Statistical analysis of hospital performance: understanding the uncertainty

Patty Solomon
University of Adelaide

NZSA 2012 Conference
Dunedin New Zealand, 30 November 2012

Somewhere between Porters Pass and Arthur's Pass

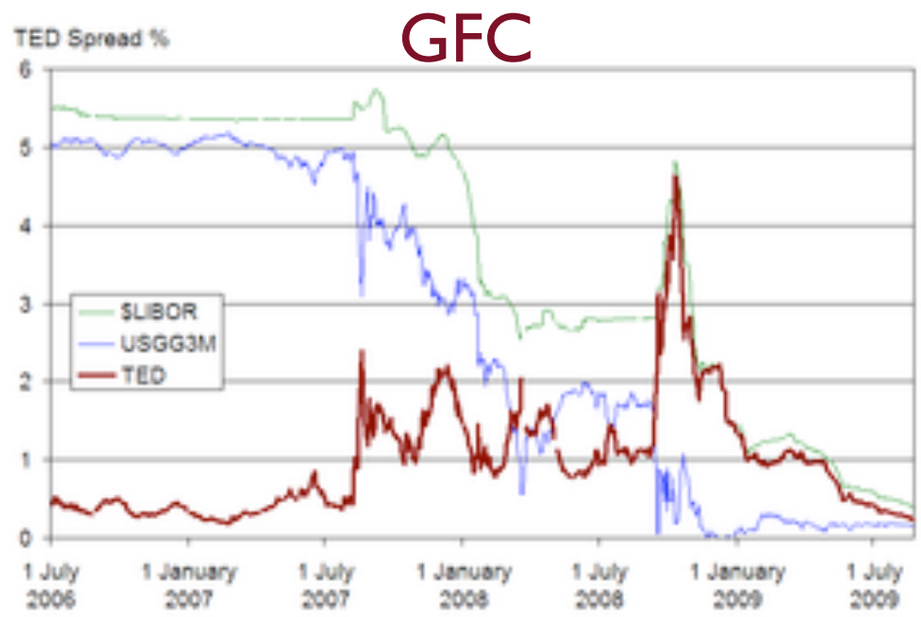
Lies, damn lies and statistics

Truths, damn truths and statistics

1st truth: statisticians are **experts** at handling uncertainty.

2nd truth: there are different **kinds** of uncertainty.

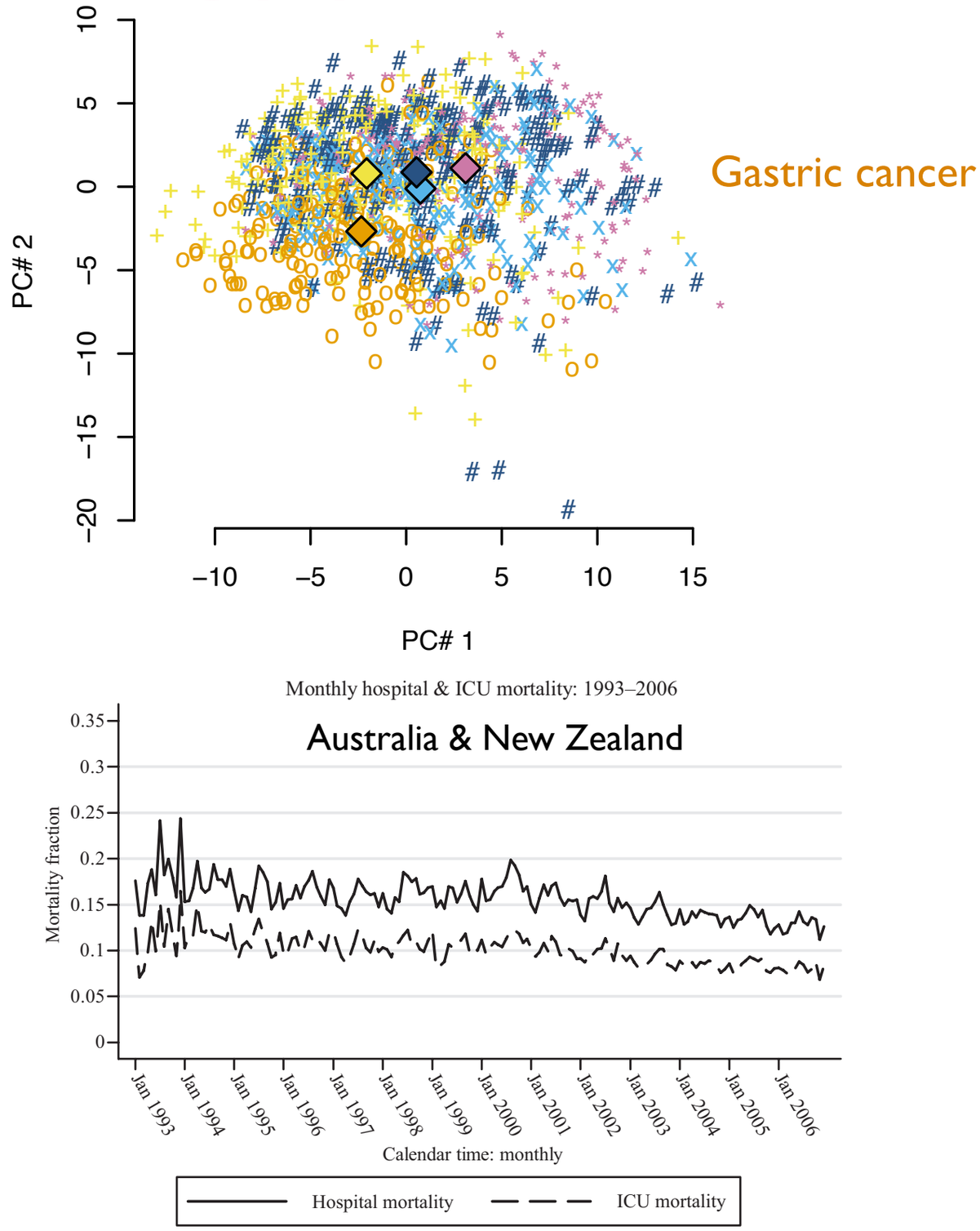
- At one extreme financial time series can be unpredictable.
- At the other extreme, survey sampling outcomes can be highly predictable.



In between, outcomes from the new "omics" technologies are surprisingly predictable.

- In DNA, RNA-seq and proteomics experiments, often > 70% of variation explained.

- In medicine and public health, the proportion of variation explained is typically high and outcomes "predictable".



3rd truth: the world is 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

http://london2012.olympics.com.au/medal-tally/sortby/gold

19/09/12 1:44 PM

You just have to choose the right performance indicator

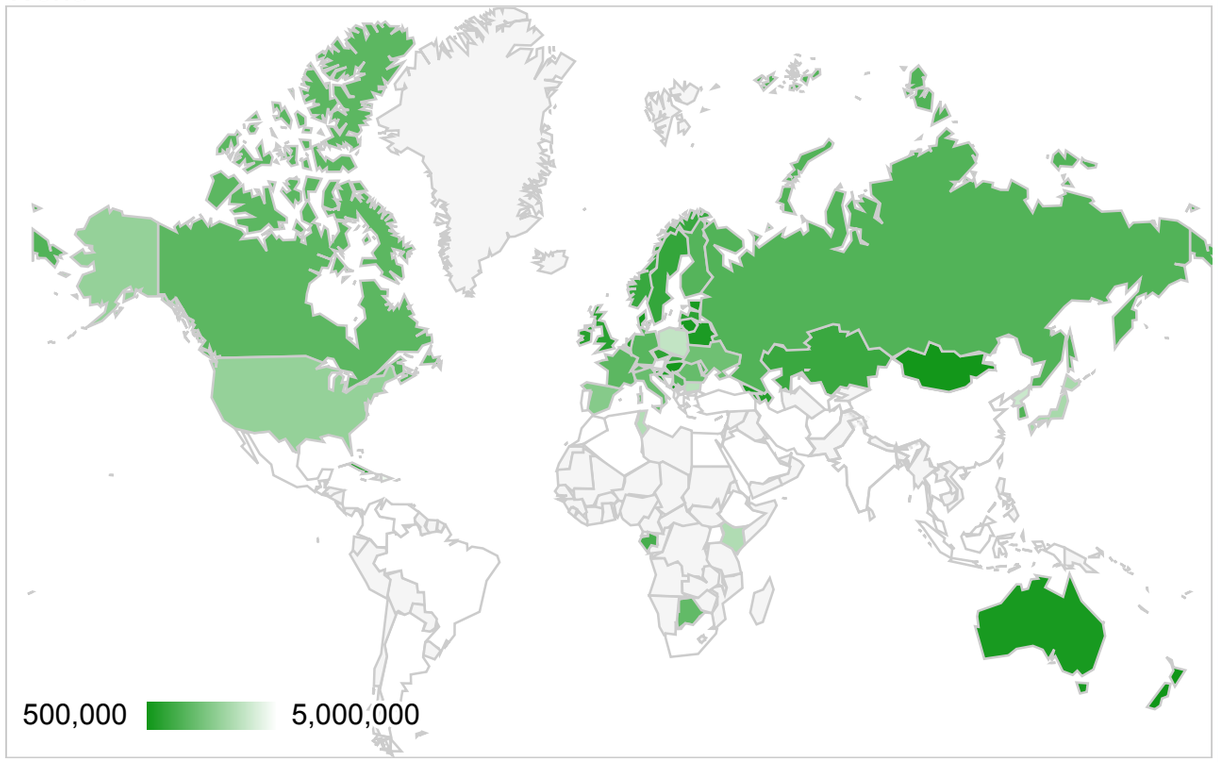
MEDALS PER CAPITA

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

World



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

4th truth: 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/31/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 the **same**, one of them will be top (or bottom) in the ranking, and not due to chance.

- In a **competitive** environment, e.g., surgical performance, there may be nothing wrong with coming last: ranks are comparative.

- The 'bottom' of the ranking may be the 'middle' of the distribution, and so on.

So let's add confidence intervals ...

J. R. Statist. Soc. A (1995)
158, *Part* 1, *pp.* 175–177

The Graphical Presentation of a Collection of Means

By HARVEY GOLDSTEIN† and MICHAEL J. R. HEALY

Institute of Education, London, UK

[Received July 1993]

Caterpillar plot:

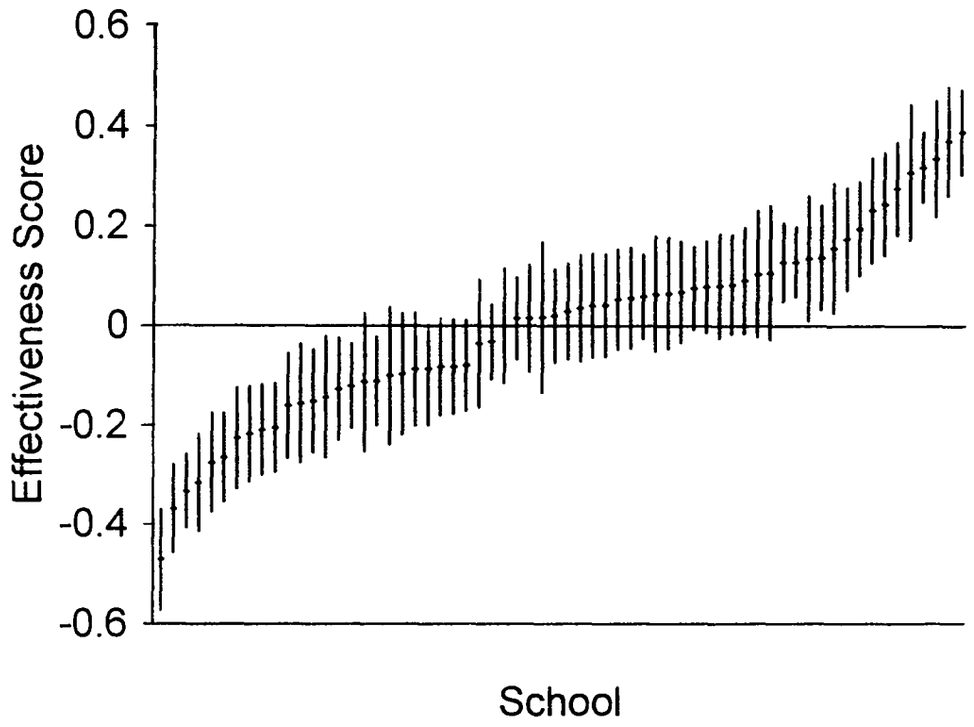
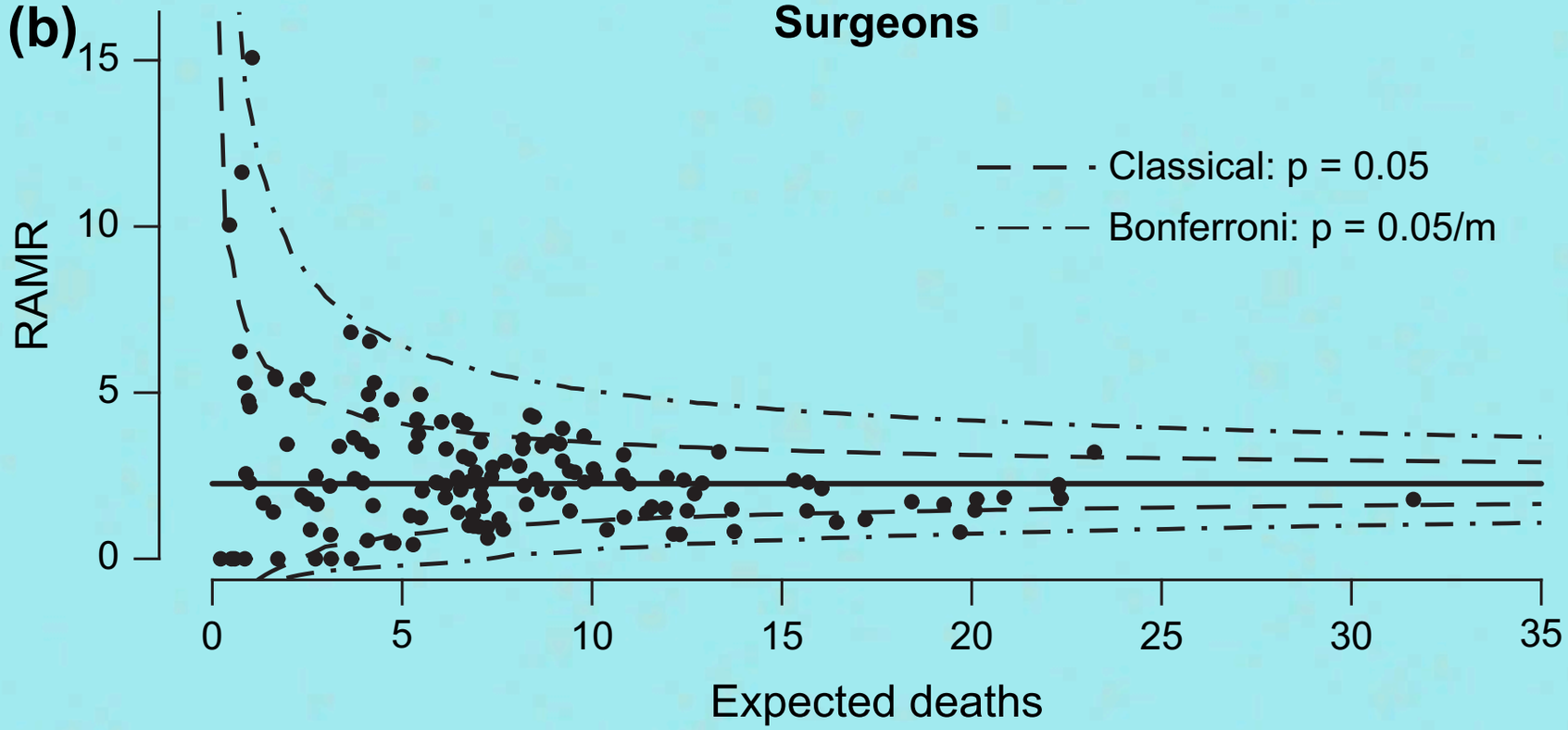


Fig. 1. Effectiveness scores for 64 schools after adjusting for intake achievement

Not helpful in picking out unusual schools.

Better to use a funnel plot

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



Surgeon-specific risk-adjusted mortality rate

Better still to use False Discovery Rate limits.

The Australian Government's response to "excess deaths" is a commitment to

Australian Government

National Health Reform

National Health Reform

Progress and Delivery

September 2011

"Under a Performance and Accountability Framework, the National Health Performance Authority (NHPA) will develop and produce Hospital Performance Reports which will report on the performance of every hospital"

So far, we have **MyHospitals**

which is … a League Table!

- We are also promised **more league tables** based on the hospital-standardised mortality ratio:

$$\text{HSMR} = \frac{\text{Observed no. deaths}}{\text{Expected no. deaths}} * 100$$

where E is obtained from a logistic regression model.

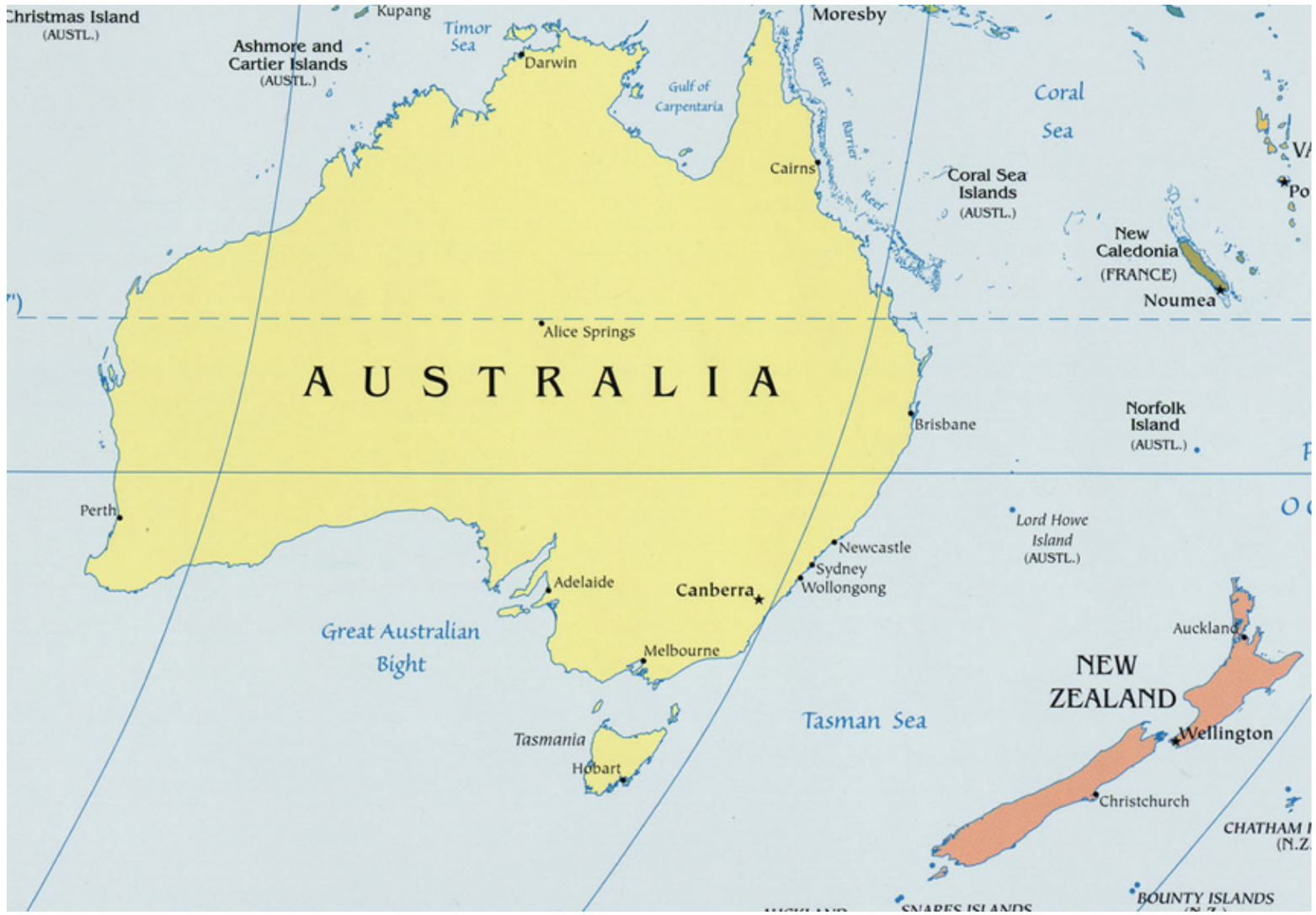
- The validity and reliability of HSMR as an effective screening tool remains in doubt: it is **not robust** and **not** been demonstrated to improve quality of care and patient outcomes. *

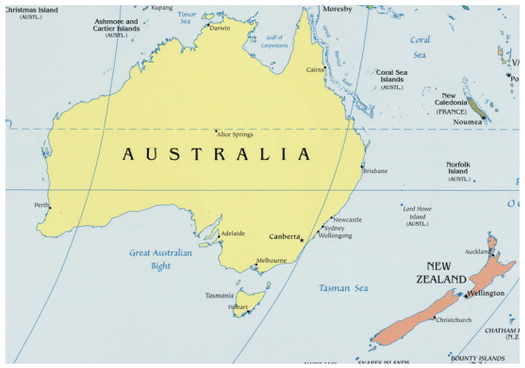
- An **unfavourable HSMR** is likely to lead to **gaming** or **inappropriate changes** to care.

* Scott *et al, Medical Journal of Australia* 2011

Our motivation: *to establish a principled statistical methodology for evaluating hospital performance using*

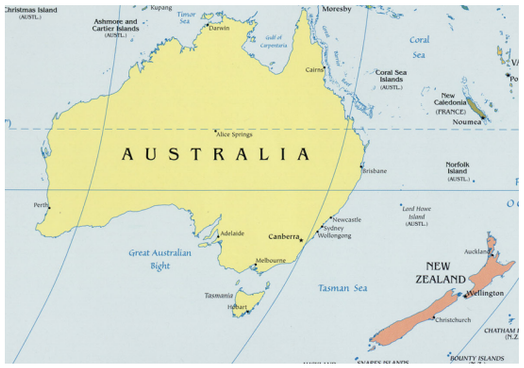
The Australian and New Zealand Intensive Care Society (ANZICS) Adult Patient Database (APD)





The ANZICS APD

- Is one of largest bi-national databases in the world.
- It collects voluntary patient-level admissions data from Intensive Care Units (ICUs) in OZ and NZ.
- 1995-2010: over 1 million individual patient admissions. In 2010, over 80% of eligible ICUs participated.
- Data collected on age, sex, patient severity score APACHE III, diagnostic category, surgical and ventilation status, hospital level, geographical locality, and more.
- APACHE = Acute Physiology And Chronic Health Evaluation score (3rd revision).
- We use **in-hospital mortality** to compare ICU performance.



The ANZICS APD

Data structure is hierarchical:

variability between ICUs

variability between patients within ICUs

⇒ A hierarchical model for mortality.

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

$$i = 1, \ldots, n_j,\ j = 1, \ldots, m$$

where $$Y_{ij} \sim \text{ Bernoulli}(p_{ij})$$

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

What is a key performance indicator?

- It is a summary statistic intended to measure the 'quality' or 'effectiveness' of a hospital's functioning.
- Whilst death could be considered the ultimate 'performance', how much should we attribute to the hospital?
- We want to compare hospitals, distinguishing 'usual' from 'unusual' performance.
- We use the log Standardised Mortality Ratio 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

→ identify extreme ICUs: 'outlier accommodation'.**

Approach II: Fit random effects distribution to *usual* ICUs to obtain a null model

→ identify divergent ICUs: 'outlier detection'

We take a classical approach to **II** *which involves three stages.*

* Ohlssen et al, *JRSS* A, 2007

**Barnett & Lewis, 1978

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

ANZICS APD: patient characteristics in 2009 and 2010

minimum 150 admissions per ICU per year*

Age in years	61.65 (18.20)		Total number of patients = 163795		
APACHE III score	51.28 (27.23)				
ICU mortality (%)	6.51				
Hospital mortality (%)	10.21				
2009-2010 patient volume	1194 (1153)				
	n (%)	Hospital mortality (%)		n (%)	Hospital mortality (%)
Ventilation			**ICU source**		
Not ventilated	94802 (57.88)	6.32	No transfer	151185 (92.30)	9.69
Ventilated	68993 (42.12)	15.56	Hospital transfer	12610 (7.70)	16.48
Gender			**ICU hospital level**		
Male	95128 (58.08)	10.31	Rural	21348 (13.03)	10.07
Female	68667 (41.92)	10.08	Metropolitan	29294 (17.88)	13.17
Patient surgical status			Tertiary	70587 (43.09)	12.74
Non-surgical	96364 (58.83)	13.86	Private	42566 (25.99)	4.06
Elective surgical	47847 (29.21)	2.36	**ICU location**		
Emergency surgical	19584 (11.96)	11.45	NT	2153 (1.31)	10.03
Patient diagnostic category			NSW	51046 (31.16)	10.53
Cardiovascular	40230 (24.56)	15.81	ACT	4014 (2.45)	9.52
Gastrointestinal	28639 (17.48)	8.92	SA	12772 (7.80)	13.71
Metabolic	11424 (6.97)	3.16	VIC	41426 (25.29)	10.28
Neurologic	18216 (11.12)	12.56	WA	3279 (2.00)	11.04
Respiratory	25057 (15.30)	13.94	NZ	9164 (5.60)	13.43
Trauma	9030 (5.51)	8.34	QLD	37337 (22.80)	7.63
Renal/Genitourinary	8612 (5.26)	4.78	TAS	2604 (1.59)	11.56
Hematological	22587 (13.79)	2.24			

*115 ICUs

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

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 AP_{ij} + \sum_{k=2}^{110} \beta_k x_{kij} + U_{0j} + U_{1j} AP_{ij}$$

with

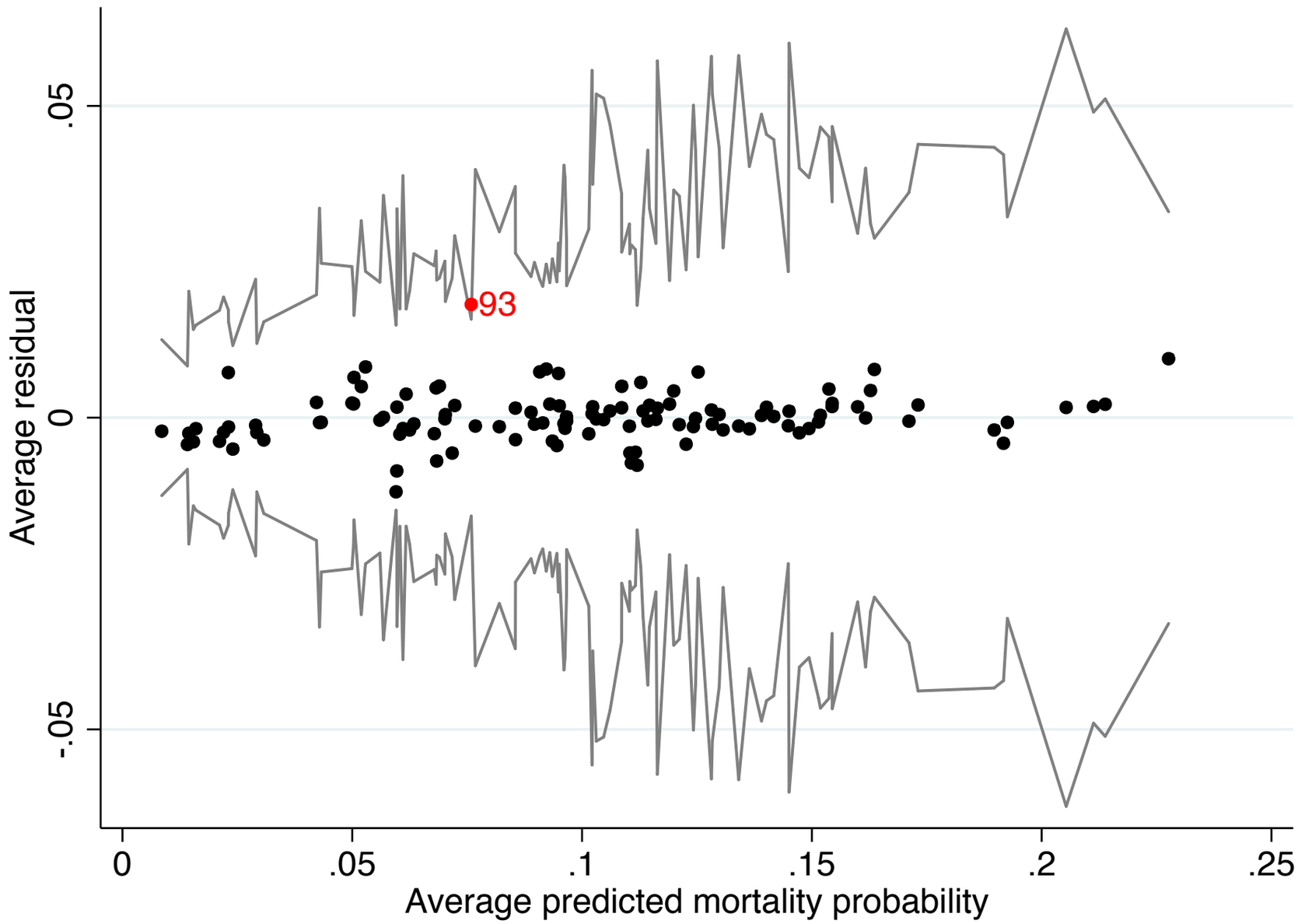
$$\boldsymbol{U}_j \sim N_2(\mathbf{0}, \Sigma)\,.$$

Random intercept will model 'unknown ICU-level variables'.

Model building: 80/20% training/test datasets.
Model fitting: in Stata v12, using AIC, ROC, etc.

Stage 1 model checking: binned residual plot

ICU-level: 115 bins

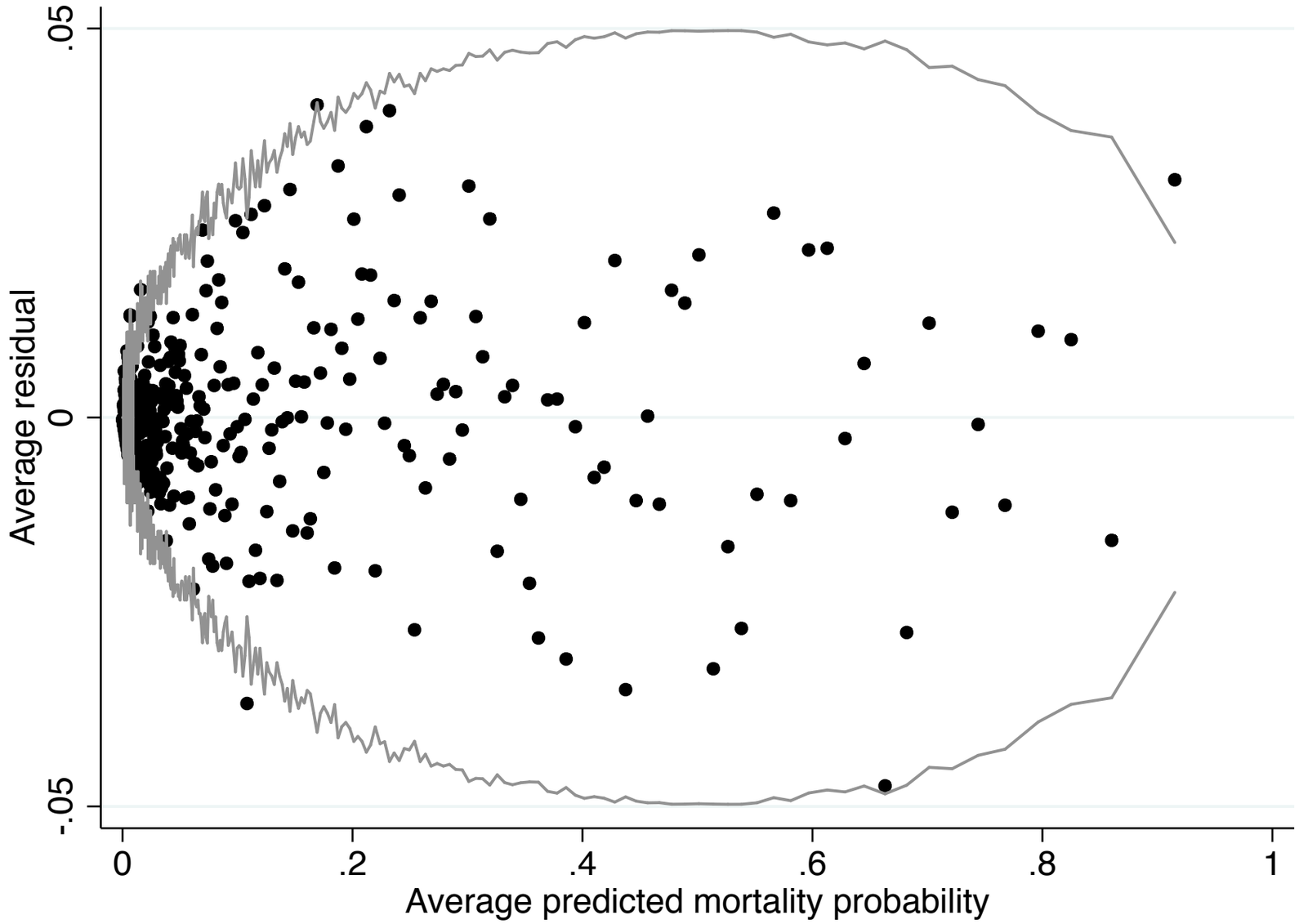


Gelman & Hill, CUP 2007:

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

Stage 1 model checking: binned residual plot

Patient-level: 404 bins



Correct adjustment for casemix is difficult.

Nevertheless, we have a good **empirical** model.

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 p_{ij}^k
- simulate outcome for each patient:

$$Y_{ij}^k \sim \text{Bernoulli}(p_{ij}^k)$$

- 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.

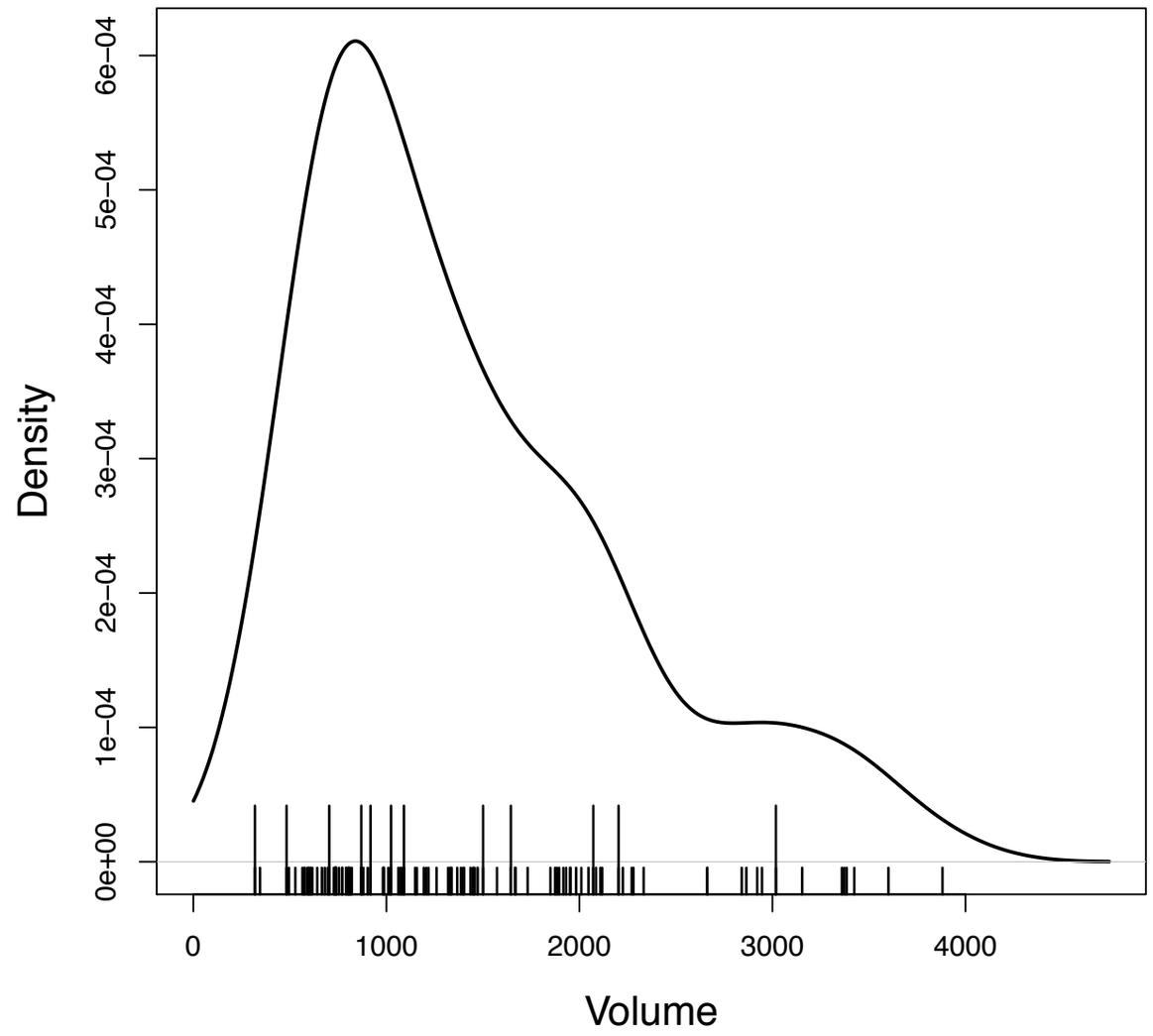
Stage I: here are the potentially unusual ICUs

$p < 0.05$ 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.

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 variance estimates at Stage 1.

Estimating the KPI from the null model:

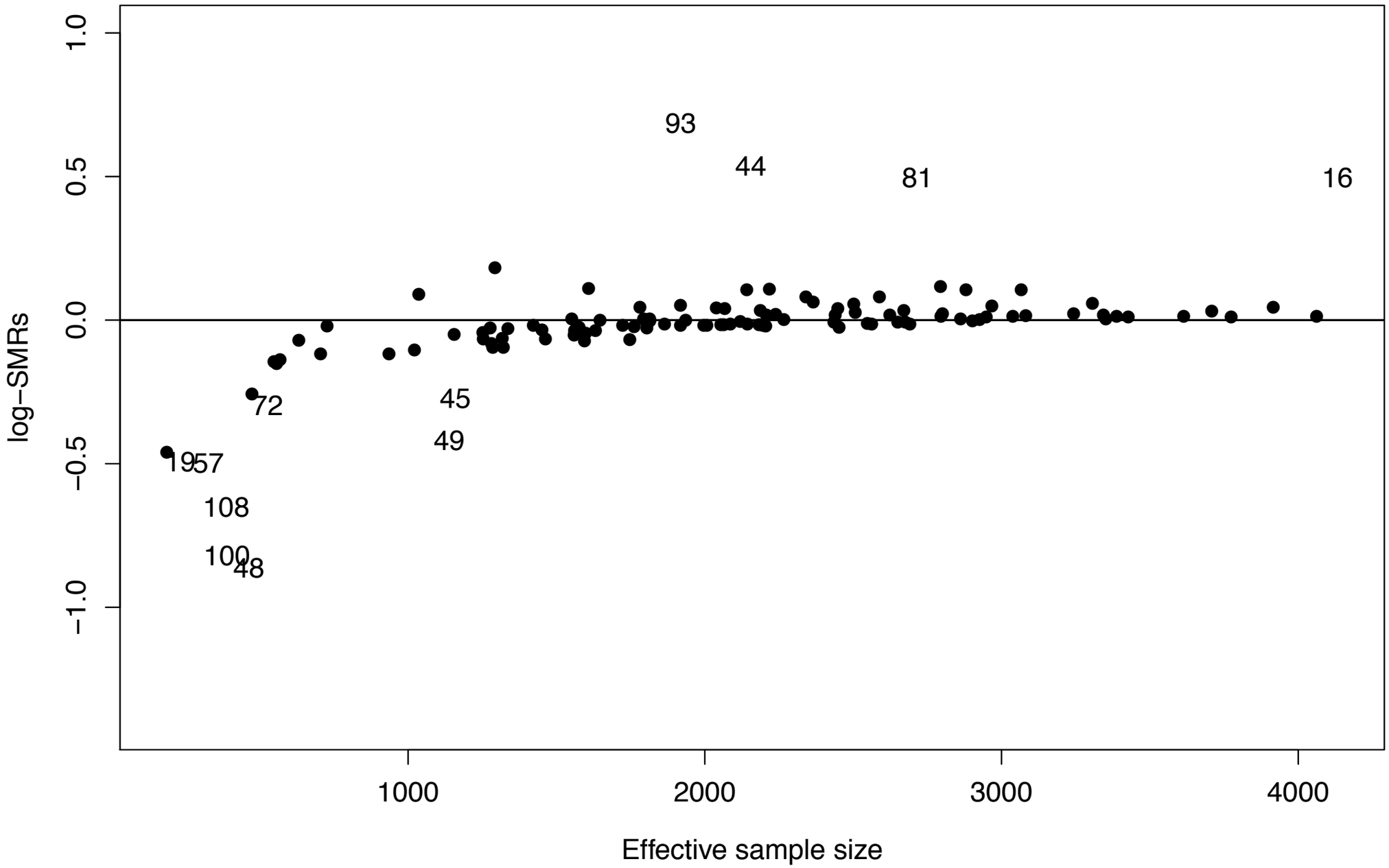
$$\log SMR_j = \log\left(\sum_{i=1}^{n_j} Y_{ij}\right) - \log\left(\sum_{i=1}^{n_j} \hat{p}_{ij}\right)$$

where

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

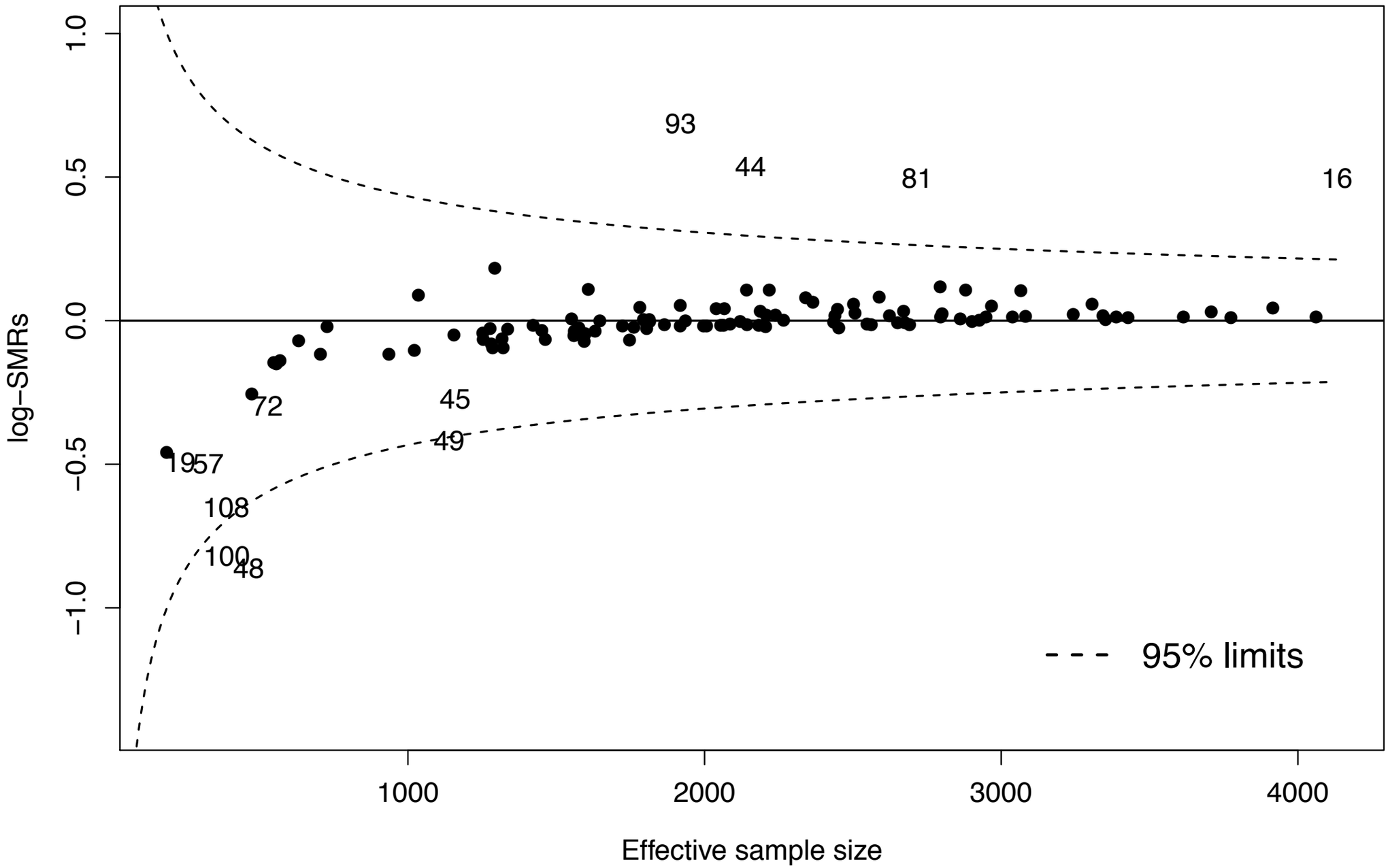
For the potentially unusual ICUs, randomly select a null ICU k and use $\hat{U}_k$

Stage 3: Funnel plot 2009 and 2010



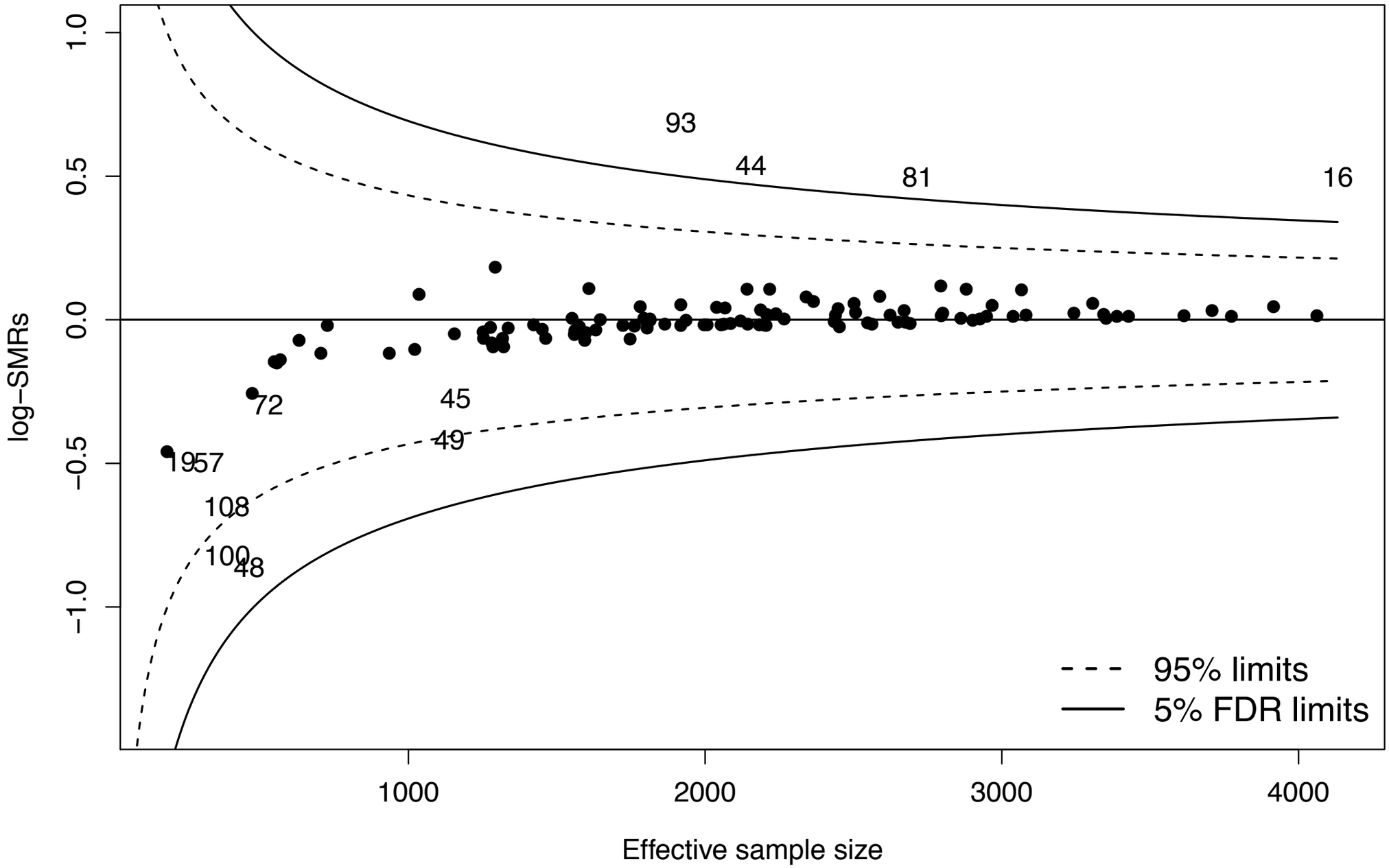
$$n_j = \frac{\sum_{i=1}^{115} \hat{\sigma}_i^2 / 115}{\hat{\sigma}_j^2}$$

Stage 3: Funnel plot 2009 and 2010



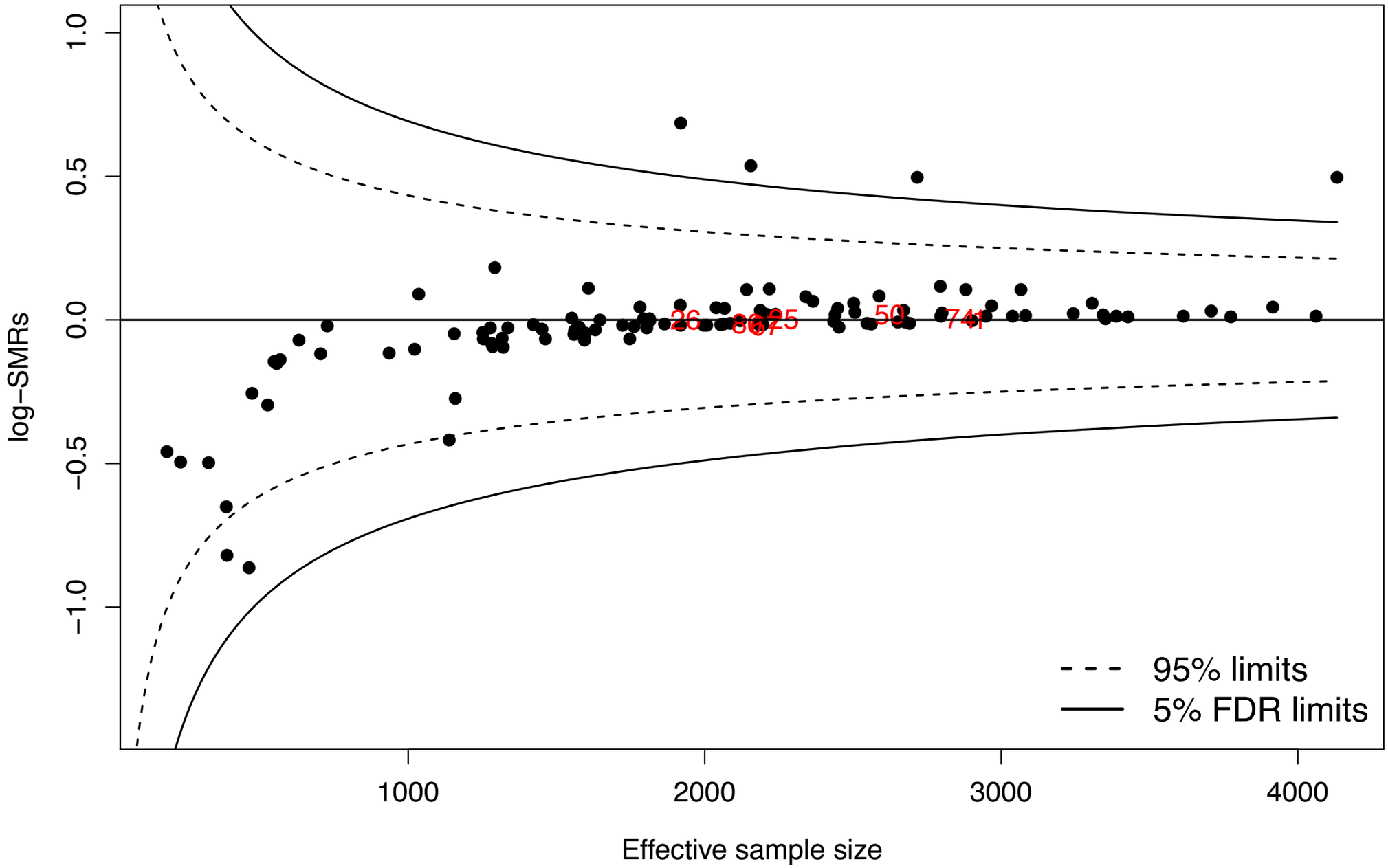
Classical limits: no adjustment for multiple hypothesis tests

Stage 3: Funnel plot 2009 and 2010



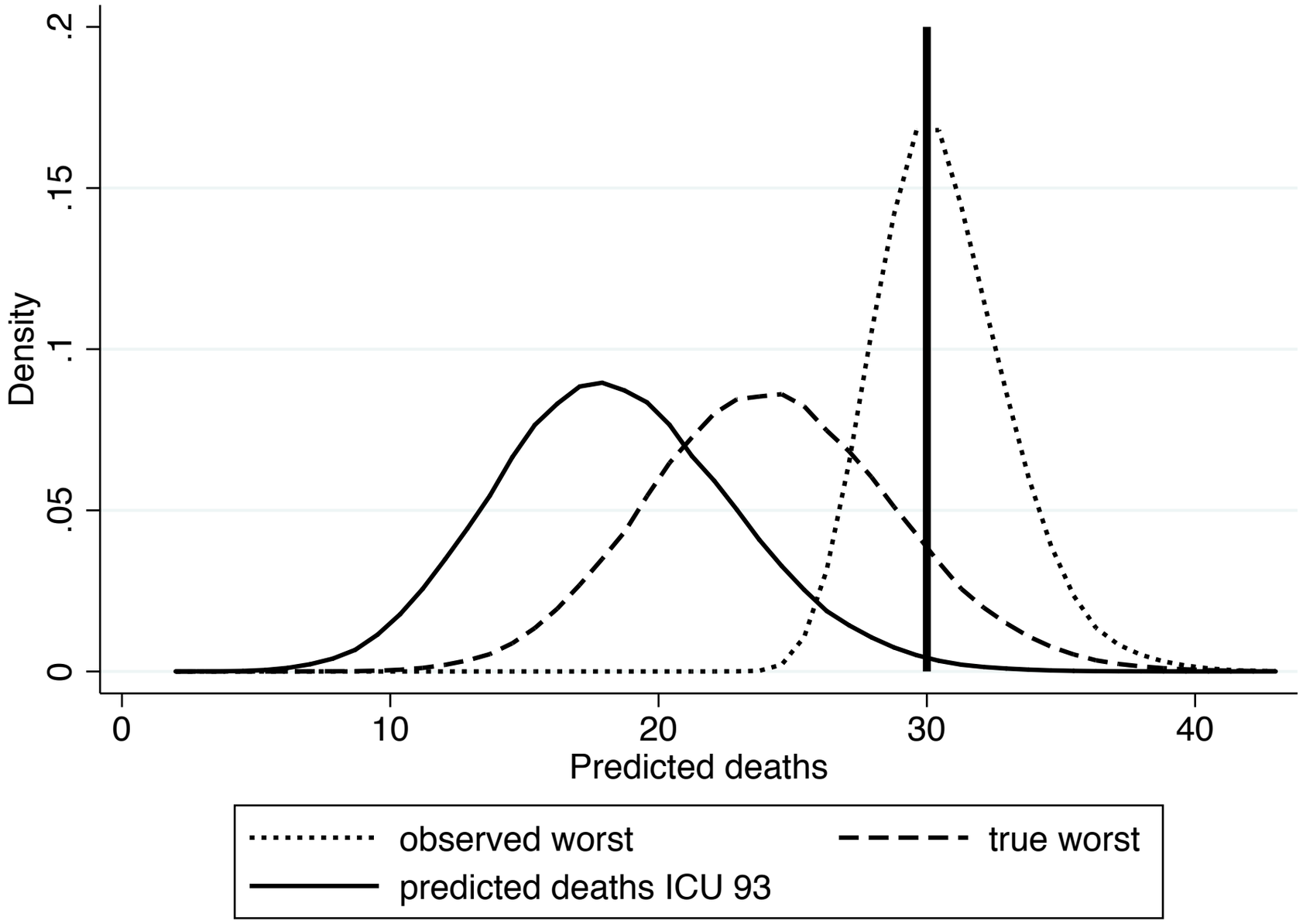
4 Private hospitals have higher than usual mortality: 1 in Vic, 3 in QLD.

Stage 3: Funnel plot 2009 and 2010



7 ‘in control’ New Zealand ICUs

Simulating the 'worst' distributions



Stage 3: we've done visualisation and adjustment for multiple comparisons

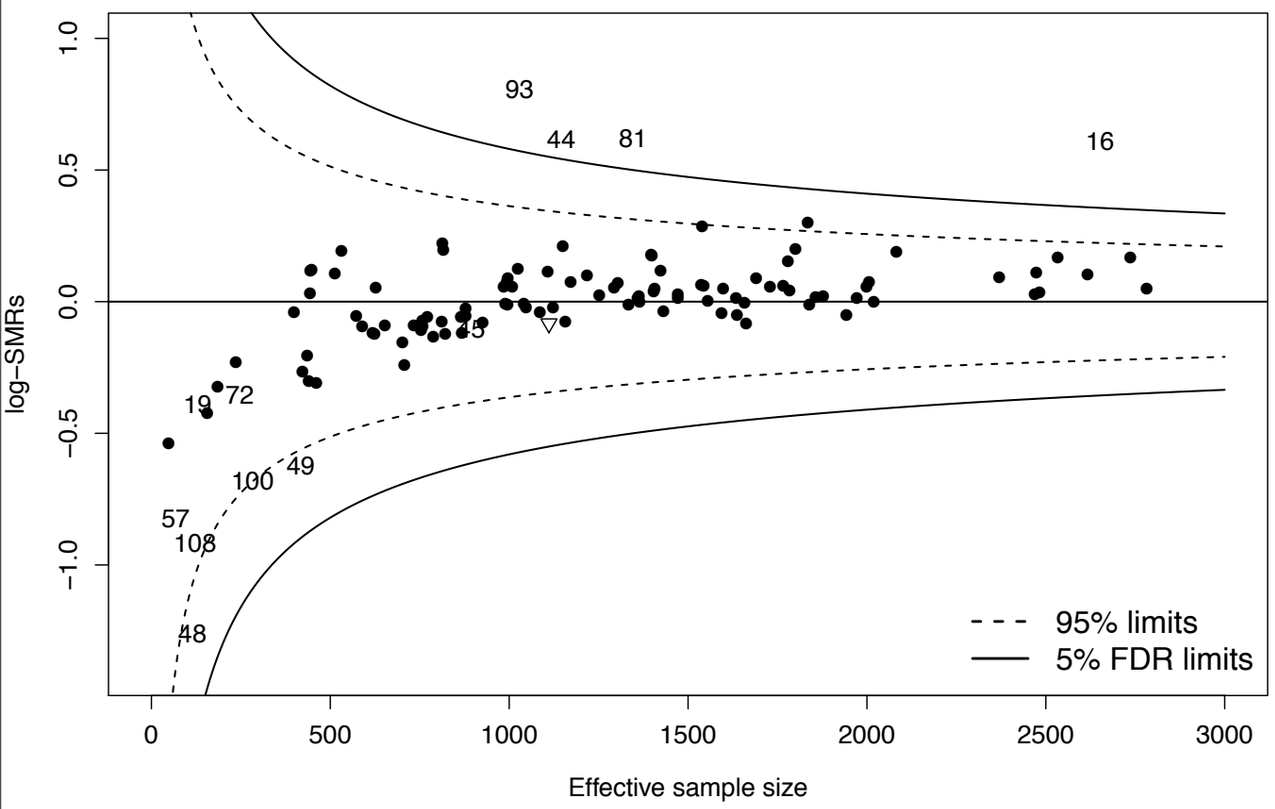
- Excess mortality in four ICUs is not explained by our (extensive) risk adjustment for "usual ICUs".
- *We postulate that these reflect real differences in "**process of care**".*
- ANZICS Centre for Outcome and Resource Evaluation (CORE) has an Outlier Management Policy which concentrates on data-quality issues.*
- Their 2010 analysis using APACHE III-J (an old algorithm, no adjustment for multiple comparisons) identified 2 rural ICUs only.

*www.anzics.com.au/core

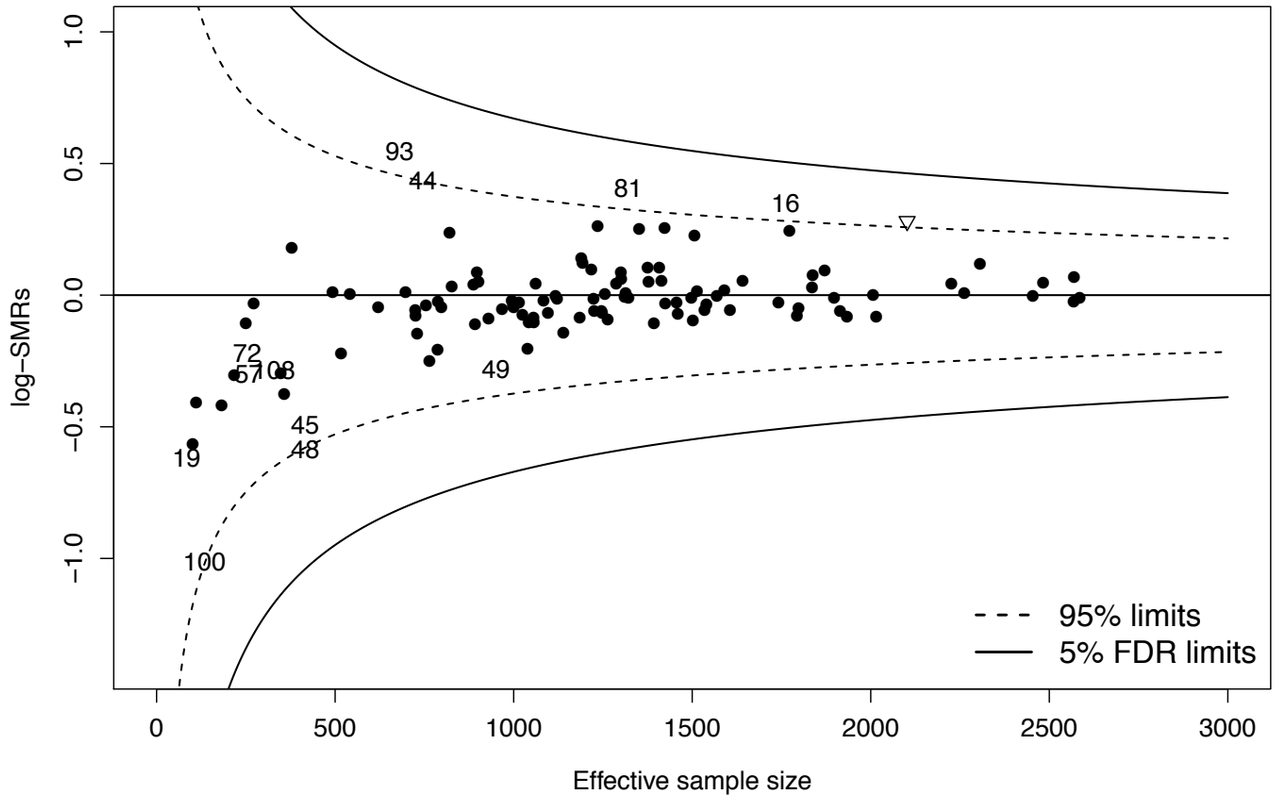
Next stage

Has ICU performance *changed* over time?

2009



2010



Could also plot both years on a single funnel, or ...

... compare years accounting for *regression to the mean**

Use an adjusted measure of ICU change from 2009 to 2010:
(widely used in the **test-retest literature** in education and psychology)

So, instead of considering

$$S_{2010,j} - S_{2009,j}$$

use

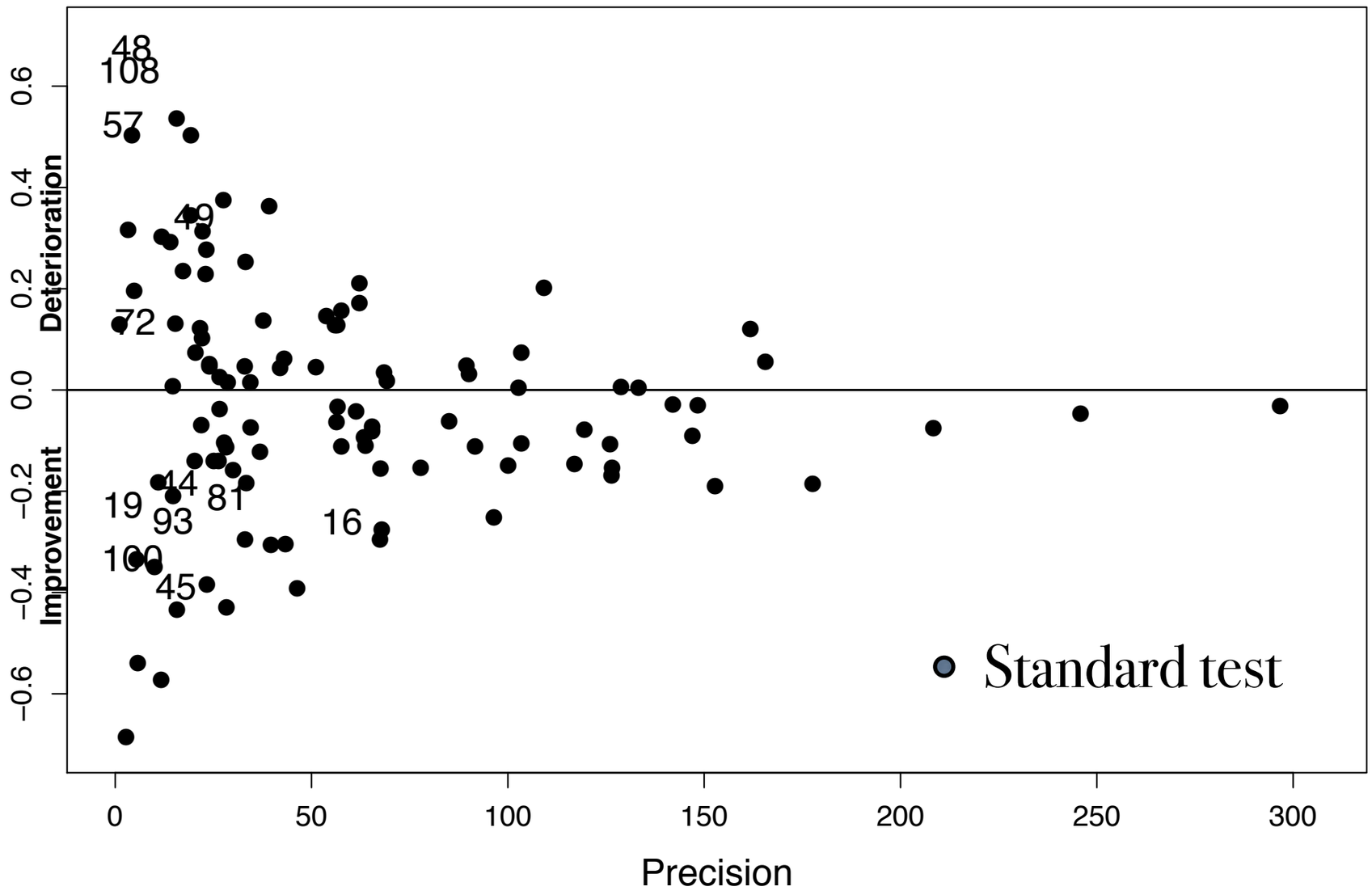
$$S_{2010,j} - E(S_{2010,j}|S_{2009,j})$$ residual change score

This tests the "surprisingness" of $S_{2010,j}$.

*Jones & Spiegelhalter, 2009

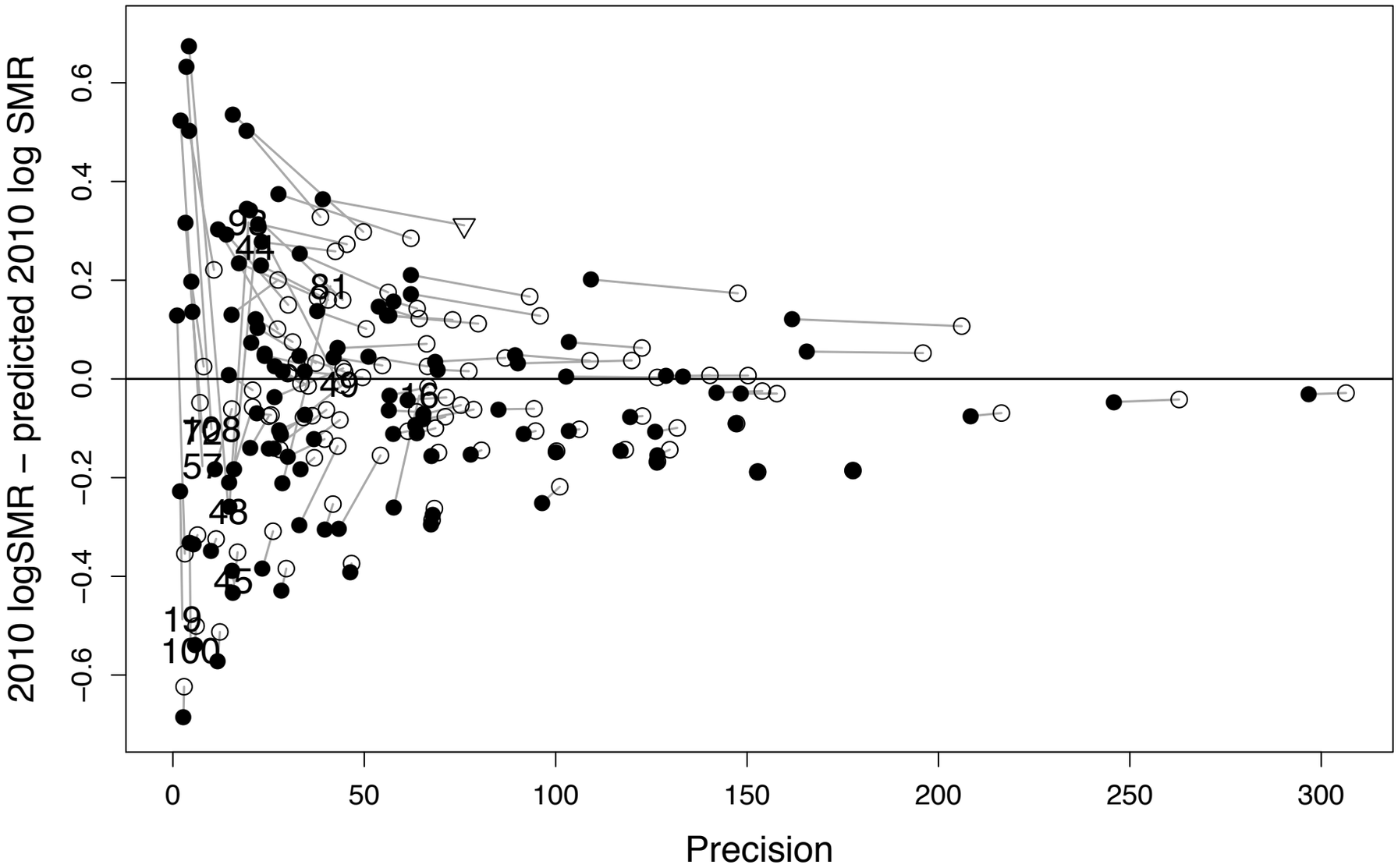
Changes in performance: 2009 to 2010

$S_{2010,j} - S_{2009,j}$



$1/SE(S_{2010,j} - S_{2009,j})$

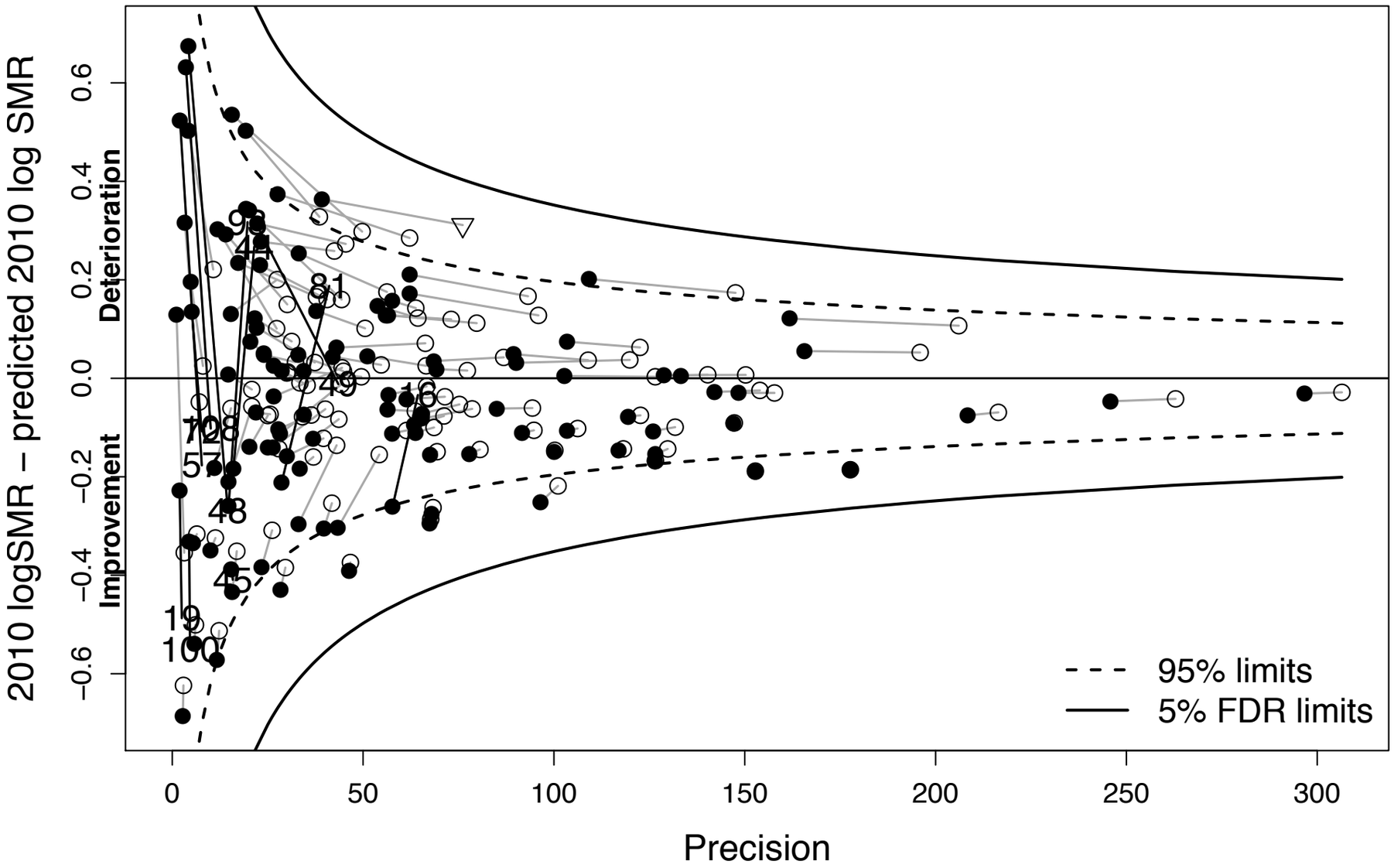
Changes in performance: 2009 to 2010



● Standard test o Adjusted test

Changes in performance: 2009 to 2010

accounting for regression to the mean



● Standard test ○ Adjusted test

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

- Differences in ICU mortality have been identified by our forensic statistical analysis.
- *Are these due to differences in "process of patient care" and therefore performance related? Or, are they due to a run of (good or) bad luck?*
- We are currently analysing 2000-2010 data which may shed light on any systemic problems in intensive care.
- A null random effects distribution representing "usual ICU mortality" is mandated.
- We have used Stata and R. ANZICS CORE use SAS, so we will be making the methodology available in SAS.

Acknowledgements

John Moran

The Queen Elizabeth Hospital

Adelaide

Jessica Kasza

University of Adelaide

Australian Research Council

ANZICS CORE