A STATISTICAL INTERLUDE...

UNDERSTANDING UNCERTAINTY IN MENTAL HEALTH QUESTIONNAIRE DATA

Andy **Fugard**

INTRODUCTION

This chapter provides a brief overview of issues to consider when interpreting mental health questionnaire data from service users. I have focused on what I think are topical areas of uncertainty. Suggestions for further reading are provided at the end.

THE KEY IDEA: REASONING FROM SAMPLE TO POPULATION

One of the key problems statistical methods solve is how to reason from your sample, for instance average changes in scores over time for a selection of service users, to the population, for example people who are likely to attend your service in future. Figure 1.4 illustrates this pictorially.

Intuitively, the larger your sample, the better you can estimate the population effect. How you sample is also important; for instance people who drop out of treatment have been shown to have worse outcomes than those who complete. This leads to a common sample bias: people who drop out are also less likely to complete follow-up questionnaires, which artificially inflates the estimated average outcome.

UNDERSTANDING SCORES FROM INDIVIDUALS

Consider a measure of symptom severity such as the difficulties subscores of the Strengths and Difficulties Questionnaire (SDQ) or from the Revised Children's Anxiety and Depression Scale (RCADS). A particular score will be the result of severity, plus noise due to factors such as:

- The complicated processes required to translate feelings into ticks on a page

- Moment-by-moment variation in feelings due to particular events that have occurred close to the time of completing the questionnaire.

Statistical methods, usually relying on analyses of large norm samples, can be used to help safely interpret scores. In all cases, clinical judgement should be central and scores triangulated with others sources of information.

The following sections illustrate some useful ideas.

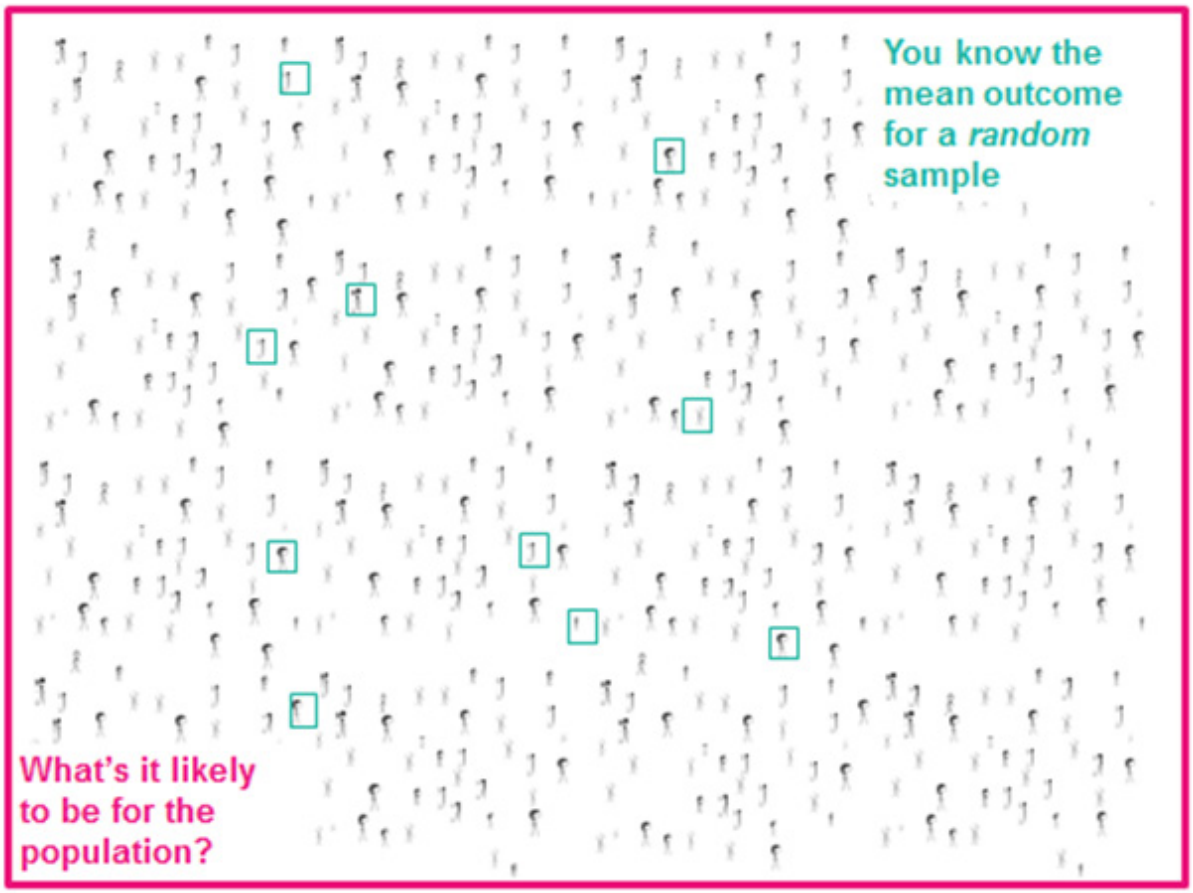

FIGURE 1.4: An illustration of the sample-to-population inference problem. Image created using stick people from XKCD (xkcd.com), which is licensed under a Creative Commons Attribution-NonCommercial 2.5 License.

CLINICAL CUT-OFFS

Some measures provide clinical cut-offs or clinical bands which can be used to help interpret scores. A score in the clinical band indicates levels of symptoms which would benefit from therapeutic input. These cut-offs are found via some kind of "gold-standard" of clinical "caseness", for instance agreement between a number of clinicians who have assessed a young person's levels of difficulties – perhaps using a structured assessment tool. In large samples where both structured assessments have taken place and measures have been completed, it is then often possible to find a score which distinguishes between people who are and those who are not experiencing clinical levels of difficulties.

For a given score, two quantities can be calculated and are often reported in test manuals:

- Sensitivity. Take all those people who were assessed to have clinical levels of difficulties by the gold standard. What proportion of these score in the clinical band on the questionnaire?
- Specificity. Take all of those people who were assessed not to have clinical levels of difficulties. What proportion of these score in the non-clinical band on the questionnaire? The quantity 1 – *Specificity* is also often used – known as the *false alarm rate.* This is how many people who are non-clinical are incorrectly classified as clinical according to the questionnaire.

Each cut-off score chosen leads to different sensitivities and specificities. Figure 1.5 below shows an example. Note how as the sensitivity improves from around 0.2 at one extreme to 1.0 at the other (i.e. we get better at detecting people with clinical problems), the specificity becomes correspondingly worse (we over-diagnose). The best cut-offs are compromises between missing possible problems and wrongly suggesting someone has a problem when they do not.

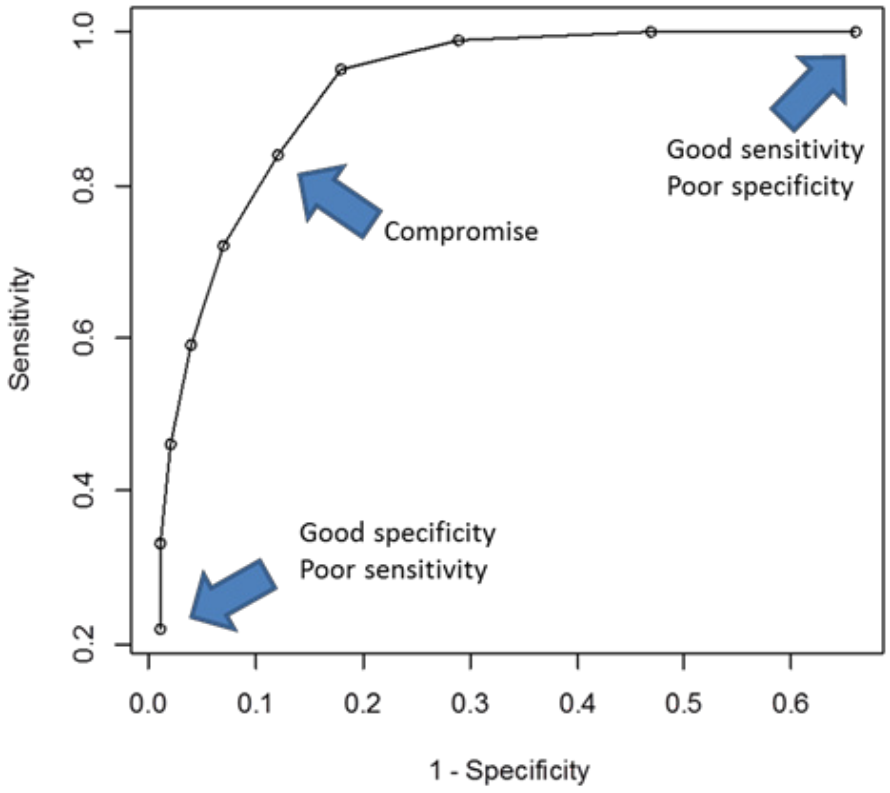

FIGURE 1.5: Sensitivity and 1 – Specificity for all possible cut-offs on a measure.

PERCENTILES

Suppose a large sample of people have completed a questionnaire, and their scores are sorted into ascending order. Percentiles provide a way to pick out individual scores along this list and are helpful landmarks to interpret symptom severity. The 0th percentile is the minimum score, the 100th percentile the maximum, and the 50th percentile is in the middle (also known as the median). So if a young person scores on the 82nd percentile of a test, then it means that their score is greater than or equal to 82% of similar young people used to calculate the norms.

Many measures provide percentiles. For example for the SDQ, these are available at http://www.sdqinfo.com/UKNorm.html from a large community sample.

STATISTICALLY RELIABLE CHANGE

Various tests such as t-tests and ANOVA can be applied to test changes in averages of groups of clients, but how about individual change?

One useful tool is the *reliable change index* and its close relation the ***reliable change criterion.*** The general idea is that scores consist of two main components: the true score reflecting whatever dimension we want to measure, and a noise component, due to reasons explained above. A change in scores over time – for instance between first and last score – is said to be reliable change at some level of confidence if it is greater than would be expected by the noise component.

The reliable change index is simply the difference between two scores (let's call them x and y) divided by the standard error of the difference (SE_{diff}): $RCI = \frac{x-y}{SE_{diff}}$

The standard error of the difference requires the following information from a large sample, for example published norm data:

- The s*tandard deviation* of scores at the beginning of treatment.
- A measure of the questionnaire's *reliability*: either Cronbach α ("alpha"), which summarises the extent to which items measure the same thing, or test-retest reliability, typically a correlation in scores measured a short time apart so that little or no "real" change would have been expected.

The formula is as follows: $SE_{diff} = SD \times \sqrt{2} \times \sqrt{1 - r}$

where SD is the standard deviation and r is the reliability. The resulting RCI is a z-score, i.e. it is normally distributed (bell curve) with a mean of 0 and an SD of 1.

This information is available for all CYP IAPT measures on the ROM web page: http://www.cypiapt.org/routine-outcome-monitoring/routine-monitoring-outcome.php

A recap on z-scores

What do these RCI values mean? One rule of thumb is that if it is over 1.96 or less than -1.96, then you can be confident (95% level) that change was statistically reliable. Figure 1.6 shows the z-score distribution pictorially. For RCI, this illustrates the likely pattern of change due to chance *alone assuming there was no change in the true score.* So, for instance, around 68% of people will have scores fluctuating around -1 to +1 even if the true score has not changed.

Example

Let's use one of the CYP IAPT measures, the depressive symptoms subscore of the RCADS. The norms for the RCADS depend on age and gender; suppose the service user is 13 years old and female. The SD is 6.26 and Cronbach ***α*** is 0.87. The standard error of the difference is then $6.26 \times \sqrt{2} \times \sqrt{(1 - 0.87)} = 3.2$. Table 1.2 shows the computed RCIs for a range of changescores.

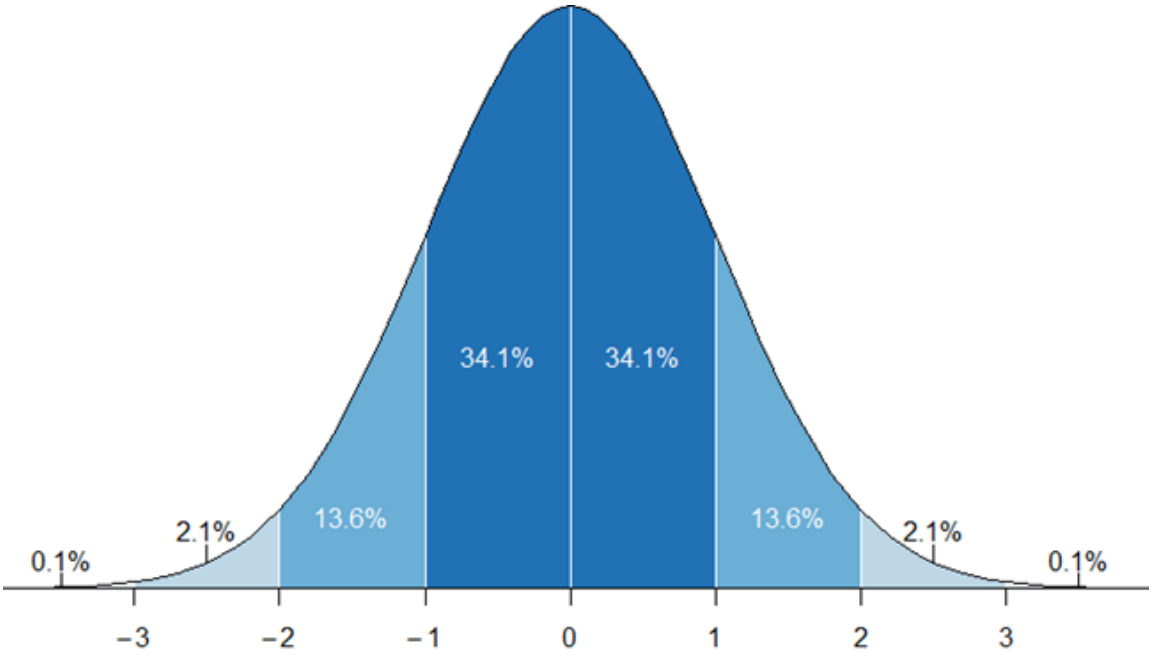

FIGURE 1.6: Distribution of z-scores. For the RCI, approximately 68% of scores will be between -1 and +1, and 95% of scores between -2 and +2 if there has been no change in the "true" score.

TABLE 1.2: Reliable change indices for a range of changescores for the RCADS depressive symptoms subscore (using norms for a 13-year-old female).

CHANGESCORE	RCI
0	0.00
1	0.31
2	0.63
3	0.94
4	1.25
5	1.57
6	1.88
7	2.19
8	2.51
9	2.82

A (95%) *reliable change criterion* is just 1.96 x SE_{diff}. For this measure this will be approximately 6, i.e. a score of 6 or more would be considered reliable change, and a score of less than 6 could well be due to measurement error. Other levels may be computed based on the z-distribution. Some examples (two-tailed) are shown in Table 1.3, for example the 90% criterion for this RCADS subscale is approximately 1.64 x 3.2 = 5.2.

TABLE 1.3: Multipliers for reliable change criteria for levels of confidence from 50% to 95%.

CONFIDENCE	MULTIPLIER
0.50	0.67
0.55	0.76
0.60	0.84
0.65	0.93
0.70	1.04
0.75	1.15
0.80	1.28
0.85	1.44
0.90	1.64
0.95	1.96

EXPECTED RECOVERY CURVES

Session-by-session measure developers are also providing curves of change over time. As more data are collected in the UK, this will be possible for CYP IAPT measures too. The idea is then that it's possible to predict the distribution of change over time and see how well a particular service user is progressing compared to others with similar problem severity. See Figure 1.7 for an illustration of how this looks.

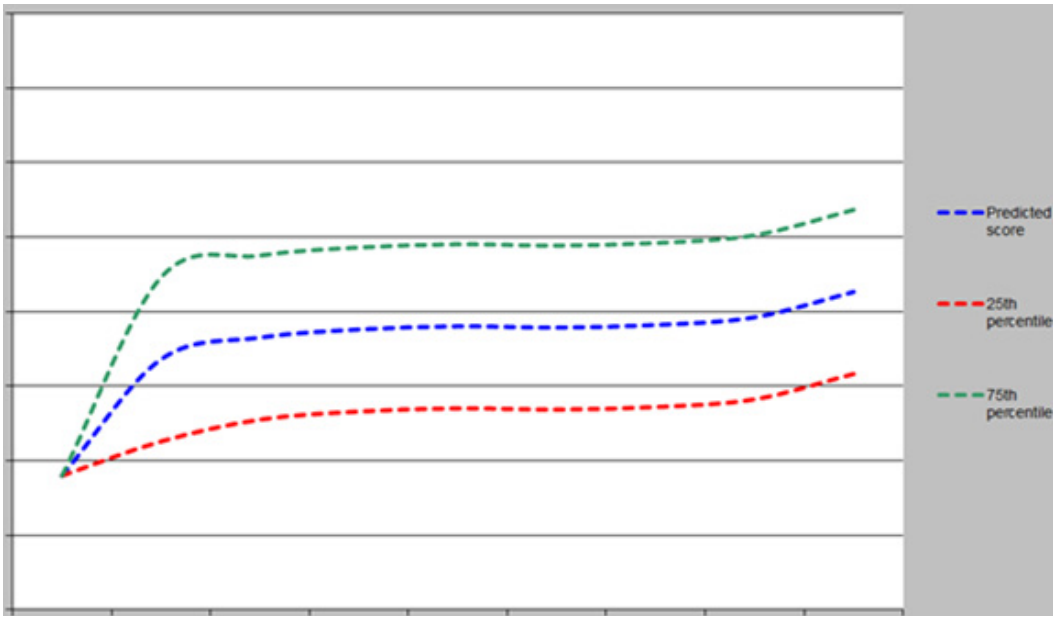

FIGURE 1.7: Expect recovery curve example for a session-by-session measure showing the 25th percentile, 75th percentile, and prediction for the average.

INSTITUTIONAL COMPARISONS

Up until now we have focused on interpreting individual scores. Increasingly outcome measures are used for team and service provider comparisons too. For some years now, Adult IAPT outcomes have been published by (what is now) the Health & Social Care Information Centre (HSCIC; see http://www.hscic.gov.uk/). Anyone can see "recovery rates" for their local adult IAPT service.

When estimating a population quantity, such as an average outcome, based on a sample, two main factors affect the precision:

1. The number of cases included in an analysis, i.e., the sample size. The larger the sample size then the more precise the estimate.
2. The spread of values, i.e., the variance or standard deviation. Greater spread means less precision.

This impact is summarised by the equation for the standard error of the mean for normally distributed (bell curve) data: $SEM = \frac{SD}{\sqrt{N}}$ where SD is the standard deviation and N is the sample size. (This, as an aside, is an example of one of the wondrous things about statistics: it's possible to come up with simple formulas which quantify best guesses about how sure one can be about the precision of inferences.)

Note that the standard deviation will tend not to change so much as the sample size increases. This gives the overall spread of values in the data. The standard error of the mean depends on the standard deviation, but it represents a level of uncertainty about what the mean is likely to be in the population based on what it is in the sample. (And recall, you always know for certain what a sample-quantity is – you just compute it.)

Figure 1.8 illustrates the implications of this for outcomes. Each point represents a service. The position on the horizontal axis shows how much data the service provides and the vertical axis is the average improvement in symptoms.

Importantly, this data is all simulated and designed so that at the population-level, there is no difference between the services' outcomes. It's clear that mean outcomes are more varied for smaller datasets than for larger ones.

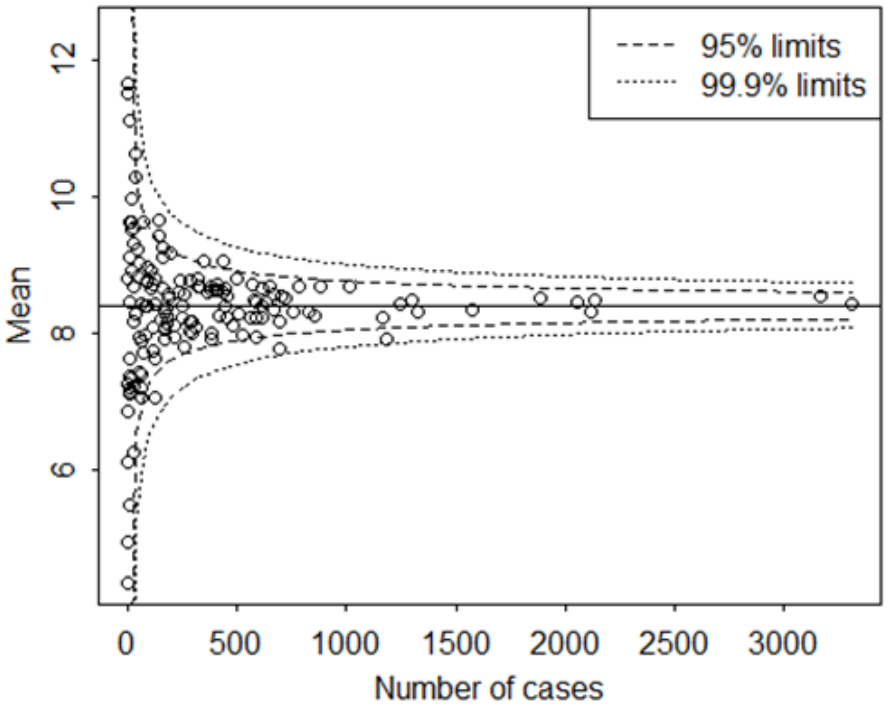

FIGURE 1.8: A "funnel plot" illustrating how sample size affects average outcomes.

This is as predicted by the standard error of the mean equation above. This equation was also used to calculate 95% and 99.9% confidence intervals for the mean. As may be seen, all points are within the 99.9% limits and most within the 95% limits. It's notable that not all are in the 95% limits – especially as this is the criterion often used to test whether there are "statistically significant" differences in data.

HELP – NONE OF THIS MADE ANY SENSE !

What if you didn't study statistics, or aren't a fan of maths? Fear not – colleagues will be able to assist! You could try asking:

- **Recent psychology graduates,** perhaps working as assistant psychologists. Graduates of UK programmes accredited by the British Psychological Society study a range of approaches. All are essential for analysing questionnaire data. Arguably analysing data is a better use of graduates' time than is data entry – a task they often seem to perform at services.
- **IT professionals** have extensive experience developing software and a strong background in mathematics. They are less likely to have an applied statistics background, however often will be able to pick it up quickly with the help of intensive one or two day courses. The statistical package R (www.r-project.org), free and driven by a powerful programming language, might be a way to entice them.
- **Other colleagues** who use terms like "regression to the mean" or "precision" in conversation. Keep an eye out. You never know who you might find. Some services have undercover physicists now working as clinical psychologists, ecologists with a knack for stats working as senior managers.

ACKNOWLEDGEMENTS

Thanks to David Trickey for helpful comments.

FURTHER READING

General introductions/refreshers to statistics

Crawley M. J., (2012) *The R book*. Wiley, Chichester.

This is the textbook to read about R and a wide range of statistical methods. Probably one for the IT professionals.

Field A., (2013) *Discovering statistics using IBM SPSS Statistics (4th edition)*. SAGE Publications, Los Angeles, London, New Dehli, Singapore, Washington D. C.

A leisurely introduction to statistics through SPSS, much loved by undergraduate psychology students.

Field A., Miles J., Field Z., (2012) *Discovering statistics using R*. SAGE Publications, Los Angeles, London, New Dehli, Singapore, Washington D. C.

As above, using R rather than SPSS for its examples.

Howell D. C., (2010) *Fundamental statistics for the behavioral sciences (8th edition)*. Wadsworth, Belmont.

Suitable for more advanced psychology graduates.

RELIABLE CHANGE

Evans C., Margison F., Barkham M., (1998) The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence-Based Mental Health,* 1, 70–72.

See also Chris Evans' very helpful webpage at http://www.psyctc.org/stats/rcsc.htm

RUNS CHARTS AND CONTROL CHARTS

Caulcutt R., (2004) Managing by fact. *Significance,* 1, 36–38.

Perla R. J., **Provost** L. P., **Murray** S. K., (2011) The run chart: a simple analytical tool for learning from variation in healthcare processes. *BMJ Quality & Safety,* 20(1), 46–51.

Provost L. P. **Murray** S., (2011) *The Health Care Data Guide: Learning from data for improvement.* Jossey-Bass, San Francisco.

REGRESSION TO THE MEAN

Ford T., **Hutchings** J., **Bywater** T., **Goodman** A., **Goodman** R., (2009) Strengths and Difficulties Questionnaire Added Value Scores: Evaluating effectiveness in child mental health interventions. *British Journal of Psychiatry,* 194, 552–558.

COMPARING INSTITUTIONAL PERFORMANCE

Fugard A. J. B., Stapley E. J., Ford T., Law D., Wolpert M., York A., (2014) Ranking mental health service outcomes is harmful: a substantive and statistical alternative. *Draft manuscript.*