EBMT European Society for Blood and Marrow Transplantation

Center comparisons with survival data

Hein Putter

Nothing to disclose

Lisbon, 20/3/2018

European Society
for Blood and Marrow
Transplantation

Center comparisons with survival data

Hein Putter

No conflict of interest

Lisbon, 20/3/2018

Benchmarking centers

- We want to compare (the performance of) several centers with respect to some benchmark
- Often (but not always) the benchmark concerns some binary (yes/no) indicator
- The benchmark could be set at the overall rate at which the indicator occurs
- For instance, consider the indicator “bad outcome” among allogeneic transplantations
- Suppose that, among all centers, this occurs about in 25% of all allogeneic transplantations

Hypothesis test

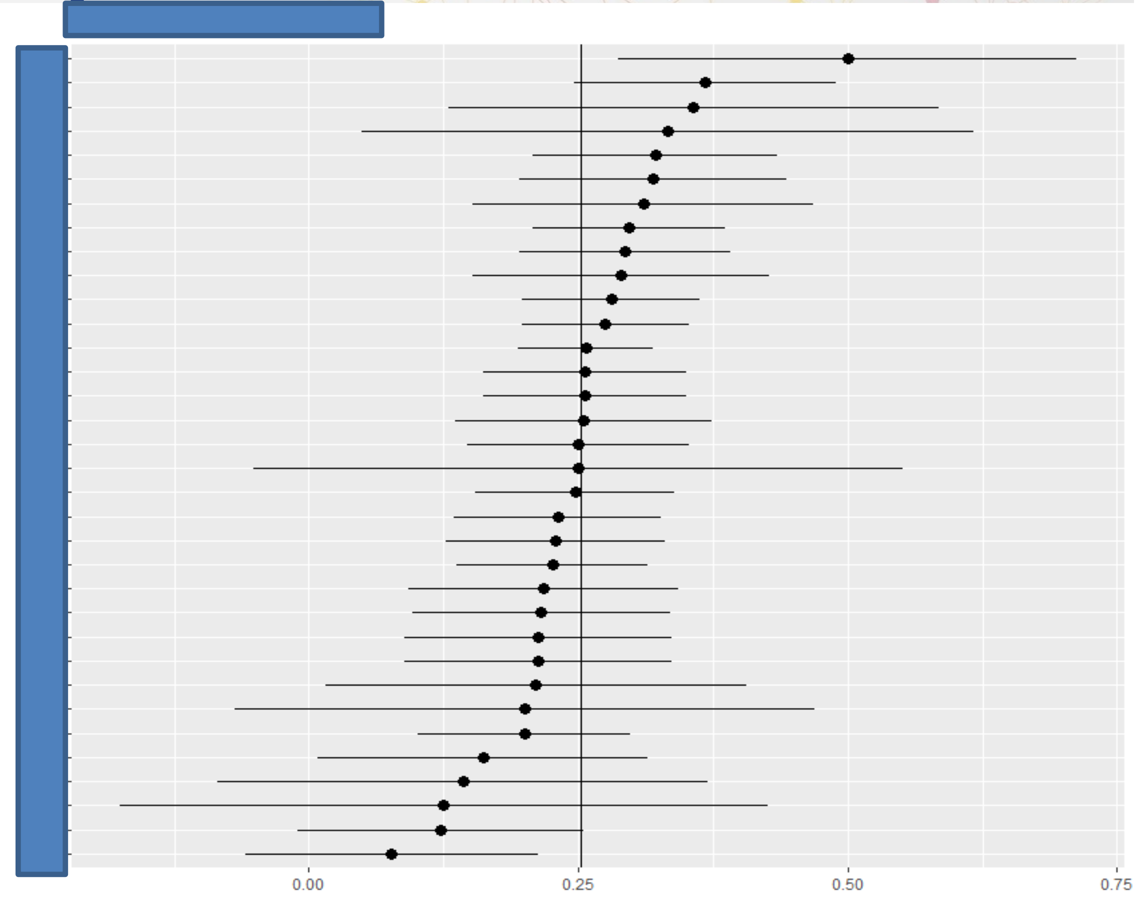
- For now, disregard differences in case mix
  - To be discussed later
- Then, for every center we can test (with $p_0 = 0.25$)

$$H_0: p = p_0 \text{ versus } H_1: p \neq p_0$$

- This is just a two-sided binomial test, and we can display results graphically in a caterpillar plot
  - Also known as league table

Caterpillar plot

- In the caterpillar plot the outcome is the estimate probability in each cente
- Shown with 95% confidence intervals
- Typically ordered by effe size (or performance)



Spiegelhalter, Statist. Med. 2005

Criticizes caterpillar plots:

… *as leading to a spurious focus on rank ordering, when it is known that the rank of an institution is one of the most difficult quantities to estimate.*

He argues that a more suitable display is the funnel plot

Funnel plot

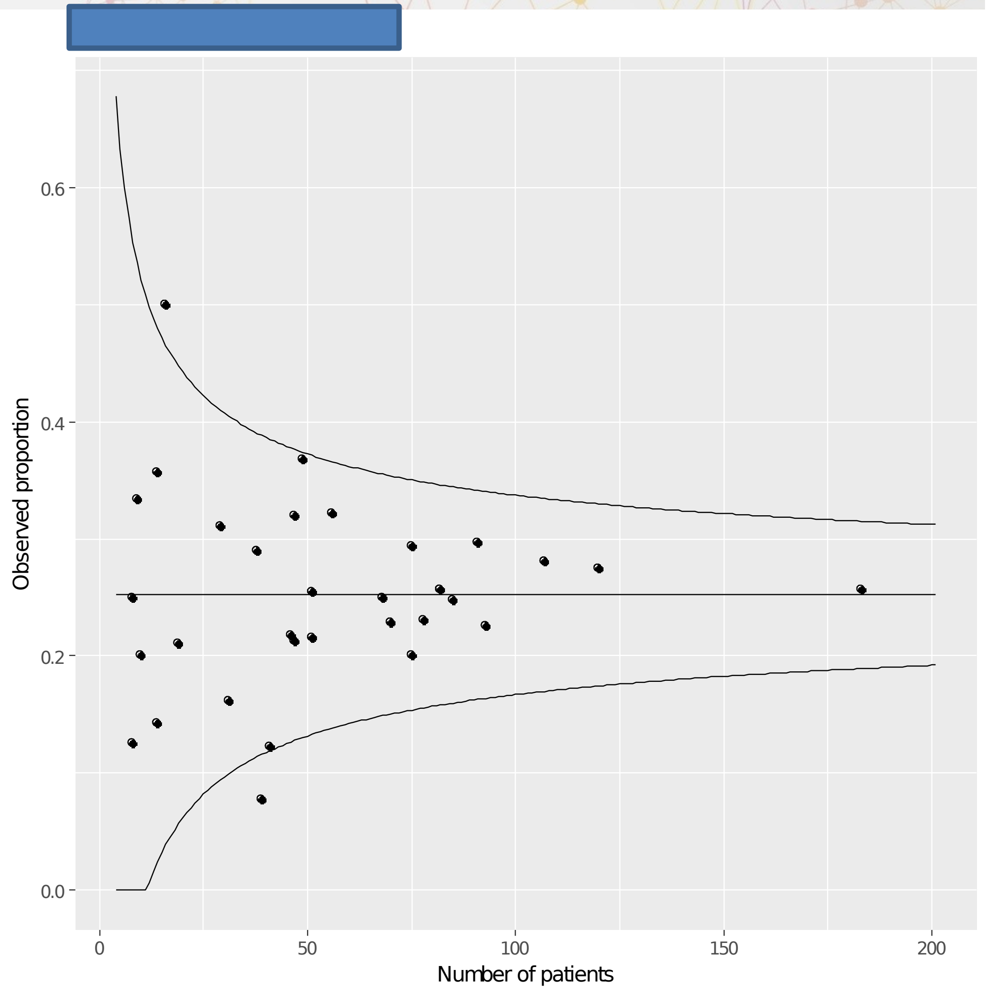
- In the funnel plot we plotted
  - x-axis: n
  - y-axis: $\hat{p} = x/n$
- Under H_0:

  $\hat{p} \sim N(p_0,\ p_0(1-p_0)/n)$
- Reject H_0 if

  $|\hat{p} - p_0| > 1.96\sqrt{p_0(1-p_0)/n}$
- Bounds do not depend on data and can be put into the plot before plotting the center results

- In the funnel plot we plotted
  - x-axis:
  - y-axis:
- Under :
- Reject if

  | > 1.96
- Bounds do not depend on data and can be put into the plot before plotting the center resu



More general framework

- x-axis: Expected (E)
  - $n\,p_0$
- y-axis: Observed/Expected (O/E)
  - $\dfrac{x}{n\,p_0} = \dfrac{\hat{p}}{p_0}$
  - Excess events over expected if center is performing according to benchmark
- Test: reject H_0 if

$$|\hat{p} - p_0| > 1.96\sqrt{p_0(1 - p_0)/n}$$

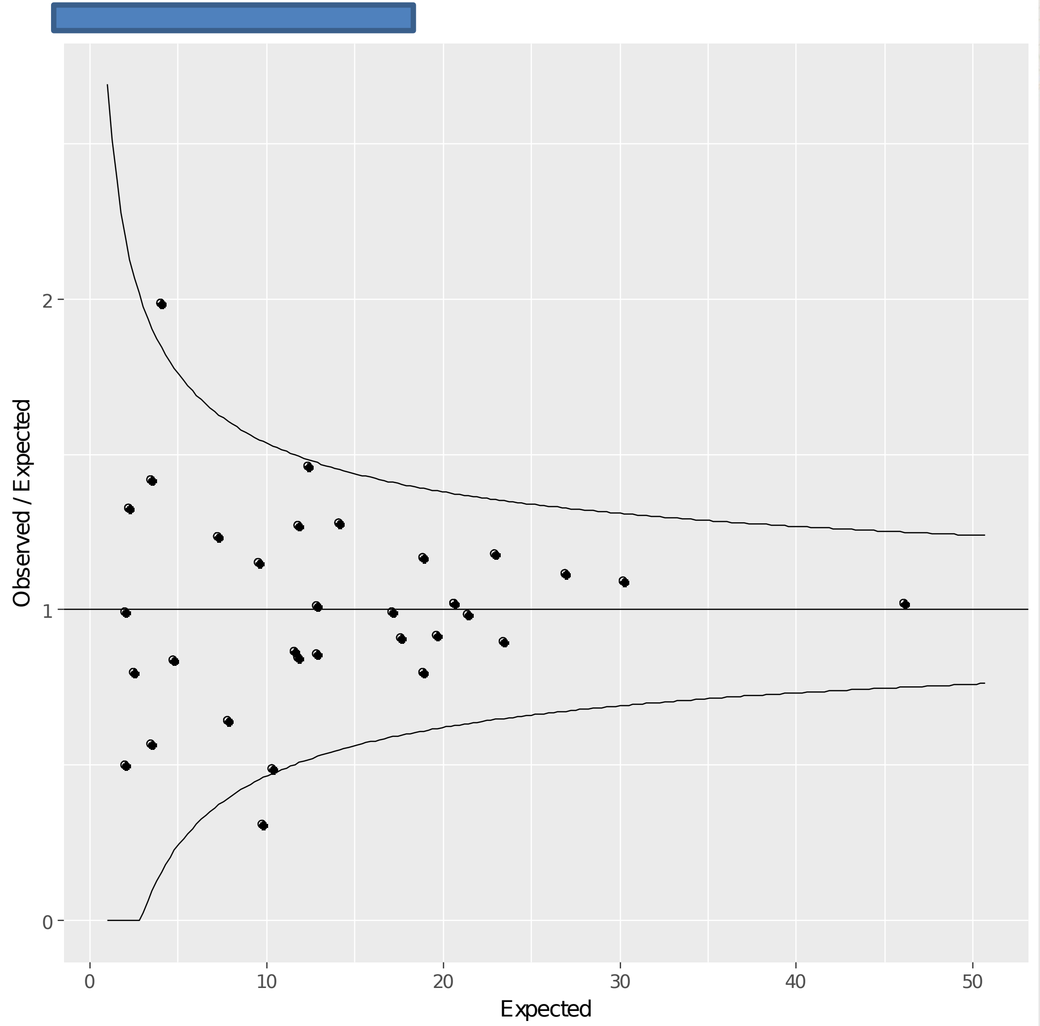
- Now becomes: reject H_0 if

$$\left|\frac{O}{E} - 1\right| > 1.96\sqrt{(1 - p_0)}/\sqrt{E}$$

- Advantage: E can be adapted to include case-mix

Funnel plot

- Same type of plot
- This time with “Expected” on the x-axis
- And “Observed/Expected” (excess) on the y-axis
- “Expected” gives a measu of the amount of informati in the data
  - The precision with which we have been able to estimate the excess



Survival outcomes

- Main question is: what are O and E in the survival setting, with censored data?
- Data: for center i, subject j, outcome (t_{ij}, d_{ij}, x_{ij})
- Underlying model, accounting for case mix x_{ij}

$$h_{ij}(t) = h_0(t)\exp(\beta^T x_{ij})$$

- Define, for center i,

$$O_i = \sum d_{ij}, E_i = \sum_j H_{ij}(t_{ij})$$

- Under H_0 we have $O_i \sim N(E_i, E_i)$

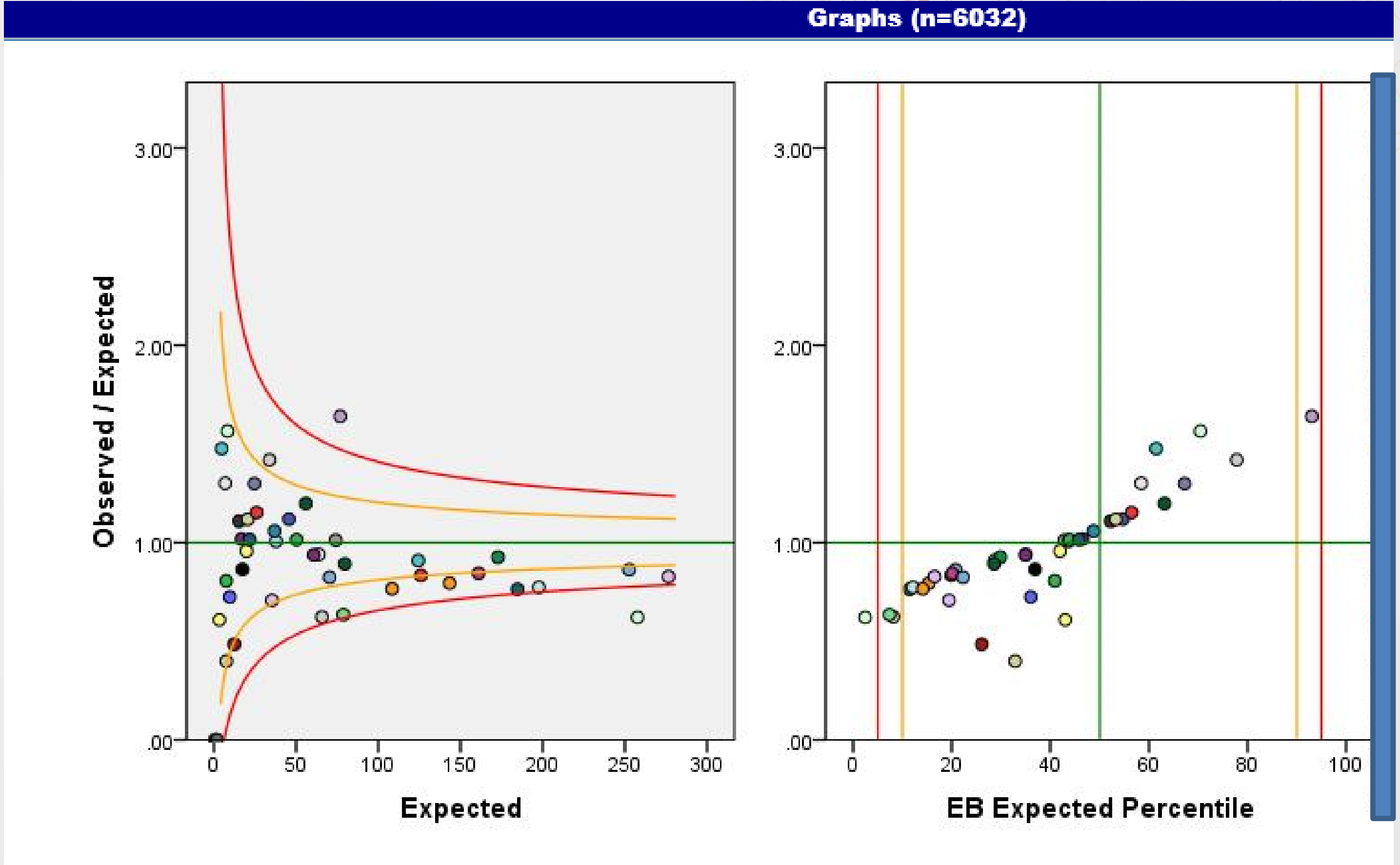
Remarks

- Justification of the approximation $O_i \sim N(E_i, E_i)$ is firmly rooted in counting process and martingale theory (Andersen, Borgan, Gill, Keiding, VII.2.2)
- Intuitive explanation of “Expected”: the number of events expected in a center, based on the number of patients, their follow-up and their patient characteristics
- Asymptotic test becomes: reject H_0 for center i if

$$\left|\frac{O}{E} - 1\right| > \frac{1.96}{\sqrt{E}}$$

- Bounds again do not depend on data, and can be put into the plot before plotting the center results
- In practice, β needs to be estimated, this is done under H_0

Result

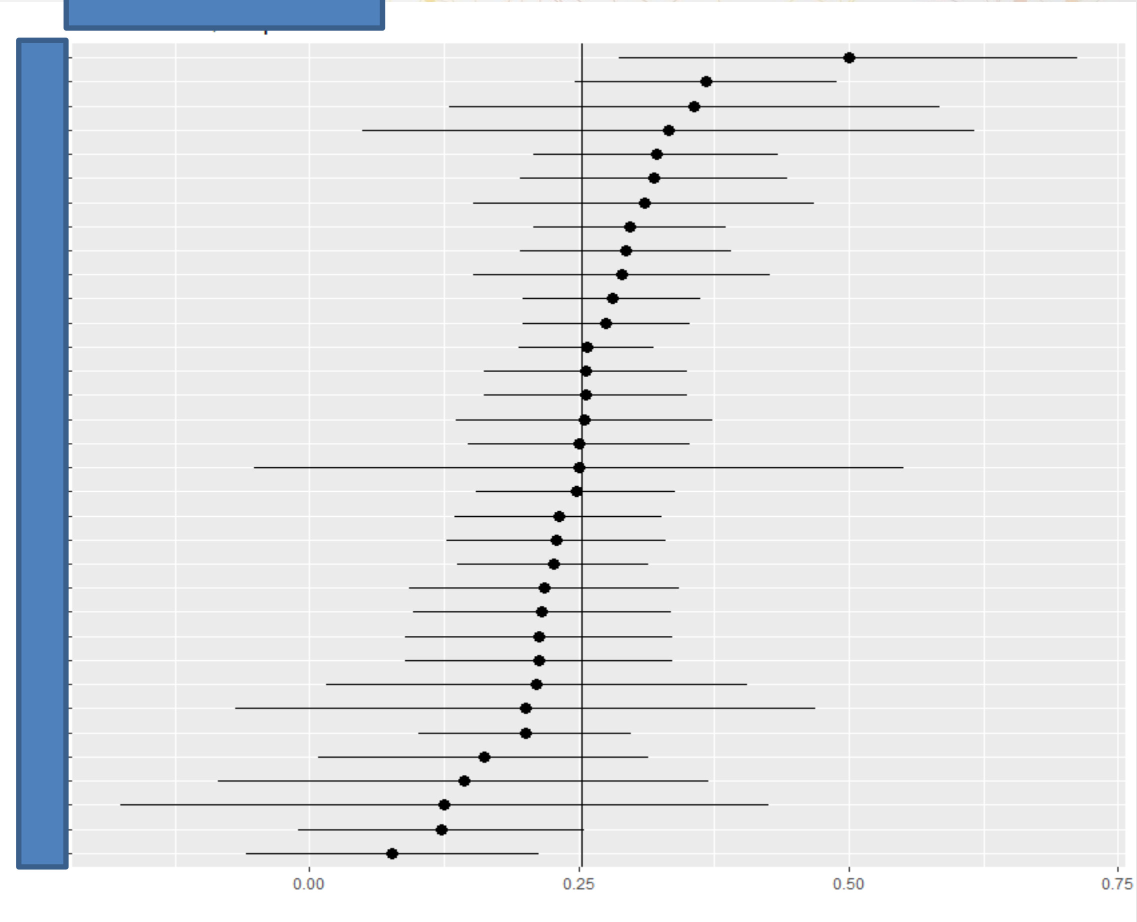


Graphs (n=6032)

Are there true differences?

- Suppose that each center has $p = 0.25$
- What would you expect?
- Could you see a picture like this?

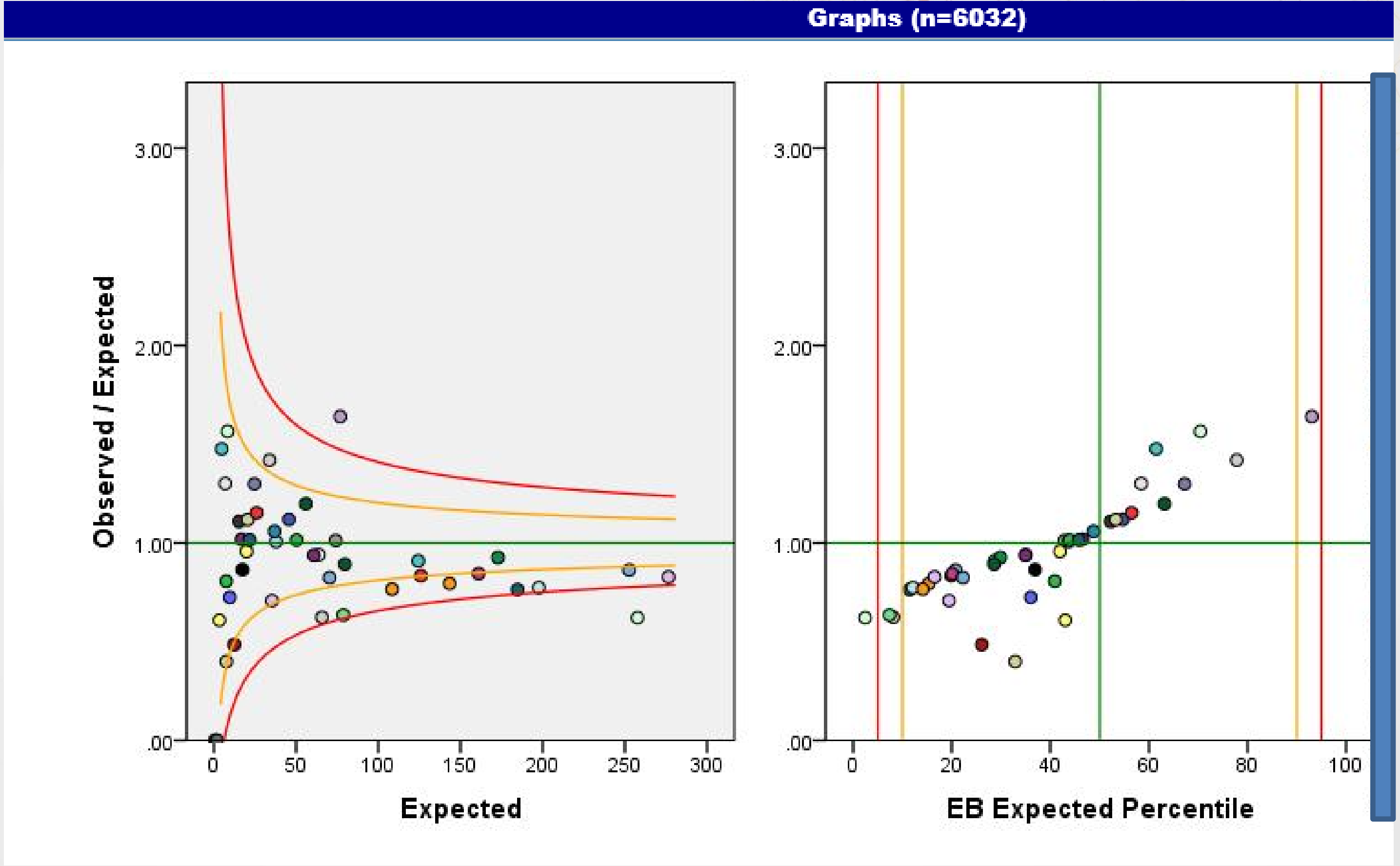
- Suppose that each cente has $p = 0.25$
- What would you expect?
- Could you see a picture like this?



Ranking

- Ranking of centers is extremely dangerous and not recommended
- Even in a situation where all centers are performing similarly, in the data there will always be a best and a worst center
- No reason at all to expect in that case that the ranking will be the same next year
- More sophisticated methods needed to disentangle random from systematic differences between centers
- Instead of ranking propose so-called Empirical Bayes (EB) percentiles
- The EB percentile gives the expected rank accounting for case mix *and* chance fluctuations
  - It gets rid of the differences that are not significant

Result

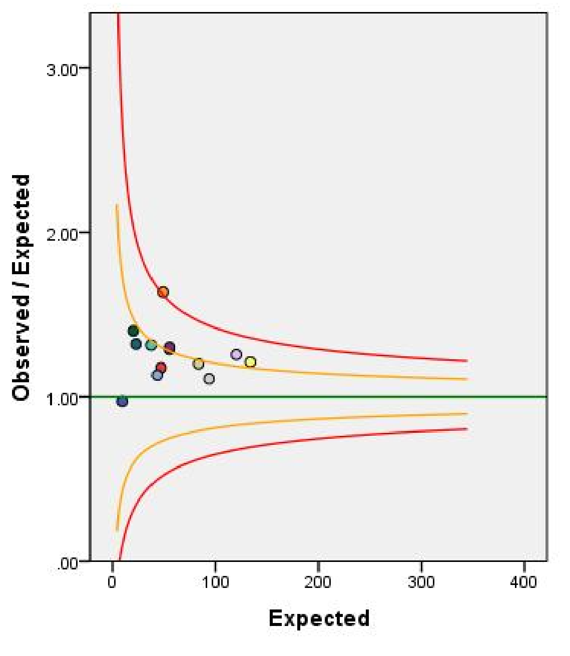

Graphs (n=6032)
Observed / Expected
3.00
2.00
1.00
.00
0
50
100
150
200
250
300
Expected
3.00
2.00
1.00
.00
0
20
40
60
80
100
EB Expected Percentile

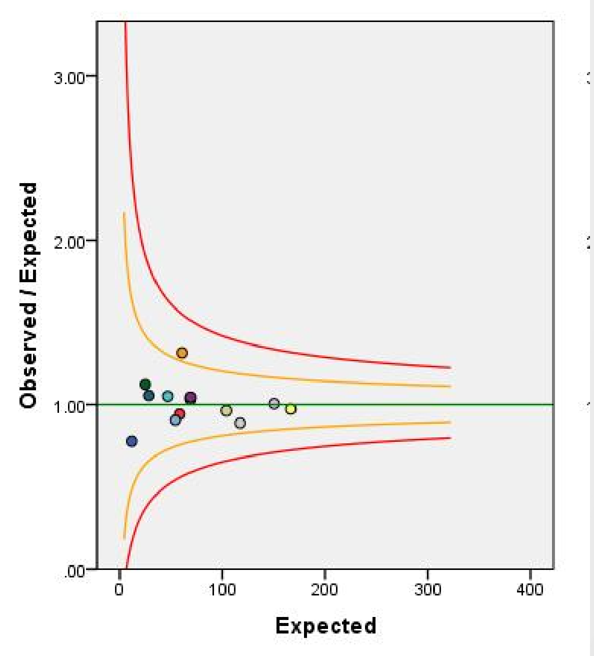
Target for benchmarking

- Target could be
  - A pre-set proportion or survival probability
  - Average in the same period of all centers, or
  - Average in the same period of all centers in the same country

ALLO, 12 months overall survival

- Center compared to other centers in Europe (left) or in its own country (right)





Case-mix correction

- In all of these outcomes we have to correct for case-mix
- This is because some specialized centers attract more serious patients
- Without correcting it would seem that these centers are performing badly
- Which variables to correct for, depends on
  - Clinical importance
  - Availability / completeness
  - Is the variable for case-mix a choice (consequence) of a center’s strategy for transplantation?
    - RIC and gender mismatch can be argued to at least partially be a decision by the treating physician and not a patient characteristic one is confronted with

Data quality

- Essential for a successful benchmarking project
- This includes
  - Completeness of the registration of those risk factors determined to be used in the case-mix models
  - For survival data: completeness of follow-up
- Possible trap: perhaps all deaths are reported in a center, but follow-up of patients alive is lagging => bias (not in favor of the center)
- One can also benchmark completeness of patient data and of follow-up
  - Using similar methodology (reverse Kaplan-Meier)
  - No case-mix correction, because completeness of data and follow-up is generally not expected to depend on patient characteristics

Some general thoughts

- The ultimate goal is improvement of patient care
- Tool for centers to get more insight into their own performance
  - How are they doing in comparison with others, after correcting for possible differences in case-mix
- Trust and transparency is essential
  - In the procedure
  - In the models used
- We must be modest in what we claim
  - Case-mix correction model will not be perfect
  - But even an imperfect case-mix correction model is a hell of a lot better than a crude comparison