BMJ Open

Improving the quality of administration of the Surgical Safety Checklist: a mixed methods study in New Zealand hospitals

Jennifer M Weller,[1,2] Tanisha Jowsey,[1] Carmen Skilton,[1] Derryn A Gargiulo,[3,4] Oleg N Medvedev,[1] Ian Civil,[5,6] Jacqueline A Hannam, Simon J Mitchell,[2,3] Jane Torrie,[2,3] Alan F Merry[2,3]

ABSTRACT

While the WHO Surgical Safety Checklist (the Checklist) can improve patient outcomes, variable administration can erode benefits. We sought to understand and improve how operating room (OR) staff use the Checklist. Our specific aims were to: determine if OR staff can discriminate between good and poor quality of Checklist administration using a validated audit tool (WHOBARS); to determine reliability and accuracy of WHOBARS self-ratings; determine the influence of demographic variables on ratings and explore OR staff attitudes to Checklist administration.

Design Mixed methods study using WHOBARS ratings of surgical cases by OR staff and two independent observers, thematic analysis of staff interviews.

Participants OR staff in three New Zealand hospitals.

Outcome measures Reliability of WHOBARS for self-audit; staff attitudes to Checklist administration.

Results Analysis of scores (243 participants, 2 observers, 59 cases) supported tool reliability, with 87% of WHOBARS score variance attributable to differences in Checklist administration between cases. Self-ratings were significantly higher than observer ratings, with some differences between professional groups but error variance from all raters was less than 10%. Key interview themes (33 interviewees) were: Team culture and embedding the Checklist, Information transfer and obstacles, Raising concerns and 'A tick-box exercise'. Interviewees felt the Checklist could promote teamwork and a safety culture, particularly enabling speaking up. Senior staff were of key importance in setting the appropriate tone.

Conclusions The WHOBARS tool could be useful for self-audit and quality improvement as OR staff can reliably discriminate between good and poor Checklist administration. OR staff self-ratings were lenient compared with external observers suggesting the value of external audit for benchmarking. Small differences between ratings from professional groups underpin the value of including all members of the team in scoring. We identified factors explaining staff perceptions of the Checklist that should inform quality improvement interventions.

INTRODUCTION

The WHO Surgical Safety Checklist ('the Checklist') is widely established in the operating room (OR) as a tool to promote

Strengths and limitations of this study

► This study builds on previous work supporting the use of the WHOBARS tool for quality improvement initiatives in administration of the WHO Surgical Safety Checklist.

► It specifically provides evidence to support to the use of the tool by OR clinicians for self-audit, which, compared to external audit, could be a pragmatic approach to quality improvement due to potential for widespread application.

► The qualitative study identified factors that influence OR staff perceptions on Checklist administration, including the key role of senior clinicians in the quality of Checklist administration, enabling staff to speak up and linking Checklist administration to patient safety.

► The extent to which self-audit of Checklist administration is feasible and can in fact lead to improved Checklist administration remains to be tested.

► Our findings may not be generalisable across countries due to organisational and cultural differences.

teamwork and sharing of important clinical information and to prevent crucial errors or omissions.[1] The global implementation of the Checklist has seen significant reductions in surgery-related complications and mortality.[2,3] Being able to reliably discriminate between teams who engage well or poorly with the Checklist could enable implementation of necessary improvements to the quality of Checklist administration to optimise patient safety.

Unfortunately, not everyone complies or engages with the Checklist as intended, even while acknowledging its importance.[4] Studies[5–10] suggest there is still scope for improvement both in compliance with administering the Checklist items, and in engagement of OR staff in the Checklist process. The influence of OR culture on engagement of clinical staff with the Checklist remains unclear. Two studies challenged the

BMJ

effectiveness of the Checklist,[11] [12] while another suggested that the Checklist works only if it is treated as more than a tick box exercise.[13]

Currently, the Checklist is widely used and is compulsory practice in many countries (eg, New Zealand). But how is it used? How reliable are OR staff at observing its use? And how can administration be improved to optimise patient safety?

To measure the quality of engagement during administration of the Checklist, we previously developed a novel tool, the WHO Behaviourally Anchored Rating Scale (WHOBARS), which measures behaviours associated with Checklist administration.[14] The WHOBARS has five domains for each phase of the Checklist (Sign In, Time Out and Sign Out). It uses a 7-point scale for each of five domains. The domains are: setting the stage; team engagement; Checklist activation; problem anticipation and process completion. The end-points of the scale are anchored by examples of poor and excellent behaviours. Below each domain is a space for observer comments. The five domains of the WHOBARS are described in the original paper.[14] (The full WHOBARS tool is provided in online supplementary appendix 1.)

In our previous study,[14] trained independent observers performed all WHOBARS ratings. However, if OR staff were able to use the WHOBARS tool to generate reliable ratings on their own performance, WHOBARS could be a useful self-regulated approach to audit and continuous quality improvement of Checklist administration. We were interested in exploring if this was possible. While ability to reliably discriminate between good and poor performance is helpful to measure improvement, it is also useful to know the accuracy of self-rated scores when compared with an external standard. If they are accurate as well as discriminating, then it would suggest an external audit is unnecessary, but if they are overly lenient or strict, an external benchmarking from time to time could be useful to drive further improvement. OR staff could potentially have different perspectives on the quality of Checklist administration, which could influence accuracy of the self-ratings.

Mitchell et al[15] write that Checklist 'implementation is complex, inconsistent and troublesome,' yet research often fails to acknowledge this. They call for research that can engage with such complexity. Our qualitative work aims to add to such research by seeking to gain an in-depth understanding of factors that influence staff perceptions of the quality of Checklist administration. Attitudes towards the Checklist may influence the quality of its administration,[16] and understanding staff attitudes and experiences regarding the Checklist may assist in designing interventions to improve its use.

We therefore used a mixed methods approach to compare perceptions of the quality of Checklist administration by independent observers and OR staff, and to explore OR staff attitudes towards and experiences of the Checklist. For the purposes of this study, OR staff included surgeons and anaesthetists (both specialists and trainees), nurses and anaesthetic assistants.

Our four research questions were:
1. Can OR staff reliably discriminate between teams who engage well or poorly with the Checklist?
2. Do OR staff self-ratings of the overall quality of Checklist administration (including all three phases) using the WHOBARS tool agree with ratings from trained, independent observers?
3. What is the influence of clinical role (anaesthetist, anaesthetic assistant, nurse, surgeon), hospital site, gender and years of clinical experience on these overall ratings?
4. What are the attitudes and experiences of OR staff in relation to the Checklist?

METHODS
This study forms part of a larger programme of research on WHOBARS and Checklist implementation. There were two phases to this study:
1. WHOBARS ratings of the quality of administration of the Checklist during surgical cases by OR staff and two independent observers, addressing research questions 1–3.
2. Interviews with selected OR staff who had participated in the WHOBARS self-rating exercise to explore in depth their attitudes and experiences in relation to the Checklist, addressing research question 4.

The study was approved by the University of Auckland Human Participants Ethics Committee (ref: UOA016558). Prestudy presentations and information sheets were offered to all OR staff and written consent obtained. Patients included adults in two OR suites and infants/children in one OR suite. Patients (and/or their legal guardians) were given information prior to the observations and asked for verbal consent. They could opt out if they did not want study personnel present during their (or their child's) surgery.

Patient and public involvement
There was no patient or public involvement in this study.

The study was conducted in 2016, in three separate OR suites in two hospitals in Auckland, New Zealand (referred to here as sites 1–3). In accordance with recommendations from the New Zealand Health Quality and Safety Commission Safe Surgery programme, prompts to the three phases of the Checklist are displayed on wall-mounted posters in all ORs. Responsibility for administering each phase of the Checklist is shared: Sign In is led by an anaesthetist, Time Out by a surgeon and Sign Out by a nurse.[17] All OR staff were naïve to WHOBARS and received no prior training on its use. The independent observers were trained in the use of the WHOBARS as previously described.[13] One of the independent observers (DG) is an academic pharmacist, and the other (CS) is a medical education researcher (trained in psychology). Neither independent observers have previously worked in ORs. One observer (DG) has prior experience of conducting research in New Zealand ORs and recognised a few OR staff.

Case selection and participants

We aimed to study 20 complete surgical cases at each of the three sites, based on sample sizes from previous studies.[14 17] Observations took approximately 2 weeks per site. All elective and acute cases (adults and children) involving surgery under general anaesthesia during normal working hours were eligible. After the list for the day was posted, observers started in the OR in which the whole team had consented. They then selected cases according to the numbers of staff in any OR with prior written consent. Only one case from any OR was observed per day to ensure inclusion of a range of OR teams. We excluded cases where any staff member or the patient withheld consent.

Rater training

Prior to the ratings, the two independent observers rated 12 training videos, used in our previous studies and previously rated by a group of trained raters. The intraclass correlation coefficient with the two independent observers from this study and the trained raters from the original validation study, across the 12 training clips, was 0.84.[14]

WHOBARS ratings

Each case was observed in its entirety by the two trained independent observers, each independently rating the five WHOBARS domains in each Checklist phase: Sign In, before induction of anaesthesia; Time Out, before skin incision and Sign Out, prior to the patient leaving the OR. OR staff also rated their team's performance after Sign Out, also using the WHOBARS tool. Demographic details (gender, age, clinical role and length of OR experience) were requested. Both OR staff and independent observers used the same WHOBARS rating scale, which is described in the introduction. It includes detailed instructions for rating (online Supplementary appendix 1).

Interviews

Subsequently, we invited OR staff in these observed cases to a semistructured interview (see online supplementary appendix 2 for interview guide). For this, we used a purposive sampling strategy. The consent form for OR rater participants included a box that participants could tick if they also consented to being contacted for an interview. Participants who ticked that box were then contacted via email (2–6 days after the OR ratings were completed) by a researcher (TJ), who had no prior relationship with the participants. Incentives and compensation were not offered to participants. We continued interviewing until we had participants from a mix of sites, different clinical roles and experience and until we had reached a point of data sufficiency, where little in the way of new ideas, opinions or concepts was arising from the interviews.[18 19] Interviews were conducted in-person or via phone (according to participant availability and preferences), by one researcher (TJ) to ensure consistency in interview method and comparability of interview data.

The researcher took detailed notes during each interview. The interviews were audio-recorded and transcribed verbatim by an external transcribing service.

Quantitative data analysis

Multiple linear regression analyses were used to explore the relationship between ratings by the independent observers and those by the staff, and to identify and evaluate relevant predictors of scores, and unique contributions of the different clinical roles to the WHOBARS item scores. Traditional reliability estimates such as Intra Class Correlation (ICCs) are not suitable for our study because they require the same raters to rate all the cases but OR staff members only rated their own case while independent observers rated all cases. Moreover, ICC cannot control for the demographic variables of the rater (eg, gender, age, professional group) which may be potential sources of error that can influence the estimates and accuracy of comparisons between independent observers and internal raters. Therefore, we used a multiple linear regression model that also included analysis of variance because this can control for effects of demographics and does not require similar cases across all raters. We used WHOBARS scores as the dependent variable and hospital site, Checklist phase and independent observer and clinical role as predictors, while controlling for demographic variables. The categorical predictors, where there were more than two, were placed into two categories (eg, nurse vs other) for the purpose of analysis.[20]

We entered demographic variables (gender, age and experience) in model 1, site in model 2, phase in model 3, independent observer in model 4 and clinical role (nurse, surgeon, anaesthetist and anaesthetic assistant) in model 5. Stepwise multiple regression was applied for the clinical role predictors (model 5). This extracts the strongest significant predictor and controls for its effect before extracting the next strongest predictor, until no significant predictors remain.

Using t-tests, we also compared the mean differences in the WHOBARS domain mean scores between independent observers and the clinical roles.

Qualitative data analysis

The qualitative methods were underpinned by an interpretivist paradigm.[21] Qualitative data were analysed using thematic analysis following Lincoln and Guba[18] and Morse and Field[19] and drawing on coding analysis as described by Saldana.[22] The interviewer wrote a summary report of her first impressions of the data and the key messages that emerged immediately after the interview. She identified recurrent phrases, concepts and themes, which formed the basis of the coding scheme. An independent academic service coded the data according to this coding scheme. Recurrent themes were identified. A series of matrix coding and coding text queries were run in QSR NVivo 10 qualitative software[23] to: (1) identify patterns between the themes; (2) identify whether particular themes were strongly supported by particular

Table 1 Number of ratings of the five WHOBARS domains in each of the three Checklist phases (Sign In, Time Out, Sign Out) by independent observer and professional role

	Sign In	Time Out	Sign Out
Independent observer (n=2)	585	590	580
Surgeon (n=71)	215	450	437
Anaesthetists (n=52)	363	365	325
Nurses (n=86)	543	719	650
Anaesthetic assistants (n=32)	275	283	210

There were some missing data from one or more Checklist phases in some of the WHOBARS ratings. Completed ratings for the five domains in the three Checklist phases would give a count of 15 ratings per case.

participant groups and (3) ensure that data-rich codes had been captured in the themes.

RESULTS
Quantitative results

We observed 60 cases but removed one because of incomplete data. The final dataset was from 243 participants across 59 different cases. Nineteen teams participated from site 1 and 20 teams participated from each of sites 2 and 3 (table 1). The sample included 104 (48.8%) males, 139 females (51.2%) and included 71 surgeons, 86 nurses, 52 anaesthetists, 32 anaesthetic assistants and 2 independent observers.

The data met assumptions of multiple linear regression: skewness and kurtosis within ±1, with no significant outliers and no evidence of multicollinearity (variance inflating factor<5). Table 2 shows the summary of the model for the multiple linear regression analysis with the WHOBARS ratings as dependent variable and demographics (gender, age, experience), site, phase, independent observers and clinical role as predictors. Demographic variables together accounted for 1.7% of the variance in WHOBARS ratings and most of this was explained by experience in the role (β=0.14). There was a significant effect for age but effect size was small, and no significant gender effect. Site explains 1.3% and phase 0.6% of the score variance. After accounting for demographics, site and phase, independent observers explained 9.2% of the variance in WHOBARS ratings. After controlling for the effects of all other non-role predictors, nurses' ratings accounted for 0.5% of variance, followed by 0.1% of the variance contributed by surgeons' ratings. Anaesthetists and anaesthetic assistants were excluded from the final model because their contribution to score variance was not statistically significant after accounting for the variance explained by other predictors (numbered 1–6 in table 2). Table 2 shows that all mean rating scores from the professional groups were significantly higher than those of the independent observers.

Table 2 also shows that individual domain mean scores produced by OR staff were significantly higher than the

Table 2 Multiple linear regression results for the WHOBARS ratings as dependent variable and demographic variables (gender, age and experience), site, phase, independent observers and clinical roles as predictors

Model	Predictor	R^2	R^2 change	Standardised β	P values	Mean
1	Demographics	0.017	0.017		<0.001	
	Gender			0.009	0.766	
	Age			−0.067	<0.001	
	Experience			0.141	<0.001	
2	Site	0.030	0.013	0.134	<0.001	
3	Phase	0.037	0.006	0.067	<0.001	
4	Independent observers	0.129	0.092	−0.35	<0.001	4.90*
5	Nurse	0.134	0.005	0.086	<0.001	5.94†
6	Surgeon	0.134	0.001	0.029	0.028	5.70†
6	Anaesthetic assistant		0.000	Excluded	0.720	5.63†
6	Anaesthetist		0.000	Excluded	0.720	5.54†
	All clinical roles		0.006		<0.001	5.75†

Mean domain ratings by independent observer and clinical role are shown in the right hand column.

*Independent Observer Mean score.

†Mean is significantly different compared with the Independent Observers' mean score for the matched cases. Mean score is Bonferroni adjusted (p<0.001).

Note: Column 1: sequential steps for models in regression analysis. Column 2: predictors of the WHOBARS scores. Column 3: R^2=multiple correlation coefficient squared reflecting sum of variances explained by current and all previous models (steps) above. Column 4: R^2 change, reflecting variance uniquely explained by the current model. Column 5: Standardised β=beta coefficient reflects how change of the predictor variable (demographics and so on) influences the outcome for the dependent variable in SD units (where there are multiple predictors, eg, Site and Phase, the largest β values are included). Column 6: p<0.05 indicates statistically significant effect. Column 7: mean rating scores.

domain mean scores of the independent observers for the same cases. However, after accounting for influence of demographics, site, phase and independent observers, inconsistency between the ratings of OR staff and those of independent observers was mainly associated with nurses and surgeons (significant predictors, explaining less than 1% of WHOBARS scoring variance). Ratings of anaesthetists and anaesthetic assistants were closer to those of independent observers than were nurses and surgeons. Even though the influence of nurse and surgeon on WHOBARS score variance was statistically significant, it explained less than 1% of variance in scoring, and the overall influence of all the predictors accounted for merely 13% of variance in scores. While there may be other factors that we have not considered in this data set that could influence WHOBARS scores, this does suggest that 87% of the variance in WHOBARS scores is in fact due to true differences in the quality of administration of the Checklist between cases, which is our primary focus of measurement.

Qualitative results

We interviewed 33 OR staff: 9 anaesthetists, 10 surgeons, 9 nurses and 5 anaesthetic assistants. Twelve were from Site 1, 13 from Site two and 8 from Site 3. OR experience ranged from 1 month to over 20 years. Interviews were 8–36 min long (average 18.9 min).

In general, participants viewed the Checklist as an excellent and increasingly embedded safety tool. Participants felt the Checklist effectively promoted interdisciplinary communication, a culture supportive of teamwork and patient safety. Dissatisfaction was usually due to poor Checklist administration rather than dissatisfaction with the Checklist itself.

Four key themes emerged: (1) Team culture and embedding the Checklist, (2) Obstacles to information transfer, (3) Raising concerns and (4) A 'tick-box' exercise. Quotes that evidence these themes are presented in online supplementary table 1.

TEAM CULTURE AND EMBEDDING THE CHECKLIST

While there was variation in the way the Checklist was used within and between teams, participants found it useful for breaking down hierarchies and enabling junior or new staff to speak up and feel valued. Almost all participants felt the Checklist improved communication between professions within the OR team, which influenced the overall culture. Participants said they liked having an allocated role in the Checklist because it made them '*feel acknowledged*' and '*important*' in the team. Thirty participants made comments about the Checklist facilitating cultural change, usually discussed in terms of '*embedding*' the Checklist and their expectations of it being used routinely (see online supplementary table 1, theme 1a).

Senior leaders

Participants viewed specialist surgeons and specialist anaesthetists as the most influential people for determining the value of the Checklist. If a specialist advocated for the Checklist, then the atmosphere in the OR was generally collegial. When specialists did not participate in the Checklist effectively, this affected the way that other staff engaged with the Checklist and with one another. Several specialists reflected on their influence on the atmosphere in the OR, suggesting that if they were assertive about following the Checklist correctly, particularly in terms of introductions and welcoming staff to assert concerns, their team functioned more effectively. In contrast, several participants from one site spoke of a specialist who was strongly resistant to using the Checklist, making it very difficult for other staff to engage with it effectively (see online supplementary table 1, theme 1b).

OBSTACLES TO INFORMATION TRANSFER

The structure and general processes of the OR at the different sites influenced where Sign In occurred. Participants reported that conducting the Sign In outside of the OR, with only some members of the surgical team involved, reduced the opportunities for information to be effectively shared between all OR staff. Participants had different ideas about what their role was and how engaged in each part of the Checklist they needed to be. Several nurse participants reported issues around effective communication when the person delivering the Checklist spoke too quietly or other staff did not stop, but carried on their own conversations or continued to work with noisy equipment.

Introducing people's names and positions in the OR was deemed a very important part of the Checklist by all participants, but especially so for new and junior staff. Despite participants' enthusiasm for introductions, participants frequently reported them as '*missing*' or '*rarely done*' (see online supplementary table 1, theme 2b).

RAISING CONCERNS

At all sites, participating anaesthetists frequently referred to the Checklist as a valued safety net, discussed in terms of minimising risk, identifying problems, preventing errors or omissions or identifying better ways of doing things. They appreciated having more than one person contributing to patient care.

Several senior nurses reported confidence in promoting effective use of the Checklist even when others made it difficult. They suggested that the Checklist created a platform to enable them to speak. Junior staff reported that when senior staff '*don't invite concerns or shut you down*', it can be nerve-racking to be assertive or find ways to discuss the issue. In such cases, junior staff would usually speak to a senior within their discipline to convey their concerns, and in some cases, they would write up an incident report.

Participants said that it can take a long time for staff to feel confident in raising their concerns. Certain staff members, they said, had a significant effect on the extent to which the Checklist created a platform for OR staff to raise any concerns. When a senior staff member was grumpy or appeared not to value the Checklist, it discouraged participants from raising concerns even if they had some. For example, a nurse with concerns about sterility did not have the courage to say anything. Nurses and anaesthetists discussed these difficulties more frequently than other participants, and comments were usually about not feeling confident to raise their concerns to the surgeon. In such cases, some participants said they used humour to ensure adherence to the Checklist.

All surgeon participants said that they felt comfortable raising any issues of concern they had with other OR staff (see online supplementary table 1, theme 3).

A 'tick-box' exercise

While participants often expressed frustration at inconsistent Checklist use, consistency could be a double-edged sword, on the one hand embedding a culture of safety but potentially also creating a monotonous habit, resulting in lack of focus or meaningful engagement. Some staff reportedly saw it as a '*tick-box*' exercise and did not genuinely engage with the Checklist or the team (see online supplementary table 1, theme 4).

DISCUSSION

Our results suggest that OR staff can reliably discriminate between teams who engage well or poorly with the Checklist, as the vast majority of variance in WHOBARS scores for the quality of administration of the Checklist was attributable to differences between cases. This suggests that WHOBARS scores from OR staff would potentially be useful for audit and quality improvement on Checklist administration. OR staff consistently rated their own Checklist performance more highly than the independent observers, although this effect was small. This suggests that it could be useful for recalibration from time to time by external trained observers. While there was a significant difference between OR professional roles, this contributed only 0.6% to the total score variance. However, reliability of self-ratings would be optimal if all members of the team were included in scoring. Furthermore, mean ratings scores from all groups suggested there was room for improvement in Checklist administration, underpinning the need for quality improvement interventions.

We found a number of factors that significantly influenced WHOBARS scores and together explained 13.4% of variance in scores. The main contributor was the two independent observers' scores which accounted for 9.2% of the variance. Other contributors were OR staff members' clinical role, age and years of experience and phase of the Checklist. The two independent raters came from different professional backgrounds (pharmacy and psychology), which could account for the greater

differences in their scores, compared with the relatively small differences between members of the OR team who share common OR experience. Staff gender and hospital site had no significant effect.

Lingard *et al*[24] write 'intervening to strengthen communicative practice among healthcare teams is complicated because such communication is rooted in the distinct and often conflicting professional identities of team members and is bounded by a culture that has been traditionally and persistently hierarchical'. The Checklist was developed in an attempt to mitigate such hierarchies, to facilitate effective timely communication, and by so doing, promote patient safety. In surveys of OR staff, Singer *et al*[25] identified a relationship between communicative practices and Checklist performance. Our qualitative findings help us to understand the factors influencing participants' perceptions of Checklist administration. Overall, participants believed that the Checklist was embedded in their hospital culture, administered reasonably well and it positively influenced their sense of belonging to a team.

However, this could be compromised by a single team member—a senior clinician or a particular personality. This could have an important effect on how the team functioned for the day. Some organisational issues impeded the potential value of the Checklist by limiting the extent to which information was shared between the whole team. Many participants associated the Checklist with speaking up, and there appeared to be different perspectives on this between the clinical roles, which may, to a certain extent, explain some of the differences in the WHOBARS ratings. Nurse participants in particular felt the Checklist provided them with an opportunity to raise their concerns, which would otherwise have been difficult. Comments from participating surgeons on the other hand suggested they did not have a problem speaking up with or without the Checklist. Participating anaesthetists frequently commented on the value of staff voicing their concerns to them, and the safety advantage of input from others to reduce the potential for error. Alidina *et al*[26] also report varying perspectives on the Checklist from different members of the surgical team, with more negative comments coming from clinicians than from other OR staff. Differing perspectives on and experiences of speaking up may be a key factor in understanding this variable buy-in from different professional groups.

Participants stressed the need for staff to actively engage with the Checklist as in a genuine conversation to stop it becoming a meaningless 'tick-box' exercise. This is in alignment with Mitchell *et al*'s 2017 findings[15] that much of the Checklist literature itself presents the Checklist as a *tool* rather than as a *process* of cultural change. Participants in our study saw the Checklist as process, and wanted to see it effectively embedded as such. A systematic review of compliance with Checklist completion and team perceptions[27] found varying compliance and perceived value of the Checklist, with what typically appeared to be a nurse-led written Checklist administration paradigm. Concerns were expressed that the Checklist could have a negative effect on

teamwork. Our research has not uncovered a negative effect of the Checklist on teamwork; rather, it has established that embedding the Checklist into routine practice is creating a cultural shift towards all OR staff feeling supported to raise concerns for the patient that they might have and towards an increased sense of value within the OR team.

Recommendations

Our findings provide some guidance for interventions to improve Checklist implementation. The WHOBARS tool provides examples of good and poor behaviours. Self-ratings of Checklist administration using this tool, by explicitly describing these behaviours, may in itself promote better quality of Checklist administration when used by OR staff. The qualitative support for the Checklist as a positive driver for improved teamwork culture, speaking up and patient safety could convince doubtful staff to reflect on how they use the Checklist. Multidisciplinary discussions on those themes identified in our interviews could provide a focus for guided reflection and subsequent change in practice. Enabling staff to speak up appears to be a key element of the Checklist and linking speaking up and patient safety should be part of any improvement intervention. The Checklist, administered as intended, has the potential to overcome many of these barriers.

Further research should investigate barriers to speaking up, including a hierarchical culture and operating within one's own task area rather than feeling responsible for the care of the patient in its entirety.[28]

Study limitations

The study occurred in one large New Zealand city. While our national approach to Checklist implementation would suggest these findings may broadly reflect practice in New Zealand,[29] the extent to which they can be generalised to other countries is unknown. Interview selection bias is also a possibility. Despite our purposive sampling strategy, staff who did not agree to participate in interviews may have had views of the Checklist that are not represented in these findings. A common limitation of a mixed method approach is that qualitative data that features specific attitudes, experiences and behaviours of a particular individual may not necessarily reflect the overall group effect estimated on a larger sample size in the quantitative part of the study.

While we have demonstrated the reliability of the WHOBARS tool for self-audit, further research would be required to determine the feasibility and acceptability of its use for this purpose.

Staff received no training in WHOBARS. Training could have further improved reliability of scores and potentially their agreement with independent observers. However, the requirement for training could affect the feasibility of staff self-ratings using WHOBARS.

CONCLUSION

Our results provide good evidence that using WHOBARS, OR staff can reliably discriminate between teams who engage well or poorly with the Checklist, which could potentially guide reflection and quality improvement. The vast majority of variance (87%) in scores between cases were attributable to differences in the quality of administration of the Checklist between cases, indicating good reliability of the WHOBARS. OR staff self-ratings were lenient compared with external observers suggesting the value of external audit for benchmarking.

Even though nurses and surgeons rated their own teams significantly higher than independent observers did, this effect was very small and accounted for a minor proportion of variance (0.6%), which did not affect the overall good reliability of scores. However, these small differences underpin the value of including all members of the OR team in the scoring.

In-depth interviews helped us to understand the dynamics behind the OR rater scores. Our qualitative findings identified several factors explaining both positive and negative staff perceptions, including the key role of senior clinicians in the quality of Checklist administration, enabling staff to speak up and linking Checklist administration to patient safety.

Interventions to improve administration of the Checklist could potentially include guided self-reflection on clinical practice using WHOBARS as a tool for self-audit and facilitated multidisciplinary discussions on the key factors influencing the quality of checklist administration, including positive leadership, speaking up and the evidence linking these factors to patient safety.

Author affiliations
[1]Centre for Medical and Health Sciences Education, University of Auckland, Auckland, New Zealand
[2]Department of Anesthesia and Perioperative Medicine, Auckland City Hospital, Auckland, New Zealand
[3]Department of Anaesthesiology, University of Auckland, Auckland, New Zealand
[4]School of Pharmacy, University of Auckland, Auckland, New Zealand
[5]Division of Surgery, Auckland City Hospital, Auckland, New Zealand
[6]Department of Surgery, University of Auckland, Auckland, New Zealand

REFERENCES

1. World Health Organization. WHO Guidelines of Safe Surgery 2009: Safe Surgery Saves Lives. http://whqlibdoc.who.int/publications/2009/9789241598552_eng.pdf (Accessed Nov, 2015).
2. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. *N Engl J Med* 2009;360:491–9.
3. Weiser TG, Haynes AB, Dziekan G, et al. Effect of a 19-item surgical safety checklist during urgent operations in a global patient population. *Ann Surg* 2010;251:976–80.
4. Sendlhofer G, Lumenta DB, Leitgeb K, et al. The gap between individual perception and compliance: a qualitative follow-up study of the surgical safety checklist application. *PLoS One* 2016;11:e0149212.
5. Mazzocco K, Petitti DB, Fong KT, et al. Surgical team behaviors and patient outcomes. *Am J Surg* 2009;197:678–85.
6. de Vries EN, Prins HA, Crolla RM, et al. Effect of a comprehensive surgical safety system on patient outcomes. *N Engl J Med* 2010;363:1928–37.
7. Vogts N, Hannam JA, Merry AF, et al. Compliance and quality in administration of a surgical safety checklist in a tertiary New Zealand hospital. *N Z Med J* 2011;124:48.
8. van Klei WA, Hoff RG, van Aarnhem EE, et al. Effects of the introduction of the WHO "Surgical Safety Checklist" on in-hospital mortality: a cohort study. *Ann Surg* 2012;255:44–9.
9. Haugen AS, Søfteland E, Almeland SK, et al. Effect of the World Health Organization checklist on patient outcomes: a stepped wedge cluster randomized controlled trial. *Ann Surg* 2015;261:821–8.
10. Mayer EK, Sevdalis N, Rout S, et al. Surgical Checklist implementation project: the impact of variable who checklist compliance on risk-adjusted clinical outcomes after national implementation: a longitudinal study. *Ann Surg* 2016;263:58–63.
11. Reames BN, Krell RW, Campbell DA, et al. A checklist-based intervention to improve surgical outcomes in Michigan: evaluation of the Keystone surgery program. *JAMA Surg* 2015;150:208–15.
12. Reames BN, Scally CP, Thumma JR, et al. Evaluation of the effectiveness of a surgical checklist in medicare patients. *Med Care* 2015;53:87–94.
13. Urbach DR, Govindarajan A, Saskin R, et al. Introduction of surgical safety checklists in Ontario, Canada. *N Engl J Med* 2014;370:1029–38.
14. Devcich DA, Weller J, Mitchell SJ, et al. A behaviourally anchored rating scale for evaluating the use of the WHO surgical safety checklist: development and initial evaluation of the WHOBARS. *BMJ Qual Saf* 2016;25:778–86.
15. Mitchell B, Cristancho S, Nyhof BB, et al. Mobilising or standing still?A narrative review of Surgical Safety Checklist knowledge as developed in 25 highly cited papers from 2009 to 2016. *BMJ Qual Saf* 2017;26:837–44.
16. Fourcade A, Blache JL, Grenier C, et al. Barriers to staff adoption of a surgical safety checklist. *BMJ Qual Saf* 2012;21:191–7.
17. Ong AP, Devcich DA, Hannam J, et al. A 'paperless' wall-mounted surgical safety checklist with migrated leadership can improve compliance and team engagement. *BMJ Qual Saf* 2016;25:971–6.
18. Lincoln YS, Guba EG. *Naturalistic inquiry*. Beverly Hills, Calif: Sage Publications, 1985.
19. Morse JM, Field PA. *Qualitative research methods for health professionals*. 2nd edn. Thousand Oaks: Sage Publications, 1995.
20. Allen P. *SPSS statistics version 22: a practical guide*. 3rd edn. South Melbourne: VIC: Victoria Cengage Learning, 2014:179–97.
21. Thorne S, Kirkham SR, O'Flynn-Magee K. The analytic challenge in interpretive description. *Int J Qual Methods* 2004;3:1–11.
22. Saldaña J. *The coding manual for qualitative researchers*. 2nd edn. Los Angeles: Sage, 2013.
23. QSR International Pty Ltd. NVivo qualitative data analysis Software. 2012:10.
24. Lingard L, Espin S, Whyte S, et al. Communication failures in the operating room: an observational classification of recurrent types and effects. *Qual Saf Health Care* 2004;13:330–4.
25. Singer SJ, Molina G, Li Z, et al. Relationship between operating room teamwork, contextual factors, and safety checklist performance. *J Am Coll Surg* 2016;223:568–80.
26. Alidina S, Hur HC, Berry WR, et al. Narrative feedback from OR personnel about the safety of their surgical practice before and after a surgical safety checklist intervention. *Int J Qual Health Care* 2017;29:461–9.
27. Wangoo L, Ray RA, Ho Y-H. Compliance and surgical team perceptions of who surgical safety checklist; systematic review. *Int Surg* 2016;101:35–49.
28. Edmondson AC. Speaking up in the operating room: how team leaders promote learning in interdisciplinary action teams. *Journal of Management Studies* 2003;40:1419–52.
29. Health Quality & Safety Commission New Zealand. Safe surgery. 2012 https://www.hqsc.govt.nz/our-programmes/safe-surgery-nz/projects/surgical-teamwork-and-communication/interventions/surgical-safety-checklist/ (Accessed 25 Oct 2017).