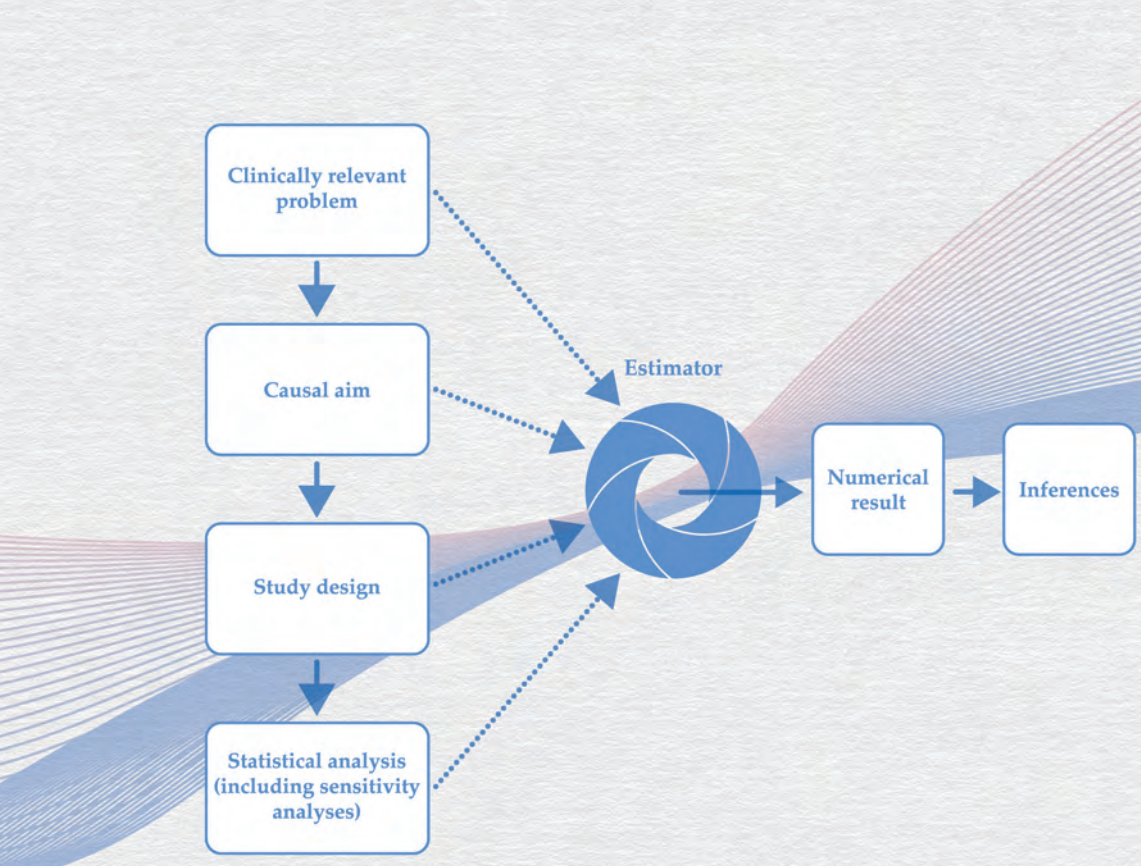
The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling

Luijken, K.

5


Clinically relevant problem
Causal aim
Study design
Statistical analysis (including sensitivity analyses)
Estimator
Numerical result
Inferences

How to assess applicability and methodological quality of studies of operative interventions in orthopedic trauma surgery

It is challenging to generate and subsequently implement high quality evidence in surgical practice. A first step towards improving the quality of surgical studies is to assess current methods to grade the strengths and weaknesses of surgical evidence and appraise current risk of bias tools for the surgical community. Here, we described items that are common to different risk-of-bias tools, how these could be used to assess operative intervention studies in orthopedic trauma surgery, and how these relate to applicability of results. We extracted information from the Cochrane risk-of-bias-2 (RoB-2) tool, Risk Of Bias In Non-randomised Studies - of Interventions tool (ROBINS-I), and methodological index for non-randomized studies (MINORS) criteria and derived a concisely formulated set of items tailored to operative interventions in orthopedic trauma surgery. The set contained nine items: population, intervention, comparator, outcome, confounding, missing data and selection bias, intervention status, outcome assessment, and pre-specification of analysis. Each item can be assessed using signaling questions and was explained using good practice examples of operative intervention studies in orthopedic trauma surgery. The set of items will be useful to form a first judgment on studies that have been included in a systematic review. Existing risk of bias tools can be used for further evaluation of methodological quality. Additionally, the proposed set of items might be a helpful starting point for peer reviewers.

1| Background

It is challenging to generate and subsequently implement high quality evidence in surgical practice[1]. In the field of orthopedic trauma surgery, it takes approximately ten years from design to execution of an RCT[2]. What is more, Oberkofler and colleagues showed that results of surgical RCTs often do not convince the surgical community of their findings due to a substantial risk of bias[3]. This is a highly undesirable situation, because a lot of effort, time, public money, and patient participation is spent on research with futile impact on surgical care[4].

To what extent a study can inform surgeons and patients depends on its applicability and methodological quality. Appraising the methodological quality of a study and judging the applicability (external validity or generalizability) of study results to clinical practice remains challenging in the field of surgical research, especially the assessment of bias (internal validity). This is reinforced by the fact that systematic reviews of operative interventions increasingly include both randomized controlled trials (RCTs) and observational studies[5,6], adding to the complexity of the assessment.

Many comprehensive risk-of-bias tools are available to assess the methodological quality of studies of interventions[7-10], such as the Cochrane risk-of-bias (RoB 2) tool for randomized trials[11] and the Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I) tool[12]. However, these tools focus on internal validity (risk of bias) aspects of a study and do not simultaneously evaluate clinical applicability of the results. The tools were often developed with a focus on studies of pharmacological interventions and may therefore not be ideally suited for studies of operative interventions. Additionally, it is convenient to assess both types of studies using a single list of items.

Here, we describe the selection of items that are common to different risk-of-bias tools, how these could be used to assess operative intervention studies in orthopedic trauma surgery, and how these relate to applicability of results. We take the perspective of a researcher who performs a systematic review and wants to make a first judgment on the applicability and methodological quality of included studies. Relevance of these items for editors, peer reviewers, and researchers will be addressed in the discussion section.

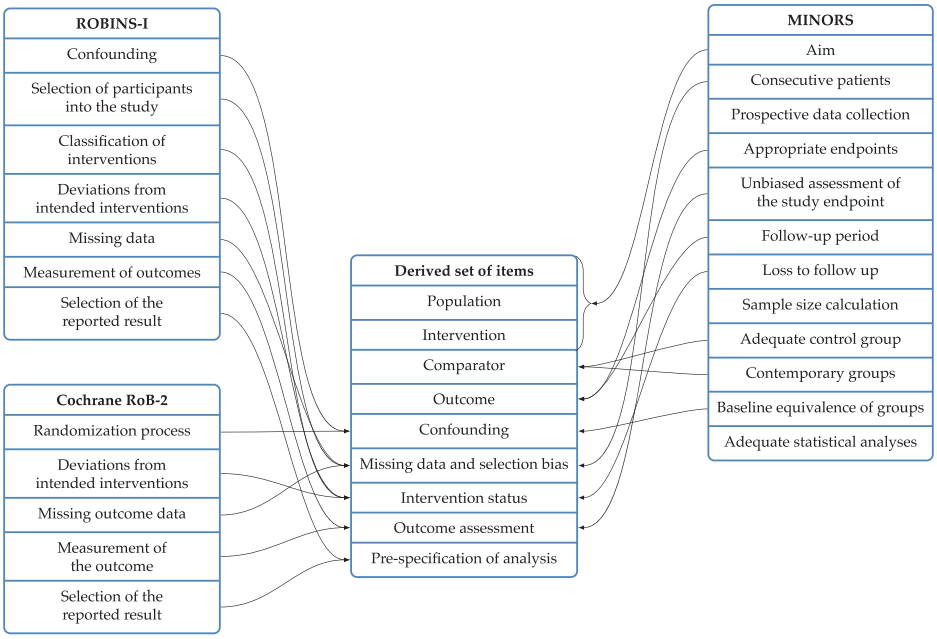


Figure 1. Flow diagram indicating which existing risk of bias tool the signaling questions in the concise set were based on.

2| Extracting items from existing risk-of-bias tools

To establish an easy-to-use set of items for assessment of applicability and methodological quality of studies of operative interventions, we extracted information from the RoB-2[11], ROBINS-I[12], and methodological index for non-randomized studies (MINORS) criteria[13] and derived a concisely formulated set of items tailored to operative interventions. It has been pointed out that the methodologically rigorous RoB-2 and ROBINS-I tools require a high level of statistical knowledge, making their implementation challenging and time consuming[14-16]. We aimed to summarize scoring items in such a way that the items were easy to use for assessment of articles of both RCTs and observational studies of operative interventions.

All signaling questions from the RoB-2, ROBINS-I, and MINORS were taken as a starting point. We identified signaling questions with overlapping topics. Based on their relevance to surgical studies, the overlapping set formed the initial key items. We then evaluated remaining signaling questions. Questions that were less relevant for studies of operative interventions, such as questions regarding time-varying exposures,

Table 1 Overview of items and overarching questions of the established set. The Appendix describes the signaling questions for each item.

Applicability		
Item	**Overarching question**	**Explanation**
Population	Is the patient population included in the study representative of the patient population defined in the PICO of the systematic review?	Patients included in the study are ideally representative of patients that would be typically encountered in the clinical practice setting for which the PICO is defined.
Intervention	Is the investigated intervention representative of the intervention defined in the PICO of the systematic review?	The studied operative intervention is ideally performed similar to the procedure that would be typically performed in the clinical practice setting for which the PICO is defined.
Comparator	Is the comparator intervention representative of the comparator defined in the PICO of the systematic review?	The comparator intervention is ideally performed similar to the procedure that would be typically performed in the clinical practice setting for which the PICO is defined.
Outcome	Is the outcome representative of the outcome defined in the PICO of the systematic review?	The outcome should be relevant to patients typically encountered in the clinical practice setting for which the PICO is defined and should be measured using an appropriate procedure at the appropriate time.
Methodology		
Item	**Overarching question**	**Explanation**
Confounding	Is there comparability of intervention groups, or are appropriate methods applied to correct for incomparability?	An important assumption needed to make causal claims about effects of operative interventions is comparability of intervention groups. This can be established through appropriate randomization (in randomized studies) or adjustment for confounding (in observational studies).
Missing data and selection bias	Were the patients included in the analysis representative of all patients included in the study and was the impact of missing data negligible?	Selection of patients based on missing values or phenomena that occurred after inclusion into the study can introduce bias in the effect of the operative intervention.
Intervention status	Was the intervention status correctly classified?	To infer the effect of a operative intervention it should be clear how crossovers and deviations from planned interventions are dealt with in the analysis.
Outcome assessment	Was the outcome correctly measured?	The study outcome should be measured such that it does not influence the estimated effect by (dis)favoring the outcome of one of the intervention groups.
Pre-specification of analysis	Were analyses prespecified and did the study adhere to the specified analysis plan?	Specifying analyses prior to analyzing the data prevents researchers from (often unintentionally) trying many approaches to fit the data, defining the research question/ hypothesis only after observing the result, and selectively reporting the findings that yield the desired result.

were discarded and questions were reformulated to be more appropriate for a surgical context (Figure 1). The set of items was further improved by user experiences in an accompanying study that assessed the applicability and methodological quality of studies from two recent systematic reviews (van de Wall, forthcoming).

The finally established set contained four items on applicability and five items of methodology (Table 1). Each item contained multiple signaling questions that help to arrive at an overarching judgement of the study quality regarding that topic, with some signaling questions specifically applicable only to RCT or observational studies. The proposed set aimed to summarize key information needed to assess the applicability and methodological quality of operative intervention studies, but the choice for items and signaling questions was surely arbitrary, and the set is opened to further elaborations and improvements.

3| Nine items to assess applicability and methodological quality

The set contained nine items: population, intervention, comparator, outcome, confounding, missing data and selection bias, intervention status, outcome assessment, and pre-specification of analysis. This set can be split in two subsets, representing applicability (first four items) and methodological quality (remaining five items).

3.1 | Items of applicability

The first four items represent the starting point of almost every clinical study, which is a clearly articulated research question (see Box 1). In a systematic review, the research question determines which original studies should be included as well as the degree to which they can provide valuable evidence. The well-known PICO acronym can be a helpful structure when defining a research question about the possible effect of an operative intervention[17].

Box 1. Well-definedness of research questions is crucial in studies of complex interventions.

Studies investigating causal effects of interventions, both randomized and non-randomized, provide scientific evidence to inform medical decisions about those interventions. Ideally, a study indicates clearly which medical decision can be informed by the findings by unambiguously defining the research question, specifying the target population, the intervention strategies that are compared, and what outcome is considered (and when).

In studies of pharmacological interventions, a research question could for instance be 'what is the effect of taking drug A compared to taking drug B on a particular outcome in a specific population?' Although this seems trivial, some parts of this question are not yet clearly defined. How is the drug administered (e.g., oral, or intravenous) and what dosages are compared? Other aspects, however, may be irrelevant, such as the hand with which a pill was taken or what shoes the individual was wearing when they took the drug. A research question should be sufficiently well defined in the sense that all *relevant* aspects are specified and thus should be addressed in the study design and analysis[18,19].

Arguably, pharmacological interventions consist of less components than operative interventions and it is more straightforward to define them precisely. Studies of operative interventions go beyond a mere description of surgical techniques; other relevant aspects include the pre- and post-surgery treatment, experience of the surgeon and team, and more. On top of that, the operative intervention itself is tailored to a particular patient[20]. Hence, defining all relevant aspects in a research question demands considerable time and effort in studies of operative interventions.

For further reading on sufficiently well-defined research questions, we refer to[18] and[19].

As an example, consider the PICO for the systematic review on operative treatment of proximal humerus fractures in the accompanying paper by van de Wall and colleagues (van de Wall, forthcoming). The PICO of the systematic review was to compare functional outcomes measured using a validated functional score for the shoulder one year after plate osteosynthesis (minimally invasive or open reduction and internal fixation) followed by 6 weeks none-weightbearing functional treatment versus one year

after initiation of conservative intervention, consisting of 6 weeks of no weight bearing, pain-guided movement and a sling if necessary, in patients with a closed, displaced, proximal humerus fractures older than 18 years.

Item 1. Population

The population defined in a research question ideally matches the patient population typically encountered in the clinical setting for which the study is conducted. In orthopedic trauma surgery, elements that define the population are, e.g., the anatomical location of the fracture, the type of fracture (e.g., open/closed, simple/multifragmentary, or combination), and age group. Fjalestad and colleagues[21] defined the relevant population as "*patients aged 60+ years with a displaced, unstable three-or four-part proximal humerus fracture of OTA group 11-B2 or 11-C2 (displaced fracture of extra-articular or articular, bifocal type) without previous shoulder injuries*". Because the population of interest was clearly reported (and its characteristics summarized in a table), the degree to which it matches to the population specified in the example PICO of the systematic review can easily be assessed.

Item 2. Intervention

Obviously, the studied operative intervention should be clearly defined. In case of an operative intervention, this entails, e.g., specification of the osteosynthesis material, surgical approach and the type and duration of the post-operative treatment regime. In case of a conservative intervention, the duration and type of conservative intervention should be clearly reported.

For example, Fjalestad and colleagues[21] defined the studied intervention as follows: "*Patients allocated to surgery were operated on within 1 week of hospital admission. The goal of surgery was anatomic reduction of the fracture and fracture stabilization [using angular stable plate] to allow for early mobilization. After surgery, patients were immobilized in a modified Velpeau bandage until self-exercises and training instructed by a physical therapist were started on the third postoperative day.*" This was accompanied by a detailed account of the operative technique and the physiotherapy protocol, such that it was clear from the description what the intervention constituted. The intervention corresponds to the intervention defined in the example PICO of the systematic review, with the exception that the post-treatment regime was extended to include strengthening exercises after 6 weeks and a recommendation of physical therapy for at least 6 months.

Other relevant aspects of the intervention are whether study hospitals routinely perform the intervention, which help to clarify whether participating surgeons are experienced in conducting the investigated procedure. For instance, Fjalestad and colleagues indicated that: "*Three surgeons performed all operations and were trained in the surgical technique before performing surgery on study participants. Surgeons 1, 2, and 3 performed 18, five, and two operations, respectively. Surgery occurred during daytime hours*"[21]. Also, a learning curve (or the absence thereof) could be relevant. For example, Knobe et al. compared helical blade nailing of the femoral head versus locked plating and reported that[22]: "*[t]hree surgeons […] were proficient in the locked plating technique and three […] were proficient with helical blade nailing. Both implants had been used by the surgeons for more than 3 years, so they would have been beyond the learning curve and they had a comparable experience level for each implant*".

Item 3. Comparator

Similar to the studied intervention, the comparator intervention should be clearly defined, and the same considerations apply. For example, Fjalestad and colleagues[21] defined the comparator intervention as follows: "*On admission to the hospital, patients were immobilized in a modified Velpeau bandage. All patients allocated to conservative treatment stayed in the hospital for at least 1 day and received the same instructions from the physiotherapist as patients allocated to surgery*", accompanied by a description of an optional closed reduction procedure. The unambiguous reporting of the conservative treatment regime allowed for assessment of the applicability of the comparator arm with respect to the comparator specified in the example PICO of the systematic review. While the conservative intervention is roughly similar to the definition of the comparator intervention in the systematic-review PICO, the optional closed reduction was not part of the systematic-review PICO.

Item 4. Outcome

Specification of a relevant study outcome consists of three parts: the outcome definition, the timepoint at which the outcome is assessed and the measurement procedure or instrument by which the outcome is assessed. For example, Fjalestad and colleagues[21] defined the primary outcome as functional outcome at one year, indicating the outcome definition and timepoint at which it was assessed. The outcome measurement was the Constant score, which is a score ranging 0 – 100 measured by self-reported pain (max. 15 points), self-reported activities of daily-living (max. 20 points), range of

motion (forward and lateral elevation, max. 10 points each, and external and internal rotation, max. 10 points each), and power (25 points)[23]. The unambiguous reporting of the outcome definition, timepoint and measurement procedure allowed for assessment of the applicability of the outcome with respect to the outcome specified in the example PICO of the systematic review.

3.2 | Items of methodology

Five methodological items are key for assessing methodological quality of a study: confounding, missing data and selection bias, classification of intervention status, outcome assessment, and pre-specification of the statistical analysis. Each of the items will be discussed below.

Item 5. Confounding

Comparability of intervention groups is essential for evaluation of effects of operative interventions and can be invoked by appropriate randomization (in randomized studies) or adjustment for confounding (in observational studies).

In randomized studies, a random allocation sequence and concealment of that allocation contribute to comparability of intervention groups, leading to comparability in observed (and unobserved) characteristics of study groups at baseline. An example of a clear description of the randomization procedure is given by Rangan and colleagues[24]: "*After obtaining informed consent and key baseline information, research associates randomly allocated patients to surgical or nonsurgical treatment using an independent remote randomization service (telephone or online access) provided by the York Trials Unit (University of York). Randomization was performed using a computer program with 1:1 allocation, stratifying by tuberosity involvement (yes or no) and using random block sizes of 4, 8, and 12.*" Based on this information, it can be assessed that the allocation sequence was random. Furthermore, inspection of baseline differences between intervention groups suggested no clinically relevant differences in observed characteristics. "*The baseline characteristics […] for randomized patients (N = 250) and those providing [Oxford Shoulder Score] data at 2 years (n = 215) were well balanced except for smoking status (there were more smokers in the nonsurgical group)*"[24].

Ideally, the allocation sequence is concealed at least until patients are enrolled in the study[25]. In case research associates or patients are aware which intervention the next enrolled patient will receive this might influence the decision to enroll (the patient) into

the study and thus limit comparability of study groups. Hence, a detailed description of the allocation procedure is needed to assess the validity of the intervention allocation.

In observational studies, allocation of intervention is no random process, and intervention groups cannot be presumed to be comparable. Therefore, a key requirement for observational studies of operative interventions is that researchers argue convincingly that intervention groups are comparable or that they provide enough detail to assess whether important clinical characteristics are sufficiently controlled for in the statistical analysis of the study[26].

An example of the former is a study by Beks and colleagues, who compared the effect of rib fixation based on a clinical treatment algorithm on intensive care unit length of stay to nonoperative intervention for both patients with a flail chest and patients with multiple rib fractures[27]. They compared groups of patients with rib fractures admitted to hospitals that either operated most patients or mostly treated patients conservatively. Allocation of emergency patients to hospitals is to a certain extent a random process, based on availability and location of the accident. When different hospitals treat patients with similar symptoms with different interventions, this allows for a natural experiment by comparing outcomes across hospitals[28]. In this example, confounding due to severe incomparability of intervention groups was deemed unlikely by design. Additionally, Beks and colleagues adjusted for a number of confounders using propensity score matching.

Indeed, when intervention groups cannot be considered to be (fully) comparable by design, statistical adjustment for measured confounders can be considered. For example, Jenkinson and colleagues adjusted for variables that are considered to be confounders, because they are known risk factors of the outcome and/or they may have contributed to the indication for a particular intervention[29]: "*The factors considered to be the most important confounders also contributing to deep-infection risk were chosen for the propensity-score algorithm. These factors included patient age, sex, time delay to debridement, fracture grade (Gustilo-Anderson grade I, II, or IIIA), evidence of gross contamination, tibial compared with nontibial site, and ASA class (1 or 2 compared with 3 or higher). These factors were chosen, based on consensus among the investigators, as the factors most important for predicting later infection but also as those most divergent between the immediate and delayed-closure groups*". Jenkinson and colleagues selected confounders based on background knowledge, in line with recommendations that specialist knowledge about the relation

between covariates and the complex intervention and/or outcome is needed to identify a set of potential confounders.

A common misconception is that confounders can be identified based on statistical criteria. In fact, statistical criteria cannot identify nor discard covariates as being confounder variables[30-34]. Of note, most statistical methods to adjust for confounding (including propensity score methods) can only adjust for measured confounding variables. After confounding adjustment, bias due to unmeasured confounding may still be present, e.g. because a confounder was measured inaccurately (or a continuous variable was dichotomized), or not measured at all[35]. A final note on confounders is that it is advisable not to interpret coefficients of confounding variables as causal effects or independent prognostic associations[36].

Item 6. Missing data and selection bias

Data are often incomplete. In some circumstances, data can be missing without substantially affecting the results. When this is the case, a study report should clarify why missingness is thought to have no effect on the study outcome, as was done for example by Portinary and colleagues[37]: "*To evaluate the impact of the emergency operations on postoperative functional status, the [activities of the daily living (ADL)] scores at the time of discharge were compared to the pre-admission ADL scores using the Chi-square test. Only patients for whom both pre-admission and postoperative ADL scores were available were included in this analysis. The subgroup analysis comparing patients with missing ADL score data with those where data was available showed no differences in terms of demographic and baseline characteristics […]. Therefore, participants without missing ADL score data were considered as a random sample of the study population. Therefore, missing data were considered to be completely at random and a complete case analysis was performed*".

In many cases, however, excluding patients for whom information on some variables is missing can introduce bias, because the missingness is related to observed or unobserved characteristics of the patients[38-40]. Beks and colleagues assumed missingness in their study was at random and described how it was dealt with[27]: "*We applied multiple imputation (25 times) to impute missing values for ASA [2.1% (7/332)], TTSS [20% (67/332)], AIS head [0.6% (2/332)], pulmonary contusion [0.6% (2/332)], pH [9.0% (30/332)], and base excess [9.0% (30/332)]. Multiple imputation was performed using the mice() algorithm in R*".

Studies should describe patterns of missing data and describe the assumed missing data mechanism. Otherwise, it is impossible to assess the potential impact of missing data and whether this was dealt with appropriately. The validity of performing a

complete case analysis cannot be assessed from a study that merely states that patients with missing values were excluded from analysis. Pointing out that few cases were missing is not a valid justification of complete case analysis, since the proportion of missing data is not directly linked to the severity of the bias that is introduced by it[41].

Apart from variables having missing values, subjects can also be missing entirely in case they are not included in the study, which could lead to selection bias. However, if those included in the study are representative of the entire set of eligible subjects, the risk of this type of bias seems small. Klei and colleagues provided a clear description why patients included in the analyses seemed representative of all patients included in the study[42]: "*Among the 116 sternovertebral fracture patients, 43 patients were excluded from further analysis (1 military patient, 14 patients who died early after admission before fracture treatment, 14 patients with either isolated upper cervical spine or lower lumbar spine fractures, and 14 patients who were lost to follow-up). The remaining 73 patients were included for further analysis*".

Sometimes patients are excluded from analysis because they do not consent to participate in the study. A comparison between patients that consented and refused to partake in the study can be done to assess the possibility that selection bias is introduced, as is described by Rangan and colleagues[24]: "*Of the 563 eligible patients, 250 (44%) consented to take part in the trial […]. The mean age of the […] participants was 66 years (range, 24-92 years), 192 (77%) were female, and 249 (99.6%) were white. These characteristics were similar to patients who refused consent (mean age, 68 years; 75% female).*"

Item 7. Intervention status

The defined PICO specifies which operative interventions are compared including their post-intervention regimens. However, deviations from these ideally unambiguously defined, yet potentially hypothetical, situations can occur in clinical practice, both in the intervention and comparator arm.

The intervention status can be incorrectly registered in the data, referred to as 'misclassification'. For instance, when data are retrieved from electronic health records, the procedure may be inaccurately registered or incorrectly extracted into the analytical dataset. A patient may falsely be recorded not to have received an operative intervention, while they actually had, or selection bias can be introduced in case of a comparison of operative interventions. However, in most cases, misclassification of surgical interventions seems unlikely.

Defining the intervention status of a patient in the final analysis is not straightforward when the patient was assigned to one intervention arm, but actually received the opposite intervention (too). This is commonly referred to as a *cross-over*. Which intervention status patients should then be assigned to depends on the aim of the study, and in particular on the intervention effect of interest. For instance, an RCT by Van der Meijden and colleagues aimed to estimate an intention-to-treat effect and assessed patients in the intervention group that they were randomized to[43]. "*One patient (2%) in the plate group and six patients (10%) in the nailing group underwent intraoperative crossover to the other treatment group and were further analyzed as part of their original treatment group according to the intention-to-treat principle*". Consequently, the result of the study no longer represents an effect of plate fixation versus intramedullary nailing on functional recovery. Rather, it represents an effect of plate fixation with optional revision using intramedullary nailing versus intramedullary nailing with optional revision using plate fixation on functional recovery. Although this interpretation is arguably less straightforward, it might be the effect of main interest in clinical practice.

Considerations regarding patients' intervention status differ slightly between RCTs and observational studies. In RCTs, a cross-over commonly refers to a patient who was assigned to a particular intervention, but then received an alternative intervention, meaning that the patient received a single intervention. In observational studies, a cross-over commonly refers to a patient who received a particular intervention first and then received the alternative intervention, meaning that the patient received both interventions. Cross-over interventions in comparisons of operative versus non-operative interventions often pose a more challenging problem than crossovers between operative interventions.

Finally, including the post-operative treatment regime for determining a patient's intervention status likely complicates matters considerably. Adherence to post-operative treatment is often less well documented and post-operative treatment options may be combined for some patients.

Item 8. Outcome assessment

Ideally, the study outcome is measured in the same way in all study patients, notably irrespective of the intervention a patient received. This can be achieved by means of a valid and reliable procedure to measure the outcome[44]. For instance, the outcome 'quality of life' can be measured using a well-established questionnaire such as the EQ5D, as was done by Banierink and colleagues[45]: "*Quality of life was assessed with the*

EuroQol 5D (EQ-5D). The EQ-5D is a brief questionnaire that measures health-related quality of life based on five dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression [17]."

Functional outcomes are ideally measured using validated instruments, too, as was done by Ochen and colleagues [46]: "*Functional outcome was assessed at least 12 months following [the operative intervention], using the Dutch language version of the QuickDASH score. The QuickDASH is a validated and shortened version of the Disabilities of the Arm, Shoulder and Hand questionnaire (DASH)*".

When the outcome is measured using a non-standardized measurement, outcome values may be less reliable. For instance, when forward or lateral elevation of the shoulder is measured by visual inspection rather than by use of a goniometer, values may be less reliable and inter-rater variability is likely increased. On top of that, an outcome assessor may (subconsciously) be affected by knowledge about the intervention that a patient received (i.e., when they are unblinded). These considerations apply to functional outcomes and self-reported outcomes, including patient reported outcome measures (PROMs), alike.

To prevent bias by unblinded outcome assessment, Nauth and colleagues designed their RCT anticipating the bias that could be introduced by differential unblinded outcome assessment of their primary outcome re-operation[47]: "*Surgeons and patients were not blinded. However, we did minimise the associated risk of bias with central and independent, although unblinded, radiographic adjudication of the primary endpoint*". A Committee adjudicated re-operations at the end of follow-up, where re-operation was defined as surgery to promote fracture healing, relieve pain, treat infection, or improve function within 24 months after the initial procedure (described in detail in the supplements of [47]).

Item 9. Pre-specification of analysis

The credibility of results can be diminished by trying many approaches to fit the data and selectively reporting the results that yield the desired outcome. When the choice to perform a statistical test depends on patterns in the data, the expected number of false positives (i.e., type I error rate) is likely inflated[48]. Similarly, the type I error rate increases when more statistical tests are conducted on the same data set, thus performing multiple statistical tests without reporting all of them in the published manuscript prohibits readers from assessing the potential for false positive findings. Although such data dredging and cherry picking has harmful consequences, these practices may well be conducted unintentionally – especially when findings (in

hindsight) are convincing and easy to explain. To overcome this problem, statistical data analysis should be prespecified as much as possible, e.g., by means of a statistical analysis plan that defines which analyses will be performed and the methods used to perform these analyses, including handling of missing data[49]. For RCTs, preregistration of the study protocol is considered the norm[50], but for observational studies, study protocols seem to be pre-specified less often, although the urgency to do this is certainly recognized[51].

To further enhance transparency, protocols can be made publicly available to allow for assessment of protocol adherence. Protocols can be preregistered at, e.g., https://clinicaltrials.gov/ (for RCTs), https://www.isrctn.com/ (both RCTs and observational studies), https://osf.io/ (both RCTs and observational studies), and protocols for systematic reviews can be preregistered on https://www.crd.york.ac.uk/prospero/ or https://osf.io. Journals such as International Journal of Surgery Protocols or the British Medical Journal Open allow for publication of study protocols.

Good examples of publicly available study protocols are a trial by Smeeing and colleagues, who compared functional outcome twelve weeks after randomization to unprotected non-weight-bearing, protected weight-bearing, or unprotected weight-bearing as tolerated in patients who underwent surgical fixation of ankle fractures[52]. The protocol is available at https://www.trialregister.nl/, NTR3727[53]. Taha and colleagues registered an observational pilot study to assess the feasibility of performing an RCT to study the effect of operative intervention of metacarpal fractures affecting the index to little finger(s) compared to non-operative intervention. The study is currently ongoing and is registered at ISRCTN (13922779).

Red flags

Apart from the aforementioned items, a study can contain aspects that set alarm bells ringing about the quality of the methodology or statistical analysis that do not fit within items 5 – 9 specifically. As a systematic reviewer or peer reviewer, it is important to be aware of this and make notes about such red flags.

4| Discussion

We proposed a concise set of items, based on existing risk-of-bias tools, to perform an initial assessment of the applicability and methodological quality of randomized and non-randomized studies into effects of operative interventions in orthopedic trauma surgery. In terms of the IDEAL Framework[54-58], this set of items is intended to assess stage 3 (assessment) and stage 4 (long-term monitoring) studies. This assessment can be done as part of a systematic review to discard studies of low quality with relative ease and to separate out higher quality studies for further scrutiny of methodological quality using available assessment tools[11,12,59,60].

In the current study, we took the perspective of a systematic reviewer, who can use the set of items to appraise studies included in a systematic review and to determine which articles can be considered for a subsequent meta-analysis. However, the proposed set of items might be a helpful starting point also when taking on different roles (Figure 2). The set of items can serve as a reference when peer reviewing an article or when informing medical decisions or policy. While the set of items is primarily derived for assessment of study reports (e.g., manuscript or published articles), it could be perceived as a starting point for researchers when they set up a study or when they report on their own research. However, given the many considerations involved in study conduct, it is advisable to consult other resources when working out a study design and analysis plan.

As systematic reviews of operative procedures increasingly include both RCTs and observational studies[5,6], it is convenient to evaluate both study types with the same set of items. Although RCTs have been described as being more internally valid than observational studies, as reflected in the traditional pyramid of evidence, it becomes increasingly apparent that randomization by design alone is insufficient as a surrogate for risk of bias[6,28,61], and revisions of the pyramid of evidence have been proposed[62]. Including both RCTs and observational studies in systematic reviews is advisable since they potentially provide complementary evidence on the effect of the studied operative intervention.

We intended to establish a set of assessment items that is easy to use with minimal loss of accuracy of the evaluation. The RoB-2 and ROBINS-I have been criticized for being time-consuming and requiring in-depth statistical knowledge, which would hinder their implementation in systematic reviews[14-16]. However, there is an evident trade-off in ease of use and rigor of the assessment. Uptake of rigorous assessment tools can be

improved both by raising awareness and training in the use of available material[63] as well as by making the existing material more accessible. Our proposal is a first step towards bridging intelligibility and scrupulosity in assessment of studies of operative interventions. We encourage further development of an assessment instrument tailored to studies of operative interventions, in particular by bringing together surgical and methodological expertise. In light of such further developments, we point out that studies of operative interventions face a methodological challenge because most studies evaluate complex interventions, but the current set of items does not explicitly address how to evaluate issues introduced by evaluations of complex interventions.

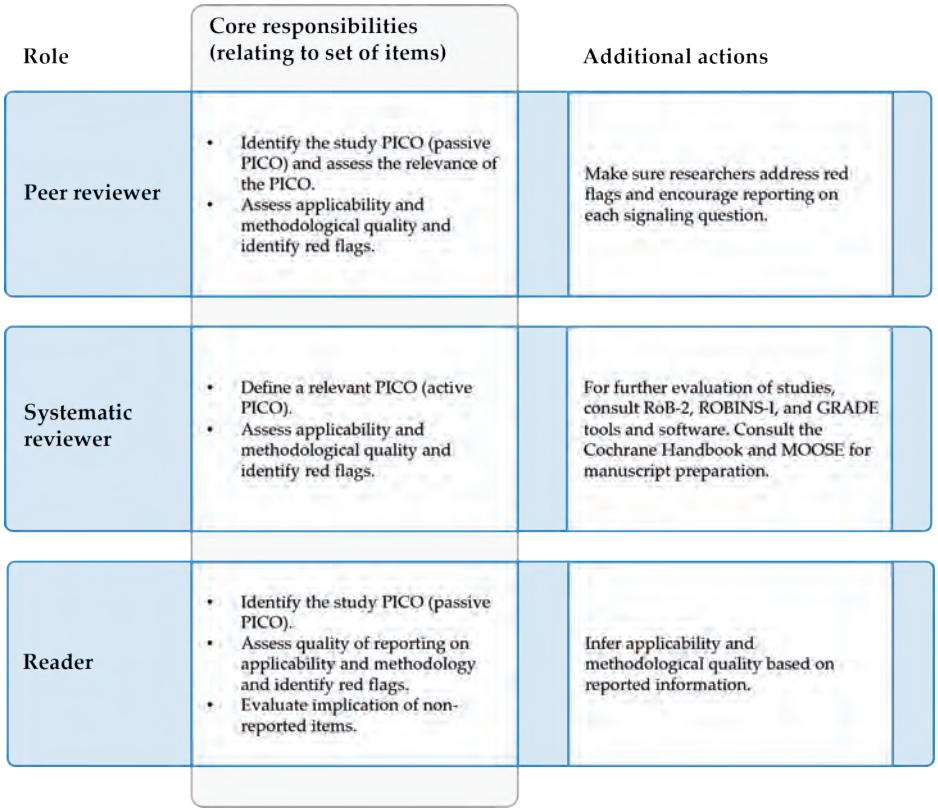


Figure 2. Schematic summary of how the concise set of items can be used by peer reviewers, systematic reviewers, and other readers appraising studies of operative interventions. The contributed value of the set of items ranges from helpful instrument to mere starting point depending on the role of the assessor. When reporting on a study it can be useful to take into account that the study can be read from these perspectives.

In an accompanying study, the set of items was applied to re-assess studies that were included in two published systematic reviews of interventions for proximal humerus factures, providing an illustration of how the proposed items can be used for assessment applicability and methodological quality of randomized and non-randomized studies into effects of operative interventions (van de Wall, forthcoming).

To conclude, the concise set of items can be used for an initial assessment of the applicability and methodological quality of randomized and non-randomized studies into effects of operative interventions. We make a call to use this set not only when performing a systematic review and meta-analysis, but to use it as a reference also when peer reviewing an article, informing medical decisions or policy, or reporting on original research.

Appendix

This file describes items and signaling questions for an assessment of applicability and methodological quality of studies of operative interventions in orthopedic trauma surgery. Researchers can decide on the scoring options for each signaling question, such as yes/no/no information or yes/possibly yes/no/possibly no/no information. Where possible, we recommend documenting quotes that explicitly address a signaling question.

PICO of the systematic review

Population: _____

Intervention: _____

Comparator: _____

Outcome: _____

Applicability

Item	Question
Population	1. Is the patient population included in the study representative of the patient population defined in the PICO of the systematic review?
	1.1. Did inclusion criteria match the patient population specified in the PICO?
	1.2. Was a relevant subgroup of participants excluded?
Intervention	2. Is the investigated intervention representative of the intervention defined in the PICO of the systematic review?
	2.1. Was the investigated intervention similar to the intervention as defined in the PICO?
	2.2. Were the participating surgeons experienced in conducting the investigated procedure?
	2.3. Was the post-operative treatment regime in the intervention arm similar to the one defined in the PICO?
Comparator	3. Is the comparator intervention representative of the comparator defined in the PICO of the systematic review?
	3.1. Was the comparator similar to the comparator as defined in the PICO?
	3.2. Were the health care professionals experienced in conducting the comparator procedure?
	3.3. Was the post-intervention treatment regime in the comparator arm similar to the one defined in the PICO?
Outcome	4. Is the outcome representative of the outcome defined in the PICO of the systematic review?
	4.1. Was the outcome measurement similar to the outcome as defined in the PICO?
	4.2. Was the timing of the outcome described and similar to the specification in the PICO?

Methodology

Item	Question
Confounding	5. Is there comparability of treatment groups, or are appropriate methods applied to correct for incomparability?
	5.1. RCT: Was the allocation sequence random?
	5.2. RCT: Was the allocation sequence concealed until participants were enrolled and assigned to interventions?
	5.3. RCT: Did baseline differences between intervention groups suggest a problem with the randomization process?
	5.4. Obs: Is there potential for confounding of the effect of the intervention in this study?
	5.5. Obs: Did the authors use an appropriate analysis method that controlled for all the important confounders?
	5.6. Obs: If 5.5. = Y or PY, were confounders that were controlled for measured adequately?
Missing data and selection bias	6. Were the patients included in the analysis representative of all patients included in the study and was the impact of missing data negligible?
	6.1. Were outcome data available for all, or nearly all, participants?
	6.2. Obs: Were intervention data available for all, or nearly all, participants?
	6.3. Obs: Were confounder data available for all, or nearly all, participants?
	6.4. If 6.1./6.2./6.3. = N or PN: were convincing arguments given for complete case analysis or were methods applied to address missing data?
	6.5. Was selection of participants into the study (or into the analysis) based on variables measured after the start of the intervention?
	6.6. Do start of follow-up and start of intervention coincide for all, or nearly all, participants?
Intervention status	7. Was intervention status correctly classified?
	7.1. Did the recorded intervention status correspond to the intervention actually received?
	7.2. Was there cross-over between interventions or non-adherence to the assigned intervention regimen that could have affected participants' outcomes?
	7.3. If 7.2. = Y or PY, was an appropriate analysis used to estimate the effect of starting and adhering to the intervention?
Outcome assessment	8. Was the outcome correctly measured?
	8.1. Was the outcome measurement a valid and reliable measurement of the outcome?
	8.2. Were outcome assessors aware of the intervention received by study participants?
	8.3. Were the methods of outcome assessment comparable across intervention groups?
Pre-specification of analysis	9. Were analyses prespecified and did the study adhere to the specified analysis plan?
	9.1. Was the analysis pre-specified, e.g., in a protocol?
	9.2. Are the reported results likely to be a selection of results of multiple analyses?
Red flags	Were there any aspects of the study or the report, **not covered by the other items**, that led to any doubt about the validity of the study? If yes, describe these in detail.

References

1. Robinson A, Johnson-Lynn S, Humphrey J, Haddad F. The challenges of translating the results of randomized controlled trials in orthopaedic surgery into clinical practice. *The Bone & Joint Journal*. 2019;101-B(2):121-123.
2. Axelrod D, Trask K, Buckley RE, Johal H. The Canadian Orthopaedic Trauma Society: lessons learned from 30 years of collaborative, high-impact research in fracture care. *The Bone & Joint Journal*. 2021;103(5):898-901.
3. Oberkofler CE, Hamming JF, Staiger RD, et al. Procedural surgical RCTs in daily practice: do surgeons adopt or is it just a waste of time? *Annals of Surgery*. 2019;270(5):727-734.
4. Chapman SJ, Aldaffaa M, Downey CL, Jayne DG. Research waste in surgical randomized controlled trials. *Journal of British Surgery*. 2019;106(11):1464-1471.
5. Houwert RM, van de Wall BJM, Groenwold RHH, Kruyt MC. A reaction to the editorial "Meta-Analyses and Systematic Reviews: JBJS Policy Revisited.". *The Journal of Bone and Joint Surgery*. 2021;103(10):849.
6. Beks RB, Bhashyam AR, Houwert RM, et al. When observational studies are as helpful as randomized trials: Examples from orthopedic trauma. *Journal of Trauma and Acute Care Surgery*. 2019;87(3):730-732.
7. Sanderson S, Tatt ID, Higgins J. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *International Journal of Epidemiology*. 2007;36(3):666-676.
8. Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar VS, Grimmer KA. A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology*. 2004;4(1):1-11.
9. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials*. 1995;16(1):62-73.
10. D'Andrea E, Vinals L, Patorno E, et al. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. *British Medical Journal Open*. 2021;11(3):e043961.
11. Sterne JA, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *British Medical Journal*. 2019;366.
12. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal*. 2016;355.
13. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. *ANZ Journal of Surgery*. 2003;73(9):712-716.
14. Jeyaraman MM, Rabbani R, Copstein L, et al. Methodologically rigorous risk of bias tools for nonrandomized studies had low reliability and high evaluator burden. *Journal of Clinical Epidemiology*. 2020;128:140-147.
15. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *Journal of Clinical Epidemiology*. 2020;126:37-44.
16. Minozzi S, Cinquini M, Gianola S, Castellini G, Gerardi C, Banzi R. Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application. *Journal of Clinical Epidemiology*. 2019;112:28-35.
17. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology*. 2011;64(4):395-400.
18. Hernán MA. Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*. 2016;26(10):674-680.

19. VanderWeele TJ. On well-defined hypothetical interventions in the potential outcomes framework. *Epidemiology (Cambridge, Mass).* 2018;29(4):e24.
20. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal*. 2008;337.
21. Fjalestad T, Hole MØ, Hovden IAH, Blücher J, Strømsøe K. Surgical treatment with an angular stable plate for complex displaced proximal humeral fractures in elderly patients: a randomized controlled trial. *Journal of Orthopaedic Trauma*. 2012;26(2):98-106.
22. Knobe M, Drescher W, Heussen N, Sellei RM, Pape H-C. Is helical blade nailing superior to locked minimally invasive plating in unstable pertrochanteric fractures? *Clinical Orthopaedics and Related Research*. 2012;470(8):2302-2312.
23. Constant C, Murley A. A clinical method of functional assessment of the shoulder. *Clinical Orthopaedics and Related Research*. 1987(214):160-164.
24. Rangan A, Handoll H, Brealey S, et al. Surgical vs nonsurgical treatment of adults with displaced fractures of the proximal humerus: the PROFHER randomized clinical trial. *Journal of the American Medical Association*. 2015;313(10):1037-1047.
25. Altman DG, Schulz KF. Concealing treatment allocation in randomised trials. *British Medical Journal*. 2001;323(7310):446-447.
26. Hernán MA, Robins JM. *Causal inference: what if.* In: Boca Raton: Chapman & Hall/CRC; 2020.
27. Beks RB, Reetz D, de Jong MB, et al. Rib fixation versus non-operative treatment for flail chest and multiple rib fractures after blunt thoracic trauma: a multicenter cohort study. *European Journal of Trauma and Emergency Surgery*. 2019;45(4):655-663.
28. Houwert RM, Beks RB, Dijkgraaf MG, Roes KC, Öner FC, Hietbrink F, Leenen LP, Groenwold RHH. Study methodology in trauma care: towards question-based study designs. *European Journal of Trauma and Emergency Surgery*. 2021;47(2):479-84.
29. Jenkinson RJ, Kiss A, Johnson S, Stephen DJ, Kreder HJ. Delayed wound closure increases deep-infection rate associated with lower-grade open fractures: a propensity-matched cohort study. *Journal of Bone and Joint Surgery*. 2014;96(5):380-386.
30. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass)*. 2009;20(4):488.
31. VanderWeele TJ. On the relative nature of overadjustment and unnecessary adjustment. *Epidemiology*. 2009;20(4):496-499.
32. Groenwold RH, Klungel OH, Grobbee DE, Hoes AW. Selection of confounding variables should not be based on observed associations with exposure. *European Journal of Epidemiology*. 2011;26(8):589.
33. Sun G-W, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*. 1996;49(8):907-916.
34. Heinze G, Dunkler D. Five myths about variable selection. *Transplant International*. 2017;30(1):6-10.
35. Altman DG, Royston P. The cost of dichotomising continuous variables. *British Medical Journal*. 2006;332(7549):1080.
36. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*. 2013;177(4):292-298.
37. Portinari M, Bianchi L, De Troia A, et al. Non-traumatic emergency abdominal surgery in nonagenarian patients: a retrospective study. *European Journal of Trauma and Emergency Surgery*. 2021:1-12.
38. Lee KJ, Tilling K, Cornish RP, et al. Framework for the Treatment And Reporting of Missing data in Observational Studies: The TARMOS framework. *arXiv preprint arXiv:200414066*. 2020.
39. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*. 2009;338.
40. Carpenter JR, Smuk M. Missing data: A statistical framework for practice. *Biometrical Journal*. 2021;63(5):915-947.
41. Groenwold RH, Dekkers OM. Missing data: the impact of what is not there. *European Journal of Endocrinology*. 2020;183(4):E7-E9.

42. Klei DS, Öner FC, Leenen LP, van Wessem KJ. No need for sternal fixation in traumatic sternovertebral fractures: outcomes of a 10-year retrospective cohort study. *Global Spine Journal*.2192568220902413.
43. Van der Meijden OA, Houwert RM, Hulsmans M, et al. Operative treatment of dislocated midshaft clavicular fractures: Plate or intramedullary nail fixation?: A randomized controlled trial. *Journal of Bone and Joint Surgery*. 2015;97(8):613-619.
44. Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable? *Injury*. 2011;42(3):236-240.
45. Banierink H, Reininga I, Heineman E, Wendt K, Ten Duis K, IJpma F. Long-term physical functioning and quality of life after pelvic ring injuries. *Archives of Orthopaedic and Trauma Surgery*. 2019;139(9):1225-1233.
46. Ochen Y, Frima H, Houwert RM, et al. Surgical treatment of Neer type II and type V lateral clavicular fractures: comparison of hook plate versus superior plate with lateral extension: a retrospective cohort study. *European Journal of Orthopaedic Surgery & Traumatology*. 2019;29(5):989-997.
47. Nauth A, Creek AT, Zellar A, Lawendy AR, Dowrick A, Gupta A, Dadi A, van Kampen A, Yee A, de Vries AC, van Otterloo AD. Fracture fixation in the operative management of hip fractures (FAITH): an international, multicentre, randomised controlled trial. *The Lancet*. 2017;389(10078):1519-27.
48. Groenwold RH, Goeman JJ, Le Cessie S, Dekkers OM. Multiple testing: when is many too much? *European Journal of Endocrinology*. 2021;184(3):E11-E14.
49. Gamble C, Krishan A, Stocken D, et al. Guidelines for the content of statistical analysis plans in clinical trials. *Journal of the American Medical Association*. 2017;318(23):2337-2343.
50. Chan A-W, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Annals of Internal Medicine*. 2013;158(3):200-207.
51. Loder E, Groves T, MacAuley D. Registration of observational studies. *British Medical Journal*. 2010;340:c950.
52. Smeeing DPJ, Houwert RM, Briet JP, et al. Weight-bearing or non-weight-bearing after surgical treatment of ankle fractures: a multicenter randomized controlled trial. *European Journal of Trauma and Emergency Surgery*. 2020;46(1):121-130.
53. Briet JP, Houwert RM, Smeeing DP, et al. Weight bearing or non-weight bearing after surgically fixed ankle fractures, the WOW! Study: study protocol for a randomized controlled trial. *Trials*. 2015;16(1):1-8.
54. Barkun JS, Aronson JK, Feldman LS, Maddern GJ, Strasberg SM, Collaboration B. Evaluation and stages of surgical innovations. *The Lancet*. 2009;374(9695):1089-1096.
55. Ergina PL, Cook JA, Blazeby JM, et al. Challenges in evaluating surgical innovation. *The Lancet*. 2009;374(9695):1097-1104.
56. McCulloch P, Altman DG, Campbell WB, et al. No surgical innovation without evaluation: the IDEAL recommendations. *The Lancet*. 2009;374(9695):1105-1112.
57. Khachane A, Philippou Y, Hirst A, McCulloch P. Appraising the uptake and use of the IDEAL framework and recommendations: a review of the literature. *International Journal of Surgery*. 2018;57:84-90.
58. Bilbro NA, Hirst A, Paez A, et al. The ideal reporting guidelines: a Delphi consensus statement stage specific recommendations for reporting the evaluation of surgical innovation. *Annals of Surgery*. 2021;273(1):82-85.
59. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*. 2008;336(7650):924-926.
60. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology*. 2011;64(4):380-382.
61. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. *British Medical Journal*. 2002;324(7351):1448-1451.
62. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *British Medical Journal Evidence-Based Medicine*. 2016;21(4):125-127.
63. Meakins JL. Evidence-based practice: new techniques and technology. *Canadian Journal of Surgery*. 2001;44(4):247-249.