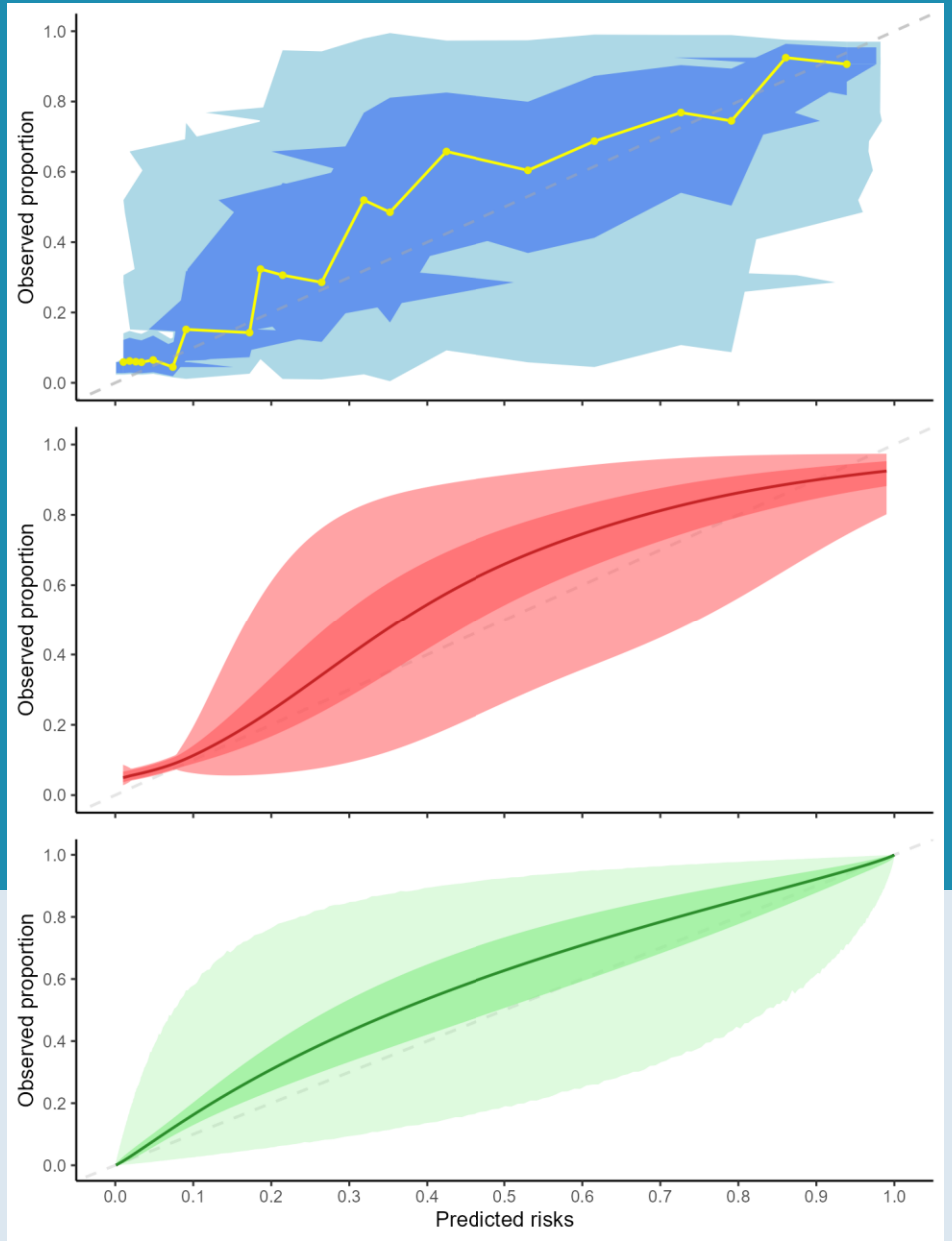
KU LEUVEN

DRAWING MULTICENTER CALIBRATION CURVES

Lasai Barreñada, Laure Wynants & Ben Van Calster

SEB
SOCIEDAD ESPAÑOLA DE BIOESTADÍSTICA

IOTA



What is calibration?

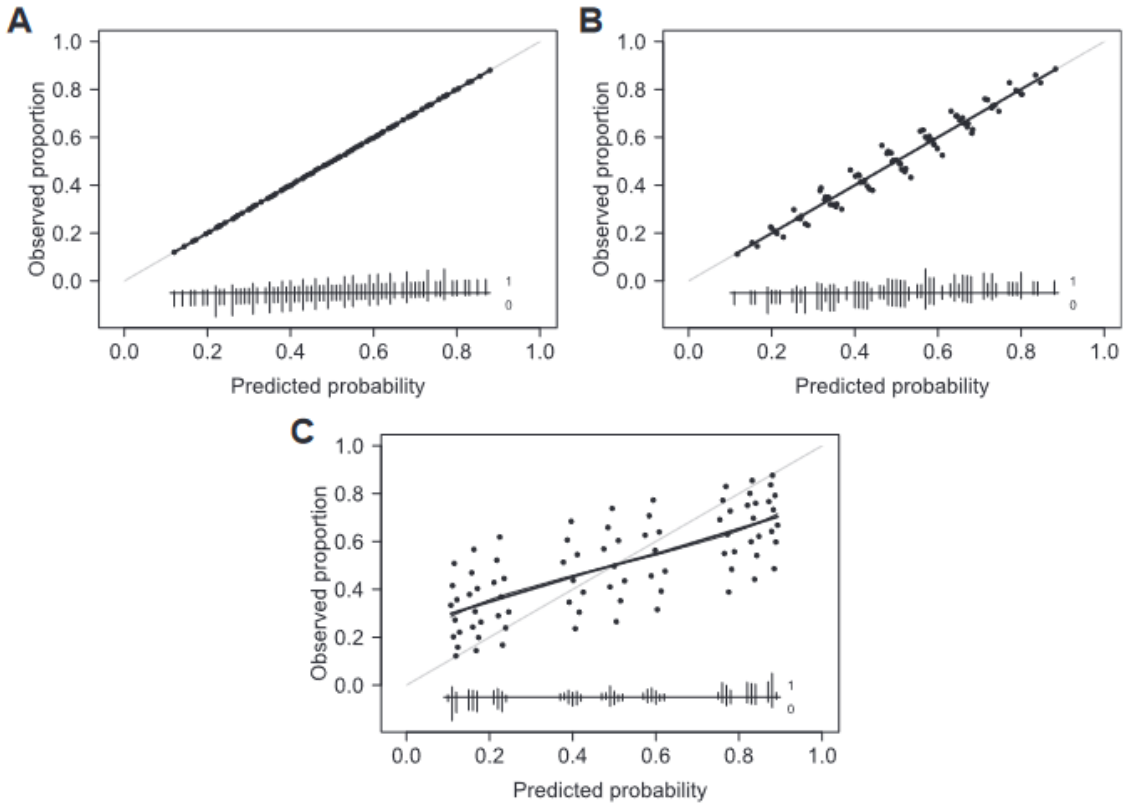
“The process of checking a measuring instrument to see if it is accurate”

Calibration in clinical prediction models

- Agreement between predicted and observed risks.
- **Challenge**: estimate observed risk for binary outcomes

Level	Definition
Mean	Observed event rate equals average predicted risk; "calibration-in-the-large"
Weak	No systematic overfitting or underfitting and/or overestimation or underestimation of risks; "logistic calibration"
Moderate	Predicted risks correspond to observed event rates
Strong	Predicted risks correspond to observed event rates for each and every covariate pattern

- Calibration often receives less attention than discrimination (AUC)
- Bad calibration can make predictions misleading.

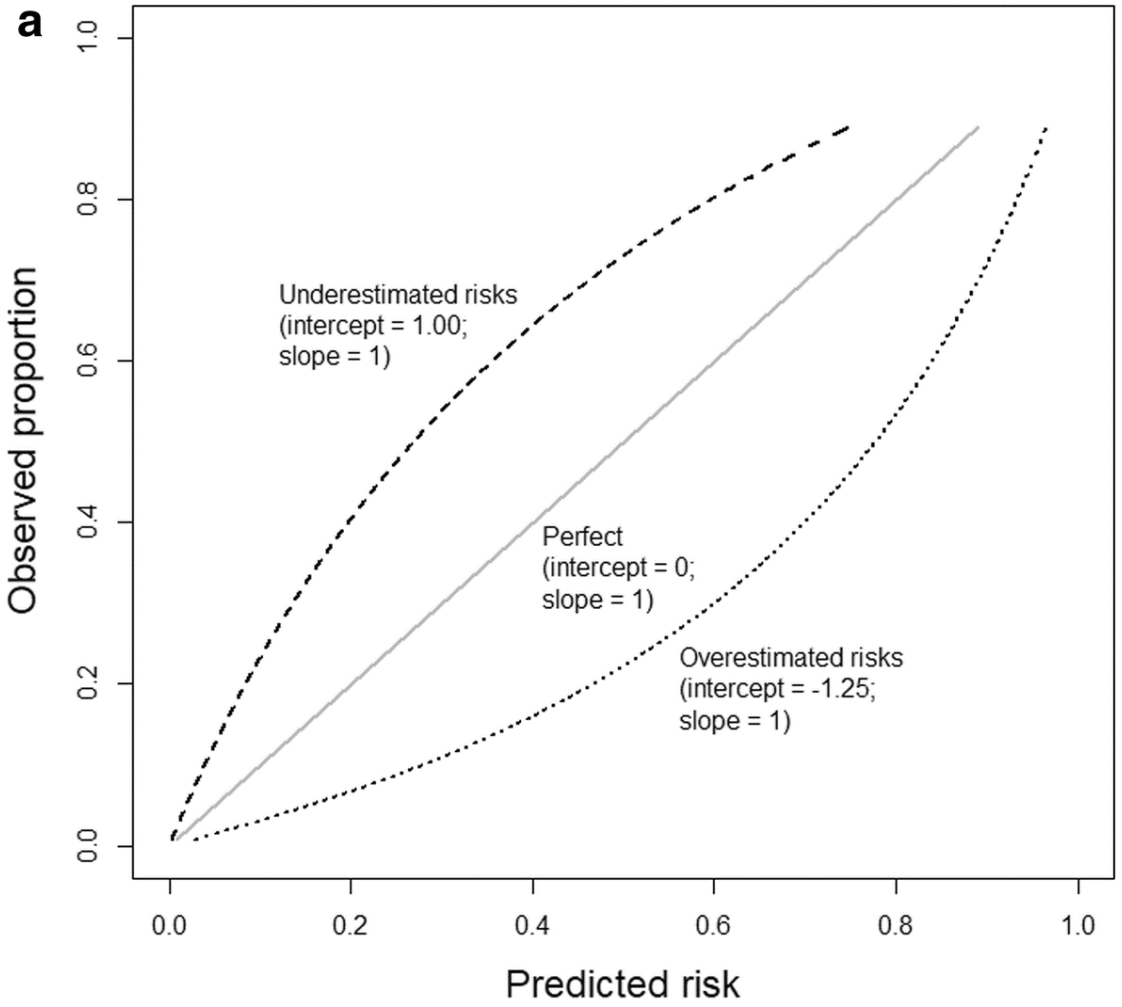


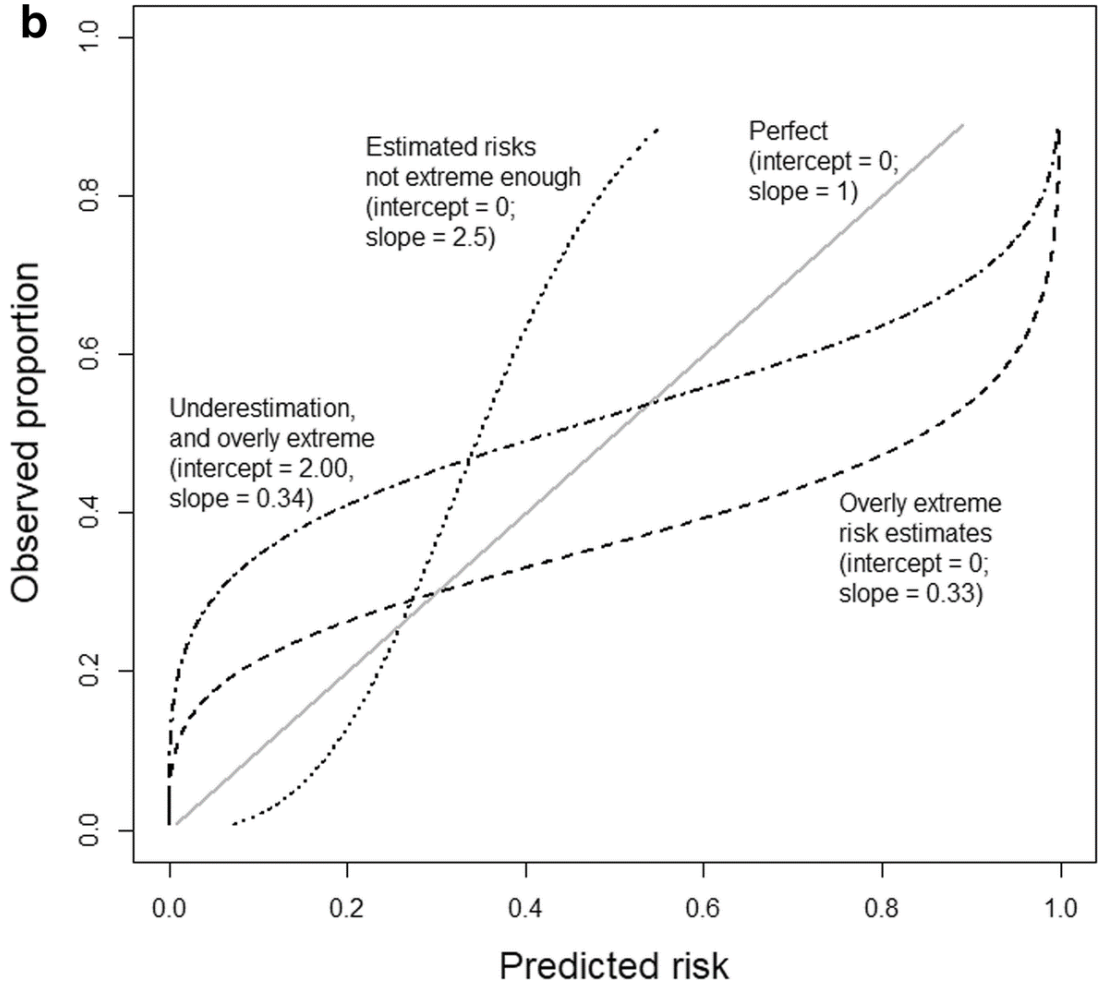
Calibration plots illustrating (A) strong calibration, (B) moderate but not strong calibration, and (C) miscalibration.

Van Calster et al., JCE, 2016
Van Calster & Vickers, Med Decis Making, 2015
Van Calster et al., BMC, Med 2019
Collins et al., BMC Med Res Meth, 2014

Calibration plot

$$Y = intercept + slope \times LP$$



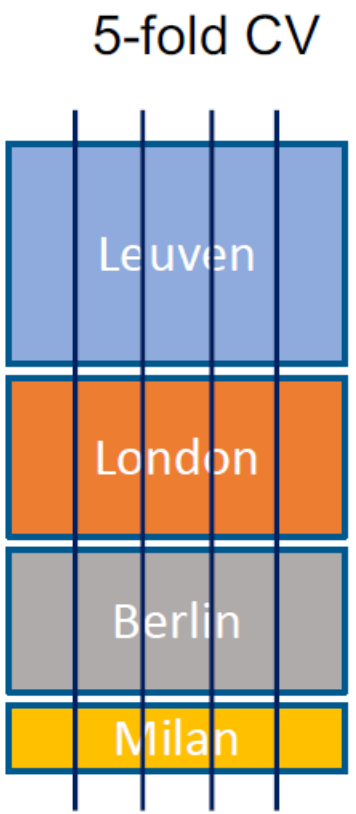


Van Calster et al., BMC Med, 2019

KU LEUVEN

Multicenter studies

- Studies that are collected simultaneously in several locations.
  - Increasingly popular.
  - Allow to obtain bigger sample sizes faster.
- More generalizability
- No independence → Random effects
- Models should be valid for all centers



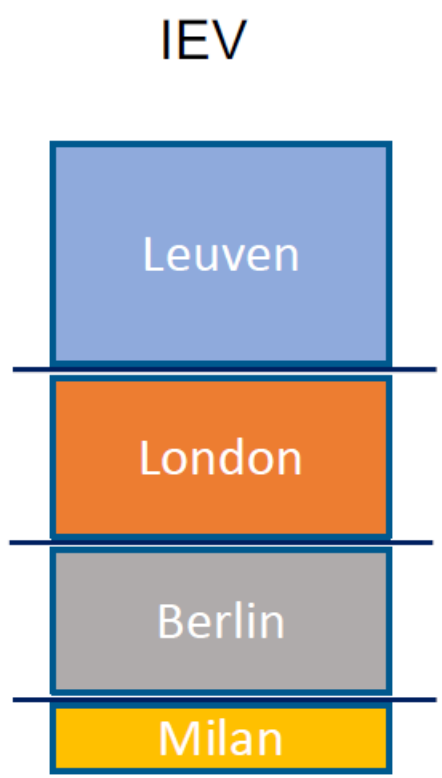
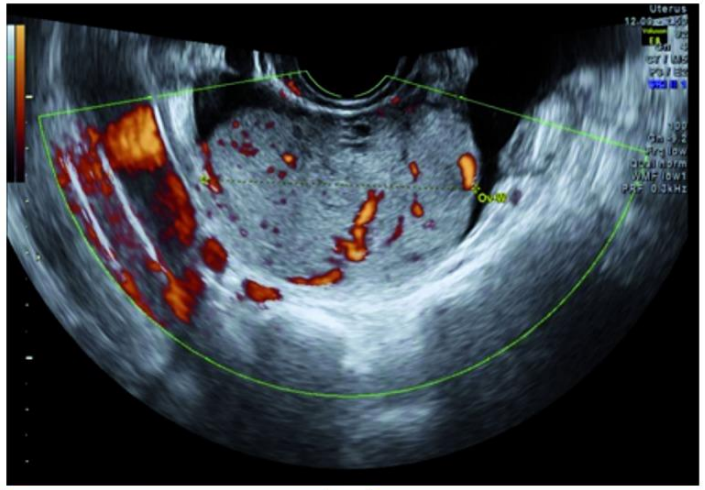
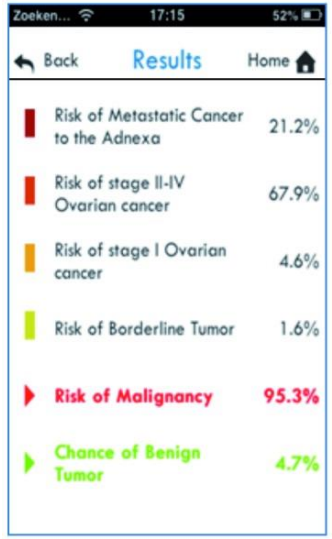
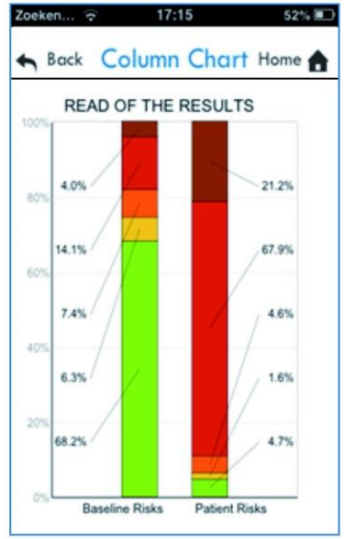


Illustration: ovarian cancer diagnosis

- ADNEX model:
  - Aim: **risk of malignancy of an ovarian mass**.
  - **Predictors**: age, type of centre (oncology centres v other hospitals) and 6 ultrasound based predictors.
  - Logistic regression model with random intercepts per center.
  - Developed with 5909 patients from 24 hospitals.
  - AUC: 0.94 (0.93-0.95)
  - Good calibration
  - **Externally validated in more than 47 studies and 17.000 tumours**
- We test performance in an independent dataset of 2196 patients from 9 centers where there were at least 100 patients.
- The distribution of predictions is very positively skewed (median 13%)

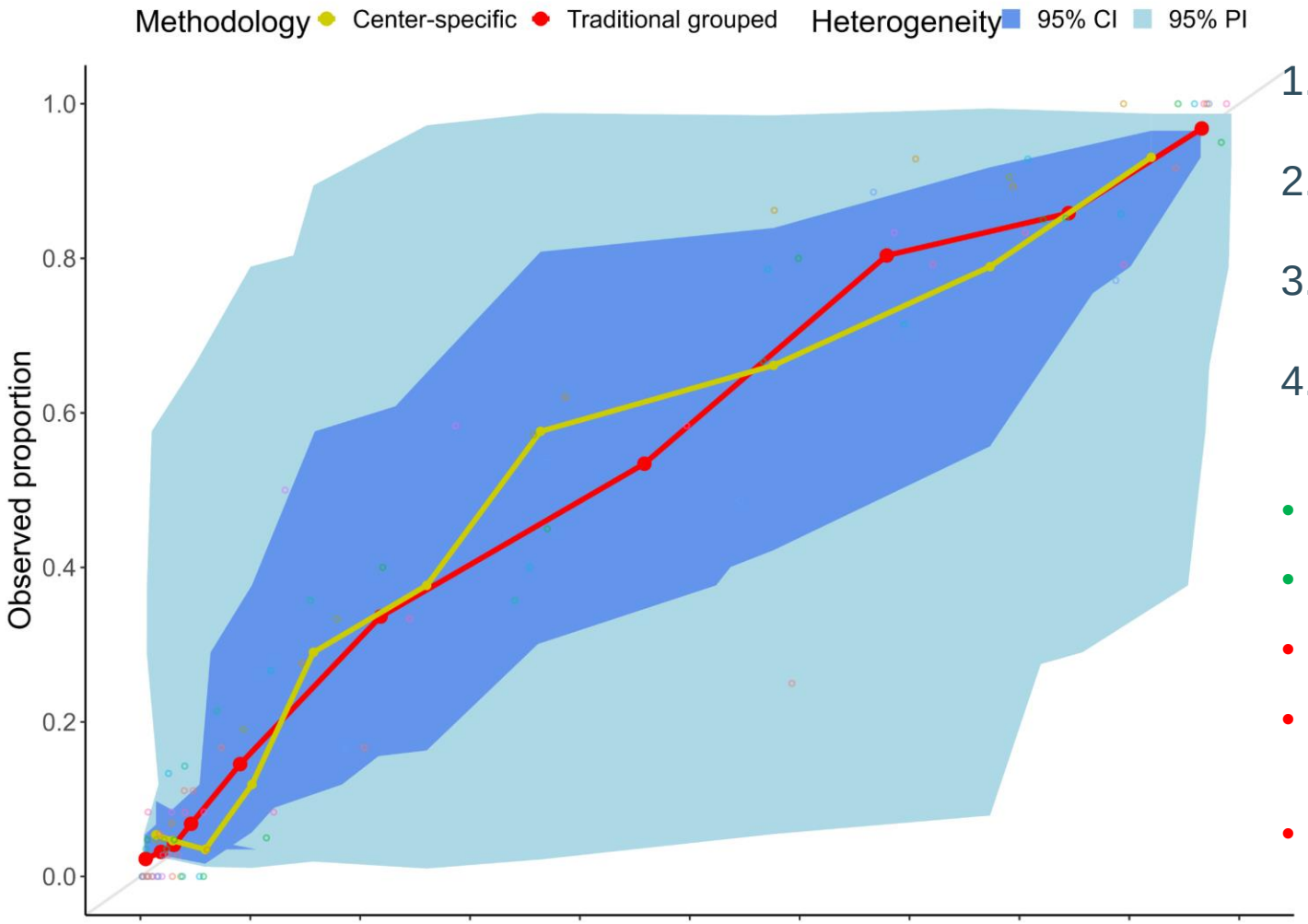






Van Calster et al., BMJ, 2014
Van Calster et al., Facts Views Vis Obgyn, 2015
Van Calster et al., BMJ, 2020
Barreñada et al., BMJ Med (Accepted), 2024

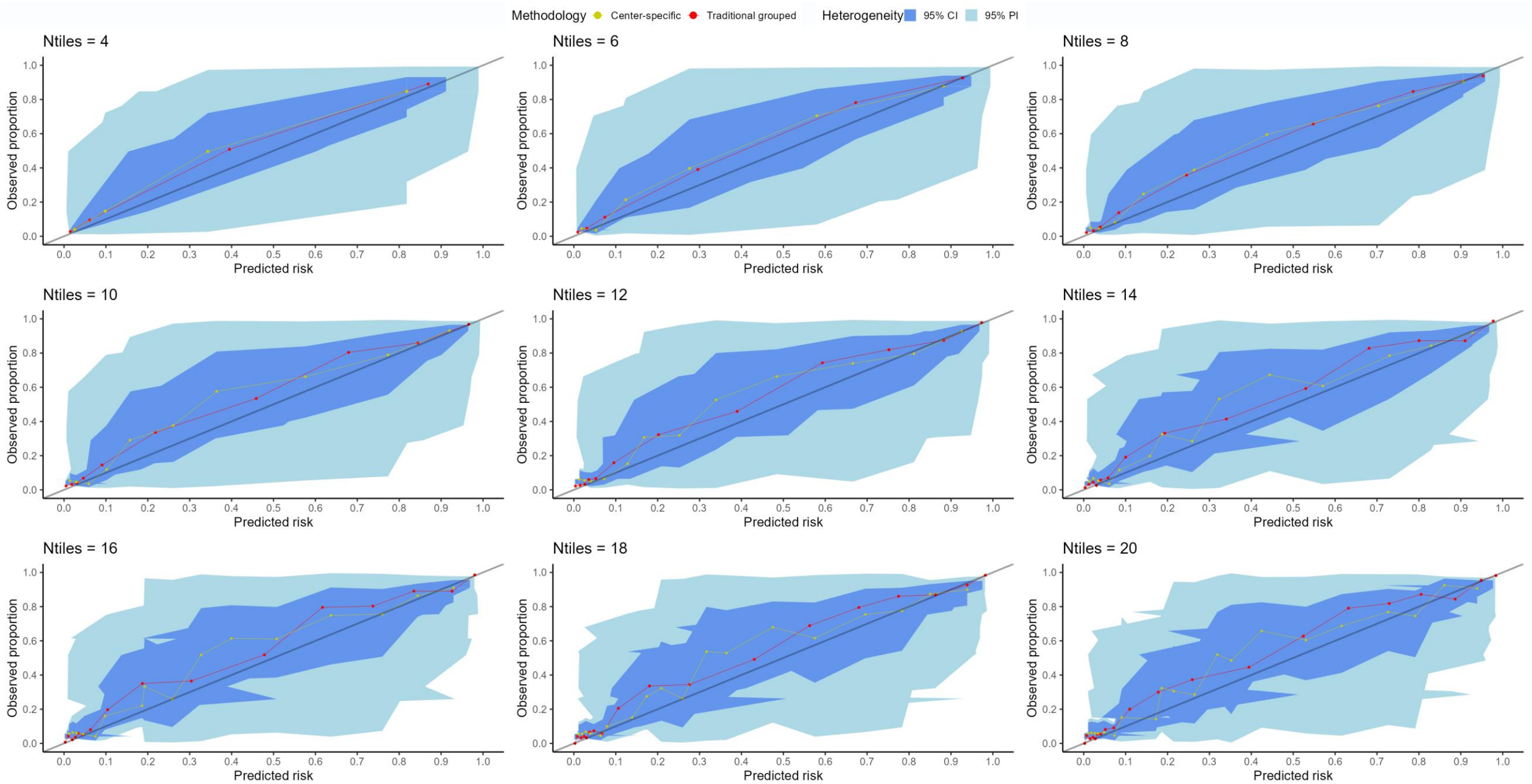
Clustered quantile (10 quantiles)



1. For each center we group the predicted probabilities by center in 10 quantiles.
2. We calculate mean outcome and mean predicted probabilities in each quantile of each center.
3. We meta-analyse the quantile of each center using bivariate random effects meta-analysis.
4. CI and PI as explained in Riley et al. (2011)

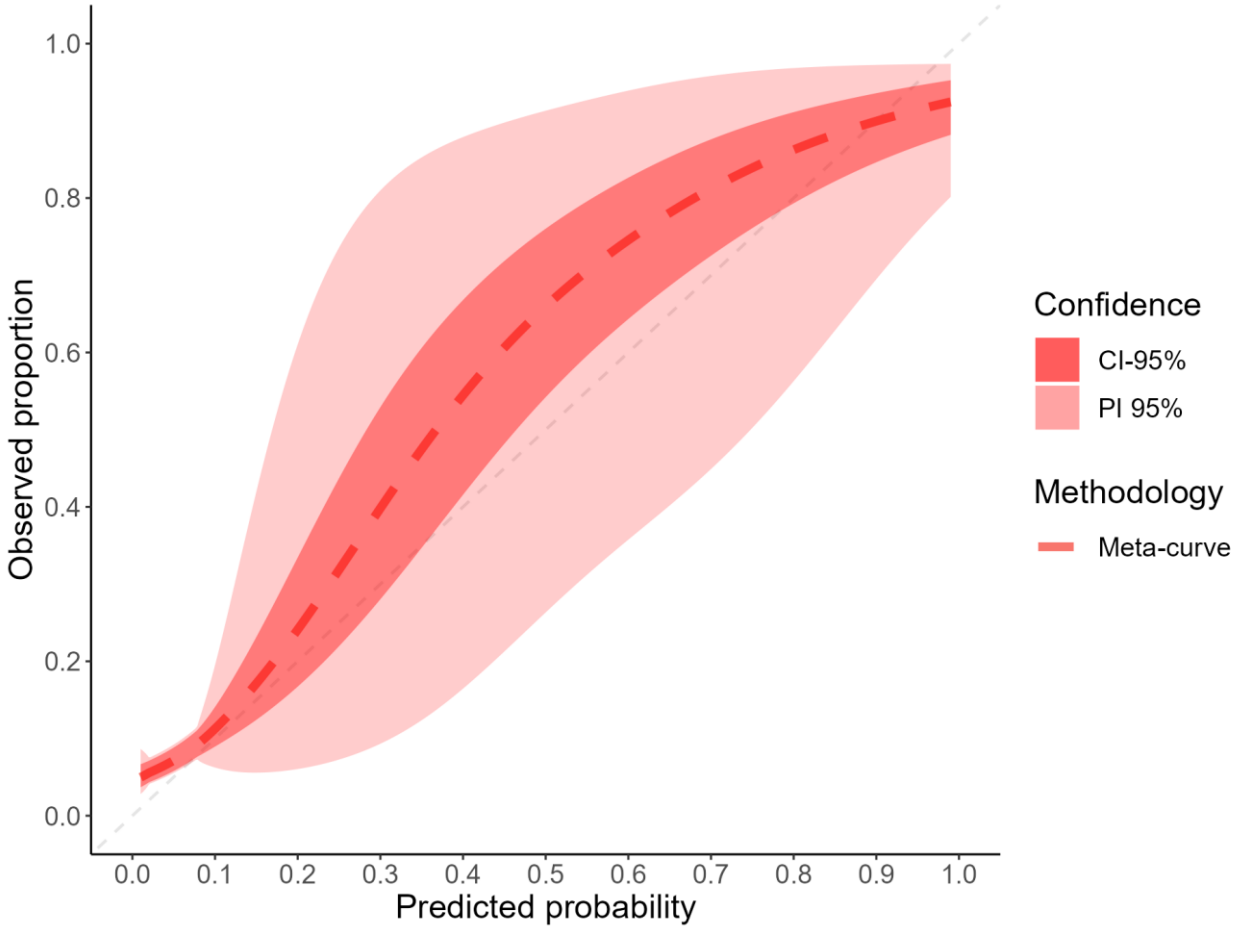
- **Model agnostic approach.**
- **Easy to compute and explain.**
- **Very dependant on number of quantiles.**
- **Can pool information of very different observations.**
- **Not possible to obtain center-specific curves.**

Riley et al., BMJ, 2011


Methodology
Center-specific
Traditional grouped
Heterogeneity
95% CI
95% PI
Ntiles = 4
Ntiles = 6
Ntiles = 8
Ntiles = 10
Ntiles = 12
Ntiles = 14
Ntiles = 16
Ntiles = 18
Ntiles = 20
Observed proportion
Predicted risk

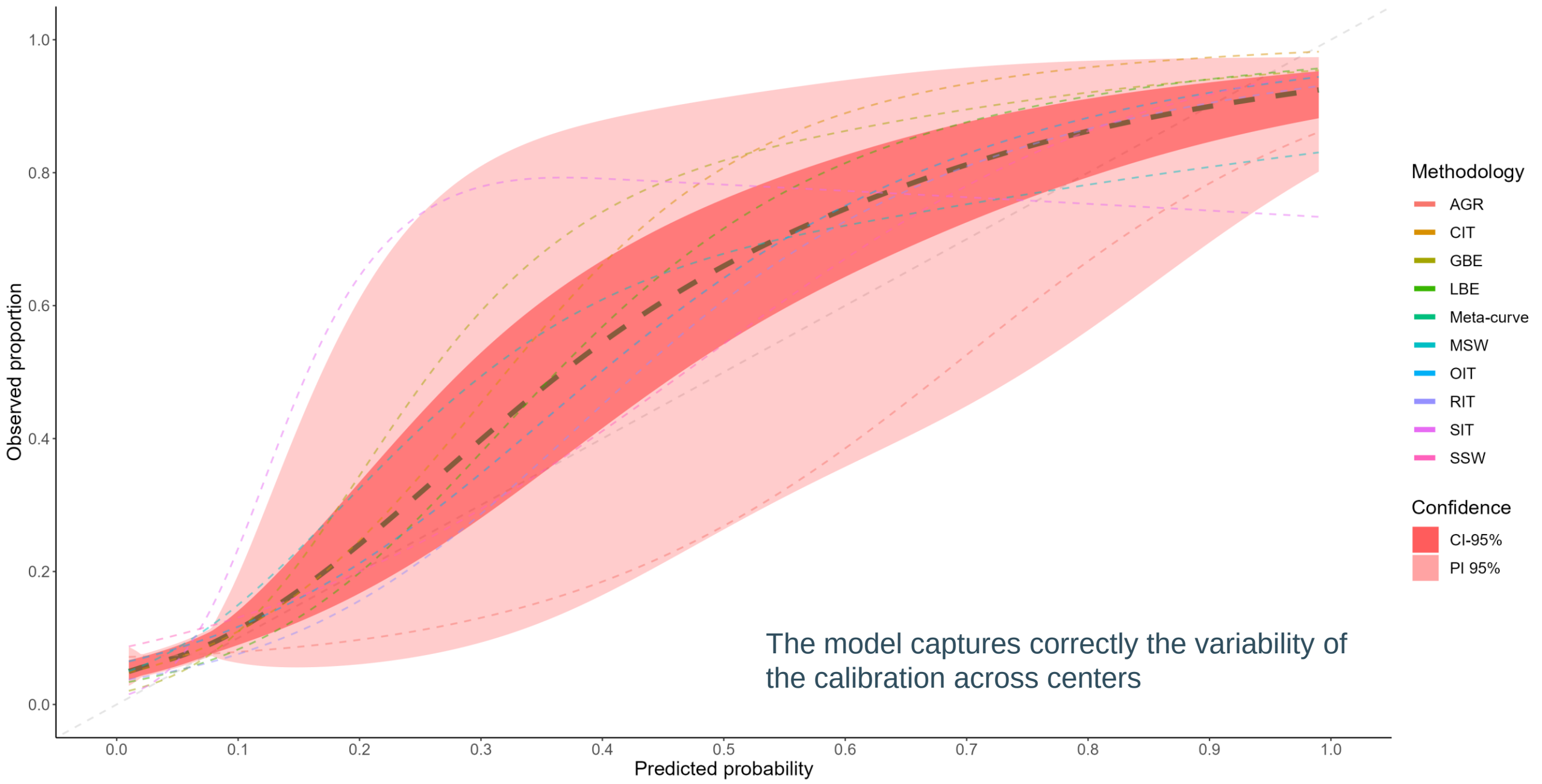
KU LEUVEN

Two-stage meta-analysis (Splines)



1. We first fit a flexible model for each center.
2. We predict the risk for a grid from 0 to 1 in each center.
3. We pool each point in the grid using random effects meta-analysis.
4. We calculate CI with the correction from HKSJ and PI based on Riley (2011)

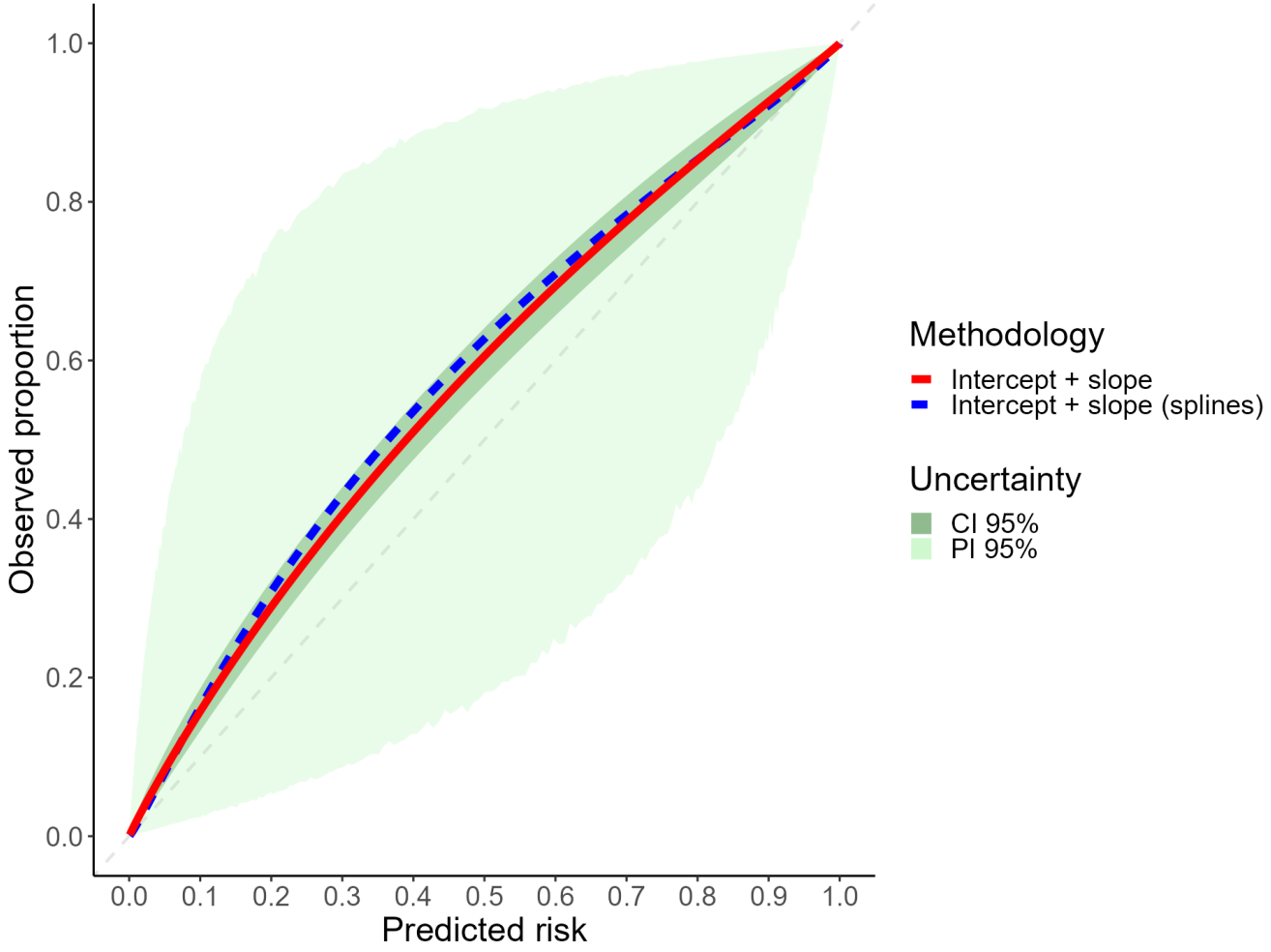
- **Small centers are accounted for but do not harm the curve.**
- **RE based confidence and prediction intervals.**
- **Depends on the model fitted in each center.**
- **Does not estimate center specific curves.**



The model captures correctly the variability of the calibration across centers

KU LEUVEN

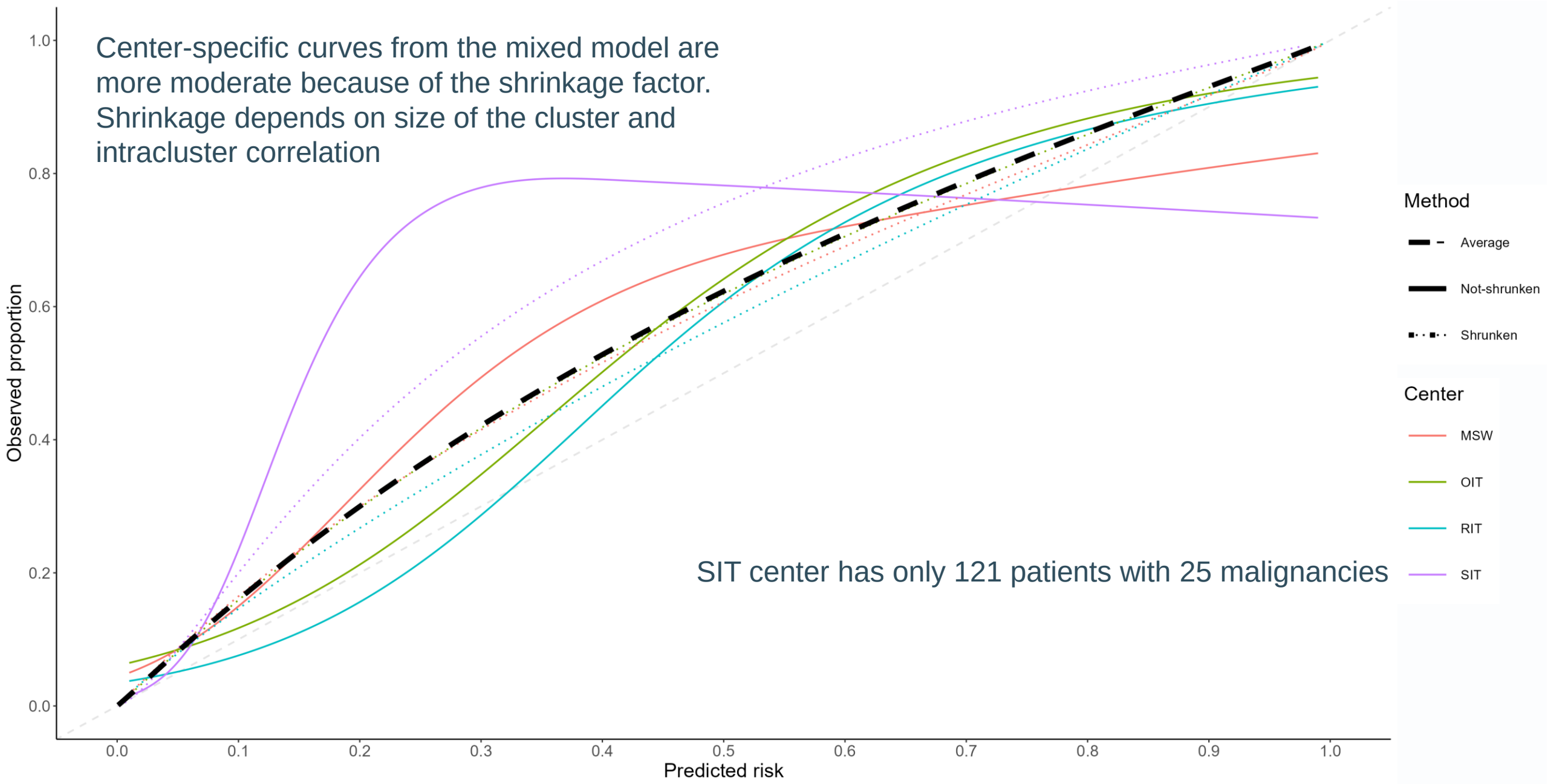
One step mixed model



- Linear mixed effects model (LME) with random intercept or slopes per center.
- Simulation based confidence intervals.

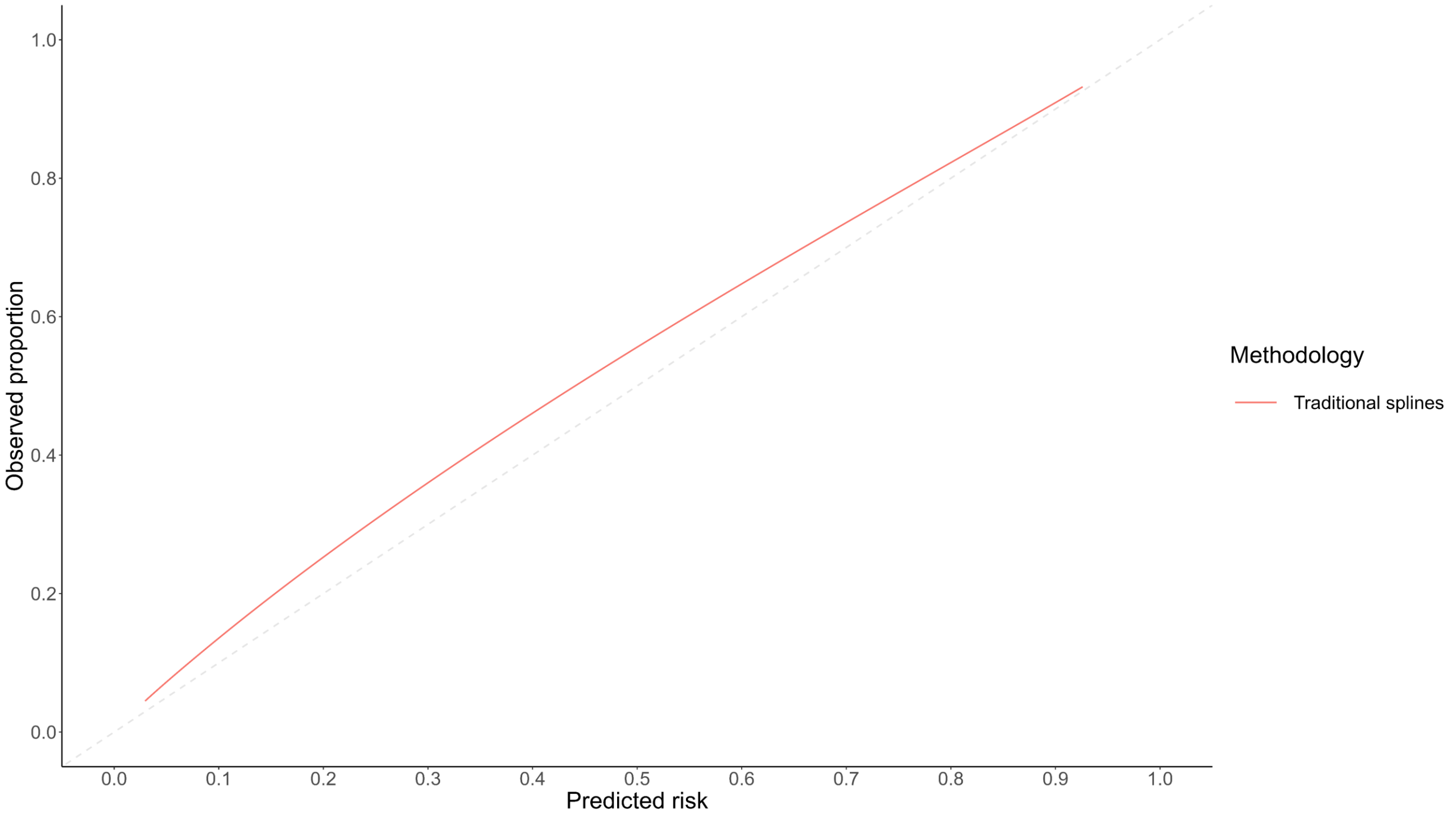
- **We can obtain shrunken center specific curves.**
- **Provides accurate calibration for centers in the validation and average center.**
- **Not model agnostic.**
- **Computationally costly.**

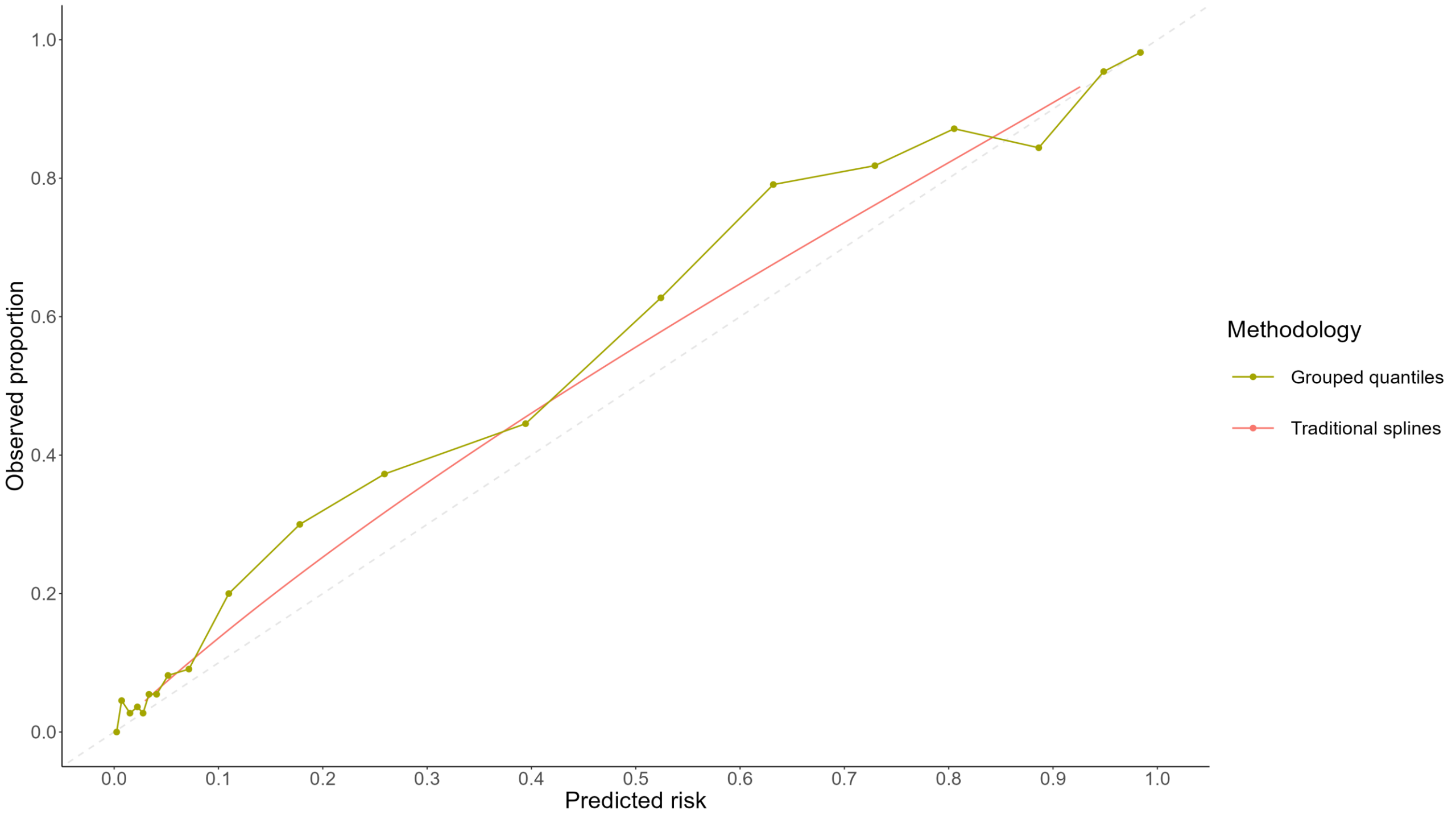
Center-specific curves from the mixed model are more moderate because of the shrinkage factor. Shrinkage depends on size of the cluster and intracluster correlation

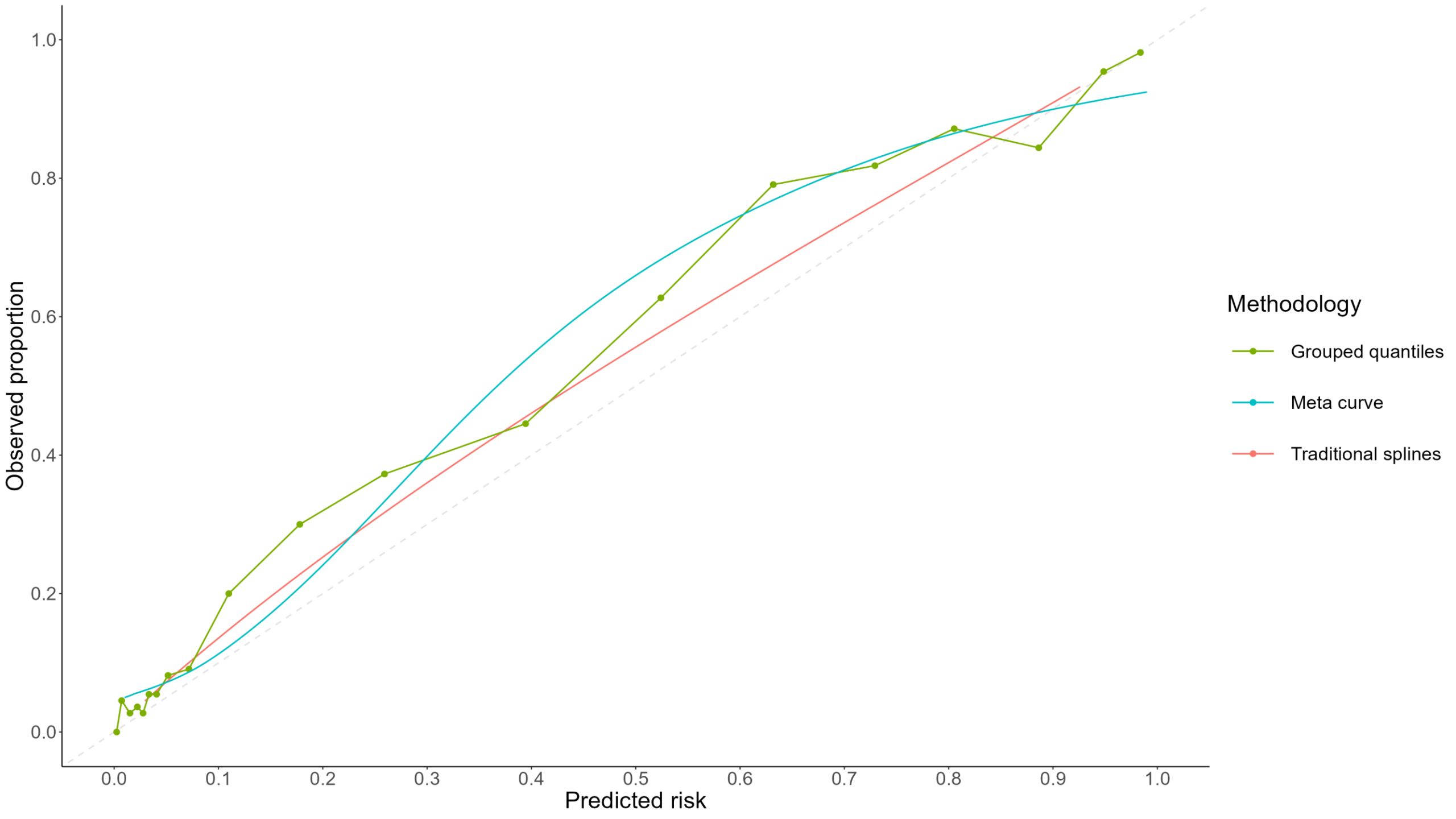


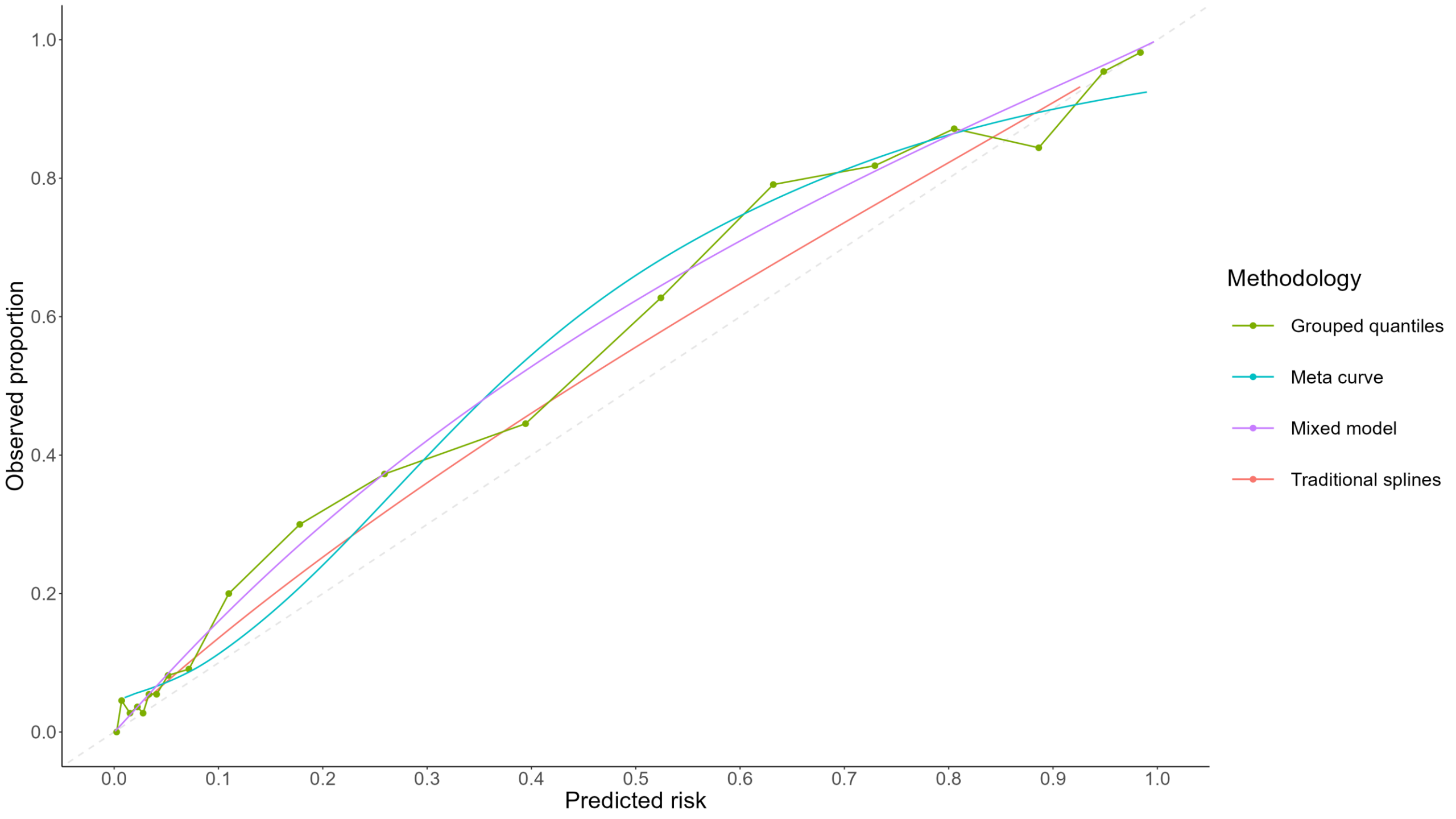
SIT center has only 121 patients with 25 malignancies

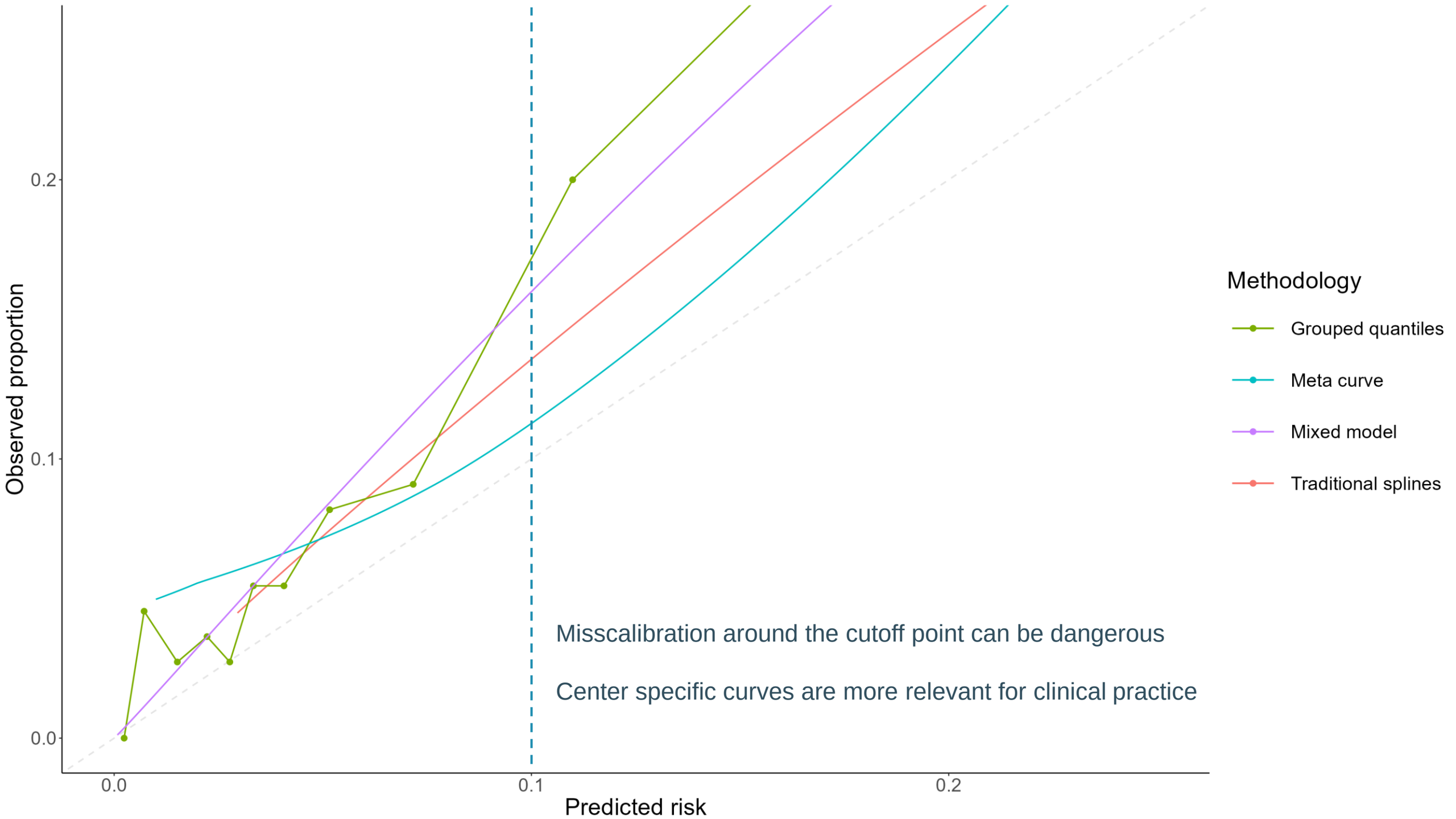
KU LEUVEN

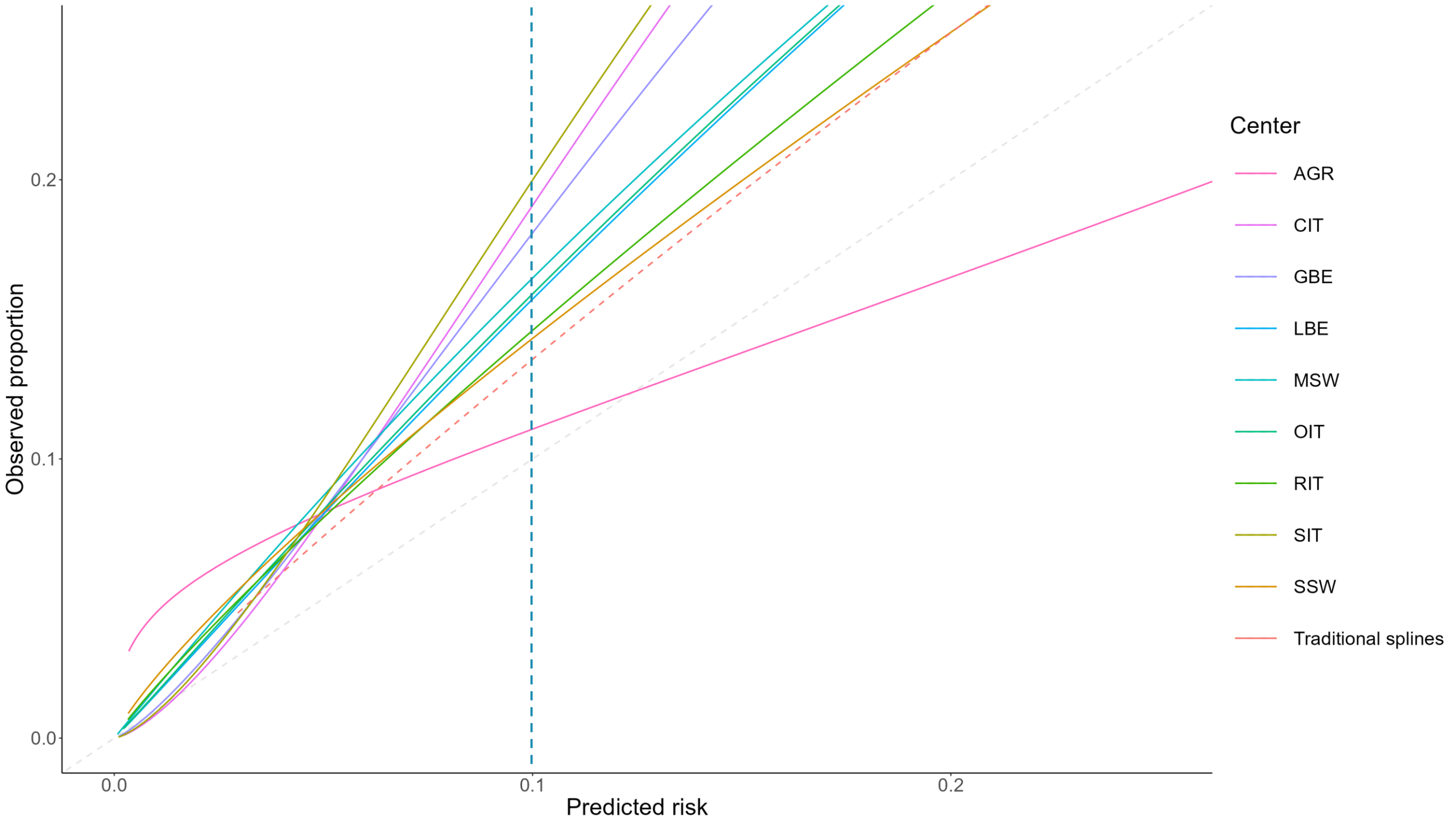

Observed proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predicted risk
Methodology
Traditional splines


Observed proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predicted risk
Methodology
Grouped quantiles
Traditional splines


Observed proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predicted risk
Methodology
Grouped quantiles
Meta curve
Traditional splines


Observed proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predicted risk
Methodology
Grouped quantiles
Meta curve
Mixed model
Traditional splines


Observed proportion
0.0
0.1
0.2
0.0
0.1
0.2
Predicted risk
Misscalibration around the cutoff point can be dangerous
Center specific curves are more relevant for clinical practice
Methodology
Grouped quantiles
Meta curve
Mixed model
Traditional splines


Observed proportion
0.0
0.1
0.2
0.0
0.1
0.2
Predicted risk
Center
AGR
CIT
GBE
LBE
MSW
OIT
RIT
SIT
SSW
Traditional splines

Overview and future

Method	Estimation of observed proportion	Strengths	Limitations
Clustered Quantile	Bivariate random effects meta-analysis of mean predicted risk and event fraction by quantile per cluster.	Model agnostic Random effects based confidence and prediction interval. All clusters present at all groups.	Groups can contain observations with very disperse predicted risks. Curves depend on number of groups
Two-stage meta-analysis	Random effects meta-analysis of estimated smooth observed proportion by cluster	Random effects based confidence and prediction interval.	Two steps process. Curve dependant on the smoother used in the center-specific models.
One Step Random slope	Random intercept and slope mixed model.	Random effects based confidence and prediction interval. Provides also shrunken curves per center. Possible to add splines	Computation time Requires estimation of 2 more coefficient per center.

- Simulation study: Varying level of clustering, size of centers, EPV, Number of clusters.
- Compare the model to a true model adding miscalibration to check whether the observed proportion are correctly estimated.
- Add the functions to CalibrationCurves package already available in CRAN.

Key references:

- **Calibration:**
  - Van Calster et al., BMC Med., 2019 → Introduction to calibration.
  - Van Calster et al., J. Clin. Epidemiol., 2016 → Deep explanation of 4 levels of calibration with simulations.
- **Random effects:**
  - Riley et al., BMJ, 2011 → How to interpret random effects meta-analysis.
  - Borenstein et al., Res. Synth. Methods, 2010 → Introduction to fixed/random effects meta-analysis.
  - Riley et al., Res. Synth. Methods, 2023 → Two or one stage meta-analysis.
- **Clustering:**
  - Debray et al., BMJ, 2023 → Reporting guidelines for clustered prediction models (TRIPOD-C).
  - Riley et al., Stat. Med., 2020 → Overview of mixed models in clustered settings.
  - Wynants et al., Diagn. progn. res., 2019 → Potential of accounting for clustering.
  - Wynants et al., Stat. Med., 2018 → Effects of ignoring clustering in prediction models.

Thank you for your attention

@BarrenadaLasai
lasai.barrenadataleb@kuleuven.be
KU LEUVEN