First-year lectures in Statistics: BMathSci(Adv)

Professor Patty Solomon

School of Mathematical Sciences
University of Adelaide

First-year lectures in Statistics: BMathSci(Adv)

- The purpose of these three (now two) lectures is to introduce you to some **real statistical applications**.
- I'll do this by talking about

1. Institutional comparisons

2. Statistics and people smuggling

Topic 1:
Institutional comparisons

- Comparing school performance.
- Comparing university performance.
- Comparing **hospital and intensive care unit performance.**

Comparing the performance of Australian and New Zealand intensive care units

ICU bedside area, the Queen Elizabeth Hospital, Adelaide

My clinical colleague: Dr John Moran, TQEH

My ex-postdoc: Dr Jessica Kasza

Comparing institutional performance is

- a statistically challenging problem
- usually done **badly**
- usually done using **league tables.**

A world plagued by league tables

LONDON 2012 MEDAL TALLY

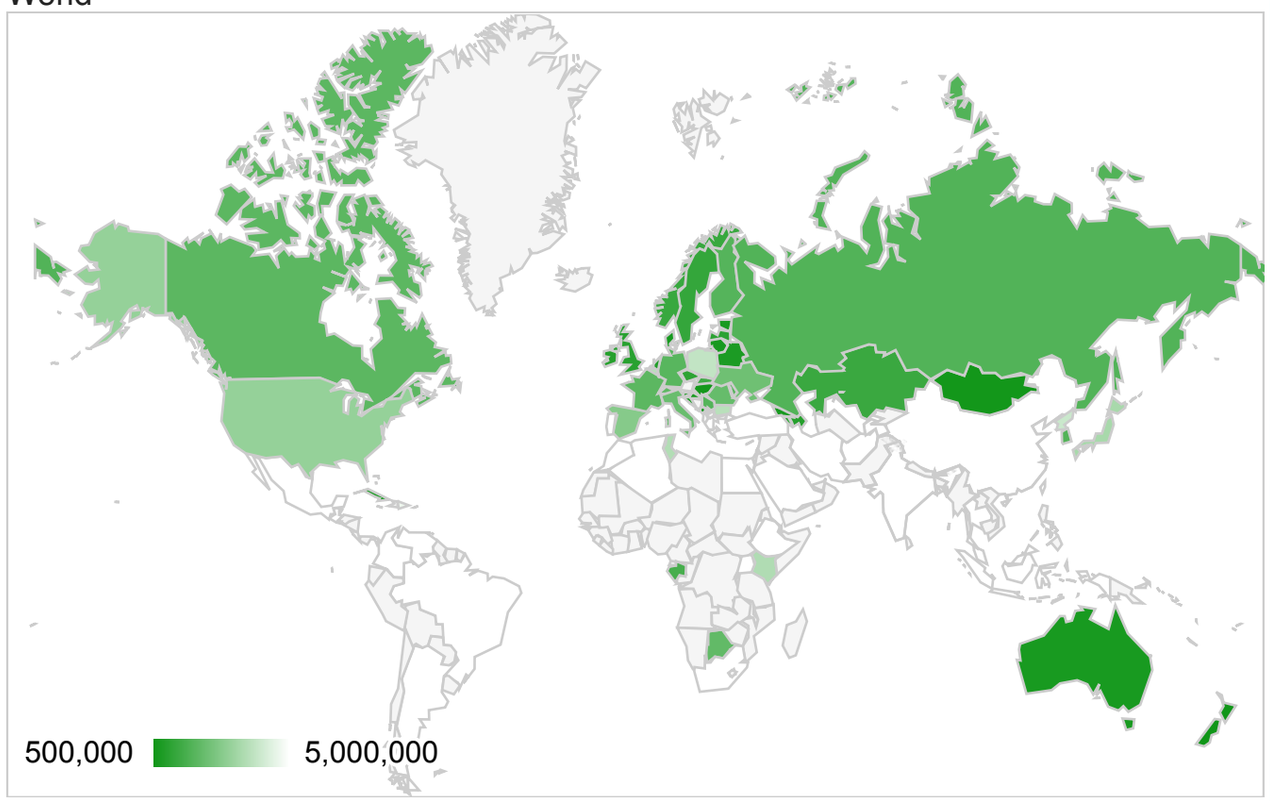
RANK	COUNTRY				TOTAL
1	United States of America	46	29	29	104
2	China	38	27	23	88
3	Great Britain	29	17	19	65
4	Russia	24	26	32	82
5	South Korea	13	8	7	28
6	Germany	11	19	14	44
7	France	11	11	12	34
8	Italy	8	9	11	28
9	Hungary	8	4	5	17
10	Australia	7	16	12	35
11	Japan	7	14	17	38
12	Kazakhstan	7	1	5	13
13	Netherlands	6	6	8	20
14	Ukraine	6	5	9	20
15	New Zealand	6	2	5	13
16	Cuba	5	3	6	14

Choose the right performance indicator …

MEDALS PER CAPITA

Olympic Glory in Proportion

Rank	Country	Medals	Population	Population per Medal
1	Grenada	1	110,821	110,821
2	Jamaica	12	2,705,827	225,485
3	Trinidad and Tobago	4	1,317,714	329,428
4	New Zealand	13	4,432,620	340,970
5	Bahamas	1	353,658	353,658
6	Slovenia	4	2,057,540	514,385
7	Mongolia	5	2,736,800	547,360
8	Hungary	17	9,962,000	586,000
9	Montenegro	1	620,029	620,029
10	Denmark	9	5,580,516	620,057
11	Georgia	7	4,469,200	638,457
12	Lithuania	5	3,192,800	638,560
13	Australia	35	22,880,619	653,731



http://www.medalspercapita.com/#medals-per-capita:2012

University league tables are popular …

Times Higher Education 100 Under 50 rankings

Click heading to sort table. Download this data

100 Under 50 rank	World University Rankings 2011-2012 position	Institution	Country	Teaching	Research	Citations	Overall score
1	53	Pohang University of Science and Technology	Republic of Korea	65.9	66.8	92.3	71.8
2	46	École Polytechnique Fédérale de Lausanne	Switzerland	55.9	40.9	95.3	66.2
3	62	Hong Kong University of Science and Technology	Hong Kong	51.4	62.6	71.0	63.0
4	86	University of California, Irvine	US	42.2	51.5	93.5	60.0
5		Korea Advanced Institute of Science and Technology	Republic of Korea	71.3	61.3	47.1	58.6
6	84	Université Pierre et Marie Curie	France	61.6	26.3	81.1	56.3
7	110	University of California, Santa Cruz	US	31.6	45.4	99.9	56.0
8		University of York	UK	43.1	50.1	71.6	55.7
9		Lancaster University	UK	38.2	43.2	75.4	53.6

http://www.guardian.co.uk/news/datablog/2012/may/21/top-100-universities-under-50

Even when large samples lead to reasonable precision, there are still problems with the concept of league tables.

Trouble with league tables

- Unless all universities are performing **precisely the same,** one of them will be top (or bottom) in the ranking, and not simply due to chance.
- In a **highly competitive environment**, e.g., surgical performance or universities, there may be nothing wrong with coming last.
- The **'bottom'** of the ranking may be the **'middle'** of the distribution, so again there may be nothing wrong with coming last.
- **Not helpful in distinguishing unusual performance.**

So let's add confidence intervals:

caterpillar plot

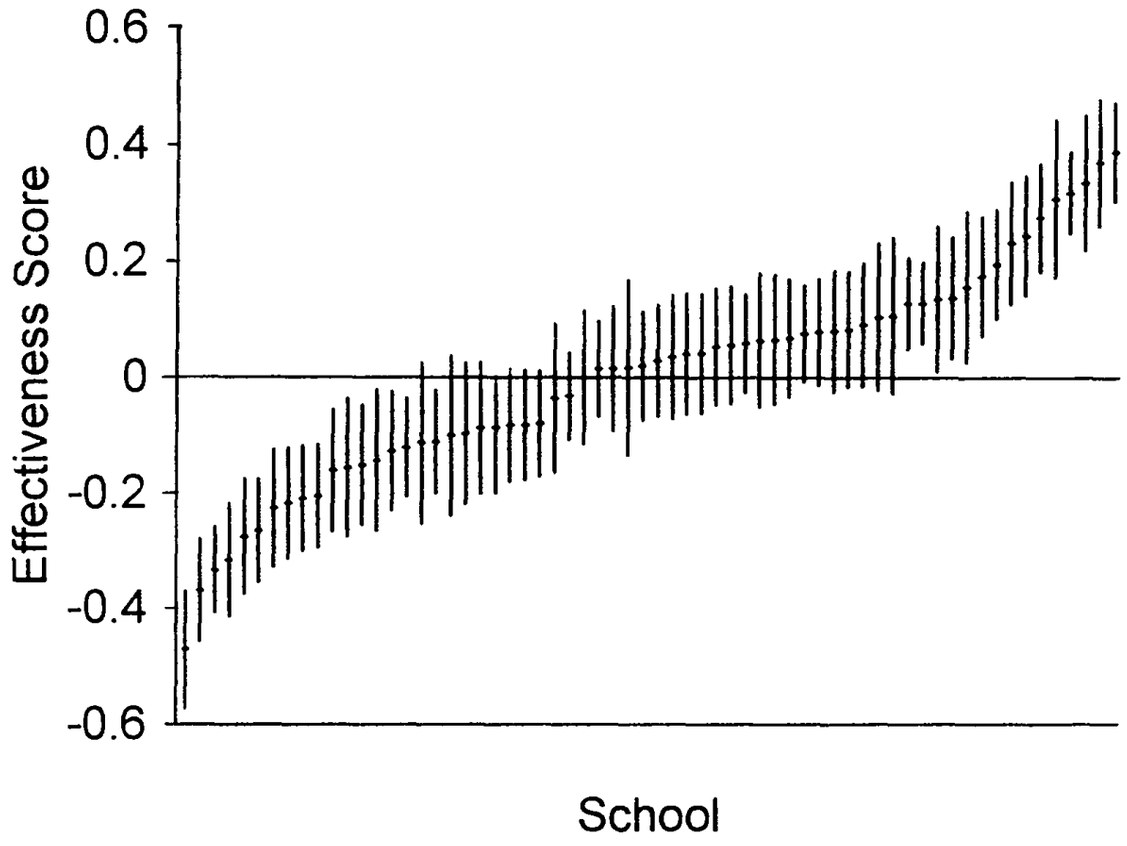


Fig. 1. Effectiveness scores for 64 schools after adjusting for intake achievement

Goldstein and Healy, JRSS A, 1995

Still not helpful in picking out unusual schools.

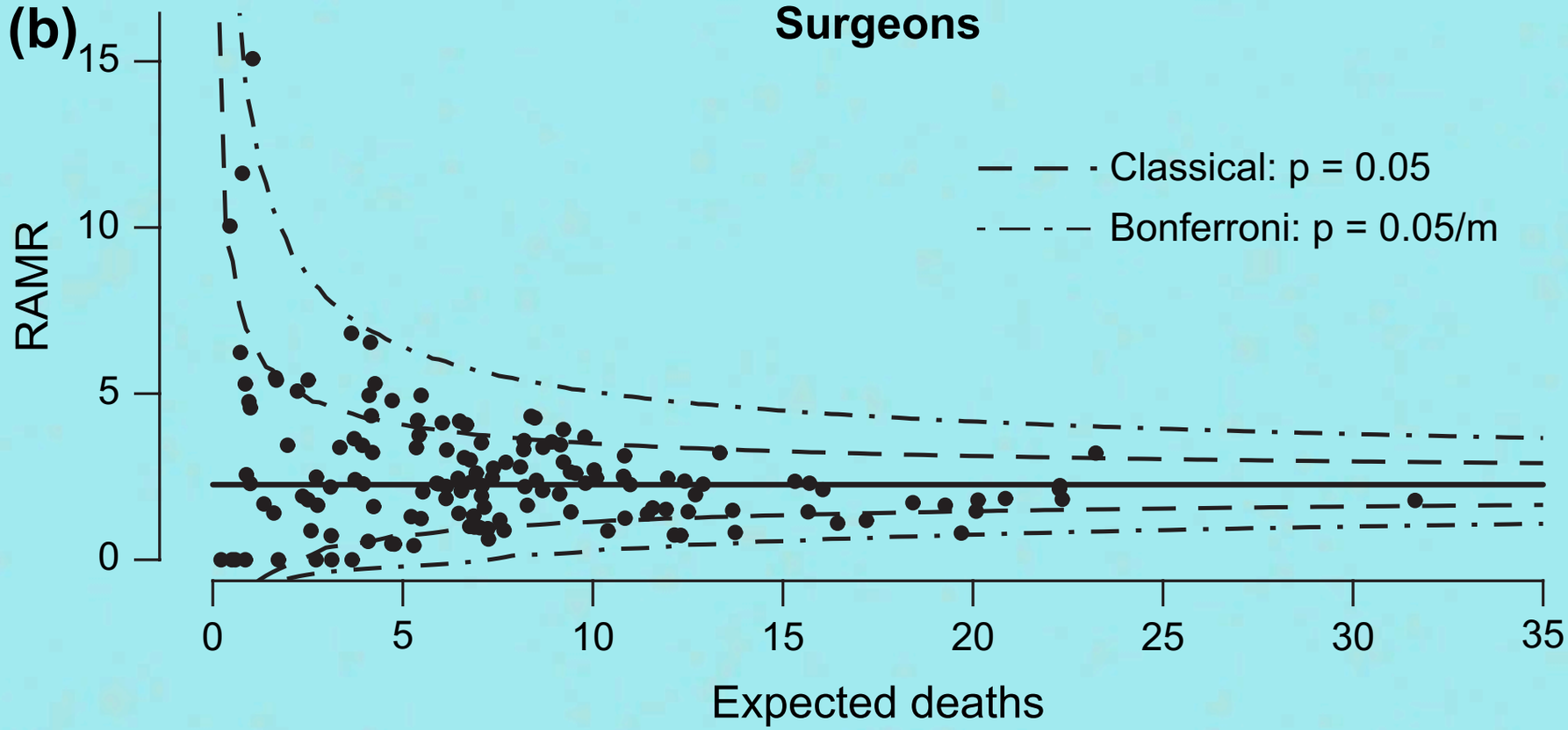
Nor does it answer the question …

Is the worst ranked school worse than we would expect the worst school to be, where the expectation is based on a null hypothesis of no difference between schools?

We also want an answer to the same question, replacing **worst** with **best.**

Better to use a funnel plot

Surgeon-specific risk-adjusted mortality rates



H.E. Jones et al. / Journal of Clinical Epidemiology 61 (2008) 232–240

Better to use False Discovery Rate thresholds.

Adjusting for multiple comparisons

Bonferroni method:

Suppose we compare 1000 null hypotheses H_0,

$$H_{01}, H_{02}, \ldots, H_{0,1000}$$

not necessarily independent and observe corresponding p-values

$$p_1, p_2, \ldots, p_{1000}$$

The Bonferroni method controls the Family-Wise Error Rate (FWER), which is the probability of of falsely rejecting even one null hypothesis, to be $\alpha \leq 0.05$.

Compare the observed p-values to the nominal threshold

$$\alpha' = \frac{0.05}{1000}$$

This controls the probability of making even one mistake, and can be at the cost of making a true discovery.

Adjusting for multiple comparisons

False discovery rate:

Suppose m independent null hypotheses are tested simultaneously, of which R are declared to be statistically significant and V are false discoveries.

Then the false discovery rate (FDR) is

$$E\left(\frac{V}{R}\right)$$

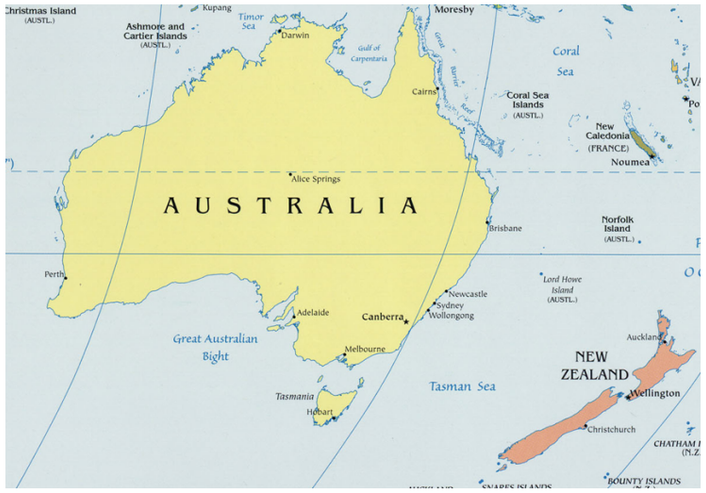
where V/R is defined to be zero when $R = 0$.

The q-value is the FDR analogue of the p-value.

It is defined as the maximum FDR for which the test may be called significant.

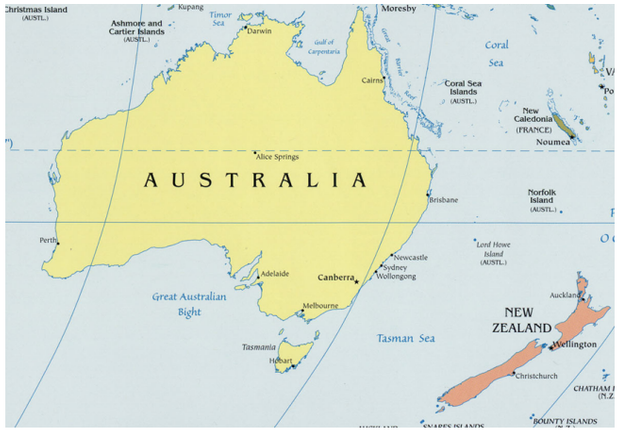
Our aim is to identify intensive care units (ICUs) with unusual performance, using

The Australian and New Zealand Intensive Care Society (ANZICS) Adult Patient Database (APD)



Statistically, the approach is one of

"horses for courses"



The ANZICS APD

- collects voluntary patient-level admissions data from ICUs in OZ and NZ;
- 1995-2010: over 1 million individual patient admissions. In 2010, more than 80% of eligible ICUs (n=157) participated;
- data collected on: age, sex, patient severity score APACHE III, patient diagnostic category, surgical and ventilation status, hospital level, geographical locality, and much more;
- **APACHE = Acute Physiology And Chronic Health Evaluation score (3rd revision);** recorded as worst during first 24 hours post admission.
- We use **in-hospital mortality** to compare ICU performance.

ANZICS APD: patient characteristics in 2009 and 2010

minimum 150 admissions per ICU per year*

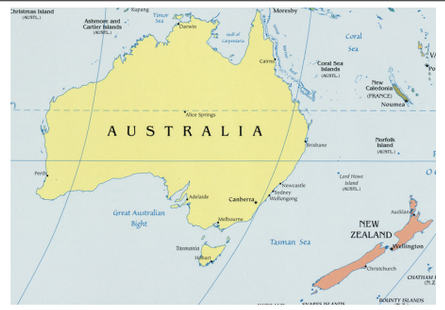
Age in years	61.65 (18.20)	
APACHE III score	51.28 (27.23)	
ICU mortality (%)	6.51	
Hospital mortality (%)	10.21	
2009-2010 patient volume	1194 (1153)	

Total number of patients = 163795

	n (%)	Hospital mortality (%)
Ventilation		
Not ventilated	94802 (57.88)	6.32
Ventilated	68993 (42.12)	15.56
Gender		
Male	95128 (58.08)	10.31
Female	68667 (41.92)	10.08
Patient surgical status		
Non-surgical	96364 (58.83)	13.86
Elective surgical	47847 (29.21)	2.36
Emergency surgical	19584 (11.96)	11.45
Patient diagnostic category		
Cardiovascular	40230 (24.56)	15.81
Gastrointestinal	28639 (17.48)	8.92
Metabolic	11424 (6.97)	3.16
Neurologic	18216 (11.12)	12.56
Respiratory	25057 (15.30)	13.94
Trauma	9030 (5.51)	8.34
Renal/Genitourinary	8612 (5.26)	4.78
Hematological	22587 (13.79)	2.24

	n (%)	Hospital mortality (%)
ICU source		
No transfer	151185 (92.30)	9.69
Hospital transfer	12610 (7.70)	16.48
ICU hospital level		
Rural	21348 (13.03)	10.07
Metropolitan	29294 (17.88)	13.17
Tertiary	70587 (43.09)	12.74
Private	42566 (25.99)	4.06
ICU location		
NT	2153 (1.31)	10.03
NSW	51046 (31.16)	10.53
ACT	4014 (2.45)	9.52
SA	12772 (7.80)	13.71
VIC	41426 (25.29)	10.28
WA	3279 (2.00)	11.04
NZ	9164 (5.60)	13.43
QLD	37337 (22.80)	7.63
TAS	2604 (1.59)	11.56

*115 ICUs



The ANZICS APD

Data structure is hierarchical:

variability between ICUs

variability between patients within ICUs

➡ A two-level hierarchical model for mortality

Let $$Y_{ij} = \begin{cases} 1 \text{ if patient } i \text{ in ICU } j \text{ dies in hospital} \\ 0 \text{ otherwise} \end{cases}$$

where $$j = 1, \ldots, m,\ i = 1, \ldots, n_j, \quad Y_{ij} \sim \text{Bernoulli}(p_{ij})$$

and $$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x_{ij}} + U_j, \quad U_j \sim N(0, \sigma^2)$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x_{ij}} + U_j, \quad U_j \sim N(0, \sigma^2)$$

- is a *random intercept logistic regression model*.
- it accommodates fact that *responses within ICUs are correlated* and provides "*shrinkage estimates*".
- Random intercepts model *unknown differences between ICUs*.
- *This is an example of a random effects model; also known as hierarchical models, nonlinear mixed models, multilevel models, variance components models, ...*
- We could fit a **fixed effects** logistic regression model, where each ICU has its own (fixed) intercept. *What would this model look like?*

We need a key performance indicator

- A KPI is a summary statistic intended to measure the 'quality' or 'effectiveness' of an ICU's functioning.
- Whilst death could be considered the ultimate 'performance', how much should we attribute to the hospital?
- We want to compare ICUs, distinguishing 'usual' from 'unusual' performance.
- We use the log Standardized Mortality Ratio (SMR) as our KPI:

$$\log SMR_j = \log \frac{\sum_{i=1}^{n_j} Y_{ij}}{\sum_{i=1}^{n_j} p_{ij}} = \log(O_j) - \log(E_j)$$

How do we identify unusual performance?*

- Approach I: Fit a random effects distribution that encompasses *all* the variation between ICUs. Then identify extreme ICUs = 'outlier accommodation'.**

- Approach II: Fit a random effects distribution to the *usual* ICUs to obtain a null model. Then identify divergent ICUs = 'outlier detection'.

We take a classical Approach II, which involves 3 Stages.

* Ohlssen et al, *JRSS* A, 2007

**Barnett & Lewis, 1978

Stage 1: find a good risk-adjusted mortality model for all 2009-2010 data

A (two-level) random coefficient logistic regression model:

$$Y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{U}_j \sim \text{Bernoulli}(p_{ij})$$

where

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 AP_{ij} + \sum_{k=2}^{110} \beta_k x_{kij} + U_{0j} + U_{1j} AP_{ij}$$

with

$$\begin{pmatrix} U_{0j} \\ U_{1j} \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_I^2 & \sigma_{I,AP} \\ \sigma_{I,AP} & \sigma_{AP}^2 \end{pmatrix}\right)$$

$\sigma_I^2, \sigma_{AP}^2$ are components of variance, and $\sigma_{I,AP}$ is the component of covariance.

Stage 1: find a good risk-adjusted mortality model for all 2009-2010 data

- For **model building**, the data were split randomly into an *80% training dataset* to fit the model, and a *20% test dataset* to estimate the prediction error (PE).
- This approach gives a valid estimate of the PE.
- In the statistics you have met so far, *the training dataset = the test dataset,* which gives an optimistic estimate of the true PE.
- This is because you are using the **same data** to **fit** the model and to **test** it.
- When not enough data available to split, use *cross-validation*.

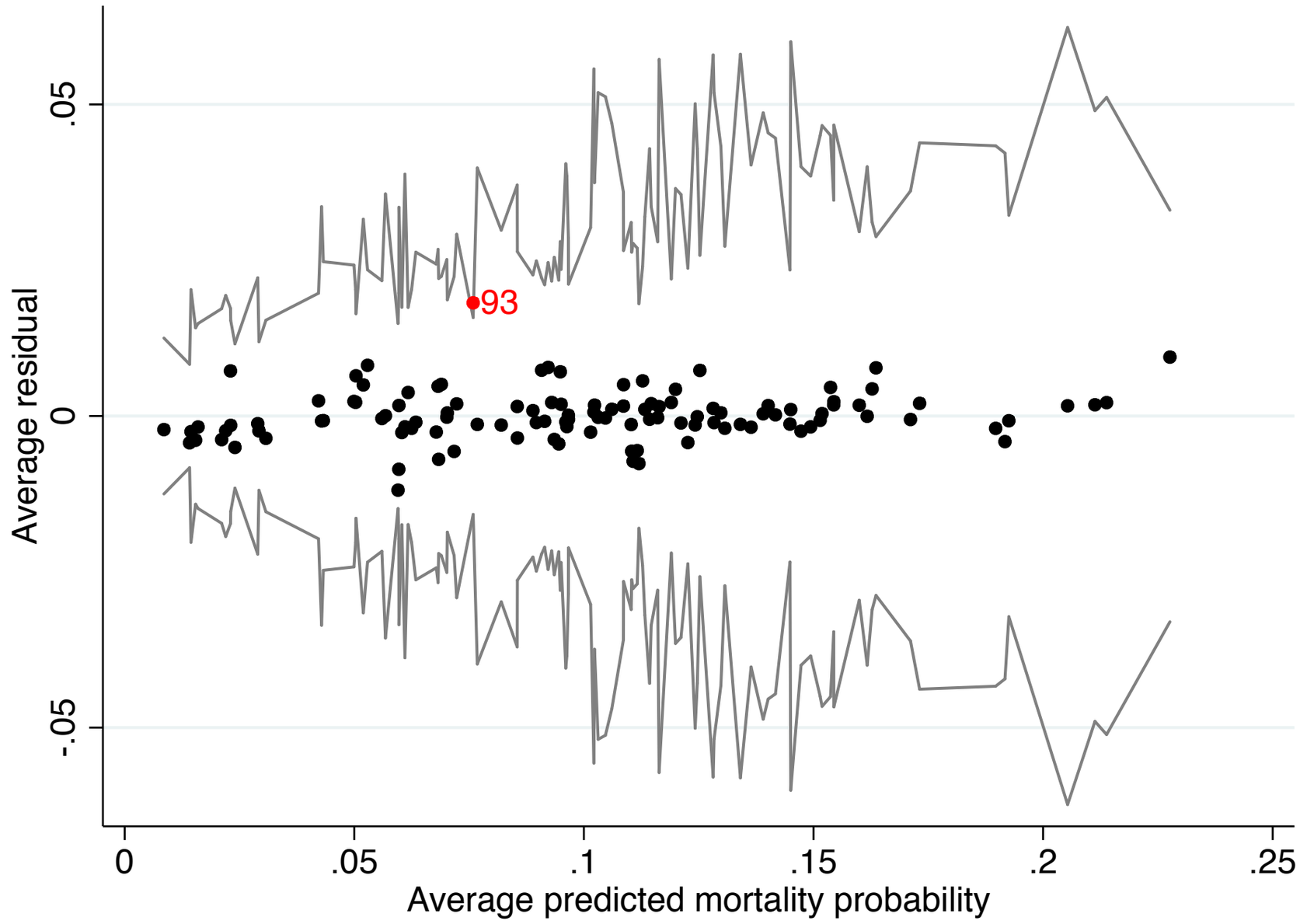
Stage 1: find a good risk-adjusted mortality model for all 2009-2010 data

- The **model fitting** was done using maximum likelihood.
- Log likelihoods approximated by numerical integration: 7-point adaptive Gaussian quadrature in Stata v12.1, xtmelogit command.
- ***R can't cope with the large dataset.***
- We started with a lot more than 112 fixed explanatory variables in the model.
- For **model selection,** variables dropped stepwise with $p<0.10$.
- **Also** used information criteria (AIC for nested models, BIC), global measures of goodness-of-fit, and binned residual plots.

Variable	Stage 1 model			
	Log odds	p-value	95% CI	
Age (per 10 year increase)	0.2352	< 0.0001	0.2165	0.2540
Age squared	0.0134	< 0.0001	0.0074	0.0194
APACHE III score (per 10 units increase)	0.5922	< 0.0001	0.5662	0.6181
APACHE III score squared	-0.0099	< 0.0001	-0.0123	-0.0075
Age $\times$ APACHE III score	-0.0208	< 0.0001	-0.0255	-0.0160
Gender (baseline male)	-0.0392	0.0519	-0.0787	0.0003
Patient Category (baseline cardiovas)				
Gastrointestinal	-0.1604	0.0088	-0.2803	-0.0404
Metabolic	-1.3120	< 0.0001	-1.5375	-1.0866
Neurologic	0.5154	< 0.0001	0.3798	0.6511
Respiratory	0.5605	< 0.0001	0.4691	0.6518
Trauma	-0.6107	< 0.0001	-0.8525	-0.3688
Renal/ genitourinary	-0.5210	< 0.0001	-0.7300	-0.3121
Hematologic	-1.4325	< 0.0001	-1.6054	-1.2596
Patient surgical status (baseline non-surgical)				
Elective surgery	-1.4800	< 0.0001	-1.6290	-1.3310
Emergency surgery	-0.4517	< 0.0001	-0.6173	-0.2862
Ventilation	0.4132	< 0.0001	0.3144	0.5120
ICU source	-0.1172	0.0043	-0.1977	-0.0367
Patient category $\times$ APACHE III score				

Stage 1 model checking: binned residual plot

ICU-level: 115 bins

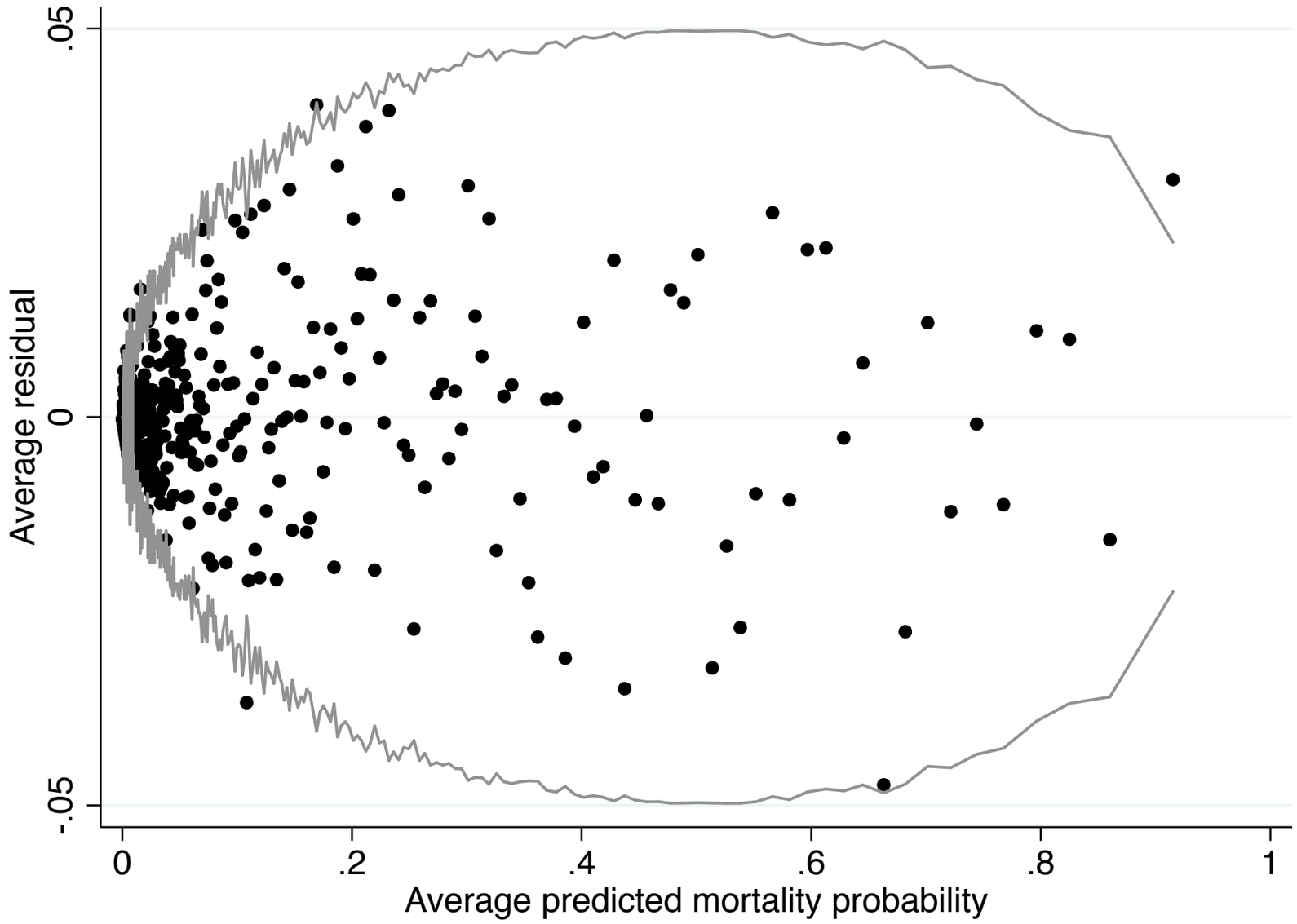


95% of binned residuals should lie within +/- 2 error bounds if model correctly specified.

Gelman & Hill, CUP 2007

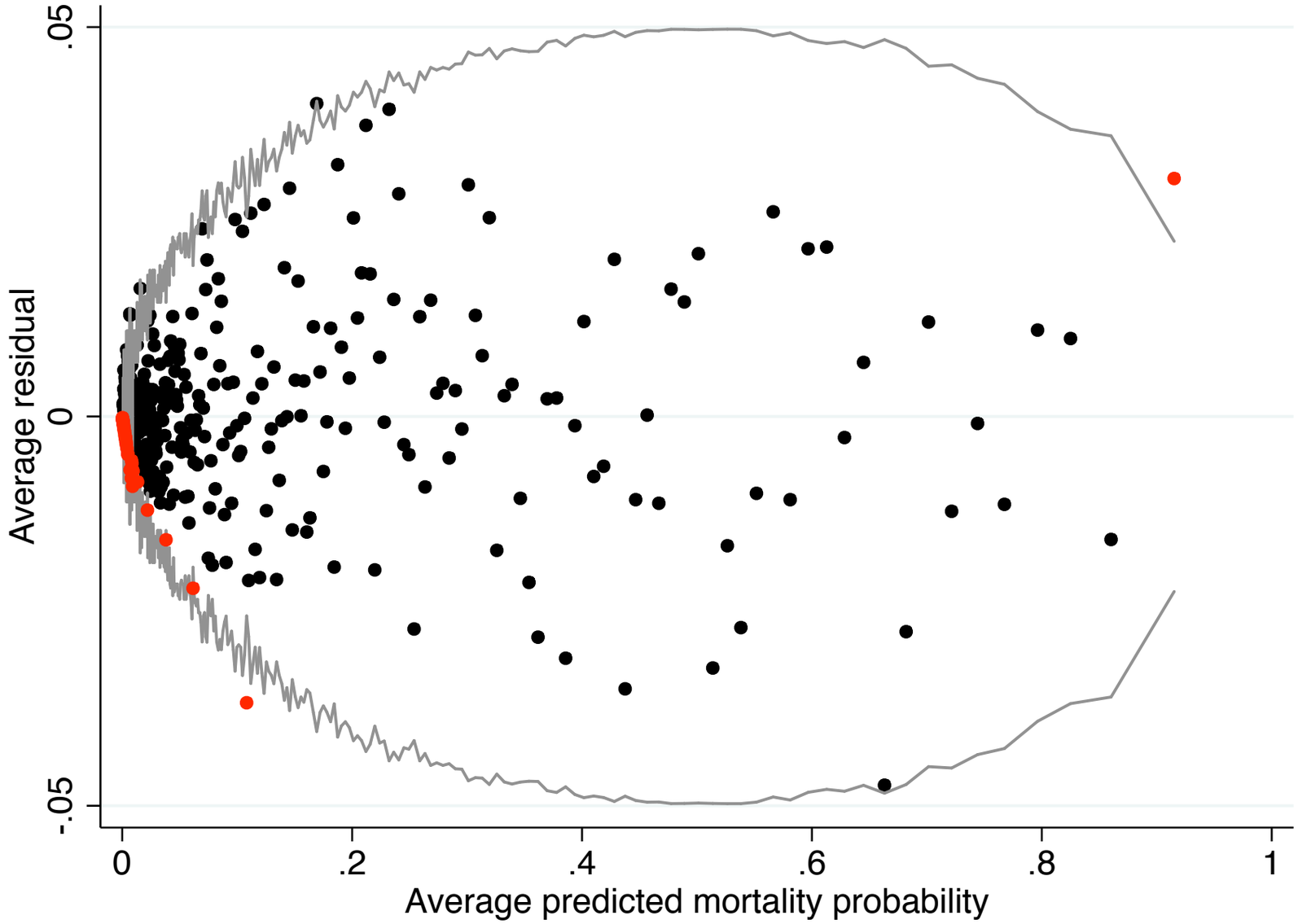
Stage 1 model checking: binned residual plot

Patient-level: 404 bins



Stage 1 model checking: binned residual plot

Patient-level: 404 bins



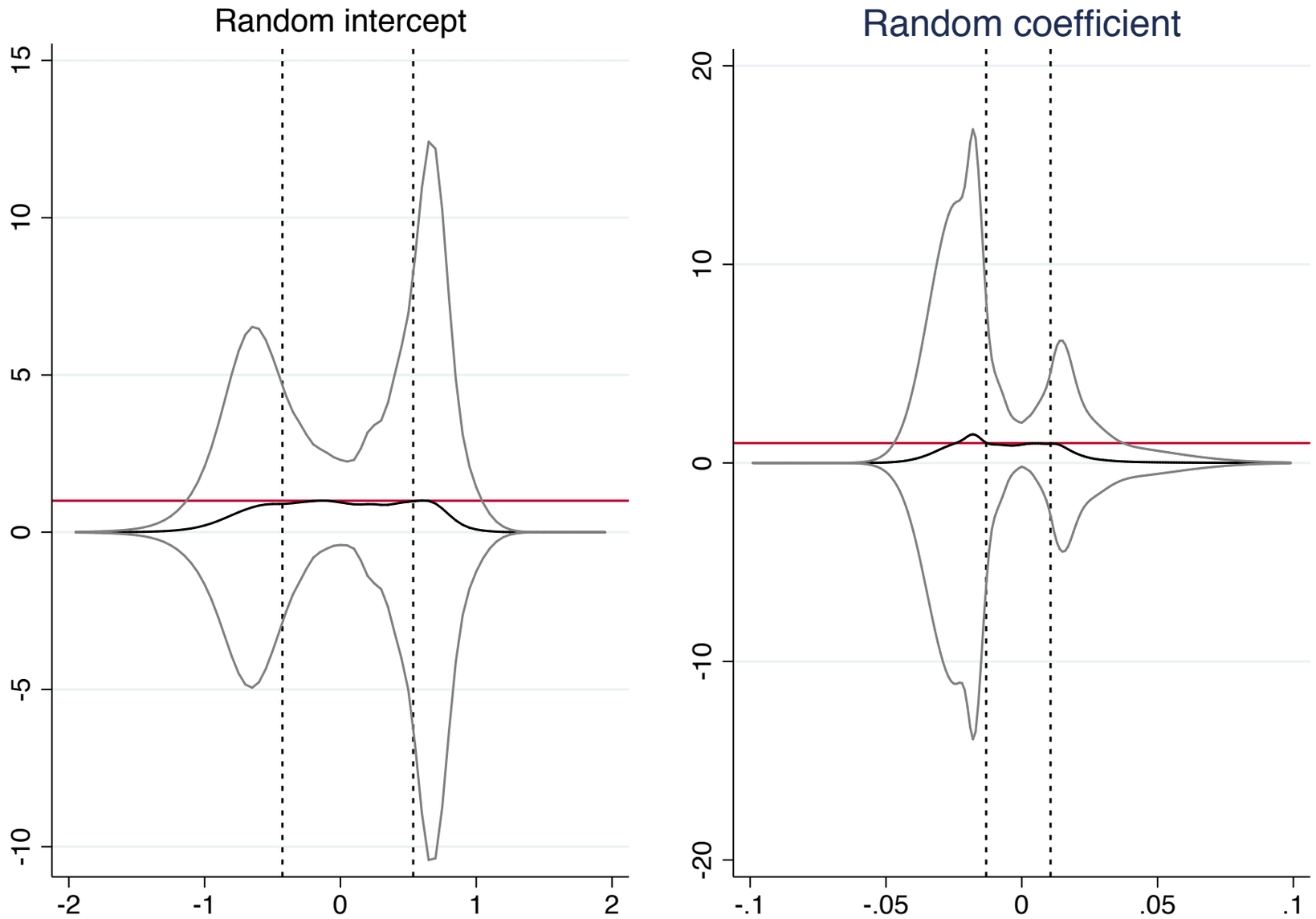
Correct adjustment for casemix is difficult.

But we have a good **empirical** model for prediction.

Stage 1 model checking: gradient function

$$\Delta(G, \boldsymbol{U}) = \frac{1}{m} \sum_j \frac{f_j(\boldsymbol{y}_j|\boldsymbol{U})}{f_j(\boldsymbol{y}_j|G)}$$

$\Delta(\hat{G}, \boldsymbol{U})$



$= 1$

Verbeke & Molenberghs *Biostatistics* 2013

"Degrees of freedom" ??

Stage I: identify potentially unusual ICUs (using approximate cross-validation)

For each ICU j and for $k = 1, \ldots, 5000$

- simulate $\boldsymbol{U}_j^k$ from fitted model, calculate $\tilde{p}_{ij}^k$
- simulate outcome for each patient:

$$Y_{ij}^k \sim \text{Bernoulli}(\tilde{p}_{ij}^k)$$

- count number of deaths: $E_j^k = \sum_{i=1}^{n_j} Y_{ij}^k$.

Calculate approximate p-value for each ICU:

$$p_j^{approx} = \frac{1}{5000} \sum_{k=1}^{5000} I_{E_j^k < O_j}$$

This measures how well the estimated model predicts O for each ICU.

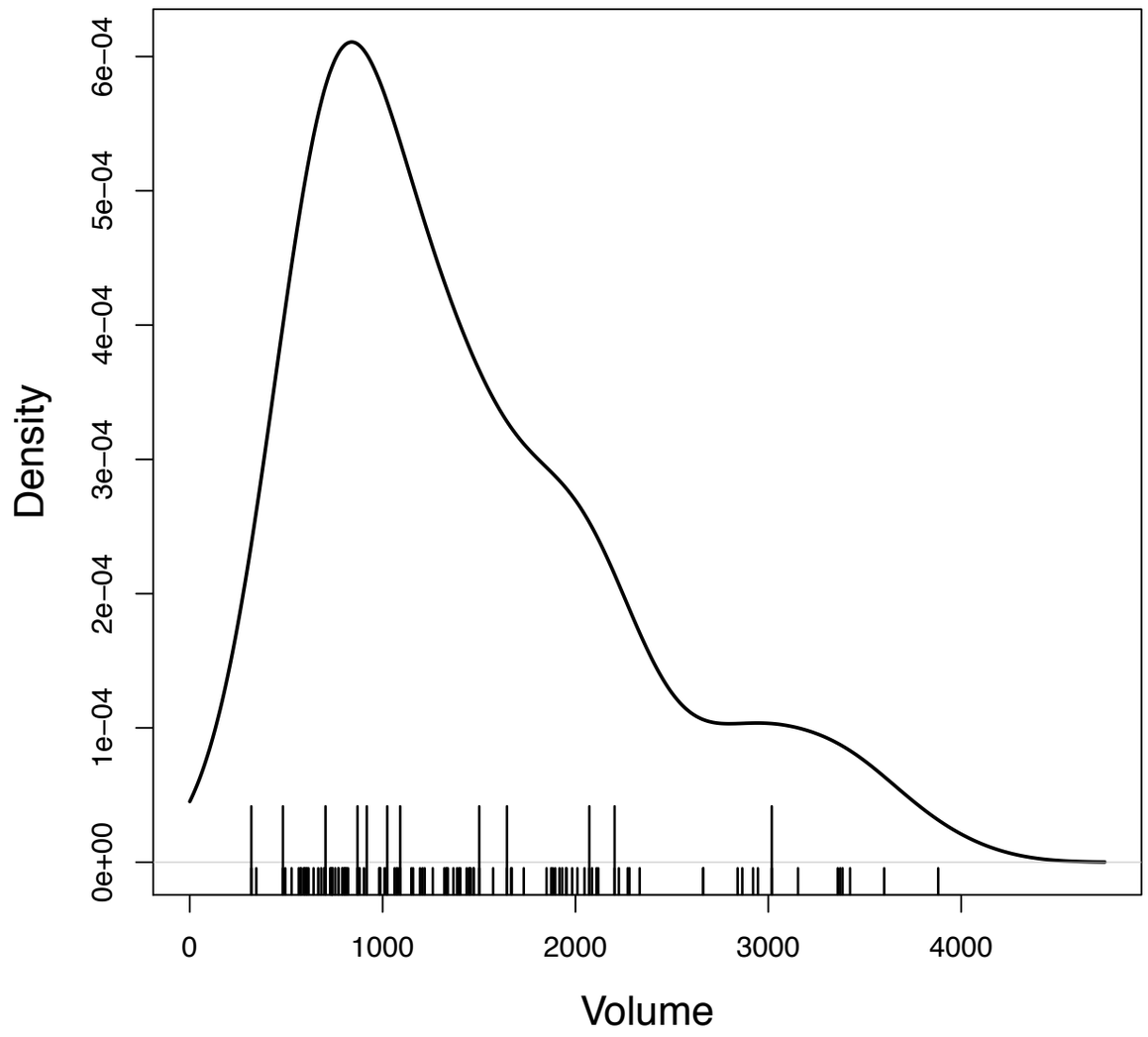
Stage I: here are the potentially unusual ICUs

$p < 0.05$ over-performing $p > 0.95$ under-performing

ICU identifier	Hospital Level	*p*-value
100	Private	0.0166
57	Private	0.0182
48	Rural	0.0202
72	Rural	0.0220
108	Private	0.0258
49	Metropolitan	0.0290
19	Private	0.0422
45	Tertiary	0.0494
93	Private	0.9658
81	Private	0.9770
44	Private	0.9874
16	Private	0.9952

(ICU identifiers are random numbers)

Kernel density plot of ICU volume 2009-2010



Large tick marks indicate volumes of 12 potentially unusual ICUs.

Stage 2: re-estimating the model

$$Y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{U}_j \sim \text{ Bernoulli}(p_{ij}) \qquad \boldsymbol{U}_j \sim N_2(\mathbf{0}, \Sigma)$$

Let

$$b_j = \begin{cases} 1 \text{ if ICU } j \text{ is identified as potentially unusual at Stage 1} \\ 0 \text{ otherwise.} \end{cases}$$

Then

$$\begin{aligned} \text{logit}(p_{ij}) &= b_j\beta_{0j} + b_j\beta_{1j}AP_{ij} + \sum_{k=2}^{110} \beta_k x_{kij} \\ &+ (1-b_j)\beta_0 + (1-b_j)\beta_1 AP_{ij} + (1-b_j)U_{0j} + (1-b_j)U_{1j}AP_{ij} \end{aligned}$$

* Separate fixed intercepts and *AP* slopes are estimated for b_j=1.
* The null RE distribution is estimated using only "in control" ICUs; the fixed effects are estimated using all ICUs.

(Langford & Lewis *JRSS A*, 1998)

Stages 1 and 2 variance components

Stage 1	$\hat{\sigma}^2$	SE
APACHE III	0.0000318	7.74×10^{-6}
Intercept	**0.0542223**	0.0115764
covariance	-0.0002500	0.0023700
Stage 2	$\hat{\sigma}^2$	SE
APACHE III	0.0000313	7.84×10^{-6}
Intercept	**0.0271328**	0.0073427
covariance	-0.0001876	0.0001879

Including all ICUs inflates the intercept variance estimates at Stage 1.

Random intercept models 'unknown ICU-level variables'.

Estimating the Key Performance Indicator from the Stage 2 model:

log Standardised Mortality Ratio

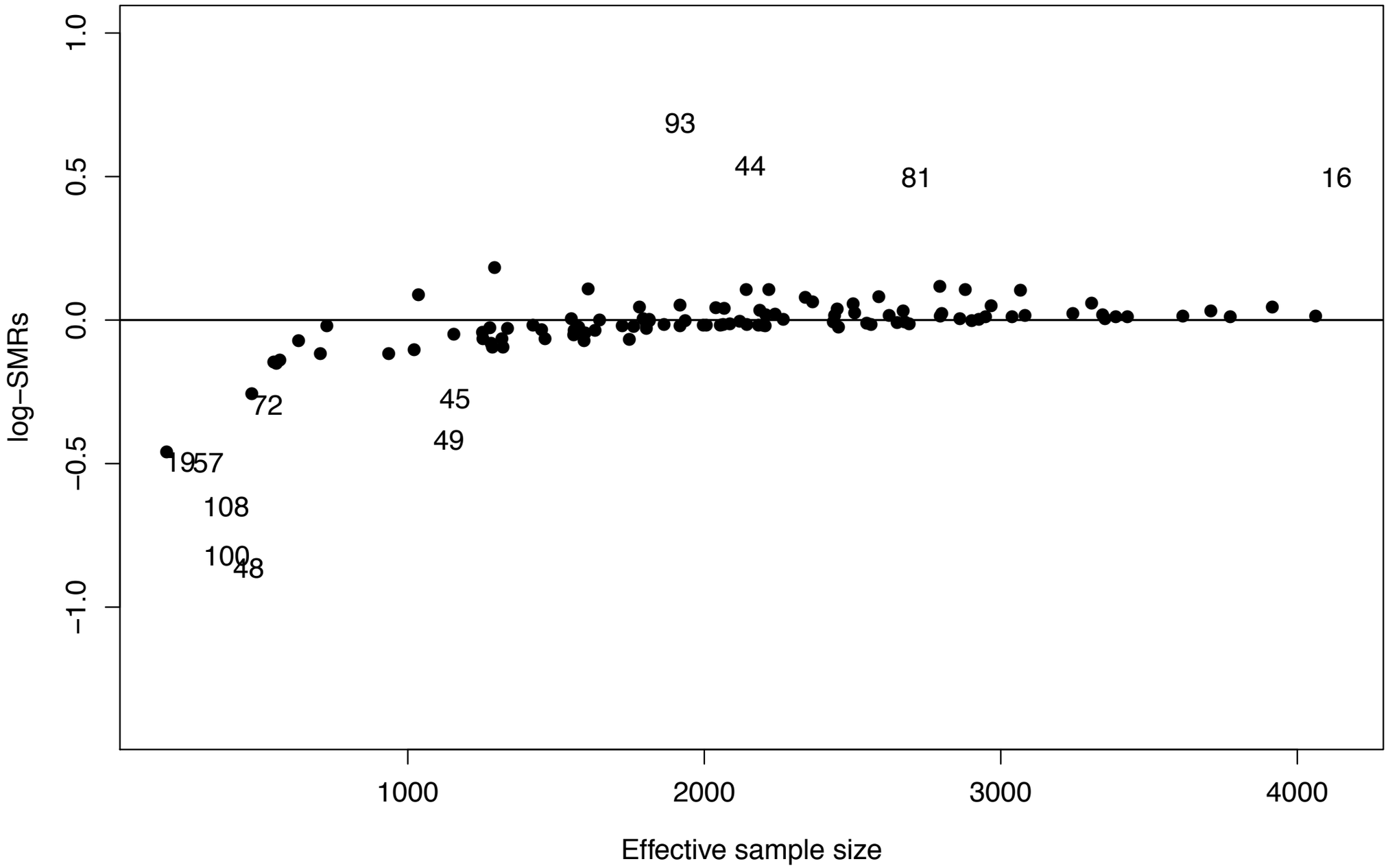
$$\log \mathrm{SMR}_j = \log \frac{\sum_{i=1}^{n_j} Y_{ij}}{\sum_{i=1}^{n_j} p_{ij}} = \log(O_j) - \log(E_j)$$

where for each patient *i* in ICU *j*,

$$\hat{p}_{ij} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{AP}_{ij} + \sum_{k=2}^{110} \hat{\beta}_k x_{kij} + \hat{U}_{0j} + \hat{U}_{1j}\mathrm{AP}_{ij}\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{AP}_{ij} + \sum_{k=2}^{110} \hat{\beta}_k x_{kij} + \hat{U}_{0j} + \hat{U}_{1j}\mathrm{AP}_{ij}\right)}$$

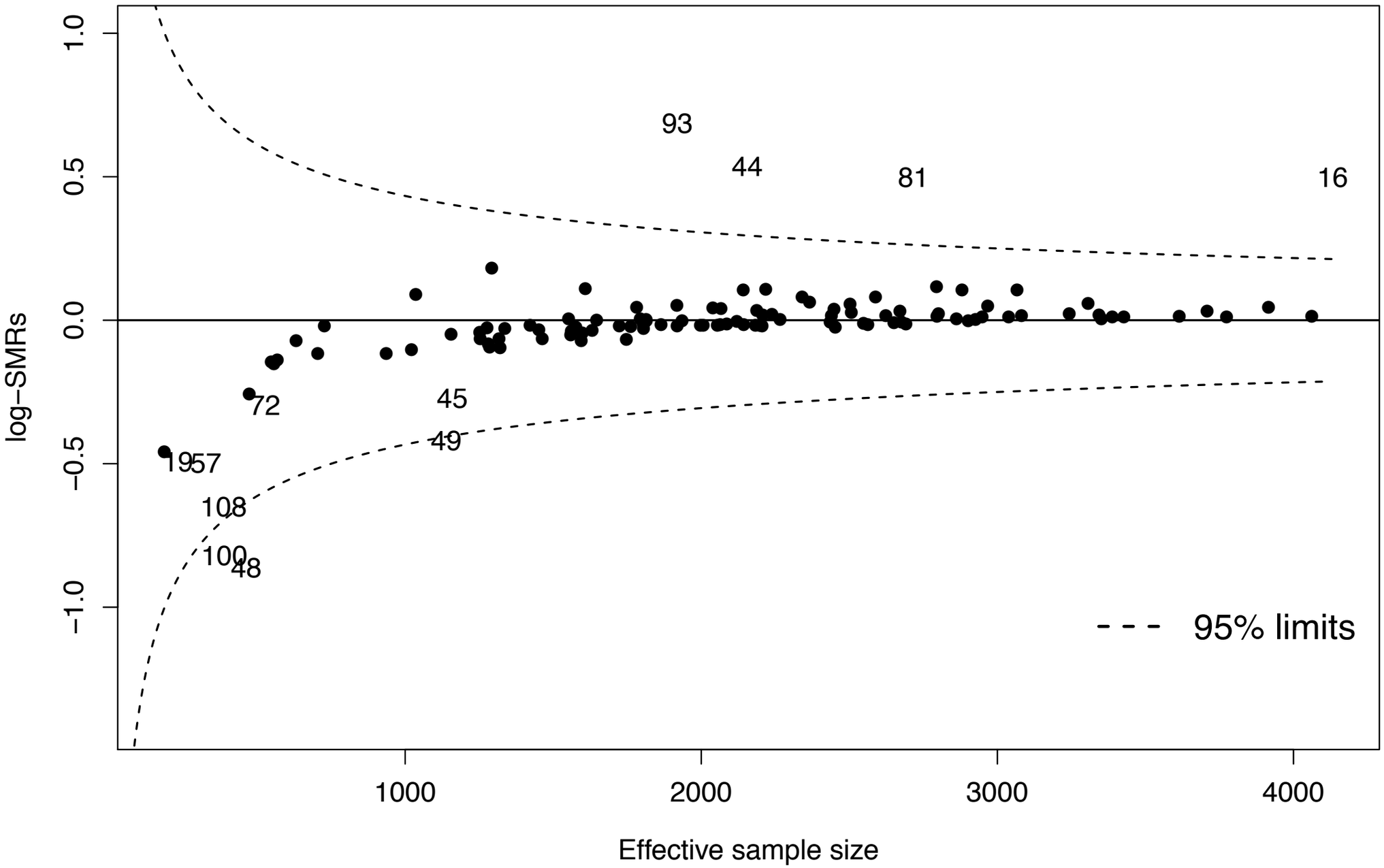
For the potentially unusual ICUs, randomly select a null ICU k and use $\hat{U}_k$

Stage 3: Funnel plot 2009 and 2010



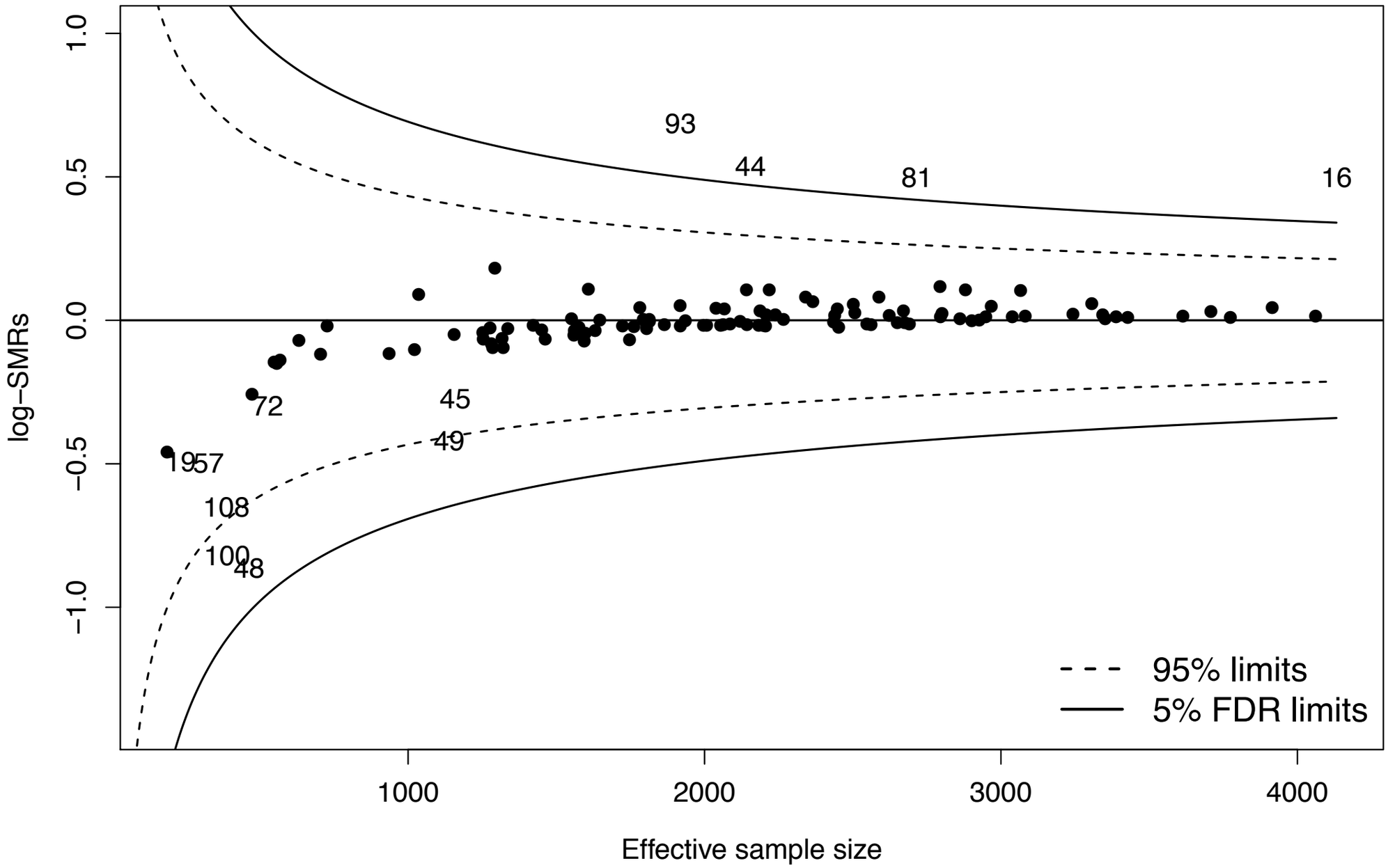
$$n_j^{eff} = \frac{\sum_{i=1}^{115} n_i \hat{\sigma}_i^2 / 115}{\hat{\sigma}_j^2}$$

Stage 3: Funnel plot 2009 and 2010

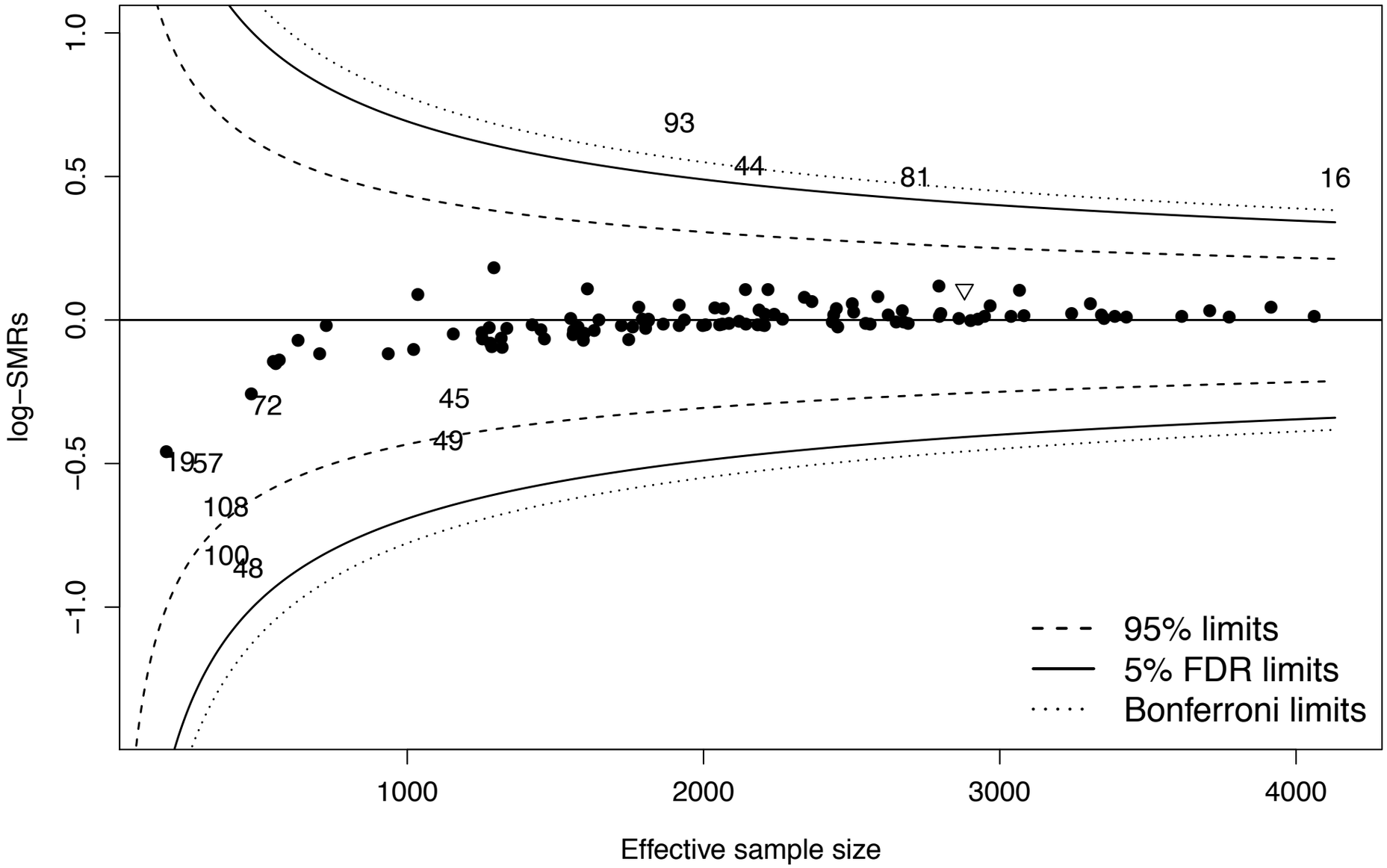


Classical limits: no adjustment for multiple testing

Stage 3: Funnel plot 2009 and 2010



Stage 3: Funnel plot 2009 and 2010



4 Private hospitals have higher than usual mortality:

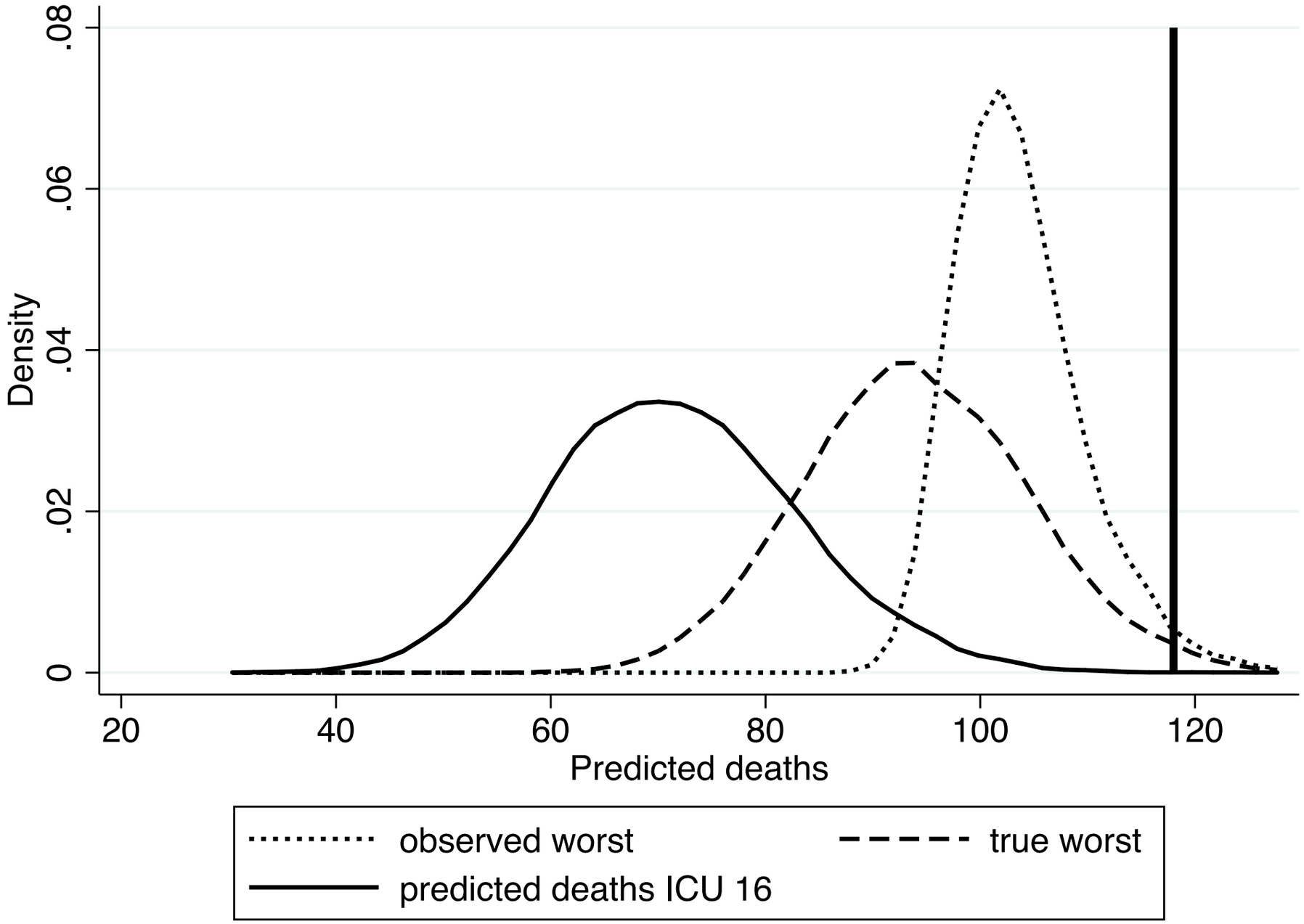
1 in Victoria, 3 in Queensland.

Kasza, Moran and Solomon, Statistics in Medicine, 2013

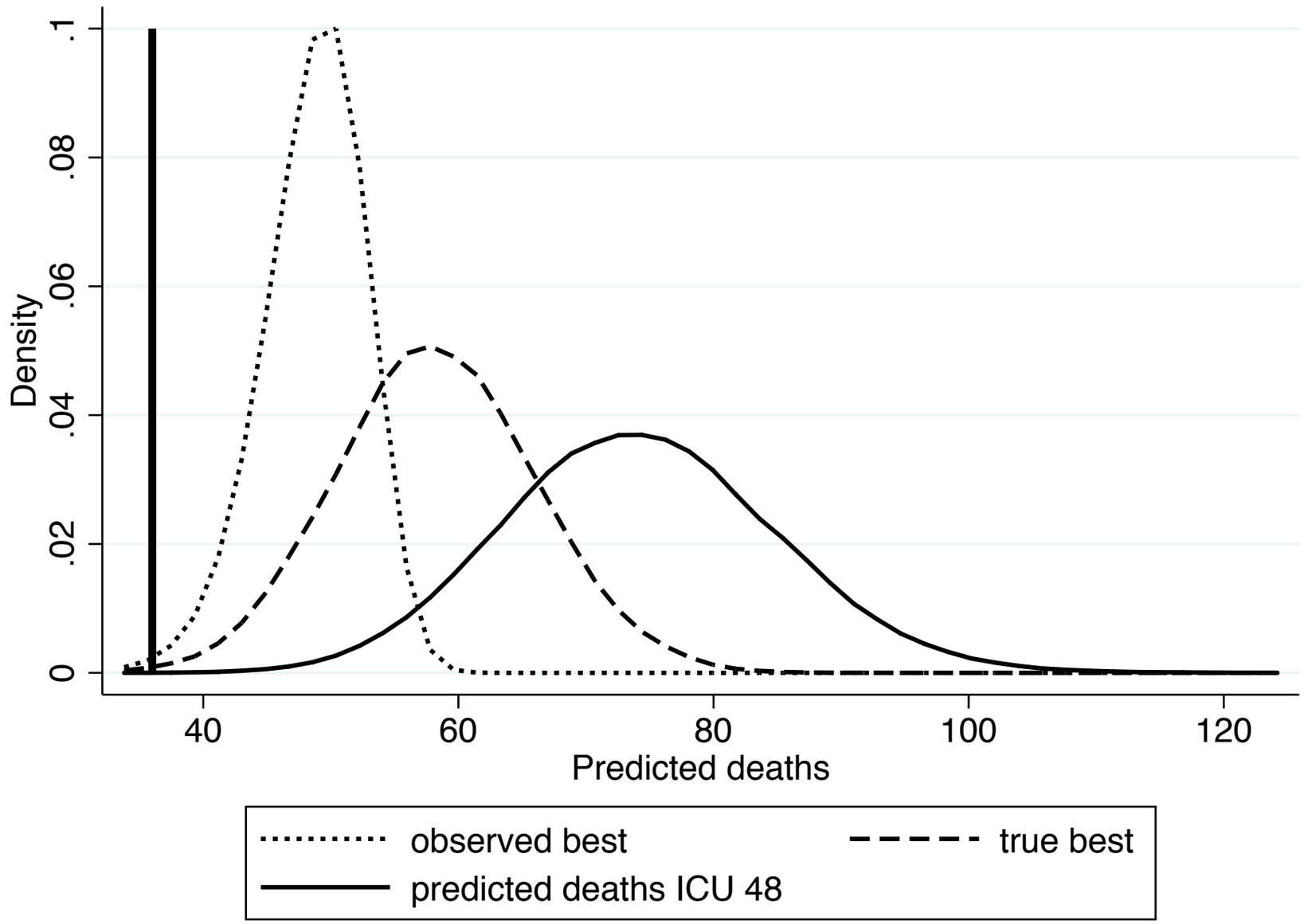
We can now answer the question

- is the **observed worst ICU** worse than would be expected if it had arisen from the **true worst ICU**, but still coming from the null random effects distribution?
- *What about the **observed best** ICU?*

Simulating the 'worst' predicted deaths distributions: ICU 16



Simulating the 'best' predicted deaths distributions: ICU 48



If you can hang on, go to the Northern Territory ...

What can we conclude about recent ICU performance in OZ and NZ?

- Four ICUs in private hospitals were identified with unusual performance in 2009 and 2010 by our three-stage analysis.
- *Are the observed differences in mortality potentially performance related?*
- Yes, and likely to be due to differences in ICU process of care.
- Three important messages are: *comprehensive risk adjustment is essential, estimation of a null model is mandated* and *the statistical analysis is complicated!*