Evaluation and Program Planning 89 (2021) 102004

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Inter-rater agreement of scores to assess quality of care in public sector primary health care facilities – A pattern of performance

Ronel Steinhöbel, Jacqueline E. Wolvaardt *, Elizabeth M. Webb

School of Health Systems and Public Health, Faculty of Health Sciences, University of Pretoria, 31 Bophelo Road, Gezina, Pretoria, 0001, South Africa

ARTICLE INFO

Keywords:
Quality
Evaluation
Rating
Primary health care

ABSTRACT

Purpose: To determine if the scores obtained from the Ideal Clinic Assessment Tool (ICAT) used to assess the quality of care in public Primary Health Care facilities in South Africa showed inter-rater agreement between self-assessments, district peer reviews and cross-district peer reviews. The ICAT scores obtained in the three types of reviews were paired as follows: self-assessments/district peer reviews, self-assessment/cross-district peer reviews and district/cross-district peer reviews. The global scores and averages of the Vital elements for the three paired reviews for 587 facilities across the country were compared using Bland-Altman plots.
Results: The Bland-Altman plots showed no inter-rater agreement between the global scores and averages of the Vital elements for the facilities in any of the paired reviews (n = 1 761 reviews). Similarly, there was no inter-rater agreement between the global scores of the three paired reviews in any of the nine provinces in the country.
Conclusion: There is still a need to continue to conduct both district and cross-district reviews despite the substantial cost of doing so. Further studies are required to determine what factors contributed to the disagreement in scores between the different types of reviews despite the preparatory training of reviewers.

1. Introduction

Tools to assess the quality of care in health facilities play an important role in continuous quality improvement (Whittaker, Shaw, Spieker, & Linegar, 2011). A recent study conducted in 137 countries found that an estimated 8.6 million people died due to lack of access to health care services and provision of poor quality of care. Of these deaths, 5 million deaths occurred due to poor quality of health care. The concept of Universal Health Coverage, that promotes access to care, cannot succeed without also providing quality health care (Kruk et al., 2018). By improving the quality of care, patient and staff satisfaction can be improved as well as the delivery of effective and efficient health care (Matsoso, Hunter, & Brijlal, 2018).

Following the global trend to improve quality, South Africa amended the National Health Act in 2013 to make provision for the establishment of the Office of Health Standards Compliance (OHSC) (Republic of South Africa. National Health Act 61 of 2003). The National Department of Health (NDoH) initiated the Ideal Clinic program in 2013 to ensure that public Primary Health Care (PHC) facilities obtain certification status by the OHSC. The program developed the Ideal Clinic Assessment Tool (ICAT) that sets out the standards required for public sector PHC facilities to provide good-quality health services (Hunter et al., 2017; Matsoso et al., 2018; Steinhobel, 2016).

The ICAT consists of 207 statements divided into 10 components and 32 sub-components. The vast majority of these statements (95%) require a ‘yes/no’ type of response e.g. ‘Facility has a functional piped water supply’. Each sub-component contains a number of elements and checklists that contain a set of measures that further defines the specific elements. Each element is assigned a specific weight category, i.e. “Vital”, “Essential” and “Important”. A score of “1” for achieved and “0” for failed is assigned to each element. In order for a facility to obtain an “Ideal Clinic” status, the facility must score a minimum of 90 % for elements weighted as “Vital”, 70 % for elements weighted as “Essential”, and 66 % for elements weighted as “Important”. If the facility has obtained the minimum score in each of the weight categories it is classified in one of the three Ideal categories, i.e. “Silver” (70–79 %), “Gold” (80–89 %) or “Platinum” (90–100 %) category (Steinhobel, 2016).

Assessing quality by making use of quality assessment tools can be conducted by staff within the organisation via self-assessments and/or peer reviews. Self-assessments are often the starting point of most assessments as it serves as preparation for peer or external reviews (Davis, 2002; Shaw, 2000). Peer reviews are conducted in addition to

* Corresponding author.
E-mail address: liz.wolvaardt@up.ac.za (J.E. Wolvaardt).

https://doi.org/10.1016/j.evalprogplan.2021.102004
Received 21 October 2019; Received in revised form 19 July 2021; Accepted 27 August 2021
Available online 17 September 2021

self-assessments as peer review is a valued, objective assessment which is effective in improving quality (Bose, Oliveras, & Edson, 2001; Evans, 2007; Grol, 1994; Maas et al., 2017). Peer review is a review that is done by a colleague within the same organisation or within local/regional organisations (Grol, 1994). One of the major concerns of peer reviews is that it adds to the administrative burden, is costly and if it is not conducted regularly, improvement might not be sustained (Bose et al., 2001; Maas et al., 2017). There is some debate regarding which type of assessment method is best to improve quality of care on the one hand while using resources effectively on the other. The literature is not conclusive in this regard (Scott, 2009), and few studies have explored the variance in scores between self-assessment and peer reviews.

Self-assessments and two types of peer reviews are conducted on South Africa's Ideal Clinic program's ICAT (download available from https://www.idealhealthfacility.org.za/) Self-assessments are conducted by the facility manager and peer reviews consist of district and cross-district peer reviews that are conducted annually by scale-up teams also referred to as the Perfect Permanent Team for Ideal Clinic Realisation and Maintenance (PPTICRM). Scale-up teams consist of staff of the district office and health facilities. Cross-district peer reviews are conducted in neighbouring districts in the same province in a subset of the 3464 public PHC facilities over a two-week period (Steinhobel, 2016). The subset of approximately 600 facilities is selected by the provinces at the start of every financial year based on the probability that the facility will achieve an Ideal status in that year. The self-assessments are conducted in the first quarter of the financial year, followed by the district peer reviews in the second quarter and the cross-district peer reviews in the third quarter.

The objective of this study was to determine whether there was inter-rater agreement of the scores obtained in self-assessments, district and cross-district peer reviews using the ICAT in order to determine whether there is a need to conduct both a district and cross-district peer review in addition to the self-assessment. Conducting peer reviews is costly in terms of the time that staff spends to conduct peer reviews as well as the cost of travel and accommodation during the cross-district peer reviews. By conducting three assessments per facility one also runs the risk of 'over' assessing facilities and not allowing the facilities enough time to implement quality improvement initiatives to correct the previously identified gaps.

2. Methods

An analytical, cross-sectional study was conducted in 2017 at 587 public PHC facilities in South Africa. The study assessed the global and average scores of the Vital elements of three types of reviews in public PHC facilities, i.e. self-assessments, district and cross-district peer reviews.

A hard copy of the ICAT was printed and scores were recorded on the forms during the reviews. The district teams for each province (which includes staff from different districts and facilities) are trained by their respective national coordinator during a one-day provincial workshop. The district teams are responsible for the training of staff at facility level. The Ideal Clinic Manual (download available from https://www.idealhealthfacility.org.za/) is used to train staff. The Manual is a step-by-step guide to achieve each of the elements and is a guide for reviewers with specific notes to assess the elements. The PPTICRM were advised to meet with the facility manager after the assessment to discuss and verify the results. The assessment scores were captured on a web-based application. The global score for the facilities is calculated by adding up all the scores assigned for all the elements and dividing it by the number of elements on the ICAT. The average score for the Vital elements is calculated by adding up only the scores for the Vital elements and dividing it by the number of elements that were weighted as Vital.

The data for the study were retrieved from the Ideal Clinic website. Inter-rater agreement was determined for the global score per facility and the average score obtained for the ten elements weighted as Vital. This decision was made as a facility must at least score 90 % for these elements in order to be classified as an Ideal Clinic. If the results indicated that there was a low inter-rater agreement between the facility global scores for the three types of reviews for the country, the inter-rater agreement for the facility global scores was also calculated per province.

Data was analysed in Stata version 15 (StataCorp. 2017. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC). Inter-rater agreement between the global scores and the average scores of the Vital elements were determined using Bland-Altman analyses. The Bland–Altman plot is a method for comparing two measurements of the same variable, where the X-axis is the *mean* of the two measurements and the Y-axis is the *difference* between the two measurements. Any anomalies can then be seen in the resultant chart. For example, if one method always gives too high a result, then all points will be above or below the zero line. The Bland-Altman plot can also show when one method overestimates high values and underestimates low values. If the points on the Bland–Altman plot are scattered above and below zero, then it is likely that there is no consistent bias of one approach versus the other (Kalra, 2017).

The scores were plotted and analysed for all the possible combinations (pairs) of the three different types of reviews, i.e. i) self-assessments and district peer review; ii) self-assessment and cross-district peer review; and iii) district and cross-district peer review. The differences between the scores of each paired review were plotted against the mean of the three scores. Percentile ranges for the difference in scores, the maximum and minimum differences in scores from one review to the next were determined and the percentage of outlier scores was calculated. For this study, an increase or decrease of five percent in the facility score was seen as an acceptable variance between the global and averages of the Vital element scores. Therefore, the percentages of facilities that showed an increase or a decrease of five per cent or more were calculated to determine the percentage of facilities that showed either improvement or deterioration from one type of review to the next.

Ethics approval was obtained from the Faculty of Health Sciences Research Ethics Committee at the University of Pretoria (Number 117/2018). A data user agreement was signed with the NDoH for permission to use the data.

3. Results

The difference between the scores of the three pairs of reviews was normally distributed for both the global and average scores of the Vital elements. Bland-Altman plots showed no inter-rater agreement for the global scores for any of the pairs of the different types of reviews (Figs. 1–3). The self-assessment/district scores have a slightly narrower reference range than the other two pairs of reviews. For the 587 self-

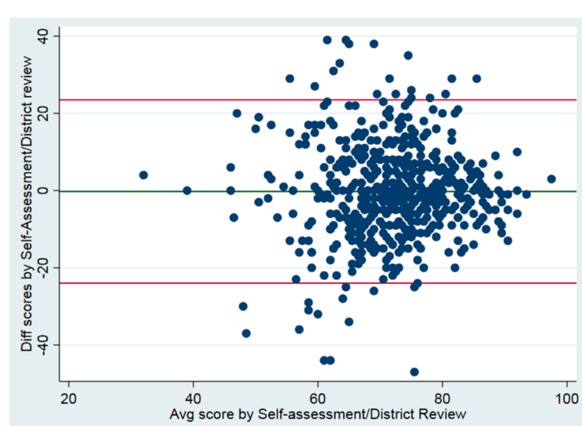

Fig. 1. Bland-Altman plot for global scores for self-assessments and district peer reviews.

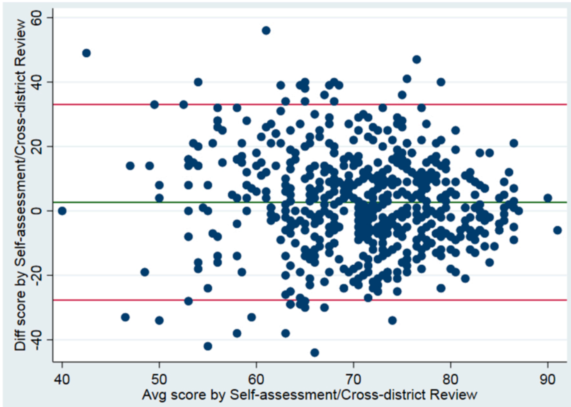

Fig. 2. Bland-Altman plot for global scores for self-assessment and cross-district peer reviews.

Fig. 3. Bland-Altman plot for global scores for district peer and cross-district peer reviews.

assessment/district reviews and the self-assessment/cross-district review, 2.5 % of the scores (n = 15) lay above the reference range and 3 % of the scores (n = 18) lay below the reference range. More outliers lay below the reference range in the district/cross-district peer reviews compared to the other paired reviews, i.e. 5 % of the scores (n = 32) while 1.5 % of the scores (n = 9) lay above the reference range.

The statistical analysis of the global scores for the three pairs of reviews is set out in Table 1. The mean of the global scores were similar, while the mean difference of the scores ranged from −0.24 to 2.91. The standard deviation (SD) of the difference between the global scores for the three types of reviews ranged from 12.11 to 15.80 indicating a wide variation or dispersion of scores. The percentile ranges show that 75 % of the scores had a difference of 7 %, 12 % and 13 % or less for the three types of reviews. There are wide differences in the individual scores of the paired reviews which contributed to the outliers. The maximum decrease in the global scores for facilities from one review to the next ranged from 39 % to 64 %. The maximum increase in the global scores for facilities was similar with an average of 45 % for the three types of reviews.

The maximum likely difference between the three types of reviews was calculated by multiplying the z-score (1.96) with the SD of the differences of the scores. The maximum likely difference ranged from 23.74 to 30.97, with the district/cross-district peer reviews recording the most difference.

From the three types of reviews the facility global scores decreased the most from the district peer reviews to the cross-district peer reviews. The percentage of facilities that performed worse in the cross-district peer reviews compared to the district peer reviews was 44 %. The percentage of facilities that performed better in the cross-district peer reviews compared to the district peer reviews was 28 % (Table 2).

Similar to the global scores, there was no inter-rater agreement for the averages of the Vital elements for any of the pairs of the different types of reviews (Figs. 4–6). The self-assessment/cross-district and district/cross-district peer reviews had the widest reference range. For the three types of review pairs the outliers were on average the same at 2 % of the scores (n = 11) with the exception of the self-assessment/district score where 5 % percent of the scores (n = 31) lay below the reference range and 1 % of the scores (n = 5) was above the reference range.

The statistical analysis of the averages of the Vital elements for the three pairs of reviews is set out in Table 3. The means for the averages of the Vital scores for the three types of reviews were similar, while the mean of the difference of the scores ranged from -0.51 to -1.50. For the global scores, the SD for the difference in scores ranged from 11.98 to 16.14. The percentile ranges show that 75 % of the scores had a difference of less than 10 % for the self-assessment/cross-district and district/cross-district paired reviews.

The maximum decrease in the average of a Vital score for a facility from one review to the next was 50 % for all the paired reviews while the maximum increase in a score from one review to the next ranged from 50 % to 60 %. Similar to the global scores, the maximum likely difference between the three types of reviews ranged from 23.74 to 31.63, with the district/cross-district peer reviews recording the most difference.

From the three types of reviews the average scores of the Vital scores

Table 2
Percentage of facilities that increased or decreased their global scores with ≥ 5 %.

Type of review	% facilities with an increase of ≥5 % in scores	% facilities with a decrease of ≥5 % in scores
Self-assessment/ district peer review	30	28
Self-assessment/cross-district peer review	29	41
District/cross-district peer review	28	44

Table 1
Statistical analysis of the global scores for three paired reviews.

Type of review	Means of paired scores	Mean difference of the scores	SD of the difference of the scores	1.96* SD of the difference of the scores	25 % Percentile of the difference of the score	75 % Percentile of the difference of the score	Maximum % increase in a facility score	Maximum % decrease in a facility score
Self-assessment/ cross-district peer review	71	2.67	15.49	30.36	−7 %	12 %	44	56
District/cross-district peer review	71	2.91	15.80	30.97	−6 %	13 %	46	64

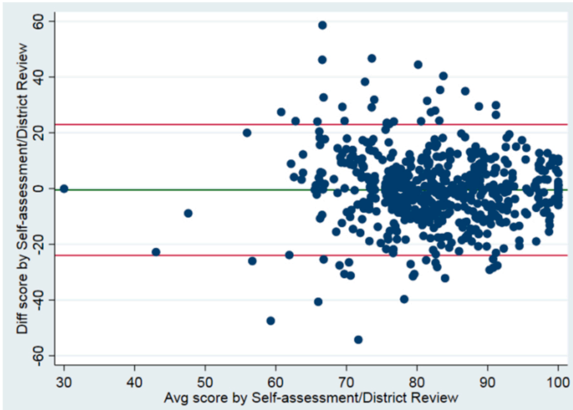

Fig. 4. Bland-Altman plot for averages of Vital for self-assessment and district peer reviews.

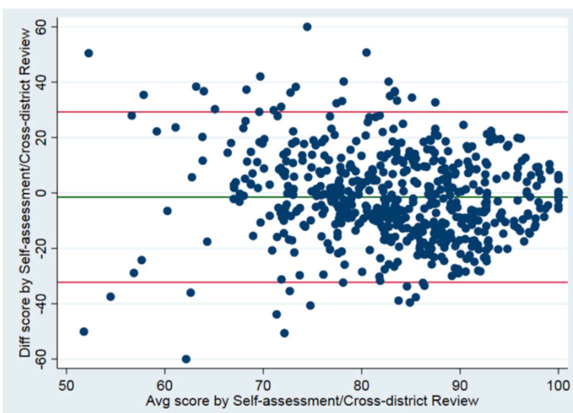

Fig. 5. Bland-Altman plot for averages of Vital elements for self-assessment and cross-district peer reviews.

Fig. 6. Bland-Altman plot for averages of Vital elements for district peer and cross-district peer reviews.

decreased the most from the district peer reviews to the cross-district peer reviews. The percentage of facilities that performed worse in the cross-district peer reviews than in the district peer reviews was 34 %. The percentage of facilities that performed better in the cross-district peer reviews than in the district peer reviews was 37 % (Table 4).

The Bland-Altman plots for the global scores of the country did not show inter-rater agreement. Therefore, further analyses were conducted to determine whether there was inter-rater agreement per province. The results of the Bland-Altman plots showed no agreement of the global scores of the three types of reviews in any of the nine provinces. There was a wide variation or dispersion of scores amongst the different provinces in the three types of reviews. For the global scores the SD for the self-assessment/district peer reviews ranged from 7.06 to 14.63, the self-assessment/cross-district peer reviews from 9.91 to 18.64 and the district/cross-district peer reviews from 9.39 to 21.00 across all nine provinces. The percentage of facilities that performed worse, according to the global scores per province from one review to the next, was the highest among the district/cross-district peer reviews. One province had 71 % of their facilities performing worse in the cross-district peer reviews than in the district peer reviews. Only three provinces had more facilities (70 %, 39 % and 33 % of facilities) that performed better in the cross-district peer reviews than in the district peer reviews.

4. Discussion

The study set out to determine whether there was inter-rater agreement of the scores obtained in self-assessments, district and cross-district peer reviews using the ICAT. There are three types of consistency (reliability): test-retest reliability (over time), internal consistency (across items) and inter-rater reliability (across researchers) (Price, Jhangiani, Chiang, 2015). In this study, the Bland-Altman plots did not show inter-rater agreement for any of the three types of reviews in the global or average of the Vital scores for the 587 PHC facilities that were analysed. The inter-rater agreement per province on the global scores for each province did not show inter-rate agreement either.

4.1. Comparison of the global and vital scores for the three types of reviews

The same assessment tool is used for all the three types of reviews but the reviews are conducted by different staff members at different times throughout the year which could explain the non-agreement between the paired reviews. A period of approximately three months lapses before the next type of review is conducted to allow facilities to implement their quality improvement plans. The absence of agreement could be due to the improvement that takes place from one review to the next. The global scores and the average scores of the Vital elements should then improve with every consecutive review. However, this was not the case as a substantial percentage of the facilities performed worse in the district (28 %) and cross-district peer reviews (41 %) than in the self-assessment. Similarly, 44 % of the facilities performed worse in the cross-district peer reviews than in the district peer reviews. The averages of the Vital scores and the global scores per province of the three types of reviews showed a similar declining trend that was observed in the country's(Price et al., 2015) global scores. There was a wide variation or dispersion of the global scores as the SD of the difference of the scores for the three types of reviews ranged from 12.11 to 15.80. The SD of the difference of the Vital scores for the three types of reviews showed a similarly wide variation (11.98–16.14). Outlier scores ranged from 1 % (n = 5) to 5 % (n = 31) for the global and averages of the Vital scores in the three types of reviews.

4.2. Factors that could have contributed to the variation in scores between the three types of reviews

The reason for the decline in the global and average of the Vital scores, the wide variances in scores and the outlier scores in all three types of reviews for the country and provinces is not clear but is possibly due to a combination of factors. Factors that could have contributed to the lack of agreement, wide variances in scores and outliers are the different types of reviews used, sustainability of quality assurance measures, standardisation of the peer review processes and validity of the ICAT (Davis, 2002; Grol, 1994; Maas et al., 2017; Scott, 2009; Wiltsey Stirman et al., 2012).

Table 3
Statistical data for the average scores of Vital elements for three pairs of reviews.

Type of review	Means of paired scores	Mean difference of the scores	SD of the difference of the scores	1.96* SD of the difference of the scores	25 % Percentile of the difference of the score	75 % Percentile of the difference of the score	Maximum % increase in a facility score	Maximum % decrease in a facility score
Self-assessment/ district peer review	83	−0.51	11.98	23.47	−10 %	0 %	50	50
Self-assessment/ cross-district peer review	83	−1.50	15.70	30.78	−10 %	10 %	60	50
District/cross-district peer review	84	−0.99	16.14	31.63	−10 %	10 %	50	50

Table 4
Percentage of facilities that increased or decreased the average scores of the Vital elements with ≥ 5 %.

Type of review	% facilities with an increase of ≥5 % in scores	% facilities with an decrease of ≥5 % in scores
Self-assessment/ district peer review	29	24
Self-assessment/cross-district peer review	39	30
District/cross-district peer review	37	34

The decline in scores from the self-assessment reviews to the district and cross-district is consistent with studies that found that self-assessment alone is not reliable or accurate as there is a general tendency for people to over-assess their own performance (Evans, 2007). One should keep in mind that self-assessments primarily allow staff to objectively appraise their work, identify learning needs, evaluate and improve performance (Davis, 2002). Although the findings of this study are consistent with others in this regard, the decline still cannot be wholly attributed to over-assessment of own performance as the overwhelming majority of items in the ICAT do not need interpretation or personal judgement.

One explanation for the decline in scores could be that the quality in some facilities could have deteriorated in the period of the reviews if quality assurance measures were not sustained. A literature review on the sustainability of quality improvement projects found that quality improvement projects are often only partially sustained even when the project was fully implemented (Wiltsey Stirman et al., 2012).

The accuracy of the peer review process further depends on how well the reviewers are prepared, organised, briefed on the purpose and taught the skills on how to conduct peer reviews (Grol, 1994). Peer reviewers must be competent and able to effectively communicate throughout the process (Bose et al., 2001; Davis, 2002; Maas et al., 2017). The outcome of the reviews are further influenced by the assessment tools that must contain explicit criteria on what is to be measured in order to avoid measurement bias (Davis, 2002; Maas et al., 2017).

4.3. Conclusion

Based on the lack of inter-rater agreement in the global scores and averages of the Vital elements of the facilities the study concludes that there is still a need to decide whether this programme of triple rating is advisable. It the decision is made to continue,the current system of triple reviews is needed to identify the factors that contributed to the lack of inter rater agreement, despite the substantial cost and human resource investment.

The results of this study merit further research to determine which factors contributed to the lack of agreement, the wide variance in the scores (reliability) and outlier scores in the district and cross-district peer review scores. Further studies will assist the NDoH and Provincial Departments of Health to identify these factors and thus the areas that require improvement to increase the quality of care and the ways in which we measure the outcomes of that care.

One possible area for further study is to verify whether the assessment tool contributed to the low inter-rater agreement. We suggest that the internal consistency of the assessment tool across items be studied through a split-half correlation method, which will result in a Pearson's correlation coefficient to be interpretedA second potential area for future research could be to identify potential gaps in how the training of reviewers is done. This strategy is supported by McLeod (2007) who advises that when the observer scores do not significantly correlate, reliability can be improved by observers in the observation techniques being used and making sure everyone agrees. However, further studies might shed light on why agreement might never be possible, which would require a new approach to evaluation and planning.

4.4. Limitations

Limitations of the study were that secondary data was used and therefore the quality of the assessments conducted and the data could not be controlled. Another limitation was that only a subset (n = 587) of the 3, 463 public PHC facilities were included in the study. Lastly, the study analysed the results for one year only.

Author statement

Ronel Steinhobel: Conceptualization, methodology, data curation, software, original draft preparation, review and approval of final manuscript.

Elize M. Webb: Supervision, reviewing, writing and editing.

Jacqueline E. Wolvaardt: Supervision, reviewing, writing and editing.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The first author (RS) is responsible for coordinating the development of the ICAT as well as the development of the web-based information system where the assessments are captured. The other authors have no competing interests to declare.

Acknowledgements

Our sincere gratitude to the NDoH of South Africa for allowing us to use the data collected on the ICAT in public Primary Health Care facilities. We appreciate the dedicated work performed by the provincial department of health managers, district managers and Primary Health Care facility managers who are responsible for implementation of the ICAT and the capturing of the results on the web-based application that was used as data source.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.evalprogplan.2021.102004.

References

Bose, S., Oliveras, E., & Edson, W. N. (2001). *How can self-assessment improve the quality of healthcare?*. Available from: Maryland: University Research Co. https://www.urc-chs.com/sites/default/files/HowCanSelf-assessmentImproveQualityofHealthcare_Sept2001.pdf.

Davis, J. D. (2002). Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. *Obstetrics & Gynecology, 99*(4), 647–651.

Evans, A. W. (2007). Reliability of peer and self-assessment scores compared with trainers' scores following third molar surgery. *Medical Education, 41*(9), 866–872.

Grol, R. (1994). Quality improvement by peer review in primary care: A practical guide. *Quality in Health Care, 3*(3), 147–152.

Hunter, J. R., Chandran, T. M., Asmall, S., Tucker, J., Ravhengani, M., & Mokgalagadi, Y. (2017). The Ideal Clinic in South Africa: Progress and challenges in implementation. In A. Padarath, P. Barron, A. Gray, & Y. Vawda (Eds.), *South African Health Review 2017* (pp. 111–121). Durban: Health Systems Trust. Available from: http://www.hst.org.za/publications/South%20African%20Health%20Reviews/HST%20SAHR%202017%20Web%20Version.pdf.

Kalra, A. (2017). Decoding the Bland–Altman plot: Basic review. *Journal of the Practice of Cardiovascular Sciences, 3*, 36–38 [serial online] [cited 2021 Jul 7] Available from: https://www.j-pcs.org/text.asp?2017/3/1/36/210855.

Kruk, M. E., Gage, A. D., Joseph, N. T., Danaei, G., García-Saisó, S., & Salomon, J. A. (2018). Mortality due to low-quality health systems in the Universal Health coverage era: A systematic analysis of amenable deaths in 137 countries. *The Lancet, 392*, 2203–2212. https://doi.org/10.1016/S0140-6736(18)31668-4

Maas, M. J., Nijhuis-van der Sanden, M. W., Driehuis, F., Heerkens, Y. F., van der Vleuten, C. P., & van der Wees, P. J. (2017). Feasibility of peer assessment and clinical audit to self-regulate the quality of physiotherapy services: A mixed methods study. *BMJ Open*. https://doi.org/10.1136/bmjopen-2016-013726

Matsoso, M. P., Hunter, J. R., & Brijlal, V. (2018). Embedding quality at the core of Universal Health Coverage in South Africa. *The Lancet Global Health*. https://doi.org/10.1016/S2214-109X(18)30323-1

McLeod, S. A. (2007). What is reliability?. *Simply psychology*. https://www.simplypsychology.org/reliability.html.

Price, P. C., Jhangiani, R., & Chiang, I. A. (2015). Psychological Measurement. *Research Methods in Psychology* (pp. 96–99). Victoria, B.C: BCcampus.

Republic of South Africa. (2003). *National Health Act 61 of 2003*. Available from: Pretoria: Government Gazette https://www.gov.za/sites/default/files/a61-03.pdf.

Scott, I. (2009). What are the most effective strategies for improving quality and safety of health care? *Internal Medicine Journal, 39*(6), 389–400.

Shaw, C. D. (2000). External quality mechanisms for health care: Summary of the ExPeRT project on visitatie, accreditation, EQFM and ISO assessment in European Union countries. *International Journal for Quality in Health Care, 12*(3), 169–175.

Steinhobel, R. (2016). PHC management. In N. Massyn, N. Peer, R. English, A. Padarath, P. Barron, & C. Day (Eds.), *District health barometer 2015/2016* (pp. 25–28). Durban: Health Systems Trust. Available from: http://www.hst.org.za/publications/District%20Health%20Barometers/District%20Health%20Barometer%202015_16.pdf.

Whittaker, S., Shaw, C., Spieker, N., & Linegar, A. (2011). Quality standards for healthcare establishments in South Africa. In A. Padarath, & R. English (Eds.), *South African health review 2011* (pp. 59–67). Durban: Health Systems Trust. Available from: www.hst.org.za/publications/South%20African%20Health%20Reviews/sahr_2011.pdf.

Wiltsey Stirman, S., Kimberly, J., Cook, N., Calloway, A., Castro, F., & Charns, M. (2012). The sustainability of new programs and innovations: A review of the empirical literature and recommendations for future research. *Implementation Science*. Available from: https://www-ncbi-nlm-nih-gov.uplib.idm.oclc.org/pmc/articles/PMC3317864/pdf/1748-5908-7-17.pdf.

Ronel Steinhobel received her bachelor's degree in Dietetics and a Master of Public health at the University of Pretoria. She worked as a dietician at public hospitals for 12 years, before a career change to work on quality in health systems. She has worked in the field of quality for 11 years and is currently employed as a deputy director at the National Department of Health in the Directorate: Quality Assurance.

Jacqueline Elizabeth Wolvaardt is an associate professor at the School of Health Systems and Public Health at the University of Pretoria. She received her bachelor's degree in nursing, her Master of Public Health and her doctorate from the same university. Her research interests are health system strengthening, action research and the undergraduate medical programme.

Elizabeth Melanie Webb is a senior lecturer at the School of Health Systems and Public Health at the University of Pretoria. She received her bachelor's degree in Genetics and Biochemistry, her honours degree in Genetics and her Master of Public Health degree from the same university. Dr Webb completed a PhD in Epidemiology, focussing on the quality of care and screening for diabetic complications at primary health care level in Tshwane, Gauteng, South Africa. Her research focuses on primary health care delivery, with a specific interest in non-communicable diseases.