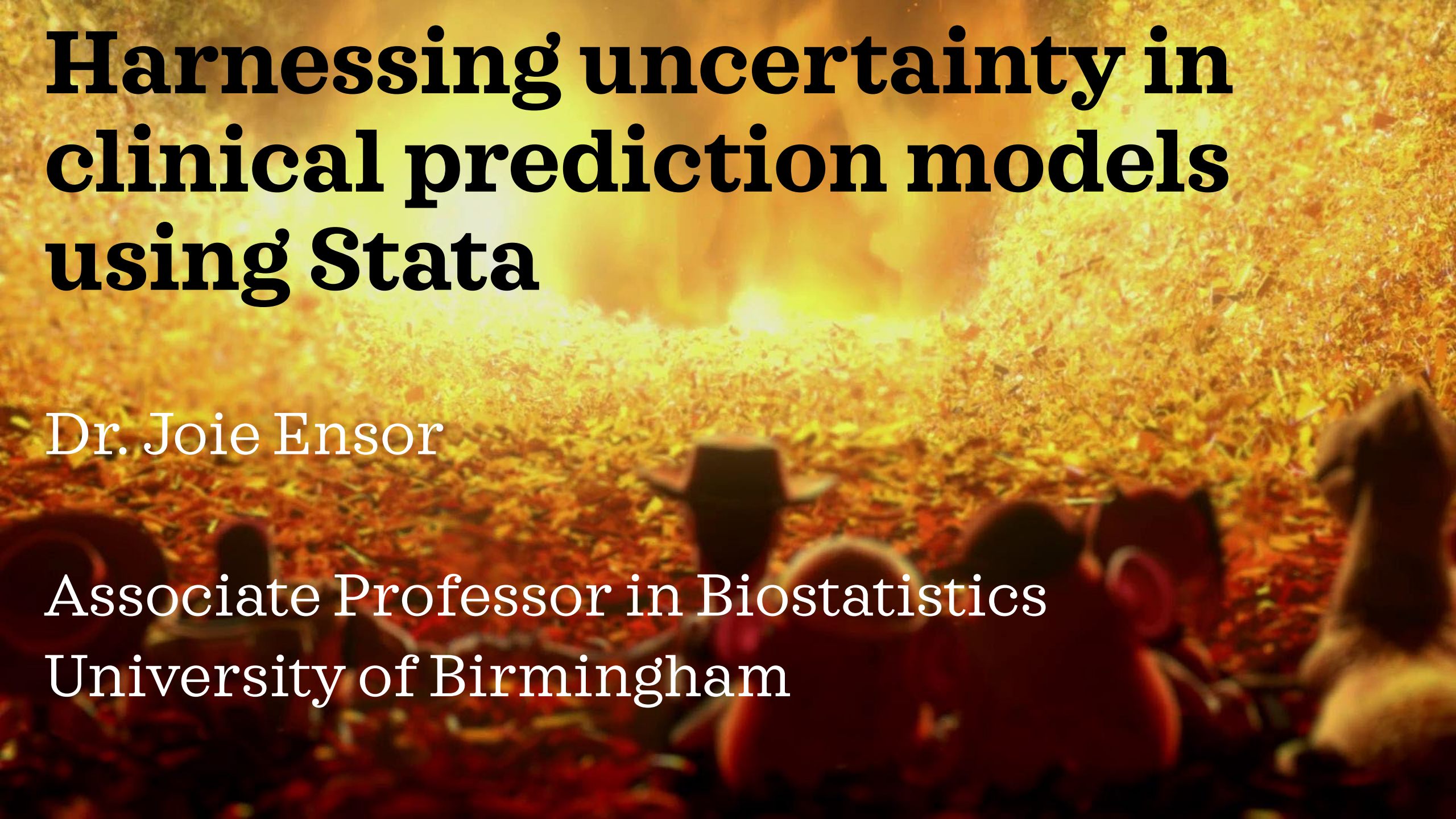
Harnessing uncertainty in clinical prediction models using Stata



Dr. Joie Ensor

Associate Professor in Biostatistics
University of Birmingham


~~clinical prediction~~ models


models

Why?

- Clinical prediction models aim to inform individual patient care

$$model \times you = your\ probability\ of\ outcome$$

- Unreliable models can lead to critical mistakes in clinical decision-making

Most published models are not fit for purpose

- Models are products of their development data

- Predictions become unreliable when:
  - Development samples are too small
  - Model complexity is large relative to outcome events
  - No adjustment for overfitting

Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biometrical Journal. 2023

Sample size is a fundamental issue

- 73% of published models use inadequate sample sizes
- Median deficit of 387 patients, IQR (-1207 to +49) (Dhiman et al.)
- Mann et al. found only 10% of studies reported a sample size calculation

Dhiman P, Ma J, Qi C, Bullock G, Sergeant JC, Riley RD, Collins GS. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. BMC Medical Research Methodology. 2023

Mann M, Collins G, Riley R, Ensor J. (2024) How are published clinical prediction model studies assessing and reporting calibration performance at external validation? doi: 10.17605/OSF.IO/74R9U


models

THE MULTIVERSE OF MODELS

- Model instability: different samples → markedly different models

- Any single model = merely one example from a multiverse of possible models

Riley RD, Pate A, Dhiman P, Archer L, Martin GP, Collins GS. Clinical prediction models and the multiverse of madness. BMC medicine. 2023

Visualising Model Instability

Predictors	Large sample	Development samples of N=100									
		1	2	3	4	5	6	7	8	9	10
Age	**1.15**	**1.63**	X	**1.65**	X	X	X	X	X	X	X
Sex	X	X	X	X	X	X	X	X	X	**0.22**	X
Systolic BP	**0.99**	X	X	X	X	X	**0.96**	X	X	X	**0.96**
Bicarbonate	**0.95**	X	X	X	**0.79**	X	X	X	X	X	X
Creatinine	**3.28**	X	X	**53.45**	X	X	**39.77**	X	X	**30.25**	**10.03**
Hemoglobin	**0.88**	X	X	X	X	X	X	X	X	X	X
Platelets	**1.00**	X	**0.99**	X	**0.99**	X	X	X	X	X	X
Potassium	**1.66**	X	**6.68**	X	X	**3.09**	X	**3.86**	**5.24**	**2.74**	X
Blood oxygen	X	X	X	X	X	X	X	X	X	X	X

Riley RD, Ensor J et al. The importance of sample size on the quality and utility of AI-based prediction models for healthcare. Forthcoming – The Lancet Digital Health 2025.

Visualising Model Instability

- Same modelling approach on different samples produces variation in models → instability in predicted probabilities

- Stability checks should become standard practice

Predictors	Large sample	Development samples of N=100									
		1	2	3	4	5	6	7	8	9	10
Age	**1.15**	**1.63**	X	**1.65**	X	X	X	X	X	X	X
Sex	X	X	X	X	X	X	X	X	X	**0.22**	X
Systolic BP	**0.99**	X	X	X	X	X	**0.96**	X	X	X	**0.96**
Bicarbonate	**0.95**	X	X	X	**0.79**	X	X	X	X	X	X
Creatinine	**3.28**	X	X	**53.45**	X	X	**39.77**	X	X	**30.25**	**10.03**
Hemoglobin	**0.88**	X	X	X	X	X	X	X	X	X	X
Platelets	**1.00**	X	**0.99**	X	**0.99**	X	X	X	X	X	X
Potassium	**1.66**	X	**6.68**	X	X	**3.09**	X	**3.86**	**5.24**	**2.74**	X
Blood oxygen	X	X	X	X	X	X	X	X	X	X	X

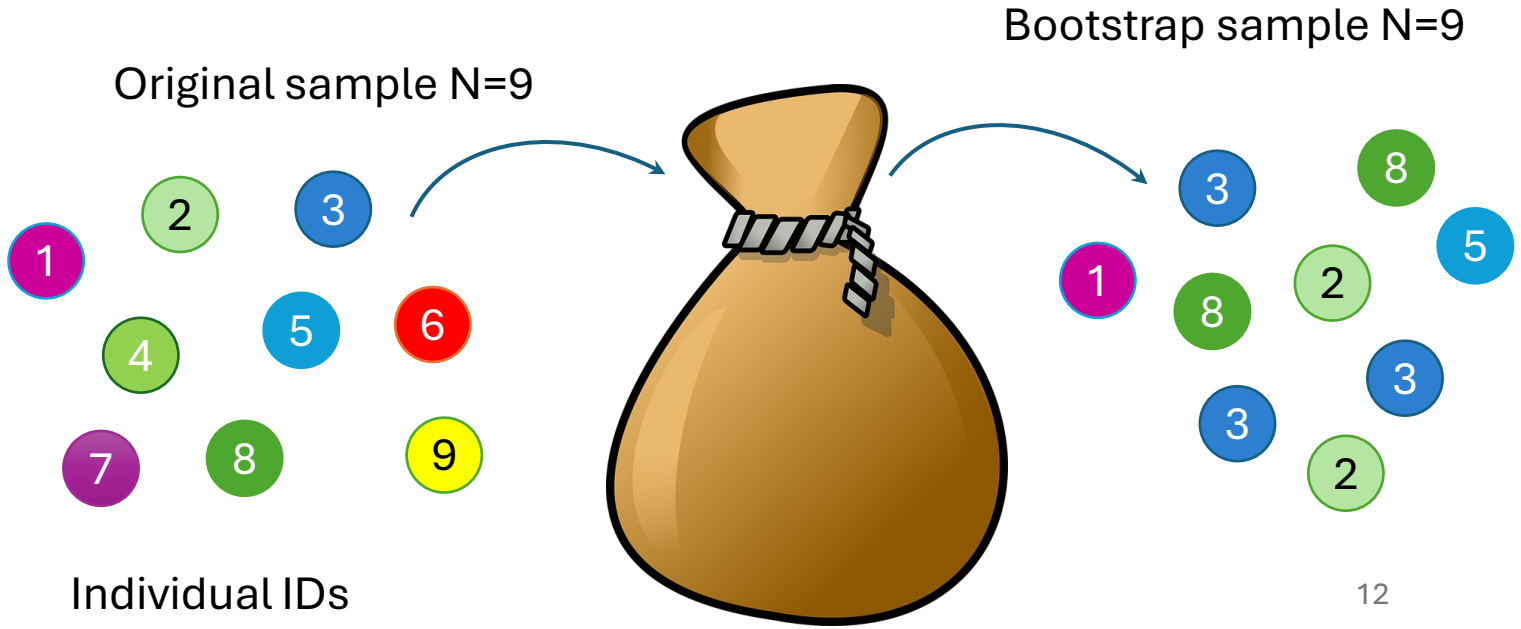
Today’s roadmap

- Internal validation for uncertainty visualisation
- Using uncertainty metrics in study design
- By the end of this talk:
  - Practical workflow for your next model study
  - Tools to assess and address model instability

Internal validation concept

- Bootstrap resampling examines impact of sampling variability

- Must match:
  - Original sample size (N)
  - Exact modelling approach



Clinical example

- **Prediction target**: Acute kidney injury in intensive care patients (48hr window)
- **Data source**: MIMIC-III database
- **Approach**: Fixed set of predictors based on evidence and clinical consensus

Logistic regression

Number of obs = **286**
LR chi2(**8**) = **12.06**
Prob > chi2 = **0.1485**
Pseudo R2 = **0.0422**

Log likelihood = **-136.80736**

AKI	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
gender	.2078414	.3450336	0.60	0.547	-.4684121	.8840949
bicarbonate_mean	-.0574709	.0439213	-1.31	0.191	-.1435552	.0286133
creatinine_mean	1.125994	.7757663	1.45	0.147	-.39448	2.646468
hemoglobin_mean	-.0822802	.0815089	-1.01	0.313	-.2420346	.0774743
bun_mean	-.01488	.0184731	-0.81	0.421	-.0510866	.0213266
potassium_mean	.5765495	.3599629	1.60	0.109	-.1289649	1.282064
sysbp_mean	-.0121554	.0107667	-1.13	0.259	-.0332578	.008947
spo2_mean	-.0270447	.0894051	-0.30	0.762	-.2022755	.1481861
_cons	1.82889	9.397801	0.19	0.846	-16.59046	20.24824

Clinical example

- **Prediction target**: Acute kidney injury in intensive care patients (48hr window)
- **Data source**: MIMIC-III database
- **Approach**: Fixed set of predictors based on evidence and clinical consensus

Discrimination statistics ...

	Estimate	SE	Lower_CI	Upper_CI
C-Statistic	0.647	0.039	0.570	0.725
Somers D	0.295	0.079	0.140	0.449

Calibration statistics ...

	Estimate	Lower_CI	Upper_CI
E/O	1.000	0.788	1.335
E-O	-0.000	-0.054	0.050
CITL	0.000	-0.297	0.297
C-Slope	1.000	0.416	1.584

Overall performance statistics ...

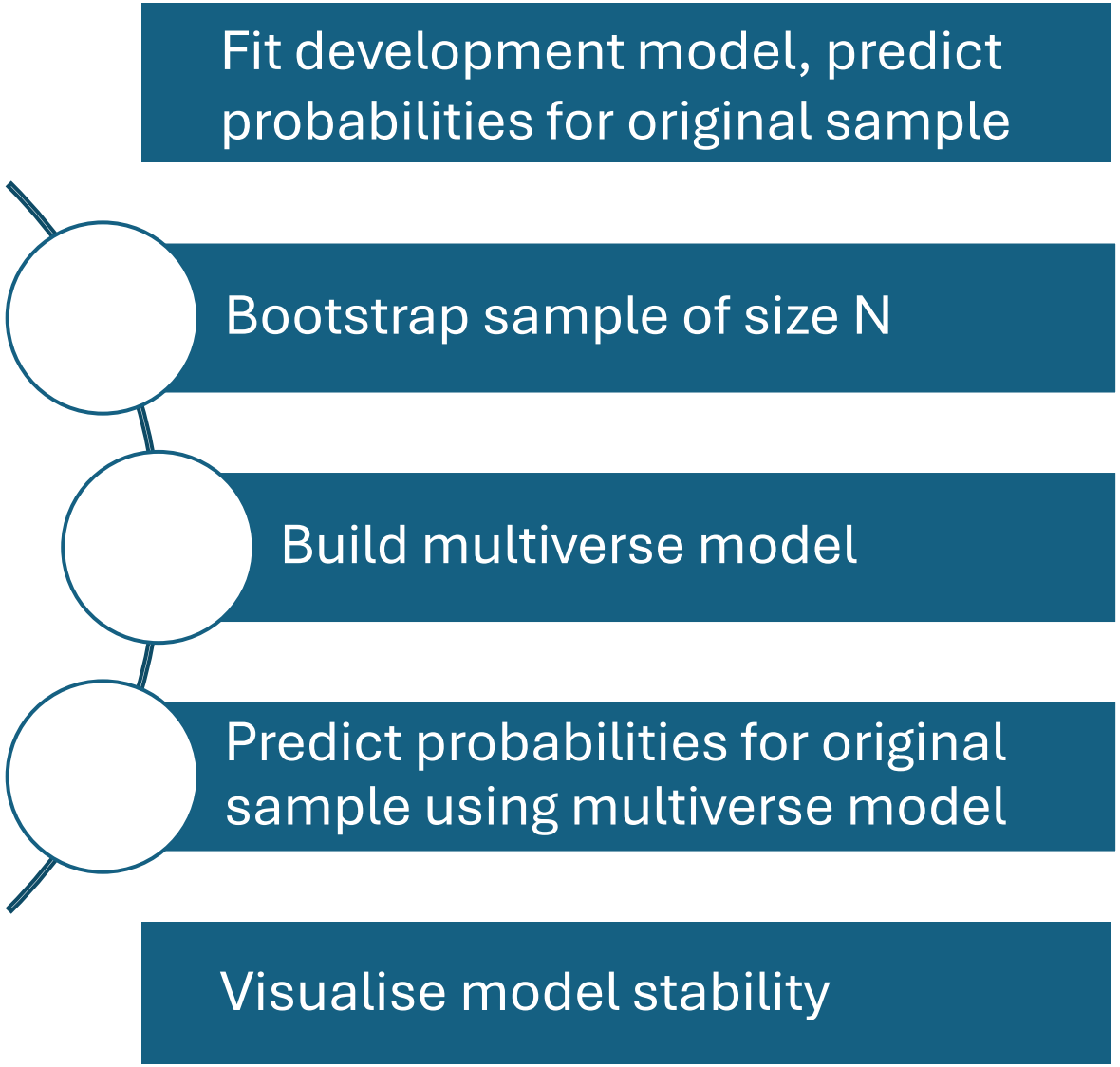
	Estimate	Lower_CI	Upper_CI
Cox-Snell R2	0.041	0.005	0.092
R2 Nagelke~e	0.065	0.008	0.143
R2 McFadde~s	0.042	0.005	0.095
Briers Score	0.153	0.125	0.184

Additional summary statistics ...

	Mean	SD	Median	LQ	UQ
LP Dist	-1.473	0.535	-1.463	-1.831	-1.109
Sample size	286.000	.	.	.	.

`pmintval` overview

- New Stata package for internal validation and uncertainty visualisation
- Returns dataset with predictions from original and bootstrap models
- Examines prediction stability across the model multiverse



`pmintval` syntax

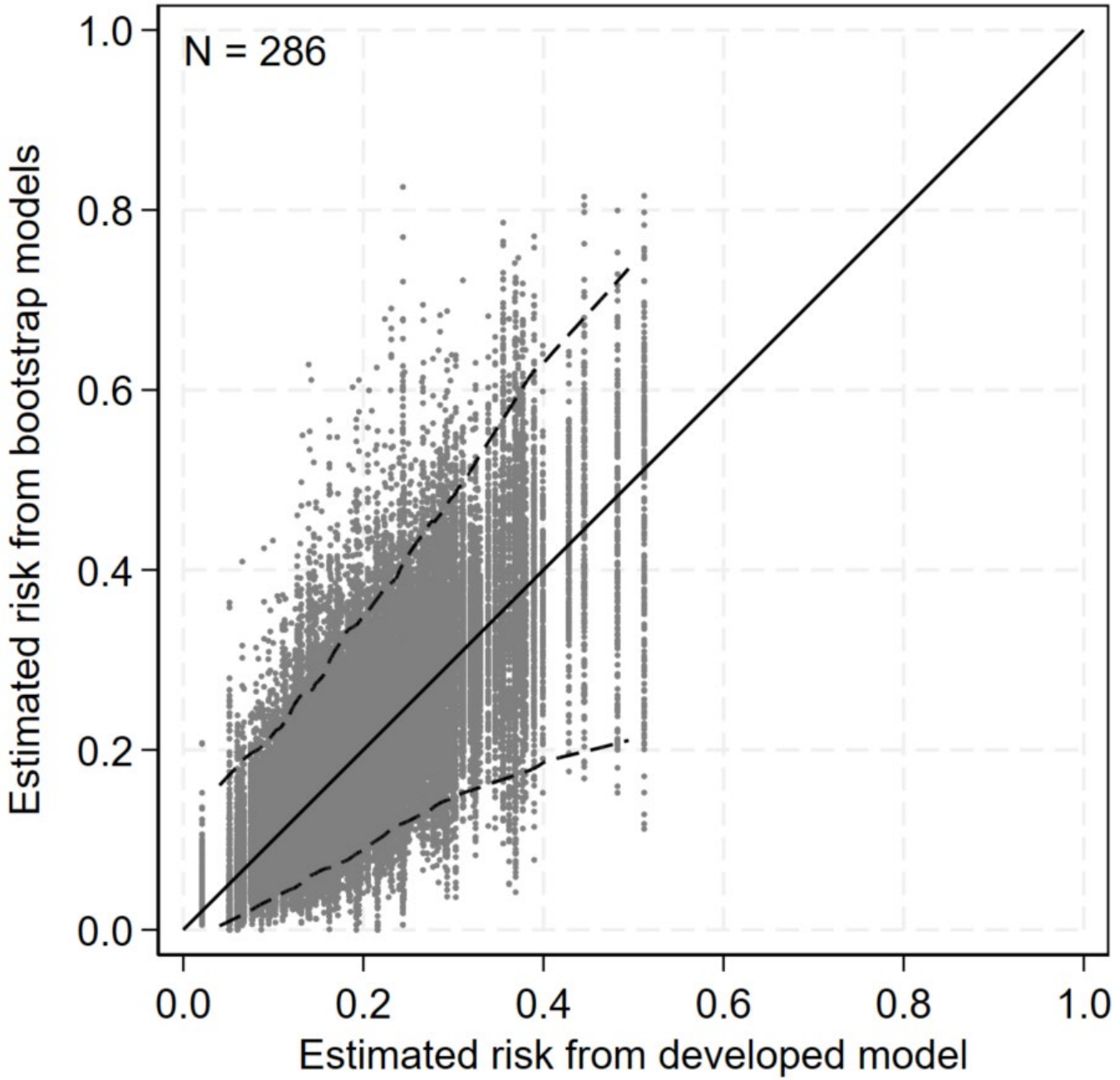
Command line

```
pmintval AKI gender  bicarbonate_mean  creatinine_mean
hemoglobin_mean  bun_mean  potassium_mean  sysbp_mean  spo2_mean,
boot(200)
```

	pr1	pr2	pr3	pr4	pr5	pr6	pr7	pr8	pr9	pr10	pr11	pr12
1	.1452606	.1554341	.1804096	.1664165	.1692847	.1834505	.1475054	.1826395	.1633522	.1548582	.152133	.1794005
2	.0457026	.057397	.0640362	.0417281	.052992	.0525907	.0436598	.043273	.0508877	.0476608	.0467994	.048035
3	.072901	.0884458	.0906838	.074332	.0881848	.079837	.0749723	.066966	.0797636	.0780905	.0743829	.0765029
4	.2357282	.2097052	.2216702	.2302375	.2448104	.2270528	.2043875	.2396001	.2448246	.2210204	.237456	.2033944
5	.1562723	.1653154	.1907796	.166422	.1511118	.1824354	.1745607	.1847174	.1878714	.1736683	.1590132	.189984
6	.2063501	.224724	.2450891	.2315639	.2354372	.247393	.223059	.2513075	.2371853	.2253135	.2178639	.2448219
7	.091362	.1113398	.1182475	.0828711	.1048185	.1035374	.084993	.0891667	.0981059	.0963472	.0909408	.0883589
8	.2109131	.2312233	.2497026	.2680111	.2438476	.251814	.2452061	.2496253	.2565619	.2428534	.2275574	.2689827

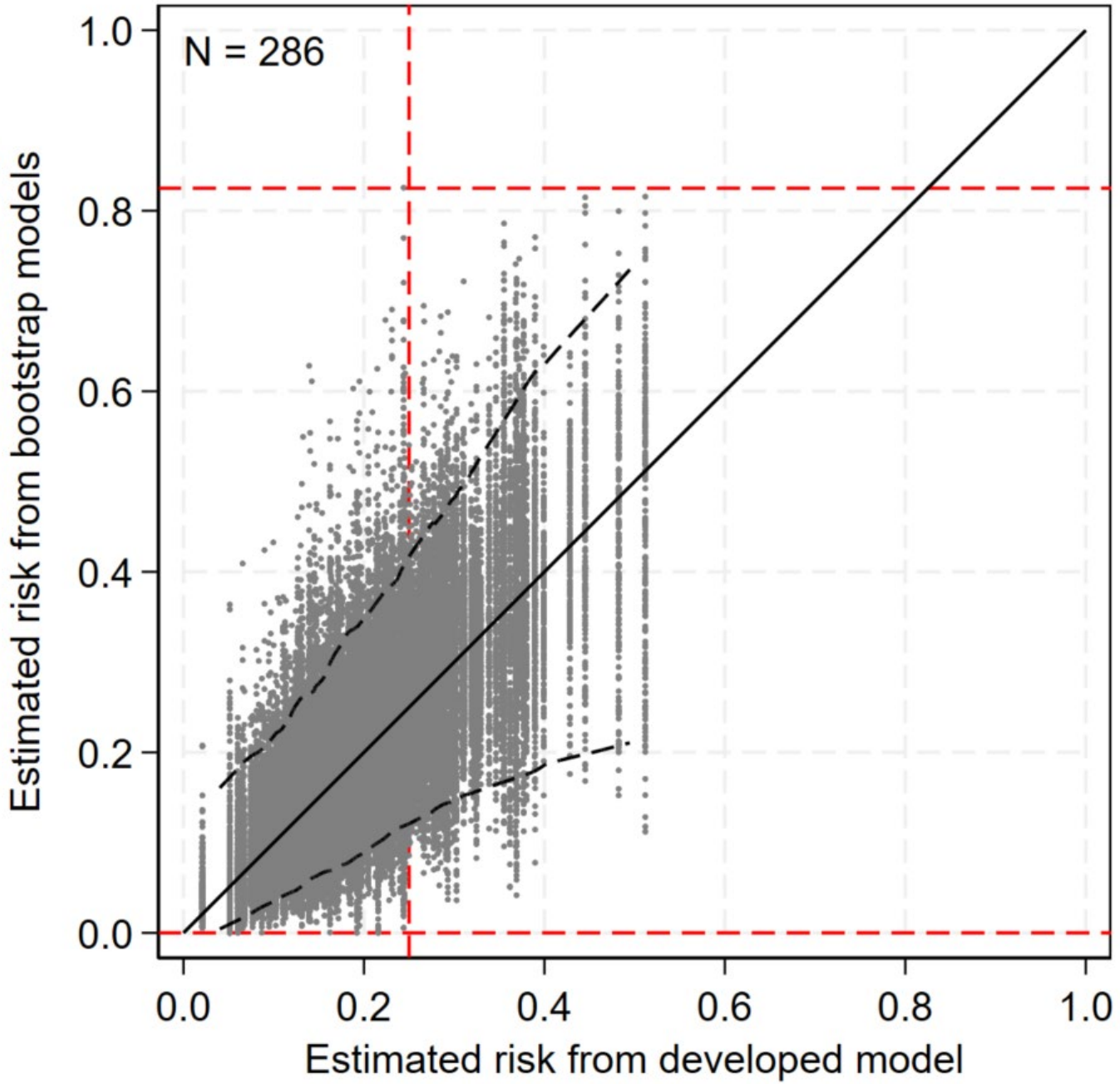
Prediction Instability Visualisation

- Illustration of potential prediction variability for individual patients



Prediction Instability Visualisation

- Illustration of potential prediction variability for individual patients
- For a patient with 25% predicted risk, true risk could range from 0% to 82.5%



Clinical Implications of Instability

- Wide prediction intervals → misleading risk communication
- Unreliable for clinical decision support
- Undermines clinician trust in the model

Option A

Option B

Certainty by design

- Traditional approaches target "average" model performance
  - `pmsampsize` package
    - Minimises overfitting & precisely estimates overall risk

- Our approach: target prediction stability for individual risk estimation

Riley, R. D., Collins, G. S., Whittle, R., Archer, L., Snell, K. I., Dhiman, P., ... & Ensor, J. (2024). A decomposition of Fisher's information to inform sample size for developing fair and precise clinical prediction models - part 1: binary outcomes. arXiv preprint arXiv:2407.09293.

`pmstabilityss` Package - Overview

- Uses information from development sample and internal validation
- Calculates required sample size for specified prediction stability
- Uses a decomposition of Fisher's information
  - Computationally quick

Riley, R. D., Collins, G. S., Whittle, R., Archer, L., Snell, K. I., Dhiman, P., ... & Ensor, J. (2024). A decomposition of Fisher's information to inform sample size for developing fair and precise clinical prediction models - part 1: binary outcomes. arXiv preprint arXiv:2407.09293.

`pmstabilityss` Package - syntax

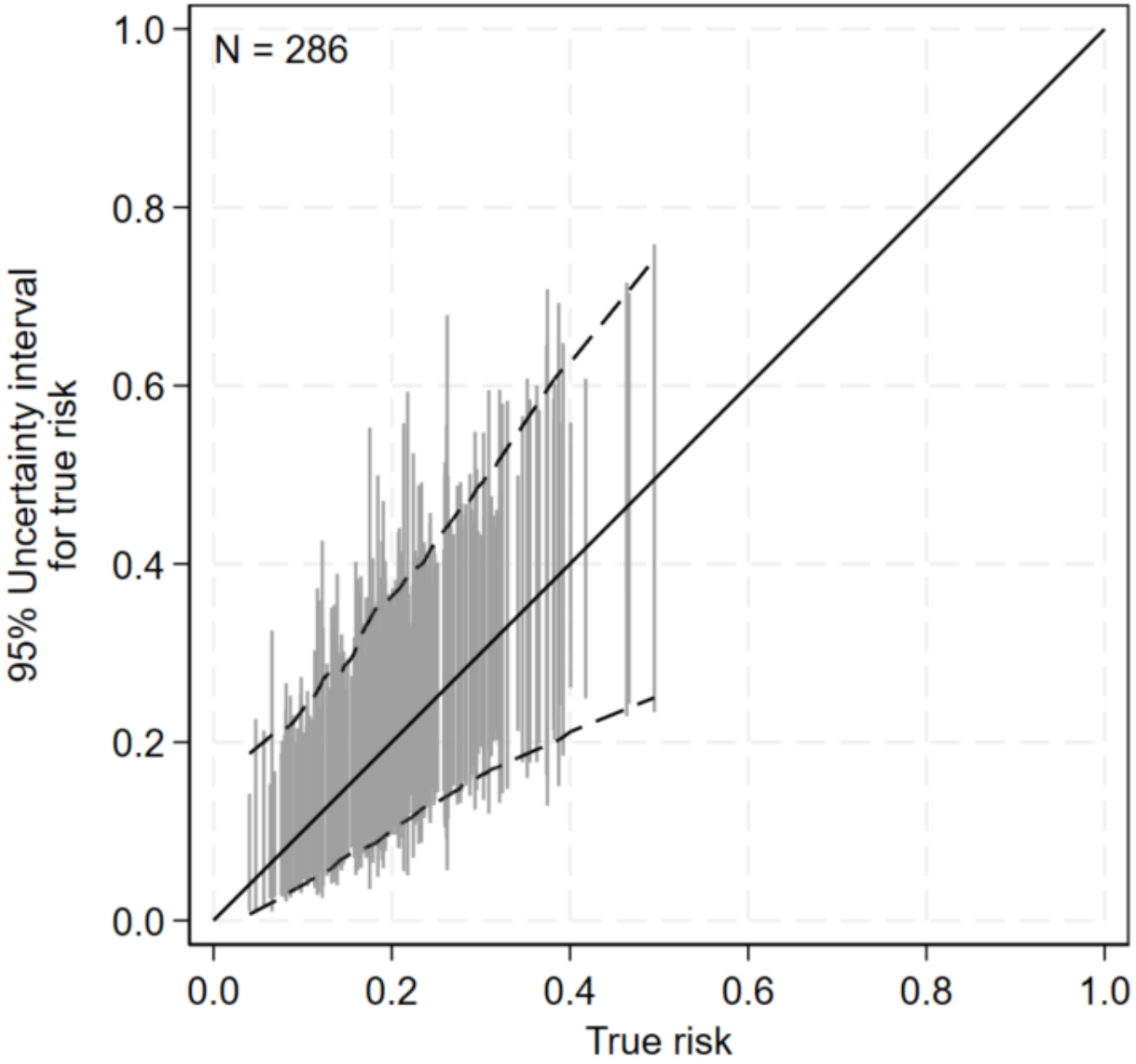
Command line

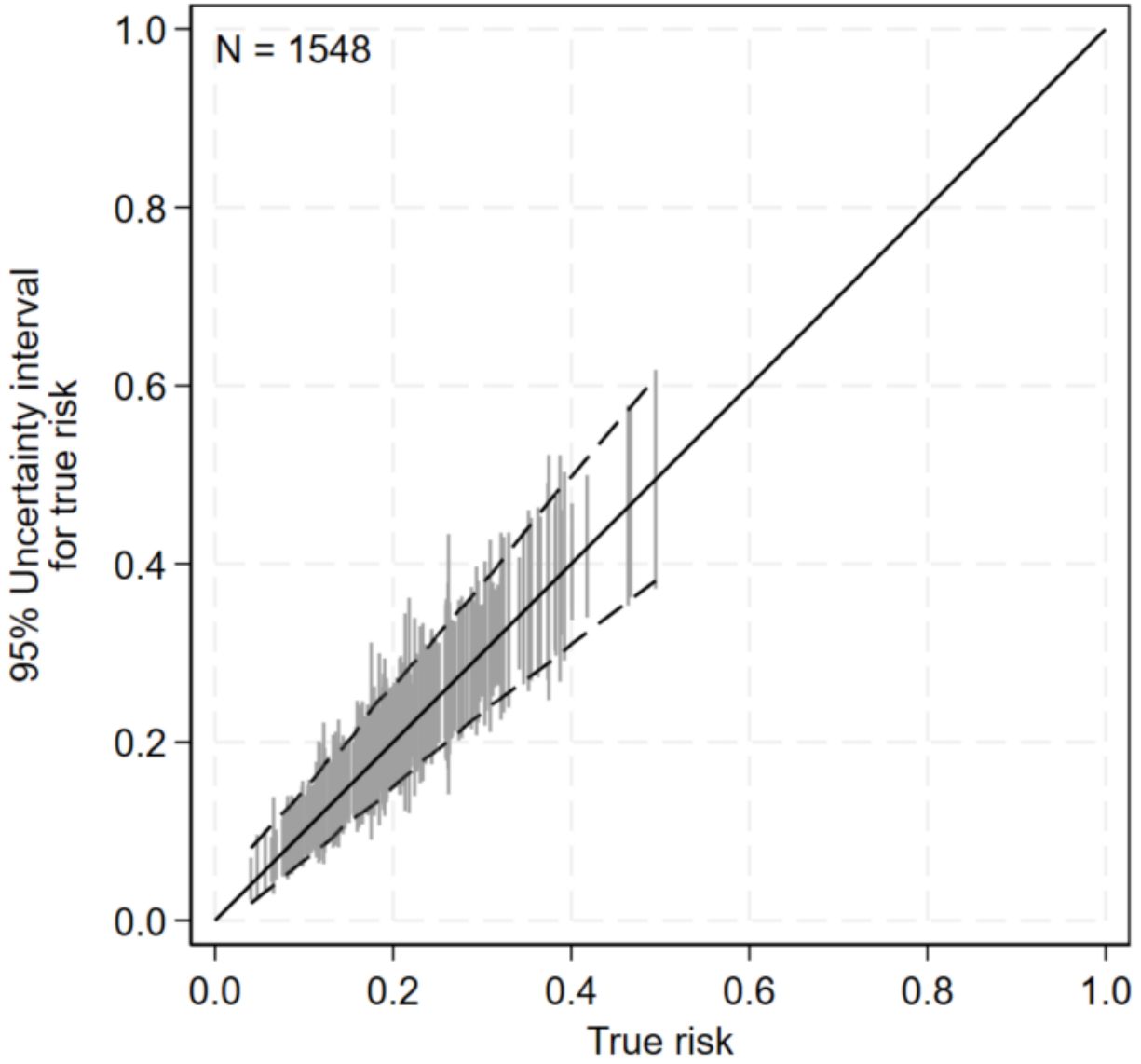
```
pmstabilityss gender  bicarbonate_mean  creatinine_mean
hemoglobin_mean  bun_mean  potassium_mean  sysbp_mean  spo2_mean,
prev(.1993) cstat(0.65) pmss lp(lin_pred)
```

```
Fixed SS of input dataset = 286

Minimum SS required by pmsampsize = 1548


Overall summary UI widths
```

N	Mean	Min	Median	Max
286	**.26**	**.13**	**.25**	**.62**
1548	**.11**	**.048**	**.1**	**.29**

Visualising Stability at Different Sample Sizes





Options to customise stability targets

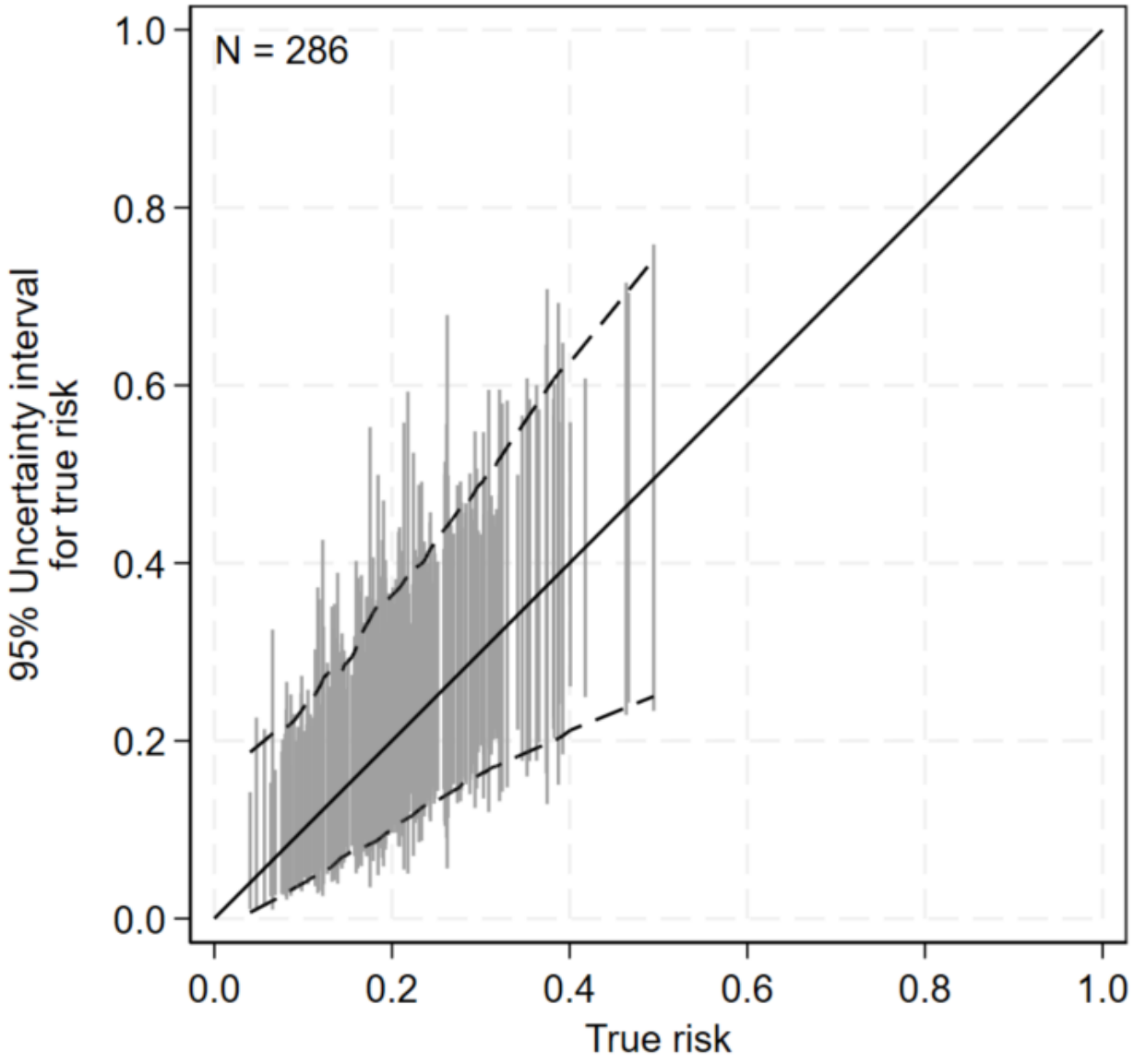
- `pciwidth()`: target acceptable prediction interval widths

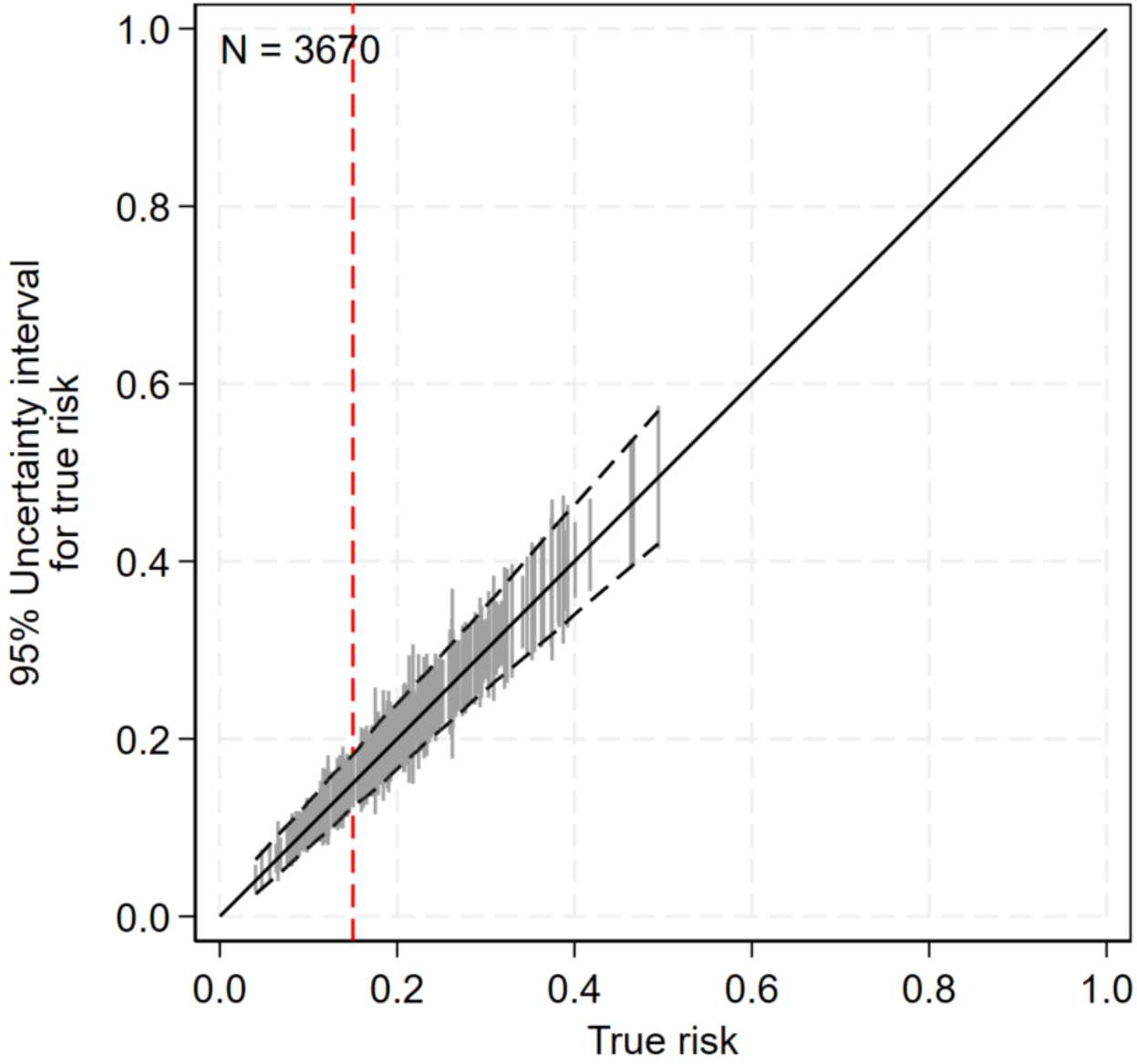
- `pcutpoints()`: cutpoints in the interval [0,1] corresponding with target widths specified

Command line

```
pmstabilityss gender  bicarbonate_mean  creatinine_mean
hemoglobin_mean  bun_mean  potassium_mean  sysbp_mean  spo2_mean,
prev(.1993) cstat(0.65) lp(lin_pred) pcut(.15 1) pci(.1 .25)
```

Specifying Stability Requirements





Specifying Stability Requirements

Fixed SS of input dataset = 286

Minimum SS required to meet target UI widths = 3670

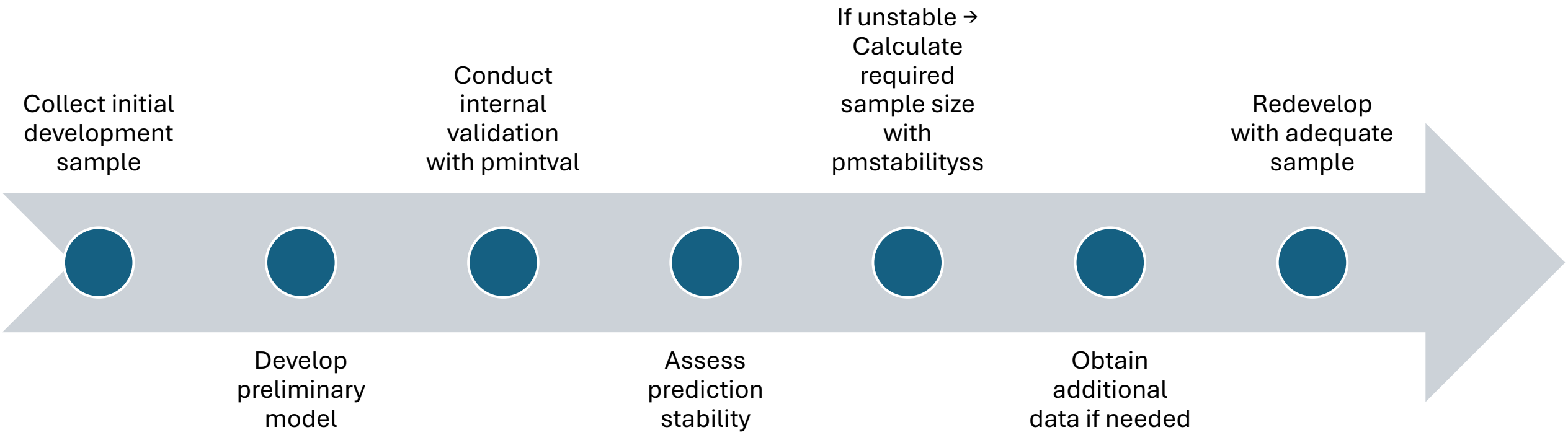
Overall summary UI widths

N	Mean	Min	Median	Max
286	**.26**	**.13**	**.25**	**.62**
3670	**.073**	**.03**	**.067**	**.19**

Summary UI widths by probability categories

N	P category	Target width	Mean	Min	Median	Max	Prop target width met
286	**.15**	**.1**	**.2**	**.13**	**.19**	**.4**	**.26**
286	**1**	**.25**	**.29**	**.15**	**.27**	**.62**	**.26**
3670	**.15**	**.1**	**.054**	**.03**	**.049**	**.1**	**1**
3670	**1**	**.25**	**.083**	**.041**	**.075**	**.19**	**1**

Complete workflow



Key takeaways

- Quantifying model uncertainty is essential for clinical models
- Balance needed between stability requirements and feasibility
- If a model cannot be developed with reasonable stability → reconsider its use

Future developments

- Upcoming features:
  - Time-to-event and continuous outcome support
  - Prospective study design with minimal inputs

- Tutorial and documentation available soon

Questions?

- Email: j.ensor@bham.ac.uk
- BlueSky: @joieensor.bsky.social
- X: @joie_ensor
- GitHub: https://github.com/JoieEnsor/pm-suite