Child and Adolescent Mental Health
a journal for all professionals working with children and young people

Child and Adolescent Mental Health 20, No. 3, 2015, pp. 155–162 doi:10.1111/camh.12086

Analysing and reporting UK CAMHS outcomes: an application of funnel plots

Andrew J. B. Fugard[1], Emily Stapley[2], Tamsin Ford[3], Duncan Law[4], Miranda Wolpert[2] & Ann York[5]

[1]Research Department of Clinical, Educational and Health Psychology, University College London, 26 Bedford Way, WC1H 0AP, UK. E-mail: a.fugard@ucl.ac.uk
[2]University College London, Anna Freud Centre, UK
[3]University of Exeter Medical School, UK
[4]Specialist CAMHS, Hertfordshire Partnership University NHS Foundation Trust, UK
[5]South West London & St George's Mental Health NHS Trust, Child and Family Consultation Centre, Richmond, UK

Background: Patient-reported outcomes measures are increasingly being used in child and adolescent mental health services (CAMHS). League tables are a common way of comparing organizations across health and education but have limitations that are not well known in CAMHS. **Method:** Parent-rated Strengths and Difficulties Questionnaire (SDQ) outcomes data from 15,771 episodes of care across 51 UK CAMHS were analysed using funnel plots, an alternative to league tables. **Results:** While most services were indistinguishable from the national average there was evidence of heterogeneous outcomes and seven services had outcomes below 99.9% limits for SDQ added-value scores. **Conclusions:** Funnel plots are powerful tools for navigating national data and can help prompt investigations using clinical theory and local service context. Examples are provided of factors to consider in these investigations. We argue that analyses of the local context are central to the valid application of funnel plots.

Key Practitioner Message

- We recommend that funnel plots are used for national analyses of CAMHS outcomes data rather than league tables.
- A funnel plot analysis of 51 UK CAMHS showed positive outcomes overall, however, there was evidence of heterogeneity across services and seven services were flagged as outliers with scores below the national average.
- Management decisions cannot rely on data analysis alone.
- Reporting should include interpretation using the local context, involving clinicians and ideally service users to help understand results.

Keywords: Patient-reported outcome measures; funnel plots; mental health; policy; national analysis; interpretation

Introduction

The use of patient-reported outcomes measures (PROMs) provides a voice to service users about the impact of the interventions provided. In child and adolescent mental health services (CAMHS), service users include children, young people, and their families. PROMs ask service users about symptoms and their severity, as well as about strengths, thus making it possible to monitor progress over the course of interventions and outcomes. Feeding back information from the service users' perspectives to clinicians has been shown to improve outcomes, especially for people who are progressing more slowly than expected or who are deteriorating (Bickman, Kelley, Breda, de Andrade, & Riemer, 2011; Lambert & Shimokawa, 2011).

PROMs have been analysed nationally in UK CAMHS since 2002 as part of the CAMHS Outcomes Research Consortium (CORC) (Wolpert, Ford et al., 2012), a not-for-profit learning collaboration (see the CORC website: www.corc.uk.net). Recently the NHS Improving Access to Psychological Therapies (IAPT; www.iapt.nhs.uk/cyp-iapt) service transformation programme has been extended to CAMHS, including the use of session-by-session PROMs (Wolpert, Fugard, Deighton, & Görzig, 2012). Such methods for collecting national data using standardized measures are important for ensuring consistency of care across the UK. However, these systems raise issues about whether it is feasible to accurately compare services in relation to outcomes.

How is performance information presented in other related fields? League tables in education are now widely consulted and are used nationally to represent school performance outcomes in the UK (Department for Education, 2013), as well as to internationally evaluate education systems (Programme for International Student

Assessment; (PISA), 2009). League tables have also been used across healthcare (Healthcare Commission, 2005). The Adult IAPT programme now publishes league tables of all services' 'recovery rates'; defined as the number of patients moving from scoring in the clinical bands of the measures to scoring in the non-clinical bands (Health & Social Care Information Centre, 2014). This means that anyone, including prospective service users and commissioners, can access the data to see how well different services are performing and potentially rank services by the proportion of patients who have recovered. A bar graph representation of the tables has also been published (Gyani, Shafran, Layard, & Clark, 2013, p. 600) and used to argue that there is great between-service variability in recovery rates.

In this paper, we will argue that league tables and related forms of display are an inappropriate method for the comparison of services even when presented with uncertainty intervals, as the vast majority of people who will read them do not have a statistical background and will tend to focus on the rank order or mean differences even if they are not statistically significantly different. A rank does not necessarily tell one much about the quality of the service, as even if all services were effective or all were failing hopelessly, one service would always have the highest and one the lowest score. We suggest an alternative method of presentation for the same information using funnel plots (Spiegelhalter, 2005), which plot the indicator of interest against the precision of the measurement, and include control limits that indicate whether a service is statistically significantly different to the national average.

The funnel plot approach has been used as a method for analysing outcomes in physical healthcare (van Dishoeck, Looman, van der Wilden-van Lier, Mackenbach, & Steyerberg, 2011). For instance, the NHS has used funnel plots in the UK to show mortality rates for individual hospitals and surgeons, in order to account for the fact that different hospitals and surgeons operate on differing numbers of patients (National Joint Registry, 2014). None of this is new to statisticians or to clinicians and managers in some medical specialties; however, the reasoning is less familiar to practitioners and commissioners in CAMHS and mental health services for other groups of people than perhaps it could be. We apply this method to national CAMHS data and provide examples of how the results may be interpreted using local contextual factors such as data quality, measures used, case mix, and therapeutic factors.

Methods

We analysed the outcomes from the current CAMHS Outcomes Research Consortium (CORC) dataset using funnel plots.

Participants

The data in this paper comes from 15,771 episodes of care submitted by 51 CAMHS from across England and Scotland. Of these 51 CAMHS, 45 are NHS CAMHS and the remainder are voluntary sector services. Demographic information about the sample can be seen in the online supplementary Table S1. See Wolpert, Ford et al. (2012) for an analysis of all-CORC average change and correlations between measures for an earlier version of this dataset.

The subset of data analysed in the current paper was taken from a larger dataset of 181,009 episodes of care collected by the CAMHS involved in CORC, of which 95,448 cases had a baseline total difficulties score on the parent/carer version of the Strengths and Difficulties Questionnaire (SDQ-Parent), but lacked follow-up data.

Measures

Parents or carers completed the SDQ-Parent (Goodman, 1999, 2001), which is a standardized and well-validated 25-item measure that can be used to assess young people's levels of difficulties in hyperactivity, emotional symptoms, conduct problems, and peer relationship problems (a higher score means more difficulties) and strengths in prosocial behaviour (a higher score means greater strengths). The SDQ also includes eight items assessing the impact of any perceived difficulties on the young person's life. A higher score on the impact supplement indicates a higher level of distress. See Table S2 for means and standard deviations for the sample for each SDQ scale.

Funnel plots may be produced for a range of different types of data, for instance normally distributed outcomes, proportions, or event frequencies. The mean differences between scores on the SDQ at baseline and at follow-up are often used as an indication of how much a CAMHS is helping the young people they see. However, as higher scores on symptoms measures tend to decrease over time irrespective of whether or not the young person has received an intervention, due to such factors as regression to the mean (a statistical artefact), attenuation (the tendency for respondents to report less on the second time of measurement; Jensen et al., 1995) and spontaneous improvement, studying the differences between scores would not enable us to determine how much of the young person's improvement was actually due to the intervention they received.

An alternative way of assessing how effective the help provided by a service that is increasingly used in routine CAMHS analyses is to calculate the SDQ added-value score (AVS; Ford, Hutchings, Bywater, Goodman, & Goodman, 2009; Youthinmind, 2009). The AVS is calculated from a formula that attempts to estimate the effectiveness of interventions and is derived from SDQ-Parent scores. The AVS compares the observed follow-up total difficulties score with the score predicted from baseline scores assuming that the child has not received any intervention. This predicted score is calculated on the basis of scores from a community sample of children who had clinically significant levels of difficulties, but most of whom had not received any intervention 6 months after their first assessment. The process is similar to standard growth charts commonly used to monitor height and weight. An advantage of the AVS is that a much lower proportion of the variance of the AVS (0.6%) is explained by characteristics of the child and their family such as diagnosis, family income, age and gender, compared to the initial (35.9%) and follow-up (24.2%) parent SDQ scores, so it also performs a type of case mix adjustment.

Standardized effect sizes indicate the extent to which a population has shifted after an intervention and are often used to compare outcomes in clinical trials and services. Raw AVS can be translated into standardized effect sizes by dividing by the standard deviation (*SD*) of the difference, which makes it easier to compare across different measures. Positive scores or effect sizes on the AVS suggest that the improvement in symptoms seen at the CAMHS is better than would have been expected had the intervention not been received, while negative scores or effect sizes suggest poorer outcomes than predicted, though not necessarily deterioration.

Recovery rates are also used as a measure of how much CAMHS are helping the young people that they see and are currently being used as a public 'Key Performance Indicator' (KPI) in Adult IAPT service league tables (Health & Social Care Information Centre, 2014). Recovery rates refer to the number of patients who have moved from scoring in the clinical band at baseline to scoring in the non-clinical band at follow-up. Patients who have moved from scoring in the non-clinical band at baseline to scoring in the clinical band at follow-up can be said to have deteriorated. SDQ scores of 17 or more are classified as lying within the clinical band.

Procedure

The SDQ-Parent was administered at each CAMHS within the first three meetings with the family. The follow-up version was then completed between four and 8 months into treatment for 66% of cases (as recommended to match the comparison sample), before 4 months for 11% of cases, and after 8 months for 23%. There was no date information for 23% of cases. Each CAMHS submitted their SDQ-Parent and demographic data in an anonymized format for collation and analysis by CORC. All analyses and graphs were produced using the free statistics package, R version 2.15.1 (R Core Team, 2012).

Funnel plots

When trying to estimate how successful an intervention is on average, we are interested in inferring the likely effect on a *population* of service users based on the *sample* of those already seen. This inference depends on the sample size (larger sample, more precise) and spread of values (smaller *SD*, more precise).

Figure 1 shows the results from a simulation of PROMs data from fictional services. Each point represents the average difference in pre–post treatment scores at a service, such that a larger number represents a better outcome. The horizontal axis shows the sample size at each service and the vertical axis shows the outcome. The simulation was designed so that there is no actual difference in outcomes between services; all the services have a population mean outcome of 0.2. However, as Figure 1 shows, there is greater variability in the sample mean for the services providing less data. Figure 1 also shows curves for 95% and 99.9% control limits, given the known population mean and *SD*, which indicate how likely the service is to have a sample estimate within these limits. As may be seen, all points are within the 99.9% limits, but some are outside the 95% limits.

There are two major causes of variation: those which affect all parts of the system (common causes), for instance imprecision inherent in self-report measurement; and those which do not affect all parts of the system (special causes), such as between-service variations in the type of care provided, or do not affect the system all of the time, such as consequences of unplanned staff changes (Provost & Murray, 2011). It is these special causes which are of particular interest and may reflect good practice or practice which could be improved. Special causes are indicated by outliers on the funnel plots.

Results

The mean AVS as a standardized effect size for the entire sample was 0.16 (95% confidence interval = 0.15–0.18). This is a statistically reliable positive effect, but small in magnitude (Cohen, 1992). The heterogeneity of outcomes from the various services may be assessed using the I^2 statistic (Higgins, Thompson, Deeks, & Altman, 2003), which is more commonly used to estimate the percentage of total variation across studies in meta-analysis due to true heterogeneity rather than chance. For the AVS, $I^2 = 94.9\%$ ($Q(54) = 708$, $p < .001$). Deeks, Higgins, and Altman (2011) provide a ‘rough guide’ to interpret these figures: 0% to 40% ‘might not be important’; 30% to 60% “may represent moderate heterogeneity”; 50% to 90% “may represent substantial heterogeneity”; and 75% to 100% is “considerable heterogeneity”. (The intervals are overlapping to highlight that these are approximate guidelines for interpretation rather than rigid classifications.) The AVS results fall into this latter ‘considerable’ category.

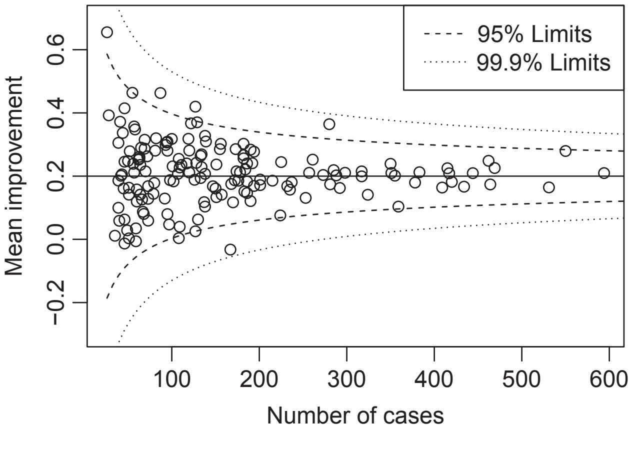

Figure 1. Plot showing average outcome for simulated services as a function of the number of cases seen. The population mean is 0.2 (shown as a horizontal line). Control limits are also plotted above and below this mean

Out of 15,771 cases, 36% moved from scoring in the clinical band at baseline to scoring in the non-clinical band at follow-up, or in other words 36% ‘recovered’ (see Table S3 for other transitions between clinical bands). Figure 2(a) shows each service in our sample ranked by their mean AVS as a standardized effect size and Figure 2(b) displays each service in our sample ranked by their recovery rates. The highest recovery rate shown in Figure 2(b) is 64%, i.e. around three in five of their patients moved from scoring in the clinical band at baseline to scoring in the non-clinical band at follow-up on the SDQ-Parent. There are many services under 30%.

Figure 3 shows the AVS plot again with 95% confidence intervals of the means, i.e. showing where the population mean is likely to be, ordered by mean rank. This highlights the uncertainty of the means, and hence rank orderings. Here the services are ordered by mean rank, but there are many different orderings depending on the ‘true value’ of the AVS.

Figure 4 displays the data from Figure 2(a) but now as a function of the sample size of the data submitted by the service. Similarly, Figure 5 plots the percentage of recovery rates data from Figure 2(b) as a function of the sample size of the data submitted by the service. Figures 4 and 5 also include control limits. If a service mean is above a particular limit, then it has a statistically significantly ‘better’ outcome compared to the national values. Similarly if the mean is below the limit, then it has a statistically significantly ‘poorer’ outcome score. The vast majority of services are within the control limits. Several services are, however, above and below the limits. For the AVS, out of 51 services, nine were above the 95% limits, seven above the 99.9% limits, 16 were below the 95% limits and seven were below the 99.9% limits. For the recovery rates data, 10 were above the 95% limits, four above the 99.9% limits, 17 were below the 95% limits and 10 were below the 99.9% limits. Importantly, there was only partial overlap between these predictions: eight were above the 95% limits for both AVS and recovery and 11 were below; four were above the 99.9% limits and five were below for both.

Discussion

We have demonstrated how the precision of measurement, affected by sample size and spread of values (*SD*), in turn affects the between-service variability in two indicators of CAMHS effectiveness. The effect size of the AVS was overall statistically significantly positive, which

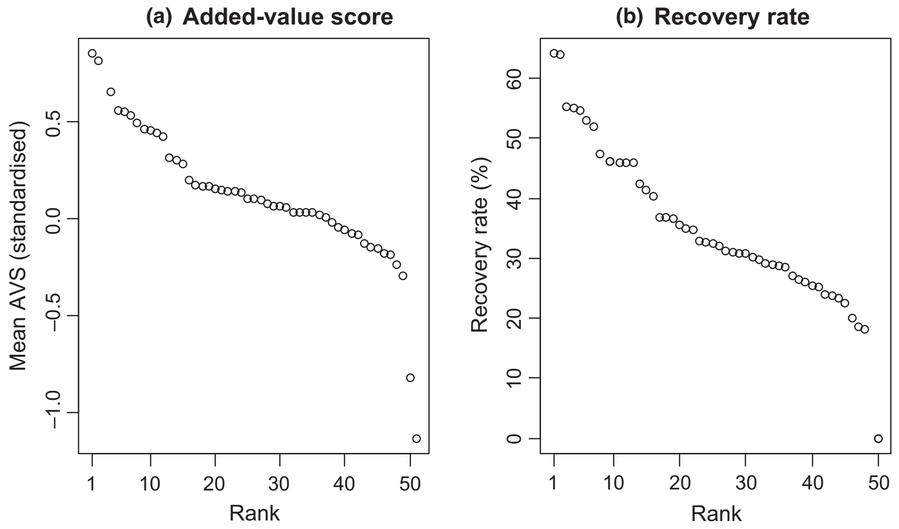

Figure 2. Services ranked by (a) mean added-value score (AVS) and (b) recovery rates

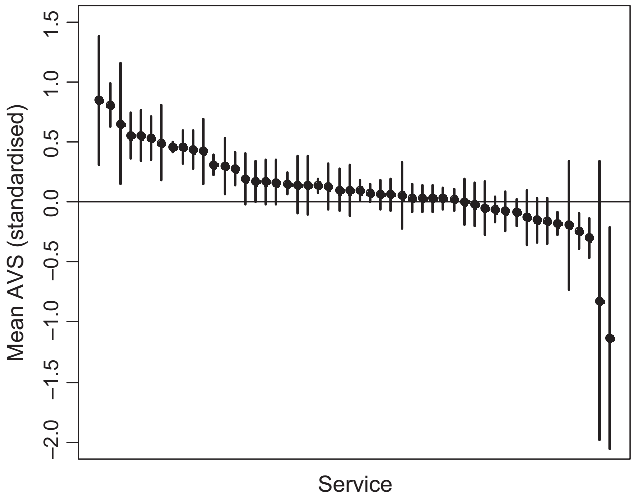

Figure 3. Caterpillar plot of mean added-value scores, showing 95% confidence intervals of the means

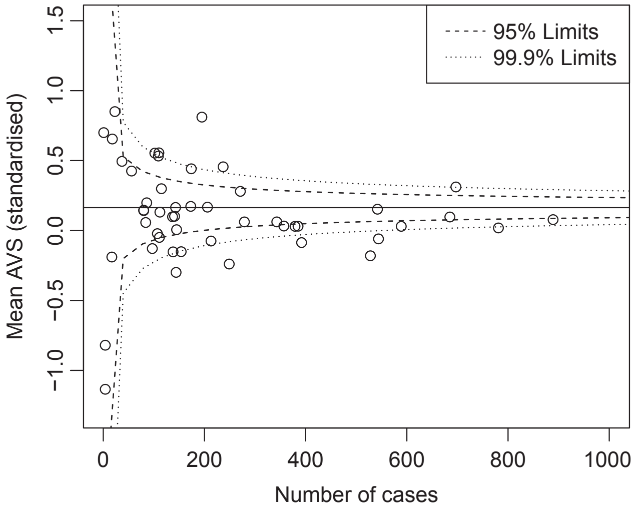

Figure 4. Each service's mean added-value score (AVS) as a function of the number of cases they submitted for analysis. Control limits are for the all-sample mean AVS of 0.16 (shown as the horizontal line) and *SD* of 1.17

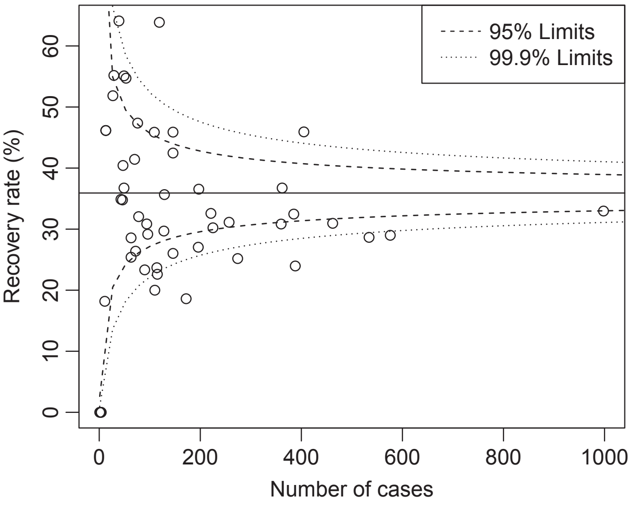

Figure 5. Each service's recovery rate as a function of the number of cases they submitted for analysis. Control limits are for the all-sample rate of 36% (shown as the horizontal line), using a method by Agresti and Coull (1998)

suggests that these children's mental health has improved more than would be expected without access to intervention. The effect size was, however, small. Around one third of those who entered treatment in the clinical band moved to 'recovery'. Since the AVS and recovery rate metrics are not perfectly correlated, different services sometimes appear as outliers depending on which metric is chosen. Finally, there was clear evidence of between-service heterogeneity in outcomes. Statistical tests of heterogeneity are more reliable than visual inspection of ranks such as those provided in the bar graph by Gyani et al. (2013) for adult IAPT. As the funnel plots show, there can be wide variation in outcomes for services that are statistically indistinguishable from each other.

Although it is statistically possible to estimate ranks using Bayesian methods (e.g. Clare Marshall & Spiegelhalter, 1998), the resulting ranks are often imprecise for the types of sample sizes encountered in clinical practice. This factor, combined with the difficulty that non-statisticians can have taking uncertainty around estimates into account, leads us to strongly discourage ranking. Although some users of data may find it easier to interpret confidence intervals (and we included an example in our analyses), these still invite a comparison of individual services with each other. Here we agree with Spiegelhalter (2005) that this between-service comparison is inappropriate and propose instead that a more important question for national analyses is whether a particular service's average outcome differs from the national average. Funnel plots are already used in surgery, but not in adult or child mental health; we think they should be.

There are some disadvantages to funnel plots. They may not be understood by the public, for instance. Given that outcomes information needs to be interpreted in the local context anyway, this is perhaps not a bad thing. Reducing a complex system of care to a single value and then expecting that single value to be used to make summative judgements about a service seems to us to be poor practice. Foley and Goldstein (2012) examine the issues of public performance data, including the use of league tables, with great care and provide as one of their recommendations that (p. 62): “Serious consideration should be given to using comparative rankings as ‘screening’ devices that are not published or made available beyond those institutions involved, but used as part of an institutional improvement programme.” Alternative analyes, including using confidence intervals, may well be more appropriate for such improvement programmes when statisticians are involved to help guide interpretations.

Knowledge of statistics is important for policy. A plot of ranks without confidence intervals is used by the Centre for Social Justice, a British think tank, to illustrate differences in clinicians' recovery rates (Callan & Fry, 2012). They use this to argue in favour of a ‘Darwinian’ approach for clinician selection such that, “only effective therapists would ‘survive’ as suppliers” (p. 35). Uncertainty in the estimates of outcomes has an impact of the precision of rank estimates, so any such ‘Darwinian’ approach would mostly be driven by chance. Similar results have been found elsewhere, for instance large heterogeneity in teacher and school effects leads to similar uncertainty in ‘value-added’ scores used in education. Massive sample sizes – unlikely to be reached in practice – would be needed to achieve adequate statistical precision for reliable results (Lockwood, Louis & McCaffrey, 2002). Ranks of the live birth rates of in vitro fertilization clinics showed great uncertainty such that only one clinic could be placed reliably in the bottom quarter of the data (Clare Marshall & Spiegelhalter, 1998). It is tempting to reject such rank-based schemes as being far from CAMHS; however, anecdotally we have already heard of counselling services that have used ranks of therapists' average outcomes to justify management decisions. We think it is important that problems with these approaches are made apparent before, not after, they are implemented.

What should happen next if a service is found to have statistically significantly higher or lower outcome scores, or recovery rates, than the national average? It is important to interpret the data in the context of the processes which generated it. Data do not speak for themselves – though performance indicators are often presented as if they do. Bullock, Little and Millham (1998) illustrate one kind of exploration we anticipate in their analysis of the long-term outcomes of young people admitted to long-stay secure treatment units. The authors predicted outcomes from variables such as whether the young people received psychotherapy and the severity of criminal offending. Cases which fell outside prediction intervals were then investigated in more detail to understand why. A similar approach could be taken using routine mental health outcomes data. It is important that the clinicians who provided the care are involved in these investigations and not only service managers or performance leads who will know less about the local clinical context. Service users who are ‘experts by experience’ should also be consulted.

Why might a service be an outlier on a funnel plot?

It can be easy to forget the range of factors that may influence results – especially in the context of an upcoming meeting with commissioners, the result of which might affect funding. Here, we provide a brief overview of the factors that we have frequently encountered when discussing outcomes analyses with service managers, clinicians and commissioners. This provides clues that we think will help the investigations of outliers on funnel plots and provide some contextualization of results. As we have emphasized, such contextualization is an integral part of analysing and reporting outcomes.

Data quality

Values can easily be inadvertently miscoded in large datasets with many dozens of different variables. We have seen missing data in follow-up data accidentally being coded as zero, which was a valid value representing an absence of symptoms. This was easy to see on a funnel plot as the service's outcomes were three standard deviations better than any other service so highly unlikely; it was less easy to see in other presentations of the data which did not highlight sample size. If results start to be used to inform payment or commissioning decisions, additional data quality issues are likely to arise due to risk of ‘gaming’ (Bevan & Hood, 2006).

Return rates

There is some evidence to suggest that service users with poorer outcomes are less likely to complete follow-up questionnaires, so outcomes are then inflated (Clark, Fairburn, & Wessely, 2008). As a result, services most efficient at data collection may show worse outcomes. This is important information in the context of Commissioning for Quality and Innovation targets, where generally CAMHS are rewarded for return rates on questionnaires. Increasingly discussions are starting around rewarding positive outcomes. Funnel plots could also be applied to return rates when the number of referrals to services is known, although IT systems can make this difficult to extract. The overall return rates, as a proportion of referrals, are unknown for the present dataset as this information was not recorded. However, it is now beginning to be extracted routinely.

What is covered in the questionnaire

Effect magnitudes tend to be smaller for measures covering a broad range of problems compared to specific measures (Lee, Jones, Goodman & Heyman, 2005).

Effects can also be artificially inflated if problems are not covered in a measure. For instance someone with social phobia might show a reduction in anxiety because of phobic avoidance; if you do not go out and see people then you will not get anxious. CYP and adult IAPT supplement use specific anxiety measures to overcome this problem (Clark, 2011; Wolpert, Fugard et al., 2012).

Chosen outcome variable

For a given questionnaire, there are different choices of variable that one could use as a measure of outcome. We have shown analyses of recovery rates and the AVS; there was overlap but also differences in which services were highlighted as outliers. Services with a higher mean score on a given outcome variable at baseline will show more regression to the mean and hence more improvement on pre–post change. This can be adjusted using the AVS though such adjustment is currently only available for the SDQ-Parent. Services with higher mean scores on a given outcome variable at baseline are also likely to have lower recovery rates, as scores need to reduce further to cross the clinical cut point.

Case mix

Case mix and the severity of presenting problems at outset can additionally have an impact on outcomes. For instance a study of outcomes of nearly 10,000 young people found that diagnostic group was a statistically significant predictor of improvement (Ogles, Carlson, Hatfield, & Karpenko, 2008). Statistical case mix adjustment can be performed to take these factors into account. However, given that the reliability of items on problem checklists varies considerably between problems (Hanssen-Bauer, Aalen, Ruud & Heyerdahl, 2007) and that selections can be incomplete, even relative to known information about a case, it is important that any adjustment is not treated as summative. Rather, we suggest auditing a random selection of case notes to explore possible moderating factors.

Organizational factors

Service 'restructuring', e.g. reducing staff numbers and asking people to reapply for their own jobs, seems anecdotally to be associated with a drop in patient-reported outcomes in UK CAMHS, but this still needs to be empirically investigated. There is relevant evidence from child welfare services in the United States of a link between organizational climate (how organizations are experienced by those who work in them) and children's outcomes. Children showed better outcomes if they were involved with services where caseworkers had a shared feeling that they were able to make a positive contribution through their work and be personally involved with and concerned about individual children and families (Glisson & Green, 2011). These factors also predict more positive attitudes towards evidence-based practice, as does lower stress (Aarons et al., 2012).

Stage in episode of care

Some services distribute questionnaires on first contact and then 6 months later, irrespective of whether any care has been received. This might sound surprising, but may be explained by the complexity of data collection which involves a range of people (e.g. assistant psychologists, administrative staff, performance leads) who might not all have access to information about how much clinical contact someone has received. In some hard-pressed services, service users may have attended only one appointment for assessment and still be on an 'internal' waiting list 6 months later. One would not expect improvement without intervention, especially if the AVS is used. One solution is simply to record stage of care, such as 'on waiting list', however, it can take time for national datasets to accommodation changes such as these.

There are also difficulties in determining what constitutes contact with a service. For example, much work done by CAMHS is 'indirect' work with other agencies which may or may not be recorded, and the fact of being in contact with services regardless of degree of direct contact may all impact on outcomes, making it difficult to determine a simple system for linking outcomes to simple service use. Another example comes from paediatric liaison work where a large improvement in psychological outcomes may be driven by physical health input, for instance pain relief, which may not be recorded on mental health datasets.

Therapeutic modality

There has been a vast quantity of research on the possible benefits of one modality over another. Routine outcomes monitoring often includes checklists for common modalities: can we assess the impact of different kinds of intervention using data from these? Clark et al. (2008) suggest not, providing an example of a client who believed he had received CBT, but what he described was clearly neither CBT nor effective. A tick-box against 'CBT' would have been misleading. On the other hand, Gyani et al. (2013) used a similar checklist to show that the type of intervention was correlated with outcomes and in the hypothesized direction: service users receiving a 'high intensity' treatment showed better outcomes than those who received a treatment categorized as 'other'. This lends support for the validity of checklists, however, should be interpreted with caution since whatever problems tend to lead to the 'other' types of intervention could also be those with worse outcomes, irrespective of the intervention used. Fidelity checklists might also help to evaluate what care was provided and again there is much research in this area. There is also overlap in therapeutic brands which has led researchers to investigate finer grained processes used in a range of approaches. This has led, for example, to the Behavioural Change Technique (BCT) Taxomomy (Michie et al., 2013) and The Taxonomy Project (Tschacher, Junghan, & Pfammatter, 2014). It might be worth considering pilots of these taxonomies in routine care.

Therapist skill

Even if the box-ticking for therapeutic modality is accurate, this does not ensure therapist competency. A study from 2007 showed that CAMHS clinicians who use CBT for fewer than one in five of their cases had limited training; nearly half had only learnt how to deliver CBT by attending one to three scientific meetings (Stallard, Udwin, Goddard, & Hibbert, 2007). This is in stark contrast to accredited training programmes, such as the British Association of Cognitive and Behavioural Therapists, which recommends one and 2 year training courses depending on the level of accreditation (www.babcp.com). CYP IAPT aims to improve this by providing postqualification training in a range of therapies including CBT and parenting programs.

One of the goals of outcomes measurement is to evaluate clinician performance. As the discussion above attempts to illustrate, it does not follow that a poor outcome for a service implies that therapists at the service are underskilled. This means that others sources of information are necessary. Video evaluation by a skilled therapist, often used during training, is probably the best way but time consuming. Alternatives are individual and peer supervision.

Goals set for treatment

Goal setting and goal-based outcomes monitoring is increasingly being used in UK CAMHS so many services have databases of the goals patients have set for therapy and a record of progress towards achieving these goals (Bradley, Murphy, Fugard, Nolas, & Law, 2013). This provides an additional source of information to aid investigation of outliers.

Conclusions

We have proposed funnel plots as a helpful way to present outcomes information from institutions, following their use in a range of other contexts in healthcare. But this is just part of how we think outcomes should be used: data alone are never enough. The data has to be interpreted and placed in the context in which it was collected. Our plea is that there is no data analysis, particularly no ranking of services' outcomes data, and importantly no management decisions, without careful investigation of the local context.

Acknowledgements

This study has not received any external funding. A. F. and E.S. previously worked for CORC. T.F., D.L. and A.Y. are unpaid members of the CORC committee. D.L. and M.W. are members of CYP IAPT Outcomes Evaluation Group. M.W. is National Informatics Lead for CYP IAPT and paid Director of CORC. A.Y. is Chair of the CYP IAPT Service Development Group. All authors were involved in drafting and critical revision of the article and contributed to conversations around how to help commissioners, managers, and clinicians understand data and make wise decisions. A.F. conceived the idea for the article and contributed to data analyses. E.S. also contributed to data analyses. Views expressed here are not necessarily those of CORC members or the Department of Health. The authors thank Roger Davies, David Foreman, James Mountford, Catriona McCallum, Peter Fonagy and Ruth Sweeting for helpful discussions, and to Thomas Booker for editorial assistance. The authors have declared that they have no competing or potential conflicts of interest.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Table S1. Information about the sample.

Table S2. Sample sizes (*N*), means (*M*) and standard deviations (*SD*) for each scale of the SDQ-Parent.

Table S3. The number of cases who had clinical and non-clinical levels of distress according to their parent or carer.

References

Aarons, G.A., Glisson, C., Green, P.D., Hoagwood, K., Kelleher, K.J., Landsverk, J.A., . . . & Schoenwald, S. (2012). The organizational social context of mental health services and clinician attitudes toward evidence-based practice: A United States national study. *Implementation Science*, *7*, 56.

Agresti, A., & Coull, B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126.

Bevan, G., & Hood, C. (2006). What's measured is what matters: Targets and gaming in the English public health care system. *Public Administration*, *38*, 629–634.

Bickman, L., Kelley, S.D., Breda, C., de Andrade, A.R., & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services*, *62*, 1423–1429.

Bradley, J., Murphy, S., Fugard, A.J.B., Nolas, S.-M., & Law, D. (2013). What kind of goals do children and young people set for themselves in therapy? Developing a goals framework using CORC data *Child and Family Clinical Psychology Review*, *1*, 8–18.

Bullock, R., Little, M., & Millham, S. (1998). Secure treatment outcomes: The care careers of very difficult adolescents. Aldershot: Ashgate.

Callan, S., & Fry, B. (2012). Commissioning effective talking therapies. London: The Centre for Social Justice.

Clare Marshall, E., & Spiegelhalter, D.J. (1998). Reliability of league tables of in vitro fertilisation clinics: Retrospective analysis of live birth rates. *British Medical Journal*, *316*, 1701–1705.

Clark, D.M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry*, *23*, 318–327.

Clark, D.M., Fairburn, C.G., & Wessely, S. (2008). Psychological treatment outcomes in routine NHS services: A commentary on Stiles et al (2007). *Psychological Medicine*, *38*, 629–634.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Deeks, J.J., Higgins, J.P.T., & Altman, D.G. (2011). Chapter 9: Analysing data and undertaking meta-analyses. In J.P.T. Higgins & S. Green (Eds.), Cochrane handbook for systematic reviews of interventions version 5.1.0 (updated March 2011). The Cochrane Collaboration. Available from: http://www.cochrane-handbook.org. [last accessed 12 December 2014].

Department for Education. (2013). *Performance tables 2012*. Available from http://www.education.gov.uk/cgi-bin/schools/performance/group.pl?qtype=GOR&superview=sec&view=aat&set=2&sort=ks4_12.ptgac5em&ord=desc&tab=150&no=J&pg=12. [last accessed 12 December 2014].

van Dishoeck, A.M., Looman, C.W., van der Wilden-van Lier, E.C., Mackenbach, J.P., & Steyerberg, E.W. (2011). Displaying random variation in comparing hospital performance. *BMJ Quality & Safety*, *20*, 651–657.

Foley, B., & Goldstein, H. (2012). Measuring success: League tables in the public sector. London: British Academy.

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.

Glisson, C., & Green, P. (2011). Organizational climate, services, and outcomes in child welfare systems. *Child Abuse and Neglect*, *35*, 582–591.

Goodman, R. (1999). The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *40*, 791–799.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*, 1337–1345.

Gyani, A., Shafran, R., Layard, R., & Clark, D.M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy*, *51*, 597–606.

Hanssen-Bauer, K., Aalen, O.O., Ruud, T., & Heyerdahl, S. (2007). Inter-rater reliability of clinician-rated outcome measures in child and adolescent mental health services. *Administration and Policy in Mental Health*, *34*, 504–512.

Healthcare Commission. (2005). *Data for 2004/2005 NHS Performance Ratings*. Available from: http://webarchive.nationalarchives.gov.uk/20090321144733/http://ratings2005.healthcarecommission.org.uk/more_information.asp. [last accessed 12 December 2014].

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560.

Jensen, P., Roper, M., Fisher, P., Piacentini, J., Canino, G., Richters, J., ... & Schwab-Stone, M. (1995). Test-retest reliability of the Diagnostic Interview Schedule for Children (DISC 2.1). Parent, child, and combined algorithms. *Archives of General Psychiatry*, *52*, 61–71.

Lambert, M.J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy*, *48*, 72–79.

Lee, W., Jones, L., Goodman, R., & Heyman, I. (2005). Broad Outcome Measures May Underestimate Effectiveness: An Instrument Comparison Study. *Child and Adolescent Mental Health*, *10*, 143–144.

Lockwood, J.R., Louis, T.A., & McCaffrey, D.F. (2002). Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability Systems. *Journal of Educational and Behavioral Statistics*, *27*, 255–270.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., ... & Wood, C.E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine*, *46*, 81–95.

National Joint Registry (2014). Surgeon and Hospital Profile. Available from: http://www.njrsurgeonhospitalprofile.org.uk. [last accessed 12 December 2014].

Health & Social Care Information Centre. (2014). Quarterly improving access to psychological therapies data set reports, England - Final Q4 2013-14 summary statistics and related information, Experimental statistics. Available from: http://www.hscic.gov.uk/pubs/iapt1314q4 [last accessed 12 December 2014].

Ogles, B.M., Carlson, B., Hatfield, D., & Karpenko, V. (2008). Models of case mix adjustment for Ohio Mental Health Consumer Outcomes among children and adolescents. *Administration and Policy in Mental Health*, *35*, 295–304.

Programme for International Student Assessment (PISA). (2009). *PISA 2009 Results: What Students Know and Can Do*. Student Performance in Reading, Mathematics and Science: Volume 1. Available from: http://www.oecd.org/pisa/pisaproducts/48852548.pdf. [last accessed 12 December 2014].

Provost, L.P., & Murray, S.K. (2011). The health care data guide: Learning from data for improvement (1st edn). San Francisco, CA: Jossey-Bass.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Spiegelhalter, D.J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine*, *24*, 1185–1202.

Stallard, P., Udwin, O., Goddard, M., & Hibbert, S. (2007). The Availability of Cognitive Behaviour Therapy Within Specialist Child and Adolescent Mental Health Services (CAMHS): A National Survey. *Behavioural and Cognitive Psychotherapy*, *35*, 501–505.

Tschacher, W., Junghan, U.M., & Pfammatter, M. (2014). Towards a taxonomy of common factors in psychotherapy-results of an expert survey. *Clinical Psychology & Psychotherapy*, *21*, 82–96.

Wolpert, M., Ford, T., Trustam, E., Law, D., Deighton, J., Flannery, H., & Fugard, A.J.B. (2012). Patient-reported outcomes in child and adolescent mental health services (CAMHS): Use of idiographic and standardized measures. *Journal of Mental Health*, *21*, 165–173.

Wolpert, M., Fugard, A.J.B., Deighton, J., & Görzig, A. (2012). Routine outcomes monitoring as part of children and young people's Improving Access to Psychological Therapies (CYP IAPT) – improving care or unhelpful burden? *Child and Adolescent Mental Health*, *17*, 129–130.

Youthinmind (2009). *An Added Value Score for Specialist Services*. Available from: http://www.sdqinfo.org/c5.html. [last accessed 12 December 2014].

Accepted for publication: 20 November 2014
Published online: 24 December 2014