Fundamental Outcome Measurement: Selecting Patient Reported Outcome Instruments and Interpreting the Data they Produce

Stephen P McKenna, PhD[1]; Alice Heaney[1]; Paul C Langley, PhD[2]
[1]Galen Research, Manchester UK; [2]University of Minnesota, Minnesota MN USA; Maimon Research, Tucson AZ USA

ABSTRACT
Over the past 40 years literally thousands of generic and disease specific patient reported outcome (PRO) instruments have been developed. While most were developed for a specific study and were never used again, there is still the question of how manufacturers and others should select a PRO instrument for a study. These studies may be clinical pivotal trials or observational tracking studies to support therapy response. Formulary committees also need to be able to interpret PRO data to make decisions about whether to accept claims for therapy response. It is possible to argue that the many different approaches to outcome measurement have resulted from the lack of agreed methodologies. However, a more likely explanation is that the authors have failed to apply the axioms of fundamental measurement when creating their measures. The result is a plethora of ordinal PRO instruments that inform little about the impact of interventions. Clinical trials rarely report PRO data. Where they do, analyses are generally restricted to (for example) changes in the experimental group's scores. Comparisons between the treatment and placebo groups or between active groups are infrequently reported, most likely due to the failure of the instrument to show differences or changes in outcome. This is unfortunate as it means no assessment is made of the value that patients gain from the intervention. This commentary is intended to make researchers and formulary committees aware of the issues that need to be addressed when selecting PRO instruments for a study or evaluating publications and claims for therapy response. The latter is crucial as reported data influence the selection of medicines and healthcare products. In the latter case a particular concern is with PRO claims embedded in simulation models.

INTRODUCTION

A patient-reported outcome (PRO) instrument assesses outcomes directly reported by the person concerned. Not all PRO instruments are the same. They differ in terms of the type of outcome they report. These can range from symptoms, functioning, health status, health-related quality of life (HRQoL), quality of life (QoL), preference measures and satisfaction with treatment. These different purposes require different types of PRO instrument. It is crucial to be clear about the reasons for administering a PRO instrument, especially when selecting one for use in a clinical study or trial. Equally importantly, where a manufacturer makes claims for their product to a formulary committee, the members of the committee should have the ability and forensic skills needed, to assess the merits of the claim in terms of the design of the PRO instrument and the metric that is being used.

If response to therapy is judged by a PRO instrument that neglects or ignores the axioms of fundamental measurement, it results in a major disservice to patients and physicians. This is an example of health technology assessment's failure to embrace measurement standards common in the physical sciences. Clinical outcome studies are most likely to use measures of health-related quality of life (HRQoL) or quality of life (QoL). The purpose of a PRO instrument should be clear from the conceptual model underlying its content. Unfortunately, few instrument developers specify the conceptual model they adopted for instrument development. Without a conceptual model it is not possible to validate the instrument as its validity should prove that the model functions as intended [1]. Consequently, PRO instruments rarely have the quality to be effective in clinical trials.

Corresponding author: Paul C Langley, PhD
Adjunct Professor, College of Pharmacy
University of Minnesota, Minneapolis, MN
Director, Maimon Research LLC; Tucson, AZ
Email: langley@maimonresearch.com
Website: www.maimonresearch.net

THE NATURE OF PRO INSTRUMENTS

PRO instruments generally collect information that clinicians consider important. This is not surprising as most instruments are authored by clinicians Health Related Quality of Life (HRQoL) instruments collect information about symptoms and functional impairments. Their use in clinical trials is somewhat unusual. If these are important for assessing change with treatment, they should already be collected for the study. Instrument authors often refer to such measures as being patient-based, but this is rarely the case. While the information generated is provided by patients this does not mean that it is of concern to them. For PRO instruments to be patient-based (or patient-centric) the content of the questionnaire should be generated from patients. If these are qualitative patient interviews, it guarantees that the content of the instrument is patient-based and of concern to them.

PRO instruments may be generic or disease-specific. The first wave PRO instruments were all generic, HRQoL multiattribute instruments, intended to be used with any type of disease. They are well known and have been widely accepted: the SF-36, EuroQol) and Nottingham Health Profile (NHP)[2 3 4]. Except for the NHP, these measures were developed in the 1970's and 1980s from pre-existing questionnaires. Surprisingly, they are

DOI: https://doi.org/10.24926/iip.v12i2.3911

still widely used in clinical trials, but with little purpose, as their ability to detect change is limited. This is not surprising as their content is generally inappropriate for a specific disease. Furthermore, they cannot cover the issues that truly matter to each specific disease population[5]. For these reasons, they are not popular with patients or health authorities. They remain in use as they are well known and have been used in many previous trials – despite failing in their purpose. Since they were developed, the science of outcome measurement has moved on substantively. Outcomes are more clearly defined and new measurement models have been introduced to produce more reliable and valid measures.

The generic HRQoL instruments also continue to be used as it is thought, mistakenly, that they allow comparisons to be made about outcomes in different conditions and between healthy and diseased populations. A healthy person may agree that they are feeling tired. However, this experience is very different from that of a respondent with rheumatoid arthritis. It is noticeable that no new generic measures are being produced.

There is also the belief that generic instruments can be used as markers for resource allocation within health systems. That is, if the assumption is made that the health budget is fixed, then the view (at least in theoretical terms) is that resources should be allocated so that the last dollar spent in a disease area yields the same benefit as the first. This leads to models focused on incremental cost per QALY claims. Unfortunately, these fail the standards of normal science as the QALY is an impossible mathematical construct and the claims are non-evaluable [6].

All the PRO instruments developed in recent years are disease-specific but few are patient-centric. Ideally, such instruments should ask relevant questions and omit issues that are not of concern to the specific population. Consequently, the new PRO instruments could have the potential to be able to detect real changes related to treatment more reliably. Unfortunately, the views and interests of clinicians continue to dominate the content of outcome measures.

Where a new PRO instrument has been developed that is patient-centric and disease-specific, there remain several additional requirements if it is to be of value in clinical trials and to health system committees evaluating products for formulary listing.

The first is that it should be acceptable to patients. This quality is more likely to occur when the questions are generated directly from relevant patients. The questionnaire needs to be carefully designed with simple response formats. This quality is rarely found in PRO instruments.

Second, the PRO instrument must be unidimensional i.e. should report on only one attribute. Where different attributes or variables are added to make a single score, the result is an invalid multidimensional composite instrument. Multiattribute data should be reported separately and presented as a profile of the different attributes measured. Unfortunately, if they are bundled together to create a single score, it is not possible to judge which of these attributes is most important to the respondents.

Third, certain qualities are traditionally required of a PRO instrument. It needs to be reliable, implying that it has little measurement error. Internal consistency is not a measure of reliability. Reliability is generally assessed by applying the questionnaire to a sample of patients on two occasions approximately two weeks apart and correlating the scores. Where the reliability coefficient is 0.7, it means that half of what the instrument measures is error. This limits its ability to detect changes in outcome. It is recommended that a minimum reliability coefficient of 0.85 is required. Reliability is a key factor in determining the validity of the measure.

Fourth, PRO instrument authors should provide evidence of construct validity. This is required to indicate that the instrument is measuring what was intended. It is also essential to confirm that the instrument's conceptual model is appropriate. Several methods can be used to determine validity. Common methods include known group validity - whether the instrument can distinguish between groups that would be expected to differ in score and convergent validity, whether the new instrument correlates as expected with a similar measure or one that assesses a related outcome. Validity is never absolutely proven but can be built on as the new measure is used in clinical studies.

Finally, PRO instruments need to be responsive, i.e. able to detect real change associated with effective interventions. Responsiveness is dependent on the quality of the content of the instrument and its reliability. Various estimates of responsiveness are used but no method is convincing. The smallest detectable difference (SDD) can be informative but for ratio or interval but not ordinal scales. It is calculated using the reproducibility and standard deviation of the measure and indicates the change in score necessary to be statistically detected in a study[6]. Few PRO instruments can detect changes with treatment as large as their SDD value and are unlikely to be effective in a clinical trial or to provide data required for formulary listing;

If a generic, multiattribute instrument is used in a simulation model that is generating incremental cost-per-QALY claims, it should come as no surprise that incremental QALY differences between products over the lifetime of the model are minimal, although as they are ordinal scores, the differences are meaningless. The unfortunate outcome is that this may result in claims for price discounting and patient access that are unrealistic[7].

DOI: https://doi.org/10.24926/iip.v12i2.3911

MEASUREMENT THEORY

Following from the work of Stevens in the 1940s four main types of measurement scale are generally recognized. However, more recently they have been referred to as a hierarchy in the use of numbers[8 9]. Nominal or categorical scales report on distinct variables such as gender (male/female) where each variable has an equal value, so no numbers are involved. No statistical analyses can be conducted on such scales. However, tests (such as the non-parametric Chi square) comparing the number of entries in different categories are possible.

Ordinal scales show the order of responses to latent variables such as satisfaction, happiness or pain. However, they do not inform on the distance between scores on the instrument. It is not valid to calculate total scores, means or standard deviations with ordinal scales and parametric statistical tests should not be used. Despite this, it is common practice to report such statistics derived from ordinal scales such as the EQ-5D-3L. Most PRO instruments yield ordinal data and consequently, are limited in how they can be used. Perhaps the best example of the misapplication of ordinal scores is in the construction of QALYs. This requires multiplying scores on an ordinal scale by time. Ordinal scales cannot support multiplication or division, let alone addition and subtraction[8].

Interval scales show both the order of items in a scale and the distance between these variables. Valid means and standard deviations can be calculated with interval scales and parametric statistical tests can be used with data they generate. Addition and subtraction are possible with interval scales, but not multiplication and division.

A ratio scale is like an interval scale but has a meaningful or convenient zero point that no values can fall below. Ratio scale data can be multiplied or divided by other variables; for example, distance travelled divided by time gives speed. Few PRO instruments measure at the ratio level. Ratio scales also require a meaningful zero point.

Traditional measurement of outcomes (classical test theory; CTT) is generally used to analyze PRO data. However, as most of these data are ordinal, only weak analyses are possible. Despite this, the data are usually treated as if they were interval or ratio data. Consequently, published findings are of little and questionable value[10].

Additional types of scales are possible by applying modern measurement techniques. Several PRO instruments have been developed using Rasch Measurement Theory (RMT), where the performance of individual items is tested rather than looking at the whole outcome instrument at the same time[1].

RASCH MEASUREMENT THEORY

RMT is of particular value in outcome instrument development because it recognizes the importance of measurement theory and produces fundamental measures[11]. Despite this, it is rarely used for instrument development. RMT has a critical role to play in measuring latent attributes. The technique employed by RMT is conjoint simultaneous measurement, independently developed in the early 1960s by Luce and Tukey, and Rasch[12 13]. In certain circumstances, RMT allows transformation of ordinal scales to interval level measurement, and in some instances bounded ratio scales[14].

The main advantages of RMT are that it creates interval scales that are unidimensional and provide fundamental measurement. This latter quality means that a score on the measure provides all the information required to judge a respondent's performance. No other measurement model can achieve these qualities. RMT has additional valuable properties. It identifies items that misfit the scale and that need to be removed. It looks at whether the response format used by the questionnaire works as intended. Local item dependency analyses identify problems associated with the relations between different items. RMT can also check whether the measure works in the same way with different groups of people (for example males and females or young and old respondents). This is called differential item functioning and is important in specifying the populations required for clinical trials. For example, in the development of the Alzheimer's Patient Partner's Life Impact Questionnaire (APPLIQue) it was shown that spousal caregivers were affected differently from other family caregivers[15 16]. Grouping all caregivers together in a clinical trial would give meaningless results.

TYPES OF PROMS

PRO instruments are generally treated as being similar to each other, making selection for use in a study an apparently straightforward task. However, different types of PRO instruments are available. The most widely used PROs assess HRQoL. These PROs address issues that are primarily of interest to clinicians. Indeed, most of them were developed by clinical experts. They address a range of symptoms (such as pain, anxiety and fatigue) and functional limitations, including ability to walk, socialize and work. Such measures should be reported as a profile as they are all different types of outcome. Each outcome should be unidimensional. Unfortunately, it is common for test developers to add together the different outcomes to produce invalid composite scores[1 17 18].

QoL measures are a different type of PRO measure. They are designed to assess issues that are of major concern to patients. Here the intention is to find out how patients value the impact of a disease and its treatment on their lives. QoL measures can also differ according to the conceptual model they employ. The most used conceptual model in QoL measurement is the patient-centric need-based approach[19]. QoL is seen as a unidimensional latent construct that measures the extent to which respondents feel that their human needs are met. While interventions may improve symptoms and functioning these variables may not change the extent to which patients are able

DOI: https://doi.org/10.24926/iip.v12i2.3911

to meet their needs. This is a holistic approach to outcome measurement. Contents of the instruments are generated directly from patients, ensuring that the measures are relevant to all respondents. Other types of PRO instruments include measures of satisfaction, utility and health status. These have different requirements for development and validation but if they are of interest must meet the required standards of fundamental measurement.

SYSTEMATIC REVIEWS OF PRO INSTRUMENTS

Considerable care should be exercised when consulting systematic reviews of PRO instruments. They are often written by people who do not have the necessary background or experience in outcome measurement. As there are no agreed requirements of measures in the literature, this leaves reviewers to decide what they consider important. As there are different types of PRO instruments, these should be developed and evaluated in different ways. Too often, older measures are considered of high quality, reflecting their often uncritical acceptance, when measures developed using modern measurement methods are ignored or inappropriately evaluated. Attempts to standardize reviews, notably the COnsensus-based Standards for the selection of Health Status Measurement INstruments (COSMIN), fail on several grounds [20]. COSMIN represents an attempt to utilize CTT criteria as the basis for grading selected outcome measures. The COSMIN checklists fail to appreciate the need to meet the axioms of fundamental measurement in instrument development as well as failing to appreciate the contribution of RMT to instrument development. The key point to recognize is that meeting the axioms of fundamental measurement must precede any statistical or psychometric analysis. Unfortunately, this is conspicuous by its absence in the COSMIN checklist and in all too many other attempts at systematic reviews of instruments utilized in disease areas.

An example of inconsistent reviews is provided by three reviews of foot and ankle questionnaires, two of which adopted the COSMIN checklist. In the first, Eechaute et al concluded that the qualities of the measures they reviewed were fair to poor [21]. Despite this they judged the Foot and Ankle Disability Index (FADI) and the Foot and Ankle Activity Measure (FAAM) the most appropriate. The second, review by Sierevelt et al, rated the Foot and Ankle Outcome Score (FAOS) and the FAAM promising outcome measures[22]. However, they also warned that these measures have shortcomings that should be considered when interpreting results in clinical settings or trials. In the third review, Jia et al reviewed fifty foot and ankle-specific instruments[23]. They reported that most of these had limited evidence of quality. They did not rate the FAOS or FAAM very highly but concluded that the Manchester-Oxford Foot Questionnaire (MOXFQ) was the most appropriate. Unfortunately, all three reviews left out important information about how the measures were developed including the item reduction process. While it would be expected that reviews would come to the same conclusion, this was clearly not the case. Rather than just accept what authors say in their instrument development articles, it is essential that published PRO instrument data claims are evaluated by the reviewers. While this would be time consuming, it would be expected to improve the quality of systematic reviews. These three reviews failed to address the issues that are critical to evaluating the quality of PRO instruments.

ISSUES TO CONSIDER WHEN SELECTING PRO INSTRUMENTS OR INTERPRETING DATA THEY GENERATE

There are several questions that should be asked for both generic and disease specific instruments when:

- selecting a PRO instrument for use in a clinical study;
- evaluating claims made using PRO instrument data;
- supporting formulary reviews for comparative product submissions; and,
- selecting PRO instruments as targets in value contracting.

The key questions are:

- What do you intend to measure?
- What does the PRO instrument measure?
- Is the instrument patient- or clinician-centric?
- How were the items in the instrument generated?
- Are the items specific to the disease being measured or are they generic?
- Has the instrument been tested with relevant respondents?
- Do the authors report the instrument's reproducibility?
- How strong is the evidence of construct validity?
- Does the instrument measure at the ordinal, interval, or ratio level?
- Did the developers apply modern measurement techniques – preferably RMT?
- Did authors report evidence of internal validity?
- Did authors report the effectiveness of the response format?
- Did authors report item fit?
- Did authors report the evaluation of local item dependency?
- Did authors report assessment of differential item functioning?
- Did authors report overall assessment of fit to the Rasch model?
- What assessments of responsiveness did authors report?

It is of interest to note that while there are a few formulary submission guidelines, for example the Academy of Managed Care Pharmacy, *Format for Formulary Submissions* (Versions 4.0 and 4.1), none ask the manufacturer making the submission to provide a review of the PRO instruments utilized in the pivotal clinical trials[24]. Of particular concern are the Consolidated Health Economic Evaluation Reporting Standards

DOI: https://doi.org/10.24926/iip.v12i2.3911

(CHEERS) developed by a taskforce set up by ISPOR, that reported in 2013[25] Questions of the measurement standards and process of instrument development for response assessment are conspicuous by their absence. This is perhaps not surprising given the commitment by ISPOR (and groups such as the Institute for Clinical and Economic Review [ICER] in the US) to invent approximate evidence through the construction of lifetime simulation models based on ordinal data generated from multiattribute generic outcome instruments [5].

The sole exception to this failure to address issues of fundamental measurement in technology assessment is the latest (Version 3) of the Minnesota (Proposed) Formulary guidelines[26]. As part of the review process proposed for formulary committees is a critical assessment of the instruments used to support both generic and disease specific product claims. While not as detailed as those proposed above, the bottom line is that the Minnesota guidelines are quite clear in recommending rejection of any claim from a PRO that does not meet the axioms of fundamental measurement, including conjoint simultaneous measurement where latent attributes such as need-based quality of life are the focus. The Minnesota guidelines propose that the only claims that should be considered are single attribute ratio or interval claims reported on separately, meeting standards for credibility, empirical evaluation and replication.

DEVELOPMENT OF A QUALITY OF LIFE OUTCOME MEASURE: THE CROHN'S LIFE IMPACT QUESTIONNAIRE (CLIQ)

The intention of the study was to develop a PRO measure suitable for assessing the impact of Crohn's Disease (CD) and its treatment in routine clinical practice and clinical trials: the Crohn's Life Impact Questionnaire (CLIQ)[27]. The conceptual model used for the study was need-based QoL. This model has been used for the development of over 30 disease-specific QoL measures. Qualitative face to face interviews were conducted with CD patients. They were asked to describe how their lives had been affected by CD. Where they mentioned specific symptoms or functional problems the interviewer probed more deeply to understand how these issues influenced need fulfilment. The interviews were recorded and transcribed. Qualitative analyses were conducted on the transcripts to identify items that would inform on the conceptual model.

A draft questionnaire was produced and was evaluated with a new group of CD patients. These interviews were intended to see whether the items were considered relevant, easy to understand and answer, and whether anything important had been omitted. Following these interviews minor changes were made to the items and some found to be unsuitable were removed. The new draft questionnaire was then tested by means of a postal test-retest survey.

Data from the survey were tested to see whether they fit the Rasch model. These analyses showed that the response format worked well. Misfitting items were discarded as were items that were too closely related (duplicated). Checks were also made for differential item functioning and local item dependency. These procedures identified a hierarchy of items that formed an interval scale. Additional analyses were then conducted to establish that the measure was reproducible and had construct validity. The results of these analyses were all published to allow readers to judge the quality of the instrument development and testing. Estimates of responsiveness can be calculated from the reproducibility and standard deviation values reported in the instrument development paper[27].

Sample items from the CLIQ are shown below. Such items differ from those included in HRQoL measures. They are focused on the issues derived from the patient interviews, are specific to CD and are patient-centric. This contrasts with HRQoL instruments that produce ordinal multiattribute scores, failing to recognize the standards of fundamental measurement.

- There is not much fun in my life.
- I feel dependent on others.
- I rarely feel clean.
- I worry about having an accident.
- I only feel comfortable at home.

CONCLUSIONS

This commentary is intended to summarize important aspects of PRO instrument development and evaluation. The development methodology described here represents modern measurement, which ensures that new measures are as responsive and meaningful as possible. Traditional approaches, such as taking items from old instruments, conducting literature searches for potential items, or asking clinicians what they think is important to a patient's QoL have no place in modern measurement. New measures are needed because these traditional methods have not developed quality PRO instruments. The lack of progress in the field of PRO development is a cautionary tale for those either selecting or developing PRO instruments.

Care is recommended when consulting PRO instrument development studies or systematic review articles. They are rarely written by specialists in PRO development. Too often recommendations are limited to the instruments that have been most used. This results in limiting progress in outcome measurement. Measures such as the SF-36 and EQ-5D-3L were developed up to 50 years ago, yet they have never managed, to detect meaningful change resulting from interventions. Their continued use is difficult to explain, despite their manifest failure to measure changes in patient value. Too often they are selected by non-experts because they recognize the name or because they have been used in previous clinical trials, observational studies or even national health surveys, with no thought to their lack of measurement properties. Adoption of modern measurement in instrument development will not be easy.

DOI: https://doi.org/10.24926/iip.v12i2.3911

Conflicts of Interest:
SPM develops QoL measures for use in clinical trials and routine clinical practice. AH develops QoL measures for use in clinical trials and routine clinical practice. PCL is an Advisory Board member and consultant to the Patient Access and Affordability Project, a program of Patients Rising.

REFERENCES

[1] Yorke J, Corris P, Gaine S et al. emPHasis-10: development of a health-related quality of life measure in pulmonary hypertension. *Eur Resp J.* 2014 43: 1106-11

[2] Ware JE, Sherbourne C. The MOS 36-item short form health survey (SF-36): I. Conceptual framework and item selection. *Med Care.* 1992; 30: 473–83

[3] EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. *Health Policy*. 1990; 16:199-208

[4] Hunt, S M; McKenna, S P; McEwen et al. A quantitative approach to perceived health status: A validation study. *J Epidem Comm Health*.1980;34 (4): 281–6

[5] Weiner E, Stewart B. Assessing individuals. Boston: Little Brown, 1984

[6] Langley P. The Great I-QALY Disaster. *Inov Pharm*. 2020;11(3): No 7
https://pubs.lib.umn.edu/index.php/innovations/article/view/3359/2517

[7] Langley P. Nonsense on Stilts - Part 1: The ICER 2020-2023 Value Assessment Framework for Constructing Imaginary Worlds. *InovPharm*.2020;11(1): No.12 https://pubs.lib.umn.edu/index.php/innovations/article/view/2444

[8] Stevens SS. (1946). On the Theory of Scales of Measurement. *Science*, 103(2684), 677-680

[9] Andrich D, Marais I. A Course in Rasch Measurement Theory. *Springer Nature*. Singapore, 2019

[10] Merbitz C, Morris J, Grip J C. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil.* 1989;70(4):308-312

[11] Bond T, Fox C. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 3rd Ed. New York: Routledge, 2015

[12] Luce R, Tukey J. Simultaneous conjoint measurement: A new type of fundamental measurement. *J Math Psychol*. 1964;1(1):1-27

[13] Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1960

[14] Langley P, McKenna S. Fundamental Measurement: The Need Fulfilment Quality of Life (N-QOL) Measure. *InovPharm*. 2021;12(2):No.6 https://pubs.lib.umn.edu/index.php/innovations/article/view/3798/2697

[15] McKenna SP, Rouse M, Heaney A et al. International Development of the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue). *Am J Alzheimer's Disease Other Dementias*. 2020; 35: 1-11

[16] Hagell P, Rouse M, McKenna SP. Measuring the impact of caring for a spouse with Alzheimer's disease: Validation of the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue). *J Applied Measurement.* 2018; 19(3): 271-282

[17] Ferner RE, Thomas M, Mercer G, et al. Evaluation of quality of life in adults with neurofibromatosis 1 (NF1) using the Impact of NF1 on quality of life (INF1-QOL) questionnaire. *Health Qual Life Outcomes*. 2017; 15(1): 34

[18] Brown A, Page TE, Daley S, et al. Measuring the quality of life of family carers of people with dementia: development and validation of C-DEMQOL. *Qual Life Res*. 2019;28(8):2299–2310

[19] McKenna S, Wilburn J. Patient value: its nature, measurement, and role in real world evidence studies and outcomes-based reimbursement. *J Med Econ*. 2018;21(5):474-80

[20] Mokkink L, Terwee C, Knol D et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMJ Med Res Methodology*. 2010; 10: 22

[21] Eechaute C, Vaes P, Van Aerschot L, et al. The clinimetric qualities of patient-assessed instruments for measuring chronic ankle instability: a systematic review. *BMC Musculoskelet Disord*. 2007;8:6

[22] Sierevelt IN, Zwiers R, Schats W, et al. Measurement properties of the most commonly used Foot-and Ankle-Specific Questionnaires: the FFI, FAOS and FAAM. A systematic review. *Knee Surg Sports Traumatol Arthrosc*. 2018;26(7):2059-2073

[23] Jia Y, Huang H, Gagnier JJ. A systematic review of measurement properties of patient-reported outcome measures for use in patients with foot or ankle diseases. *Qual Life Res*. 2017;26(8):1969-2010

[24] The Academy of Managed Care Pharmacy. Format for Formulary Submissions – Guidance on Submission of Pre-approval and Post-approval Clinical and Economic Information and Evidence, Version 4.1. AMCP. 2019

[25] Husereau D, Drummond M, Petrou S, et al. Consolidated health economic evaluation reporting standards (CHEERS)—Explanation and elaboration: A report of the ISPOR health economic evaluations publication guidelines good reporting practices task force. *Value Health*. 2013;16:231-50

[26] Langley P. Value Assessment, Real World Evidence and Fundamental Measurement: Version 3.0 of the Minnesota Formulary Submission Guidelines. *InovPharm*. 2020;11(4): No 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/3542/2613

[27] Wilburn J, McKenna SP, Twiss J, et al. Assessing quality of life in Crohn's disease: development and validation of the Crohn's Life Impact Questionnaire (CLIQ). *Qual Life Res*. 2015; 24: 2279–88