UNIVERSITY OF LEEDS

This is a repository copy of *Improving domain definition and outcome instrument selection: Lessons learned for OMERACT from imaging*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/178072/

Version: Accepted Version

Article:
D'Agostino, MA, Beaton, DE, Maxwell, LJ et al. (17 more authors) (2021) Improving domain definition and outcome instrument selection: Lessons learned for OMERACT from imaging. Seminars in arthritis and rheumatism. ISSN 0049-0172

https://doi.org/10.1016/j.semarthrit.2021.08.004

© 2021, Elsevier. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/.

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Title:

Improving domain definition and outcome instrument selection: Lessons learned for OMERACT from imaging

Authors:

Maria Antonietta D'Agostino[a,b,c,] Dorcas E Beaton[d], Lara J Maxwell[e], Sam Michel Cembalo[f], Alison Maria Hoens[f,g,h], Catherine Hofstetter[f], Codruta Zabalan[f,i], Paul Bird[j], Robin Christensen[k,l], Maarten de Wit[f], Andrea S Doria[m], Walter P Maksymowych[n], Win Min Oo[o,p] Mikkel Østergaard[q,r], Teodora Serban[s], Victor S Sloan[t,u], Lene Terslev[q,r], Marion A van Rossum[v], Philip G Conaghan[w], Maarten Boers[x,y]

Affiliations

[a]Università Cattolica del Sacro Cuore

[b]Rheumatology UOC, Fondazione Policlinico Universitario Agostino Gemelli, IRCSS, Rome, Italy

[c]UVSQ, Inserm U1173, Infection et inflammation, Laboratory of Excellence INFLAMEX, Université Paris-Saclay, Montigny-le-Bretonneux, France

[d]Institute for Work & Health and Institute for Health Policy Management and Evaluation, University of Toronto, Toronto

[e]Faculty of Medicine, University of Ottawa, Ottawa, Canada

[f]OMERACT Patient Research Partner

[g]University of British Columbia Faculty of Medicine Department of Physical Therapy, Canada

[h]Patient Partner, Arthritis Research Canada

[i]Romanian League against Rheumatism

[j]University of New South Wales, Sydney, Australia

[k]Section for Biostatistics and Evidence-Based Research, the Parker Institute, Bispebjerg, Frederiksberg Hospital, Copenhagen, Denmark

[l]Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, Denmark.

[m]The Hospital for Sick Children, Medical Imaging Department, University of Toronto, Toronto, Canada

[n]Department of Medicine, University of Alberta, Canada

[o]Rheumatology Department, Institute of Bone and Joint Disease, Kolling Institute, Sydney University, Sydney, Australia

[p]Department of Physical Medicine and Rehabilitation, University of Medicine, Mandalay, Mandalay, Myanmar

[q]Copenhagen Center for Arthritis Research, Center for Rheumatology and Spine Diseases, Rigshospitalet, Glostrup, Denmark

[r]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

[s]La Colletta Hospital, Rheumatology Department, ASL3 Genovese, Genoa, Italy

[t]Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ, USA

[u]The Peace Corps, USA

[v]Amsterdam Rheumatology and Immunology Center | Reade and Emma Children's Hospital Amsterdam University Medical Centers, Amsterdam, The Netherlands

[w]Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, and NIHR Leeds Biomedical Research Centre, UK

[x]Department of Epidemiology and Data Science, Amsterdam Public Health Research Institute, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[y]Amsterdam Rheumatology and Immunology Center, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Correspondence to:

Maria Antonietta D'Agostino, MD, PhD

Professor of Rheumatology
UOC of Rheumatology, Agostino Gemelli University Polyclinic Foundation IRCCS,
Catholic University of Sacred Heart, Largo Francesco Vito 1, 00168 Roma, Italy
Phone: +39 06 30157807
Email: mariaantonietta.dagostino@unicatt.it

Highlights

- Clear definition of the domain we want to measure is a necessary prerequisite to the selection of a good instrument.

- Measurement instruments, whatever the domain they strive to measure, are always the result of the interaction between the instrument (technique+ application + scoring system) and a domain of interest, and it is always the score that is used to represent the domain.

- Imaging outcome measurement instruments for research are not the imaging techniques themselves (i.e. ultrasound or X-ray), but the result of a formalized interpretation (through a scale or score)

- Clear identification of the sources of variability that can directly influence the instrument and therefore the measurement of the domain of interest should be clearly identified before endorsing any outcome measurement instrument

Abstract

Objectives

Imaging is one of the most rapidly evolving fields in medicine. Unfortunately, many imaging technologies have been applied as measurement instrument without rigorous evaluation of the evidence supporting their truth, discriminatory capability and feasibility for that context of use.

The Outcome Measures in Rheumatology (OMERACT) Filter 2.1 Instrument Selection Algorithm (OFISA) is used to evaluate such evidence for use of an instrument in a research setting. The objectives of this work are to: (1) define and describe the key conceptual aspects that are essential for the evaluation of imaging as an outcome measurement instrument and (2) describe how these aspects can be assessed through OFISA.

Methods

Experts in imaging and/or methodology met to formalize concepts and define key steps. These concepts were discussed with a team of patient research partners with interest in imaging to refine technical and methodological aspects into comprehensible information. A workshop was held at OMERACT2020 and feedback was incorporated into existing OMERACT process for domain and instrument selection.

Results

Three key lessons were identified: (1) a clear definition of the domain we want to measure is a necessary prerequisite to the selection of a good instrument, (2) the sources of variability that can directly influence the instrument should be clearly identified, (3) incorporating these first two lessons into OFISA improves the quality of every instrument selection process.

Conclusions

The incorporation of these lessons in the updated OMERACT Filter (now 2.2) will improve the quality of the selection process for all types of outcome measurement instruments.

Keywords:

Imaging, Outcome measurement instruments, OMERACT, patient reporting outcomes, ultrasound, MRI, rheumatic and musculoskeletal diseases, methodology.

Introduction

Imaging tests are among the most rapidly evolving fields within medicine. In the last 40 years, management of diseases has been transformed by the rapid expansion of sophisticated new technologies offering a large range of options for identifying, monitoring and predicting pathological processes. Unfortunately, many of these technologies, including imaging, have been used as measurement instruments and disseminated into daily practice or employed as endpoints in clinical trials without rigorous evaluation of the evidence supporting their truth, discriminatory capability and feasibility for that context of use.

In 1992, Outcome Measures in Rheumatology (OMERACT) was established to improve the quality and consistency of measurement in clinical research by developing "core outcome sets (COS)" that recommend a minimum set of measures that should always be used in clinical trials (and observational studies) in different rheumatological disorders. This ensures valid and comparable results between trials, reduces research bias, and benefits clinical decision makers.

Since 2014, OMERACT has updated and clarified the two stages needed to form a COS. The first is to develop "core domain" sets to define "what" should be measured in a COS. A core domain may be a body function (e.g. inflammation), a consequence of a disease (e.g. structural damage), or the impact of a disease (e.g. on quality of life). This means that in a "core domain" set, the domain of interest needs to be clearly identified, and its definition is to be provided. Once this stage is completed, measurement instruments can be selected and included in a "Core Outcome Measurement" set (i.e. "how" to measure the core domain). Together the core domain set and the core outcome measurement set form a complete COS. (1-5).

Importantly, regardless of the nature of the chosen instrument (patient-reported, clinician-reported, performance-based, laboratory findings or observations through imaging), the selection of any outcome measurement instrument should follow the same evidence-based decision-making process. This means gathering enough evidence to demonstrate that the instrument can truthfully assess that domain, discriminate between situations of interest and be feasible to use in the context of clinical studies in a defined population (6). The score obtained by such measurement represents the outcome of interest (for example, a synovitis score). OMERACT requires identifying at least one instrument for each domain of interest.

<u>Imaging as a measurement instrument</u>

Firstly, it is important to understand the mechanisms by which different available imaging techniques lead to the production of an image. Images are the result of the detection of an interaction between a form of energy and organs, tissues or cells. The same device usually produces the energy, detects the interaction and translates it into an image (i.e. makes it visible). Domains measured by imaging fall under the broad category of biomarkers in the core area of pathophysiological manifestations of the disease in the OMERACT framework, such as body function and structure, organ function (1,2). These domains are

usually physiological or pathological manifestations of a health condition (i.e. disease-centered domains), and importantly this can include both reversible (e.g. inflammation) and irreversible (e.g. structural damage) manifestations which can occur independently or concurrently.

The ability of an imaging technique to detect and measure a structure or process is firstly dependent on its technical characteristics (i.e. can this form of energy visualize the organ or process under study?) and on its technical performance (i.e. is the signal-to-noise ratio appropriate?). Table 1 shows some examples of imaging techniques used in medical imaging, the physical principles generating the energy and the tissue interaction.

Table 1. The mechanism by which different techniques lead to an image.

Imaging technique	Physical principles	Tissue interaction	Example result
Radiography	X-ray radiation	Absorption by Ca++ in tissue	Images of bone
Ultrasound greyscale	Sound waves	Reflection and refraction by tissues	Images of soft tissue
Ultrasound Doppler	Sound waves	Reflection and refraction by moving blood cells	Images of blood flow movement/presence
MRI	Magnetic gradient	Energy released by relaxation of protons in the cells of tissues	Images of tissue based on water content

For example, X-ray radiation is applied in conventional radiography (CR) and reflects the degree to which the X-Ray radiation is blocked or attenuated by different tissues within the body. X-ray radiation can be harmful; therefore, the applied energy should be minimized. Since the generation of the image (i.e. different attenuation between tissues) and the correct discrimination between structures (i.e. spatial and contrast resolution) are dependent on the energy of the X rays, only dense tissues are perfectly visualized. CR is best for visualization of bone. Furthermore, as the X-rays pass through the tissue, a bi-dimensional image of a tridimensional structure is obtained. Thus, the image also depends on the angle of incidence of the X-ray beam and the position of the tissue. The use of computed tomography (CT) overcomes the problems related to the incidence of the X-Ray beam, but not those related to the radiation.

A second important aspect is the correct interpretation of the image by humans or computers. One of the key challenges in using imaging as measurement instrument, especially in research, is the complex interaction between the technical characteristics of the imaging technique, the setting in which it is applied (e.g. clinical practice or research), and the interpretation of the acquired data. In clinical practice, images are usually interpreted in a qualitative, or semi-quantitative way. The imaging examiner/reader records the presence or absence of certain pre-established features that suggest a diagnosis. In sequential examinations, the examiner/reader can detect presence of absence of change (improvement or deterioration of a given feature), and true quantification can be limited to recording the physical dimensions of a feature (e.g. the diameter of a pulmonary node with X-Ray or CT) but also to the use of

more sophisticated measurements obtained by post-processing of the images such as with magnetic resonance imaging (MRI) (e.g. apparent diffusion coefficients in diffusion-weighted MRI, T2 relaxation times in T2 mapping MRI).

To use imaging techniques as measurement instruments in research, a formal scale or score is needed. Manifestations or abnormalities visualized by imaging can be quantified either by a nominal (present/absent) or an ordinal (semi-quantitative) scale of subjective observer-based judgments; or by an interval scale of units that assess a physical dimension (length, area, volume, velocity, uptake, etc.).

Thus, an imaging outcome measurement instrument for research is not about the imaging technique itself (i.e. ultrasound, MRI or X-ray), but about the result of a formalized interpretation (through a quantitatively or continuous scale, or binary or semi-quantitative scoring systems) of an image acquired through a standardized application of the technique to tissues, organs, structures and processes.

The instrument (imaging technique + application + scoring system) yields results that represent the imaging domain (or domain of interest) for a research study (in OMERACT, this is primarily a clinical trial).

In many ways this definition of an instrument also applies to other types of instruments, including patient-reported outcomes (PROs) measures where the items and response options along with any instructions are the equivalent to the technique used (i.e. ultrasound, MRI, etc.). The specific questions/items and their interpretation are the scoring system (or instrument).

Measurement instruments, whatever the domain they strive to measure, are always the result of the interaction between the instrument (technique + application + scoring system) and a domain of interest (for example joint inflammation, fatigue or work productivity), and it is always the resulting score that is used to represent that domain. In the application of imaging as measurement instrument, the interaction between the imaging technique, the way in which the image is acquired, the domain of interest and the reader interpretation may influence the result of measurement.

In 2008, a methodological work performed by D'Agostino focused on evaluating the possible sources of variability influencing an imaging technique, ultrasound, when used as a diagnostic test, and underlined the need to evaluate and standardize the complex relationship between the instrument and the domain under study, especially in the absence of a gold standard (7). The methodology proposed in such setting was then applied within the OMERACT Ultrasound Working Group to develop imaging outcome measurement instruments based on ultrasound (8) and within several European Alliances for Rheumatology (EULAR) initiatives culminating into developing recommendations on how to report ultrasound studies in rheumatic diseases (9), and contributing to the efforts made by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative to evaluate risk of bias in reliability studies (10).

Following these concurrent initiatives, it was pointed out that the complex relationship between the technique, its application, and the scoring system produced, has rarely been evaluated in existing outcomes measurement instruments especially in more technically driven instruments such as imaging instruments.

The objective of this work, therefore, is to critically define and describe two key conceptual aspects identified as essential for the development and evaluation of imaging outcome measurement instruments that can be used as lessons learned for other types of instruments as well:

- Clear definition of the domain we want to measure as a necessary prerequisite to the selection of a good instrument.
- Clear identification of the sources of variability that can directly influence the measurement of the domain of interest and therefore the correct development and application of the scoring system.

An additional objective, constituting the third lesson, is that the application of the first two lessons to other instruments, improve the quality of instrument selection for all clinical and biological outcome assessments and that such selection can be performed through the application of the OMERACT Filter 2.1 Instrument Selection Algorithm (OFISA).

Methods

A group of five experts in imaging and/or methodology met over the last three years to address the key points needed for evaluating the quality of imaging outcome measurement instruments and to identify a possible stepwise approach to the aforementioned objectives.

Several face-to-face and online meetings were performed to formalize concepts and define the key steps needed for evaluating imaging instruments aligning with the algorithm process proposed by OMERACT for selecting measurement instruments (4). The group obtained a stepwise approach that was subsequently discussed with a team of patient research partners with interest in imaging (OMERACT Patient Research Partner Task Force on Imaging Outcomes) to refine technical and methodological aspects into useable information for imaging experts, patients, members of the OMERACT community with little knowledge of imaging, and also methodologists outside OMERACT.

Training tools were then developed by the task force to support the uptake and discussion of these lessons at an OMERACT Workshop held in October, 2020 (videos on each lesson available at: https://omeract.org/instrument-selection/), where 62 persons (50 clinician/researchers and 12 patients) participated in iterative clarification and refinement of the lessons and their application to the outcome measurement instrument selection. Feedback from these groups was then integrated, and the resulting key points were incorporated into OMERACT processes for Core Domain Sets and Core Outcome Measurement Instrument Selection processes in the updated OMERACT Filter 2.2. (11)

Results

Lesson 1. Detailed definition of the domain of interest is the foundation for instrument selection

The first step in the identification of the most appropriate instrument returns to the last step of the domain selection process (3): there needs to be a clear definition of the domain we are interested in, and this must come before the OMERACT instrument selection process begins. Without a definition that offers enough clarity of the domain, and of the range of variety expected in a population, it will be difficult to

evaluate if a candidate outcome measurement instrument will be able to measure the domain of interest (e.g. inflammation). Sometimes definitions are deliberately broad in nature. For example, "pain" is a broad and complex domain that may require more detail – for example, pain intensity, duration or frequency, or whether it is intermittent or constant pain (12). In the same way, "inflammation" can be described using more specific attributes: tissue inflammation, soluble mediators of inflammation, organ or structure (i.e. joint) inflammation, etc.

Therefore, researchers should first specify which precise characteristic or aspect of a "broad" domain they are interested in measuring — we propose to call this the "target" domain — then he/she should verify that a candidate instrument is able to measure such "target domain".

No imaging technique is able to assess all anatomical structures and all pathophysiological manifestations. Once the appropriate imaging technique is chosen, it is necessary to verify whether the result (i.e. the image) obtained has an appropriate relationship with the target domain under study.

Therefore, a stepwise process is recommended to yield a detailed definition of the target domain.

a) In what **Core Area** are we working? As stated above most imaging outcomes fall within the core area of (pathophysiological) manifestations of the disease.
b) **What** do we want to measure? This requires good understanding and clarification of the domain of interest or **Broad Domain**: i.e. the manifestation of interest within the Core Area of Manifestations/Abnormalities;
c) **What** are we focusing on? This requires specification and conceptual ("theoretical") definition of the **Target Domain**: i.e. a structure or process within the broad domain that we want to assess;
d) **What** can we actually measure? This requires identification and "operational" definition of the measurable **Domain Components** ("elementary components"): i.e. the (parts of the) process or structure that can actually be measured by the imaging technique.

The choice of the best imaging technique requires verification that the technique is able to visualize the process/structure within the manifestation of interest (target domain) and to detect 'enough' domain components. This point implies a comprehensive description of the domain and the elements of that domain that the instrument is able to detect and measure. Finally, the instrument needs to have accurate and consistent definitions of the target domain and its elementary components before it can be selected for a core set. **Figure 1** shows this stepwise process.

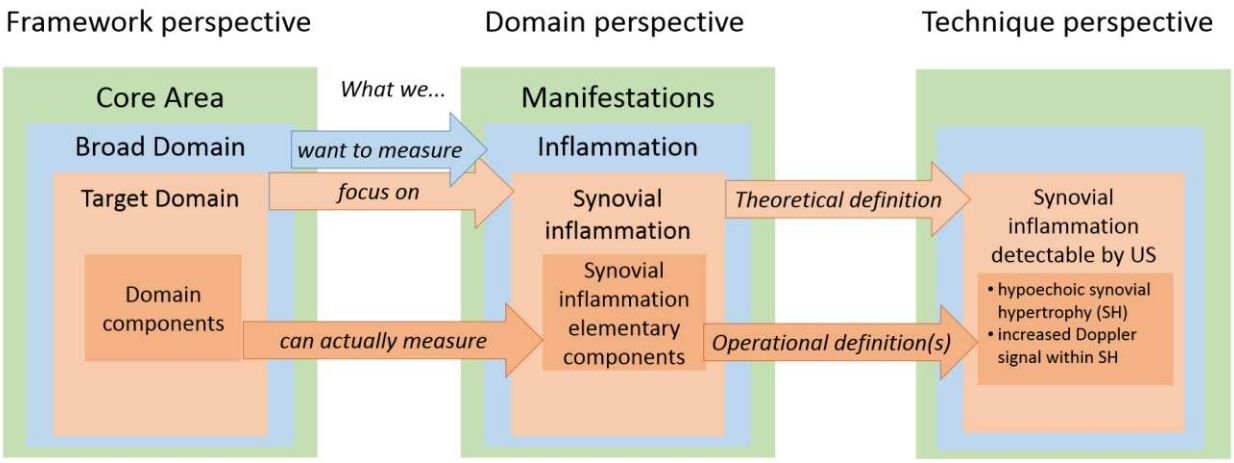

Figure 1. A Process for the identification of the layers of the domain identification and definition. On the left is shown the recommended framework for describing the domain of interest (i.e., broad domain), what we want to focus on measuring (i.e., target domain) and the components of that target domain that need to be measured (i.e., domain components). In the middle is shown what this framework might look like from the perspective of the domain, in this case the broad domain of inflammation leads to the more specific target domain of synovial inflammation. Finally, on the right is the technique perspective showing that ultrasound is the technique of interest for assessing the synovial inflammation (i.e., ultrasound-detected synovitis) and the elementary component definitions needed to operationalize the assessment of the target domain are described (i.e., hypoechoic synovial hypertrophy).

This 'boxes' approach provides an easy template to start the assessment of face (theoretical definition) and content validity (operational definitions) of an imaging instrument.

The domain definition template in Appendix 1 shows examples of broad and target domains as well as of domain components and clarifies the concept of theoretical and operational definitions.

In the instrument development process, face and content validity - both theoretical and operational definitions - (i.e. clarity, accuracy and comprehensiveness) should first be tested and found to be acceptable under ideal conditions. This implies searching the literature for definitions, formulating initial definitions and, based on these, developing possible scoring systems, optimizing and agreeing through a consensus process (such as Delphi) among experts, and testing for validity and reliability on static images or video clips and on patients. Through this, candidate scoring systems can be developed. Often, this process is iterative before the final instrument (definitions + scoring system) is complete and ready for a full evaluation using OFISA. Therefore, from the perspective of a researcher searching for an imaging outcome measurement instrument for a core set (for example an ultrasound scoring system for synovitis), the selection should be based on the assessment of the literature on the development of the instrument, and the documentation that the above steps were performed (8).

<u>Lesson 2. Understanding the sources of variability that influence outcome measurement instruments</u>

Variability and errors are inherent to every measurement instrument. Measurement errors to be considered are random measurement errors (noise), which affect reliability of the instrument, and systematic measurement errors (bias) which affect validity (15-17).

The knowledge of possible sources of variability affecting the measurement instrument and of the errors they may generate is an essential step in the choice and validation of an imaging outcome measurement instrument. As stated above, one of the main challenges in the development and proper use of imaging outcome measurement instruments is the complex interaction between the technical characteristics of the imaging technique, the setting in which it is applied, and the influence of several factors on the interpretation of the acquired data. All of these create sources of variability.

In imaging, variability is firstly related to the technical performance of the imaging instrument, as well as the quality and implementation of the imaging acquisition protocol (i.e. proper choices for radiation dose, MRI sequences, ultrasound setting, patient and probe positioning (18-19). Next, variability is related to the quality of the agreed definition(s) of what should be measured and assessment of severity of the studied lesion(s), the precise details of the scoring methodology that are available as well as knowledge transfer tools (20-22). For example, the measurement of synovitis on ultrasound in peripheral joints is different from the assessment of bone marrow edema, by MRI, in complex joints like the sacroiliac joints. Each assessment is based on different methods. Various methods have been used. A significant problem is a lack of description of some of the methods used which leads to variation in interpretation and application of such methods. Therefore, lack of knowledge on how to correctly score is also an important source of variability. The quality of training of the operator or reader is also of utmost importance, and standards or targets set for what constitutes acceptable scoring proficiency should be clearly defined in advance and consensually agreed upon by experts.

Finally, other sources of variability may appear during the examination, due to the system set up (e.g. artifacts due to the superposition of imaged structures in X-Rays), the patient (e.g. movement during an MRI examination, physical characteristics such as body mass index or deformity), or both (e.g. the positioning of the ultrasound probe influencing the trajectory of the ultrasound beams; position and movement of a patient during an ultrasound examination) (23-24). These interactions and their associated variability need to be accounted for before a score based on a given technique can be developed, accepted and evaluated as an imaging outcome measurement instrument in research.

The same is true for instruments based on PROs. Looking at the literature on questionnaire-based instruments that assess at the impact of paper and pencil versus computer interfaces, similar concerns over undesired sources of variability in scores can be noticed. Therefore, the much more explicit attention to sources of variability in imaging outcome measurement instruments is beneficial for other types of instruments as well, such as clinically assessed outcomes (i.e. joint counts, enthesitis assessments) or PROs. (13). In all cases, what we want is to measure the true state (presence/absence/extent) of the manifestation, in the absence of errors.

Table 2 shows in details the mains sources of variability affecting imaging outcome measurement instruments and the errors they generate, as well as possible solutions to be applied to reduce such

variability. **Table 3** shows how the same variability aspects, and the same errors, may affect instruments based on PROs, as well as possible solutions to apply.

Source	Specifics	Error type	Example	Solution
Target Domain	Theoretical and operational definitions	Systematic	Biased definitions	improve definitions through repeated exercises
Imaging Method	Technical characteristics	Systematic	Biased equipment (incorrect calibration)	Use quality equipment, proper calibration
	Machine parameters	Systematic	Bias in probes, settings, doses, etc.	Use quality equipment, proper procedure
	Scoring manual	Systematic	Unclear understanding of scoring method	Clear description of scoring framework
Reader/Technician	Definition of target domain	Systematic	Biased understanding of target domain	Evidence-based agreement on definitions
	Acquisition	Random	Unclear protocol, unreliable application, lack of training	Improve protocol, institute training
	Interpretation	Random	Unreliable interpretation	Establish reference standards, e.g. atlas, training
Patient/Disease	Physical Characteristics	Systematic	Improper positioning due to joint deformities	Adapt procedure
	Disease characteristics	Random	Variability in spectrum of disease/detection threshold	Use mean of multiple measurements, appropriate patient selection

Table 2. Sources of variability affecting Imaging Outcome Measurement Instruments

Source	Specifics	Error type	Example	Solution
Target Domain	Theoretical and operational definitions	Systematic	Biased definitions, domain captured by instrument differs from the target domain	Improve definitions through repeated discussions, repeated exercises
PRO Approach	Technical characteristics	Systematic	Biased wording: positive versus negative wording	Look at wording of questionnaires
	PRO approach parameters	Systematic	Biased scoring: computer- versus paper-based	Test to estimate bias, use only one consistent approach
Gathering Information	Acquisition	Random	Unclear instructions for completion of questions, e.g., how to answer if you are using an adaptive device	Improve protocol, pilot test instructions to assess understanding
	Environment	Random/ Systematic	Noisy environment may cause poor attention to details	Provide consistent, quieter area for completion of questionnaires
	Guessing	Systematic	When questions are unclear people will guess	Pilot test questionnaire to ensure good understanding
Patient/Disease	Physical Characteristics	Systematic	Joint damage may preclude change in function	Be sensitive to patient population and adaptive techniques
	Disease characteristics	Random	People with longer term disease more likely to adapt techniques, e.g., cane use	If this is an issue, consider it at selection

Table 3. Sources of variability affecting Patient Reported Outcome Measurement Instruments

Sources of variability have the effect of either decreasing precision or leading to a mis-estimation of the outcome. For example, if one examiner consistently over-estimates the amount of synovitis during an ultrasound examination, his/her results will probably differ from those of another examiner. Reducing the impact of the source of variability is important and becomes part of operational guidelines for the application of an imaging instrument. Possibilities include having two independent assessments and using the average of measurements (which cannot avoid ascertainment bias), using only one highly skilled examiner for the whole trial, or training all examiners until their inter-rater agreement is sufficient. Variability of imaging data acquisition should be minimized through standardization of the protocol and operator training, and variability of imaging data interpretation through the use of consensual definitions, pre-reading or pre-interpretation, calibration sessions, utilization of imaging atlas, and through several steps of reliability sessions.

Imaging outcome measurement instruments make these sources of variability quite transparent. This step is still based on the evaluation of the influence of each potential source of variability in test-retest and reliability studies allowing assessing the impact of these sources on the instrument performance. Articulating efforts to identify and mitigate unwanted variability will be part of the very early stages in developing and selecting an instrument because they form the basis for using imaging as outcome measurement instrument. The same exercise of identifying and reducing sources of variability is important for any type of outcome.

Lesson 3. How OFISA should be applied in the selection of imaging measurement instruments for core outcome measurement sets

OFISA is a staged decision-making tool to guide core set developers through evidence-based decisions on the selection of outcome measurement instruments. OFISA encompasses the traditional elements of the original OMERACT Filter (6): Truth, Discrimination and Feasibility. In OFISA these are applied in a new order where Truth is divided into two criteria: i. Truth-1 evaluating what the instrument is be able to capture in the target population (traditionally the properties of Face and Content Validity), and ii. Truth-2 which corresponds to the more traditional evaluation of whether the instruments score relates to other instruments (Construct (and rarely, Criterion) validity). In between the two is the Feasibility assessment, a very practical review of the usability of the instrument (burden, equipment needs, access, scoring framework and complexity, availability of validated knowledge transfer tools).

In OFISA, a decision is made after the assessment of domain match (Face and Content Validity) and feasibility as to whether to continue to assess the measurement properties of an instrument in depth. Failure to meet the Truth-1 and Feasibility criteria in OFISA means that the instrument assessment should not progress for further consideration as nothing can repair poor content in a questionnaire or an insurmountable hurdle in access to a particular pathophysiological process with an imaging technique.

In practice for imaging instruments, the issue of problematic feasibility only truly becomes apparent when the instrument is tested for reliability between readers and repeated testing demonstrates repeated failure to attain adequate reliability. For these reasons these two steps go together, even if the feasibility aspects may be evaluated also later in the evaluation process. The next step in assessment is Truth-2

where Construct (and wherever possible, Criterion) validity is assessed. OFISA then continues to look at the elements of discrimination, typically test-retest reliability studies in large samples, responsiveness (Longitudinal Construct validity), and discrimination between groups such as we would find in a clinical trial and establishing thresholds of meaning.

The current OFISA, developed for clinical and patient-centered outcome instruments, with minor adjustments can be therefore applied to complex instruments such as imaging measurement instruments and in so doing improve the process for all types of instruments. **Figure 2** shows where the first 2 lessons should be included in OFISA.

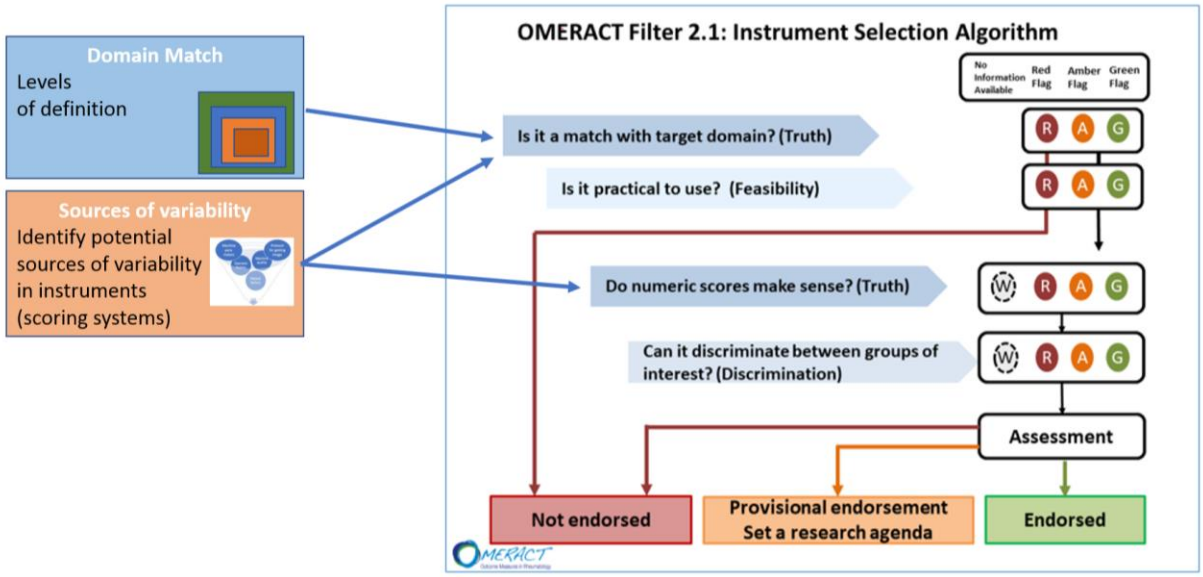

Figure 2. Where lessons 1 and 2 should be evaluated in OFISA

Discussion

In the process of reviewing outcome measurement instruments that are used in the broad field of imaging and how they move through the OMERACT instrument selection process, 2 core lessons emerged that are applicable to all OMERACT instruments and all OMERACT selection domains.

First, a good instrument selection process must be preceded by a clear definition of what we want to measure and of what we are able to measure. The new template (**Figure 1**) for detailed domain definitions should help to improve this process and considered the end product of the creation of a core domain set.

Second, sources of variability (noise and bias) can, and should, be identified at the domain definition level or when a candidate instrument is identified. Noise (artifact) can affect reliability and should be minimized by proper procedures, training, and use of reliability exercises or repeated measurements. Bias can affect validity, and can be minimized through use of high-quality equipment, proper calibration, high quality

definitions of the target lesions, and proper training in the application of the instrument, and defined study design. Identifying sources of variability is now part of the instrument selection process, adding it onto the domain definition template.

Finally, the OMERACT OFISA has been updated in Filter 2.2 to include the work presented here. The changes to OFISA associated with identifying sources of variability will also help in identifying factors that could be influencing the score obtained on an outcome instrument.

The adjustments to the instrument selection process described above, complement and clarify the work made by Easson et al (14), in reviewing key aspects of measurement properties for imaging outcomes. It addresses a gap identified when we tried to apply a PRO-oriented instrument selection framework to an imaging outcome. The lessons work both ways: we learned that elements prominent in imaging outcomes also apply to other types of measurement, and updated our methodology accordingly; and the imaging field, with its strong focus on application in patient care, can benefit from our methodology to properly define outcome domains, and to appraise and select instruments for research. Although this has not yet been tested, we feel that this update to the OMERACT instrument selection process prompted by imaging makes it robust for other biomarkers such as soluble and tissue biomarkers, but it is also of relevance to clinician observed outcomes such as pulmonary function testing or joint counts. Our work reinforces the idea that the underlying principles are the same regardless of the way in which information is obtained: it is necessary to gather enough validity evidence to build a case for using that instrument to represent the target domain in a core domain set.

Conclusion

This work has reinforced the need for a detailed definition of the target domain within the broad domain of interest before the instrument selection can begin. It also broadens the number of measurement properties examined in the OMERACT Filter 2.2 to include examination of sources of variability and the solutions applied to reduce that, and to include comparison to gold standard(s) (Criterion validity evidence). OMERACT instrument selection processes have been made clearer by reviewing it in the context of imaging outcome measurement instruments. The involvement of several stakeholders in this conceptual methodological process, including patient research partners, has improved understandability through the use of tools and educational videos to support new changes. The new OMERACT Filter 2.2 Instrument Selection Algorithm described in the OMERACT Handbook incorporates these improvements for all outcome measurement instruments under consideration for OMERACT endorsement.

Acknowledgements

PGC is supported in part through the NIHR Leeds Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

This article was prepared by Victor Sloan in his personal capacity. The views expressed are his own and do not reflect the views of the Peace Corps or the US Government.

Thank you to the workshop participants including the facilitators, content experts and reporters:

Catherine Hofstetter

Dorcas Beaton

Tarimobo Otobo

Sofia Ramiro

Bev Shea

Karine Toupin April

Désirée van der Heijde

Victor Sloan

Lyn March

Lene Terslev

Simon Décary

Francesca Ingegnoli

Jeffrey Chau

Lee Simon

Walter Maksymowych

Mikkel Østergaard

Alexa Meara

Robin Christensen

Maarten de Wit

Maria Stoenoiu

Anne Boel

Maarten Boers

Robert Landewe

Dan Zhao

Enrico De Lorenzis

Vibeke Strand

Mariana Ivanova

Win Min Oo

Deb Constien

Annette Mckinnon

Kei Ikeda

Philip Conaghan

Michael Backhouse

Sarah Mackie

Alison Hoens

Grayson Schultz

Zahi Touma

Peter Tugwell

Marion van Rossum

Bing Bingham

Tom Buttel

Annelies Boonen

Andrea Doria

Sabrina Mai Nielsen

Corrado Campochiaro

Raouf Hajji

Charles Goldsmith

Ihsane Hmamouchi

Thasia Woodworth

Niti Goel

Mikael Boesen

Sam Michel Cembalo

Ingrid de Groot

Codruta Zabalan

Manouk de Hooge

Declarations of interest

Dorcas Beaton, Paul Bird, Maarten Boers, Sam Michel Cembalo, Philip Conaghan, Maria Antonietta D'Agostino, Maarten de Wit, Alison Hoens, Catherine Hofstetter, Lara Maxwell, Win Min Oo, Marion van Rossum, Teodora Serban, Lene Terslev, Codruta Zabalan report they have nothing to disclose.

Andrea Doria reports: Advisory Board(s) International Prophylaxis Study Group (not for profit), OMERACT SIG in MRI in JIA (not for profit), Research Support Baxalta-Shire (Research Grant), Novo Nordisk (Research Grant), Terry Fox Foundation (Research Grant), PSI Foundation (Research Grant), Society of Pediatric Radiology (Research Grant), Garron Family Cancer Centre (Research Grant).

Robin Christensen reports: The Parker Institute is grateful for the financial support received from public and private foundations, companies and private individuals over the years. The Parker Institute is supported by a core grant from the Oak Foundation; The Oak Foundation is a group of philanthropic organizations that, since its establishment in 1983, has given grants to not-for-profit organizations around the world. Prof. Christensen is a founding member of the Technical Advisory Group of OMERACT, an organization that develops outcome measures in rheumatology and receives arms-length funding from 12 companies.

Walter P. Maksymowych is Chief Medical Officer of CARE Arthritis, a company that develops and validates imaging-based scoring systems for application in clinical trials and basic and clinical research of arthritis disorders.

Victor Sloan reports stock for service as member of Board of Directors of Oncopath Genomics LLC outside the submitted work and as CEO of Sheng Consulting LLC I provided clinical research consulting services to several pharmaceutical companies involved in developing new therapies for autoimmune disease.

Mikkel Østergaard reports grants, personal fees and non-financial support from AbbVie, grants, personal fees and non-financial support from BMS, personal fees from Boehringer-Ingelheim, personal fees from Eli Lilly, personal fees and non-financial support from Janssen, grants, personal fees and non-financial support from Merck, personal fees and non-financial support from Pfizer, personal fees and non-financial support from Roche, grants, personal fees and non-financial support from UCB, grants and personal fees from Celgene, personal fees from Sanofi, personal fees from Regeneron, grants, personal fees and non-financial support from Novartis, personal fees from Gilead, outside the submitted work.

Funding:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745-53

2. Boers M, Beaton DE, Shea BJ, Maxwell LJ, Bartlett SJ, Bingham CO, et al. OMERACT Filter 2.1: Elaboration of the Conceptual Framework for Outcome Measurement in Health Intervention Studies. J Rheumatol. 2019 Aug 1;46(8):1021.

3. Maxwell LJ, Beaton DE, Shea BJ, Wells GA, Boers M, Grosskleg S, et al. Core Domain Set Selection According to OMERACT Filter 2.1: The OMERACT Methodology. J Rheumatol. 2019 Aug 1;46(8):1014.

4. Beaton DE, Maxwell LJ, Shea BJ, Wells GA, Boers M, Grosskleg S, et al. Instrument Selection Using the OMERACT Filter 2.1: The OMERACT Methodology. J Rheumatol. 2019 Aug 1;46(8):1028.

5. Beaton DE, Boers M, Tugwell P, Maxwell LJ. (2021). Chapter 36: Assessment of Health Outcomes. In: Firestein GS, Budd RC, Gabriel SE, Koretzky FA, McInnes IB, O'Dell JR (eds). *Firestein and Kelley's Textbook of Rheumatology*, 11th Ed. Vol.1, pp. (522-535). Philadelphia, USA: Elsevier.

6. Boers M, Brooks P, Strand V, Tugwell P. The OMERACT Filter for outcome measures in rheumatology. J Rheumatol. 1998. 25:198-9.

7. D'Agostino MA. Evaluation de l'apport de l'échographie des enthèses (mode B et Doppler puissance) au diagnostic de spondylarthrite. PhD Thesis in Clinical Research, Methodological Innovation and Public Health, 2008. 238 pages. Available from: https://www.theses.fr/2008PA11T046

8. Terslev L, Naredo E, Keen HI, Bruyn GAW, Iagnocco A, Wakefield RJ, et al. The OMERACT Stepwise Approach to Select and Develop Imaging Outcome Measurement Instruments: The Musculoskeletal Ultrasound Example. J Rheumatol First Release April 15 2019; doi:10.3899/jrheum.181158

9. Costantino F, Carmona L, Boers M, Backhaus M, Balint PV, Bruyn GA, et al. EULAR recommendations for the reporting of ultrasound studies in rheumatic and musculoskeletal diseases (RMDs). Annals of the Rheumatic Diseases. Published Online First: 22 January 2021. doi: 10.1136/annrheumdis-2020-219816

10. Mokkink, L.B., Boers, M., van der Vleuten, C.P.M. et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med Res Methodol 20, 293 (2020). https://doi.org/10.1186/s12874-020-01179-5

11. Handbook Chapter 5: OMERACT Instrument Selection (2021)

12. Hawker G, Davis A, French M, Cibere J, Jordan JM, March L, et al. Development and preliminary psychometric testing of a new OA pain measure – an OARSI/OMERACT initiative. Osteoarthritis and Cartilage.2008. 16 (4): 409-14. https://doi.org/10.1016/j.joca.2007.12.015.

13. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of Electronic and Paper-and-Pencil Administration of Patient-Reported Outcome Measures: A Meta-Analytic Review. Value Health. 2008;11(2):322-33. doi: 10.1111/j.1524-4733.2007.00231.x

14. Easson AM, Tomlinson G, Doria AS. Chapter 4: Measurements: Validity, Reliability, and Responsiveness. p.54-86 In: Doria AS, Tomlinson G, Beyene J, et al. Research methods in Radiology: a practical guide. New York : Thieme Publishers. 2018

15. Obuchowski NA. Special Topics III: bias. Radiology. 2003 Dec;229(3):617-21

16. Obuchowski NA, Subhas N, Schoenhagen P. Testing for interchangeability of imaging tests. Acad Radiol. 2014 Nov;21(11):1483-9

17. Obuchowski NA, Mazzone PJ, Dachman AH. Bias, underestimation of risk, and loss of statistical power in patient-level analyses of lesion detection. Eur Radiol. 2010 Mar;20(3):584-94.

18. Obuchowski NA, Remer EM, Sakaie K, Schneider E, Fox RJ, Nakamura K, Avila R, Guimaraes A. Importance of incorporating quantitative imaging biomarker technical performance characteristics when estimating treatment effects. Clin Trials. 2021 Apr;18(2):197-206.

19. Obuchowski NA, Mozley PD, Matthews D, Buckler A, Bullen J, Jackson E. Statistical Considerations for Planning Clinical Trials with Quantitative Imaging Biomarkers. J Natl Cancer Inst. 2019 Jan 1;111(1):19-26.

20. Obuchowski NA, Mazzone PJ, Dachman AH. Bias, underestimation of risk, and loss of statistical power in patient-level analyses of lesion detection. Eur Radiol. 2010 Mar;20(3):584-94.

21. Obuchowski NA, Subhas N, Polster J. Statistics for Radiology Research. Semin Musculoskelet Radiol. 2017 Feb;21(1):23-31. doi: 10.1055/s-0036-1597252. Epub 2017 Mar 2. PMID: 28253530.

22. Obuchowski NA. Reducing the number of reader interpretations in MRMC studies. Acad Radiol. 2009 Feb;16(2):209-17. doi: 10.1016/j.acra.2008.05.014. PMID: 19124107.

23. Dachman AH, Obuchowski NA, Hoffmeister JW, Hinshaw JL, Frew MI, Winter TC, Van Uitert RL, Periaswamy S, Summers RM, Hillman BJ. Effect of computer-aided detection for CT colonography in a multireader, multicase trial. Radiology. 2010 Sep;256(3):827-35. doi: 10.1148/radiol.10091890. Epub 2010 Jul 27. PMID: 20663975

24. Obuchowski NA. Reference values: no need for confusion. J Thorac Cardiovasc Surg. 2009 Jun;137(6):1572-3. doi: 10.1016/j.jtcvs.2009.02.031. PMID: 19464491.

Appendix 1: OMERACT template for detailed domain definitions from the technique perspective

Term, description	Example	Example	Example
Core Area One of the Core Areas as defined in Boers, 2019[1]. Each core area in the framework has a specific function, and together they contain the whole "universe" of domains (concepts) that one could conceivably measure to assess the effects of an intervention.	Pathophysiological manifestation	Pathophysiological manifestation	Life Impact
Broad Domain General term, description of a concept.	Inflammation	Pain	Pain
Target Domain Focused description of the domain that can be placed in the OMERACT Onion.	Synovitis (= Synovial inflammation)	Intensity of pain	Pain impact on daily activities
Working definition of target domain A more precise definition that can be used to develop or assess instruments. This might fit in as the definition used in the Delphi survey.	Ultrasound-detected synovitis (= hypoechoic synovial hypertrophy which may exhibit Doppler signal with synovium)	The daily average of the intensity of the sensation of pain expressed on a range from no pain to worst pain imaginable	A sense of the degree to which people are impacted by pain in terms of the accomplishment of daily activities and roles other than paid work
Domain components What components are essential to capture the content of the target domain? <u>Theoretical definition</u>: The conceptual definition of the domain components according to the technique used to measure it.	*Theoretical definition:* Ultrasound-detected synovitis is characterized by hypoechoic synovial hypertrophy and Doppler signal	*Theoretical definition:* Consider a 24-hour window of pain intensity during an average day; exclude special activities or during special events	*Theoretical definition:* Consider the impact of the pain on selfcare, leisure, social roles at home (e.g. parenting), but not work roles. General impact over whole day rather than specific time of day (e.g. morning or

Term, description	Example	Example	Example
Operational definition: The definition of the domain components described in terms specific to the technique used to measure them.	*Operational definition:* Ultrasound synovial hypertrophy is defined as a hypoechoic thickening of the synovium	*Operational definition:* 24-hour average pain intensity defined over a range that incorporates no pain to worst pain sensation imaginable	night) *Operational definition:* Impact of pain defined over a range that incorporates no impact on ability to unable to perform a task over a range of items as described above

1. Boers M, Beaton DE, Shea BJ, Maxwell LJ, Bartlett SJ, Bingham CO, et al. OMERACT Filter 2.1: Elaboration of the Conceptual Framework for Outcome Measurement in Health Intervention Studies. J Rheumatol. 2019 Aug 1;46(8):1021.