Ranking-based Data Integration for Robust Outcomes Prediction and Biomarker Detection

Statistical Analyses of Multi-Outcome data (June 30, 2026)

Nicholas C. Henderson

University of Michigan, Department of Biostatistics

Multi-Study Data Integration

- Many studies focus on refining the clinical utility of newer biomarkers.
  - Such studies often have modest sample sizes.
- While limited information may be available about these biomarkers, there is usually **substantial information** available about outcomes in the same disease context.
  - Published and validated risk models.
  - Large disease registries.
  - Artificial Intelligence (AI) **"foundation models"**

Aim 1: Borrowing External Information to Improve Risk Model Estimation

- **Leverage external sources** to improve the performance of a risk prediction model for a smaller study of interest.
- Inferences in **a study of interest** can be improved by leveraging existing information that is **"external"** to the study.
- **Directly combining** these sources can be challenging due to inherent differences in study designs and patient populations

Borrowing External Information in the Age of AI

EUROPEAN UROLOGY FOCUS 10 (2024) 1011–1018

available at www.sciencedirect.com
journal homepage: www.europeanurology.com/eufocus

Prostate Cancer

Artificial INtelligence to Support Informed DEcision-making (INSIDE) for Improved Literature Analysis in Oncology

Article

A pathology foundation model for cancer diagnosis and prognosis prediction

https://doi.org/10.1038/s41586-024-07894-z

Received: 16 November 2023

Accepted: 1 August 2024

Published online: 4 September 2024

Check for updates

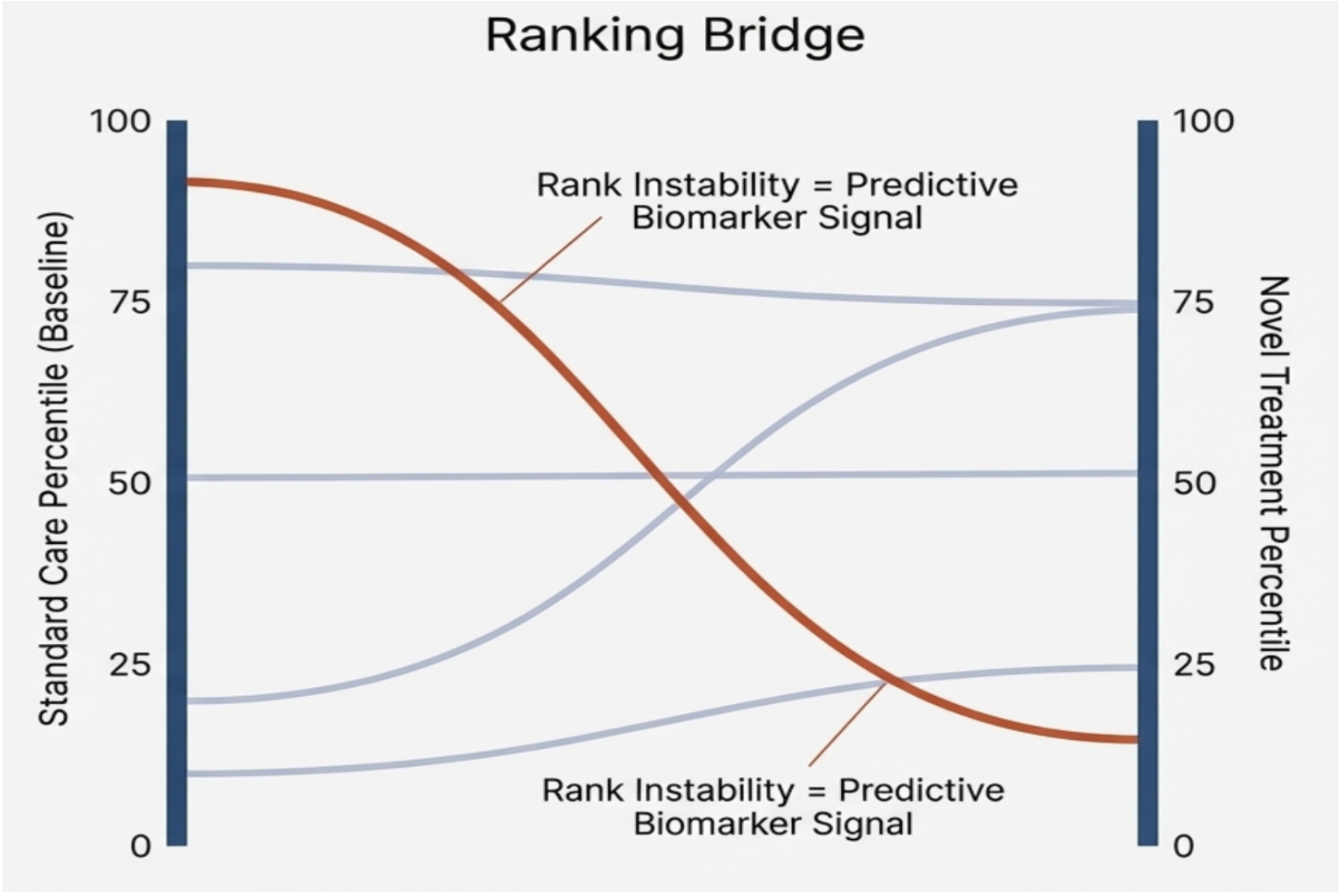
Xiyue Wang[1,2,24], Junhan Zhao[1,3,24], Eliana Marostica[1,4], Wei Yuan[5], Jietian Jin[6], Jiayu Zhang[5], Ruijiang Li[2], Hongping Tang[7], Kanran Wang[8], Yu Li[9], Fang Wang[10], Yulong Peng[11], Junyou Zhu[12], Jing Zhang[5], Christopher R. Jackson[1,13,14], Jun Zhang[15], Deborah Dillon[16], Nancy U. Lin[17], Lynette Sholl[16,18], Thomas Denize[16,18], David Meredith[16], Keith L. Ligon[16,18], Sabina Signoretti[16,18], Shuji Ogino[16,19,20], Jeffrey A. Golden[16,21], MacLean P. Nasrallah[22], Xiao Han[15], Sen Yang[1,2] ✉ & Kun-Hsing Yu[1,16,23] ✉

Histopathology image evaluation is indispensable for cancer diagnoses and subtype classification. Standard artificial intelligence methods for histopathology image

Aim 2: Precision Medicine and Contrasting External Risk Models

- Evaluating **how** external risk models **differ** can also provide useful information.
- Compare risk models across different studies involving **more than one treatment**.
- Look at **changes in risk rankings** across different treatment contexts.
- Aim for greater robustness by mainly integrating the **risk ranking information** across studies.

Contrasting Risk Rankings



Aim 1: Borrowing External Information to Improve Risk Model Estimation

Challenges of Integrating External Evidence

- **Outcome Incompatibility**: External and internal risk models may utilize different primary outcomes
  - Continuous vs. Binary outcomes
  - Progression-free survival (PFS) vs. Overall survival
  - PFS vs. "real-world" PFS.
  - PSA progression vs. Overall survival
- **Study Heterogeneity**
  - Differences in patient populations
  - Differences in standard of care over time

Challenges of Integrating External Evidence

- Many established risk models are not **well-calibrated** to generating particular survival probabilities.
- **Motivation**: The way many established risk models **order patients** is informative and is **more stable** across contexts.
- For example, in advanced **prostate cancer**:
  - Gleason score, visceral metastasis, disease volume, PSA levels will be consistent risk factors.

Internal Study Data Structure

- We have individual-level observations from an "internal study"

$$D_I = \{(Y_i, \mathbf{x}_i); i = 1, \ldots n_I\}, \qquad \mathbf{x}_i = (\mathbf{z}_i, \mathbf{b}_i)$$

- $\mathbf{z}_i$ - "conventional" covariates
- $\mathbf{b}_i$ - "novel" covariates
- **Goal**: Build a risk model for outcomes Y_i using all covariate information $\mathbf{x}_i$.

External Prognostic Model(s)

- Assume an external prognostic model can compute a **risk score** using the conventional covariates $\mathbf{z}_i$.
- Denote the external risk score function with

$$f_E(\mathbf{z}_i)$$

- **Larger values** of $f_E(\mathbf{z}_i)$ are interpreted as predicting **larger values** of $E(Y_i)$.
- $f_E(\mathbf{z}_i)$ does not need to be well-calibrated to the distribution of internal-dataset outcomes.
  - OncotypeDX recurrence score
  - Nottingham prognostic index
  - Bellmunt risk score
  - An older, historical risk model

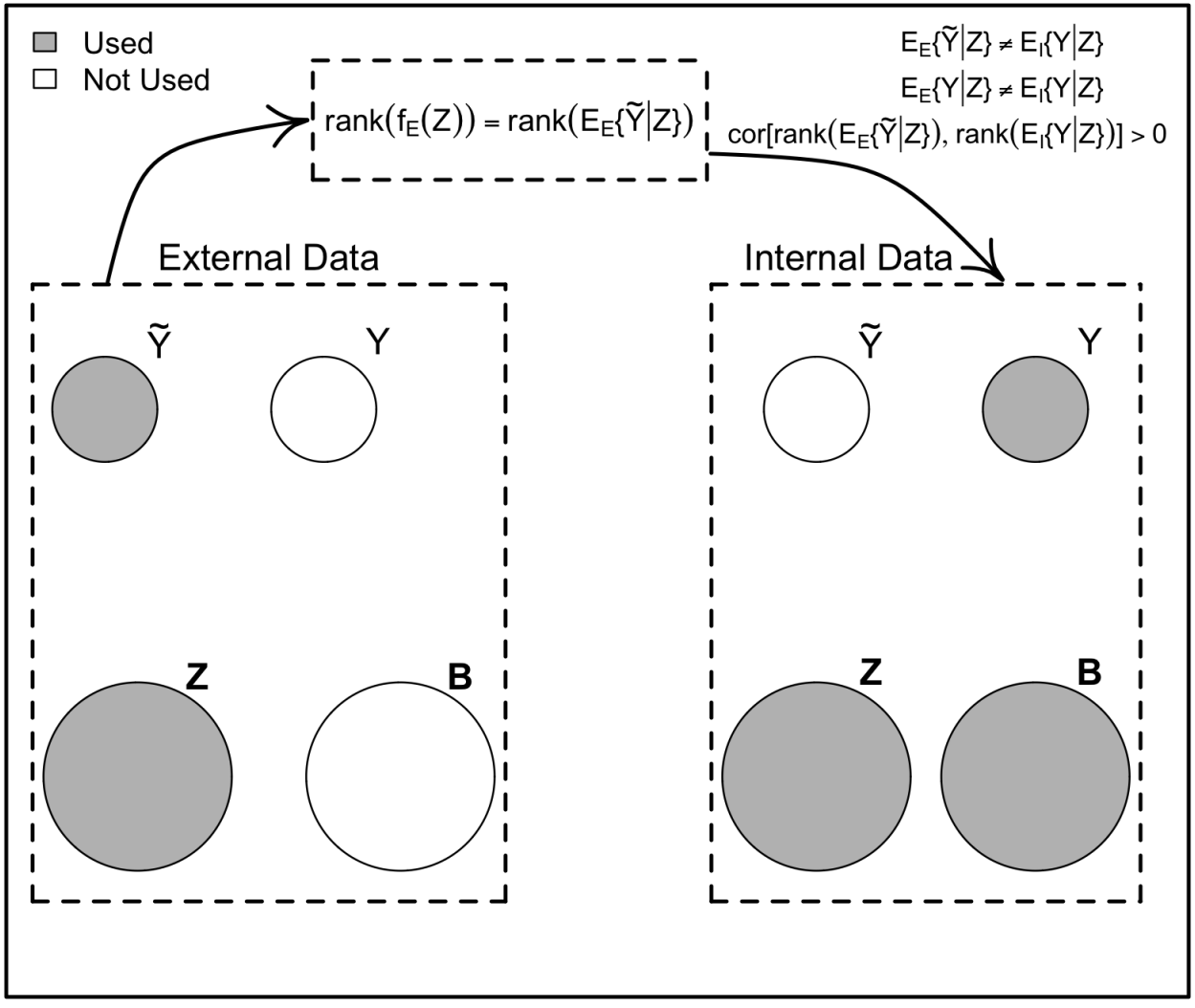
Transportability

- What are the relevant **"transportability assumptions"** for using $f_E(\mathbf{z}_i)$ in the context of the internal study?
- A risk score for the internal study would ideally use either

$$f_I(\mathbf{z}_i) = E_I(Y_i \mid \mathbf{z}_i) \quad \text{or} \quad g_I(\mathbf{x}_i) = E_I(Y_i \mid \mathbf{x}_i)$$

- We **will not** assume that $f_E(\mathbf{z}_i) = f_I(\mathbf{z}_i)$.
- Do assume that $f_E(\mathbf{z}_i)$ and $f_I(\mathbf{z}_i)$ (or $g_I(\mathbf{z}_i)$) have some degree of **nonparametric association**.

Transportability of Rankings



Transportability and External Risk Model Rankings

- How to harness nonparametric association between:
  - external risk scores
  - risk scores from internal model (for a fixed β)
- Start with the **rankings** from the external risk model:

$$\begin{aligned} r_i^E(\mathbf{z}_i) &= \text{rank}\{f_E(\mathbf{z}_i)\} \\ &= \sum_{j=1}^{n} I\Big(f_E(\mathbf{z}_i) \geq f_E(\mathbf{z}_j)\Big) \end{aligned}$$

Internal Risk Model

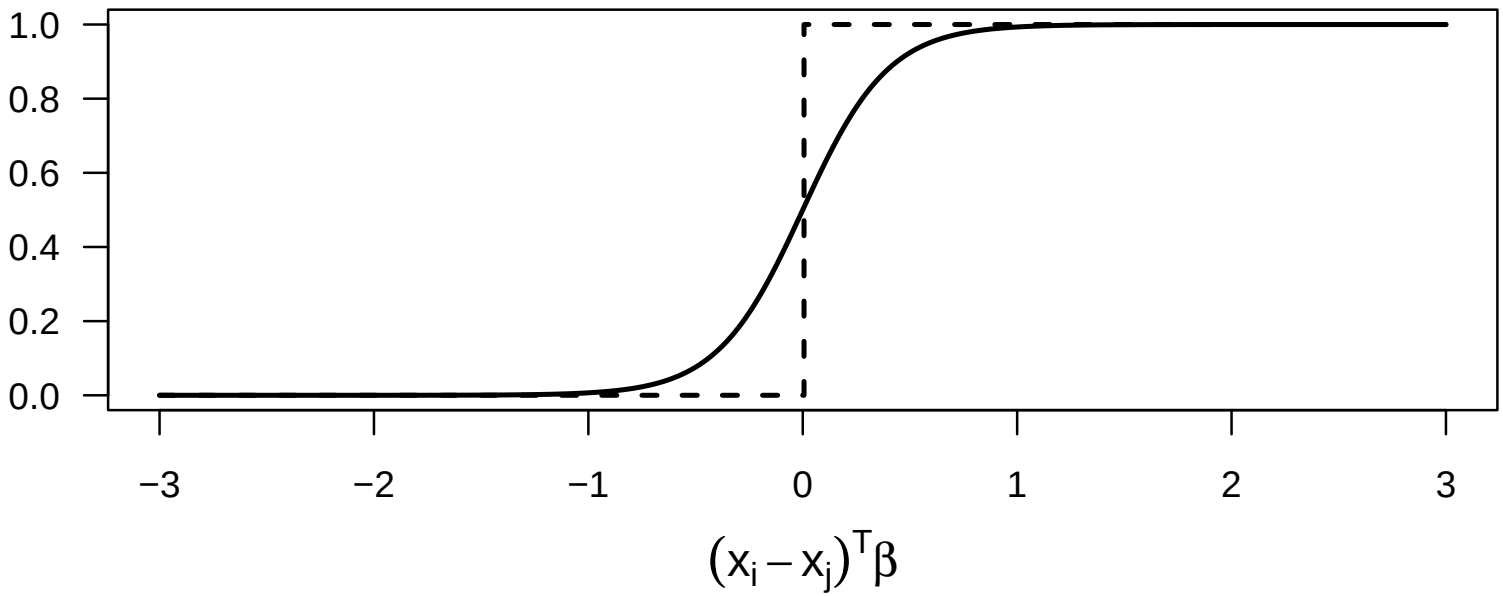
- For an internal risk regression model with parameters β, ranking according to risk is the same as ranking according to β.
  - These rankings can themselves be thought of as **model parameters**.
- Refer to these as **"ranking parameters"**:

$$\psi_i(\beta) = \text{rank}(\mathbf{x}_i^T \boldsymbol{\beta}) = \sum_{j=1}^{n} I\Big((\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\beta} \geq 0\Big)$$

- **Interpretation:** For a proposed β of the internal risk model, $\psi_i(\boldsymbol{\beta})$ is the ranking the internal risk model would assign to individual i.

Internal Risk Model

- **Strategy**: When estimating $\boldsymbol{\beta}$:
  - Compare $\psi_i(\boldsymbol{\beta})$ with the external rankings $r_i^E(\mathbf{z}_i)$.
  - Use smooth approximation: $\psi_i^\nu(\boldsymbol{\beta}) = \sum_{j=1}^n g_\nu\Big((\mathbf{x}_i - \mathbf{x}_j)^T\boldsymbol{\beta}\Big)$
  - $g_\nu(x) = 1/\{1+\exp(-x/\nu)\}$



Connecting Ranking Parameters with External Rankings

- Ranking parameters $\psi_i(\boldsymbol{\beta})$ represent rankings assigned by internal risk model.
- How to connect these with external ranking information?
- **Spearman's** (smooth) rank association:

$$D^{\nu}_{Sp,0}(\boldsymbol{\beta}; \mathbf{r}^E) = \frac{1}{4n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} r_i^E(\mathbf{z}_i) g_\nu\Big((\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\beta}\Big)$$

- **Kendall's** (smooth) τ association

$$\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} I\Big(r_i^E(\mathbf{z}_i) > r_j^E(\mathbf{z}_j)\Big) g_\nu\Big((\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\beta}\Big)$$

Estimating the Internal Risk Model

- Penalize the **negative rank association** between ranking parameters and external rankings and **minimize**

$$\ell_{\lambda,\alpha}(\beta_0, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 + \frac{\alpha}{2}\underbrace{\sum_{j=1}^{p}\beta_j^2}_{L_2 \text{ penalty}} - \lambda \log \underbrace{D_\bullet^\nu(\boldsymbol{\beta}; \mathbf{r}^E)}_{\text{rank association}}$$

- Rank association can be expressed as:

$$D_\bullet^\nu(\boldsymbol{\beta}; \mathbf{r}^E) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} g_\nu\left((\mathbf{x}_i - \mathbf{x}_j)^T\boldsymbol{\beta}\right)$$

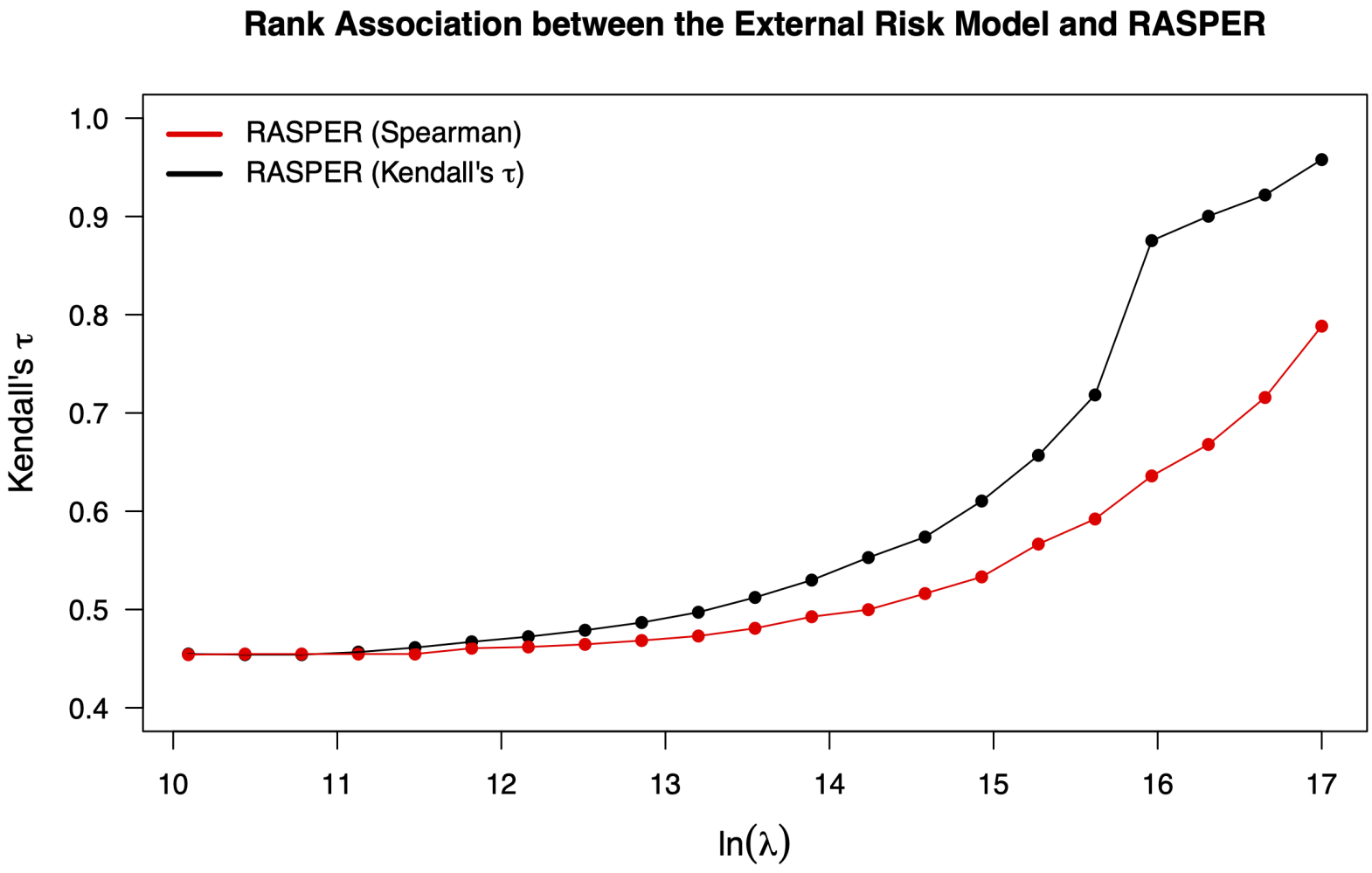
  - Spearman $w_{ij} = r_i^E(\mathbf{z}_i)/n^3$
  - Kendall's τ, $w_{ij} = 2I(r_i^E(\mathbf{z}_i) > r_j^E(\mathbf{z}_j))\big/(n(n-1))$

RASPER

- Refer to the vector of regression coefficients $\hat{\boldsymbol{\beta}}_{\lambda,\alpha}$ which minimize the ranking penalized objective function as:
- the **R**anking-**AS**sociated **PE**nalized **R**egression (RASPER) estimate

Varying the penalty term

- The rank association with the external model provides a natural range $\lambda_l \leq \lambda \leq \lambda_u$, for λ.



Hyperparameter Selection (LOO)

- The most effective approach is typically **leave-one-out** (LOO) cross-validation.
  - This is typically true for modest sample sizes (i.e., $n \leq 100$).

Simulation Studies

- RASPER performs well relative to more "direct borrowing" methods when:
  1. Distance between internal and external risk scores is large,
  2. **Rank correlation** between internal and external risk scores is **high**,
  3. The true internal model is **linear** and the external model is **nonlinear**.

Simulation Study

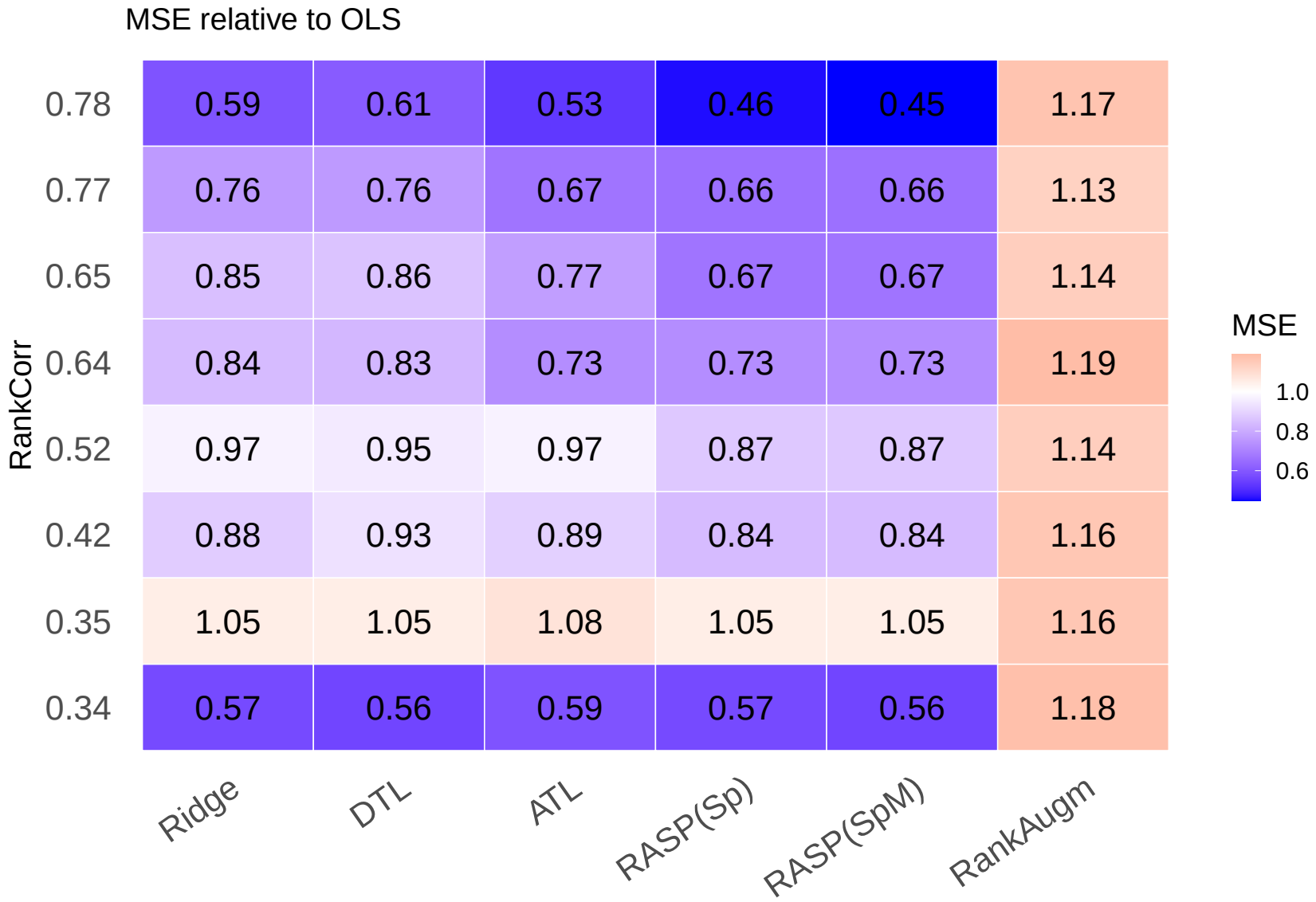
- True internal risk model is **linear**.
  - 4 conventional covariates
  - 2 novel covariates
- The mean of the external risk model is **nonlinear**

$$\mu_E(\mathbf{z}_i) = \Big(1 + f_1(z_{i1}) + \theta_2 f_2(z_{i2}) + \theta_3 f_1(z_{i1}) f_2(z_{i2})\Big) \times \exp\Big\{ -\theta_4 I(z_{i3} < 2) + 10\theta_4 I(z_{i3} \geq 2) - \theta_5 z_{i5} \Big\}$$

- For methods that require the external model to be linear, use an approximation of this nonlinear function

$$\tilde{\mu}_E(\mathbf{z}_i) = \sum_{j=1}^{p} \tilde{\beta}_{j,E} z_{ij}$$

Simulation Results (n=50)



Prognostic Model for ICI-treated Prostate Cancer Patients

- **Immune checkpoint inhibitors** (ICIs) have transformed cancer treatment in past 10-20 years.
- Use of ICI in treating **prostate cancer** has been more limited.
- Only patients with certain molecular signatures are treated with ICI
  - Mismatch Repair Deficiency (MMRd)
  - Microsatellite Instability (MSI)
  - Tumor Mutational Burder (TMB)

Prognostic Model for ICI-treated Prostate Cancer Patients

- To construct this prognostic model, we will use the **MSK-CHORD** data resource (Jee et al., 2024)
- Harmonized, real-world clinicogenomic data source
  - Large group of patients (n = 24,490)
  - Five cancer types
  - **Prostate cancer** (n = 3,211)
- Even though the number of prostate cancer is substantial, the number of **ICI-treated** patients is much smaller.
  - Only include ICI treatment for mCRPC patients.
  - Only include ICI treatment post chemotherapy

Prognostic Model for ICI-treated Prostate Cancer Patients

- After restricting the analysis to post-chemotherapy mCRPC patients with complete data
  - **n = 79** patients remaining for our analysis
- "Conventional" covariates to include:
  - **PSA** at start of ICI treatment
  - Presence of **visceral metastasis** at start of ICI treatment
  - **ECOG** performance status
  - **Time to progression** on chemotherapy prior to ICI

External Prognostic Model

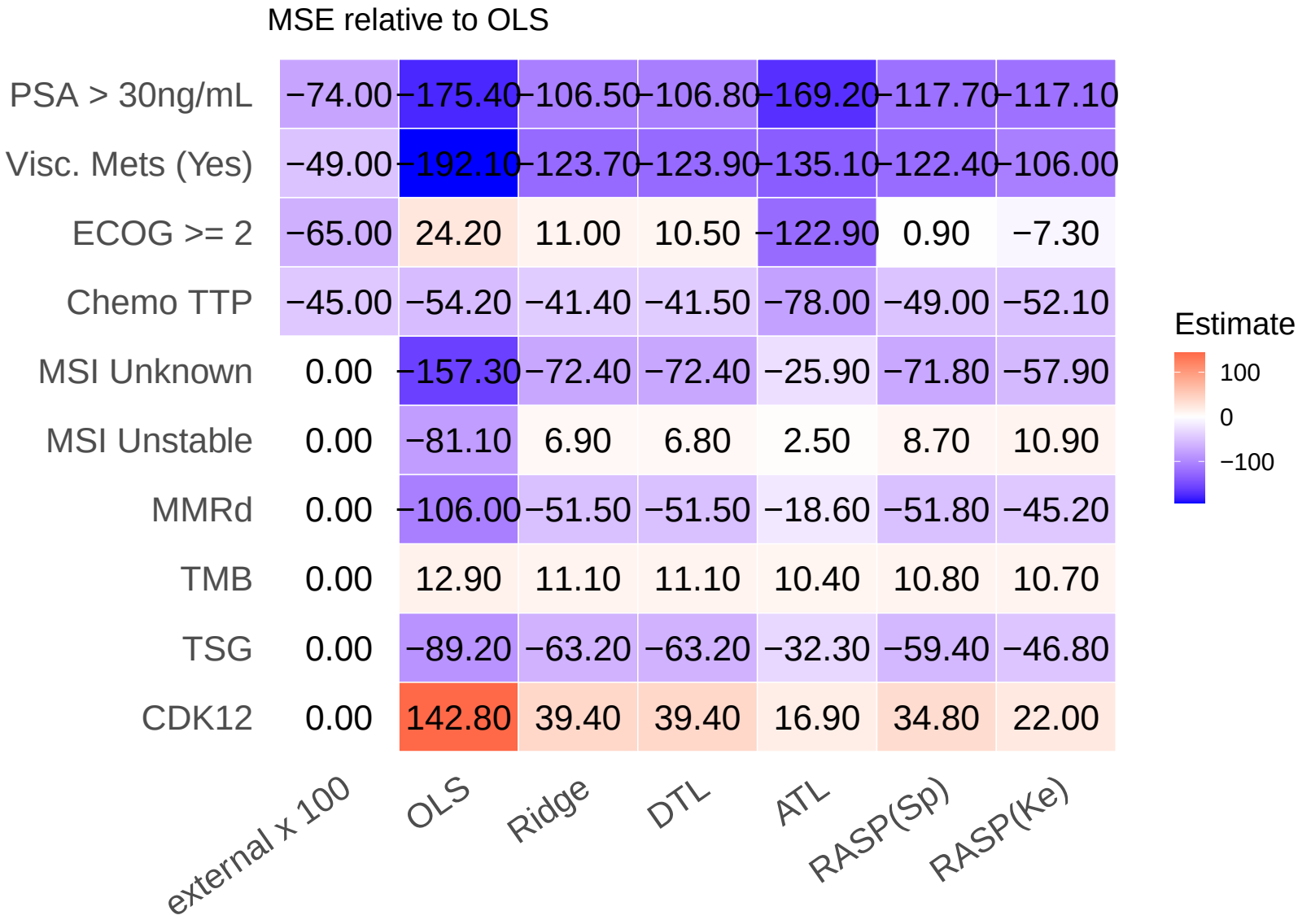
- Use a recently published nomogram (Suzuki et al. (2025)) designed to predict survival for (mCRPC) patients, **after** first-line chemotherapy.
- This risk score is:

$$
\begin{aligned}
f_E(\mathbf{z}_i) = 0.90 &+ 0.74 I(\mathrm{PSA}_i > 30 \tfrac{\mathrm{ng}}{\mathrm{ml}}) \\
&+ 0.49 I(\text{visceral metastasis}) \\
&+ 0.65 I(\mathrm{ECOG}_i \geq 2) \\
&- 0.45 \min\left\{2, \frac{\text{days to progression on prior chemo}}{180}\right\}.
\end{aligned}
$$

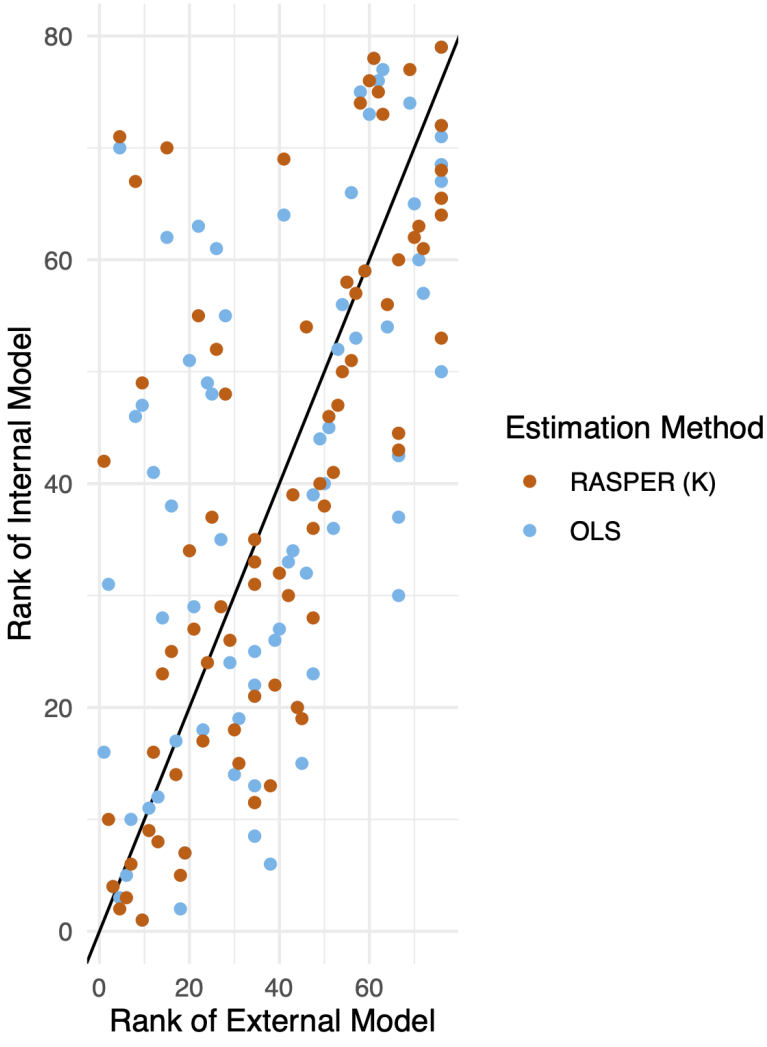
Novel Covariates

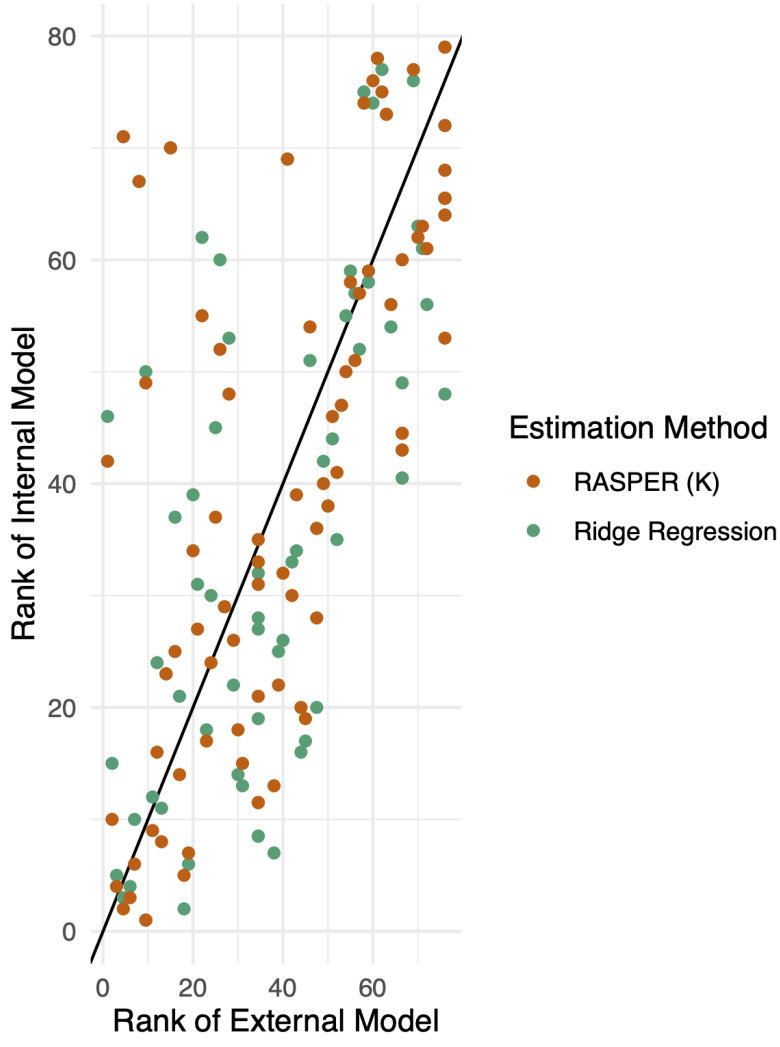
- **Goal**: Update the external model with the following **novel covariates**:
  1. **MMRd** (Yes/No)
  2. **MSI status** (stable, unstable, unknown)
  3. **Tumor mutational burden** (TMB)
  4. Alteration in a **tumor suppressor gene** (TSG) either *TP53*, *PTEN*, or *RB1*
  5. Alterations in the *CDK12* gene

Coefficient Estimates from an Internal 3-year RMST Model

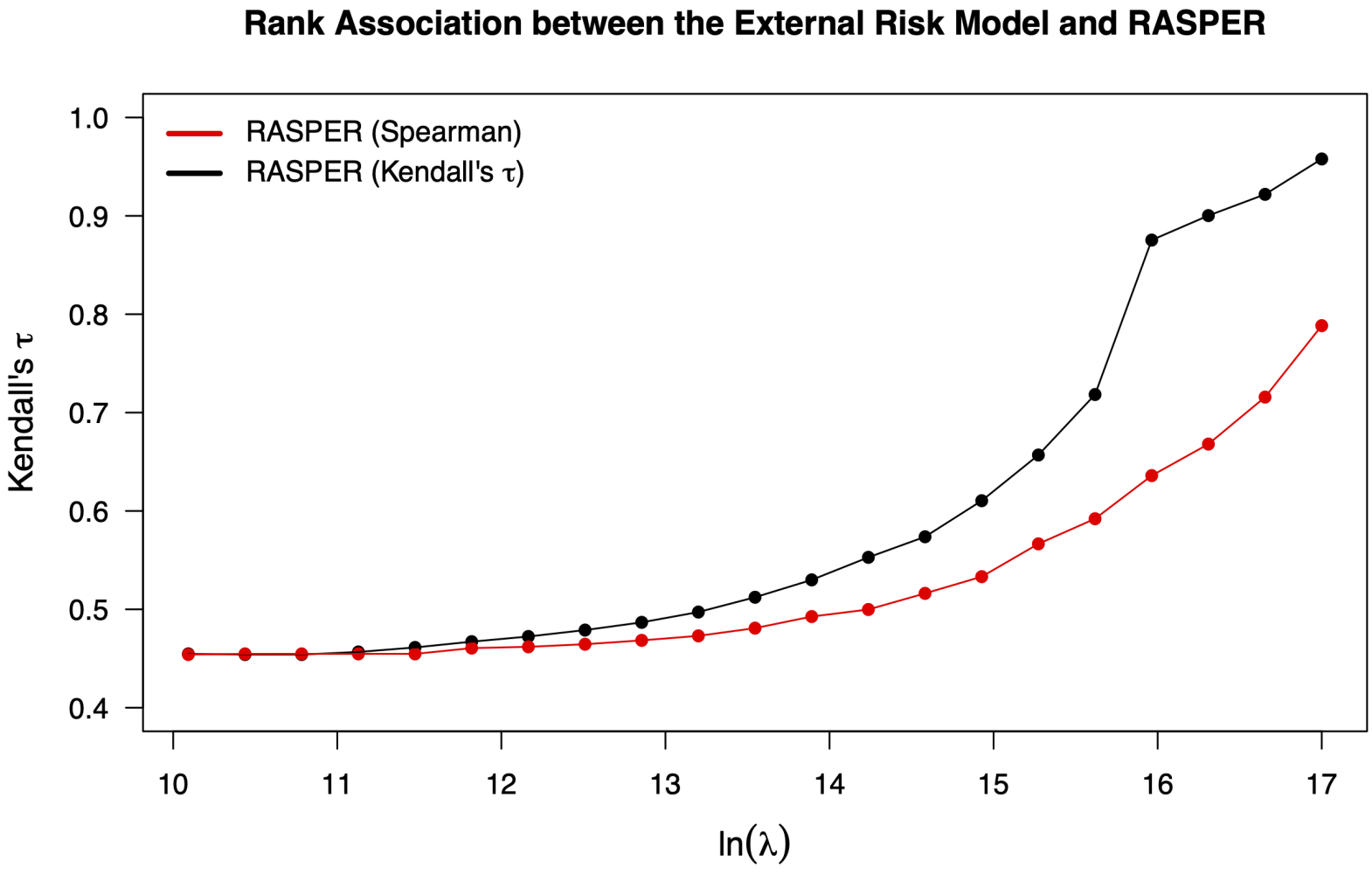


Ranking Comparisons





Rank Correlation with Suzuki (2025) model

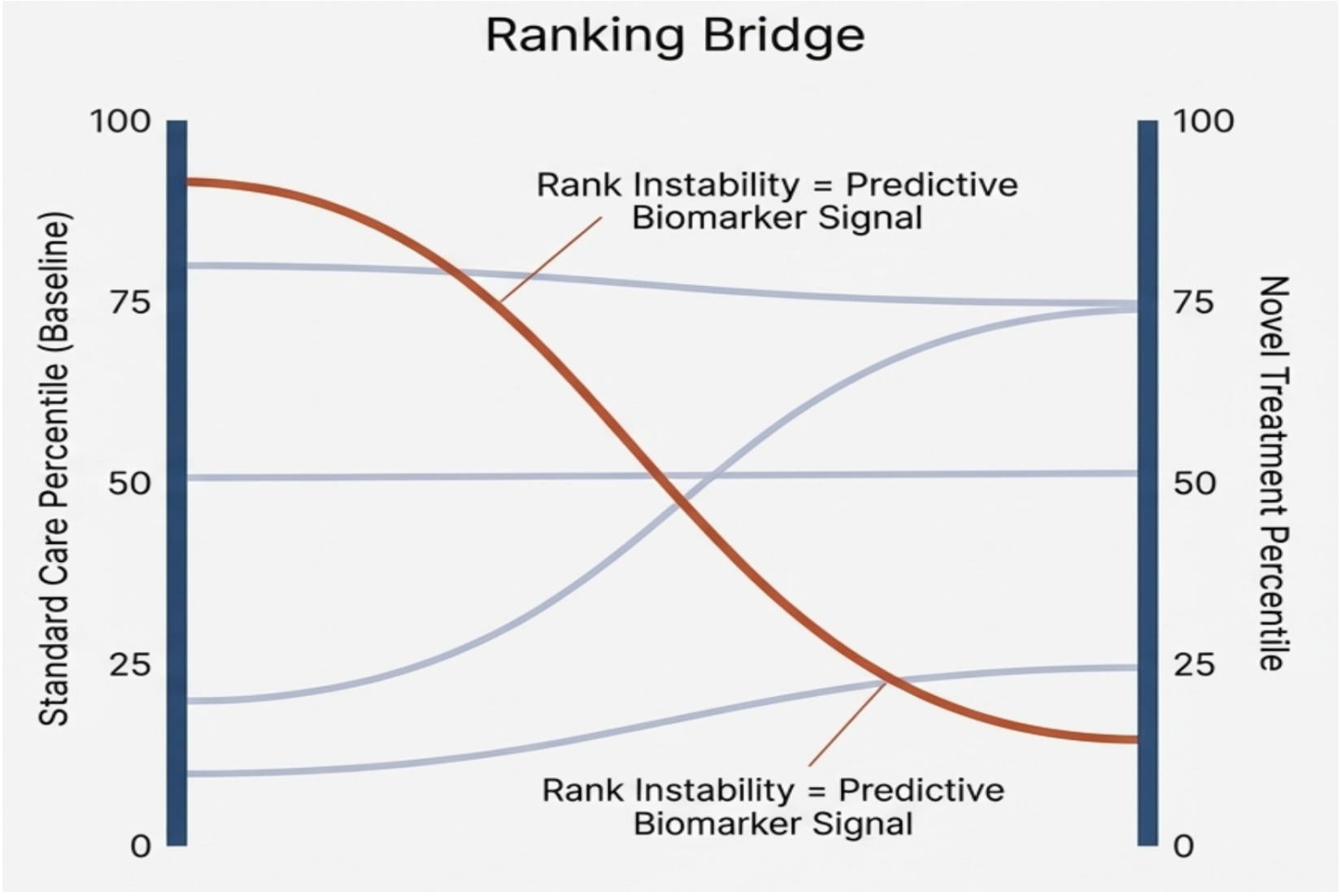


Aim 2: Precision Medicine and Contrasting External Risk Models

Precision Medicine and Real-World Evidence

- Major challenges when using **real-world evidence sources** to identify predictive biomarkers:
  - Observed treatments often under a **single standard of care**
  - Limited **direct comparisons** of two or more treatments.
- Comparing risk models across **different contexts** can be useful.
  - Look at changes in risk rankings across different treatment contexts.
  - Covariates associated with large changes in risk rankings: **potential predictive biomarkers**.

Ranking Discrepancies across contexts



Different Treatments across Different Contexts

- Consider **hypothetical outcomes** from the following four configurations

$$\text{Internal/Treatment: } Y_i = f(\mathbf{x}_i) + h(\mathbf{x}_i) + g(\mathbf{x}_i) + \varepsilon_i, \quad (1)$$
$$\text{Internal/Control: } Y_i = f(\mathbf{x}_i) + h(\mathbf{x}_i) + \varepsilon_i, \quad (2)$$
$$\text{External/Control: } Y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad (3)$$
$$\text{External/Tretment: } Y_i = f(\mathbf{x}_i) + g(\mathbf{x}_i) + \varepsilon_i, \quad (4)$$

- Often **only observe** outcomes from settings (1) and (3)
- $g(\mathbf{x})$ **treatment effect variation** function
- $h(\mathbf{x})$ **context variation** function

Contrastive RASPER (CRASPER)

- Consider our **observed outcomes**

$$\begin{aligned} \text{Internal/Treatment: } Y_i &= f(\mathbf{x}_i) + h(\mathbf{x}_i) + g(\mathbf{x}_i) + \varepsilon_i, \\ \text{External/Control: } Y_i &= f(\mathbf{x}_i) + \varepsilon_i \end{aligned}$$

- We can directly estimate $f(\mathbf{x})$ from the **External/Control** context.
- Suppose $f(\mathbf{x})$ and $f(\mathbf{x}) + h(\mathbf{x})$ have high**rank association**.
- Then, we can use the approximation

$$f(\mathbf{x}_i) + h(\mathbf{x}_i) \approx c_0 \times \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_r \tag{5}$$

where $\mathbf{x}^T \hat{\boldsymbol{\beta}}_r$ has very high **rank association** with $\hat{f}(\mathbf{x})$.

CRASPER

- Plug in $c_0 \times \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_r$ for $f(\mathbf{x}_i) + h(\mathbf{x}_i)$.
- We then minimize:

$$\ell(\beta_0, c_0, \boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \mathbf{x}_i^T \hat{\beta}_r - c_0 \mathbf{x}_i^T \theta \right)^2 + \alpha_1 \sum_{j=1}^{p} |\theta_j| + \frac{\alpha_2}{2} \sum_{j=1}^{p} \theta_j^2$$

- $\hat{\boldsymbol{\beta}}_r$ - Han's maximum rank correlation estimator for external risk scores $\hat{f}(\mathbf{x}_1), \ldots, \hat{f}(\mathbf{x}_n)$.
- **Parameter Constraint**: $c_0 \geq 0$.

CRASPER

- Generates an **interpretable, sparse model** for the **treatment effect variation** function $g(\mathbf{x})$
- Should primarily be used for **detecting** potential **predictive covariates**.
  - These covariates would need to be later validated.
- Relies strongly on the assumption of **ranking transportability** across study contexts.

RASPER Summary

- **Borrowing ranking information** has several advantages:
  - Rankings are often more transportable
  - Outcomes do not have to be "matched" across studies
- Borrowing is done by introducing the notion of **ranking parameters**.
- Penalization is done by looking at association between **ranking parameters** with **external ranks**.

Resources

- ArXiv paper describing RASPER in more detail:
  - **"Robust Updating of a Risk Prediction Model by Integrating External Ranking Information"**. https://arxiv.org/abs/2603.10389
- RASPER package https://github.com/nchenderson/rasper