Quasi-experimental, reproducible intervention to improve observer agreement in the classification of distal radius fractures and development of Computerized Adaptive Test

Fabiano Caumo, MD
Pedro Gaspar Soares Justo, MD
Henrique Ayzemberg, MD
Bruno Melo Joao Ricardo Vissoci Ana Paula Bonilauri Ferreira, DDS, PhD
Ricardo Pietrobon, MD, PhD

Abstract

Introduction

Although the AO classification is considered one of the main pillars behind the education of orthopedic residents, observer agreement has been reported to be poor in a number of previous publications (Andersen et al. (1996); Belloti et al. (2008); Kural et al. (2010); Kucuk et al. (2013)). Although well established, these findings are disconcerting as classifications are supposed to guide treatment indications, and a substantion lack of observer agreement might mean that a large percentage treatment variability might be explained by a simple lack of understanding of the fracture radiological appearance, biomechanics and, as a consequence, treatment. Despite our knowledge of the cognition behind medical diagnosis having substantially improved over the past couple decades, however, to our knowledge these fields have remained largely without an intersection.

The accurate usage of a scale can be evaluated determining how reproducible are the answers from different observers at different times (Garbuz et al. (2002)). According to this criteria, although AO scale is widely accepted, several studies have found low agreement rates. (Belloti et al. (2008); Kural et al. (2010); Kucuk et al. (2013)). These low agreement rates may be reflex of the use of wrong cognitive schema or wrong heuristics.

Previous studies have found that physicians tend to use heuristics instead of evidence based instruments such as scales (Ferreira et al. (2010)) in their daily practice. Since medical education is based on several schemas where knowledge is built for future use (Regehr and Norman (1996)), learning a coherent cognitive schema that avoids cognitive overload may facilitate the understanding of scales (Ruiter, Kesteren, and Fernandez (2012); Ferreira et al. (2010)), thus enhancing its agreement rates and consequently its accurate daily usage.

The objective of this study was therefore to investigate the intra and interobserver reliability of the AO classification system for distal radius fractures with and

without the aid of a cognitive schemata obtained through a cognitive task analysis.

Methods

Institutional Review Board

We obtained approval from the Research Ethics Committee of the São José Municipal Hospital prior to the initiation of this project (protocol number 1240.567). All participants were provided with informed consent prior to enrollment in the study.

Participants

A total of 14 orthopedic residents participated in this study, 6 were first-year, 4 were second-year, and 4 were third-year residents. Their average age was 29, 13 being males.

Image selection

Two second-year and one fourth-year residents that were not enrolled directly with the study and an orthopedic surgeon, specialist in hand selected 20 images containing fractures with a wide pattern variation chosen to cover the full spectrum of the AO classification for distal radius fractures. As imagens eram na incidência ântero-posterior e perfil. As imagens foram obtidas de arquivos dos Hospital São Municipal São José (Joinville – SC). Qualquer sinal de identicação de pacientes foi removido. Imagens radiográficas mal posicionadas que poderiam gerar problemas na interpretação foram excluídas. Imagens de baixa qualidade ou com artefatos ou outros defeitos técnicos também foram excluídas.

AO classification

The AO classification is based on the degree of bone fracture, serving as the basis for both treatment indication and outcome assessment (Belloti et al. (2008)). It divides fractures into three main categories: extra-articular, partial articular and complete articular. These three groups are organized in progressive order of severity in relation to their anatomical severity, treatment complexity and prognosis. Fractures from Group A (extra-articular) do not affect the radiocarpal joint surface at all, while Group B (partial articular fracture) does affect the radiocarpal joint leaving a section of the articular surface remaining connected to the diaphysis. Fractures in Group C (complete articular fracture) constitutes a complete separation between the involved articular surface and the diaphysis.

These three main groups are then subdivided into three subgroup, therefore constituting a total of 27 different types of fractures. These fractures vary in relation to how stable they are, their degree of comminution, how reduceable they are and the localization of their fragments (Müeller, 1987). https://www2.aofoundation.org/wps/portal/surgery?showPage=diagnosis&bone=Radius&segment=Distal)

Situated schemata extraction using Cognitive Task Analysis

For the purposes of our paper, we define a situated schema as the collection of concepts and situations (e.g., narratives) that an expert hand surgeon relates to each classification category. In order to extract the situated schema from our expert hand surgeon (HA), we used the following sequence.

First, the AO classification was presented to the hand surgery expert in an electronic format combining text and graphics for each classification category. Second, we asked the surgeon to "think aloud" about what they thought when finding a case in their daily practice. After an initial description, we specifically asked the expert to discuss any diagnostic, biomechanical or related therapeutic decision if it had not yet been previously mentioned. We also encouraged the expert to provide any narratives that might occur to him while thinking aloud about each classification category. The entire process was recorded in a video.

Second, the video was analyzed and a graph constituted by nodes and edges was built using Graphviz. Each node represented either a concept or a situation, while edges connected relationships among diagnostic, biomechanical and therapeutic nodes.

Study logistics and procedures

Baseline evaluation At baseline, all participants independently classified all 20 images according to the AO classification. All participants simultaneously gathered in a single room, being instructed not to look at each other's responses or discuss any cases. Each resident received a directory with all images to be classified. All responses were provided in paper sheets, which were subsequently transcribed to a database. Residents were allowed to check the Web for the classification.There was no time limit for making decisions. Study authors did not participate as study subjects. The directory with all images was deleted at the end of the sesssion in order to decrease the odds of recall bias in subsequent evaluations.

Pre-intervention, thirty-day evaluation After 30 days, each study participant received a new directory with the same 20 images, but in a different order. All other procedures were executed exactly as described for the baseline session.

Intervention The educational intervention was constituted by weekly sessions where participants completed 15 exercises related to the diagnosis, biomechanics and therapeutic planning of radius fractures. Os residentes, após terem respondido cada exercício, eram apresentados à resposta correta, juntamente com uma explicação justificando tal resposta. Os exercícios estavam organizados, na Plataforma online Edx, em "blocos" de 15 perguntas cada, ou seja, foram criados 4 blocos correspondentes as quatro semanas de intervenção. Os blocos de perguntas foram programados para serem "realised" semanalmente, isto é, os residentes tinham acesso somente a um bloco de perguntas por semana. The full spectrum described by the AO classification was covered based on cognitive schema from our expert hand surgeon (HA) as described in the previous session - Situated schemata extraction using Cognitive Task Analysis.

Post-intervention, sixty-day evaluation After the 4-week intervention period, all participants classified the same 20 images following the same protocol.

Outcome measurement

Intra-observer agreement was measured by comparing ratings by the same participant between baseline and the thirty 30-day assessment. Baseline inter-observer agreement was measured at the 30-day assessment. The pre-post intervention evaluation was conducted by comparing the 30-day pre-intervention assessment with the 60-day post-intervention assessment.

Data analysis

All data were extracted directly from MySQL and MongoDB databases connected to the Open edX platforms. Data sets were then merged, also undergoing an exploratory graphical analysis to verify distributions, percentages, means and frequencies/percentages as well as rates of missing data.

Apenas observadores que haviam completado um determinado grupo de observações (dia 1, dia 30 ou dia 60) foram considerados na análise. Porcentagens de concordância assim como valores de Kappa Fleiss foram reportados. Kappa Fleiss é uma medida de concordância para variáveis categóricas que leva em consideração a possível concordância ao acaso. (**???**) Por fim, a comparação entre os valores de Kappa pré e pós intervenção (dias 30 e 60, respectivamente) foram estimadas através da computação de erros padrão e intervalos de confiança (95%) utilizando bootstrap.

Results

Resultados descritivos

Quando todos os 11 observadores completando as 3 avaliações foram considerados, não houve nenhuma instância em que houvesse concordância completa entre todos, mesmo quando houve uma maior tolerancia em relação a subclassificações. No entanto, a porcentagem de concordância entre os 14 observadores completando a primeira avaliação com uma tolerância em relação à subclassificação, houve uma concordância em 70% das avaliações. Esta concordância caiu para 50% e 45% nas avaliações dos dias 30 (14 observadores) e 60 (15 observadores, pós-intervenção), respectivamente.

Em relação à comparação das observações do mesmo observador, comparando os dias 1 e 30, a porcentagem de concordância foi de 11.2% sem tolerância e 78.6% com tolerância.

Valores de kappa

Valores de Kappa Fleiss para a concordância entre observadores amostra do dia 1 foram de 0.225 ($p < 0.001$), 0.212 ($p < 0.001$) para o dia 30 e 0.214 ($p < 0.001$) para o dia 60. Não houve uma diferença estatisticamente significativa entre as concordâncias do dia 30 (pré-intervenção) e dia 60 (pós-intervenção). A concordância entre o mesmo observador nos dias 1 e 30 demonstrou um valor de Kappa Fleiss de -0.004.

Discussão

Até onde sabemos, esse é o primeiro estudo avaliando uma intervenção na tentativa de melhorar o grau de concordância entre observadores para a classificação da AO para o terço distal do rádio. Nossos resultados mostraram uma concordância estável durante o estudo, não tendo sido alterada em decorrência da intervenção. Também encontramos que o grau de concordância é melhorado quando a classificação é simplificada através da retirada das subclassificações.

O baixo grau de concordância entre observadores nesse estudo está alinhado com a literatura sobre classificações de fraturas ortopédicas (Andersen et al. (1996); Belloti et al. (2008); Kural et al. (2010); Kucuk et al. (2013)). Essa baixa concordância se dá em grande parte pela complexidade das classificações, o que dá margem a interpretações diversas, especialmente por profissionais em treinamento e portanto com menos experiência (Kreder et al. (1996); Arealis et al. (2014)). Apesar de que a intenção de se criar uma classificação que seja clinicamente detalhada é inicialmente interessante, a alta carga cognitiva exigida dos profissionais que a irão utilizar tende a fazer com que ela perca a sua praticidade. Esforços deveriam ser realizados portanto para a criação de escalas

que sejam mais dinamicamente adaptadas a profissionais com diferentes graus de experiência na interpretação radiográfica. Por exemplo, profissionais que trabalhem em pronto socorros deveriam utilizar uma escala mais simplificada, enquanto sub-especialistas deveriam utilizar escalas mais detalhadas. O grau de detalhamento em cada uma destas subescalas seria definido através de estudos que identifiquem o grau de concordância obtido na prática clínica diária.

Apesar da nossa intervenção ter sido baseada em mecanismos bem estabelecidos na literatura sobre esquemas cognitivos (Regehr and Norman (1996); Ruiter, Kesteren, and Fernandez (2012)), não houve uma melhora da concordância como nós havíamos hipotetizado. Esquemas cognitivos situados hipotetizam que o cérebro raciocina não apenas através de informações armazenadas no próprio cérebro, mas utilizando fatores ambientais como tecnologias, contatos sociais, entre outros fatores (Van Merriënboer and Sweller (2010)). Causas para a não melhora provavelmente se devem ao baixo tempo de exposição em relação à intervenção , e também ao fato de que esta exposição não ocorreu em um contexto clínico, mas sim em um ambiente educacional artificial. Resultados superiores talvez pudessem ter sido encontrados se a intervenção educacional pudesse ter ocorrido durante a prática clínica diária, especificamente no momento em que os participante estivessem atendendo pacientes com fraturas de radio distal.

Apesar de o nosso artigo ser, até onde sabemos, o primeiro a conduzir uma intervenção na tentativa de melhorar o grau de concordância entre observadores, o nosso estudo tem limitações. Primeiro, a nossa intervenção utilizou um desenho pré-pós ao invés de um estudo randomizado. Na ausência de randomização, nós, portanto, não podemos fazer afirmações sobre relações causais entre a intervenção e a ausência de impacto sobre o grau de concordância. Estudos randomizados requerem no entanto amostras significativamente maiores, o que pode levar a barreiras logísticas em relação à sua execução. Segundo, esse estudo foi restrito a um grupo de participantes de uma única instituição, o que limita a sua generalizabilidade. Enquanto a participação de múltiplas instituições é sempre desejável, sociedades profissionais em ortopedia ainda não estão logisticamente organizadas para permitir estudos de maior escala, o que dificulta a sua realização. Por último, intervenções mais prolongadas e contextualizadas na prática diária teriam sido desejáveis, mas como nas limitações anteriores, a sua execução é limitada por fatores logísticos.

Em conclusão, nós não recomendamos que intervenções educacionais curtas e descontextualizadas da prática clínica sejam utilizadas no aprendizado de classificações complexas. No entanto, a simplificação de tais classificações deveria ser considerada, levando a uma personalização da escala a ser utilizada a grupos de profissionais com graus de experiência diferentes. No que diz respeito a estudos futuros, recomendamos a utilização de estudos randomizados que permitam investigações causais, assim como a contextualização e personalização das intervenções educacionais.

Andersen, Dennis J, William F Blair, Curtis M Stevers Jr, Brian D Adams, George Y El-Khouri, and Eric A Brandser. 1996. Classification of distal radius

fractures: an analysis of interobserver reliability and intraobserver reproducibility. *The Journal of hand surgery* 21, no. 4: 574–82.

Arealis, Georgios, Ilias Galanopoulos, Vassilios S Nikolaou, Andrew Lacon, Neil Ashwood, and Christos Kitsis. 2014. Does the cT improve inter-and intra-observer agreement for the aO, fernandez and universal classification systems for distal radius fractures? *Injury*.

Belloti, João Carlos, Marcel Jun Sugawara Tamaoki, Carlos Eduardo da Silveira Franciozi, João Baptista Gomes dos Santos, Daniel Balbachevsky, Eduardo Chap Chap, Walter Manna Albertoni, and Flávio Faloppa. 2008. Are distal radius fracture classifications reproducible? Intra and interobserver agreement. *Sao Paulo Medical Journal* 126, no. 3: 180–85.

Ferreira, Ana Paula Ribeiro Bonilauri, Rodrigo Fernando Ferreira, Dimple Rajgor, Jatin Shah, Andrea Menezes, and Ricardo Pietrobon. 2010. Clinical reasoning in the real world is mediated by bounded rationality: implications for diagnostic clinical practice guidelines. *PloS one* 5, no. 4: e10265.

Garbuz, Donald S, Bassam A Masri, John Esdaile, and Clive P Duncan. 2002. Classification systems in orthopaedics. *Journal of the American Academy of Orthopaedic Surgeons* 10, no. 4: 290–97.

Kreder, Hans J, Douglas P Hanel, Michael Mckee, JESSE Jupiter, GARY McGillivary, and MARC F Swiontkowski. 1996. Consistency of aO fracture classification for the distal radius. *Journal of Bone & Joint Surgery, British Volume* 78, no. 5: 726–31.

Kucuk, Levent, Mert Kumbaraci, Huseyin Gunay, Levent Karapinar, and Oguz Ozdemir. 2013. Reliability and reproducibility of classifications for distal radius fractures. *Acta orthopaedica et traumatologica turcica* 47, no. 3: 153–57.

Kural, Cemal, Ibrahim Sungur, Ibrahim Kaya, Akin Ugras, Ahmet Ertürk, and Ercan Cetinus. 2010. Evaluation of the reliability of classification systems used for distal radius fractures. *Orthopedics* 33, no. 11: 801.

Regehr, Glenn, and Geoffrey R Norman. 1996. Issues in cognitive psychology: implications for professional education. *Academic Medicine* 71, no. 9: 988–1001.

Ruiter, Dirk J, Marlieke TR van Kesteren, and Guillen Fernandez. 2012. How to achieve synergy between medical education and cognitive neuroscience? An exercise on prior knowledge in understanding. *Advances in health sciences education* 17, no. 2: 225–40.

Van Merriënboer, Jeroen JG, and John Sweller. 2010. Cognitive load theory in health professional education: design principles and strategies. *Medical education* 44, no. 1: 85–93.