BJR

## ONCOLOGY

# The inter-rater reliability of the diagnosis of surgical site infection in the context of a clinical trial

**J. Nuttall,**
**N. Evaniew,**
**P. Thornley,**
**A. Griffin,**
**B. Deheshi,**
**T. O'Shea,**
**J. Wunder,**
**P. Ferguson,**
**R. L. Randall,**
**R. Turcotte,**
**P. Schneider,**
**P. McKay,**
**M. Bhandari,**
**M. Ghert**

*McMaster University, Ontario, Canada*

## Objectives

The diagnosis of surgical site infection following endoprosthetic reconstruction for bone tumours is frequently a subjective diagnosis. Large clinical trials use blinded Central Adjudication Committees (CACs) to minimise the variability and bias associated with assessing a clinical outcome. The aim of this study was to determine the level of inter-rater and intra-rater agreement in the diagnosis of surgical site infection in the context of a clinical trial.

## Materials and Methods

The Prophylactic Antibiotic Regimens in Tumour Surgery (PARITY) trial CAC adjudicated 29 non-PARITY cases of lower extremity endoprosthetic reconstruction. The CAC members classified each case according to the Centers for Disease Control (CDC) criteria for surgical site infection (superficial, deep, or organ space). Combinatorial analysis was used to calculate the smallest CAC panel size required to maximise agreement. A final meeting was held to establish a consensus.

## Results

Full or near consensus was reached in 20 of the 29 cases. The Fleiss kappa value was calculated as 0.44 (95% confidence interval (CI) 0.35 to 0.53), or moderate agreement. The greatest statistical agreement was observed in the outcome of no infection, 0.61 (95% CI 0.49 to 0.72, substantial agreement). Panelists reached a full consensus in 12 of 29 cases and near consensus in five of 29 cases when CDC criteria were used (superficial, deep or organ space). A stable maximum Fleiss kappa of 0.46 (95% CI 0.50 to 0.35) at CAC sizes greater than three members was obtained.

## Conclusions

There is substantial agreement among the members of the PARITY CAC regarding the presence or absence of surgical site infection. Agreement on the level of infection, however, is more challenging. Additional clinical information routinely collected by the prospective PARITY trial may improve the discriminatory capacity of the CAC in the parent study for the diagnosis of infection.

Cite this article: *Bone Joint Res* 2016;5:347–352.

**Keywords:** Infection; Reliability; Endoprosthetic reconstruction

## Article focus

- To determine the level of inter- and intra-rater agreement in the diagnosis of surgical site infection in the context of a clinical trial.

## Key messages

- There is substantial agreement among the members of the PARITY CAC regarding the presence or absence of surgical site infection.
- Agreement on the level of infection, however, is more challenging.
- Additional clinical information routinely collected by the prospective PARITY trial may improve the discriminatory capacity of the CAC in the parent study for the diagnosis of infection.

- J. Nuttall, MD,
- N. Evaniew, MD,
- P. Thornley, MD, Orthopaedic Resident, McMaster University and Hamilton Health Sciences, Hamilton, Ontario, Canada
- A. Griffin, MSc, Clinical Research Manager, University Musculoskeletal Oncology Unit | Mount Sinai Hospital, Toronto, Ontario, Canada
- B. Deheshi, MD, FRCSC, MSc, Orthopaedic Surgeon, Assistant Professor, Department of Surgery, McMaster University and Hamilton Health Sciences, Hamilton, Ontario, Canada
- T. O'Shea, MD, FRCPC, MPH, Infectious Diseases Consultant Associate Professor, Department of Medicine, McMaster University and Hamilton Health Sciences, Hamilton, Ontario, Canada
- J. Wunder, MD, Orthopaedic Surgeon, Professor Division of Orthopaedic Surgery, University of Toronto, Toronto, Ontario, Canada
- P. Ferguson, MD, FRCSC, MSc, Orthopaedic Surgeon, Associate Professor, Division of Orthopaedic Surgery, University of Toronto, 600 University Avenue, Suite 476(G) | Toronto, M5G 1X5, Canada
- R. L. Randall, MD, FACS, Orthopaedic Surgeon Professor, Department of Orthopaedics, University of Utah, 2000 Circle of Hope, Suite 4260 | Salt Lake City, 84112-5550, USA
- R. Turcotte, MD, FRCSC, Orthopaedic Surgeon, Professor, Division of Orthopaedic Surgery, McGill University, Montreal General Hospital, 1650 Cedar Avenue, Room B5.159.6, Montreal, QC, H3G 1A4, Canada
- P. Schneider, BSc, Research Coordinator, Department of Clinical Epidemiology and Biostatistics, McMaster University, 293 Wellington Street North, Suite 110, Hamilton, ON, L8L 8E7, Canada
- P. McKay, BSc, Orthopaedic Research Program Manager, Department of Clinical Epidemiology and Biostatistics, McMaster University, 293 Wellington Street North, Suite 110, Hamilton, ON, L8L 8E7, Canada
- M. Bhandari, MD, FRCSC, PhD, MD, FRCSC Orthopaedic Surgeon, Professor, Department of Clinical Epidemiology and Biostatistics and Department of Surgery, McMaster University, 293 Wellington Street North, Suite 110 | Hamilton, ON, L8L 8E7, Canada
- M. Ghert, MD, FRCSC, Orthopaedic Surgeon, Associate Professor, Department of Surgery, McMaster University, 711 Concession Street, Surgical Offices B3 169A, Hamilton, ON, L8V 1C3, Canada

Correspondence should be sent to
M. Ghert;
email: ghert@hhsc.ca

doi: 10.1302/2046-3758.58.BJR-2016-0036.R1

*Bone Joint Res* 2016;5:347–352.
*Received: 6 May 2016;*
*Accepted: 7 June 2016*

## Strengths and limitations

- There is no universal standard for a kappa value that describes acceptable or unacceptable levels of agreement
- The information recorded for the clinical vignettes may have suffered from multiple collection and reporting biases inherent in the construction of a patient's chart.
- This study is an important step towards calibration for the PARITY trial CAC.

## Introduction

In large clinical trials, Central Adjudication Committees (CACs) are often used to minimise the variability in outcomes assessment associated with having multiple investigators at multiple sites.[1,2] CACs typically consist of three or more physicians who are experts in their field, and are experienced in clinical research.[3,4] CACs operate according to pre-defined Adjudication Charters, and they assess outcome events according to predefined criteria. Use of CACs in large clinical trials significantly increases data quality and reduces research inaccuracy by minimising between-site variability and biased outcome assessment.[1,2,5]

The Prophylactic Antibiotic Regimens in Tumour Surgery (PARITY) trial is a multicentre, blinded, randomised controlled trial, using a parallel two-arm design to investigate whether long-duration post-operative antibiotic regimens (five days) will decrease the rate of deep surgical site infection among patients being surgically treated for a lower extremity bone tumour in comparison with short-duration post-operative antibiotics (24 hours).[6]

The primary outcome for the PARITY trial is the rate of surgical site infections in each arm within one year of surgery, an end point that relies heavily on subjective clinical judgement.[7] Experts agree that a universal benchmark test for diagnosing deep post-operative infection does not currently exist.[8] Multiple reports have investigated the role of serologic, microbiological, and diagnostic imaging tests, but the treating surgeon's interpretation of these findings, in combination with history and physical examination, is essential.[7] In prior studies of surgical site infection for endoprosthetic reconstruction for bone tumours, diagnostic criteria have been variably defined.[9] Our systematic review identified that only two studies[10,11] reported using the Centers for Disease Control (CDC) criteria for a deep post-operative wound infection, and few reported on commonly used serologic and laboratory parameters.[10] Given the variability in defining the primary outcome, it is essential to demonstrate that the methods to be used by the PARITY CAC will allow the CAC to agree reliably on the presence or absence of surgical site infection.[1,5]

### Study questions

- What is the level of inter-rater and intra-rater agreement of the PARITY CAC in the diagnosis of surgical site infection?
- Does the use of the CDC criteria for surgical site infection increase or decrease the level of agreement of the CAC?
- What is the optimal size of the PARITY CAC?
- Do formal consensus meetings change the results significantly from the majority opinion of independent reviewers?

## Materials and Methods

**Study design and setting.** This study assessed the inter- and intra-rater agreement of determining post-operative wound infections following wide surgical excision and endoprosthetic reconstruction for lower extremity primary bone tumours.

**Participants.** The CAC for the PARITY trial is a five-member panel of orthopaedic surgeons and an infectious disease specialist. The PARITY CAC members were asked to adjudicate 29 non-PARITY cases of lower extremity endoprosthetic reconstruction. Cases were identified without randomisation from the databases of two separate sarcoma centres. These cases were selected by the PARITY group by consensus to represent a range of potential clinical situations that were likely to be encountered in the PARITY trial. For all cases, a comprehensive search for all contemporaneously recorded clinical data, including laboratory results, clinical notes, and radiographic studies, was undertaken. All clinical data at the time of diagnosis of infection, if applicable, were exhaustively collected. Short clinical vignettes, including all relevant recorded information from each patient's electronic and physical records, were presented in a survey format (example case in supplementary material). Ethical approval for collection of these data was granted by the research ethics boards of the involved hospitals.

**Description of experiment.** All reviewers were first asked to review the CDC criteria for surgical site infection[12] and to familiarise themselves with this classification system (see supplementary material). In order to determine inter-rater reliability, all members of the CAC were asked independently to determine the presence or absence of infection for each case according to the CDC criteria for surgical site infection. Each reviewer was asked to specify whether or not the case was familiar to him or her. Reviewers were also able to select 'unable to assess' as an option if they felt that there was inadequate information contained within the clinical vignette to reach a conclusion as to the presence or absence of infection. However, reviewers were encouraged to reach a decision if at all possible. Intra-rater reliability was determined by repeating the survey using identical cases three months after initial responses were collected. A consensus meeting was held and all cases were discussed between all committee members until a consensus opinion on each case was determined. No communication about the cases or classification system occurred between reviewers before the consensus meeting. All discussions were continued until a consensus position was reached for all cases. At the conclusion of the consensus meeting, the opinions of the reviewers regarding their experiences with the application of the classification system were collected.

**Variables, outcome measures, data sources and bias.** The results of each round of the survey were collected for each reviewer, who was asked to identify if the case was recognisable to them. After the first round of the survey, inter-rater reliability was calculated. In order to reflect the primary outcome measure of the PARITY study, the CDC categories of 'superficial incisional surgical site infection', 'deep incisional surgical site infection', and 'organ/space surgical site infection' were aggregated into a single category of ‘infection’ and compared with the single category of 'no infection'. This binary outcome measure was analysed independently for inter-rater reliability. ‘Unable to assess’ was an available category for the reviewers to select throughout the study and was incorporated into all statistical analyses. The first round of survey data was also used to simulate CAC groups of varying sizes using combinatorial analysis. After the second round of the survey, the paired responses of each reviewer were analysed for intra-rater reliability. After concluding both rounds of the survey, a final CAC consensus meeting was held. Reviewers were asked to review the clinical information presented in each vignette and come to a consensus opinion for both the CDC classification of surgical site infection and the binary outcome measure. All reviewers were blinded to the results of the initial round of the survey until the final CAC meeting. These results were compared with the majority opinion for each case established in the first round of the survey.

All cases from the first round of the survey that did not reach strict consensus (all five members in agreement) on both the CDC level of infection and the aggregated binary outcome measure were referred to the CAC consensus meeting for discussion. Discussions were timed and minutes taken to observe qualitatively patterns that may have led to the lack of agreement.

**Statistical analysis.** The primary outcome to be investigated in each round of the survey was categorical, therefore agreement was calculated using the Fleiss kappa statistic.[13] Landis and Koch[14] proposed the following standards for strength of agreement for the kappa coefficient: poor (0.01 to 0.20); slight (0.21 to 0.40); fair (0.41 to 0.60); moderate (0.61 to 0.80); and substantial (0.81 to 1.00). The calculation was repeated excluding responses in cases where the reviewer identified that the case was known to them.

An *a priori* power calculation estimated that 29 cases will provide 80% power to exclude a kappa of 0.40 or lower, if the true underlying kappa is 0.90 or higher, at an alpha of 0.05. With five reviewers evaluating 29 cases, confidence intervals (CI) around kappa were estimated at +/– 0.115.[15,16]

Intra-rater agreement was calculated using the Cohen kappa statistic.[17] Combinatorial analysis of all possible combinations of CACs of sizes ranging from two to five members was analysed using the Fleiss kappa statistic (or Cohen’s kappa statistic in the case of two member panels).

All statistical analyses were in Microsoft Excel (Microsoft, Redmond, Washington) using the Real Statistics Resource Pack data analysis add-in.[18]

## Results

**What is the level of inter- and intra-rater agreement of the PARITY CAC in diagnosing surgical site infection?** All five panelists (full consensus) or four of five panelists (near consensus) agreed on the presence or absence of surgical site infection in 20 of the 29 cases. The Fleiss kappa value was calculated as 0.44 (95% CI 0.35 to 0.53), which is consistent with moderate agreement. Absolute agreement between reviewers was 80% for presence or absence of infection, and 75% using the CDC criteria (see supplementary material). Following the second round of adjudication, intra-rater agreement was substantial (0.67, 95% CI 0.87 to 0.47) for presence or absence of infection, and moderate for CDC criteria (0.57, 95% CI 0.75 to 0.39) (Table I).

**Does use of the CDC criteria for surgical site infection increase or decrease the level of agreement of the CAC?** When asked to classify each clinical vignette according to the CDC criteria for deep or organ/space surgical site infection, panelists reached a full consensus in 12 out of 29 cases and near consensus in five of 29 cases. This corresponded to a kappa of 0.35 (95% CI 0.28 to 0.42), which is fair. When all possible responses were examined, the greatest statistical agreement was observed in the outcomes of no infection, 0.61 (95% CI 0.49 to 0.72), consistent with substantial agreement, and any infection, 0.58 (95% CI 0.47 to 0.70), indicating moderate agreement (Table I). When responses from CAC members who identified that they were familiar with the case were excluded, the level of agreement identified by this analysis did not change significantly from the data above with a kappa of 0.29 (95% CI 0.22 to 0.36).

**What is the optimal size of the PARITY CAC?** Analysis of CAC size showed a stable maximum Fleiss kappa of 0.46 (95% CI 0.50 to 0.35) at CAC sizes of three members or more. No significant additional increase in kappa values was observed when CAC sizes larger than three members were constructed (Fig. 1).

**Do formal consensus meetings change the results significantly from the majority opinion of independent reviewers?** In total, 62% of all cases required consensus discussion. The average discussion time to reach a consensus was three minutes; this ranged from one minute to seven minutes. The consensus of the panel differed from the majority individual opinion in 17% of all cases. In this minority of cases, the most common conclusion

**Table I.** Comparison of inter- and intra-rater agreement of the PARITY Central Adjudication Committees (CAC) in the classification of post-operative infection using the Centers for Disease Control (CDC) criteria for surgical site infection and aggregated 'any infection' and 'no infection' categories with 95% confidence intervals

| | **Kappa** | **Significance** |
|---|---|---|
| Inter-rater reliability | Fleiss kappa | |
| Infection *vs* no infection | 0.43 (p < 0.001, 0.34 to 0.52) | Moderate agreement |
| No infection | 0.59 (p < 0.001, 0.47 to 0.70) | Moderate agreement |
| Infection | 0.46 (p < 0.001, 0.35 to 0.58) | Moderate agreement |
| Unable to assess | 0.07 (p = 0.22, 0.00 to 0.19) | Slight agreement |
| CDC | 0.35 (p < 0.001, 0.28 to 0.42) | Fair agreement |
| No infection | 0.59 (p = 0.01, 0.47 to 0.70) | Moderate agreement |
| Superficial incisional infection | 0.15 (p < 0.001, 0.03 to 0.26) | Slight agreement |
| Deep incisional infection | 0.24 (p < 0.001, 0.12 to 0.35) | Fair agreement |
| Organ/space infection | 0.27 (p < 0.001, 0.16 to 0.39) | Fair agreement |
| Unable to assess | 0.07 (p = 0.22, 0.00 to 0.19) | Slight agreement |
| Intra-rater reliability | The Cohen kappa | |
| Infection *vs* no infection | 0.68 (0.47 to 0.87) | Substantial agreement |
| CDC | 0.57 (0.39 to 0.75) | Moderate agreement |

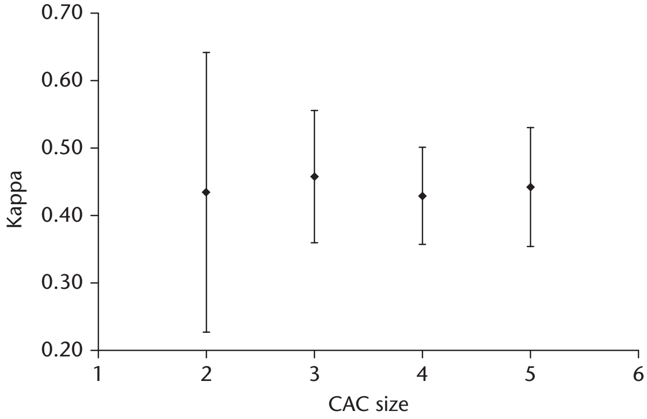

Fig. 1

Flow chart showing the comparison of Fleiss kappa agreement of increasing central adjudication committees (CAC) size following combinatorial analysis of individual respondent's classifications of surgical site infection. Error bars depict 95% confidence intervals of kappa values. A maximum stable kappa value can be seen at CAC sizes of greater than three.

was that the CAC was unable to assess the presence or absence of infection or reach an agreement on categorisation based on CDC criteria due to inadequate clinical information.

## Discussion

The complication of surgical site infection after endoprosthetic reconstruction of the lower limb in orthopaedic oncology is a major source of morbidity for this group of patients. The PARITY trial is a large multicentre trial investigating the utility of long-duration antibiotics in decreasing the rate of surgical site infection. In order to investigate the outcome of infection in PARITY, and any large surgical trial, reliably, investigators must be able to agree on whether or not a study participant has developed a surgical site infection. The data in this study provide evidence that the use of the current CAC will provide a reasonable level of agreement when asked to determine the presence or absence of infection. The subclassification of level of infection according to the CDC criteria, however, is less reliable. These data overall support the current PARITY CAC protocol which will require the committee to reach a consensus only as to the presence or absence of surgical site infection.

The study has a number of limitations. First, there is no universal standard for a kappa value that describes acceptable or unacceptable levels of agreement.[15] Though the agreement can be quantified statistically, there is no consensus in the literature as to an acceptable cut-off value for level of agreement, and this probably varies according to context. Second, although the cases used in this study were carefully selected to be representative of the type of cases that will be encountered in the PARITY trial, the clinical data were selected retrospectively from existing records at multiple cancer centres and may not actually be representative. The information recorded for the clinical vignettes may have suffered from multiple collection and reporting biases inherent in the construction of a patient's chart. These limitations likely contributed to the five out of 29 cases where the panelists' consensus position was that they were unable to assess the presence or absence of infection due to inadequate available clinical information. However, this study demonstrates a minimum level of agreement of the CAC in the determination of all possible categorisations, including 'unable to assess'. It is unlikely that the level of agreement would decrease with the provision of additional clinical information. It should be noted that the PARITY trial will collect data prospectively based on standardised reporting requirements. Finally, it should also be noted that some of the reviewers expressed that there was a learning curve associated with the implementation of the CDC criteria regarding the survey cases. As this research also served as a training exercise for the PARITY CAC, the acclimatisation of the reviewers may have decreased the inter- and intra-rater agreement observed in the study.

The completion of this study should minimise the impact of the learning curve on the final outcomes assessment of the PARITY trial.

Two previous studies have used the CDC classification system for post-operative surgical site infection, however, the system remains largely invalidated.[8] Our conclusion from this study suggests that using the full classification system decreases the level of agreement between reviewers to 'fair'. The decrease in the kappa value associated with increasing the level of subclassification is consistent with results from other nested classification systems in the orthopaedic literature.[19-21] When these independent assessments were pooled in the final consensus meeting, it was noted that critical pieces of information, such as cultures or laboratory results that would be required for classification under the CDC system, were either not recorded or not routinely collected in standard patient follow-up. As such, the results of the study do not support using the full CDC classification system to differentiate between levels of post-operative infection when chart data are reviewed retrospectively. It is possible that in the context of a prospective study such as the PARITY trial, additional information could be routinely collected to meet the requirements of the CDC classification system. However, the clinical importance of differentiating between levels of infection remains uncertain. Superficial infections may not require re-operation for irrigation and debridement, nonetheless there is a possibility that these infections may progress to deep infections and therefore should be encapsulated in studies investigating infection as an outcome.

The results of this study suggest that there is a high level of agreement between reviewers in determining the presence or absence of infection. For outcomes with a high degree of agreement, Walter et al[4] suggested that only three adjudicators are required. Larger committees increase the expense and logistical difficulty, but often do not yield significant statistical benefit. The Study to Prospectively Evaluate Reamed Intramedullary Nails in Tibial Fracture (SPRINT) investigators demonstrated that reducing their adjudication committee from six to four members would not have altered their study findings and would have reduced incremental adjudication costs by nearly $100 000.[3] Total time saving would have been approximately 18 000 man hours. Use of an odd number of committee members is considered to be particularly important for resolving disagreements and achieving consensus.[3] Although the methodology used in this study differs from that used by Walter et al, a similar conclusion was reached. There is no significant benefit to inter-reviewer agreement when CAC sizes of greater than three are used. On a practical level, a reasonable number of additional members may allow for consistent full quorum participation as the PARITY study progresses.

This study did not demonstrate that formal consensus meetings would change the interpretation of the majority of cases sent to the CAC for review. Nonetheless, a substantial minority of case outcomes were changed by the consensus meeting. The reviewers involved indicated that the interactive discussion of cases subjectively improved the reliability of their interpretation by uncovering systematic errors in the application of the CDC criteria that would otherwise bias results. The meeting also created a universal standard for interpretation of cases sent to the CAC in situations where the CDC criteria could be interpreted differently by different reviewers. The ability to reflect on the overall protocol and the information required for the PARITY CAC was also considered valuable.

The use of CACs is intended to reduce the amount of bias that is inherent in large multicentre trials where multiple investigators may make independent decisions as to the study outcome. This study demonstrates that a small CAC can use the CDC guidelines for post-operative surgical site infection to determine the presence or absence of infection reliably. There is a substantial decrease in reliability when CAC members are asked to differentiate between the categories of the CDC guidelines. Consensus meetings are an efficient and effective method of managing disagreement among CAC members.

## Supplementary material

A figure showing an example of a clinical vignette and tables showing the sum of reviewer responses for each reviewed case and Centers for Disease Control criteria for a deep post-operative wound infection can be found alongside this paper at http://www.bjr.boneandjoint.org.uk/

## References

1. **Barringhaus KG, Zelevinsky K, Lovett A, Normand SL, Ho KK.** Impact of independent data adjudication on hospital-specific estimates of risk-adjusted mortality following percutaneous coronary interventions in massachusetts. *Circ Cardiovasc Qual Outcomes* 2011;4:92-98.
2. **Vannabouathong C, Saccone M, Sprague S, Schemitsch EH, Bhandari M.** Adjudicating outcomes: fundamentals. *J Bone Joint Surg [Am]* 2012;94-A:70-74.
3. **Simunovic N, Walter S, Devereaux PJ, et al.** Outcomes assessment in the SPRINT multicenter tibial fracture trial: adjudication committee size has trivial effect on trial results. *J Clin Epidemiol* 2011;64:1023-1033.
4. **Walter SD, Cook DJ, Guyatt GH, King D, Troyan S.** Outcome assessment for clinical trials: how many adjudicators do we need? Canadian Lung Oncology Group. *Control Clin Trials* 1997;18:27-42.
5. **Carson P, Fiuzat M, O'Connor C, et al.** Determination of hospitalization type by investigator case report form or adjudication committee in a large heart failure clinical trial (β-Blocker Evaluation of Survival Trial [BEST]). *Am Heart J* 2010;160:649-654.
6. **Ghert M, Deheshi B, Holt G, et al.** Prophylactic antibiotic regimens in tumour surgery (PARITY): protocol for a multicentre randomised controlled study. *BMJ Open* 2012;2:e002197.
7. **Parvizi J, Zmistowski B, Berbari EF, et al.** New definition for periprosthetic joint infection: from the Workgroup of the Musculoskeletal Infection Society. *Clin Orthop Relat Res* 2011;469:2992-2994.
8. **Alexander JW, Solomkin JS, Edwards MJ.** Updated recommendations for control of surgical site infections. *Ann Surg* 2011;253:1082-1093.
9. **Racano A, Pazionis T, Farrokhyar F, Deheshi B, Ghert M.** High infection rate outcomes in long-bone tumour surgery with endoprosthetic reconstruction in adults: a systematic review. *Clin Orthop Relat Res* 2013;471:2017-2027.

10. **Hardes J, von Eiff C, Streitbuerger A, et al.** Reduction of periprosthetic infection with silver-coated megaprostheses in patients with bone sarcoma. *J Surg Oncol* 2010;101:389-395.
11. **Morii T, Yabe H, Morioka H, et al.** Post-operative deep infection in tumour endoprosthesis reconstruction around the knee. *J Orthop Sci* 2010;15:331-339.
12. **Mangram AJ, Horan TC, Pearson ML, Silver LC, Jarvis WR.** Guideline for Prevention of Surgical Site Infection, 1999. Centers for Disease Control and Prevention (CDC) Hospital Infection Control Practices Advisory Committee. *Am J Infect Control* 1999;27:97-132.
13. **Fleiss JL, Levin B, Cho Paik M.** *Statistical Methods for Rates and Proportions. Wiley Series in Probability and Statistics.* Hoboken: John Wiley & Sons, 2003.
14. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174.
15. **Donner A, Eliasziw M.** A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Stat Med* 1992;11:1511-1519.
16. **Sim J, Wright CC.** The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257-268.
17. **Cohen J.** Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213-220.
18. **Zaiontz C.** Real Statistics Using Excel. http://www.real-statistics.com Release 3.8 (date last accessed 8 June 2016).
19. **Hammond KE, Dierckman BD, Burnworth L, Meehan PL, Oswald TS.** Inter-observer and intra-observer reliability of the Risser sign in a metropolitan scoliosis screening program. *J Pediatr Orthop* 2011;31:e80-e84.
20. **Ihejirika RC, Thakore RV, Sathiyakumar V, et al.** An assessment of the inter-rater reliability of the ASA physical status score in the orthopaedic trauma population. *Injury* 2015;46:542-546.
21. **Lippe J, Spang JT, Leger RR, et al.** Inter-rater agreement of the Goutallier, Patte, and Warner classification scores using preoperative magnetic resonance imaging in patients with rotator cuff tears. *Arthroscopy* 2012;28:154-159.

**Funding Statement**

- Funding has been received from the Canadian Institutes of Health Research (CIHR) and Physicians' Services Incorporated Resident Research Grant, both of which are related to this article. Dr Evaniew is funded in part by a CIHR Doctoral Research Award.
- Professor Bhandari reports funding received from Smith and Nephew, Stryker, Amgen, Zimmer, Moximed, Bioventus, Merck, Eli Lilly, Sanofi, Conmed, Ferring; none of which is related to this article.
- Dr Ghert reports funding received from Wright Medical, none of which is related to this article.
- R. L. Randall reports funding received from Biomet which is not related to this article.

**Author Contribution**

- J. Nuttall: Study design, Data acquisition, Data analysis, Manuscript preparation and editing
- N. Evaniew: Study conception, Study design, Data acquisition, Data analysis, Manuscript editing and review
- P. Thornley: Study design, data acquisition, manuscript editing and review
- A. Griffin: Data acquisition, Manuscript editing and review
- B. Deheshi: Study design, Data acquisition, Manuscript editing and review
- T. O’Shea: Study design, Data acquisition and analysis, Manuscript editing and review
- J. Wunder: Data acquisition, Manuscript editing and review
- P. Ferguson: Data acquisition, Manuscript editing and review
- R. L. Randall: Study design, Data acquisition, Manuscript editing and review
- R. Turcotte: Study design, Data acquisition, Manuscript editing and review
- P. Schneider: Study design, Data acquisition, Manuscript editing and review
- P. McKay: Study design, Data acquisition, Manuscript editing and review
- M. Bhandari: Study design, Data acquisition, Manuscript editing and review
- M. Ghert: Study conception, Study design, Data analysis, Manuscript editing and review, Study supervision

**ICMJE conflict of interest**

- None declared