low) are best formulated by an investigator who has a thorough understanding of the research topic and desires to study one specific unanswered question. The number of cases available for inclusion should be assessed and be adequate. Studies that are based on relatively few cases may be useful for evaluating rare conditions but may not lead to firm conclusions. Case-control studies are also retrospective in nature and rely heavily on previously collected data. Thus, when choosing an area of interest, it is critical that a database is easily accessible. Finally, once the study design is complete, the research team should perform an assessment of its ability to complete the study, the amount of time it will take, and whether the desired outcome will be achieved.

FORMULATING A RESEARCH QUESTION: THE PICOT FORMAT

The first step in conducting a research study is to pose a study question, and it is arguably the most important step. Spending adequate resources to develop a clear and relevant question will "determine the research architecture, strategy, and methodology."[72] The research question should be framed in a manner that is easily understood. A poorly designed question can hinder your research efforts, making it difficult for readers to interpret the results, and ultimately, jeopardize publication.

One way to enhance the organization and clarity of the research question is to use the PICOT format.[73] When using the PICOT format, one frames the study question in terms of the population of interest, the intervention, the comparator intervention, outcomes, and the time frame over which the outcomes are assessed.[73] In case-control studies, the population should be specific, addressing key eligibility criteria such as type of patient, geographic location, and qualifying disease or condition. The intervention is one or more exposure variables under investigation, and the comparator is often the absence of those factors. The outcome is the proportion of cases exposed to the variables under question compared with the controls. The data collected are usually reported as odds ratios. It is worth mentioning that the PICOT format is generally most useful for comparative studies or studies of association between exposure and outcome.[73]

Consider the following example: a researcher wants to investigate whether a traumatic anterior shoulder dislocation can increase the risk of severe shoulder osteoarthritis (OA) developing in later life. Marx et al.[74] designed an elegant, case-control study to evaluate this question. Their cases (n = 80) comprised patients who had had either a hemiarthroplasty or total shoulder arthroplasty for OA of the shoulder. They chose this group in that each had severe OA requiring replacement surgery, the diagnosis was easily confirmed at the time of surgery, and the sample of patients was easily identifiable. They excluded patients with rheumatoid disease, avascular necrosis, cuff tear arthropathy, and other systemic causes of severe shoulder pain.

Marx et al.[74] chose a group of patients undergoing total knee replacement (n = 280) for OA of the knee without OA of the shoulder as the control group because this group of patients had similar age, gender, and comorbidity distributions and was also easily identifiable. Subjects were then asked whether they had ever had a shoulder dislocation. The findings of this study were that the risk of shoulder OA developing was 19.3 times greater if there had been a shoulder dislocation in earlier life. The reader is encouraged to read this study as an elegant example of a clinical case-control study.

IDENTIFYING POTENTIAL RISK FACTORS

In general, the investigator has already identified potential risk factors, or exposures, that may have an association with the outcome when considering the study design. In the first example, Marx et al.[74] drew upon their clinical experience when asking the question about a possible association between shoulder dislocations and later development of shoulder arthritis. In other cases there may be one or more of a list of potential risk factors that may have an association with the identified outcome. In either scenario it is vital that a reasonably complete database is identified that can be easily searched for potential risk factors. It does little good to formulate a research question only to find that the necessary data are either difficult to obtain, incomplete, or simply not available.

IDENTIFYING THE CASES

When designing a case-control study, investigators begin by selecting patients with the outcome of interest, the case patients. The enrollment criteria for the case patients must be well-defined and as specific as possible. Criteria may include age, gender, and/or geographic location. The investigators must specify how the presence or absence of the desired outcome to be studied is established (e.g., clinical symptoms, physical examination findings, imaging studies, or laboratory studies).[69,75] It is preferable also to define the time period of collection because diagnostic criteria can change over time. Detailed descriptions of the case participants will aid in determining the validity of the study results.[75] For example, in a study looking at the fracture risk associated with nonsteroidal anti-inflammatory drugs, acetylsalicylic acid, and acetaminophen, the investigators identified fracture cases through the National Hospital Discharge Register in Denmark between January 1, 2000, and December 31, 2000.[70]

IDENTIFYING APPROPRIATE CONTROLS

The next step is to identify the controls—that is, the group of individuals who are reasonably similar to the cases but in whom the outcome of interest has not occurred. The controls are usually selected from the same population as the cases so that the only difference between the 2 groups is the exposure to the putative risk factors.[69,75] Similar to case assessment, the method of control selection should be clearly documented.[75] For example, in a study examining the risk of ACL tearing based on ACL volume, tibial plateau slope, and intercondylar notch dimensions as seen on MRI, investigators compared the MRI findings of 27 patients who had had a noncontact ACL injury with controls who had an intact ACL and were matched by age, gender, height, and weight.[76]

Sometimes, the rarity of the disease under investigation may limit the total number of cases identified; in these situations, statistical confidence can be increased by selecting more than 1 control per case.[69] Typically, the ratio of controls to cases should not exceed more than 4:1 or 5:1.[69]

DATA COLLECTION

Once the appropriate cases and controls have been selected, the investigators look back in time to examine the relative frequency of exposure to variables in both groups. The collection of data may involve a chart review or patient interviews. Whenever possible, data collection should be done by study personnel who are blinded to patient status—that is, whether the patient is a case or control.[77] This will limit the possibility that the information is collected differently based on patient status.[77] These data will then allow the calculation of a measure of association between the exposure variables and the outcome of interest. A flow diagram of how the data should be collected is shown in Fig 4.

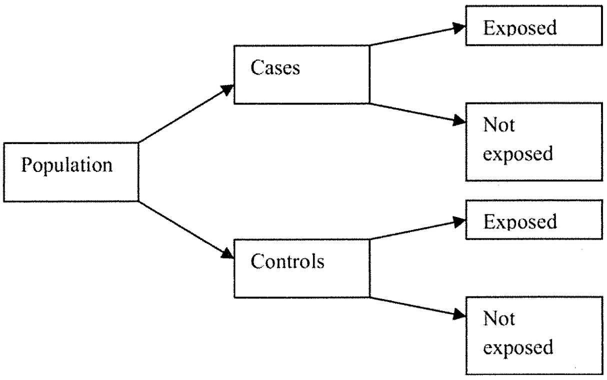


FIGURE 4. Diagrammatic representation of a case-control study. Investigators begin by identifying cases and suitable controls. In a retrospective manner, the cases are compared with controls for the presence of risk factors or past exposures.

It is important to keep track of your study's raw data throughout its progress to ensure accuracy and to strengthen the reporting of your case-control study.[73] A database of this information helps facilitate the process. The success of your study also depends on a qualified and experienced research team, because you will simply not have enough time to complete the project on your own.[69] One particularly important study personnel to include on your team is a research coordinator. This person is responsible for organizing the trial and communicating with the principal investigator, providing details on patient recruitment, data submission, and any problems experienced.[69]

INSTITUTIONAL REVIEW BOARD

Most case-control studies involve the collection of personal patient data. As a result, an approval by the institutional review board and ethics committee will likely be required before beginning your study. In general, the application usually involves specifying the details of your research, including the question, methodology, statistical analyses, and outcomes of interest. It will also request a copy of the informed consent form that will be read and signed by patients before study participation. Finally, it may ask for a description of the estimated study budget.

STATISTICAL CONSIDERATIONS

Power Analysis

Sample size is an important consideration when designing your case-control study. An appropriate

Table 9. *Basic 2 × 2 Table Illustrating How to Calculate an Odds Ratio*

	Disease	
	Yes	No
Exposure		
Yes	a	b
No	c	d

NOTE. The odds ratio is given by (a/c) ÷ (b/d) or ad/bc.

sample size ensures that your study is "powered" to detect a difference when there is one.[69] Details about the sample size calculation should be reported in the final publication as well.[75] Although information on how to calculate sample size is beyond the scope of this chapter, investigators are advised to consult epidemiologic or statistical textbooks for further details. This leads to another critical consideration when conducting a case-control study: the use of biostatisticians, who will be responsible for the appropriate statistical analyses. If necessary, involving a biostatistician early on in the planning phases of your study may be helpful.

Data Analysis

In case-control studies, a measure of association between the exposure(s) and the target outcome is usually reported as an odds ratio.[69] This refers to the odds of an event occurring in the exposed group compared with the odds of the same event happening in the unexposed group.[69] The final value can range from 0 to infinity. One is the neutral value, which means that there is no difference between the 2 groups.

Table 9 illustrates a basic 2 × 2 table. The odds ratio is given by (a/c) ÷ (b/d) or ad/bc.

For example, in a case-control study looking at the risk factors for plantar fasciitis, the investigators found an odds ratio of 3.6 for those who reported that they spent the majority of the day on their feet.[78] This indicates that the odds of weight bearing for most of the day is 3.6 times higher in patients diagnosed with plantar fasciitis than in those who do not have the disease.

LIMITING BIAS IN THE CONDUCT OF A CASE-CONTROL STUDY

As alluded to previously, case-control studies are retrospective in nature and often rely on patients' recollections to identify exposure, making them susceptible to recall bias.[69] This occurs when patients with an adverse outcome have a different likelihood of recalling past exposures than those who have not had an adverse outcome.[69] It is often difficult to limit recall bias in case-control studies. One way is to study past exposures that are objective and can be easily confirmed. If exposure data are being collected by study personnel through patient interviews, the assessors should also be blinded to the status of the patient (i.e., whether the patient is a case or a control) so that the information is not collected differently.[77] For example, Marx et al.[74] used glenohumeral dislocation as the exposure variable. They believed that it would be very unlikely for patients to incorrectly recall whether they had ever had a shoulder dislocation in the past. Furthermore, they attempted to confirm the exposure data by contacting patients and eliciting additional information such as date of dislocation, mechanism of injury, and number of recurrences.

Another important source of bias is from confounders—that is, a variable that is associated with both the exposure and the outcome. In case-control studies, the control group is selected so that it is ideally similar to the cases, except for the exposure status. However, any control group is at risk for an unequal distribution of prognostic factors compared with the cases, which can lead to biased results.[77] Careful selection of appropriate control patients is an important way to limit the effects of confounding variables. In the study by Marx et al.,[74] for example, the authors chose a group of patients who had undergone total knee arthroplasty because they were similar to the cases with respect to age, health, and mental status. They also identified prior surgery for recurrent shoulder dislocation as a potential confounding variable in the study. As a result, they conducted a subgroup analysis by excluding patients with prior surgeries, which is another way to strengthen the reporting of the case-control study.[75]

REPORTING A CASE-CONTROL STUDY

When preparing the manuscript for publication, it is important to maintain adequate transparency of your study.[75] The reporting of your case-control study should be detailed enough to allow readers to assess its strengths and weaknesses.[75] Investigators are strongly encouraged to refer to the STROBE statement[75] for further details on the reporting of observational studies to improve the overall quality of the final manuscript. Table 7 in section 5 (pp 24-25) also provides a checklist of items from the STROBE statement to include in the publication.

In the manuscript, the "Introduction" section needs to address the reasons for the study and the specific objectives and hypotheses.[75] Next, the "Methods" section should provide details on the study's processes. The goal is to provide sufficient information so that the readers can judge whether the methods were able to provide reliable and valid answers.[75] In case-control studies, it is important to document the eligibility criteria for study participants, including the method of case ascertainment and control selection.[75] All study outcomes should be clearly specified as well, including the diagnostic criteria.[75] Furthermore, you should describe the statistical methods used in the study and how the sample size was calculated.[75]

The "Results" section is a factual account of the study's outcomes, which means that it should not reflect the author's views and interpretations.[75] Data should be provided on the recruitment and description of study participants. It is also important to explain why patients may not have participated in the study or why they were excluded if applicable; this allows the readers to judge whether bias was introduced into the study.[75] The main outcomes should be documented, including the numbers in each exposure category and the statistical analyses.

In the final stages of the manuscript, the "Discussion" section addresses the issues of validity and meaning of the study.[75] A structured approach has been suggested by the STROBE statement, which involves presenting the information in the following manner: (1) summarize key findings, (2) provide possible explanations or mechanisms, (3) compare current outcomes with the results from previous studies, (4) list study limitations, and (5) specify the clinical and/or research implications of current study findings.[75]

CONCLUSIONS

In the hierarchy of evidence, case-control studies represent Level III evidence.[69] However, despite some methodologic limitations associated with case-control studies, they can be very useful in informing many research questions, particularly when they are well-designed and -reported.

Kevin Chan, M.D.
Kevin P. Shea, M.D.
Mohit Bhandari, M.D., Ph.D., F.R.C.S.C.

SECTION 7

Study Designs: Case Series, Level IV Evidence

Although RCTs provide the highest level of evidence, they are also the most expensive studies to conduct.[79] As patients become more educated, it is also more difficult to enroll patients because they are looking for specific treatments and are less willing to risk being in a control group. Even in cases where it is not clear what the best treatment is for a given patient population, patients want to be increasingly involved in an informed decision-making process and may opt for a more aggressive treatment strategy to maximize the possibility of improving function.

We cannot ignore the ethical question posed by many surgeons when considering randomized trials.[80] Is it ethical for a surgeon to offer a patient no treatment when no surgical treatment is considered inferior treatment by the surgeon? If the surgeon is uncertain whether one treatment is better than the other (clinical equipoise),[80] then patients can be enrolled. If the surgeon is certain or not completely uncertain that the treatment to be studied is superior, then ethically, the surgeon should not perform the inferior treatment (control treatment) on a patient. A case series may be the preferred type of study in this instance.

The difficulties with RCTs have resulted in more case series being performed.[80-82] In a recent statement, the editors of *Arthroscopy* acknowledged that case series are the most common type of article in their journal.[80] However, they pointed out that not all case series are alike, and when properly performed, case series can improve patient care and add to our clinical knowledge.

There are 2 common types of case series designs. These are studies with prospectively collected data with retrospective data analysis (Fig 5A) and retrospective reviews of cases (Fig 5B). Before initiating any study involving patients, approval from the institutional review board or ethics board is required.[83] For investigators in

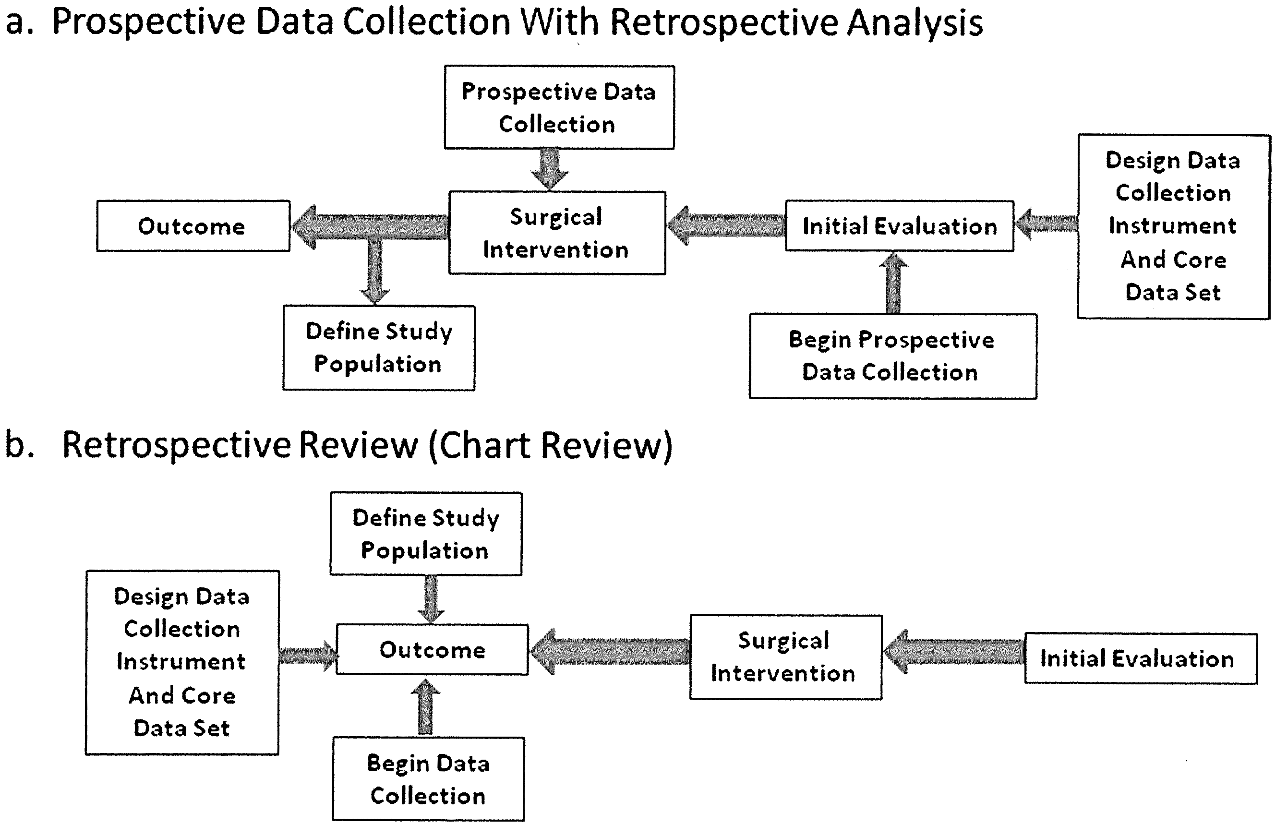

a. Prospective Data Collection With Retrospective Analysis
Prospective Data Collection
Surgical Intervention
Outcome
Define Study Population
Initial Evaluation
Design Data Collection Instrument And Core Data Set
Begin Prospective Data Collection
b. Retrospective Review (Chart Review)
Define Study Population
Design Data Collection Instrument And Core Data Set
Outcome
Surgical Intervention
Initial Evaluation
Begin Data Collection

FIGURE 5. Prospective data collection versus retrospective chart review. In a study with retrospective analysis of prospectively collected data, data collection is started preoperatively. For retrospective chart review, data are collected postoperatively.

the United States, it is also important to address Health Insurance Portability and Accountability Act privacy guidelines before initiating any study.[84] This chapter will describe each study and provide suggestions of how to improve the quality of the study. With proper planning and study design, case series can provide an important addition to the orthopaedic literature.

RETROSPECTIVE REVIEW OF CASES

A retrospective review involves gathering data that have been collected for reasons other than research. These are most commonly seen in the literature describing uncommon pathologies or procedures. By reviewing past cases, these procedures can be added together over several years and studied.

Most retrospective review studies involve review of patient charts.[82] With the advent of electronic medical records, these studies may be easier to perform and more complete than in the past. The main weakness of these studies is that there was no consistent data collection plan for all patients before the study was initiated. The quality of data is dependent on the patients' charts and dictations, which may have changed over the years of the studies. Many studies use retrospective data to determine the patient population and then prospectively collect outcome data on these patients. These studies cannot show improvements over time, but they can give a general overview of the outcome of specific procedures.

The first step in designing a retrospective review is to establish the patient population. After the specific procedure that the study will be based on has been chosen, patients are identified. Most chart reviews are done by searching billing software for specific procedure codes. This provides an overall list of patients; however, all procedures should be verified by the operative notes. It is also important to have 2 sources to search for patients. For example, studies have used billing software and the physician's personal log.[85] This is especially important if there are multiple physicians and the study covers multiple years. In addition, it should not be assumed that all physicians code their procedures the same.

Table 10. *Keys to a Quality Retrospective Chart Review*

1. Determine how study patients will be identified. It is important to identify all patients with intervention that is being studied. If only a small subset of patients is used, the data may not represent the actual outcome.
2. Define in detail the inclusion/exclusion criteria and strictly enforce them. No patient should be removed from the study unless he or she has a specific exclusion criterion.
3. Have strict guidelines for data collection. Do not allow a data point to be assumed negative just because it is not in the chart.
4. Define what is an acceptable level of complete data, as well as which data points are mandatory, before data collection.

Strict inclusion and exclusion criteria will improve the quality of the study. If patients aged under 18 years are included in the study, additional consents may be required from one's ethics committee. Gender should rarely be used as an inclusion/exclusion criterion in orthopaedic studies. A specific time frame for which patients will be included must be established. This time frame should be based on when the procedure was performed in a similar manner over time. In addition, if changes in postoperative or rehabilitation protocols were noted, this should either be noted as a data point or accounted for by the time period selected.

After the inclusion/exclusion criteria have been determined, it must be determined what data points will be collected from the charts. These data points should be points that are absolute and do not need to be interpreted by the chart reviewer (Table 10). These data points should also be points that can be expected to be found in the majority of patient charts. For instance, if knee pain is a data point, then the most consistent way to gather these data would be a yes/no selection. Many charts may list it on a scale of 1 to 10, but some may list it as mild to extreme. If data are not presented in the same format, it is better to dichotomize the data rather than trying to make 2 scales fit. In addition, reviewers should not assume values for data. For example, if pain is not mentioned, it cannot be assumed that pain is absent. This would have to be left as a blank data point. For data that may be kept by other departments (i.e., radiology), the availability of the data needs to be determined. If radiographs are necessary for the study, they must be available for all patients in the study. In some institutions radiographs may only be kept for 10 years, and old radiographs may be destroyed or archived in difficult-to-access sites.

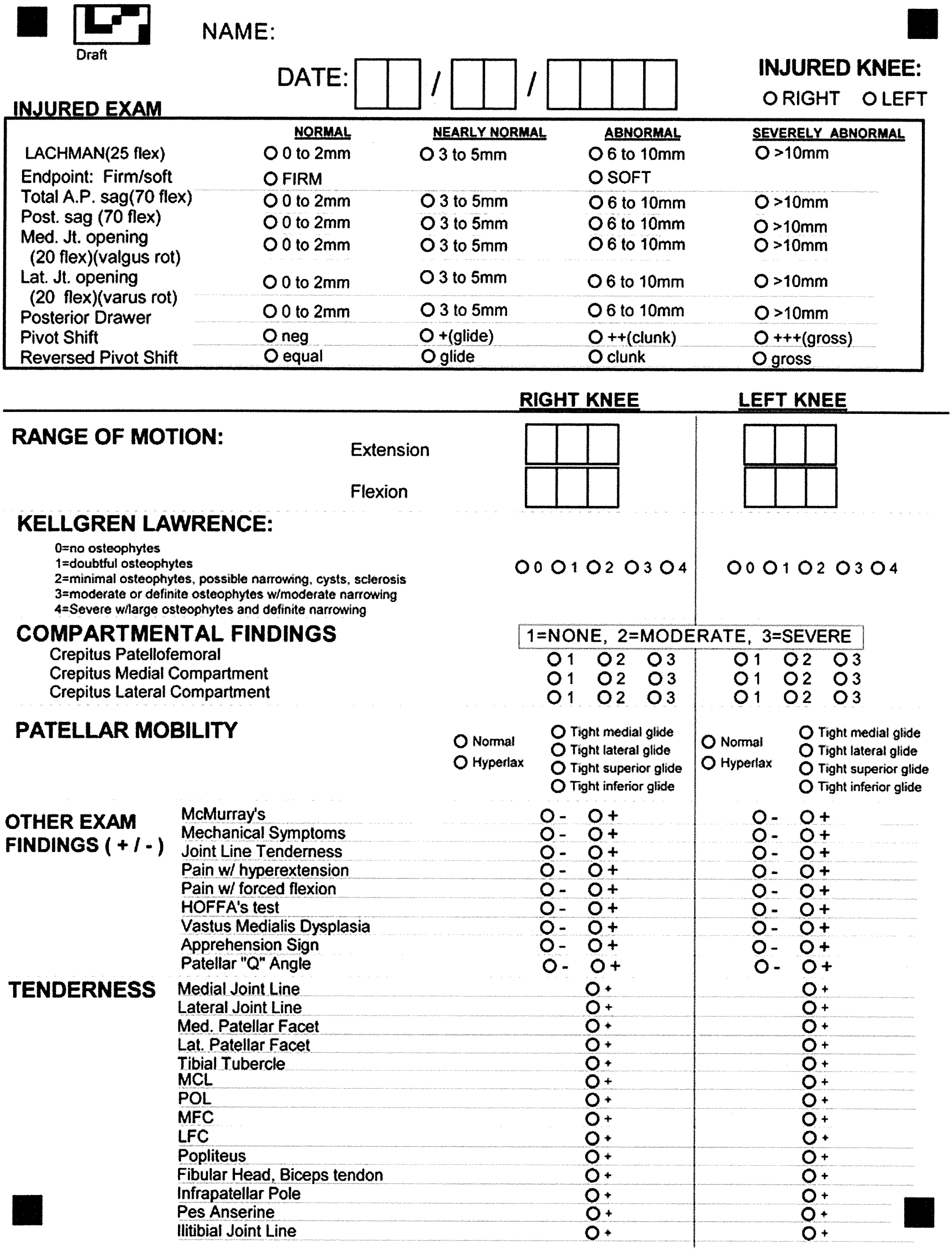
It is common in a retrospective review for one person to collect all the data from the charts. This means the data are collected consistently; however, it may also add a single reader's bias. For data collection, we suggest that a specific data collection form be designed before initiation of data collection to reduce any bias. This form will define data points and direct data collection. This will also reduce the need to interpret nonspecific data. Before starting data collection, the investigators must decide what percentage of data points are needed to include patients in the study. For example, if the data sheet has 20 data points to collect and 80% is the level of data that must be collected to be included in the study, then any patient who is missing more than 4 data points would not be included. However, some data points should be mandatory, especially if they involve the question the study aims to answer.

STUDIES WITH PROSPECTIVELY COLLECTED AND RETROSPECTIVELY ANALYZED DATA

To track longitudinal patient outcomes data, patients complete questionnaires before intervention and then at specific time points. These data are commonly stored in a research database and are considered prospectively collected. These data are collected on consecutive patients with a predetermined survey instrument that is completed by all patients. These studies suggest clinical course and response to intervention.

Data collection instruments are developed to cover all procedures done on a specific joint. For example, we have a knee arthroscopy outcome instrument (Figs 6-8), a shoulder instrument, and a hip arthroscopy instrument. All patients who are seen in the clinic complete one of these instruments. In addition, physical examination findings, surgical findings, and treatments are recorded. At defined time points, follow-up questionnaires are collected from patients. Because the study question is not designed at the beginning of data collection, the available data for the study will be limited to the specific instruments that were implemented and collected prospectively. Before starting prospective data collection, the data instruments should be carefully developed. You should develop these based on what you think will be important in 2 years, 5 years, 10 years, and beyond. Leaving off one key data element can limit the productivity of your research database. A thorough review of the literature will also help determine which data points are important.

The data instrument should be a comprehensive assessment of outcomes after a treatment, which includes a generic measure of health-related quality of life, a condition-specific measure of function, and a measure of patient satisfaction with outcome.[86,87] Any scoring sys-



Draft

NAME:

DATE: □□ / □□ / □□□□

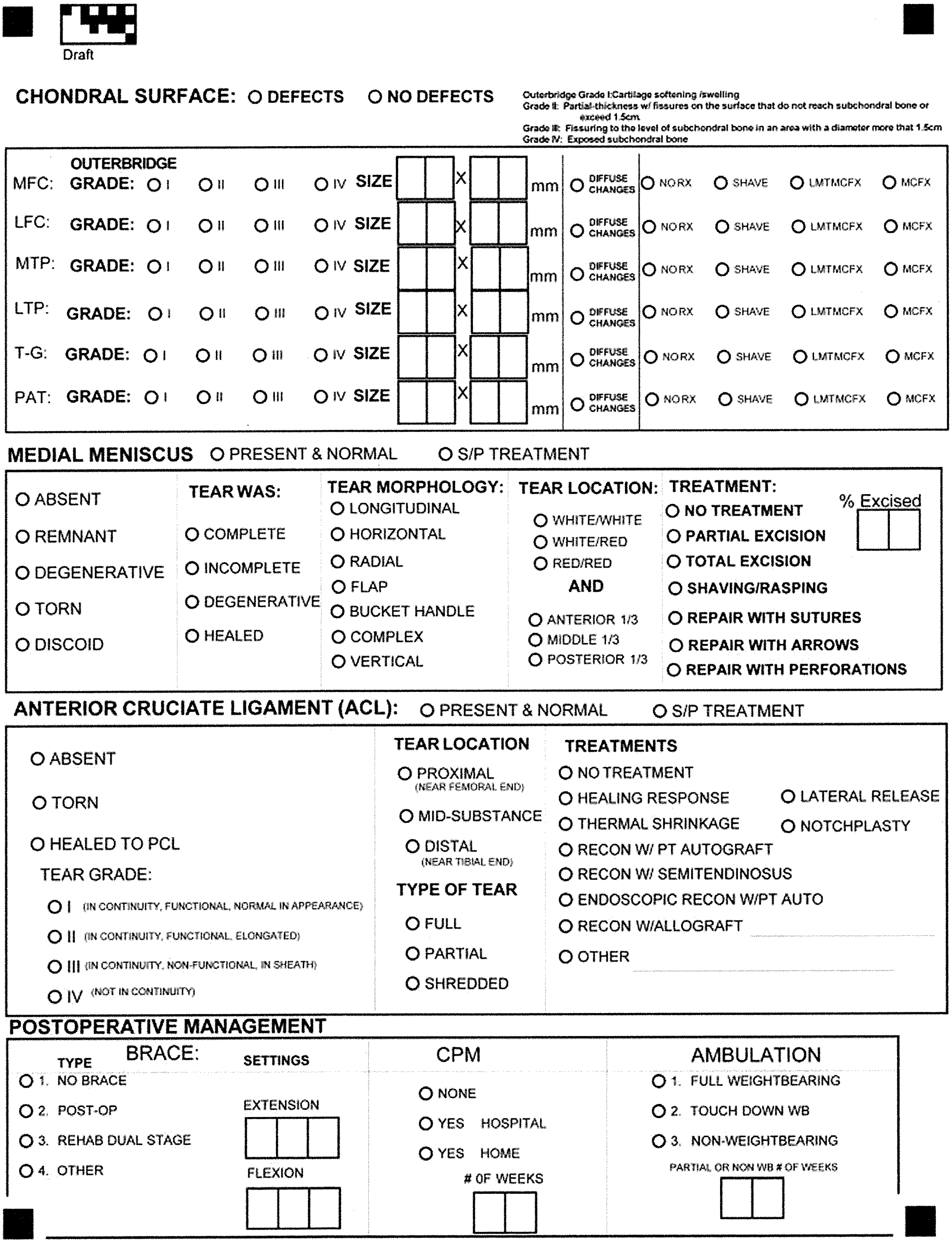
INJURED KNEE:
○ RIGHT ○ LEFT

INJURED EXAM

	NORMAL	NEARLY NORMAL	ABNORMAL	SEVERELY ABNORMAL
LACHMAN(25 flex)	○ 0 to 2mm	○ 3 to 5mm	○ 6 to 10mm	○ >10mm
Endpoint: Firm/soft	○ FIRM		○ SOFT	
Total A.P. sag(70 flex)	○ 0 to 2mm	○ 3 to 5mm	○ 6 to 10mm	○ >10mm
Post. sag (70 flex)	○ 0 to 2mm	○ 3 to 5mm	○ 6 to 10mm	○ >10mm
Med. Jt. opening (20 flex)(valgus rot)	○ 0 to 2mm	○ 3 to 5mm	○ 6 to 10mm	○ >10mm
Lat. Jt. opening (20 flex)(varus rot)	○ 0 to 2mm	○ 3 to 5mm	○ 6 to 10mm	○ >10mm
Posterior Drawer	○ 0 to 2mm	○ 3 to 5mm	○ 6 to 10mm	○ >10mm
Pivot Shift	○ neg	○ +(glide)	○ ++(clunk)	○ +++(gross)
Reversed Pivot Shift	○ equal	○ glide	○ clunk	○ gross

		RIGHT KNEE	LEFT KNEE
RANGE OF MOTION:	Extension	□□□	□□□
	Flexion	□□□	□□□
KELLGREN LAWRENCE:	0=no osteophytes 1=doubtful osteophytes 2=minimal osteophytes, possible narrowing, cysts, sclerosis 3=moderate or definite osteophytes w/moderate narrowing 4=Severe w/large osteophytes and definite narrowing	○0 ○1 ○2 ○3 ○4	○0 ○1 ○2 ○3 ○4
COMPARTMENTAL FINDINGS		1=NONE, 2=MODERATE, 3=SEVERE	
	Crepitus Patellofemoral	○1 ○2 ○3	○1 ○2 ○3
	Crepitus Medial Compartment	○1 ○2 ○3	○1 ○2 ○3
	Crepitus Lateral Compartment	○1 ○2 ○3	○1 ○2 ○3
PATELLAR MOBILITY		○ Normal ○ Hyperlax ○ Tight medial glide ○ Tight lateral glide ○ Tight superior glide ○ Tight inferior glide	○ Normal ○ Hyperlax ○ Tight medial glide ○ Tight lateral glide ○ Tight superior glide ○ Tight inferior glide
OTHER EXAM FINDINGS (+ / -)	McMurray's	○ - ○ +	○ - ○ +
	Mechanical Symptoms	○ - ○ +	○ - ○ +
	Joint Line Tenderness	○ - ○ +	○ - ○ +
	Pain w/ hyperextension	○ - ○ +	○ - ○ +
	Pain w/ forced flexion	○ - ○ +	○ - ○ +
	HOFFA's test	○ - ○ +	○ - ○ +
	Vastus Medialis Dysplasia	○ - ○ +	○ - ○ +
	Apprehension Sign	○ - ○ +	○ - ○ +
	Patellar "Q" Angle	○ - ○ +	○ - ○ +
TENDERNESS	Medial Joint Line	○ +	○ +
	Lateral Joint Line	○ +	○ +
	Med. Patellar Facet	○ +	○ +
	Lat. Patellar Facet	○ +	○ +
	Tibial Tubercle	○ +	○ +
	MCL	○ +	○ +
	POL	○ +	○ +
	MFC	○ +	○ +
	LFC	○ +	○ +
	Popliteus	○ +	○ +
	Fibular Head, Biceps tendon	○ +	○ +
	Infrapatellar Pole	○ +	○ +
	Pes Anserine	○ +	○ +
	Ilitibial Joint Line	○ +	○ +

FIGURE 6. Example of questions from a physician-completed knee objective data collection form.



Draft

CHONDRAL SURFACE: ○ DEFECTS ○ NO DEFECTS

Outerbridge Grade I:Cartilage softening /swelling
Grade II: Partial-thickness w/ fissures on the surface that do not reach subchondral bone or exceed 1.5cm
Grade III: Fissuring to the level of subchondral bone in an area with a diameter more that 1.5cm
Grade IV: Exposed subchondral bone

MFC:	OUTERBRIDGE GRADE:	○ I	○ II	○ III	○ IV	SIZE □□ X □□ mm		○ DIFFUSE CHANGES	○ NORX	○ SHAVE	○ LMTMCFX	○ MCFX
LFC:	GRADE:	○ I	○ II	○ III	○ IV	SIZE □□ X □□ mm		○ DIFFUSE CHANGES	○ NORX	○ SHAVE	○ LMTMCFX	○ MCFX
MTP:	GRADE:	○ I	○ II	○ III	○ IV	SIZE □□ X □□ mm		○ DIFFUSE CHANGES	○ NORX	○ SHAVE	○ LMTMCFX	○ MCFX
LTP:	GRADE:	○ I	○ II	○ III	○ IV	SIZE □□ X □□ mm		○ DIFFUSE CHANGES	○ NORX	○ SHAVE	○ LMTMCFX	○ MCFX
T-G:	GRADE:	○ I	○ II	○ III	○ IV	SIZE □□ X □□ mm		○ DIFFUSE CHANGES	○ NORX	○ SHAVE	○ LMTMCFX	○ MCFX
PAT:	GRADE:	○ I	○ II	○ III	○ IV	SIZE □□ X □□ mm		○ DIFFUSE CHANGES	○ NORX	○ SHAVE	○ LMTMCFX	○ MCFX

MEDIAL MENISCUS ○ PRESENT & NORMAL ○ S/P TREATMENT

○ ABSENT
○ REMNANT
○ DEGENERATIVE
○ TORN
○ DISCOID

TEAR WAS:
○ COMPLETE
○ INCOMPLETE
○ DEGENERATIVE
○ HEALED

TEAR MORPHOLOGY:
○ LONGITUDINAL
○ HORIZONTAL
○ RADIAL
○ FLAP
○ BUCKET HANDLE
○ COMPLEX
○ VERTICAL

TEAR LOCATION:
○ WHITE/WHITE
○ WHITE/RED
○ RED/RED
AND
○ ANTERIOR 1/3
○ MIDDLE 1/3
○ POSTERIOR 1/3

TREATMENT:
○ **NO TREATMENT**
○ **PARTIAL EXCISION**
○ **TOTAL EXCISION**
○ **SHAVING/RASPING**
○ **REPAIR WITH SUTURES**
○ **REPAIR WITH ARROWS**
○ **REPAIR WITH PERFORATIONS**

% Excised □□

ANTERIOR CRUCIATE LIGAMENT (ACL): ○ PRESENT & NORMAL ○ S/P TREATMENT

○ ABSENT
○ TORN
○ HEALED TO PCL

TEAR GRADE:
○ I (IN CONTINUITY, FUNCTIONAL, NORMAL IN APPEARANCE)
○ II (IN CONTINUITY, FUNCTIONAL, ELONGATED)
○ III (IN CONTINUITY, NON-FUNCTIONAL, IN SHEATH)
○ IV (NOT IN CONTINUITY)

TEAR LOCATION
○ PROXIMAL (NEAR FEMORAL END)
○ MID-SUBSTANCE
○ DISTAL (NEAR TIBIAL END)

TYPE OF TEAR
○ FULL
○ PARTIAL
○ SHREDDED

TREATMENTS
○ NO TREATMENT
○ HEALING RESPONSE
○ THERMAL SHRINKAGE
○ RECON W/ PT AUTOGRAFT
○ RECON W/ SEMITENDINOSUS
○ ENDOSCOPIC RECON W/PT AUTO
○ RECON W/ALLOGRAFT ______
○ OTHER ______

○ LATERAL RELEASE
○ NOTCHPLASTY

POSTOPERATIVE MANAGEMENT

BRACE:

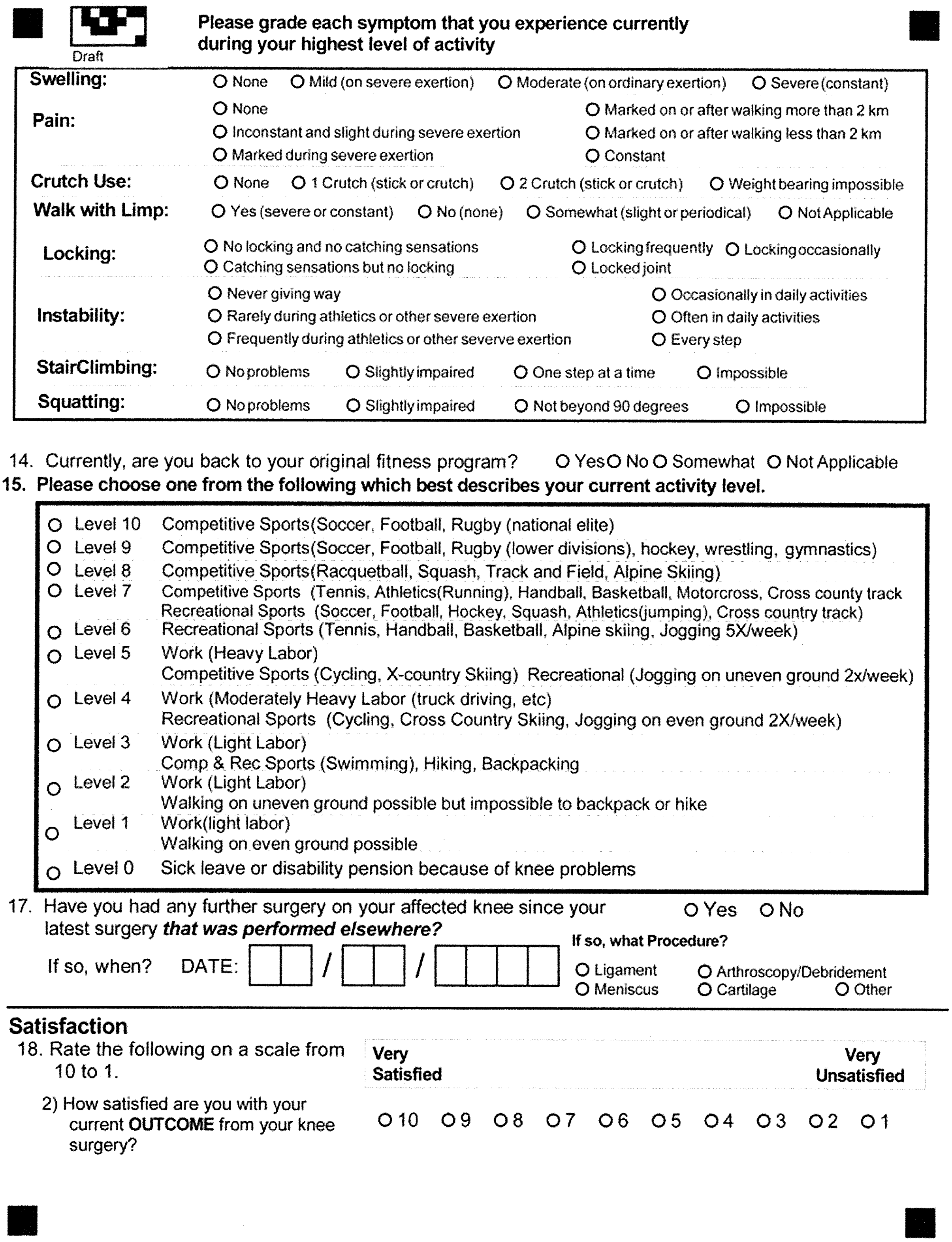
TYPE
○ 1. NO BRACE
○ 2. POST-OP
○ 3. REHAB DUAL STAGE
○ 4. OTHER

SETTINGS
EXTENSION □□□
FLEXION □□□

CPM
○ NONE
○ YES HOSPITAL
○ YES HOME
0F WEEKS □□

AMBULATION
○ 1. FULL WEIGHTBEARING
○ 2. TOUCH DOWN WB
○ 3. NON-WEIGHTBEARING
PARTIAL OR NON WB # OF WEEKS □□

FIGURE 7. Example of questions from a physician-completed knee surgery data collection form.



Draft

Please grade each symptom that you experience currently during your highest level of activity

Swelling:	○ None ○ Mild (on severe exertion) ○ Moderate (on ordinary exertion) ○ Severe (constant)
Pain:	○ None ○ Inconstant and slight during severe exertion ○ Marked during severe exertion ○ Marked on or after walking more than 2 km ○ Marked on or after walking less than 2 km ○ Constant
Crutch Use:	○ None ○ 1 Crutch (stick or crutch) ○ 2 Crutch (stick or crutch) ○ Weight bearing impossible
Walk with Limp:	○ Yes (severe or constant) ○ No (none) ○ Somewhat (slight or periodical) ○ Not Applicable
Locking:	○ No locking and no catching sensations ○ Catching sensations but no locking ○ Locking frequently ○ Locking occasionally ○ Locked joint
Instability:	○ Never giving way ○ Rarely during athletics or other severe exertion ○ Frequently during athletics or other severve exertion ○ Occasionally in daily activities ○ Often in daily activities ○ Every step
StairClimbing:	○ No problems ○ Slightly impaired ○ One step at a time ○ Impossible
Squatting:	○ No problems ○ Slightly impaired ○ Not beyond 90 degrees ○ Impossible

14. Currently, are you back to your original fitness program? ○ Yes ○ No ○ Somewhat ○ Not Applicable

15. Please choose one from the following which best describes your current activity level.

○	Level 10	Competitive Sports(Soccer, Football, Rugby (national elite)
○	Level 9	Competitive Sports(Soccer, Football, Rugby (lower divisions), hockey, wrestling, gymnastics)
○	Level 8	Competitive Sports(Racquetball, Squash, Track and Field, Alpine Skiing)
○	Level 7	Competitive Sports (Tennis, Athletics(Running), Handball, Basketball, Motorcross, Cross county track Recreational Sports (Soccer, Football, Hockey, Squash, Athletics(jumping), Cross country track)
○	Level 6	Recreational Sports (Tennis, Handball, Basketball, Alpine skiing, Jogging 5X/week)
○	Level 5	Work (Heavy Labor) Competitive Sports (Cycling, X-country Skiing) Recreational (Jogging on uneven ground 2x/week)
○	Level 4	Work (Moderately Heavy Labor (truck driving, etc) Recreational Sports (Cycling, Cross Country Skiing, Jogging on even ground 2X/week)
○	Level 3	Work (Light Labor) Comp & Rec Sports (Swimming), Hiking, Backpacking
○	Level 2	Work (Light Labor) Walking on uneven ground possible but impossible to backpack or hike
○	Level 1	Work(light labor) Walking on even ground possible
○	Level 0	Sick leave or disability pension because of knee problems

17. Have you had any further surgery on your affected knee since your latest surgery *that was performed elsewhere?* ○ Yes ○ No

If so, when? DATE: □□ / □□ / □□□□

If so, what Procedure?
○ Ligament ○ Arthroscopy/Debridement
○ Meniscus ○ Cartilage ○ Other

Satisfaction

18. Rate the following on a scale from 10 to 1.

Very Satisfied **Very Unsatisfied**

2) How satisfied are you with your current **OUTCOME** from your knee surgery?

○10 ○9 ○8 ○7 ○6 ○5 ○4 ○3 ○2 ○1

FIGURE 8. Example of questions from a patient-completed knee subjective data collection form.

TABLE 11. *Keys to Prospectively Collected Data and Retrospective Analysis Studies*

1. When building your data collection instruments for prospective data collection, use validated outcome scores including a condition-specific score, a quality-of-life scale, and a measure of patient satisfaction.
2. Maintain your database. Set rules and protocols to ensure quality data are maintained.
3. We recommend that patient follow-up forms be mailed on an annual basis. You may only need 2-year and 10-year follow-up, but annual data collection allows you to keep in touch with patients and also provide data on improvement or decline over time.
4. Identify the study group from the database using inclusion and exclusion criteria. If patients refuse to participate, this must be respected.

tem that will be used as part of the data instruments must have been tested to determine whether it can measure change after an intervention. Because the instruments are picked before data collection, it is very important that a valid, reliable, and responsive score is used to collect data (see section 9). If an untested score is used, when the data are analyzed and the results are poor, one will not know whether patients did not improve or whether the instrument was not able to detect the improvement over error. Valid and reliable questionnaires will ensure quality data collection.

If a database is set up for the collection of data over a series of years, steps must be taken to ensure that the data are of the highest quality (Table 11). Standard operating procedures for data collection, data entry, and data verification must be developed and implemented. In addition, Health Insurance Portability and Accountability Act guidelines must be followed in the United States when collecting and storing data.[84] Data audits should also be performed annually or every other year. For a database to be useful, it must be filled with accurate, quality data.

Based on what data have been collected, a study can be designed. Just as in a retrospective study, the inclusion/exclusion criteria are crucial in these studies. However, in studies where all data are collected prospectively, incomplete data should not be exclusion criteria. In addition, with large numbers, more variables can be studied. With regression analysis of large groups, independent predictors of the outcome can be determined.

Once the inclusion/exclusion criteria have been determined and the variables of interest are determined, the database can be queried to extract the data. A control group may also be identified by use of the same inclusion/exclusion criteria but with a previously performed technique or with surgery that did not include the procedure of interest. Care must be taken when identifying a control group. It must be of equal trauma to the patient, equal recovery, and equal rehabilitation. If the control group does not quite match up, it is better to proceed without a control group.

When data have been queried and put into a spreadsheet, continuous data should be analyzed for normal distribution. After this, data can be analyzed by use of the proper statistical tests.[88] These tests will be discussed in further chapters. We encourage all researchers who are starting to perform clinical studies to obtain input from a statistician. It is also helpful to have an independent statistician review all data analysis at the completion of the study.

PRESENTATION OF DATA

When presenting data from these studies, it is crucial to fully describe how data were collected. There are many examples in the literature of studies that are described as retrospective reviews but it is unclear how the patients were identified and how additional data were obtained. Readers are more likely to consider your study if it is easy to understand the study design. It is also very important to include the numbers of patients in the study. This should start with the total number of patients who had the procedure completed. Then, the number of patients who fit the inclusion/exclusion criteria should be listed. Failures should also be reported. Failures should be adequately defined. If patients are followed up to an endpoint, this should also be defined. If it is unclear why patients are considered failures, then it will be difficult for readers to understand the study outcome. Regarding follow-up, the number of patients available for follow-up should be reported. Then, the percentage of those patients in whom follow-up was obtained should be reported. In some studies, having follow-up is considered an inclusion criterion. If the readers do not know whether these data are on 80% of the patients or on 30% of the patients, then it is again difficult for them to interpret the study outcome.

CONCLUSIONS

Case series are common in the literature today (Table 12). Many of these studies use prospectively collected data, which increases the quality of these studies. As more physicians begin to monitor their patients

TABLE 12. *Summary*

1. Case series, when done properly, are important additions to the literature.
2. Prospective data collection allows for quality research with minimum selection bias. It also allow physicians to track all of their patients over time. This provides a means of patient feedback and improving patient care.

for quality-of-care purposes, more of these prospective database studies will be completed. If these studies are well-designed and well-executed and the analysis is done properly, then they provide important information to the literature. Depending on the individual clinical setting, this type of study could become the research study of choice.

Karen K. Briggs, M.P.H.
Robert F. LaPrade, M.D., Ph.D.

SECTION 8

Special Designs: Systematic Reviews and Meta-Analyses

Health care professionals are increasingly required to base their practice on the best available evidence derived from research studies. However, these studies may vary in quality and produce conflicting results. It is therefore essential that health care decisions are not based solely on 1 or 2 studies but rather take into account the range of information available on that topic.[89] Health care professionals have traditionally used review articles as a source of surmised evidence on a particular topic because of the explosion of medical literature and scarcity of time.

Review articles in the medical literature are traditionally presented as "narrative reviews," in which experts in a particular field provide a summary of evidence. There are several key disadvantages to the use of traditional narrative reviews. The validity of a review article is dependent on its methodologic quality.[89] Authors of narrative reviews often use informal, subjective methods to collect and interpret studies and are therefore prone to bias and error.[90] Reviewers can disagree on issues such as what types of studies to include and how to balance the quantitative evidence they provide. Selective inclusion of studies to reinforce preconceived ideas or promote the author's view on a topic also occurs.[89,90] Furthermore, traditional reviews often ignore sample size, effect size, and research design and are rarely explicit about how studies are selected, assessed, and analyzed.[90] In doing so, they do not allow readers to assess the presence of potential bias in the review process.[90]

In contrast to narrative reviews, systematic reviews apply "scientific strategies in ways that limit bias to the assembly, a critical appraisal, and synthesis of relevant studies that address a specific clinical question."[89]

WHAT IS A SYSTEMATIC REVIEW?

Systematic reviews are scientific investigations conducted with a specific methodology using independent studies as "subjects."[91] They synthesize the results of multiple primary investigations using established strategies aimed at limiting random error and bias.[91] Strategies include a comprehensive search of relevant articles using explicitly defined and reproducible criteria. In a systematic review, primary research designs and study characteristics are appraised, data are synthesized, and results are interpreted.[91]

Systematic reviews can be quantitative or qualitative in nature. In a qualitative systematic review, results of primary studies are summarized without being statistically combined. Quantitative reviews, on the other hand, are known as meta-analyses, which use statistical methods to combine the results of 2 or more studies.[91]

Current evidence-based practice guidelines are based on systematic reviews appropriately adapted to local circumstances and values. Economic evaluations compare the costs and consequences of different courses of action. The knowledge of consequences available for these comparisons is often generated by systematic reviews of primary studies.[91] In this manner, systematic reviews play a key role in clinical decision making by allowing for an objective appraisal of knowledge accumulated from the robust and

Table 13. *Features of a Systematic Review*

Key Points
Systematic reviews address a specific topic or problem
Systematic reviews assemble, critically appraise, and synthesize results of primary studies
Systematic reviews are prepared using explicit methods that limit bias and random error
Systematic reviews can help clinicians keep abreast of the overwhelming amount of medical literature
Systematic reviews can help predicate clinical decisions on research evidence
Systematic reviews are often more efficient and accurate than single studies

NOTE. Adapted from Cook et al.[91]

increasingly productive search for solutions to medical problems.[90] The features of a systematic review are listed in Table 13.

RATIONALE FOR CONDUCTING SYSTEMATIC REVIEWS

Quantity of Information

Over 2 million articles are published annually in the biomedical literature.[92] Decision makers of various types are inundated with an unmanageable amount of information. Systematic reviews are needed to refine this cumbersome amount of information. Practitioners and clinicians can use systematic reviews in place of an overwhelming volume of medical literature to keep informed.[91] In addition, through critical exploration, evaluation, and synthesis, systematic reviews are able to separate insignificant and unsound medical information from salient critical studies that should be incorporated into the clinical decision-making process.[92]

Integration

Systematic reviews integrating critical biomedical information are used by various decision makers. Research investigators need systematic reviews to summarize existing data, refine hypotheses, estimate sample sizes,[91] recognize and avoid pitfalls of previous investigations, and describe important secondary or adverse effects and covariates that may warrant consideration in future studies.[92] Without systematic reviews, researchers may miss promising leads or embark on studies inquiring into questions which have been previously answered.[91] Information encompassed within systematic reviews is also used by health policymakers to formulate guidelines and legislation regarding the use of certain diagnostic tools and treatment strategies as well as optimizing outcomes using available resources.[91,92] As previously discussed, systematic reviews are used by clinicians. Single studies rarely provide definitive answers to clinical questions. Systematic reviews can help practitioners solve specific clinical problems by ascertaining whether findings can be applied to specific subgroups, as well as keeping practitioners literate in broader aspects of medicine.[91,91] Lastly, systematic reviews shorten the time between medical research discoveries and clinical implementation of effective diagnostic or treatment strategies.[92]

Efficiency

Conducting a systematic review is usually more efficient, less costly, and quicker than embarking on a new study. It can also prevent pursuing research initiatives that have already been conducted.[92] Lastly, pooled results from various studies can give a better estimate of outcomes.

Generalizability

By using different eligibility criteria for participants, definitions of disease, methods of measuring exposure, sample sizes, populations, study designs, and variations of a treatment, multiple studies addressing the same question provide an interpretative context not available in any individual study.[92] Pooled results from these studies are more generalizable to the population than any individual study.[92]

Consistency

Systematic reviews can determine consistency among studies of the same intervention or among different interventions. Assessments of whether effects are in the same direction or of the same magnitude can also be made. Lastly, systematic reviews can help ascertain consistency of treatment effects across different diseases with a common underlying pathophysiology and consistency of risk factors across study populations.[92]

In addition to establishing consistencies, systematic reviews can be used to assess inconsistencies and conflicts in data.[92] Effectiveness of treatments in particular settings or only among certain subjects can be explored and assessed. Furthermore, findings from certain studies that stand alone because of uniqueness of the study population, study quality, or outcome measure can be explored.[92]

Increased Power and Precision: One of the most commonly cited reasons for conducting systematic reviews is the increase in power. Meta-analyses and pooled results yield increased statistical significance by increasing the sample size. The advantage of increasing power is particularly relevant to conditions of relatively low event rates or when small effects are being assessed.[92] Quantitative systematic reviews also allow for increased precision in estimates of risk or effect size. Meta-analyses show that increasing sample size from temporally consecutive studies results in a narrowing of confidence intervals.[92,93]

Accuracy: In contrast to traditional views, systematic reviews apply explicit scientific principles aimed at reducing random and systematic errors of bias and therefore lead to better and more accurate recommendations.[91] Furthermore, the use of explicit methods allows for an assessment of what was done and yields a better ability to replicate results in the future and understanding of why results and conclusions of reviews differ.

ELEMENTS OF A SYSTEMATIC REVIEW

A review is considered "systematic" if it is based on a clearly formulated question, identifies relevant studies, appraises the studies' quality, and summarizes evidence using an explicit and predetermined methodology (Table 13).

Step 1: Framing the Research Question

A good systematic review has a well-formed, clear question that meets the FINER (feasible, interesting, novel, ethical, and relevant) criteria.[94] Feasibility of the question is largely dependent on the existence of a set of studies that can be used to evaluate the question. The research question should describe the disease or condition of interest, the population, the intervention and comparison treatments, and the outcome(s) of interest.[94,95]

Step 2: Identifying Relevant Publications

Systematic reviews are based on a comprehensive and unbiased search of completed studies.[96] To capture as many relevant citations as possible, a wide range of medical, environmental, and scientific databases should be searched.[95] The Center for Review and Dissemination has compiled a comprehensive resource list for researchers undertaking systematic reviews.[97] The process for identifying studies to be included in the review and the sources for finding these studies should be established before conducting the review, such that they can be replicated by other investigators. Depending on the subject matter, MEDLINE, AIDSLINE, CINAHL, EMBASE, and CANCERLIT, among other databases, can be used. In addition, a manual review of the bibliographies of relevant published studies, previous reviews, evaluation of the Cochrane Collaboration database, and consultation with experts can also be undertaken.[94]

Criteria for Including and Excluding Studies: Before one conducts a systematic review, a rationale should be provided for including or excluding studies. Criteria for including or excluding studies typically specify the period in which the studies were published, the targeted population, the disease or condition of interest, the intervention of interest, acceptable control groups, an accepted length of loss to follow-up, required outcomes, and whether blinding should be in place. Though these are typical, other criteria can also be specified.[94] The criteria for inclusion and exclusion should be established before conducting the review.[95]

Once the criteria are established, each potentially eligible study should be reviewed for eligibility independently by 2 examiners. Any discrepancies should be settled by a third examiner or by consensus between the 2 examiners.[94] When determining eligibility, the examiners should be blinded to the dates of publication, authors of the study, and results to ensure an unbiased selection.[94]

Collecting Data From Eligible Studies: Predesigned forms should be created, which include variables such as eligibility criteria, design features, population included in the study, number of individuals in each group, intervention, and primary and secondary outcomes, as well as outcomes in subgroups.[94] The data should be abstracted individually by 2 independent assessors. As with the inclusion and exclusion of studies, if the 2 assessors disagree, a third assessor should settle the discrepancy or a consensus process may be used.[94]

Often, it is difficult to ascertain whether studies are eligible because published reports may or may not adequately describe important information such as design features, risk estimates, and standard deviations.[94] It is usually not appropriate to calculate risk estimates and confidence intervals based on crude data from observational studies because sufficient information may not be available for potential confounders. To attain adequate information, efforts should be made to contact the investigators and retrieve necessary information.[94]

Step 3: Assessing Study Quality

The greatest drawback to a systematic review is that the results can be no more reliable than the quality of the studies on which they are based.[94] If individual studies are of poor quality, this poses a significant risk to the overall quality of the systematic review. A simple procedure to ensure this is to create relatively strict criteria for good study design when establishing the inclusion and exclusion criteria. This is of particular importance when using observational studies. It is often difficult to conduct RCTs in evaluating public health interventions at the community level.[95] Therefore systematic reviews assessing the safety of such interventions need to include evidence from a broader range of study designs.[95] When using data from observational studies, results should be adjusted for potential confounding variables to ensure that results of meta-analyses are not confounded.[94]

Quality is a multidimensional concept that can relate to design, conduct, and analysis of the trial. Quality of a primary investigation can be affected by the presence of bias, which consequently affects internal validity. Assessing the quality of the studies included is currently debated.[98] Quality scores can combine information on several features in a single numerical value. Numerous quality checklists exist. However, caution must be exercised in their application because scores, and thus quality estimates, may differ across varying checklists. On the other hand, a component approach examines key dimensions individually.[98]

Incorporating study quality into meta-analysis can entail excluding trials that fail to meet some standard of quality. Although this may be justified, it can also lead to excluding studies that may contribute valid information.

Step 4: Meta-analysis—Summarizing the Evidence

Once all studies to be included have been identified and the data abstracted, a summary estimate and confidence interval may be calculated.[94] Methods for calculating the summary estimate and confidence interval, as well as principles of meta-analyses, are discussed in the next section. It is important to note that different approaches to calculating these estimates will yield different results.

Step 5: Presenting the Findings

Three types of information are typically included in systematic reviews. First, characteristics of each study are presented in tables. These often include study sample size, number of outcomes, length of follow-up, methods used in the study, and characteristics of the population studied. Second, results of individual studies are displayed. These can include risk estimates, confidence intervals, or *P* values.[94] Finally, the meta-analysis summary estimate, confidence interval, and subgroup and sensitivity analyses are presented. All information should be presented clearly in tables and figures.

META-ANALYSIS

Principles

After a systematic review, data from individual studies may be pooled quantitatively by use of established statistical methods. A useful definition of meta-analysis is given by Huque as "a statistical analysis that combines or integrates the results of several independent clinical trials considered by the analyst to be 'combinable.'"[90] The rationale for conducting meta-analysis is that combining individual studies provides an increased sample size, which improves the statistical power of the analysis and the precision of the estimates of treatment effects.[89]

Meta-analysis is a 2-stage process. First, it involves calculation of a measure of treatment effect with its 95% confidence interval for each individual study. This is accomplished by use of summary statistics such as odds ratios, relative risks, and risk differences. Second, an overall treatment effect is calculated as a weighted average of the individual summary statistics.[89] It should be noted that data from individual studies are not simply averaged. Instead, results are weighted. Higher weight is given to studies that provide more information.[89]

Heterogeneity

Combining the results of individual studies may not be appropriate if the results differ greatly. There are several ways to ascertain whether the results are heterogeneous and therefore inappropriate to combine.[99] First, individual studies can be reviewed to determine whether there are substantial differences in the study design, study population, interventions, or outcomes.[94] Second, the investigator can examine the results of individual studies; if some trials report a benefit whereas others report a significant harm, then the results are most likely heterogeneous. Statistical approaches exist to facilitate the establishment of heterogeneity in results.[94]

Tests of homogeneity assume that the results of individual studies are the same (the null hypothesis).

The test is used to determine whether the data refute this null hypothesis (the alternate hypothesis). A χ^2 test is commonly used. If the P value is greater than or equal to .10, then the data support the null hypothesis and the studies are homogeneous. If the P value is less than .10, then the null hypothesis is rejected and the study findings are considered to be heterogeneous. All meta-analyses should report the P value.[94,100]

If this test shows homogeneous results, then the differences between the studies can be attributed to sampling variation. In this case a fixed-effects model is used to combine the results. If the test indicates that heterogeneity exists between study results, then a random-effects model should be used to combine results.[99,100]

A major limitation to this approach is that statistical tests often lack power to reject the null hypothesis and studies appear to be homogeneous when they are not. There is no statistical solution to this problem.[94] Therefore a discussion of heterogeneity and its potential effects should always accompany summary estimates.[94,100]

Methods

Treatment Effects: The primary goal of meta-analysis is to calculate a summary effect size. If the outcome is binary (e.g., disease *v* no disease), then odds ratios or relative risks should be used. If the outcome is continuous (e.g., blood sugar measurement), then mean differences should be used.[89,100]

Odds ratio is defined as "the ratio of the odds of the treatment group to the odds of a control group."[89] Odds are calculated by dividing the number of patients in the group who achieve a certain endpoint by the number of patients who do not. Risk, in contrast to odds, is calculated as the number of patients in the group who achieve the stated endpoint divided by the total number of patients in the group.[89] Relative risk is the ratio of the 2 risks. An odds ratio or relative risk greater than 1 indicates increased likelihood of the stated outcome being achieved in the treatment group. Correspondingly, a relative risk or odds ratio of less than 1 indicates decreased likelihood of outcome in the treatment group. A ratio of 1 indicates no difference between the 2 groups. All estimates of relative risk and odds ratio should be accompanied by confidence intervals.[89]

Fixed- Versus Random-Effects Models: There are various statistical approaches to calculate a summary effect. These approaches are thoroughly discussed by Cooper and Hedges.[101] The choice of statistical method is dependent on the outcome measure and presence of heterogeneity. The fixed-effects model calculates the variance of a summary estimate based on the inverse of the sum of weights of each individual study.[94] The random-effect model adds variance to the summary effect in proportion to the variability of the results of the individual studies.[94] The confidence interval around the summary measure is usually greater in the random-effects model, and therefore the summary effects are less likely to be significant. Many journals now require authors to use the random-effects model because it is considered the most conservative.[94] It is also quite reasonable to use both a random- and fixed-effects model and present both estimates.

Confidence Intervals: Confidence intervals should accompany each summary measure. Intervals are commonly reported with 95% confidence but can be reported with 90% or 99% confidence.[89] A 95% confidence interval is the range within which the true treatment effect will lie with 95% certainty. The width of a confidence interval dictates precision; the wider the interval, the less the precision.

Many formulas exist to calculate the variance of summary risk estimates. The variance of the summary estimate is used to calculate the 95% confidence interval around the summary estimate ($\pm 1.96 \times \sqrt{}$variance).[94]

Assessment of Publication Bias

Publication bias occurs when published studies are not representative of all studies that have been conducted.[94] If reasons that studies remain unpublished are associated with their outcome, then meta-analyses combining the published results will be seriously biased. Hypothetically, with a treatment that has no actual effect on a disease of interest, studies that show a benefit may be published whereas studies that suggest harm may not be published. In this case a meta-analysis combining only published results would depict a beneficial impact.[102]

There are 2 main ways to circumvent the effects of publication bias. First, unpublished studies should be identified and included in the summary estimate. Unpublished studies can be identified by contacting investigators in the field and reviewing abstracts, meeting presentations, and doctoral theses. However, including unpublished studies can be problematic. It is often difficult to identify unpublished studies, and when identified, it is often difficult to extract relevant data, such as inclusion and exclusion criteria, or determine the quality of methods.[94] Efforts should be made, in these circumstances, to contact the investigators.

The extent of potential publication bias can be estimated. This estimate can then be reflected in the systematic review's conclusions. Publication bias exists when unpublished studies yield different results from published studies. Unpublished studies are likely to be smaller than published studies and likely to have found no association between risk factor or intervention and the outcome of interest. If there is publication bias, there should exist an association between a study's sample size (or the variance of the outcome estimate; smaller studies tend to have larger variance) and findings. This association can be measured by use of the Kendall τ.[94] A strong correlation between sample size and findings would suggest a publication bias.[94]

Alternatively, a funnel plot can also be indicative of publication bias. In the absence of publication bias, a plot of the standard error versus log of outcome measure (i.e., odds ratio and relative risk) should have a funnel or bell shape (Fig 9A).[103] When publication bias is present, the plot is asymmetrical and truncated in a corner (Fig 9B).[103]

When substantial publication bias is present, summary estimates should not be calculated. If little publication bias is present, summary estimates should be interpreted with caution. All meta-analyses should contain a discussion of potential publication bias and its effect on the summary estimates presented.[94,102]

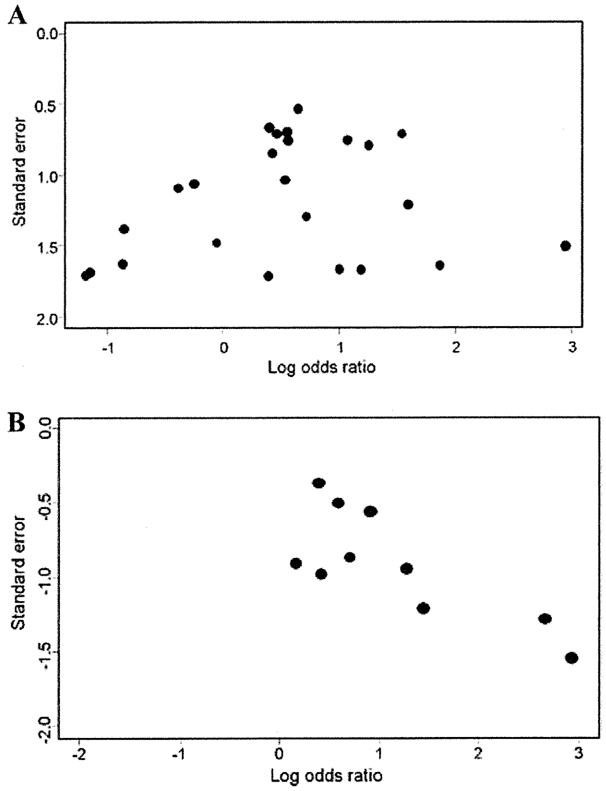


FIGURE 9. (A) Funnel plot that does not suggest publication bias.[22] (B) Funnel plot suggestive of publication bias.[22]

Subgroup and Sensitivity Analyses

Subgroup Analysis: The main aim of a meta-analysis is to produce an estimate of the average effect seen in trials of a particular treatment. The direction and magnitude of this overall effect are intended to guide clinical decision making. However, practitioners are presented with a problem when asked to use an average effect on specific groups of patients because the effect of a given treatment is likely to be different across different groups of patients.[104] It may, therefore, be possible to use data from all of the studies or some subset of the studies included in the systematic review.[94] Although meta-analyses offer a reliable basis for subgroup analyses, they are not exempt from bias and the results of such analyses should always be interpreted with caution.[104]

Sensitivity Analysis: The robustness of findings of a meta-analysis should be examined through sensitivity analyses.[100] An analysis can entail an assessment of the influence of methodologic quality and the presence of publication bias.[102] Quality summary scores or categorical data on individual components can be used to explore the methodologic quality. Simple stratified analyses and meta-regression models are useful for exploring associations between outcome effects and study characteristics.[98]

IMPLEMENTATION AND COCHRANE COLLABORATION

Despite the considerable amount of resources spent on clinical research, relatively little attention has been given to ensuring that the findings of research are implemented in routine clinical practice.[105] There are many strategies for intervention that can be used to promote behavioral changes among health practitioners (Table 14).

The Cochrane Collaboration, an international organization, has facilitated informed decision making in health care by preparing, maintaining, and promoting the accessibility of systematic reviews on the effects of health care interventions.[106] These reviews are available in the Cochrane handbook, updated and modified in response to new evidence.[106] Because Cochrane reviews have greater methodologic rigor and are more frequently updated than systematic reviews published in paper-based journals, they present an excellent resource to be used in clinical practice.

CONCLUSIONS

Systematic reviews are scientific investigations conducted with a specific methodology using independent

TABLE 14. *Interventions to Promote Behavioral Change Among Health Professionals*[105]

Consistently effective interventions	
	Educational outreach visits
	Computerized or manual (e.g., mail) reminders
	Multifaceted interventions (a combination of audit and feedback, reminders, local consensus processes, and marketing)
	Interactive educational meetings (participation of health care practitioners in seminars or workshops that include discussions)
Interventions of variable effectiveness	
	Audit and feedback (e.g., summary of clinical performance)
	Local consensus processes (focus groups, discussion with experts and local practitioners)
	Patient-mediated interventions (aimed at changing performance of health care providers)
Interventions that have little or no effect	
	Educational materials (distribution of information pamphlets or best practice guidelines)
	Didactic educational meetings

NOTE. Adapted from Bero et al.[105]

studies. They appraise study characteristics, synthesize the results of multiple primary investigations using established strategies, and interpret results in a manner aimed at limiting bias. Strategies include a comprehensive search of relevant articles using explicitly defined and reproducible criteria. A detailed explanation of the steps involved in conducting a systematic review was discussed in this chapter. Systematic reviews often use meta-analyses to combine results of the eligible studies to increase sample size, which improves the statistical power of the analysis and the precision of the estimates of treatment effects. Meta-analysis is a 2-stage process, which involves calculating a treatment effect with its 95% confidence interval for each individual study and, if appropriate, calculating an overall treatment effect as a weighted average of the individual summary statistics. The specifics of conducting a meta-analysis were also discussed.

Zahra N. Sohani, M.Sc.
Jón Karlsson, M.D., Ph.D.
Mohit Bhandari, M.D., Ph.D., F.R.C.S.C.

SECTION 9

Special Designs: Survey and Reliability Studies

Survey and reliability studies are a valuable element of outcomes research. With survey studies, it is very important to determine which questions to ask. Each question should be viewed as a possible measure of general or joint-specific health, and each data point collected should be potentially valuable in addressing the research question about how the patient is progressing.[107] This is an important first step to addressing what questions should be on future survey studies so that these surveys can contribute valuable information to patient assessment. The psychometric properties of a survey should also be established.[107-110] This will determine whether the questions are valid, reliable, and responsive. If a survey is not tested for these parameters, it will be unclear whether the survey is measuring what it is supposed to measure, whether it is accurate, and whether it can measure change. Without these, the results of the survey may come under question.

Reliability studies help determine what are accurate and consistent measurements and whether these measures can be consistently interpreted.[111,112] These studies are commonly done on radiographic measurements. The reliability of a radiographic measurement allows for clinicians to compare their measurement with those of other centers given the measurement reliability.

SURVEY STUDIES

When approaching a topic of study, it is essential to keep in mind your research purpose and the specific questions that you are trying to answer with the use of your research instrument. Each question on the instrument should be a measure that addresses your research questions. If you are asking a question that is not useful in terms of answering the research questions, then it should be removed. The only exceptions would be information collected to control for population factors such as gender, age, smoking status, and so on. Too many questions on an instrument may interfere with obtaining complete responses and thereby interfere with the research purpose.

Before embarking on designing a new questionnaire, a thorough review of the literature is important to determine what scales are currently being used and whether your question has been previously addressed. This is also helpful in determining what the best method of addressing the question might be, namely what are the unanswered questions in the field of study, what is known, what are the problems with current research, and what would be clinically useful.

The usual method of administration in orthopaedics is by questionnaire. Questionnaires are less expensive, can be standardized, and can measure subjective data at different points in time to determine the outcomes of an intervention.[113] The design of the questions on your study instrument is important for collection of useful data. If you are using a standardized, scannable questionnaire that will be entered into a database, it is necessary to consider the fields on the form and how the data will be analyzed. For example, if you are asking a question about the types of symptoms that a patient has, he or she is given a list of 10 possible symptoms, all of these symptoms are entered into 1 field, and then these data will have to be re-coded for use by the analysis software. In addition, the lack of a response for any given symptom may indicate that patients do not have the symptom or they did not answer the question, even though they do have this symptom. However, if you ask the question as, "Do you have the following symptoms?" and each symptom listed is a field with a possible yes or no response, then these data will not have to be re-coded for analysis and it is clear whether patients have the symptom or not if they mark yes or no or if they missed the question and did not answer yes or no.

The wording of the questions is also important to obtain complete and meaningful answers. If the respondent does not understand the question, he or she might not answer the question. Worse, if the respondent misunderstands the question and answers it incorrectly, then the data might be skewed and your results will not make sense and/or be useful. Questions should also be asked in a neutral manner. It is important to avoid words of judgment. For example, the question "On average, how many days per week do you abuse alcohol?" prejudges the respondent's answer, whereas the question "On average, how many days per week do you drink more than 4 alcoholic drinks?" allows the respondent to answer with less judgment. The wording should allow respondents to answer honestly but not encourage them to exaggerate. The wording should also be at a reasonable educational level that takes into consideration the spectrum of your study population. Most often, a sixth-grade reading level will allow for sufficient information to be collected but still be accessible to most of the respondents.

TABLE 15. *Keys to Survey Research*

1. Questionnaires should be designed so that they are easy to read, understand, and complete. Poorly designed questionnaires will lead to incomplete data.
2. Questions and outcome scores should be reliable, valid, and responsive.
3. Define a clear data collection protocol. A good operating procedure will improve the quality of the data collection and the data.

The format of your questionnaire can help the respondent to complete the forms (Table 15). For example, if you have a question that asks the respondent to rate limitation on a 1 to 10 scale followed by a question that asks the respondent to rate activity on a 0 to 10 scale, it might be confusing whether 10 represents the most limitation or the most activity. It is better to clearly separate these questions to visually cue the respondent to recognize that 10 represents the most activity in one case and the most limitation in another. In addition, the order of the questions is important to help the respondent feel comfortable with the study. Questions about personal or potentially embarrassing information should come later in the questionnaire and preferably not on the first page. For example, questions about income, sex, and/or recreational or prescription drug use should not be at the beginning of the questionnaire. Finally, the font should be large enough for visually impaired respondents to see, and answers should be visually consistent so as not to frustrate the respondent or make him or her dizzy.

The use of outcome instruments whose psychometric properties have been vigorously established is essential. Psychometric evaluation is the epidemiologic field that studies the development and validation of outcome instruments and questionnaires.[107-110,114] The important psychometric properties of an outcome instrument include reliability, validity, and responsiveness.[107-110] Reliability refers to the reproducibility of an outcome measure, either between subjects (test-retest reliability) or between observers (interobserver reliability). Validity questions whether an outcome instrument actually measures what it intends to measure. Components of validity include content validity ("face" validity and floor/ceiling effects), criterion validity (how an instrument compares to an accepted "gold standard" instrument), and construct validity (whether the instrument follows expected noncontroversial hypotheses). Responsiveness assesses change in the instrument's value over time or treatment.

Test-retest reliability is determined by comparing an original questionnaire and a second postoperative questionnaire given in a short time span when no change in clinical status has occurred. The intraclass correlation coefficient (ICC) is determined for each component score. An ICC greater than 0.70 is considered acceptable.[110] The standard error of the measurement is also calculated as described previously.[115] This value will be used to determine the 95% confidence interval for individual scores, which provides an estimate of where the actual score may lie. To further define this interval, the minimum detectable change is calculated to determine the smallest change that can be considered a true difference after measurement error and noise have been taken into account.[115] Noise would be changes in the score due to factors other than changes due to the intervention.

Content validity is determined by the floor and ceiling effect of the score. Preoperative scores are used to establish content validity. Floor effects (scale = lowest possible) and ceiling effects (scale = highest possible) will be determined for each component. Floor and ceiling effects of less than 30% were considered acceptable.[110]

Criterion validity is determined by the correlation of the score with a gold standard. The definition of a gold standard is a score that has been validated for the population you are studying. It is common in orthopaedics to use the Short Form (SF)-12 or SF-36 as a gold standard because it has been extensively studied.[116] The Pearson correlation coefficient should be used for the continuous variables that are normally distributed and the Spearman ρ should be used for nonparametric data.

Construct validity is determined by developing 5 to 10 hypotheses or constructs that are noncontroversial and considered true by many surgeons—for example, "patients with severe pain have lower activity level." These constructs are developed by consensus and tested in the study population. Construct validity tests the score to make sure that score can measure what it claims to measure. If it is a functional score, then it should pass tests that are considered true differences in function.

Responsiveness to change is assessed by comparing the initial scores with scores after an intervention. The time between the initial and follow-up scores should be long enough for the intervention to have made a difference. For example, you would not measure function after ACL reconstruction 2 days after the surgery. Effect size is calculated as (mean postoperative scale – mean preoperative scale)/standard deviation of preoperative scale. Standardized response mean is calculated as (mean postoperative scale – mean preoperative scale)/standard deviation of change in scale. Small effects are considered greater than 0.20, moderate effects are considered greater than 0.50, and large effects are considered greater than 0.80.[110]

The data collection protocol should include the times at which each measure will be collected, with clear specifications for ranges of acceptable collection times. For example, if an outcomes measure will be collected before surgery and at 1 year and 2 years after surgery, it should be clearly stated that the measure can be collected within 1 month before surgery and within 1 month before or after the 1-year and 2-year marks. The collection protocol should also detail who will collect the data. If the data are being collected by 1 or more observers, then the protocol should detail what their qualifications are and how they will collect the data. If any tools are used to collect the data, then they should be described and the methodology for collection should be described. If calculations are required, the method of calculation should be described, especially if a specific type of software is being used.

RELIABILITY STUDIES

Reliability of a measurement defines whether the measurement is dependable and reproducible. Reliability studies answer the following questions: (1) Is this measurement consistent? (2) Is this measurement free of error? Intraobserver reliability and interobserver reliability compare the scoring of tests by the same observer and by different observers, respectively.[111] Intraobserver reliability tests the ability of 1 observer to duplicate his or her test responses on the same test at different points in time, when no intervention for the disease has taken place and/or there has been no progression. Interobserver reliability tests the ability of more than 1 observer to give similar responses on the same test. Reliability is a measure of reproducibility, not of accuracy, and the different ways to measure reliability each provide insight into reproducibility.

When initiating a study that included objective measurements, such as alpha angle in the hip, it is important to include a study of the reliability of the measurement (Table 16). Although it may be quoted in the literature, it is important for readers to know the reliability of the measurement in individual practice settings. When designing the reliability arm of the study, it is important to consider who will be the observers in the study. It is assumed that an observer with more training—for example, a senior physician with 10 years of surgical experience—would be more reliable in measuring than a resident. If the measurement you are testing is commonly measured by residents and senior physicians, it will be important to include both in your study.

When identifying the group of patients to be used in the reliability study, you must ensure that the group covers all levels of the scores. For example, if you are testing the reliability of scoring Kellgren-Lawrence grades on knee radiographs, you do not want to only include those with severe OA. This will make it easier for the observers. Make sure there are a number of patients with every grade in the study group.

After the group has been identified, the object to be tested, for example, the anteroposterior radiograph of the knee, should be de-identified. It should be numbered and any information on the patient should be removed, including his or her name. If the patient has been examined by the observer, the observer should not know that the patient is in the reliability study. Using a random-number generator, you can determine the order in which the radiographs will be observed. If you have 20 radiographs, then use a random-number generator between 1 and 20, and this will provide you with the order of the first reading. For the second reading, again randomize the order for the samples.

Determining the number of observers is usually based on clinical practicality. To have 5 different observers go to a patient's room and measure his or her hip range of motion may be impractical for the patient and the clinic. To determine the sample size needed for the number of subjects, the acceptable level of reliability for the measure must be known.[111,117] In addition, sample size may be different depending on whether the measurement represents continuous or categorical data.[117]

Table 16. *Inter-Rater Versus Intrarater Reliability*

Inter-rater reliability of a measure will define whether the agreement between 2 observers is acceptable.

Intrarater reliability of a measure will define whether there is agreement between 2 observations by the same observer at different times.

Table 17. *Survey Studies*

1. Survey studies should be based on clinically relevant questions.

2. Psychometric properties of outcome scores are important and should be determined before using the score.

3. Important information can be obtained through survey research. The data represent a valuable research tool, and this also allows physicians to track patients and improve patient care.

For continuous variables, such as degrees, inches, and so on, the ICC is used to measure reliability. The ICC is a ratio of the variability among subjects to the overall variability observed in the data. A score of 0 to 0.4 indicates poor reliability, a score of greater than 0.4 to 0.75 indicates fair or moderate reliability, and a score of greater than 0.75 indicates excellent reliability.

For categorical data, the κ coefficient is used to report reliability. The κ coefficient measures the observed agreement compared with the possible agreement beyond chance. For more complicated models, a statistician should be consulted.

CONCLUSIONS

The goal of reliability and survey studies is to measure patient health accurately (Table 17). Specifically, information about how patient health will be affected and how long it will be sustained are essential factors for improving patient care in the future. The key to obtaining this valuable knowledge is good measurement grounded in an understanding of what and how health is being measured.[110] Reliability, validity, and responsiveness studies of disease-specific outcomes instruments provide researchers the tools they need to make accurate measurements of patient health. Survey studies that use these outcomes measures can provide surgeons with the clinically relevant patient information they need to improve function and activity levels for patients with varying types of orthopaedic disease. Survey and reliability studies are therefore valuable tools in the process of continually improving patient care.

Karen K. Briggs, M.P.H.
Kira Chaney-Barclay, M.P.H.
Robert F. LaPrade, M.D., Ph.D.