_{i,bk}$ for ln(OR) of tx k vs. b

$$r_{ik} \sim Bi(p_{ik}, n_{ik}), \quad i = 1, 2, \ldots N; \; k = 1, 2, \ldots K$$

$$logit(p_{ib}) = \log\left(\frac{p_{ib}}{1 - p_{ib}}\right) = \mu_i, \quad i = 1, 2, \ldots N; \; k = b = 1, 2, \ldots K$$

$$logit(p_{ik}) = \log\left(\frac{p_{ik}}{1 - p_{ik}}\right) = \mu_i + \delta_{i,1k}, \quad i = 1, 2, \ldots N; \; k = 2, 3, \ldots K; b < k$$

Fixed Effect

$$logit(p_{ik}) = \mu_i + d_{i,bk}, \quad i = 1, 2, \ldots N; \; b = 1, 2, \ldots K; \; k = 2, 3, \ldots K; \; b < k$$

$$logit(p_{ib}) = \mu_i + \beta x_i, \quad i = 1, 2, \ldots N; \; b = 1, 2, \ldots K$$

$$logit(p_{ik}) = \mu_i + d_{i,1k} + \beta x_i, \quad i = 1, 2, \ldots N; \; k = 2, 3, \ldots K; \; b < k$$

Random Effect

$$\delta_{i,bk} \sim N(d_{bk}, \sigma^2)$$

$$logit(p_{ik}) = \mu_i + \delta_{i,bk} + \beta x_i, \quad i = 1, 2, \ldots N; \; b = 1, 2, \ldots K; \; k = 2, 3, \ldots K; \; b < k$$

IPDfromKM: Reconstruct Individual Patient Data From Published Kaplan-Meier Survival Curves

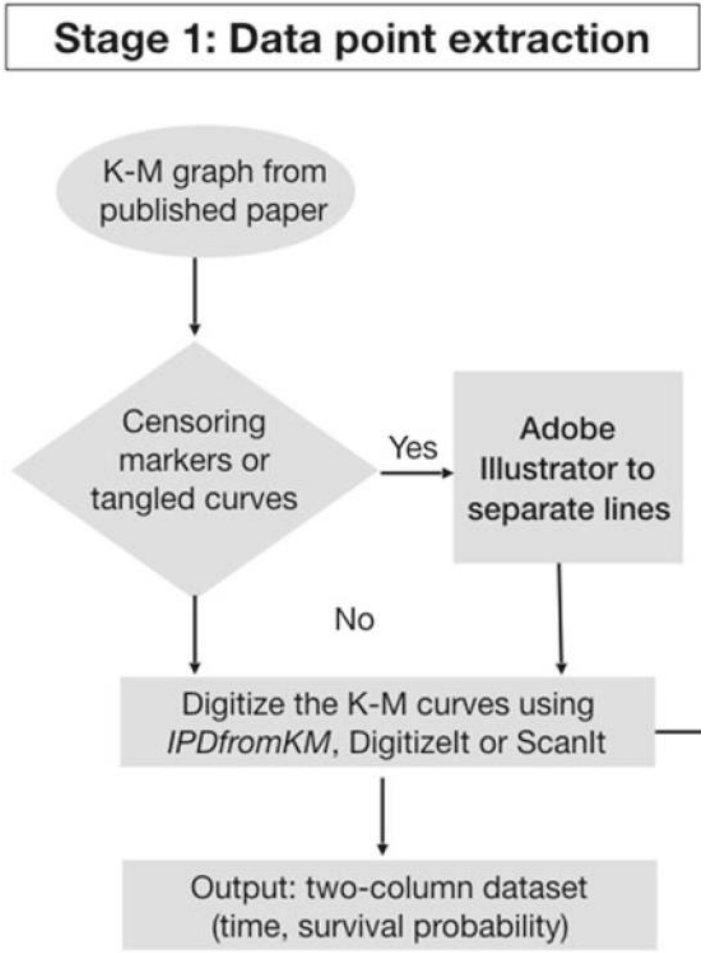
- Overview of the method IPDfromKM
- Two-stage modified-iKM algorithm for IPD (individual patient data) reconstruction
  - Extract coordinates from published K-M curves
  - IPD reconstruction and accuracy assessment
    - ➢ By R package IPDfromKM
    - ➢ By Shiny application
- Some Results from simulation study
- Concluding remarks

Liu, N.; Zhou, Y.; Lee, J.J. IPDfromKM: Reconstruct individual patient data from published Kaplan-Meier survival curves. BMC Med. Res. Methodol. **2021**

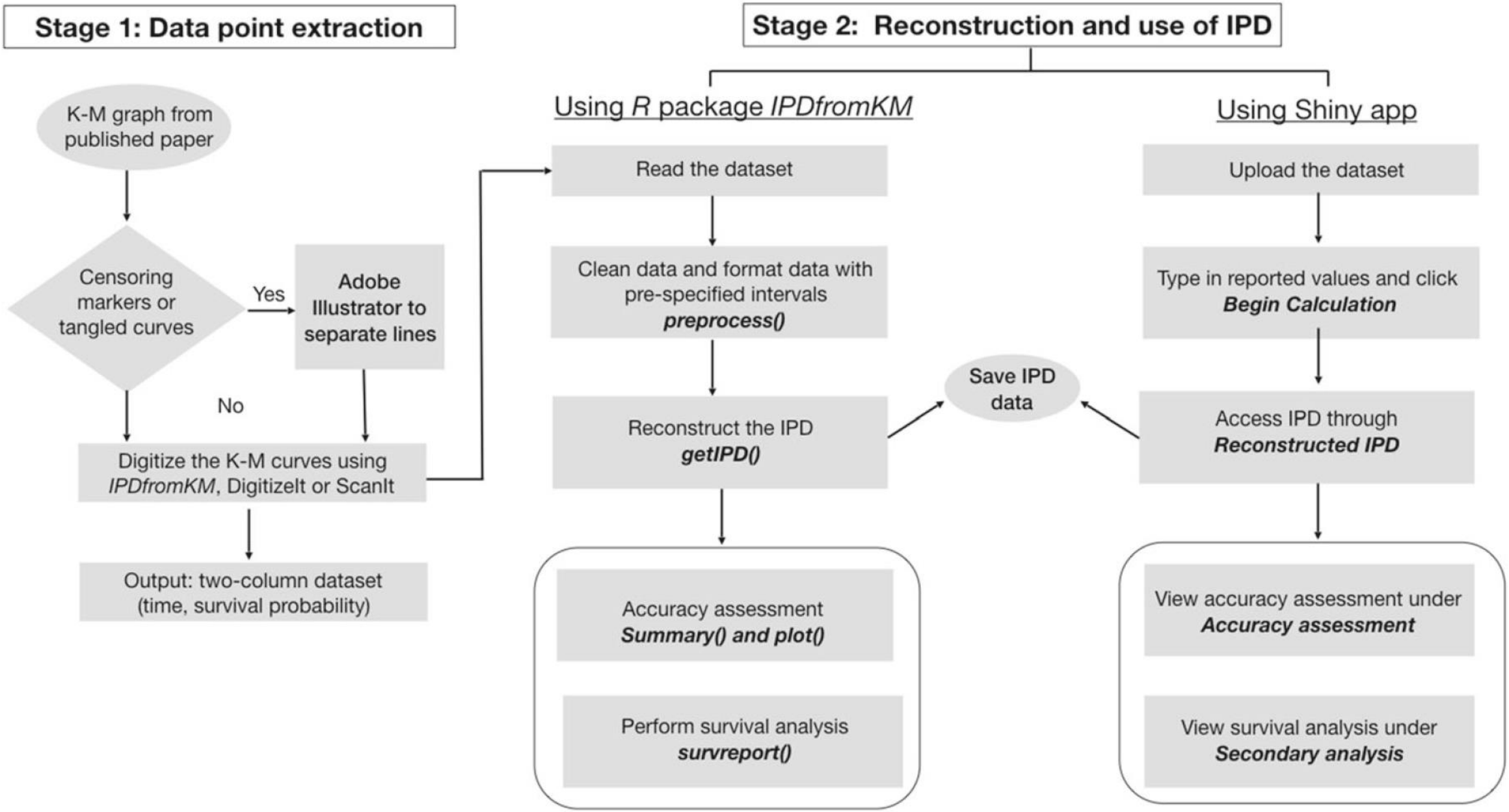
Overview of the Method

- Individual patient data (IPD) is critical for secondary analysis on published survival data.
- Researchers often lack access to IPD.
- Our software is an easy-to-use, two-stage approach to reconstruct IPD from published survival curves with "All-in-One" feature.
  - Extract raw data coordinates from published K-M curves
  - Reconstruct IPD from the extracted data coordinates
  - Visualize the reconstructed IPD
  - Assess the accuracy of the reconstruction
  - Perform secondary analysis on the basis of the reconstructed IPD
- > 100 citations

Flowchart of the IPD reconstruction platform



Flowchart of the IPD reconstruction platform



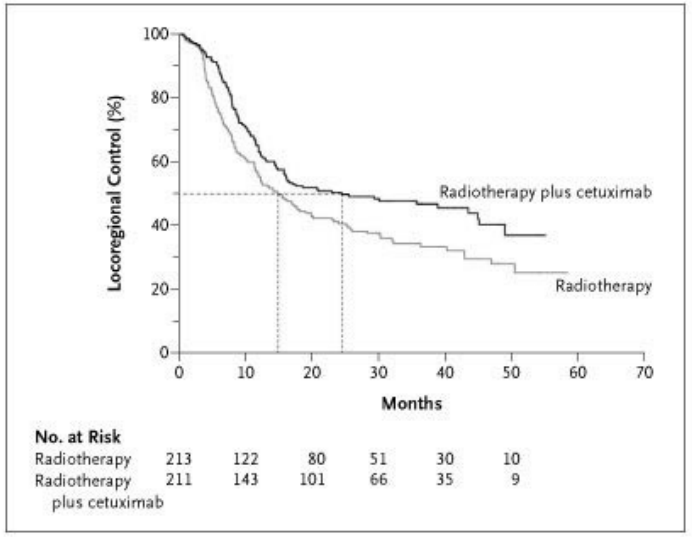
Coordinates Extraction in Stage 1

- Main challenge of digitizing K-M curves comes from tangled lines and censoring markers.
- There are several options in this stage:
  - Adobe Illustrator can be used to separate the tangled together curves before extracting data points
  - Using R package or Shiny application we developed
  - Other commonly used software like DigitizeIt(https://www.digitizeit.de) or ScanIt (https://www.amsterchem.com/scanit.html)
  - Video tutorials can be found in our Shiny page (https://biostatistics.mdanderson.org/shinyapps/IPDfromKM/)

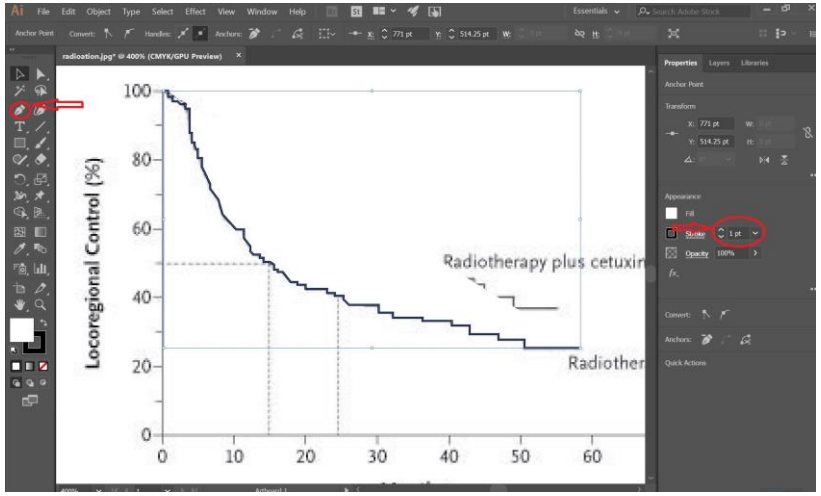
Quality data extraction is the key of accurate IPD reconstruction!

Evenly distributed & sufficiently large number of points yield better results!

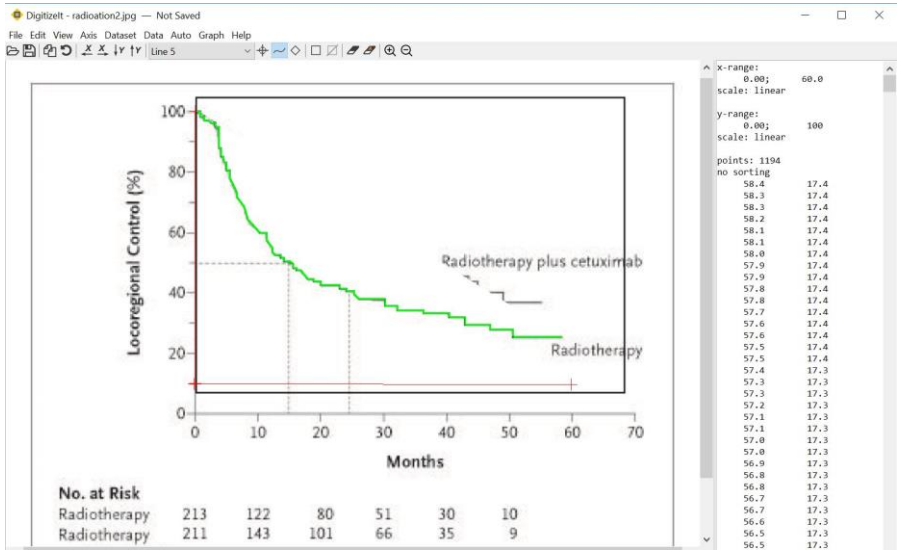
One Example Using Adobe and DigitizeIt to Extract Data



Step1: the original curve



Step2: use Adobe to separate out one curve



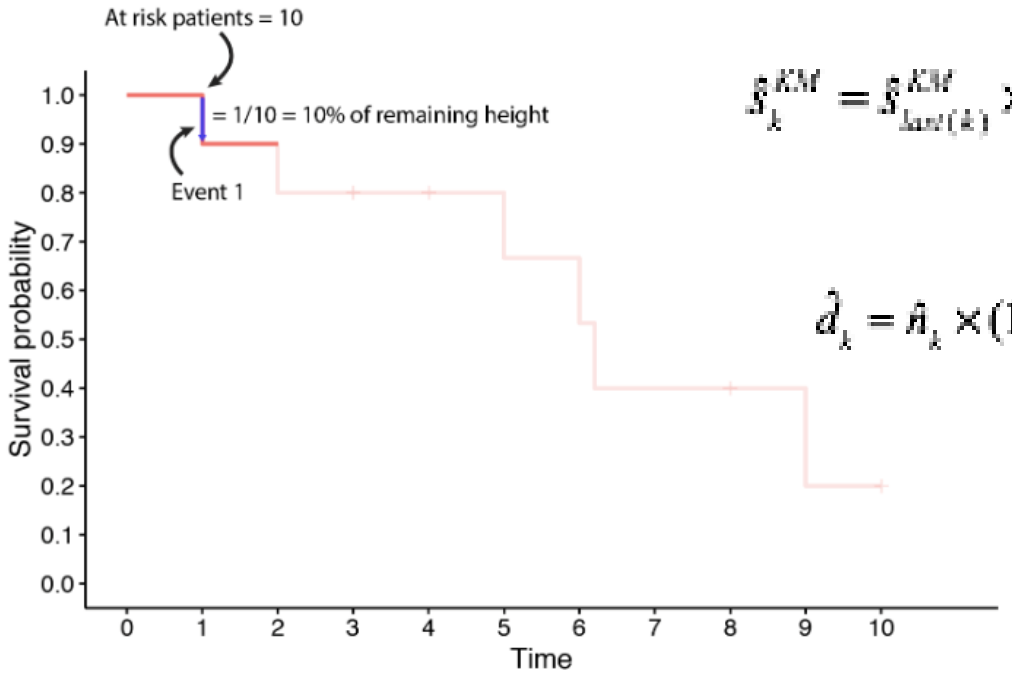
Step3: use DigitizeIt to extract coordinates into the N × 2 table (time, survival probability)

```
> data <- read.csv(file='radio.csv',sep=",",head=T)
> head(data)
   Line.2.x Line.2.y
1 0.2667810 99.78312
2 0.2931803 99.78312
3 0.3254957 99.78312
4 0.3618950 99.78312
5 0.4112606 99.78312
6 0.4611677 99.78312
> dim(data)
[1] 2524    2
```

Step4: read the table in R

R Package IPDfromKM

- Based on an iterative numerical algorithm published by Guyot *et al.* (2012) with several improvements and extensions.



$$\hat{S}_k^{KM} = \hat{S}_{last(k)}^{KM} \times \left(1 - \frac{\hat{d}_k}{\hat{n}_k}\right)$$

$$\hat{d}_k = \hat{n}_k \times \left(1 - \frac{\hat{S}_k^{KM}}{\hat{S}_{last(k)}^{KM}}\right) = \hat{n}_k \times \frac{\hat{S}_{last(k)}^{KM} - \hat{S}_k^{KM}}{\hat{S}_{last(k)}^{KM}}$$

- Main assumption: a constant censoring rate within each interval.
- Strategy: iteratively adjust the censored patients number within each time interval, until the reported number of patients at risk equals the estimated number.

Functions in R Package IPDfromKM

Table 1 Overview of the user visible functions in *IPDfromKM*

Function	Description	Object returned
`getpoints`	Extract raw data coordinates from published K-M curves.	A data frame containing the x- and y- coordinates of the K-M curve of interest.
`preprocess`	Preprocess the read-in data coordinates.	A list including cleaned data ready for reconstruction and a "riskmat" table displaying the index of read-in points within each time interval.
`getIPD`	Estimate the IPD.	A list including the reconstructed IPD.
`survreport`	Perform survival analysis on reconstructed IPD.	K-M curve, cumulative hazard, times for targeted survival probabilities.
`plot`	Plot the object returned by getIPD().	K-M curves and number at risk for both reconstructed IPD and read-in data.
`summary`	Summarize objects returned by getIPD().	Descriptive results for accuracy assessment and survival analysis on reconstructed IPD.

Please consult the documentation (e.g., help("preprocess")) for function arguments and detailed return types

A Build-in Data Set as an Example

- Radiation data was extracted from published K-M curves from a two-arm randomized controlled trial using ScanIt.
- Two treatment groups of neck cancer patients: 213 in the radiotherapy group (referred to as “radio”) and 211 in the radiotherapy plus cetuximab group (referred to as “radio_plus”).

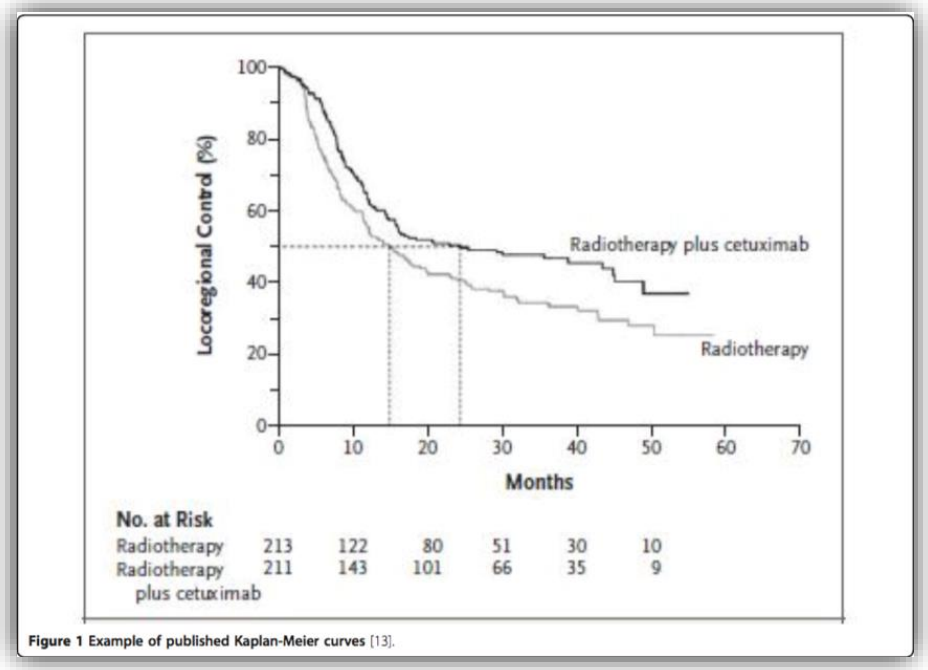


Figure 1 Example of published Kaplan-Meier curves [13].

```
> library(IPDfromKM)
> names(Radiationdata)
[1] "radio"           "radioplus"       "trisk"
"nrisk.radio"     "nrisk.radioplus"
> Radiationdata$trisk
[1]  0 10 20 30 40 50 60 70
> Radiationdata$nrisk.radio
[1] 213 122  80  51  30  10
> Radiationdata$nrisk.radioplus
[1] 211 143 101  66  35   9
```

Prepare Data Using Preprocess( ) Function

- Sort the data by time
- Make monotonicity adjustment to ensure that survival probability decreases over time
- Perform step control

```
> pre_radio <- preprocess(dat=Radiationdata$radio,
+     trisk=Radiationdata$trisk,
+     nrisk=Radiationdata$nrisk.radio,maxy=100)
Total points read from Kaplan-Meier curve= 145
Points left after preprocess = 143
The indexes for each reported interval
  interval lower upper t.risk n.risk
1        1     1    42      0    213
2        2    43    66     10    122
3        3    67    88     20     80
4        4    89   109     30     51
5        5   110   129     40     30
6        6   130   143     50     10
```

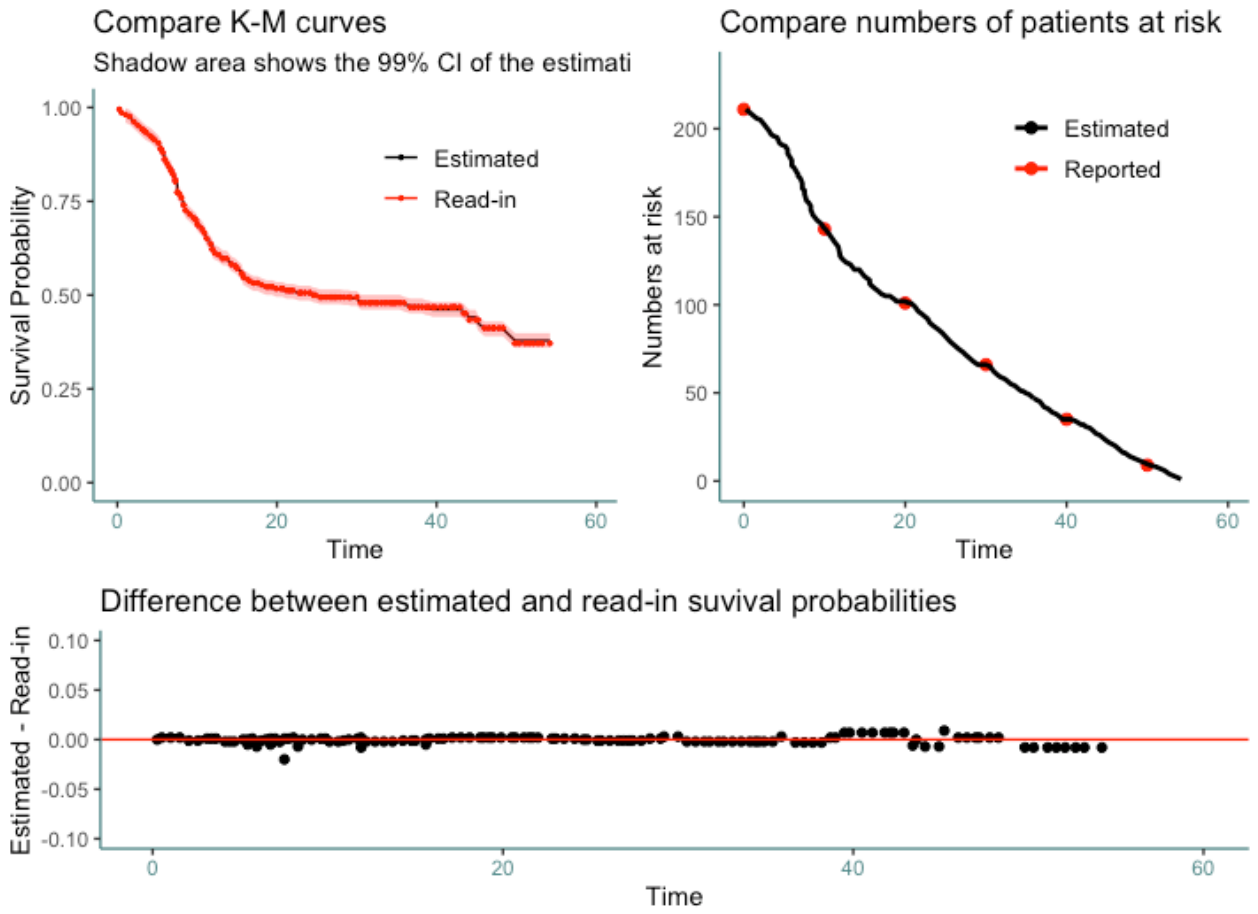
```
> pre_radio <- preprocess(dat=Radiationdata$radioplus,
+     trisk=Radiationdata$trisk,
+     nrisk=Radiationdata$nrisk.radioplus,maxy=100)
Total points read from Kaplan-Meier curve= 136
Points left after preprocess = 135
The indexes for each reported interval
  interval lower upper t.risk n.risk
1        1     1    36      0    211
2        2    37    63     10    143
3        3    64    86     20    101
4        4    87   109     30     66
5        5   110   127     40     35
6        6   128   135     50      9
```

Reconstruct IPD

```
> est_radio <- getIPD(prep=pre_radio,
+                     armID=0,tot.events=NULL)
> summary(est_radio)

The read in and estimated survival parameters:
  interval lower upper trisk nrisk nrisk.hat censor.hat
1        1     1    36     0   211       211          6
2        2    37    63    10   143       143          5
3        3    64    86    20   101       101         31
4        4    87   109    30    66        66         28
5        5   110   127    40    35        35         22
6        6   128   135    50     9         9          9
  event.hat
1        62
2        37
3         4
4         3
5         4
6         0

Evaluation of the survival probability estimation precision:
                  ResultSummary   value
1        Total number of patients 211.000
2          Root Mean Square Error   0.004
3             Mean Absolute Error   0.003
4              Max Absolute Error   0.009
5 Kolmogorov-Smirnov test P-value   0.549
> plot(est_radio)
```



Accuracy Assessment

Summary of the K-M Curve Estimation

Summary of Estimation	
Total number of patients	211
Read in points	135
Estimation Precision	
Root MSE	0.004
Mean absolute error	0.003
Max absolute error	0.009
Kolmogorov-Smirnov Test	
Test statisitics D	0.098
P value	0.549

RMSE, mean absolute error and max absolute error measure the differences between the estimated survival probabilities and the read-in values.The KS test compares the distributions of the estimated and read-in K-M curves.

- RMSE measures the difference in survival probabilities calculated using reconstructed data and original data.
- Mean absolute error and the Max absolute error assess the precision of the estimation.

Rule of thumb

RMSE ≤ 0.05

mean absolute error ≤ 0.02

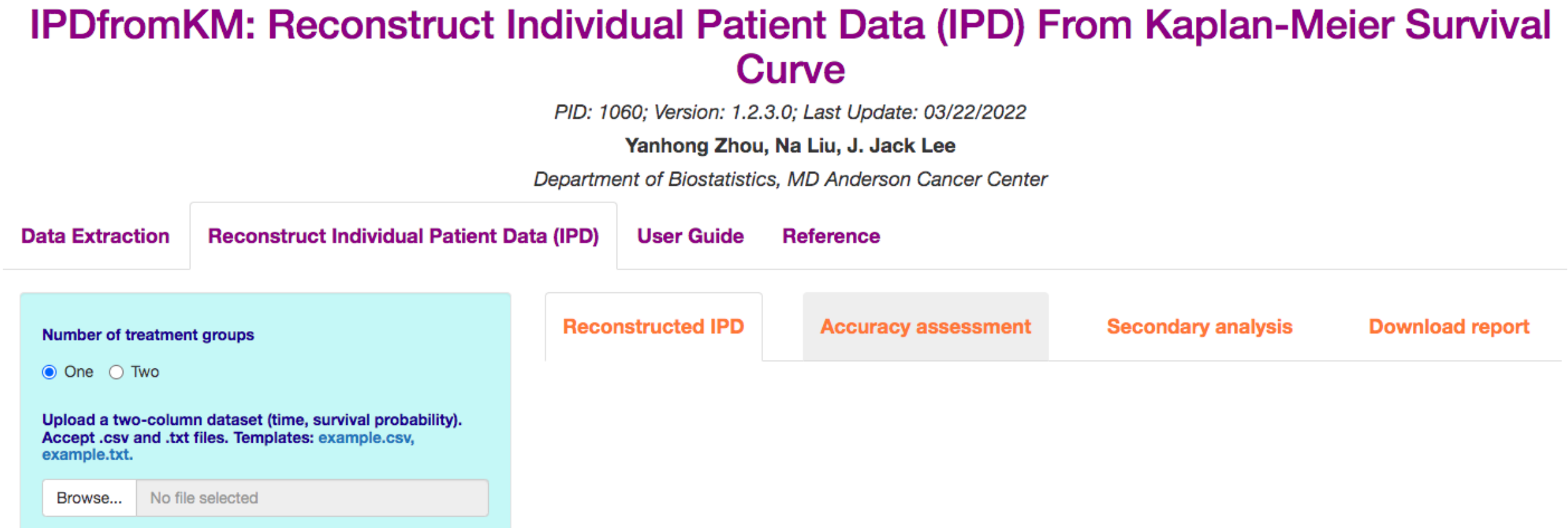
max absolute error ≤ 0.05

} Well reconstructed data

- Kolmogorov-Smirnov test is performed to compare the distributions of the read-in and the estimated survival curves. A large P-value is desired.

Shiny Application

- Main panels in the Shiny page: Data Extraction (used to extract data points), Reconstruct Individual Patient Data (IPD) (used to reconstruct IPD), and User Guide (providing video tutorials).
- https://trialdesign.org/one-page-shell.html#IPDfromKM
  https://biostatistics.mdanderson.org/shinyapps/IPDfromKM/



Shiny Application

All-in-one!

What do you need to input to the Shiny application?

1. Image file (.png or .jpeg) with KM curves or .csv or .txt files with coordinates data
2. Risktime(trisk)
3. Number of patients at risk(nrisk)
4. Total number of patients reported (optional when information for nrisk is provided).
5. Total number of events reported (optional, but having it will improve accuracy)

What can you do in the Shiny application:

1. Extract data points from published K-M curves
2. Reconstruct IPD using the modified-iKM algorithm
3. Assess the accuracy of the reconstruction
4. Perform survival analysis using the reconstructed IPD.
5. Download reconstructed IPD dataset
6. Generate a concise report for the IPD reconstruction and analysis

Comparing AD vs. IPD for Evidence Synthesis

- Aggregate Data
  - Advantage
    - Readily available in all reported trials
    - Easy to work with summary statistics, e.g., log(HR) and its SE
  - Disadvantage
    - Very limited on what can be done. Constrained by the different analysis method of each trial.
    - Possible heterogeneity in the inclusion analysis, summary, and reporting
      - Some trials adjusted by covariates and some not
      - Some use Cox model, others use RMST.

- Individual Patient Data
  - Advantage
    - Most granular raw data for each patient including treatments, outcomes, and covariates.
    - With raw data, can apply most appropriate method, e.g., Cox model or RMST.
    - Allow arm-based and contrast-based analysis for the network meta-analysis to compare the relative treatment effect across different treatments.
  - Disadvantage
    - Not readily available
    - If IPD is generated from KM, only survival time can be generated.
      - No matched covariates can be generated for the same patient.

http://prisma-statement.org/?AspxAutoDetectCookieSupport=1

https://www.ipdma.co.uk/

Meta-Analysis Methods

- Based on AD
  - Many standard meta-analysis methods available
- Based on IPD
  - One-stage analysis
    - Lump all data together and analyze them as if in one trial
    - Add trial specific considerations such as random effect for trials
  - Two-stage analysis
    - First stage: within trial analysis to generate summary data
    - Standard meta-analysis based on summary data
- Based on hybrid data
  - Some trials have IPD and some don't.
    - Not always possible to generate IPD from summary data, e.g., no KM curves reported.
  - Can apply two-stage analysis
  - Multilevel modeling of AD and IPD
    - Trial level and patient level
      - Frequentist
      - Bayesian hierarchical model

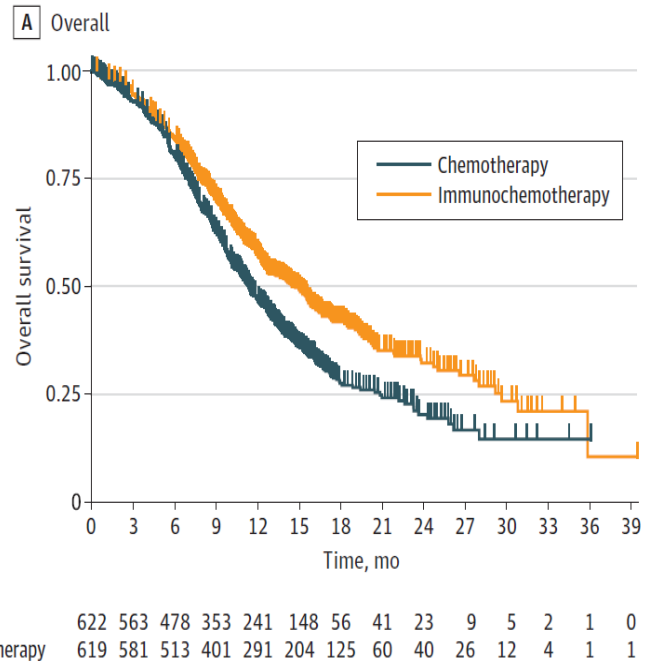
Effectiveness of Immune Checkpoint Inhibitors in Patients With Advanced Esophageal Squamous Cell Carcinoma A Meta-analysis Including Low PD-L1 Subgroups

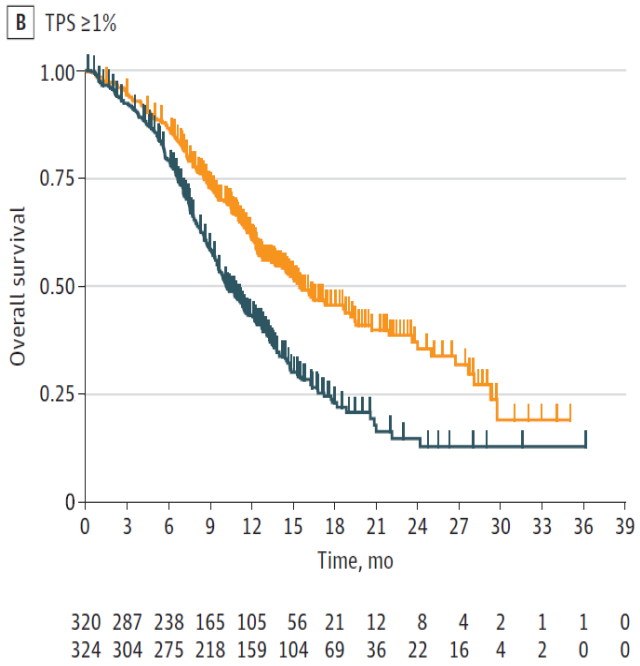
- Derive survival data for subgroups with low PD-L1 expression from trials comparing ICIs with chemotherapy in esophageal squamous cell carcinoma and perform a pooled analysis.
  - Apply IPDfromKM to extract IPF from KM curves of all-comer populations, subgroups with high PD-L1, and those with low PD-L1 (when available)
  - In studies with unreported curves for subgroups with low PD-L1 expression, KMSubtraction was used to impute survival data.
- 9 trials and 4752 patients
  - No significant benefit in OS for ICI vs chemo in patients who had a PD-L1 tumor proportion score (TPS) < 1% (HR, 0.91; 95%CI, 0.74-1.12; P = .38)
  - A significant but modest OS benefit for ICI vs chemo in the subgroup with a PD-L1 combined positive score (CPS) < 10 (HR, 0.77; 95%CI, 0.62-0.94; P = .01)

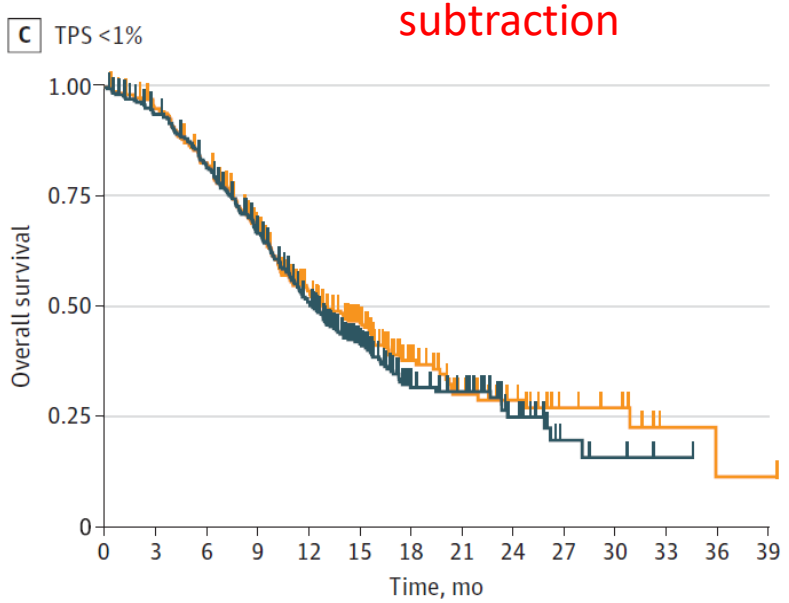
Yap et al. JAMA Oncol. 2023;9(2):215-224. Zhao, KMSubtraction reconstruction KM, BMC Med Res Methodology 2022

- KMSubtractionMatch() utilizing raw reconstructed time-to-event outcomes to match patients from subgroups among the overall cohort.
  - Minimal cost bipartite matching with the Hungarian algorithm was adopted aiming to match patients from the overall and subgroup cohort by minimizing follow-up time differences between matched patients.
  - Matching through Mahalanobis distance matching (Mahalanobis) and nearest neighbor matching by logistic regression (Logisitic) with the MatchIt package
  - Patients with events and censorships were matched separately.

Obtain by subtraction







No. at risk

Chemotherapy	622	563	478	353	241	148	56	41	23	9	5	2	1	0
Immunochemotherapy	619	581	513	401	291	204	125	60	40	26	12	4	1	1

Chemotherapy	320	287	238	165	105	56	21	12	8	4	2	1	1	0
Immunochemotherapy	324	304	275	218	159	104	69	36	22	16	4	2	0	0

No. at risk

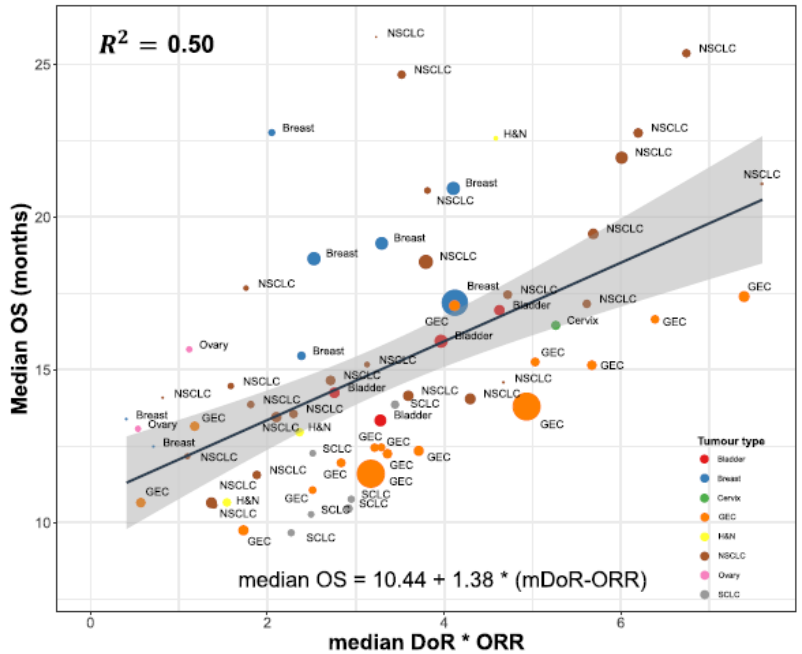
Chemotherapy	302	276	239	188	136	92	35	29	15	5	3	1	0	0
Immunochemotherapy	295	275	237	183	132	100	56	24	18	10	8	2	1	1

Comprehensive evaluation of surrogate endpoints to predict overall survival in trials with PD1/PD-L1 immune checkpoint inhibitors plus chemotherapy

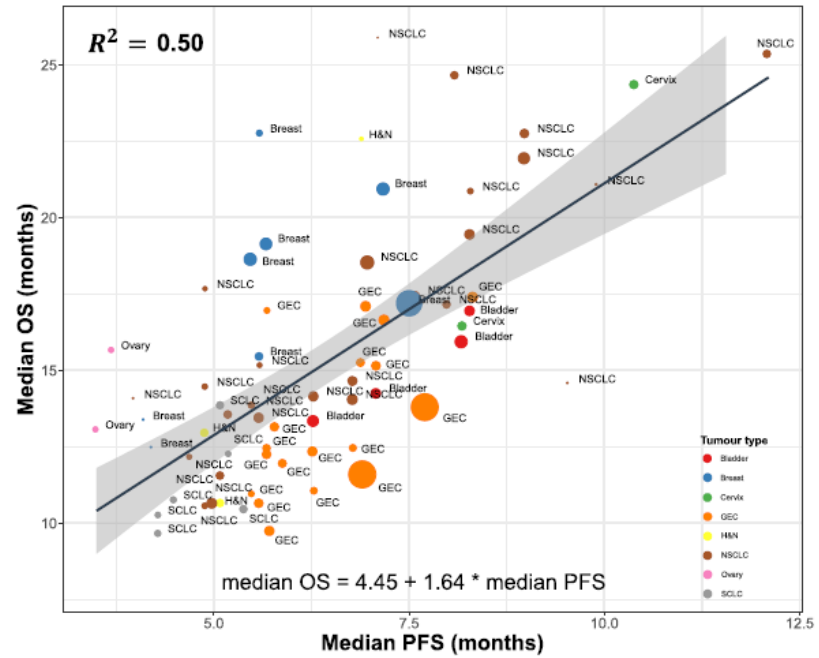
- Systematic review to identify RCT of anti-PD1/PD-L1 drugs plus CT versus CT alone. Performed
  - arm-level analysis to evaluate predictors of median OS (mOS)
  - comparison-level analysis for OS hazard ratio (HR) estimations. Linear regression models weighted by trial size
- 39 RCTs involving 22,341 patients (17 non-small cell lung, 9 gastroesophageal and 13 in other cancers)
  - Overall, ICI plus CT improved OS (HR = 0.76; 95%CI: 0.73–0.80).
  - In the arm-level analysis, the best mOS prediction was obtained with a new endpoint that combines median duration of response and ORR (mDoR-ORR) and with median PFS (R2 = 0.5 both).
  - In the comparison-level analysis, PFS HR showed a moderate association with OS HR (R2 = 0.52).
  - Early OS read-outs were highly associated with final OS outcomes (R2 = 0.80).
- Conclusions
  - The association between surrogate endpoints and OS is moderate-low.
  - Early OS read-outs showed a good association with final OS HR
  - mDORORR endpoint could help to better design confirmatory trials after single-arm phase II trials.

Villacampa et al. *Cancer Treatment Reviews 116 (2023) 102542*

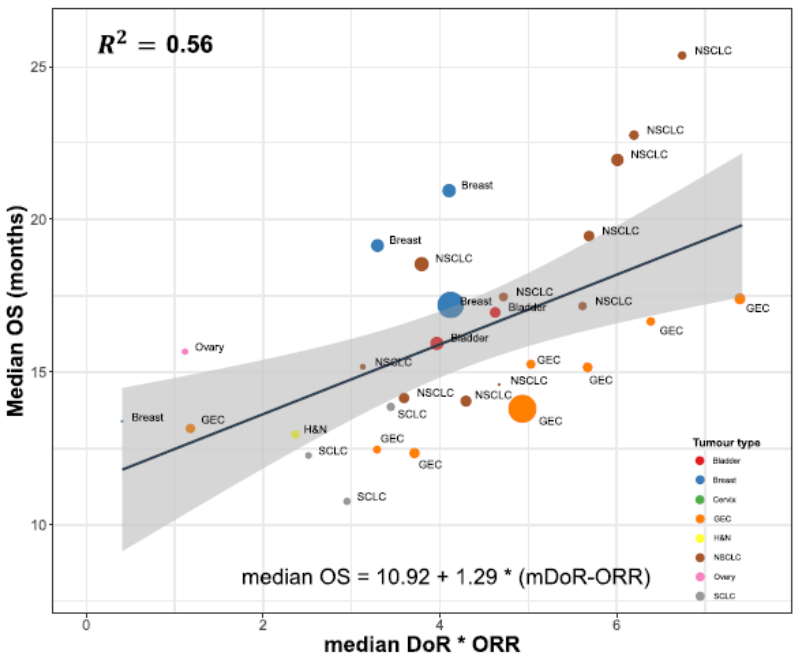
B (ITT population)



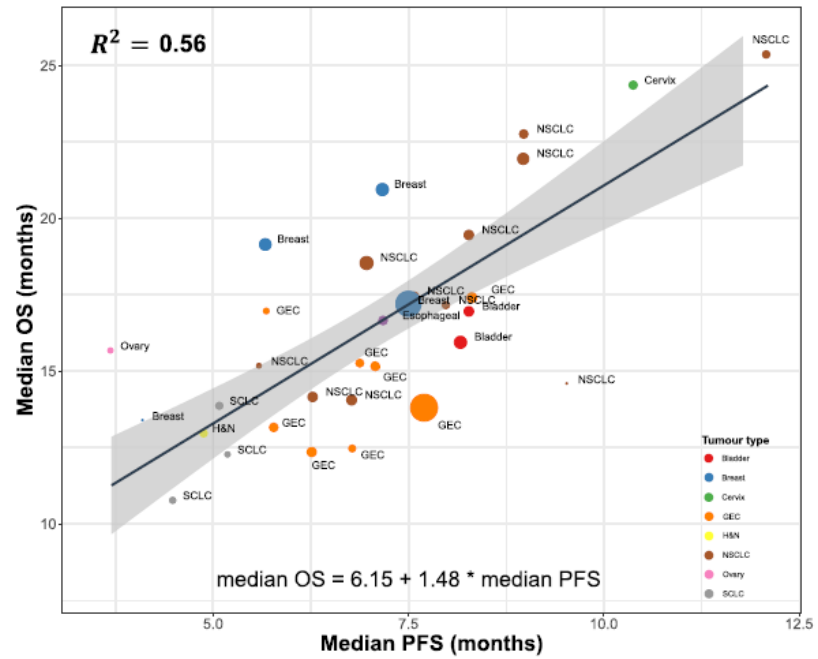
C (ITT population)



D (ICI + CT arm)



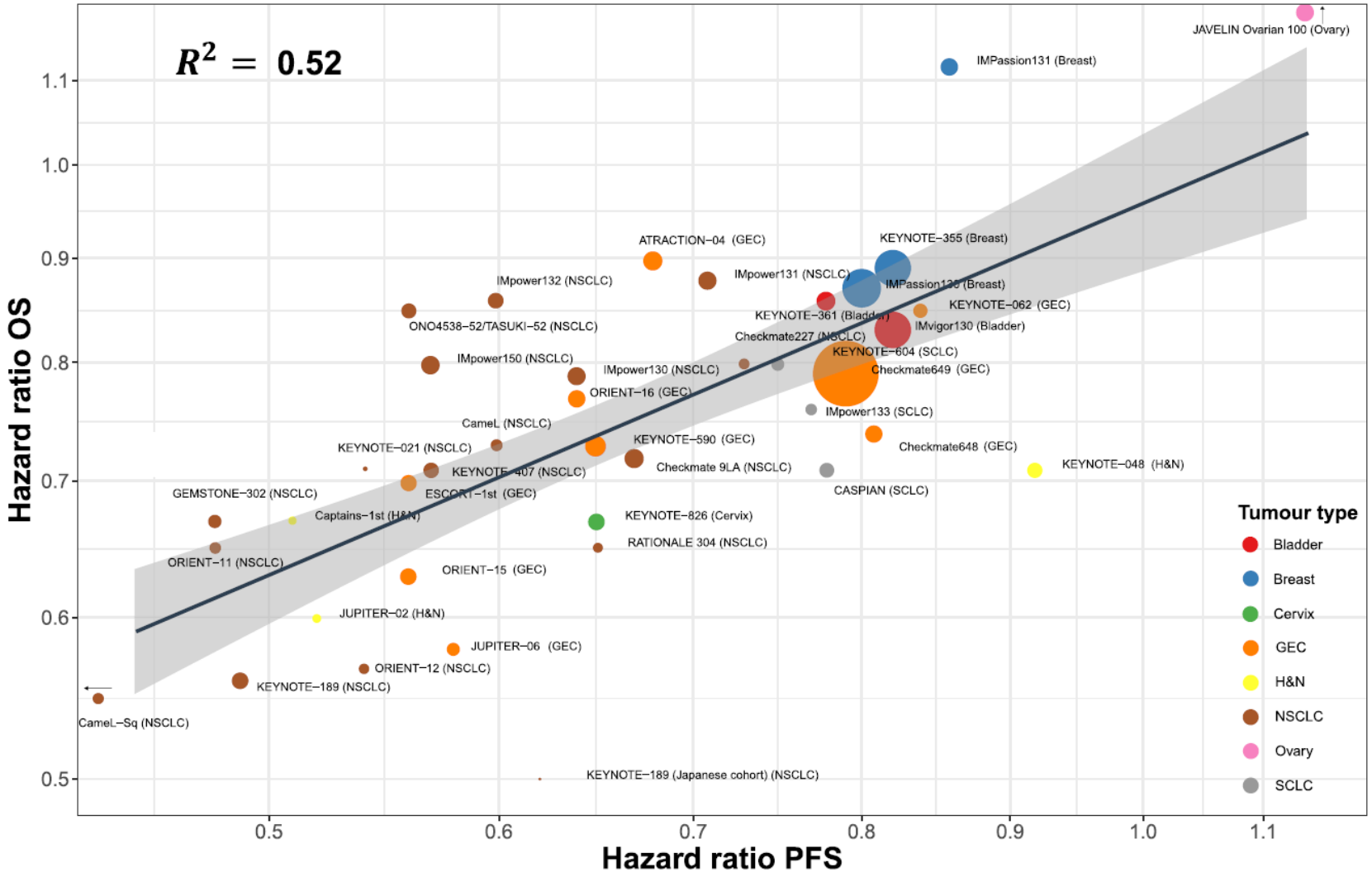
E (ICI + CT arm)

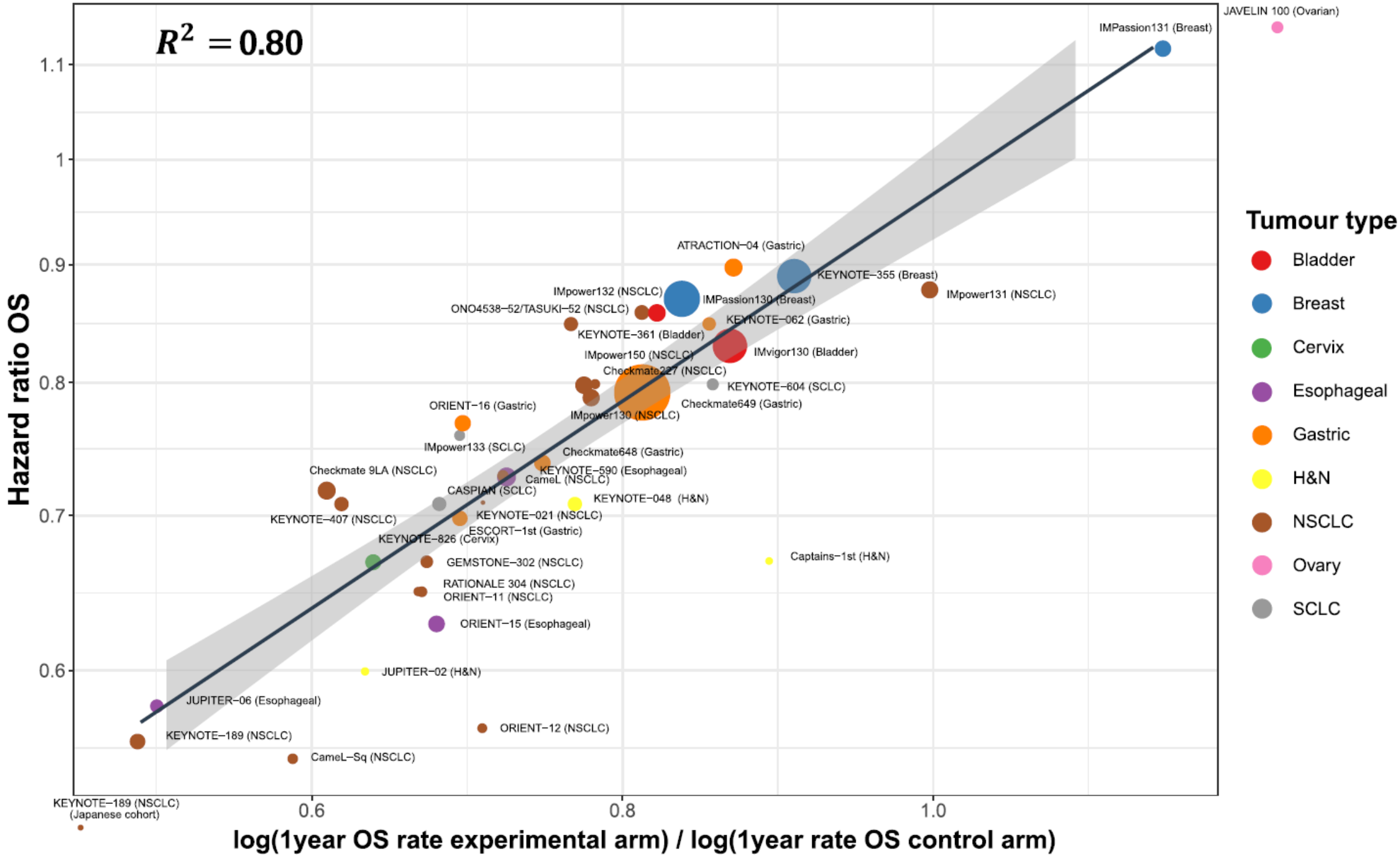


A

Comparison-level analysis (OS HR)	Overall population		By tumour type (ITT)		
	ITT population	PD-L1 population	NSCLC/ SCLC	Gastro-esophageal	Others
PFS hazard ratio	0.52	0.38	0.28	0.35	0.67
Ratio 1-year PFS rates	0.26	0.38	0.03	0.71	0.48
ORR risk ratio * mDoR risk ratio	0.23	0.15			
DCB risk ratio	0.23	0.37			
ORR risk ratio	0.17	0.19			
mDoR-ORR risk ratio	0.14	0.14			
mDoR risk ratio	0.11	0.01			
CR risk ratio	0.05	0.03			

B




$R^2 = 0.80$
Hazard ratio OS
log(1year OS rate experimental arm) / log(1year rate OS control arm)
1.1
1
0.9
0.8
0.7
0.6
0.6
0.8
1.0
JAVELIN 100 (Ovarian)
IMPassion131 (Breast)
ATRACTION−04 (Gastric)
KEYNOTE−355 (Breast)
IMpower131 (NSCLC)
IMpower132 (NSCLC)
IMPassion130 (Breast)
ONO4538−52/TASUKI−52 (NSCLC)
KEYNOTE−062 (Gastric)
KEYNOTE−361 (Bladder)
IMvigor130 (Bladder)
IMpower150 (NSCLC)
Checkmate227 (NSCLC)
KEYNOTE−604 (SCLC)
Checkmate649 (Gastric)
ORIENT−16 (Gastric)
IMpower130 (NSCLC)
IMpower133 (SCLC)
Checkmate648 (Gastric)
KEYNOTE−590 (Esophageal)
Checkmate 9LA (NSCLC)
CameL (NSCLC)
CASPIAN (SCLC)
KEYNOTE−048 (H&N)
KEYNOTE−407 (NSCLC)
KEYNOTE−021 (NSCLC)
ESCORT−1st (Gastric)
KEYNOTE−826 (Cervix)
Captains−1st (H&N)
GEMSTONE−302 (NSCLC)
RATIONALE 304 (NSCLC)
ORIENT−11 (NSCLC)
ORIENT−15 (Esophageal)
JUPITER−02 (H&N)
JUPITER−06 (Esophageal)
ORIENT−12 (NSCLC)
KEYNOTE−189 (NSCLC)
CameL−Sq (NSCLC)
KEYNOTE−189 (NSCLC)
(Japanese cohort)
Tumour type
Bladder
Breast
Cervix
Esophageal
Gastric
H&N
NSCLC
Ovary
SCLC

Combining AD and IPD (1)

- NMA combines direct and indirect comparisons to assess the effect of multiple treatments.
  - The inclusion of indirect comparisons may introduce confounding, leading to heterogeneity and inconsistency in the analysis
  - Heterogeneity and inconsistency are especially problematic when the distribution of effect modifier varies across studies
- Identifying the sources of heterogeneity and inconsistency in NMA solely based on aggregate data (AD) is often challenging.
- The use of individual participant data (IPD) is widely considered beneficial in evidence synthesis, as it improves data quantity, quality, and standardization of outcomes.
  - IPD allows for more flexibility in investigating effect modifiers, making it valuable for addressing inconsistencies in NMA.

Debray et al. An overview of methods for network meta-analysis using individual participant data: when do benefits arise?, SMMR 2018

Combining AD and IPD (2)

- Meta-analyses may yield biased estimates due to
  - Imbalances in patient characteristics, known as effect modifiers, which can impact the relative effectiveness of compared interventions
- Meta-regression has long been utilized to address such biases and enhance precision in meta-analyses.
- Meta-regression typically consists of conducting linear regression of the study results as a function of an effect modifier, both in the aggregate.
  - Two potential limitations
    - A limited number of data points to reliably estimate trends
    - The risk of ecological fallacy
- Individual patient data (IPD) can help overcome the limitations of aggregate data (AD) meta-regression mentioned above.
- There are different approaches to employing IPD and AD in conducting meta-regression
  - Two-stage approaches where adjusted AD are created using IPD
  - One-stage approaches that integrate IPD and AD together.

Kanters et al. Comparing the use of aggregate data and various methods of integrating individual patient data to network meta-analysis and its application to first-line ART, BMC Medical Research Methodology 2021

Combining AD and IPD (3)

- Methods are needed to combine randomized and nonrandomized evidence, addressing issues like confounding and lack of comparator arm in single-arm trials (SATs).
- Extensions to meta-analysis assume exchangeability of baseline response parameters across trials, allowing for synthesis of SATs and RCTs in a NMA context
  - Using a mixture of IPD and AD and a contrast-based approach to NMA
  - Arm-based NMA methods parametrize absolute treatment effects across arms, incorporating SATs into a synthesis of RCTs in pairwise meta-analysis and NMA.
- Under a Bayesian framework, these methods use SATs data to provide information on the pooled treatment effect and between-study heterogeneity parameters via
  - A prior distribution
  - Can be down-weighted according to commensurability with RCT data
- Pros and Cons:
  - Pros: Valuable when SATs are the only evidence available.
  - Cons: Compromising randomized treatment allocation and introducing bias in RCT treatment effect estimates,
- Alternative methods match similar trial arms based on characteristics, directly entering matched pairs into NMA or estimating relative treatment effects from SATs using matched trial data.

Singh et al. Bayesian network meta-analysis methods for combining individual participant data and aggregate data from single arm trials and randomised controlled trials, BMC Medical Research Methodology 2022

Concluding Remarks

- In addition to AD, IPD can be very useful in synthesizing evidence across multiple trials.
- Network meta-analysis combining AD and IPD is a powerful approach for evidence synthesis. However, many challenges exist to obtain efficient and unbiased estimates.
- More methods for combining AD and IPD need to be developed for time-to-event data.
  - Some trials have AD only and some other trials have IPD
  - Same trial with mixture of AD and IPD

Bayesian Methods in Meta-analysis of Survival Data

Shouhao Zhou

Division of Biostatistics and Bioinformatics
Pennsylvania State University

A joint work with Chenqi Fu, Don Berry, Xuelin Huang,
Nicholas Short, Farhad Ravandi, Lydia Hanna, Daniel Barker

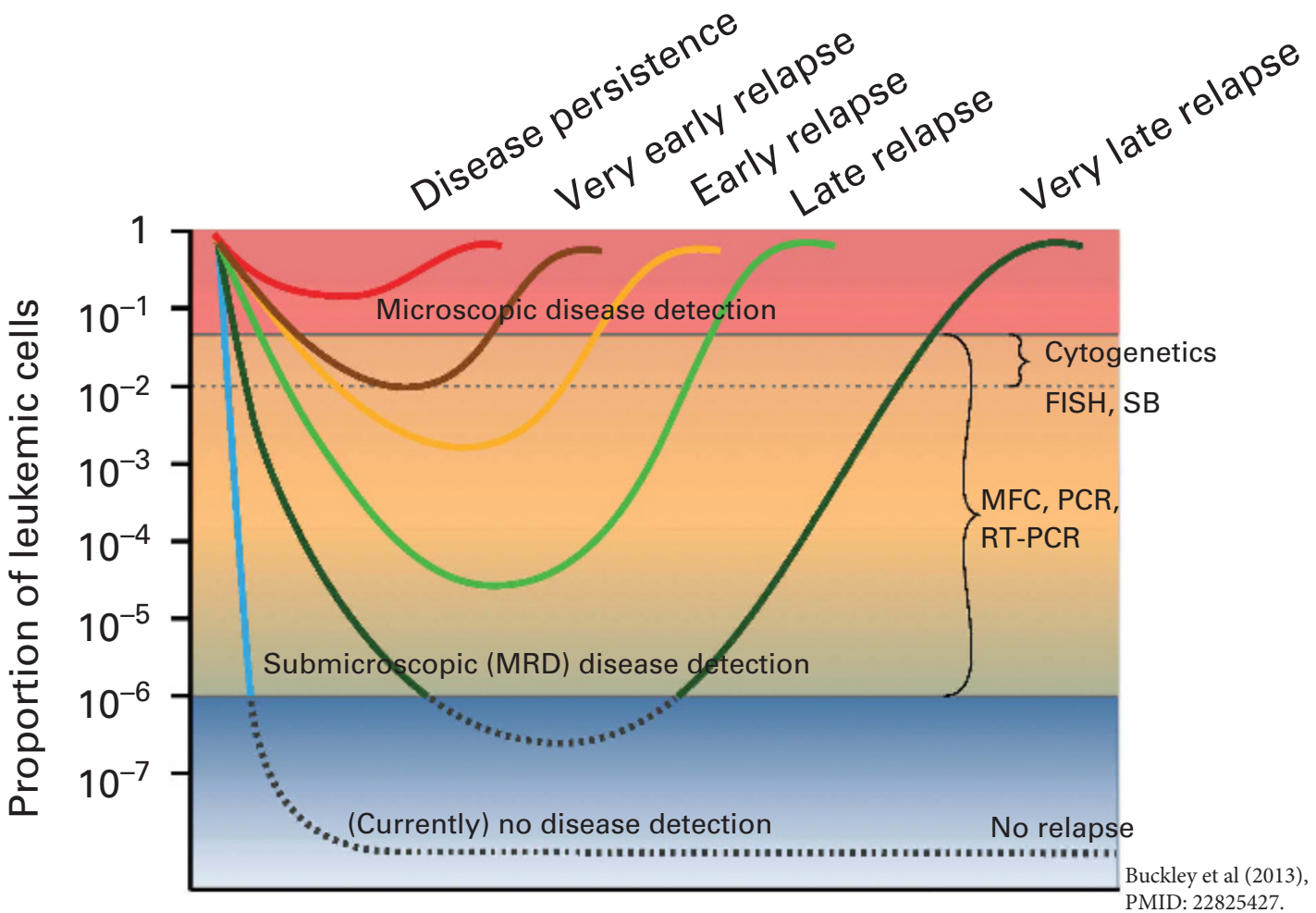
SCT annual meeting
Baltimore, MD
May 24, 2023

No relevant disclosures

Table of contents

Meta-analysis of Reconstructed Survival Data (MARS)

What is measurable residual disease?



Buckley et al (2013), PMID: 22825427.

Hypothesis:

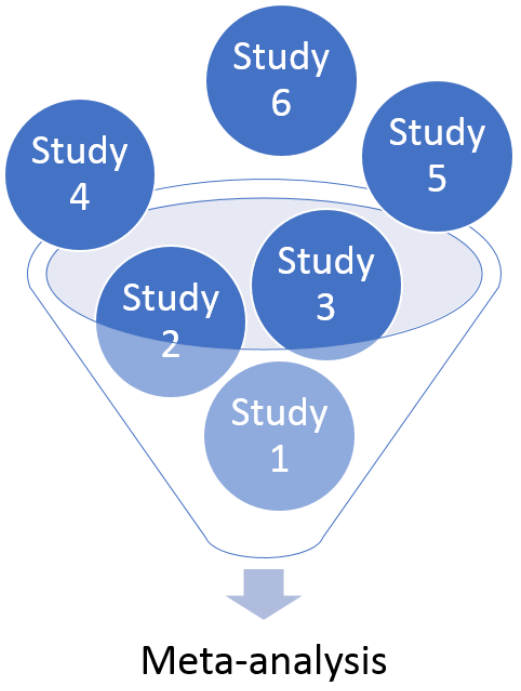
Residual tumor cells (tumor burden) $\longrightarrow$ Tumor recurrence

Role of MRD in treatment of hematologic malignancies

- FDA guideline in drug development (FDA 2020)
  - prognostic surrogate marker for drug development
  - endpoint in clinical trials
  - risk classification and patient selection
- Call for meta-analysis with special considerations
  - comprehensive literature search of both positive and negative results, to limit selection bias
  - efficient and robust statistical methods, to demonstrate the effect of MRD
- Application: MRD in AML (Short et al. 2020)
  - DFS: 64 studies (16 + 36 + 12) with 8,033 patients

Systematic review and Meta-analysis

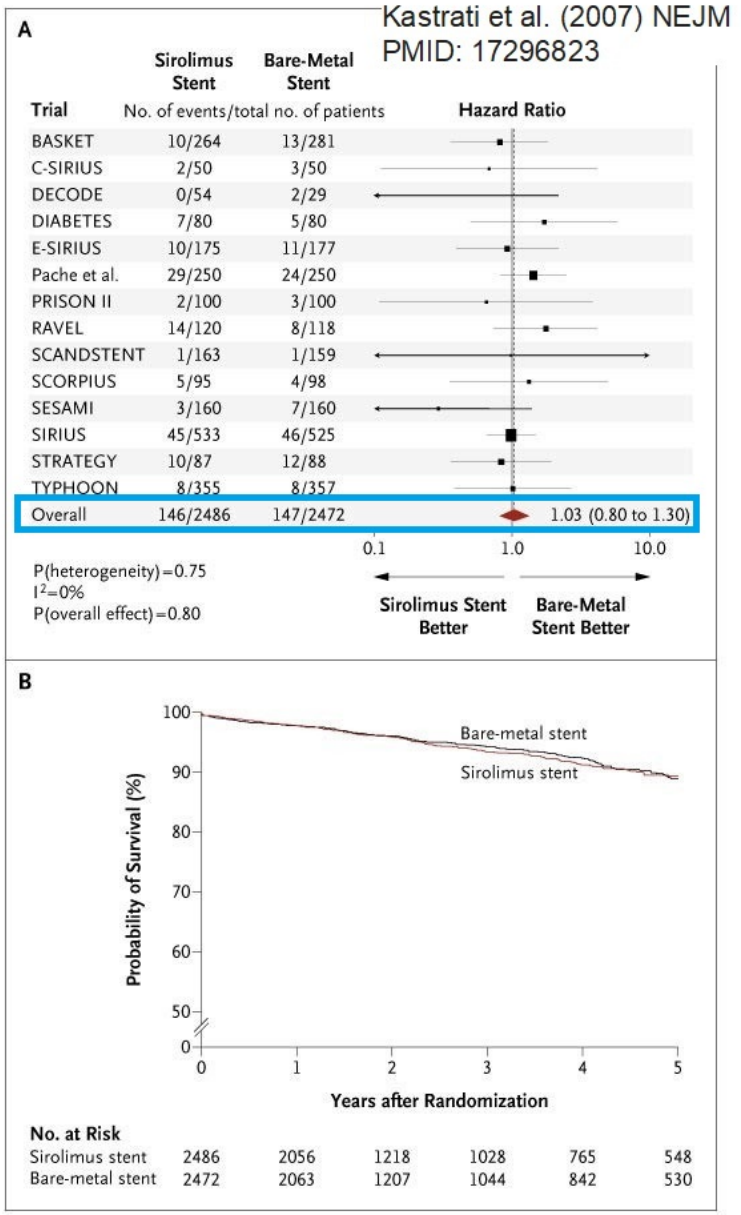
- Systematic Review
  - summarize the scientific evidence related to treatment, causation, diagnosis, or prognosis of a specific disease.
  - review and summarize already-existing studies: ~~new data~~.
- Meta-analysis
  - a powerful tool to integrate information
  - large amount of data from multiple studies/sources
  - increased precision to address research question



goal: valid statistical inference of population characteristics

Meta-analysis Reporting

- Standard approach in meta-analysis for survival outcome uses aggregate study-level hazard ratio (HR) estimates
  - DerSimonian and Laird method (random-effects model) for overall effect
- Nonstratified Kaplan–Meier curves by intervention group



Existing methods

- Individual patient data (IPD) meta-analysis
  - gold standard
  - homogeneous data analysis and model assumption check across studies: one-stage vs two-stage approaches
  - difficult to obtain data
- aggregate data (AD) meta-analysis
  - publicly available data
  - assuming proportional hazards
  - computationally efficient: fixed-effect model vs random-effects model
  - limited information extraction $\longrightarrow$ limited conclusion

Motivating Example

- 64 studies with 8,033 patients
- Only 44 reported HRs
  - Others reported KM plots or survival rates at specific time points
- Proportional hazards (PH) assumption was violated in some studies

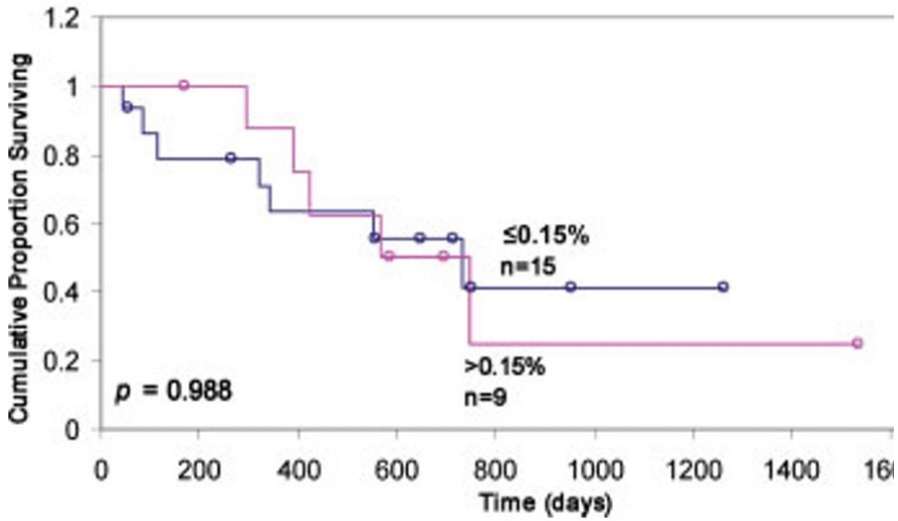
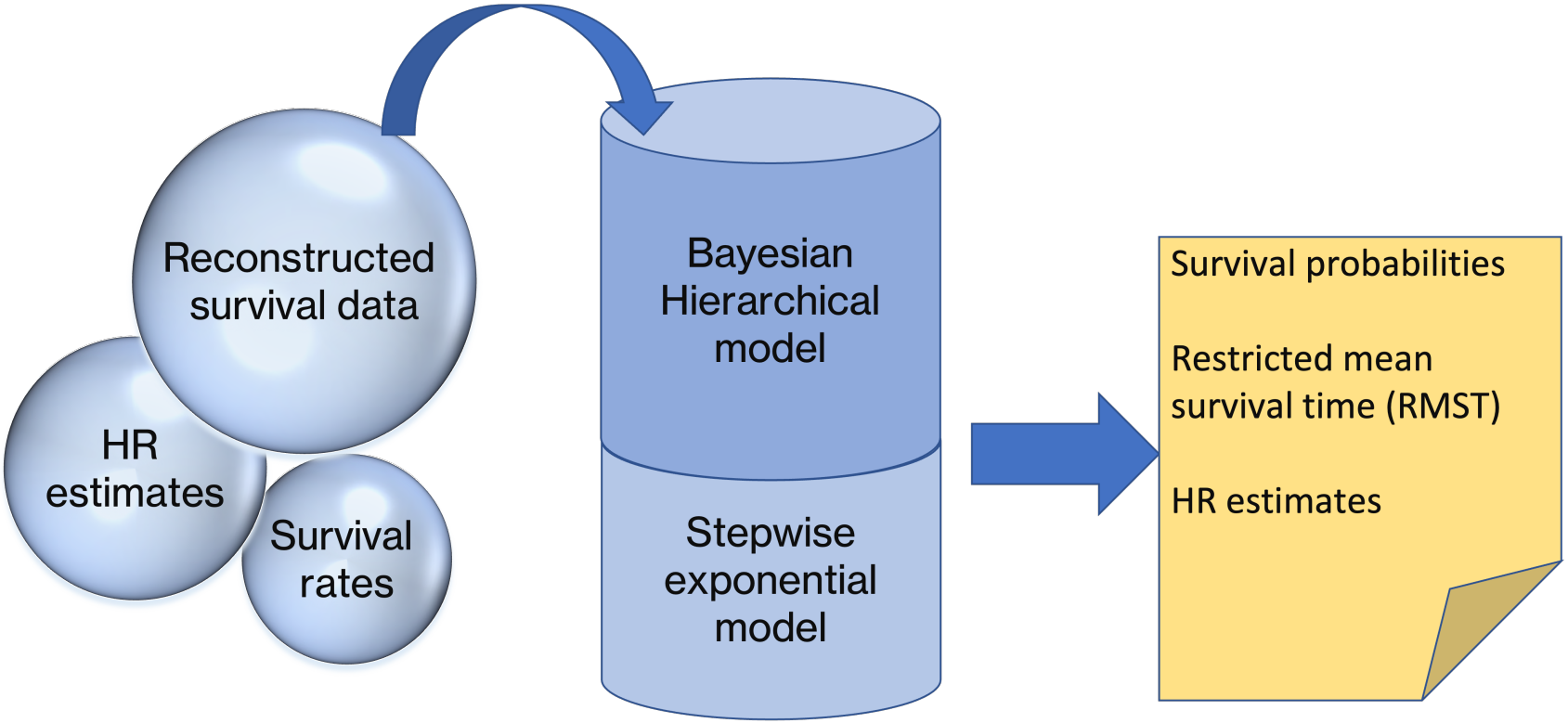


Figure: Detection of MRD using five-color MFC in AML identifies patients with high risk of relapse.

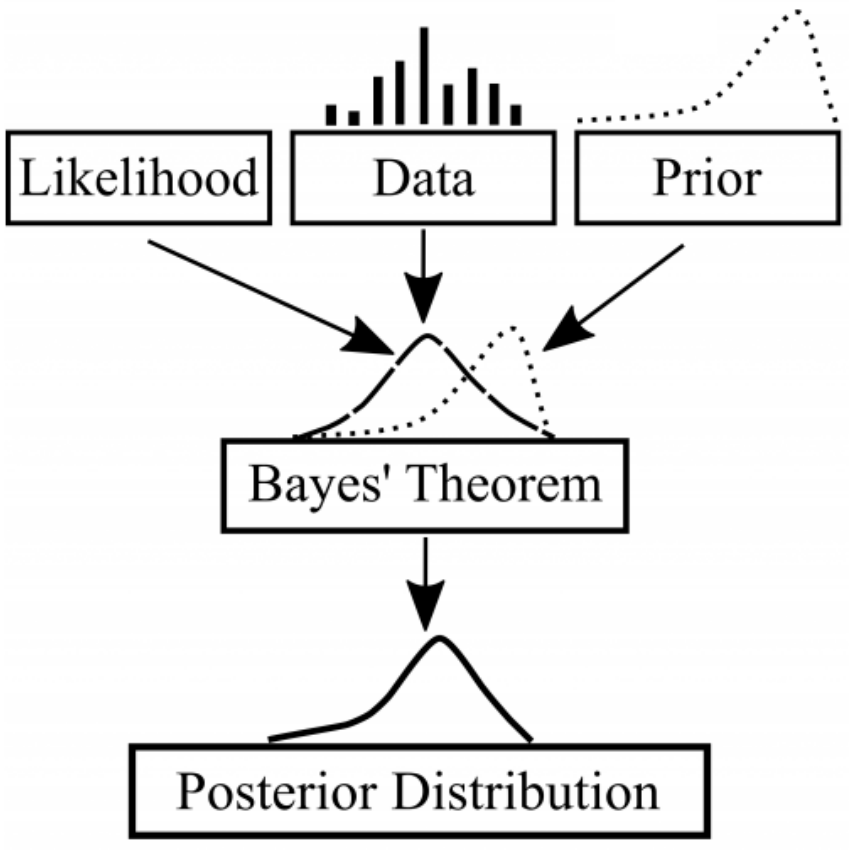
Source: Al-Mawali et al. PMID: 18727068.

MARS

- Meta-Analysis of Reconstructed Survival data (MARS)
  - Cover different data types
  - Flexible modeling framework: relax PH constraints
  - Enrich meta-analysis reporting



Bayesian Approach



Source: Rathi 2018

- Incorporate between study heterogeneity
- Integrate more studies to reduce selection bias and extractor bias
- Relax PH assumption
- Deliver flexible and effective estimation

Data type I: Reconstructed survival data

- Survival data at AD level reconstructed from the Kaplan-Meier curves (2 steps)
  1. acquire the positions of censoring and event points on images of KM survival curves
  2. perform iterative numerical computation to convert 2-dimensional position data into time-to-censoring and time-to-events (Guyot et al. 2012)

- Advantages
  - information available on baseline hazard function and time-varying hazard ratio
  - alternative to IPD in modeling and reporting

- **IPDfromKM** (Liu, Y. Zhou, and Lee 2021)

survtime	censor	group
5	1	2
28	0	1
39	0	1
7	1	2
9	1	2
45	0	1
3	0	2
34	0	1
45	0	1
8	0	2

Data type I: Reconstructed survival data

- Step-wise exponential function for time trend
  - Split the time scale into J disjoint segments with each interval $I_j = (t_{j-1}, t_j]$ for $j = 1, 2, \ldots, J$, where $0 = t_0 < t_1 < \cdots < t_J$.
  - Within each interval, the baseline hazard rate and the piecewise log HR are assumed to be constant, denoted by λ_j and β_j for $j = 1, 2, \ldots, J$.

- The hazard rate for the i^{th} subject in the k^{th} study is modeled as

$$h_{k,i}(t) = h_k(t, X_{k,i}) = \lambda_j \exp\{\alpha_k + \theta_k(t) X_{k,i}\}, \quad t \in I_j, \tag{1}$$

The study-specific log HR $\theta_k(t)$ is decomposed into the time-dependent effect β_j and an inter-study effect μ_k

$$\theta_k(t) = \beta_j + \mu_k, \quad t \in I_j. \tag{2}$$

Bayesian Hierarchical Model

- The Poisson presentation of survival data (Johansen 1983)

- The likelihood for the reconstructed survival data is given by

$$L_1 = \prod_{k=1}^{K_1} \prod_{i=1}^{n_k} h_{k,i}(t_{k,i})^{d_{k,i}} \exp\left(-\int_0^{t_{k,i}} h_{k,i}(u)du\right)$$

  where $d_{k,i} = 1$ for an event and $d_{k,i} = 0$ for censoring.

Data Type II: Hazard ratio estimates

- The observed constant log HR $\hat{\theta}_k$ for k^{th} study as an estiamte for weighted average of time-dependent effects β_j's in addition to a study-specific effect μ_k,

$$L_2 = \prod_{k=K_1+1}^{K_1+K_2} \mathcal{N}(\hat{\theta}_k \mid \bar{\theta}_k, \hat{se}(\hat{\theta}_k)^2)$$

with the HR aligned with study duration

$$\begin{aligned} \bar{\theta}_k &= \frac{1}{T_k}\int_0^{T_k} \theta_k(u)du \\ &= \mu_k + \frac{1}{T_k}\{\sum_{l=1}^{j-1} \beta_l |I_l| + \beta_j(T_k - t_{j-1})\}, \quad T_k \in I_j, \end{aligned}$$

where $\theta_k(t)$ is defined in equation (2), $|I_j| = t_j - t_{j-1}$, and T_k is the duration of the kth study.

Data Type III: Survival rate estimates

- The reported survival rates at time t for group $X = x$ are treated as the realizations of the survival function derived from equation (1)

$$L_3 = \prod_{k=K_1+K_2+1}^{K_1+K_2+K_3} \prod_{x=0}^{1} \mathcal{N}(\text{logit}(\hat{S}_k(t,x)) \mid \text{logit}(S_k(t,x)), \frac{\hat{se}^2(\hat{S}_k(t,x))}{\hat{S}_k(t,x)^2(1-\hat{S}_k(t,x))^2})$$

where

$$\begin{aligned} S_k(t, X = x) &= \exp\left\{-\int_0^t h_k(u,x)du\right\} \\ &= \exp\{-\sum_{l=1}^{j-1} \lambda_l \exp\{\alpha_k + (\mu_k + \beta_l)x\}|I_l| \\ &\quad -\lambda_j \exp\{\alpha_k + (\mu_k + \beta_j)x\}(t - t_{j-1})\}, \quad t \in I_j. \end{aligned}$$

Reporting measures

- Survival probabilities for treatment group $X = x$: for $t \in I_j$,

$$S(t, X = x) = \exp\{-\sum_{l=1}^{j-1} \lambda_l \exp\{\beta_l x\}|I_l| - \lambda_j \exp\{\beta_j x\}(t - t_{j-1})\}.$$

- Restricted mean survival time (RMST): the expected value of survival time up to time τ (Royston and Parmar 2013)

$$RMST(\tau, X = x) = E(\min(T, \tau)) = \int_0^\tau S(u, x)du.$$

Consistency

Theorem 1 (Consistency)

Let $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)^T$ and $\log \boldsymbol{\lambda} = (\log \lambda_1, \log \lambda_2, \ldots, \log \lambda_J)^T$.
Denote $\hat{\boldsymbol{\beta}}$ and $\log \hat{\boldsymbol{\lambda}}$ as MAP estimates generated by MARS.
Under some regularity conditions and identifiability conditions,
$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \log \hat{\boldsymbol{\lambda}} \end{pmatrix}$ converges to $\begin{pmatrix} \boldsymbol{\beta} \\ \log \boldsymbol{\lambda} \end{pmatrix}$ in probability as $K \to \infty, n \to \infty$.

Theorem 1 states that MARS yields consistent survival probability estimation.

Efficiency

Theorem 2 (Relative efficiency)

Let $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\beta}}^*_{IPD}$ be the estimates generated from MARS and IPD meta-analysis, respectively.
Under some regularity conditions, we have the following

1. for any fixed K and n, $Var(\hat{\boldsymbol{\beta}}^*) \geq Var(\hat{\boldsymbol{\beta}}^*_{IPD})$;
2. if $Var(\hat{\boldsymbol{\beta}}_k)$ coincides with the variance of MLE of $\boldsymbol{\beta}_k$ for $k \in \mathcal{K}_2$, as $K, n \to \infty$ and $Kn^{-1/2} \to 0$, the asymptotic variances satisfy

$$\lim_{\substack{n\to\infty, K\to\infty \\ Kn^{-1/2}\to 0}} Var(\hat{\boldsymbol{\beta}}^*) = \lim_{\substack{n\to\infty, K\to\infty \\ Kn^{-1/2}\to 0}} Var(\hat{\boldsymbol{\beta}}^*_{IPD}).$$

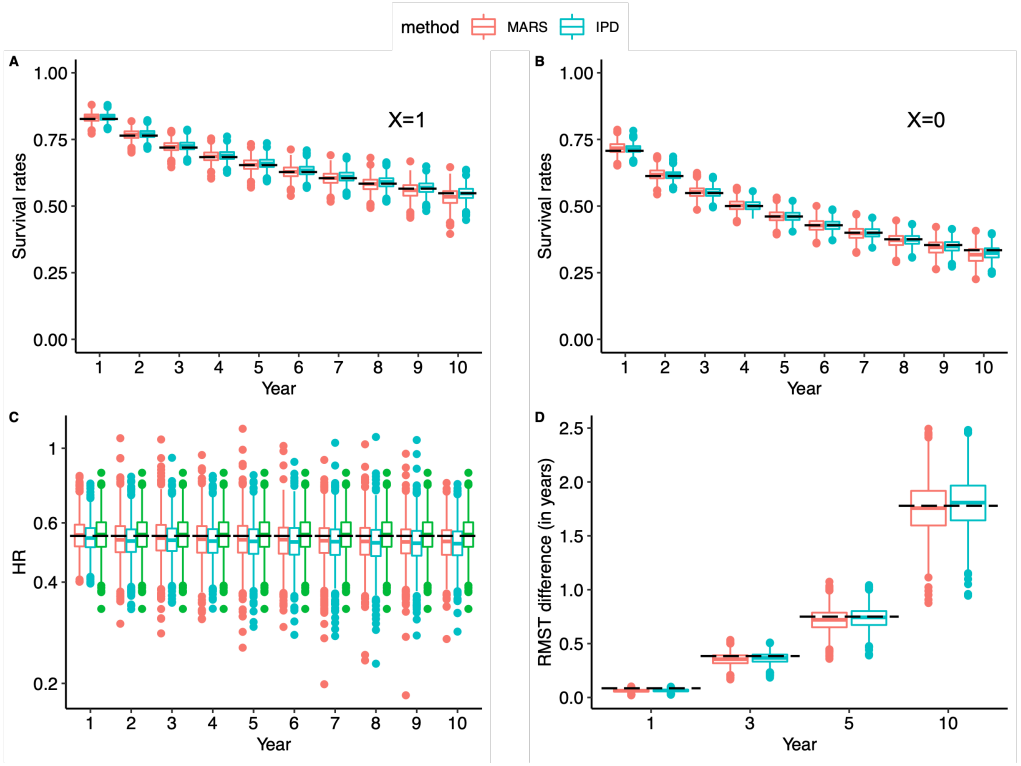
Theorem 2 lists some conditions that MARS can achieve asymptotic full efficiency with type I and II data.

Simulation setup I

- Case 1: proportional hazards
  - IPD meta-analysis: 13 studies raw data
  - Standard AD meta-analysis: 6 studies of HR estimates
  - MARS: 6 studies of reconstructed survival data, 3 studies of HR estimates and 4 studies of survival rates
- Survival time is assumed to be right-censored with an exponential distribution
- The sample size of each study is from 50 to 150
- The follow-up time is from 80 to 120 months and we partitioned time by year
- Results are based on 1000 replications

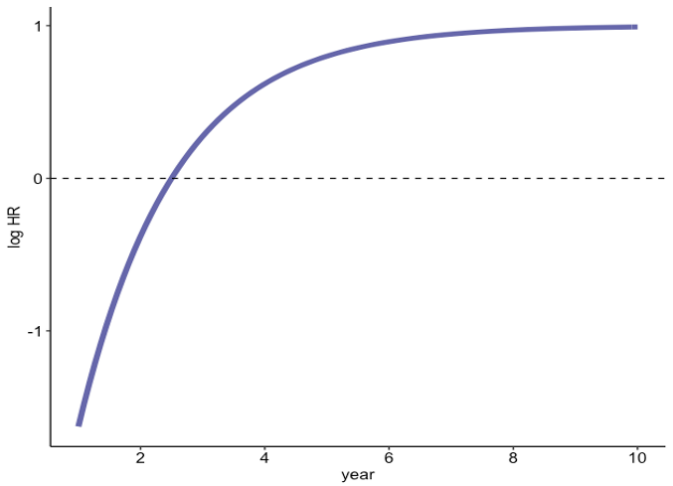
Simulation result I

- MARS provided similar point estimates of survival probabilities to IPD meta-analysis (panel A, B)
- The piecewise HRs estimated by MARS were approximately constant (panel C)



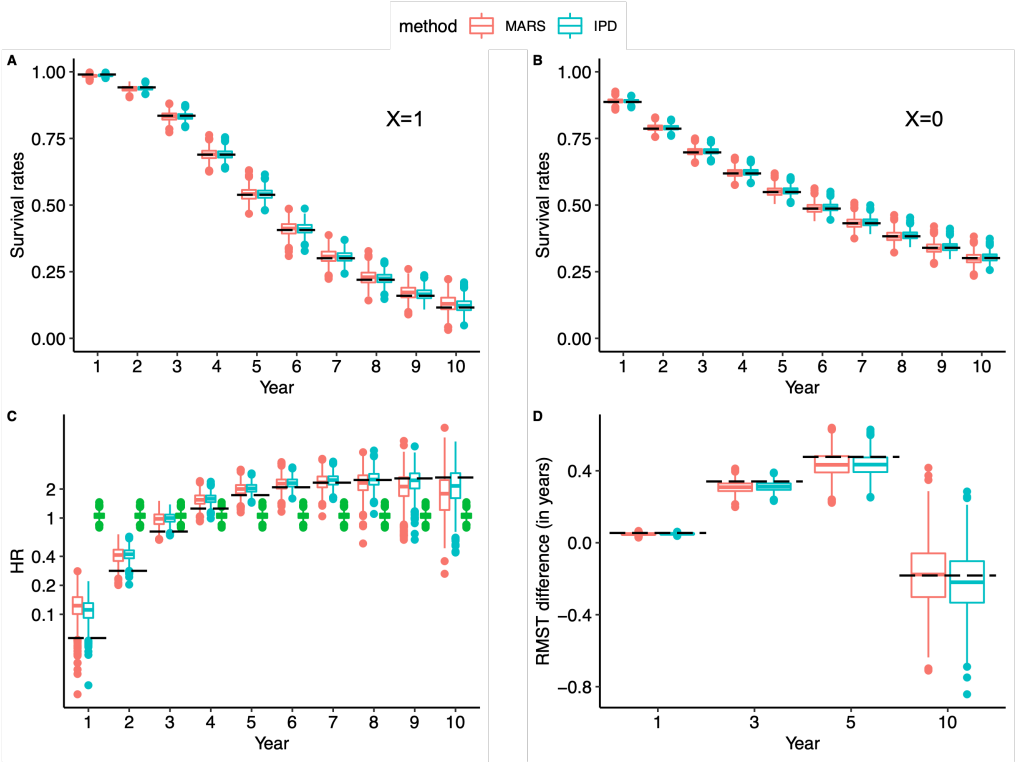
Simulation setup II

- Case 2: time-varying hazard ratios
  - IPD meta-analysis: 20 studies raw data
  - Standard AD meta-analysis: 12 studies of HR estimates
  - MARS: 9 studies of reconstructed survival data, 7 studies of HR estimates and 4 studies of survival rates



Simulation result II

- MARS provided comparable estimates of survival rates to IPD meta-analysis (panel A, B)
- The piecewise HRs had slightly inferior performance to estimate the piecewise HRs (panel C)



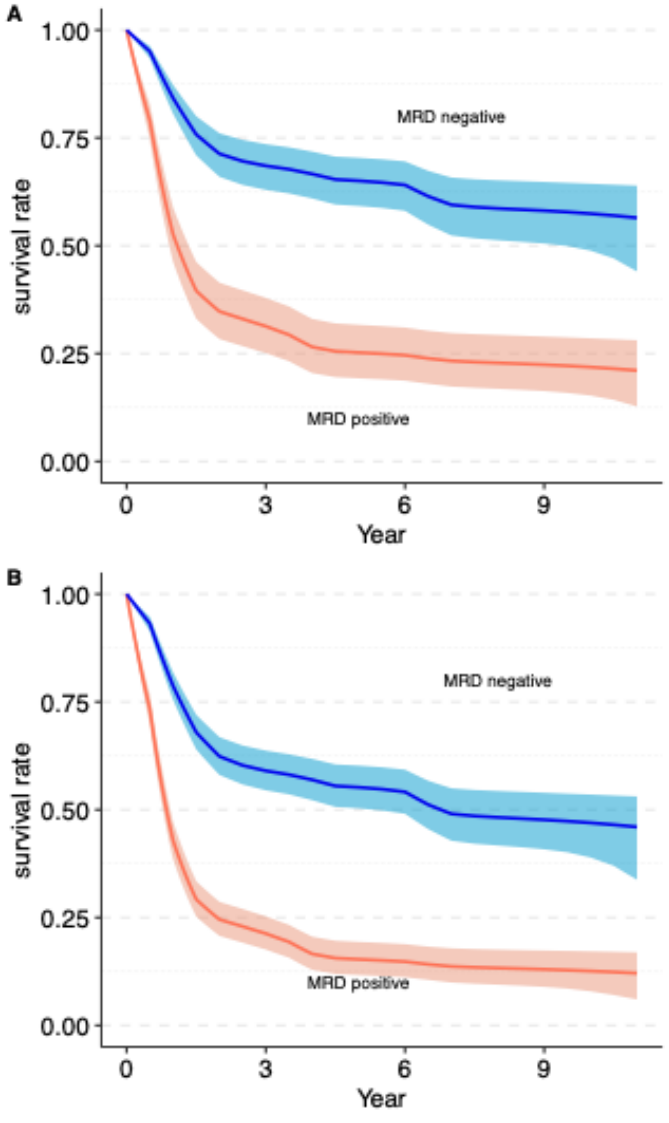
Application - MRD in AML

- Endpoint: disease-free survival (DFS)
- 64 studies with 8,033 AML patients
  16 studies of reconstructed survival data, 36 studies of HR estimates, and 12 studies of survival probabilities
- Use time segments of 0-6 months and every 6 months of follow-up thereafter to a maximum of 11 years for a total of 22 intervals

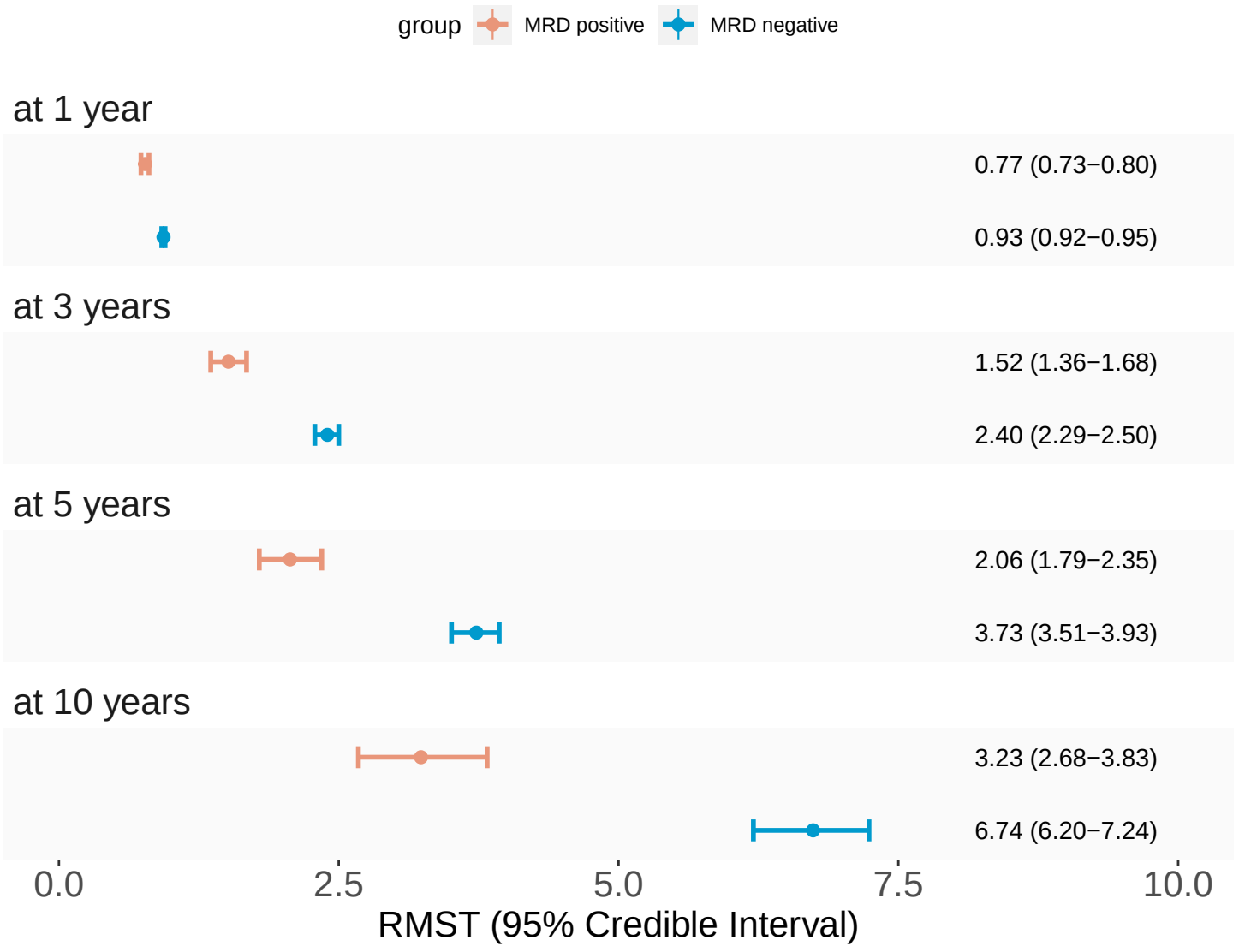
Survival probabilities

- The MRD negativity is associated with superior DFS (A)
  - Estimated DFS at 5 years
    MRD-negative:
    65% (95% CrI, 59%-70%)
    MRD-positive:
    25% (95% CrI, 19%-32%)

- Selection bias: using only reconstructed survival data (B)
  - underestimated survival probabilities
  - Narrower credible intervals



RMST

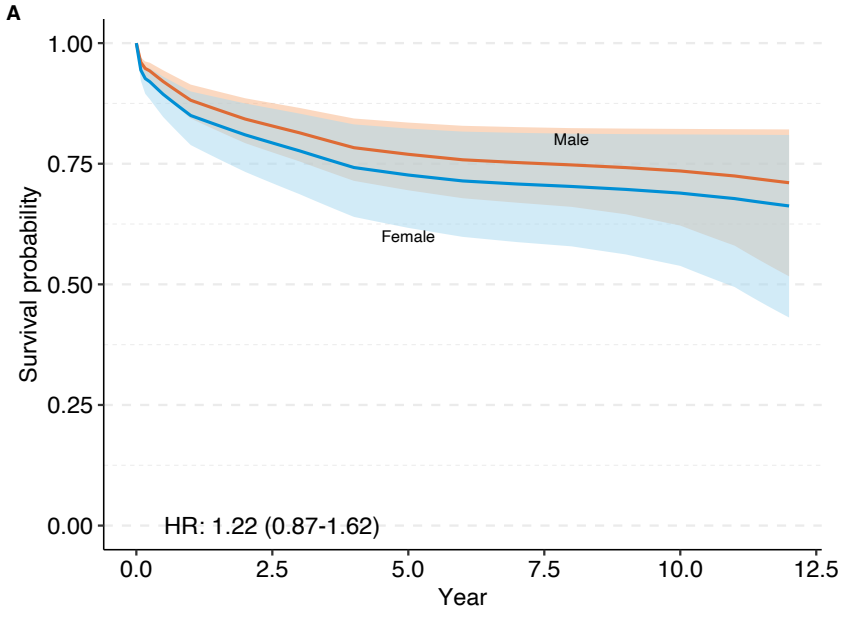

group
MRD positive
MRD negative
at 1 year
0.77 (0.73−0.80)
0.93 (0.92−0.95)
at 3 years
1.52 (1.36−1.68)
2.40 (2.29−2.50)
at 5 years
2.06 (1.79−2.35)
3.73 (3.51−3.93)
at 10 years
3.23 (2.68−3.83)
6.74 (6.20−7.24)
0.0
2.5
5.0
7.5
10.0
RMST (95% Credible Interval)

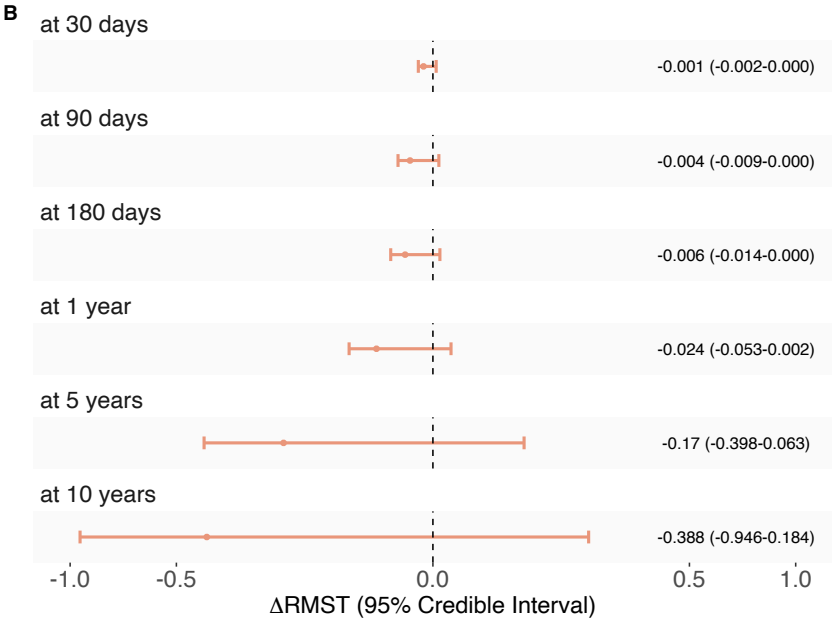
Subgroup analysis

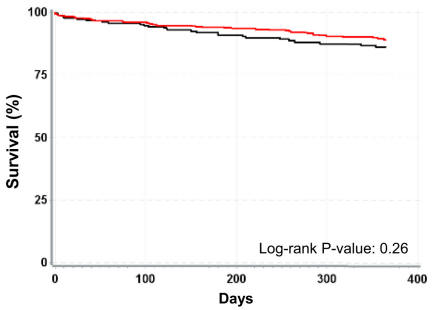
Subgroup	Disease-Free Survival (DFS)	
	Diff in 5-year survival rates	Diff in 5-year RMST (years)
Overall		
	0.40 (0.36-0.44)	1.66 (1.47-1.86)
Age		
Adult	0.39 (0.34-0.43)	1.62 (1.42-1.83)
Pediatric	0.59 (0.53-0.65)	2.37 (2.10-2.63)
Mixed	0.57 (0.40-0.68)	2.30 (1.67-2.72)
MRD detection time		
Induction	0.37 (0.33-0.41)	1.55 (1.35-1.75)
Consolidation	0.57 (0.51-0.63)	2.28 (2.02-2.54)
After consolidation	0.61 (0.55-0.66)	2.41 (2.16-2.66)
MRD detection method		
MFC	0.37 (0.31-0.42)	1.55 (1.31-1.78)
PCR(WT1)	0.59 (0.52-0.65)	2.37 (2.07-2.64)
PCR(gene)	0.60 (0.54-0.65)	2.40 (2.15-2.64)
NGS	0.55 (0.43-0.64)	2.23 (1.77-2.60)
Cytogenetics/FISH	0.45 (0.24-0.58)	1.85 (1.07-2.37)
Others	0.54 (0.42-0.63)	2.19 (1.74-2.55)
AML subtype		
CBF	0.36 (0.32-0.41)	1.53 (1.33-1.73)
Non-CBF	0.63 (0.57-0.69)	2.51 (2.25-2.77)
Specimen source		
Bone marrow	0.38 (0.34-0.42)	1.59 (1.40-1.78)
Peripheral blood	0.65 (0.59-0.71)	2.59 (2.32-2.85)
Mixed	0.57 (0.47-0.65)	2.31 (1.92-2.64)

Consistent beneficial effect of achieving MRD negativity.

Application: Gender disparity in endovascular surgery outcomes

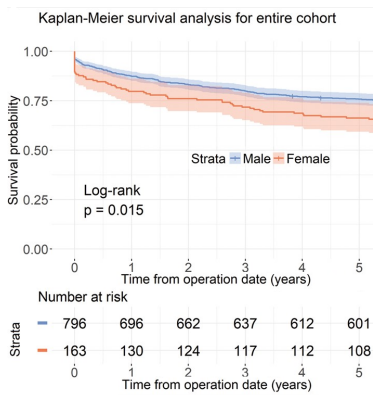


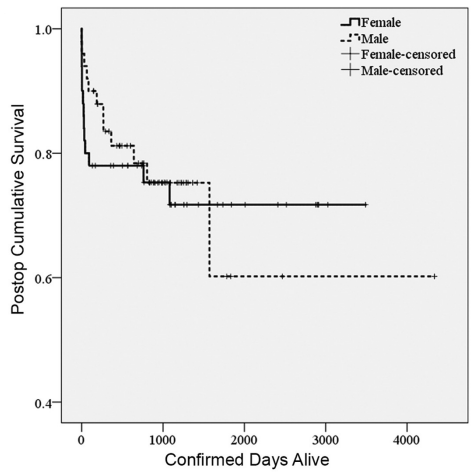




Days	0	100	200	300	365
Female	279	206	170	140	129
SE	0	0.014	0.019	0.023	0.025
Male	576	438	370	321	295
SE	0	0.008	0.011	0.014	0.015

1-year survival: women 89.6%, men 91.7%





Application: Risk of dementia by sleep duration

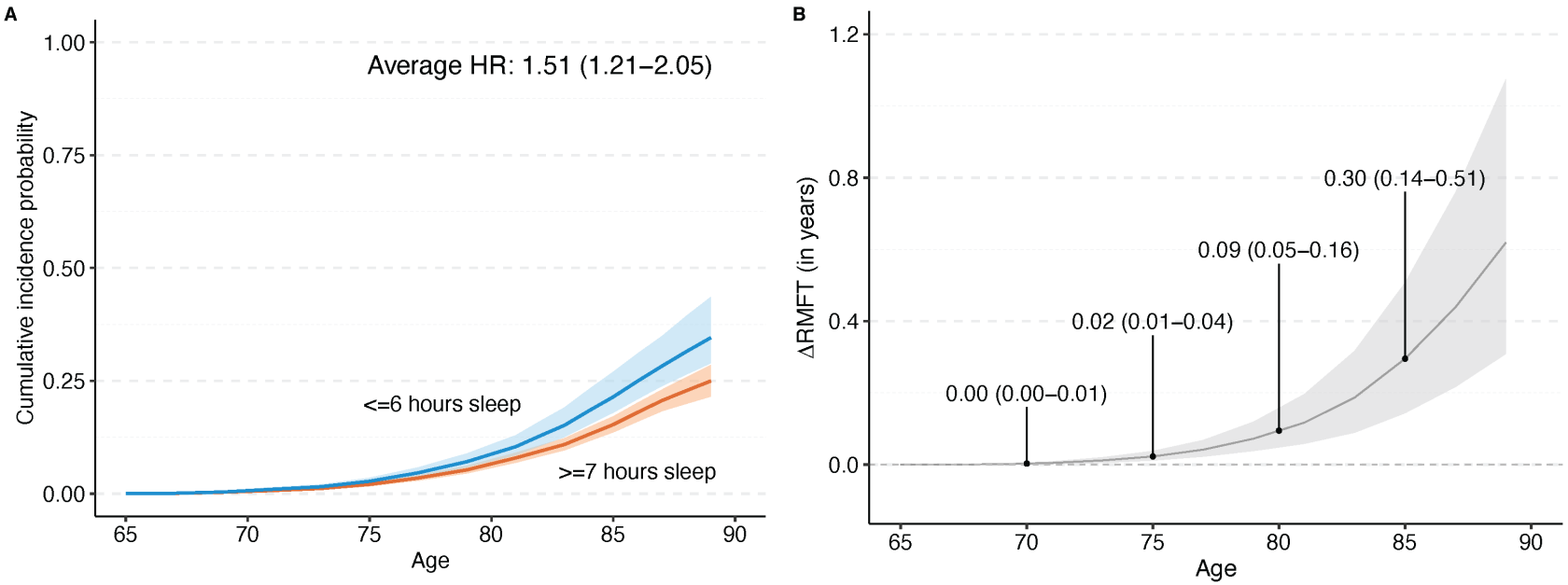


Figure: (A) Cumulative incidence probabilities, with 95% CrIs and averaged hazard ratio (HR), estimated using reconstructed survival data and hazard ratios. (B) Difference of restricted mean failure time (ΔRMFT), with 95% CrIs.

Summary

- MARS was developed to
  - improve robustness in meta-analysis of survival data
  - reduce selection bias
  - enrich scientific reporting
- Properties
  - Asymptotically consistent
  - Relatively efficient
  - General-purpose
- Promising small-sample empirical performance
  - comparable performance to IPD meta-analysis in both PH and non-PH scenarios

References

FDA (2020). *Guidance for Industry: Hematologic Malignancies: Regulatory Considerations for Use of Minimal Residual Disease in Development of Drug and Biological Products for Treatment*. Tech. rep.

Guyot, Patricia et al. (2012). "Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves". In: *BMC medical research methodology* 12.1, p. 9.

Johansen, Søren (1983). "An extension of Cox's regression model". In: *International Statistical Review/Revue Internationale de Statistique*, pp. 165–174.

Liu, Na, Yanhong Zhou, and J Jack Lee (2021). "IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves". In: *BMC Medical Research Methodology* 21.1, pp. 1–22.

Rathi, Ankit (2018). *Bayesian Statistics for Data Science*. https://medium.com/rathi.ankit/.

Royston, Patrick and Mahesh KB Parmar (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". In: *BMC medical research methodology* 13.1, pp. 1–15.

Short, Nicholas J et al. (2020). "Association of measurable residual disease with survival outcomes in patients with acute myeloid leukemia: a systematic review and meta-analysis". In: *JAMA oncology* 6.12, pp. 1890–1899.