Predictive modeling and simulation methods for quality assessment and benchmarking in critical care

Niels Peek

Dept. of Medical Informatics

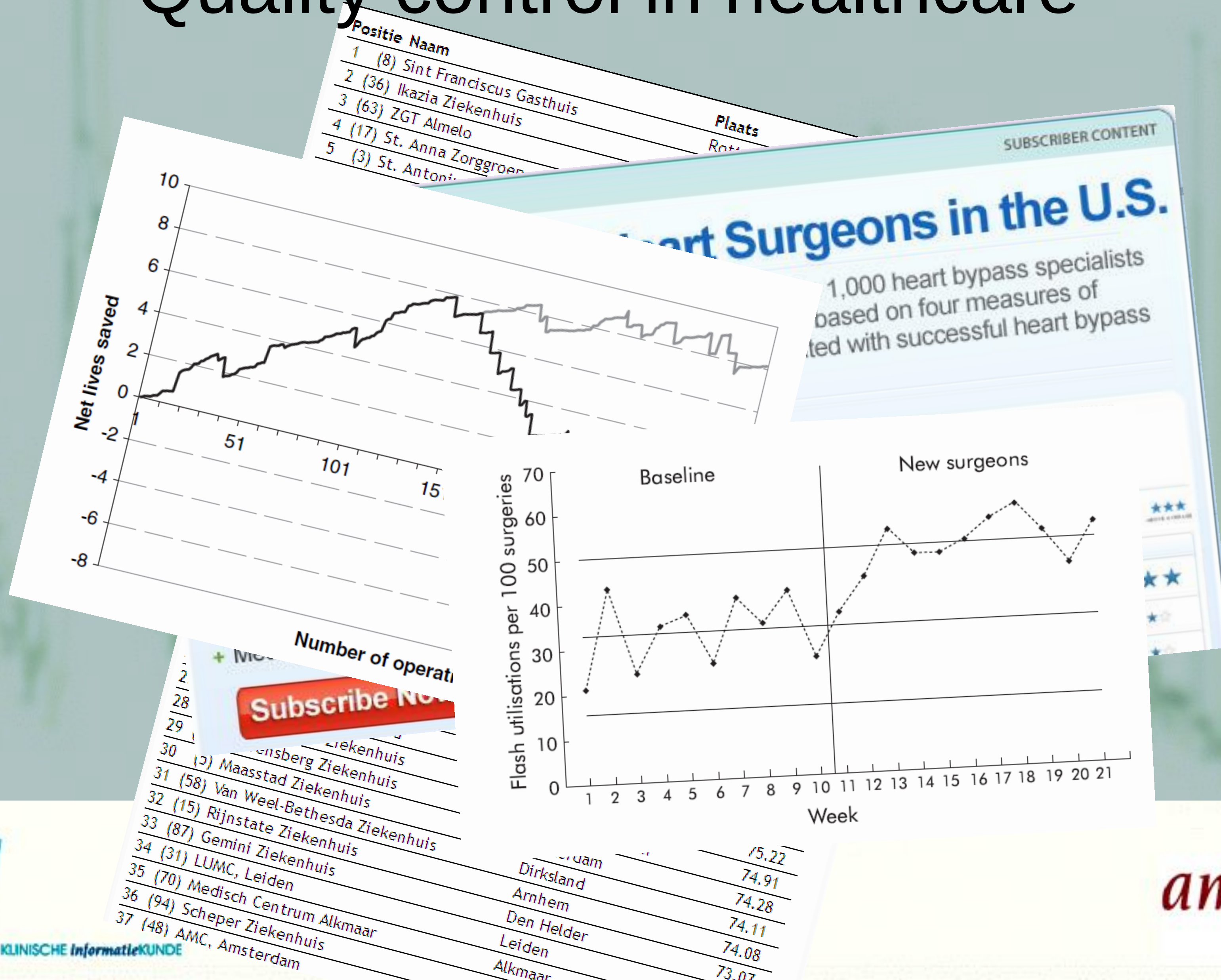
Academic Medical Center
University of Amsterdam

The Netherlands

PDMS Conference Bern, Jan 24th, 2014

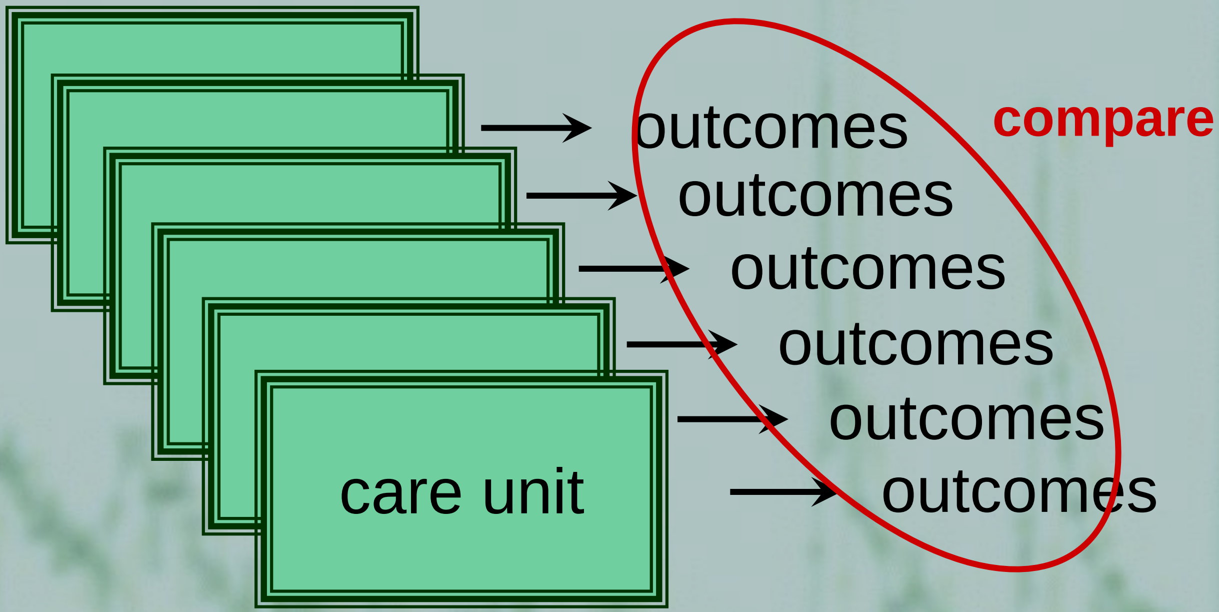
KLINISCHE InformatieKUNDE

Quality control in healthcare



Quality control in healthcare

benchmarking (comparative audit)



Quality control in healthcare

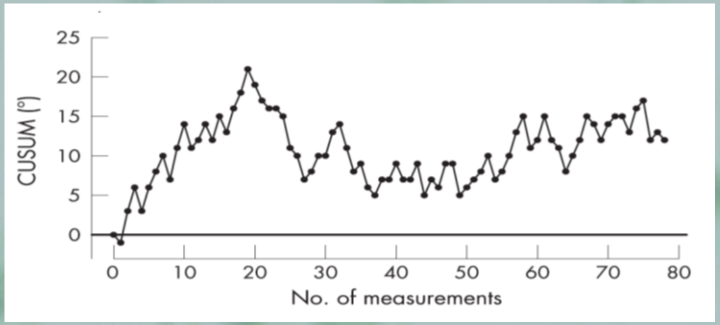
statistical process control

care unit → outcomes

outcomes

outcomes

outcomes



What this talk is about

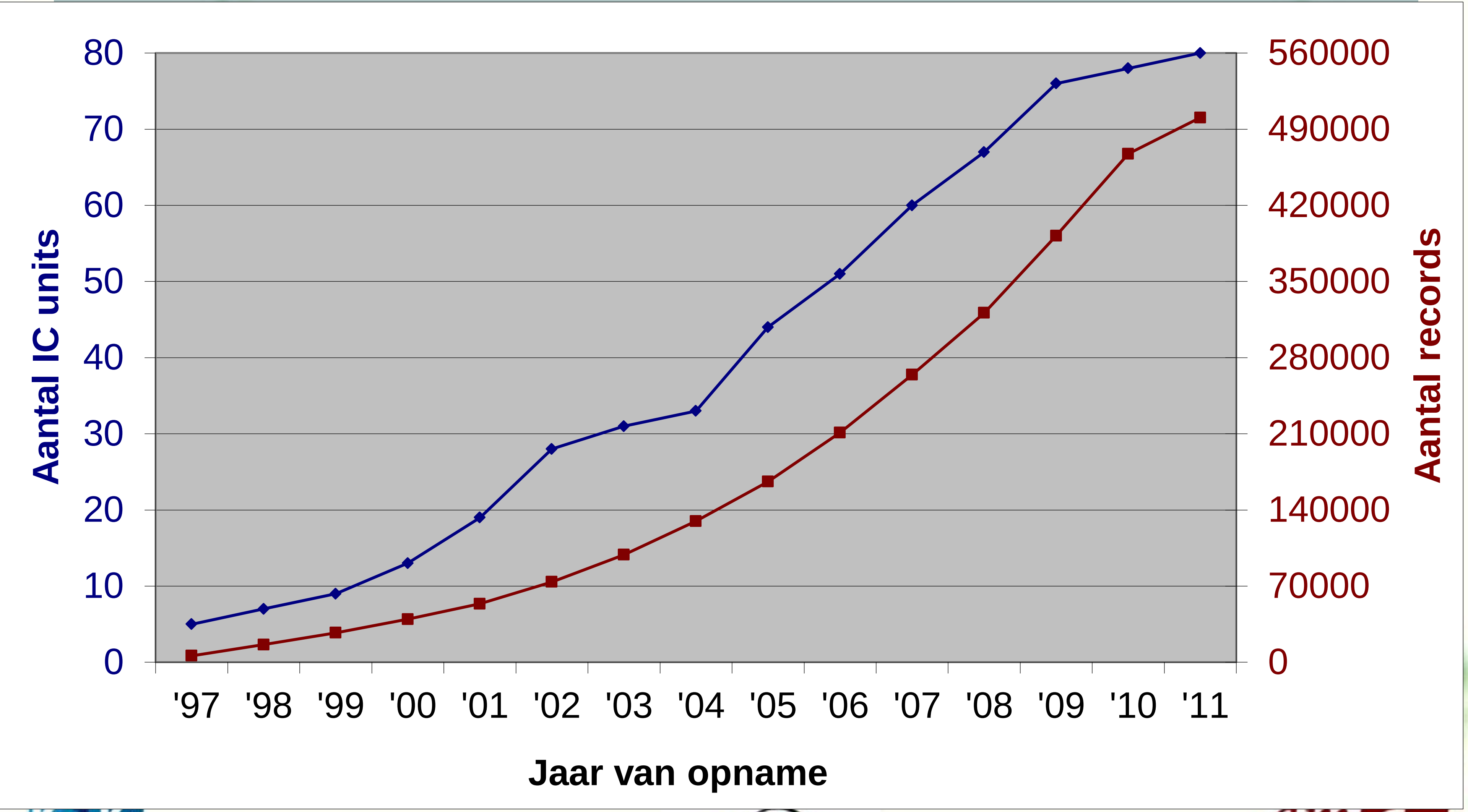
Can we make reliable decisions based on outcome data?

Yes, but we need to know the properties of the quality assessment methods

They can be studied using statistical simulation methods based on resampling

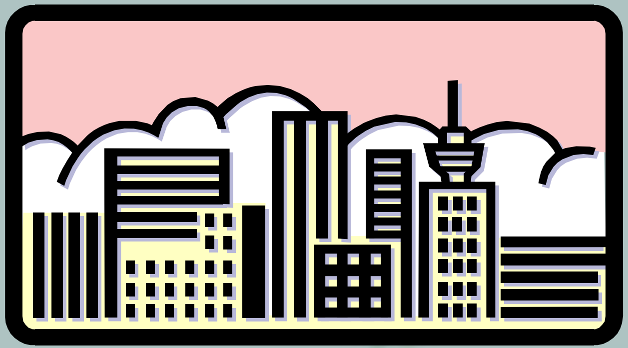
KLINISCHE InformatieKUNDE

The NICE registry

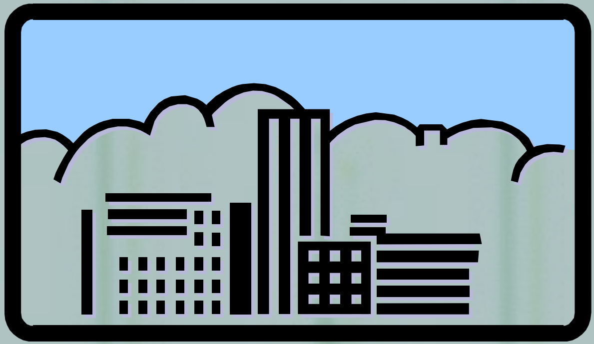


Comparing outcome statistics

A



B



Mortality: **25%** **15%**

Problem: "mix" of admitted patients varies between ICUs

- e.g., urban vs. rural areas
- surgical vs. medical admissions

Case mix variation in the NICE Db

- Admission types
  - Medical 34.1% 6.9% – 81.8%
  - Emergency surgery 14.1% 3.5% – 40.5%
  - Elective surgery 51.8% 16.6% – 92.7%
- Mean age: 62.2 55.4 – 67.1
- Chronic diagnoses: 24.5% 8.6 – 51.8

KiK KLINISCHE InformatieKUNDE

NICE

amc

Case-mix correction

patient
case mix

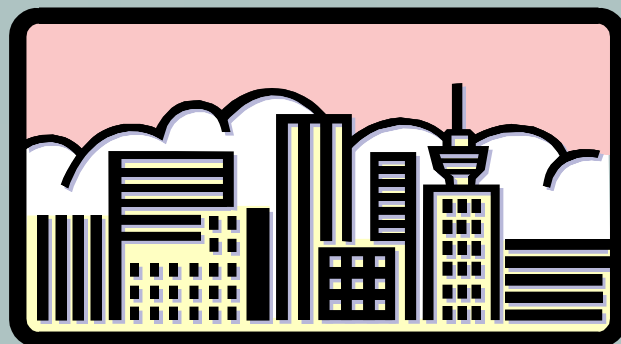
prognostic
model

expected
outcomes

compare

Standardized mortality ratio (SMR)

A



B

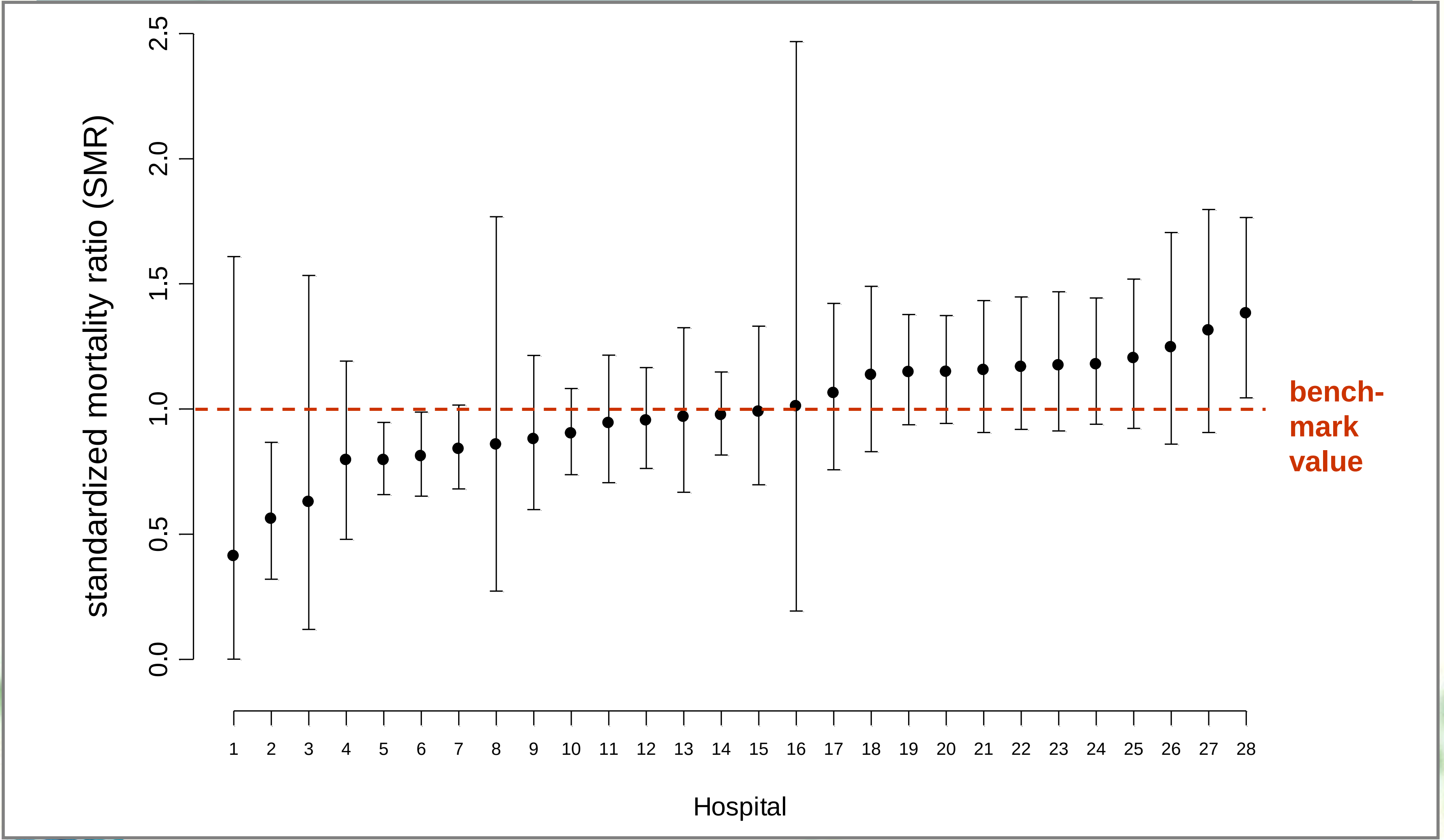
	A	B
Observed mortality:	**25%**	**15%**
Expected mortality:	**30%**	**12%**
SMR:	**0.83**	**1.25**

KiK KLINISCHE InformatieKUNDE

NICE

amc

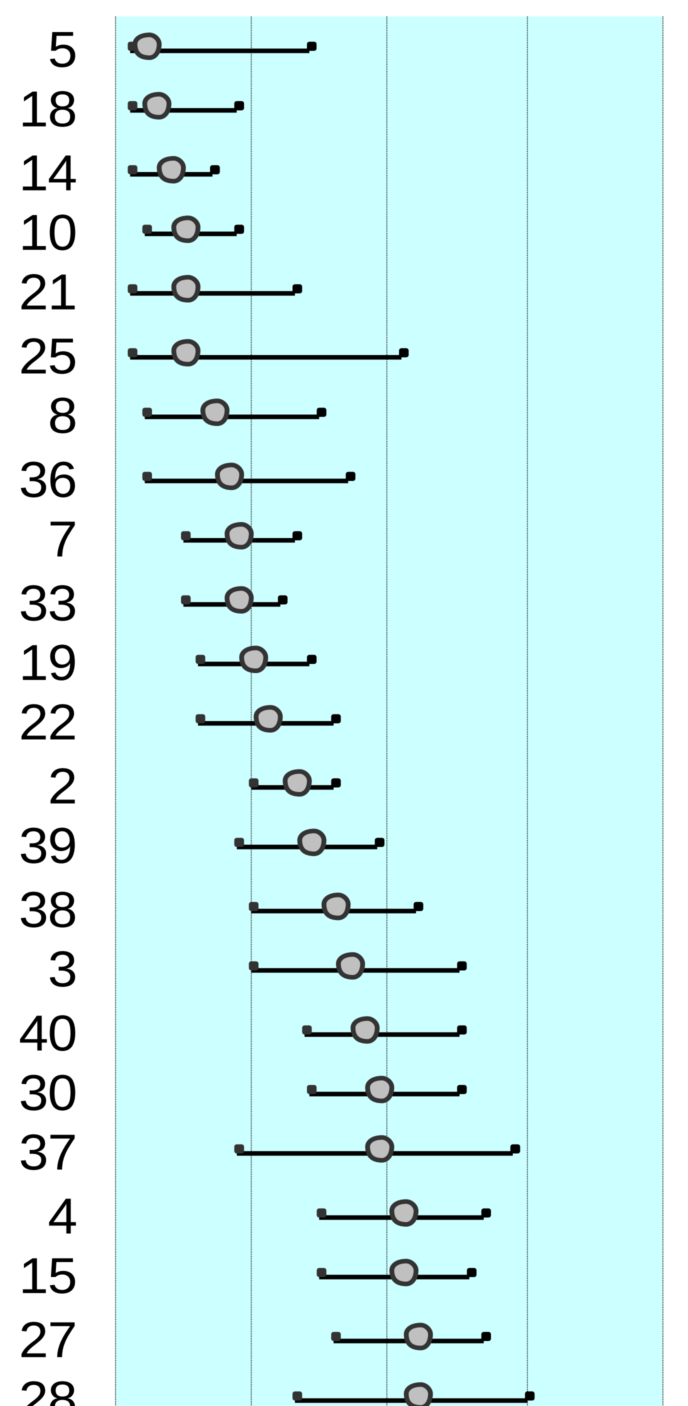
Institutional profiling



A pitfall with ra

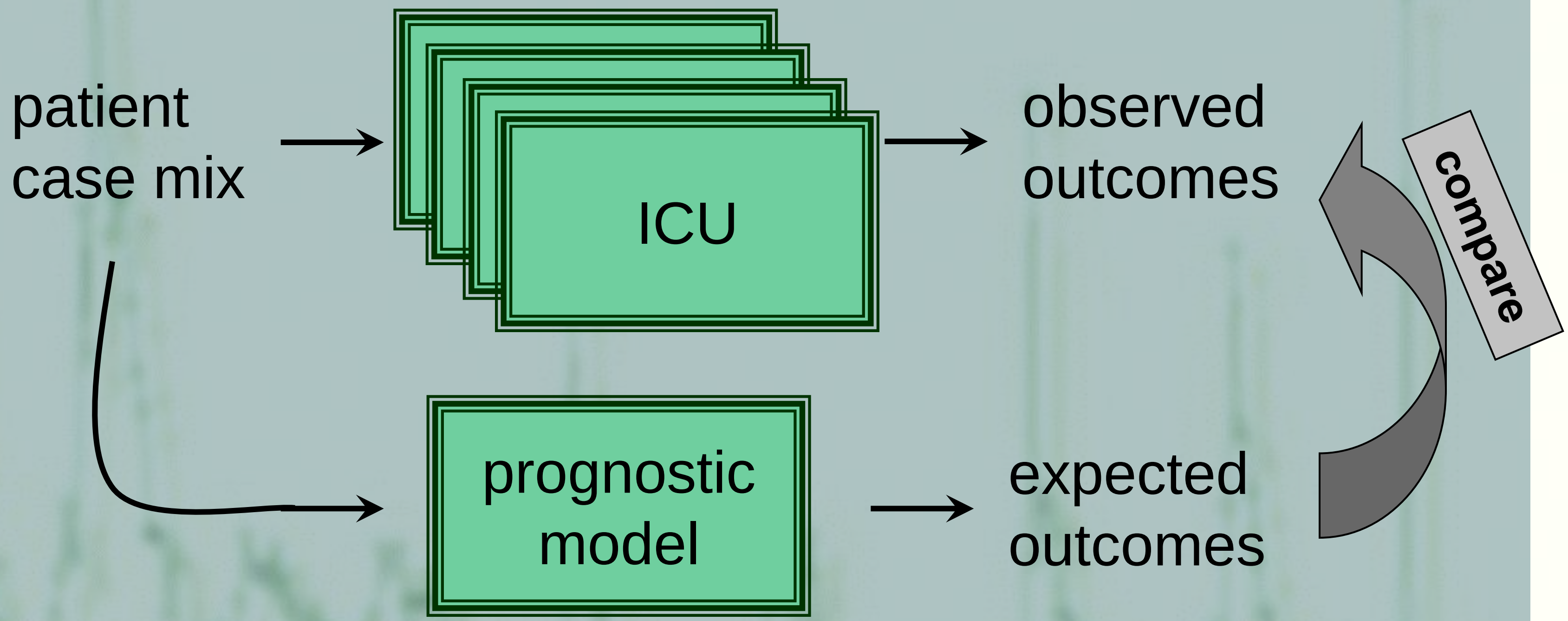
- League table ranks are based on SM
finite samples
- Therefore ranks are statistical quant
subject to uncertainty
- The uncertainties in ranks can be as
resampling methods, such as boots

Example: ICU ranks with 95% CIs
(40 ICUs, 10,000 bootstrap samples)



KLINISCHE InformatieKUNDE

How do we know that the model is right?



Compute measures of discrimination (e.g. area under ROC curve) and calibration (e.g. Brier score) els

BMJ | BMJ|Journals | BMJ

BMJ

CRITICAL CARE

IMPACT FACTOR 3.12

Poster presentation

SAPS 2 is a better score than APACHE II to predict mortality in the ICU

G Nobre, M Kalichsztein, J Kezen, F Braga, G Almeida, G Penna, P Kurtz, P Araujo, R Vegni, M Freitas and C Valdez
Casa de Saúde São José, Rio de Janeiro, RJ, Brazil

from 26th International Symposium on Intensive Care and Emergency Medicine
Brussels, Belgium. 21–24 March 2006

Critical Care 2006, **10**(Suppl 1):P408 doi:10.1186/cc4755

The electronic version of this abstract is the complete one and can be found online at: http://ccforum.com/supplements/10/S1

Published 21 March 2006

Background

The prediction of mortality in the ICU is very important to evaluate the quality of the care for our patients. The two most used scores that are used are the APACHE II and the SAPS 2, but there are conflicting results in the literature regarding which of them is the best predictor tool.

PubMed
Add article to my folders
Download to citation manager

pressure and heart rate) in the first 24 hours in intensive care to calculate each patient's risk of dying. Any score that uses data collected over 24 hours is affected by the quality of care provided [2 3] — the very thing that units are

KLINISCHE InformatieKUNDE

ELSEVIER

Journal of Clinical Epidemiology 60 (2007) 491–501

Journal of Clinical Epidemiology

External validation of prognostic models for critically ill patients required substantial sample sizes

- Random subsamples were taken from the NICE Db of varying size (n= 250 / 500 / 750 / 1,000 / 2,500 / 5,000)
- Each time the model validation and comparison process was performed
- This was repeated for 500 times
- Differences in performance between Apache II, SAPS II and MPM_{24} II were small …
- … and with small sample sizes (up to n=1,000) the results were extremely variable

Does it matter which model we use?

The impact of different prognostic models and their customization on institutional comparison of intensive care units*

Ferishta Bakhshi-Raiez, MSc; Niels Peek, PhD; Robert J. Bosman, MD; Evert de Jonge, MD, PhD; Nicolette F. de Keizer, PhD

Objectives: To evaluate the influence of choice of a prognostic model and the effect of customization of these models on league tables (i.e., rank-order listing) in which intensive care units (ICUs) are ranked by standardized mortality ratios using Acute Physiology and Chronic Health Evaluation (APACHE) II, Simplified Acute Physiology Score (SAPS) II, and Mortality Probability Model II (MPM_{24}II).

Design: Retrospective analysis of prospectively collected data on ICU admissions.

Setting: Forty Dutch ICUs.

Patients: A data set from a national registry of 86,427 patients from January 2002 to October 2006.

Interventions: The league tables associated with the different models were compared to evaluate their agreement. Bootstrapping was used to quantify the uncertainty in the ranks for ICUs. First, for each ICU the median rank and its 95% confidence interval were identified for each model. Then, for a given pair of models, for each ICU the median difference in rank and its associated 95% confidence interval were computed. A difference in rank for an ICU for a given pair of models was considered relevant if it was statistically significant and if one of the models would categorize this ICU as a performance outlier (excellent performer or very poor performer) while the other did not.

Measurements and Main Results: For 20 ICUs, there was a significant difference in rank (2–19 positions) between one or more pairs of models. Three ICUs were rated as performance outliers by one of the models, while the other excluded this possibility with 95% certainty. Furthermore, for ten ICUs, one or more pairs of models classified these ICUs as performance outliers while the other model did not do so with certainty. Regarding the agreement between the original models and their customized versions, in all cases the median change in rank was three positions or less and the models fully agreed with respect to which ICUs should be classified as performance outliers.

Conclusions: Institutional comparison based on case-mix adjusted league tables is sensitive to the choice of prognostic model but not to customization of these models. League tables should always display the uncertainty associated with institutional ranks. (Crit Care Med 2007; 35:2553–2560)

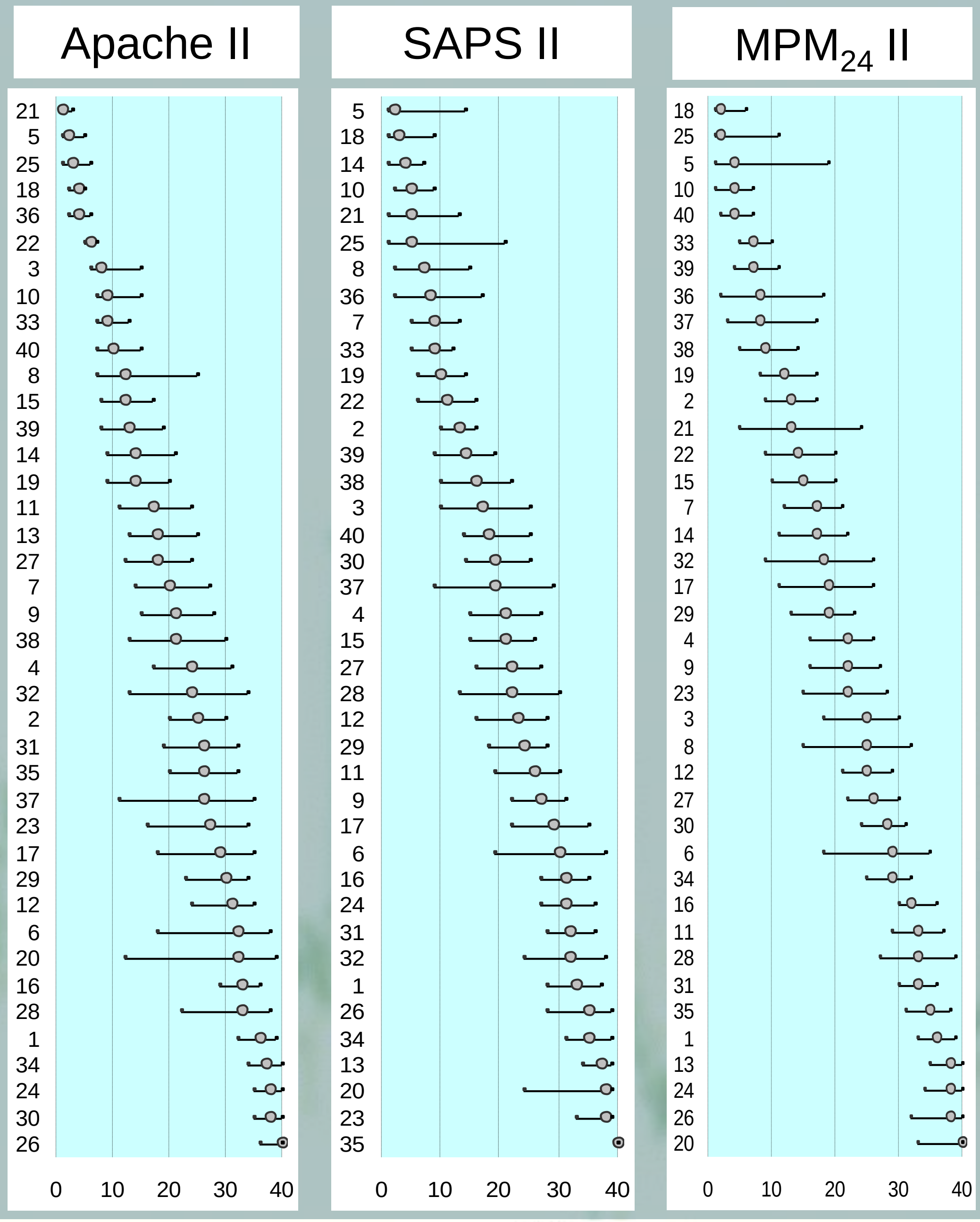
Key Words: healthcare evaluation mechanisms; benchmarking; Simplified Acute Physiology Score II; Acute Physiology and Chronic Health Evaluation II; Mortality Probability Model II

The impact of different prognostic models and their customization on institutional comparison of intensive care units*

Ferishta Bakhshi-Raiez, MSc; Niels Peek, PhD; Robert J. Bosman, MD; Evert de Jonge, MD, PhD; Nicolette F. de Keizer, PhD

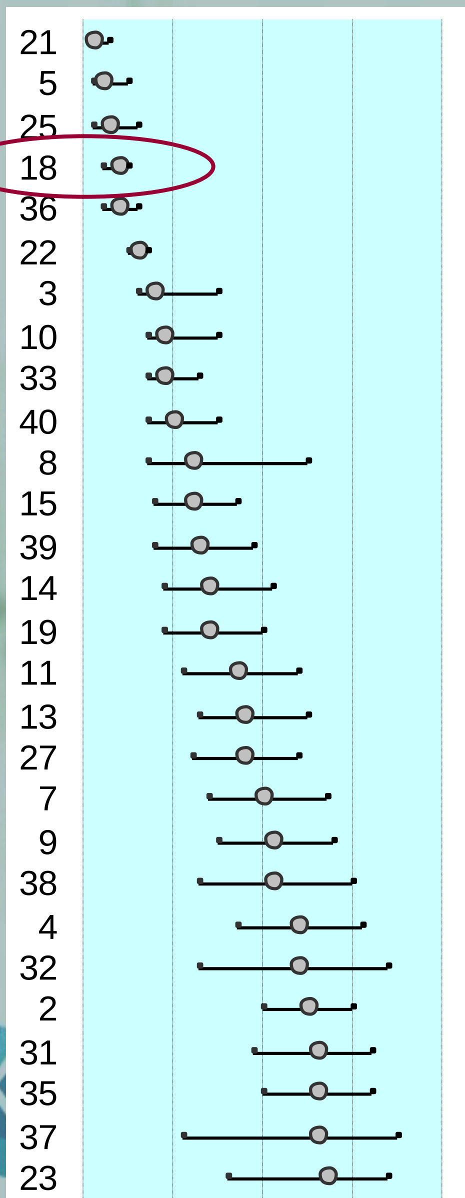
- League tables were constructed using case-mix correction with Apache II, SAPS II and MPM_{24} II
- 95% confidence intervals were constructed for the ranks of individual ICUs, using bootstrap sampling (10,000 replications)

KLINISCHE InformatieKUNDE

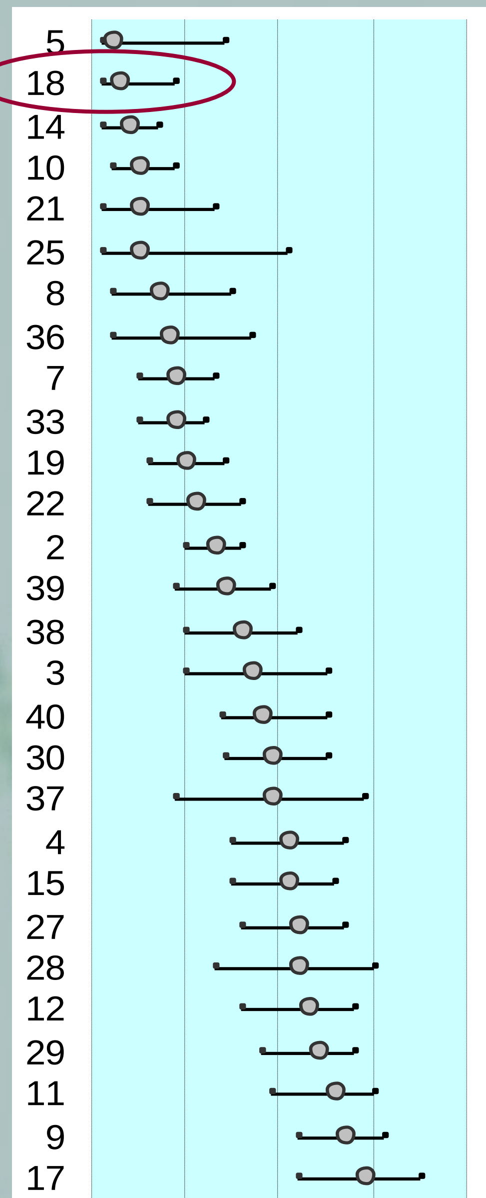


KLINISCHE InformatieKUNDE

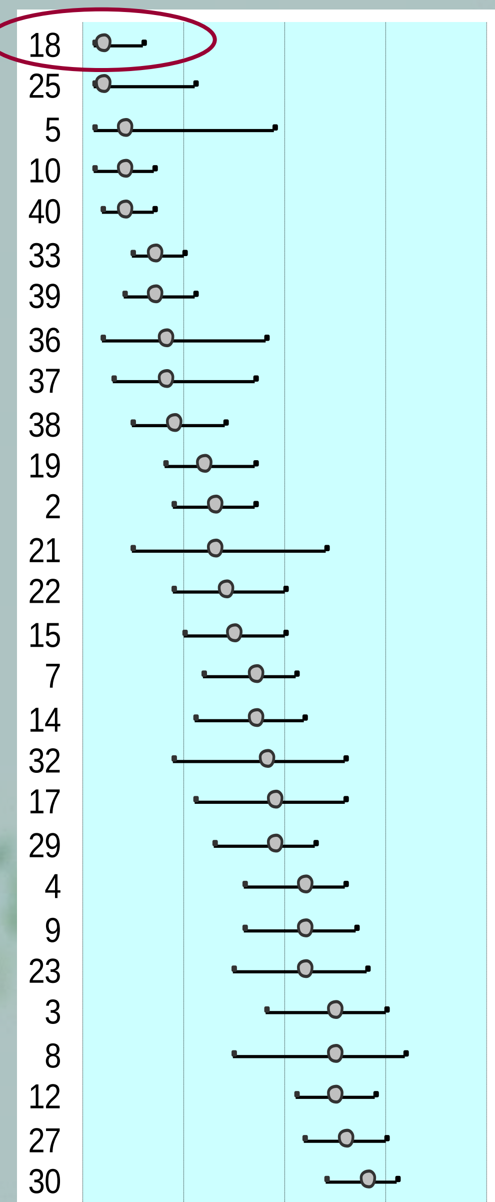
Apache II

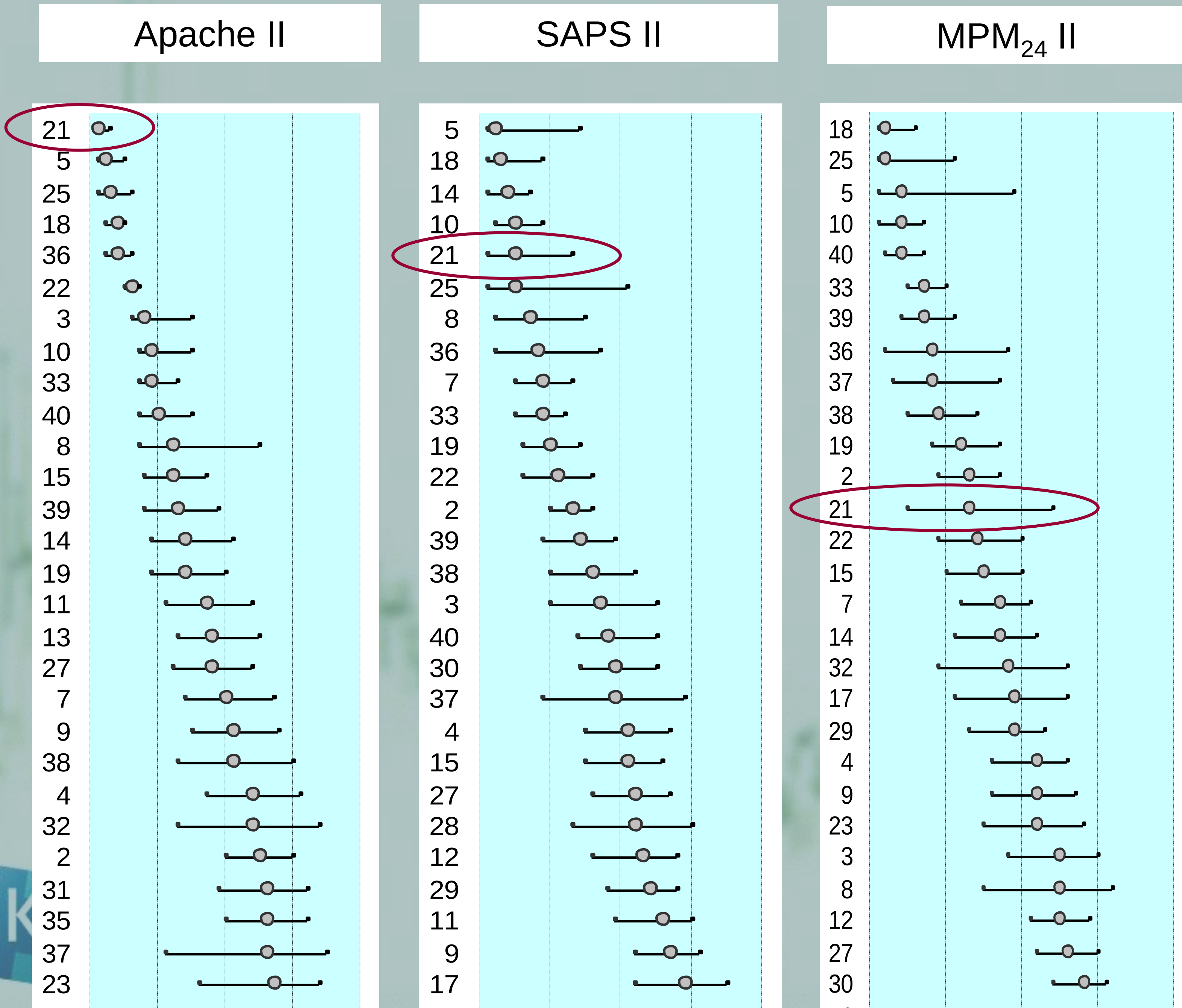


SAPS II

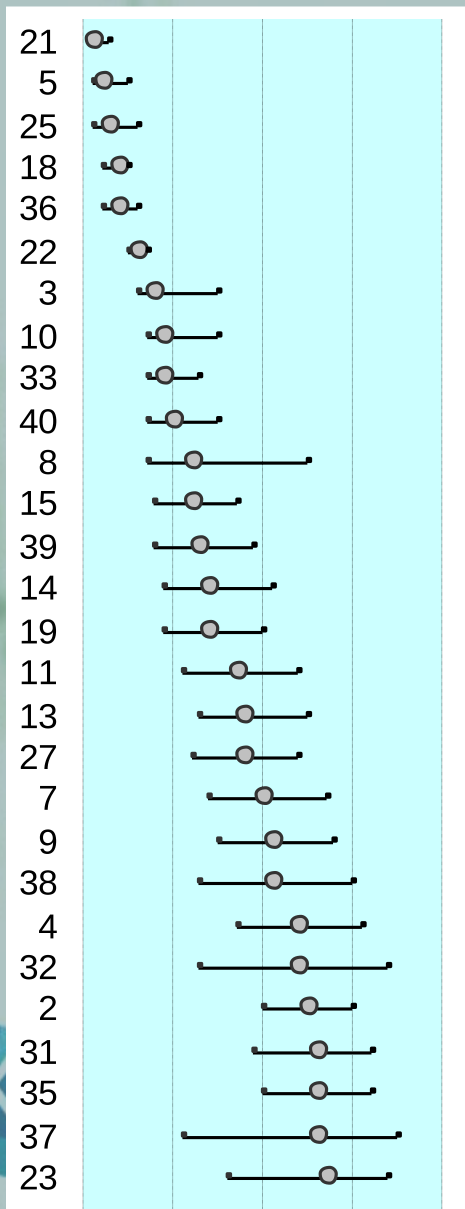


MPM_{24} II

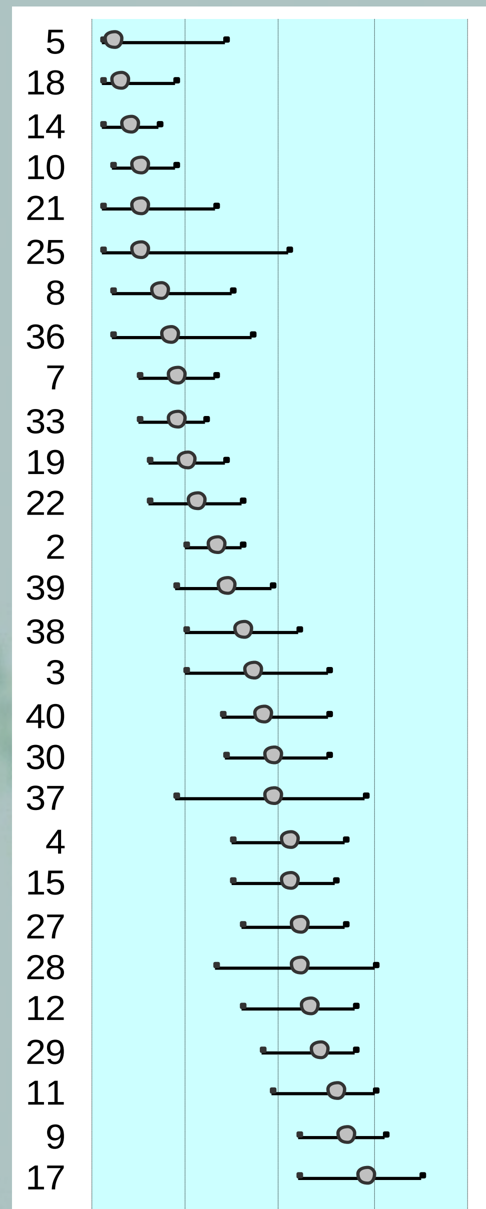



Apache II
21
5
25
18
36
22
3
10
33
40
8
15
39
14
19
11
13
27
7
9
38
4
32
2
31
35
37
23
SAPS II
5
18
14
10
21
25
8
36
7
33
19
22
2
39
38
3
40
30
37
4
15
27
28
12
29
11
9
17
MPM24 II
18
25
5
10
40
33
39
36
37
38
19
2
21
22
15
7
14
32
17
29
4
9
23
3
8
12
27
30

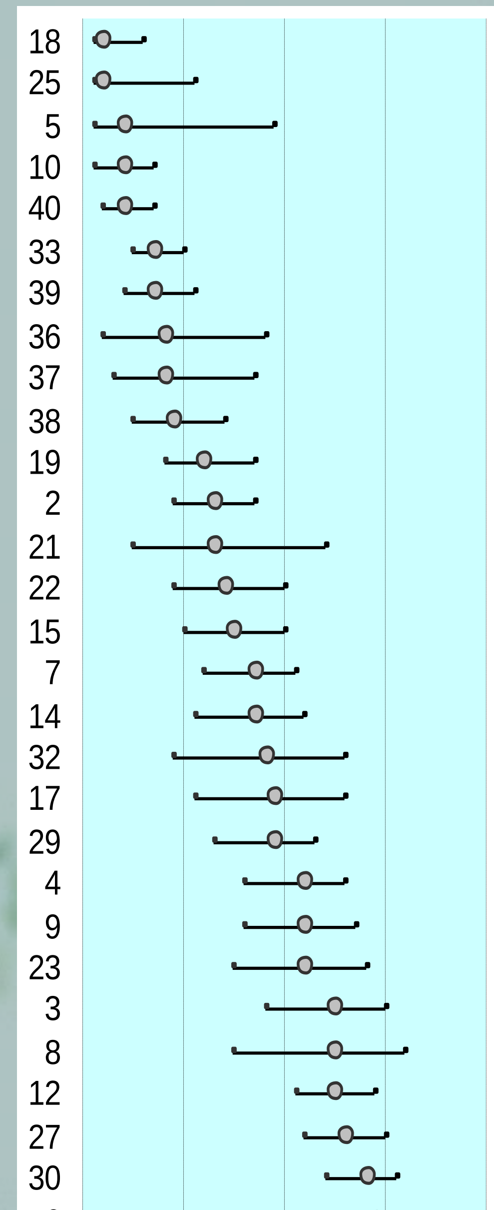
Apache II

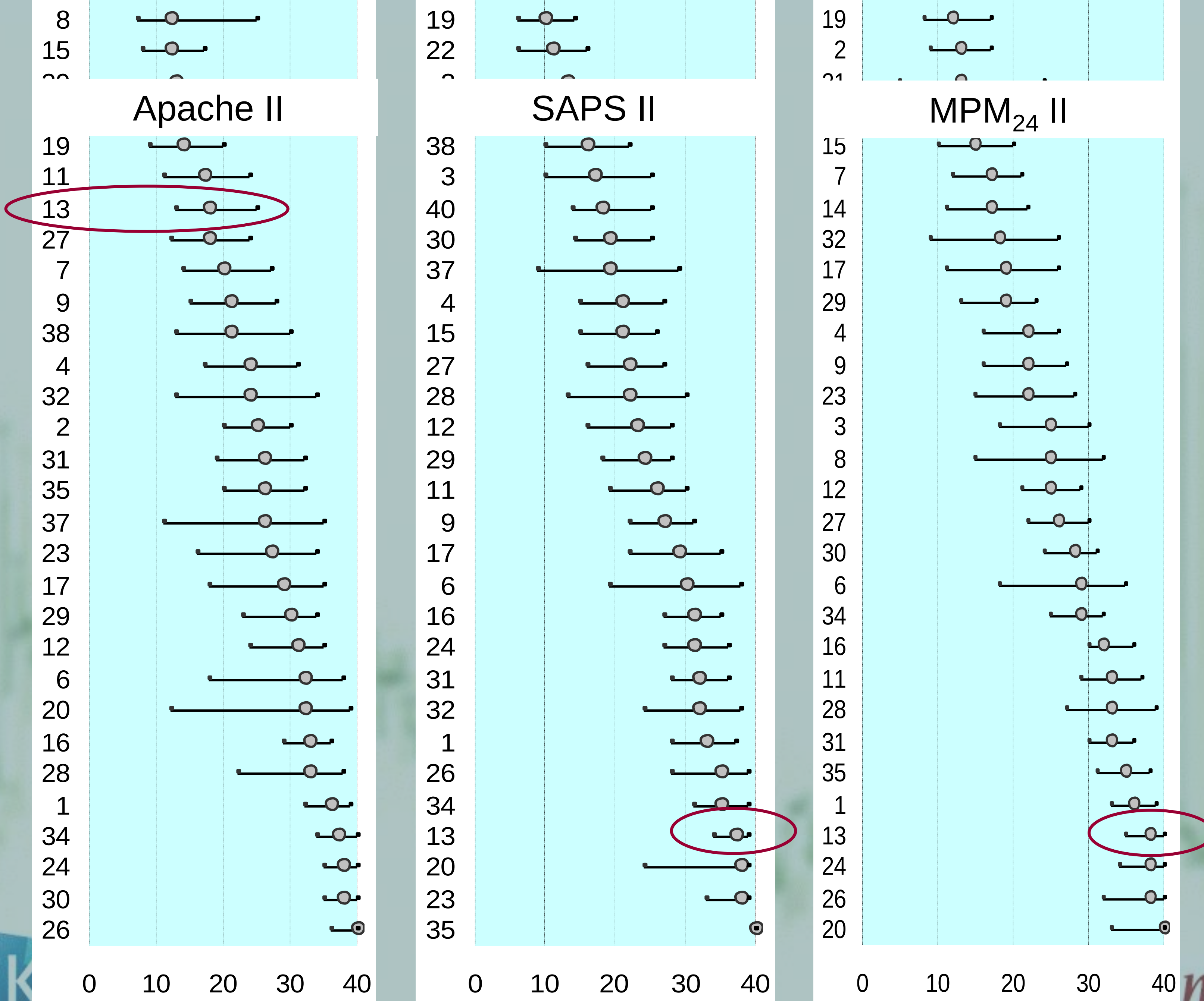


SAPS II



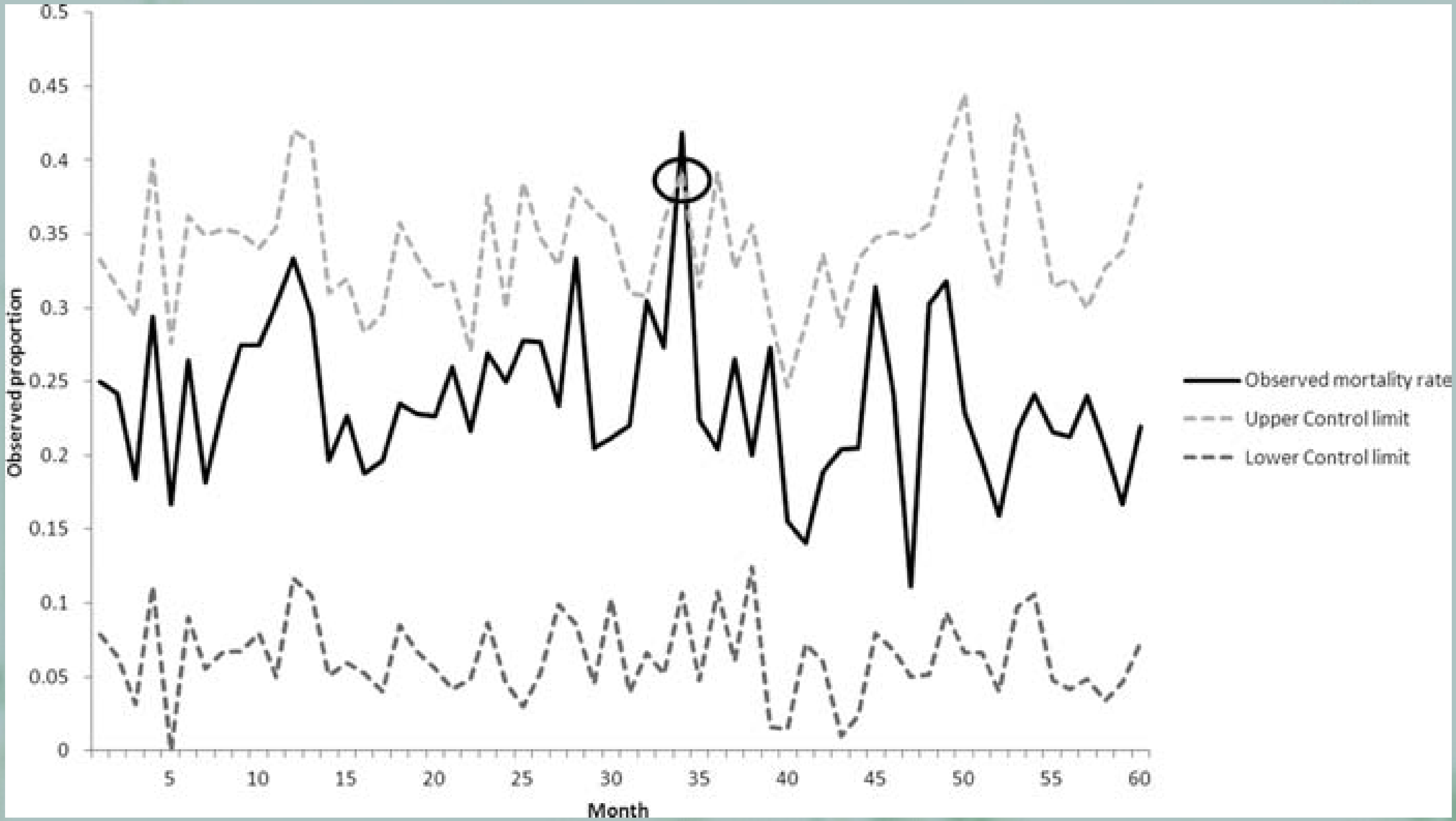
MPM_{24} II





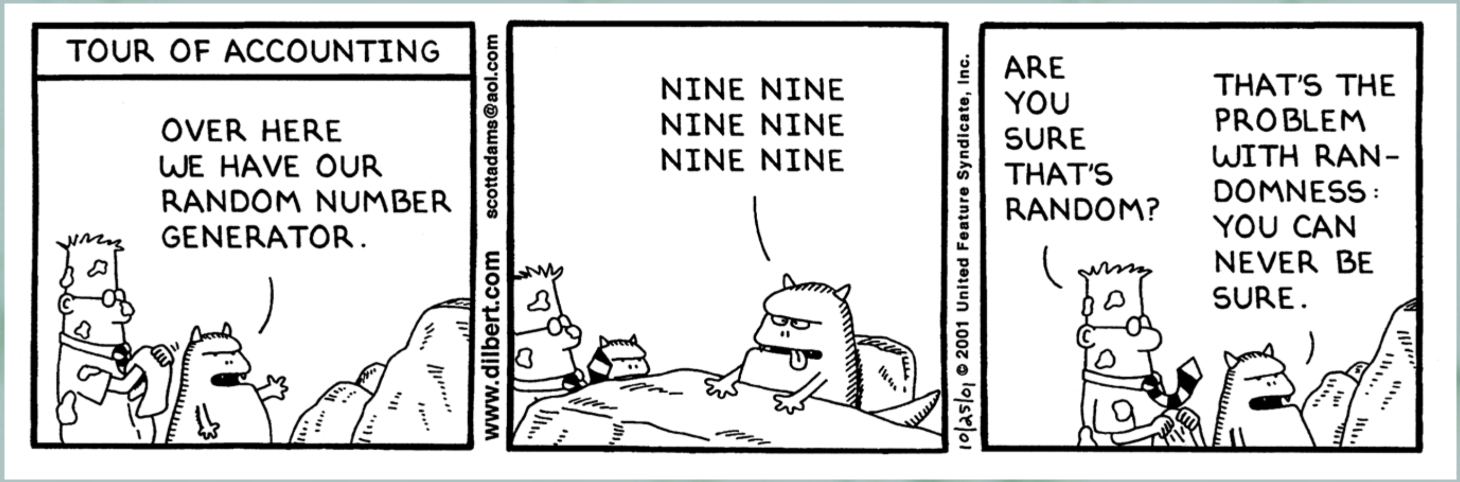
Apache II
8
15
19
11
13
27
7
9
38
4
32
2
31
35
37
23
17
29
12
6
20
16
28
1
34
24
30
26
0 10 20 30 40
SAPS II
19
22
38
3
40
30
37
4
15
27
28
12
29
11
9
17
6
16
24
31
32
1
26
34
13
20
23
35
0 10 20 30 40
MPM_{24} II
19
2
15
7
14
32
17
29
4
9
23
3
8
12
27
30
6
34
16
11
28
31
35
1
13
24
26
20
0 10 20 30 40

Control charts

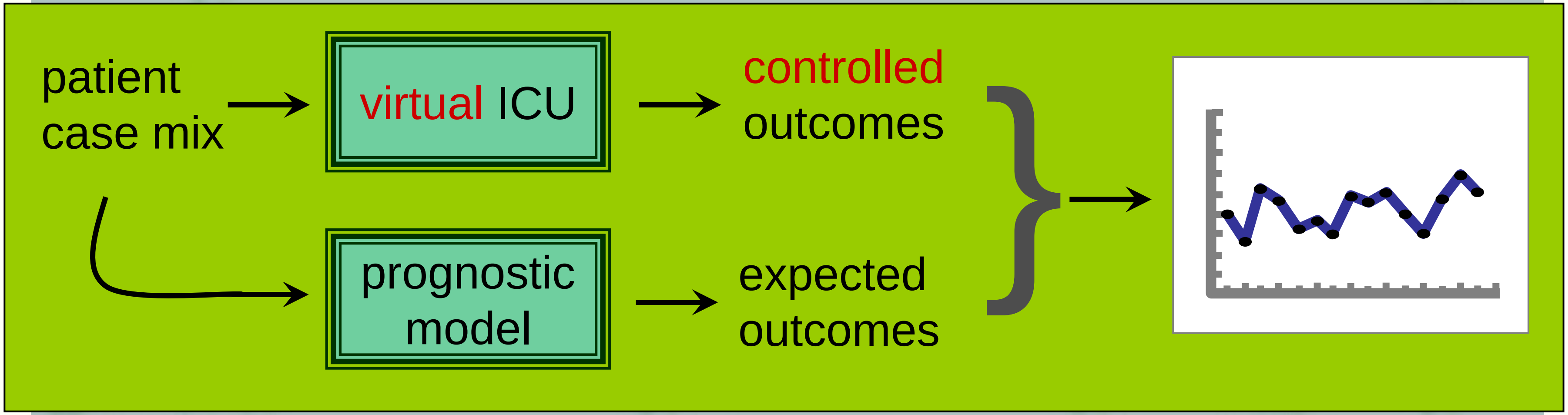


Control charts – common pitfalls

- projection of own beliefs to the data
- excessive distrust of extreme data points
- “trend happiness”



Studying the properties of control charts



Study design

- ICU admission records randomly drawn from the NICE Db and grouped by month
- Survival was determined by random number generation and Apache IV predicted probabilities
- Apache IV probabilities were artificially increased before determining survival (i.e. poor quality of care)
- 7 different types of control chart applied
- 32 different scenarios investigated, varying mortality increase factor, # adm/month, and baseline risk
- 5,000 repetitions per scenario

KLINISCHE InformatieKUNDE

Results

- Efficiency of control charts to detect increased mortality was moderate
  - 4 months to detect a doubling in mortality in an ICU with 50 adm/month
- Better performance with higher patient volumes
- Risk-adjusted EWMA has shortest time-to-signal, on average

KLINISCHE InformatieKUNDE

Summary and conclusions

- The quality of medical care can be assessed by summarizing patient outcomes
- Commonly-used methods are benchmarking and control charts
- A frequent pitfall is to neglect the uncertainty in ranks and other quality statistics
- The properties of quality assessment tools can be studied using statistical simulation methods based on resampling

KLINISCHE InformatieKUNDE

Thank you for your attention

Contact:

Niels Peek
n.b.peek@amc.uva.nl

Acknowledgements: Nicolette de Keizer, Ameen Abu-Hanna, Evert de Jonge, Daniëlle Arts, Ferishta Bakshi-Raiez, Twan Koetsier, Ilona Verburg, Peter van der Voort, Robert-Jan Bosman, David Cook.