Journal of Public Health: From Theory to Practice (2019) 27:29–35
https://doi.org/10.1007/s10389-018-0921-9

ORIGINAL ARTICLE

# Comparing single-level and multilevel regression analysis for risk adjustment of treatment outcomes in common mental health disorders

Lisanne Warmerdam[1] · Edwin de Beurs[1] · Marko Barendregt[1] · Jos Twisk[2,3]

Received: 1 February 2018 / Accepted: 11 April 2018 / Published online: 24 April 2018


## Abstract

**Aim** The aim of this paper is to compare single-level and multilevel regression analysis to obtain risk-adjusted outcomes from mental health care providers.

**Subject and methods** The study population consisted of adult patients receiving treatment for common mental health disorders. The outcome was self-reported symptom level at post-test. Risk adjustment models were developed using single- and multilevel regression analysis. In the multilevel approach, a random intercept for each provider was included. The intraclass correlation coefficient was used to estimate the proportion of variability in treatment outcome between providers. Spearman correlation coefficient of ranks was used to compare results between the two approaches.

**Results** The effects of most casemix variables on outcomes were similar for the two models. The ranking of providers in both methods was also quite similar ($\rho = .99$). The multilevel model estimated that 5.4% of total variability in adjusted post-test scores was explained by the provider factor.

**Conclusions** The findings of risk adjustment of mental health outcomes are quite robust for the use of single-level or multilevel regression analysis in the current study. However, given the small but significant amount of variation in outcomes that is attributable to providers, the multilevel approach is recommended for dealing with outcomes when patients are clustered within providers.

**Keywords** Risk adjustment · Mental health care · Multilevel analysis · Outcomes

## Introduction

Over recent years, worldwide interest in using outcome studies to evaluate performance of mental health services has been increasing (Hermann et al. 2007; Moran and Jacobs 2015; Rosen et al. 2009). To ensure a fair comparison between mental health care providers with regard to their treatment outcomes (e.g., number of deaths or % recovery), risk adjustment needs to be applied. Risk adjustment is a statistical procedure to control for differences in casemix when comparing groups on outcomes of interest (Iezzoni 2013).

Traditionally, there are two methods for adjusting outcome results; direct and indirect standardization. Both direct and indirect standardization involves the calculation of expected outcomes, which are compared to the number of observed outcomes (Israels 2013). Direct and indirect standardization are both techniques which require many observations. Accordingly, parametric approaches have been developed, such as regression analyses. Regression modeling is the most commonly used method for risk-adjusting outcomes. It's a powerful tool for making fairer comparisons among providers with different types of patients. Comparison of observed to expected outcomes is central to performance profiling (Iezzoni 2013). The existing single-level regression models, most often used in outcome studies, treat all patients as independent observations and ignore the fact that these patients are

✉ Lisanne Warmerdam
lisanne.warmerdam@sbggz.nl

[1] Stichting Benchmark GGZ (the Dutch Benchmark Foundation in Mental Health Care), Rembrandtlaan 46, 3723 BK Bilthoven, the Netherlands

[2] Department of Epidemiology & Biostatistics, VU University Medical Center, Amsterdam, the Netherlands

[3] Department of Health Sciences, VU University, Amsterdam, the Netherlands

grouped within providers. However, patients receiving therapy within the same provider may be correlated, violating one of the basic assumptions of traditional regression analysis.

Multilevel models account for the correlations among observations from the same provider and give an estimate of that correlation. Multilevel models consider the providers involved in a study as a random sample from the population of all providers, and divide the random variability of data into variability between patients within providers and between providers. The use of multilevel modeling is recommended when observations are clustered within more than one level; e.g., individual patients (level 1) are clustered within providers (level 2) (Burgess et al. 2000; Goldstein and Spiegelhalter 1996). One of the advantages of multilevel modeling is that it allows the use of covariates measured at any of the levels. It enables one to explore the extent to which differences in outcomes between providers are the result of factors such as organizational structure, particular therapies, or possibly some patient characteristics.

The use of multilevel models to estimate provider outcomes represents a statistical approach which is still relatively new but is becoming more common. The field of risk adjustment in mental health care is also relatively new, and testing this approach in this field deserves attention. In a previous paper, we described a casemix model using single-level regression analysis. We demonstrated that for the majority of providers, risk adjustment did not markedly change their ranking on clinical outcomes compared to unadjusted outcomes (Warmerdam et al. 2017). The goal of this study is to evaluate single-level and multilevel regression analysis for risk adjustment when comparing clinical outcomes of mental health care providers.

## Methods

### Data

Since 2010, the Dutch foundation for benchmarking mental health care (Stichting Benchmark GGZ, SBG) has gathered outcome data on a nationwide basis. SBG provides information regarding treatment outcomes for their stakeholders (providers, patients, and financiers) with the goal of monitoring, comparing and, ultimately, improving the quality of mental health care. At this moment, almost 90% of the providers are submitting data to SBG. Since 2010, the number of treatments with pre- and post-test outcome data are on the rise, ranging from about 10% in 2010 to almost 50% of all completed treatments in 2016. The data for the current study stem from an adult population aged 18 years and older with a variety of common mental health diagnoses. The goal of treatment was to recover from the disorder primarily by reducing symptoms. Patients with severe mental disorders were excluded.

Between June 2014 and May 2015, data of 143,128 completed treatment trajectories were submitted to SBG. For 37,191 treatments (26%), pre- and post-test scores were available to calculate treatment outcome. Information of demographic and clinical variables of patients is also collected to enable casemix correction. To be included in this study, the following two criteria were applied to providers; 1) information of all casemix variables is present in at least 80% of the pre–post measured treatments, and 2) the provider has at least $N = 30$ measured treatments with information about casemix variables. This resulted in a final dataset consisting of $N = 31,849$ measured treatments from 85 different providers, with complete pre- and post-test data and complete casemix variables.

### Measurement of outcome

The following self-report questionnaires for psychopathology are used to assess treatment outcome in this study: The Symptom Checklist-90 (SCL-90; Arrindell and Ettema 2003; Derogatis 1975a), the Brief Symptom Inventory (BSI; de Beurs 2008; Derogatis 1975b), the Outcome Questionnaire (OQ-symptomatic distress subscale; de Jong et al. 2007; Lambert et al. 2004), the Depression Anxiety Stress Scales (DASS; Lovibond and Lovibond 1995), the Clinical Outcomes in Routine Evaluation (CORE-problems subscale; Barkham et al. 2005), and the Short Symptoms List (Korte Klachten Lijst; KKL; Lange and Appelo 2007). All these questionnaires are valid and reliable generic instruments for the assessment of severity of psychopathology. To bring scores on the various questionnaires to a common metric, pre- and post-test scores have been normalized into T-scores with a mean of 50 and a standard deviation of 10 (de Beurs 2010) and transformed in order to have a normal distribution of scores. Higher T-scores correspond to more self-reported symptoms. The outcome in this study is the T-score on the post-test.

### Patient characteristics

Information about the following demographic and clinical variables were gathered by providers during the intake phase of the treatment and used in the development of the casemix model: year of birth (used for estimating age), gender, social economic status (SES), and urbanization level. SES and urbanization have five levels, with higher levels representing higher social economic status and lower urbanization. Age, urbanization, and SES were treated as continuous variables in all analyses. Clinical variables were: the pre-test T-score on the outcome measure, the score on the Global Assessment of Functioning (GAF; Hall 1995), and primary diagnosis according to DSM-IV axis I and axis II (American Psychiatric Association 1994).

The GAF represents axis V in the DSM-IV (American Psychiatric Association 1994); this is a single item rating scale for the severity of illness and functional impairment in psychiatry, with acceptable psychometrics (Hall 1995). Theoretically, the score on the GAF varies between 0 (dead) and 100 (perfect health). Scores on the GAF were recoded into three categories (severe problems: 1—50, moderate problems: 51–60, some problems: 61–100). The primary diagnoses were grouped into 26 main categories of the DSM-IV and dummy recoded.

### Statistical analyses

The development of the risk adjustment model is described elsewhere more elaborately (Warmerdam et al. 2017). In short, simple and multiple regression analyses were used to select casemix variables that best predicted post-test level of symptoms. Split sample validation was applied to validate the model. The coefficients of the casemix variables in the final risk adjustment model are displayed in Table 2. Based on the regression coefficients of single-level multiple regression analysis, a predicted post-test score for each observation was calculated. A relative performance factor (RPF) was determined for each patient as the ratio between the actual post-test score and the predicted post-test score. RPFs were aggregated to the provider level to get a mean RPF for each provider. To obtain the risk-adjusted post-test scores for each provider, the mean RPF was multiplied by the observed post-test score across all providers (Nuttall et al. 2013). Providers with scores significantly higher (worse) or lower (better) than the national mean were identified as high outliers or low outliers respectively. Finally, risk-adjusted post-test scores were ordered to obtain providers' ranking based on the single-level regression model.

The same analyses were performed with multilevel analysis. The difference between the prediction based on a multilevel analysis and a single-level analysis is that each provider has its own specific intercept, while in a single-level regression one intercept is used for all providers. In addition, the regression coefficients can slightly differ between the two methods, due to the adjustment for the correlated observations within the provider. To estimate the variability in post-test scores between providers, the intraclass correlation coefficient (ICC) was calculated, adjusted for casemix variables. The ICC represents the percentage of variance explained by the clustering of patients within providers. The ICC is calculated as the estimated between-provider variance divided by the sum of the estimated between-provider and within-provider variance (Snijders and Bosker 1999).

In contrast to classical single-level regression, the multilevel model estimates 'provider effects' using different intercepts for each provider. These provider effects (also known as second-level residuals) represent the distance between estimated post-test scores and the overall estimated post-test score. In order to compare providers post-test outcome with the overall mean, provider effects were ordered from the smallest to the largest and graphically presented with their 95% confidence intervals (CI). Providers with effects significantly less than zero performed better than the overall mean, whereas providers with effects significantly higher than zero performed worse (Goldstein and Healy 1995). To compare the ranking of the providers based on a single–level regression analysis and a multilevel regression analysis, a Spearman's rank correlation coefficient (ρ) was calculated. All analyses were performed using SPSS version 19.

## Results

### Sample characteristics

Differences between the sample ($n = 31,849$) and the excluded patients ($n = 111,279$) on casemix variables were tested with effect sizes (Eta squared for continuous variables and Cramer's V for nominal variables). The highest value found was V = 0.05 for the primary diagnoses, with an overrepresentation of eating disorders in the included group. All other effect sizes were below 0.05.

Table 1 reports descriptive information of the sample. Patients were mostly female (62.8%), and the mean age was 38.7 (± 13.0) years. Most patients suffered from a mood disorder (33.9%) or an anxiety disorder (25.2%). According to scores on the GAF, half of the patients had a moderate level of functioning in daily life. The sample as a whole scored $T = 50.4$ ($SD = 9.5$) at the pre-test; at post-test, the mean symptom level was $T = 41.4$ ($SD = 10.7$). On average, patients improved almost a standard deviation (Cohen's d = .90).

### Estimation results

According to the single-level regression model, the following variables were most strongly related with the post-test score: T-score at pre-test, level of functioning (GAF), age, SES, and the presence of a personality disorder or somatoform disorder (see Table 2). These variables accounted for 28% of the variance in post-test scores, of which the T-score at pre-test explained by far the largest part of this variance (partial $R^2 = 22.9\%$). Most variables that were significantly related to the post-test scores in the single-level regression model are also significant predictors in the multilevel model. There is one exception; according to the multilevel model, the presence of a somatoform disorder is no longer significantly related to symptom level at post-test (Table 2). The ICC (adjusted for casemix variables) in the multilevel model was 5.4% ($p < .001$) which means that 5.4% of the variance in post-test outcomes is due to differences between providers.

**Table 1** Descriptive statistics ($n = 31,849$)

| Variable | Level | % | Variable | % |
|---|---|---|---|---|
| Gender | Male | 37.2 | Other mood disorder | 3.7 |
| Age (M, SD) | | 38.7 (13.0) | Panic disorder | 4.7 |
| SES | Low 1 | 24.0 | Social phobia | 2.8 |
| | 2 | 20.1 | Obsessive–compulsive disorder | 2.0 |
| | 3 | 18.1 | Posttraumatic stress disorder | 7.2 |
| | 4 | 18.9 | Acute stress disorder | 0.2 |
| | High 5 | 19.0 | Generalized anxiety disorder | 3.2 |
| Urbanization | Urban 1 | 25.2 | Other anxiety disorder | 4.7 |
| | 2 | 25.9 | Developmental disorder | 11.3 |
| | 3 | 23.0 | Substance dependency/abuse | 0.5 |
| | 4 | 16.2 | Psychotic disorder | 0.8 |
| | Rural 5 | 9.6 | Sexual disorder | 0.6 |
| Score on pre-test (M, SD) | | 50.4 (9.5) | Sleeping disorder | 0.2 |
| Score on post-test (M, SD) | | 41.4 (10.7) | Somatoform disorder | 8.4 |
| Change score | | 9.1 (10.1) | Impulse control disorder | 1.2 |
| GAF | < 51 | 28.2 | Dissociative disorder | 0.1 |
| | 51-60 | 49.1 | Eating disorder | 4.1 |
| | > 60) | 22.6 | Personality disorder type A | 0.1 |
| Depressive disorder first episode | | 14.3 | Personality disorder type B | 3.5 |
| Depressive disorder recurrent | | 12.2 | Personality disorder type C | 4.1 |
| Dysthymic disorder | | 2.9 | Personality disorder (not otherwise specified) | 5.3 |
| Bipolar disorder | | 0.8 | Other disorders | 0.5 |

M = median, SD = standard deviation

## Comparison of the models

Provider effects acquired by the multilevel model are presented in Fig. 1, with their 95% confidence interval. The figure shows that 21 providers (25%) had an effect significantly lower (better) than zero. Their effects ranged from −5.97 to −0.98. Another 21 providers (25%) had effects significantly higher (worse) than zero (ranging from 1.08 to 3.49). The remaining providers ($n = 43$, 51%) scored not significantly different from zero.

Using the classical single-level approach, risk-adjusted post-test score were found to be significantly lower (better)

**Table 2** Regression coefficients casemix variables in the final risk adjustment model

| | Single-level regression analysis | | Multilevel analysis | |
|---|---|---|---|---|
| | B | SE | B | SE |
| Intercept | 12.08* | .34 | 12.78* | .42 |
| Pre-test score | 0.53* | .01 | 0.53* | .01 |
| GAF < 51 | 3.37* | .15 | 3.31* | .17 |
| GAF = 51–60 | 1.61* | .13 | 2.01* | .14 |
| GAF > 60 | REF | | REF | |
| Somatoform disorder | −1.36* | .19 | −0.30 | .20 |
| Personality disorder type B | 2.00* | .28 | 1.42* | .27 |
| Personality disorder type (not otherwise specified) | 2.00* | .23 | 1.19* | .23 |
| Age | 0.03* | .00 | 0.03* | .00 |
| SES | −0.26* | .04 | −0.24* | .04 |

* < .001; REF = reference group

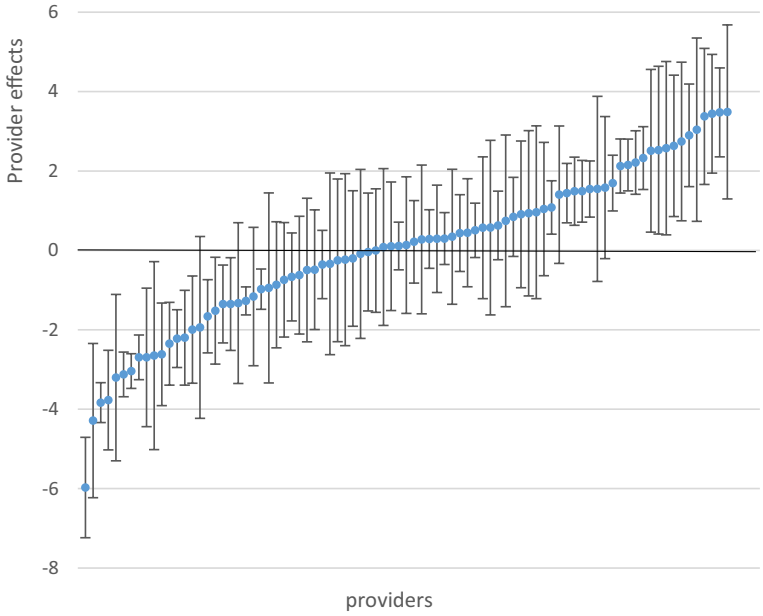


**Fig. 1** Provider effects and their confidence intervals obtained by the multilevel model

than the mean for 16 providers (19%) and significantly higher (worse) for 29 providers (34%). Forty providers (47%) did not deviate from the mean. The multilevel model confirmed that the 16 providers identified by the single-level approach as positive outliers performed better than the mean. However, another five providers were identified as positive outliers in the multilevel model. From the 29 providers among the negative outliers identified by the single-level model, eight were found to perform not differently from the mean using the multilevel model.

The rankings of providers according to the single-level and the multilevel approach correlated with $r = 0.99$ (Fig. 2).

## Discussion

The present study showed similar findings for the two risk-adjustment methods (single-level regression and multilevel regression) in ranking of mental health care providers on self-reported clinical outcomes (Spearman's $\rho = 0.99$). The current study showed further that the effects of most casemix variables on outcomes are comparable between the single and the multilevel model. The multilevel model showed that 5.4% (ICC) of the differences in the adjusted post-test scores were attributable to differences between providers. In the study of Moran and Jacobs (2015), where a clinician rating scale (HoNOS) was used to measure functional level of patients, the ICC lies between 2% and 11% (respectively, linear and ordered probit model). As far as we know, no other studies using multilevel analysis in the field of risk adjustment on mental health outcomes have been published. The low ICC in this study demonstrates the low degree of correlation among observations within a provider. When observations are more highly correlated, the difference of standard errors between single-level and multilevel regression becomes greater. In our

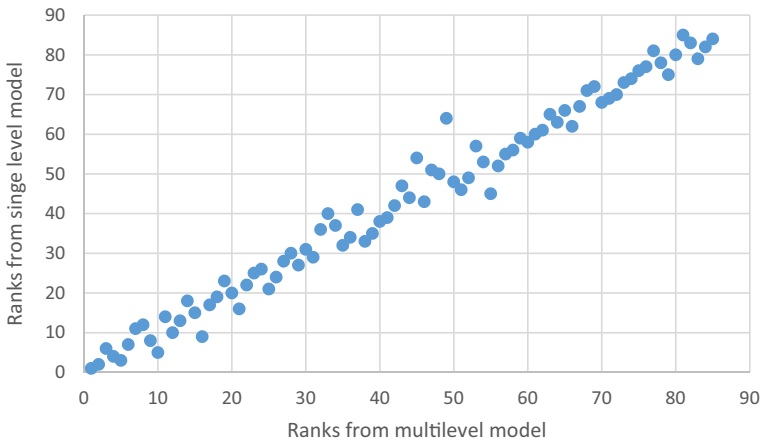


**Fig. 2** Scatterplot of ranks obtained by single-level linear regression and multilevel regression

study, differences of standard errors between the two models are minimal, consistent with the low ICC.

Based on the analysis of provider effects (group-level residuals), providers were divided into three performance groups: better than average, equal, and worse than average. Identification of outliers resulted in fewer negative outliers using the multilevel model compared to the single-level model. On the other hand, the multilevel model identified more positive outliers, and more providers scoring not different from the mean.

In general, multilevel models including random effects are slightly more conservative in identifying outliers compared with traditional approaches such as standard regression. Provider-specific estimates of providers with few observations are shrunk towards the mean of the population (Arling et al. 2007). This guards against extreme outcomes from providers with a small number of cases. On the other hand, single-level regression does not allow individual providers to be influenced by the group as a whole, and might be preferable if the purpose of a model is to determine which providers perform significantly better or worse. Therefore, it is important to consider the purpose of using a certain method.

A limitation of this study is that no variables relating to provider characteristics such as resources, providers' choice of therapeutic approach, or any other variables that might influence outcomes, were examined. These provider-level factors might have an effect (either positive or negative) on outcomes that could not be detected in this study, which included only variables associated with outcome at the patient level.

The goal of risk adjustment is to make outcomes more comparable and to find explanations for the differences in outcomes between providers. It is argued that differences in the quality of care *within* providers (e.g., locations, teams of therapists) may be greater than differences *between* providers (Lovaglio 2012). In this case, multilevel analysis is highly suited to unravel these outcomes at the various levels. In the Netherlands, the diversity among providers is substantial, ranging from large integrated mental health care providers to smaller specialized mental health institutions. Both types of mental health institutes may be subdivided in various locations, and within these locations various departments or teams may exist. This study is limited by including two levels (the patient level and the provider level) in the analysis, while including more levels could provide more insight into sources of variation in outcomes at these different levels.

This study contributes to the current knowledge about risk adjustment of mental health outcomes by comparing a single-level with a multilevel approach. By using the hierarchical structure of this dataset, we can make inferences about the influence of different levels on outcomes. Given the significant (albeit small) amount of variation attributable to providers, the multilevel approach is the most appropriate method for dealing with outcomes when patients are clustered within providers. In this study, the two approaches reveal little difference in predictive validity. Furthermore, the use of multilevel modeling does not lead to a very different ranking of providers. This means that the findings of risk adjustment of mental health outcomes in the current study are quite robust for the used methodologies.

## Compliance with ethical standards

**Ethical approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

**Informed consent** The Dutch Central Medical Ethical Committee (CCMO) has ruled that Dutch Law regarding research with humans does not apply to the collection of anonymized information and, consequently, providing SBG with this data and analyzing anonymized data for the present study does not require additional informed consent from participants.

**Conflicts of interest** The authors declare that they have no conflicts of interest.

## References

American Psychiatric Association (1994) Diagnostic and statistical manual of mental disorders, 4th edition, (DSM-IV). American Psychiatric Association, Washington DC

Arling G, Lewis T, Kane RL, Mueller C, Flood S (2007) Improving quality assessment through multilevel modeling: the case of nursing home compare. Health Serv Res 43:1177–1199. https://doi.org/10.1111/j.1475-6773.2006.00647.x

Arrindell WA, Ettema JHM (2003) SCL-90: Herziene handleiding bij een multidimensionale psychopathologie indicator. Swets & Zeitlinger, Lisse

Barkham M, Gilbert N, Connell J, Marshall C, Twigg E (2005) Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings. Br J Psychiatry 186:239–246. https://doi.org/10.1192/bjp.186.3.239

Burgess JF, Christiansen CL, Michalak SE, Morris CN (2000) Medical profiling: improving standards and risk adjustment using hierarchical models. J Health Econ 19:291–309. https://doi.org/10.1016/S0167-6296(99)00034-X

de Beurs E (2008) Brief symptom inventory handleiding. PITS B.V, Leiden

de Beurs E (2010) De genormaliseerde T-score, Een "euro" voor testuitslagen. MGv 65:685–696

de Jong K, Nugter MA, Polak MG, Wagenborg JEA, Spinhoven P, Heiser WA (2007) The outcome questionnaire (OQ-45) in a Dutch population: a cross-cultural validation. Clin Psychol Psychother 14:288–301. https://doi.org/10.1002/cpp.529

Derogatis LR (1975a) The symptom checklist-90-R. Clinical Psychometric Research, Inc, Baltimore

Derogatis LR (1975b) The brief symptom inventory. Clinical Psychometric Research, Inc, Baltimore

Goldstein H, Healy MJR (1995) The graphical presentation of a collection of means. J R Stat Soc Ser A Stat Soc 158:175–177. https://doi.org/10.2307/2983411

Goldstein H, Spiegelhalter DJ (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc Ser A Stat Soc 159:385–443. https://doi.org/10.2307/2983325

Hall RC (1995) Global assessment of functioning: a modified scale. Psychosomatics 36:267–275. https://doi.org/10.1016/S0033-3182(95)71666-8

Hermann RC, Rollins CK, Chan JA (2007) Risk-adjusting outcomes of mental health and substance-related care: a review of the literature. Harv Rev Psychiatry 15:52–69. https://doi.org/10.1080/10673220701307596

Iezzoni LE (2013) Risk adjustment for measuring healthcare outcomes, 4rd edn. Health Administration, Chicago

Israels A (2013) Methods of standardisation. Statistics Netherlands: The Hague/Heerlen. https://www.cbs.nl/nr/rdonlyres/0579a307-01ef-43ca-8e29-c45b46334903/0/2013x3702.pdf

Lange A, Appelo M (2007) Korte Klachtenlijst, handleiding. Bohn Stafleu van Loghum, Houten

Lambert MJ, Gregersen AT, Burlingame GM (2004) The outcome questionnaire 45. In: Maruish M (ed) The use of psychological testing for treatment planning and outcomes assessment: instruments for adults. Lawrence Erlbaum Associates, Mahwah, pp 191–234

Lovaglio PG (2012) Benchmarking strategies for measuring the quality of healthcare: problems and prospects. Sci World J 2012:1–13. https://doi.org/10.1100/2012/606154

Lovibond SH, Lovibond PF (1995) Manual for the depression anxiety stress scales Second edition. Psychology Foundation, Sydney

Moran V, Jacobs R (2015) Comparing the performance of English mental health providers in achieving patient outcomes. Soc Sci Med 140:127–135. https://doi.org/10.1016/j.socscimed.2015.07.009

Nuttall D, Parkin D, Devlin N (2013) Inter-provider comparison of patient-reported outcomes: developing an adjustment to account for differences in patient case mix. Health Econ 24:41–54. https://doi.org/10.1002/hec.2999

Rosen AK, Chatterjee S, Glickman ME, Spiro A, Seal P, Eisen SV (2009) Improving risk adjustment of self-reported mental health outcomes. J Behav Health Ser Res 37:292–306. https://doi.org/10.1007/s11414-009-9196-9

Snijders TAB, Bosker RJ (1999) Multilevel analysis: an introduction to basic and advanced multilevel modeling. Sage Publishers, London

Warmerdam L, Barendregt M, de Beurs E (2017) Risk adjustment of self-reported clinical outcomes in Dutch mental health care. J Public Health 25:311–319. https://doi.org/10.1007/s10389-017-0785-4