Clin Orthop Relat Res (2022) 480:1259-1262
DOI 10.1097/CORR.0000000000002260

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Not the Last Word

Published online: 24 May 2022

Not the Last Word: My Perfectly Reliable, and Perfectly Worthless, Fracture Classification System

Joseph Bernstein MD[1]

One sure path to glory in orthopaedics is to invent a popular fracture classification system. Some surgeons, otherwise lost to history, are remembered only for their eponymous classifications. Others among the true greats—Charles Neer and Joseph Schatzker come to mind—are known to younger surgeons less for their prodigious achievements but more as cartographers of shoulder and tibia fractures, respectively.

Despite the praise given to inventors of popular fracture classification systems, criticism comes too, especially regarding the lack of reliability. In one highly cited report, Sidor et al. [7] evaluated the reliability of the Neer classification of proximal humeral fractures. They found that experts disagreed with each other so often—and disagreed with themselves so often, when shown the films for a second time—that one journal [4] proposed banning this system from its pages altogether.

To remedy this deficiency of reliability, and to take my best and only shot at orthopaedic immortality, I invented what I immodestly will call the Bernstein Classification of Femur Fractures [3]: type R, signifying those of the right femur, and type L—no surprise—for those on the left. (Perhaps I should include the name of one of my co-authors, Jeff Silber, and designate this the "B-S classification.")

This silly division of femur fractures illustrates the limitations of reliability as measure of merit. Sure, a modicum of reliability is necessary for clinical use, but the true value of a fracture classification depends on the extent to which the individual categories have unique pathophysiology, treatment requirements, prognosis, or other distinct, defining features.

Consider the ancient physician contemplating a set of patients with excessive urination. He named this condition "diabetes," using the Greek word for "siphon" in homage to the presenting complaint. He soon noticed that some of the patients produced sweet urine and others a more dull-tasting variety and further classified the diabetes as "mellitus" or "insipidus," using the Latin words for "sweet" and "tasteless," respectively. Is this a valid classification? Well, for those without a refined palate, the methods of classification might not be very reliable. Still, now that we know something about the pancreas and the pituitary, we also know that the deep grammar of this ordering is spot-on.

Or consider the more modern orthopaedic surgeon contemplating patients with possible labral tears of the shoulder. Imagine she attempts to classify them on her physical examination. Here, too, the reliability of the classification assignment, at least based clinical tests, may be lacking [5], and yet nobody would doubt the categories "labrum intact" and "labrum torn" are meaningfully distinct entities.

I'd take this defense one step further. Many of the studies that purport to invalidate fracture classification systems on reliability grounds (one of mine

A note from the Editor-in-Chief: We are pleased to present to readers of Clinical Orthopaedics and Related Research® *the next* Not the Last Word. *The goal of this section is to explore timely and controversial issues that affect how orthopaedic surgery is taught, learned, and practiced. We welcome reader feedback on all our columns and articles; please send your comments to eic@clinorthop.org.*

The author certifies that there are no funding or commercial associations (consultancies, stock ownership, equity interest, patent/licensing arrangements, etc.) that might pose a conflict of interest in connection with the submitted article related to the author or any immediate family members.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.
The opinions expressed are those of the writer, and do not reflect the opinion or policy of *CORR*® or The Association of Bone and Joint Surgeons®.

J. Bernstein ✉, University of Pennsylvania, 424 Stemmler Hall, Philadelphia, PA 19104, USA, Email: orthodoc@uphs.upenn.edu

[1]Department of Orthopaedic Surgery, University of Pennsylvania, Philadelphia, PA, USA

Not the Last Word

included [1]) are methodologically flawed because they do not recapitulate the real-world environment. In a typical study, a researcher will gather some radiographic images, corral a couple of colleagues, and with luck, emerge soon thereafter with a list of responses that can be compared among the readers. Yet there is no assurance that the presented patients match the distribution of pathology seen in clinical practice. Further, the reviewers in the study are likely too tired and insufficiently motivated to provide a diagnosis with the rigor they'd apply when managing their own patients.

A new standard must be set. All reliability studies must be based on a large series of consecutive patients, to mitigate spectrum bias. Images should be presented no more than two per day over the course of weeks, with plenty of time given for a response. Last, the evaluators must have some skin in the game. It may be hard to make a study volunteer as attentive as the surgeon sitting for a certifying examination, say, but to the extent that fracture classification demands effort and attention—and it does—an apathetic reviewer will flounder. My favored remedy, which I hope can get past an institutional review board one day, is to have the evaluators agree to the publication of their names and the quality of their performance relative to the others. The prospect of public shaming concentrates the mind wonderfully.

I would further add that the evaluators in the study should be allowed to indicate the degree of confidence for their responses. They should be allowed to say whether they would, as a treating physician, insist on additional images. Last, they should report the individual perceptions and measurements that lead to their responses. Measurements matter. For instance, in the case of the Neer classification, the amount of displacement determines the response: Fractures displaced 9 mm get placed in one Neer category and those displaced 11 mm in another. Needless to say, binary labeling of a continuous distribution, especially one that is imprecisely measured, might impede reliable assessment.

In the meantime, the power of fracture classification systems with imperfect reliability may be improved by crowdsourcing the responses [6]. Crowdsourcing is based on the idea that a group of even ordinary evaluators putting their heads together can outperform the experts. (The classic example is guessing the weight of an ox: The mean value of a crowd of fairgoers was closer to the measured value than the estimate of an experienced butcher.) In the realm of femoral neck fracture displacement, my colleagues and I were able to show that a classification derived by the arithmetic synthesis of a group of three individual responses was more reliable than the classification assignments of any single reader [2].

In short, good classifications are based on insightful distinctions between categories. At times, even good classifications are marred by poor reliability. In some instances, that lack of reliability might be remedied by methods such as crowdsourcing. In other instances, we must wait for the development of better diagnostic techniques.

The enduring popularity of some sweet fracture classification systems, like those of Neer and Schatzker, might be based on the underlying validity of the scheme. Other more insipid systems might endure simply because of inertia, or surgeons' preference for shibboleths and code words. Sadly, we have yet to devise a means of applying this classification of classifications reliably.

Mark S. Vrahas MD

Levin/Gordon Distinguished Chair in Orthopaedics and Professor and Chair, Department of Orthopaedics

Cedars Sinai Health System

I agree with Dr. Bernstein's observations about reliability studies. Plenty of studies impugn the reliability of almost every classification you can think of, but these flaws do not necessarily diminish their value. The Schatzker classification was originally proposed based on orthogonal radiographs. At least seven studies have evaluated intra- and interobserver reliability based on plain radiographs with kappa statistics ranging from 0.57 to 0.91. Adding two-dimensional CT scans yielded interobserver kappa values ranging from 0.46 to 0.75. Adding three-dimensional reconstructions yielded values between 0.596 and 0.85 [8]. Despite this wide variation in reported reliability, the Schatzker classification remains extensively used.

Communication is one of the functions of classification, and even though we know that the Schatzker classification does not tell the full story and that there are many outliers, it does provide a well-understood starting point.

The world is imperfect, but trying to make sense of it requires making meaningful groups for comparison. Classification is necessary for research. Although I am thankful that it is no longer necessary to taste urine to distinguish between diabetes mellitus and insipidus, the classification created by the ancient Greek physician with a taste for urine probably prompted the search for a deeper mechanism. As our knowledge and technology improve, our understanding of fractures improves. A classification just provides a starting point.

Not the Last Word

Dr. Bernstein suggests interesting methodological approaches for evaluating classification systems that would likely be effective. However, I worry that these approaches will generate another slew of papers evaluating fracture classifications while providing no more value than traditional approaches. Ultimately, a fracture classification is only valuable in the context of the study where it was used. We are interested in outcomes, and many factors other than the fracture anatomy contribute to results. No classification is perfect, but the task is deciding whether the researchers used a classification to group cases in a way that allows us to extrapolate their findings to gain some greater understanding.

The flaw in fracture classifications is not their reliability but rather how they are used. Too often surgeons use fracture classification to dictate management rather than as a guide. Fracture classification only points to studies in which the classification was used. Even if there existed a fracture classification that was perfectly reliable, it would only be useful if it was applied in a study that provided guidance on how to manage a patient.

Dr. Bernstein's L-R classification is perfectly reliable, and if he did a study that demonstrated some differences in outcomes between patients with L and R fractures, it would provide useful guidance. Then again, the patient might have two left feet.

Stuart A. Green MD

Clinical Professor, Orthopaedic Surgery

University of California, Irvine

Any worthwhile classification system organizes disparate information into discrete, useful bundles. The Gustillo-Anderson open fracture classification system, for example, groups limb injuries by wound size, predicting the likelihood of later infection in each category with reasonable accuracy, leading to its rapid, widespread acceptance. Joseph Bernstein's L-R hip fracture classification system clearly organizes information into discrete bundles: left hip and right hip. However, his system is not original. A primordial trilobite created laterality distinction 521 million years ago.

To fashion a useful trauma classification system, a developer must have an objective in mind. Many classification systems organize the subject matter by complexity, with the higher numbered groups caused by sequentially greater mechanical energy and resulting in more severe anatomic discombobulation. Such systems typically explain the mechanism of fracture displacement and based on that, they seek to guide treatment.

One problem with all classification systems is that they must be memorized to pass specialty board certifying exams. However, once board-certified, trauma surgeons can themselves appear ignorant by misidentifying a Schatzker V as a Schatzker IV. Moreover, new variants and subgroups for existing classification systems pop up regularly, like mushrooms after summer storms. For this reason, a wise surgeon soon learns to omit precise monikers for virtually all fracture patterns and simply use simpler adjectives like "comminuted" or "severely comminuted" in operative reports and other permanent documentation.

Nevertheless, the human mind seeks to bring order and predictability to seemingly chaotic events. For this reason, I present here, for the first time, the Stuart Green Trauma Classification System (SGTCS), developed after half a century of clinical orthopaedic practice. My classification system covers not only fractures and dislocations, but all manner of traumas. The system predicts, with unerring accuracy, the ultimate prognosis from any bodily injury or combination thereof, regardless of the causative mechanical energy. The SGTCS's usefulness becomes immediately apparent upon first application, allowing an orthopaedist prognostic clairvoyance to a degree never before thought possible.

The SGTCS has two categories and several subcategories. The surgeon first places every injury into one of two broad categories, compensable or noncompensable, by asking, "Is there a person or entity to sue for pain and suffering?" If so, you are dealing with a compensable injury. If not, the trauma victim falls into the noncompensable category.

Subdivisions of compensable injuries include worker's compensation and tort. Typically, worker's compensation injuries receive monetary compensation fixed by statute, whereas tort injuries are more open-ended. For this reason, compensable torts further subdivide into "little policy" and "big policy" accidents. For example, an automobile trauma caused by a person with little policy coverage will produce less permanent pain and suffering than the identical injury caused by a big policy entity, like a municipal bus.

As a rule, noncompensable injuries cause less pain and suffering than compensable injuries, regardless of etiology. Here too, subcategories exist. Nonrecreational injuries occur with nobody to blame, for instance, tripping over the rug at home, breaking a femur. Recreational injuries are the result of hobby-type activities and, therefore, rarely cause permanent pain or suffering, regardless of the magnitude of injury.

Now I ask, dear reader, have you ever seen a more reliable or worthy

trauma classification system in your professional career? I think it's the only one you'll ever need.

References

1. Bernstein J, Adler LM, Blank JE, Dalsey RM, Williams GR, Iannotti JP. Evaluation of the Neer system of classification of proximal humeral fractures with computerized tomographic scans and plain radiographs. *J Bone Joint Surg Am*. 1996;78:1371-1375.
2. Bernstein J, Long JS, Veillette C, Ahn J. Crowd intelligence for the classification of fractures and beyond. *PLoS One*. 2011;6:e27620.
3. Bernstein J, Monaghan BA, Silber JS, DeLong WG. Taxonomy and treatment–a classification of fracture classifications. *J Bone Joint Surg Br*. 1997;79:706-709.
4. Burstein AH. Fracture classification systems: do they work and are they useful? *J Bone Joint Surg Am.* 1993;75:1743-1744.
5. Meserve BB, Cleland JA, Boucher TR. A meta-analysis examining clinical test utility for assessing superior labral anterior posterior lesions. *Am J Sports Med*. 2009;37:2252-2258.
6. Ranard BL, Ha YP, Meisel ZF, et al. Crowdsourcing–harnessing the masses to advance health and medicine, a systematic review. *J Gen Intern Med*. 2014;29:187-203.
7. Sidor ML, Zuckerman JD, Lyon T, Koval K, Cuomo F, Schoenberg N. The Neer classification system for proximal humeral fractures. An assessment of interobserver reliability and intraobserver reproducibility. *J Bone Joint Surg Am*. 1993;75:1745-1750.
8. Zeltser DW, Leopold SS. Classifications in brief: Schatzker classification of tibial plateau fractures. *Clin Orthop Relat Res*. 2013;471:371-374.