MAIMON WORKING PAPERS No. 18 JULY 2022

EVIDENTIARY STANDARDS FOR PATIENT-CENTERED CORE IMPACT (PC-CIS) VALUE CLAIMS

Paul C. Langley, Adjunct Professor, College of Pharmacy, University of Minnesota, Minneapolis MN

Abstract

Proposals for a patient centered core impact set (PC-CIS) are of little relevance to formulary and health system decisions, let alone patients and providers, unless the elements included in the data set meet the standards of normal science and fundamental measurement. Adhering to these standards will have the effect of focusing on the adequacy of proposed core impact measures, with a filter in place to accept only those that meet the standards not only of the physical sciences but also mainstream economics. and health economics. Fortunately, we are well aware of what the criteria for acceptance and rejection of the core impacts within disease states should be in terms of their required attributes and their relevance for supporting evaluable value claims. Care must be taken to delineate the core impact elements: separately identifying those that are purely clinical from core patient centric impacts, which is turn should be separated from impacts defined in terms of drug utilization and resource utilization. The purpose of this brief commentary is to set out the required standards for core impact patient-centric value claims and the framework for evaluating those claims. The critical issue for patient-centered core impacts is to recognize the constraints imposed by the standards of fundamental measurement for target patient populations within disease areas; unless these constraints are recognized we will fail. The leads to the role of Rasch or modern measurement theory calibration as the framework for patient centric measures of latent traits or attributes. From these perspectives PC-CIS is premature; until we have agreed criteria for measurement for impact or outcomes for clinical, patient-centric and resource utilization as a core set of disease specific instruments, it seems pointless to push forward to a wider scope when the present evidentiary foundation is so weak.

INTRODUCTION

Value claims for pharmaceutical products and devices can only be understood if we are clear about the standards that must be applied in the development, application and evaluation of instruments to capture response to therapy within disease and therapy areas [1]. Certainly, within this potential profile of value claims, patient-centric value claims can play a key role in informing decision makers; but these must not be seen in isolation from the purely clinical claims for a product or device and impact claims in terms of drug utilization and other elements of resource utilization. Value claims will only make sense if they are relevant to decision makers and are captured in formulary submission guidelines.

We have to move away from what are often nebulous statements on the PC-CIS assumed need for patient centered core impact claims that comprise *a patient derived and patient prioritized list of impacts a disease and/or its treatment have on patients (and /or their family and caregivers)* [2]. We must be more specific and avoid a hit list of 'broad and inclusive' short-term and long-term so-called patient-centric health outcomes and other related implications. That way madness lies; nothing will be accomplished. If the focus is on value claims and their evaluation to support pricing and access it is preferable to focus on 'core' claims that are manageable and consistent with the standards of normal science and fundamental measurement. It is more manageable to insist on the application of standards for instrument design for patient-centric measures that reflect patient experience and patient needs. Certainly, clinical consideration and experience are elements in patient experience, but if we are to capture these it should be through specific clinical outcomes, while recognizing that we need holistic measures of patient outcomes. This does not mean that we should consider multiattribute HRQoL instruments such as the EQ-5D-3L/5L. Apart from their limited application in the symptoms covered and response levels for disease states, they are only ordinal composite scores and cannot, therefore, support response claims or quality adjusted life year (QALY) claims [3]. Attempts to justify the place of the EQ-5D-5L, for example, as a critical and breakthrough contribution to measuring health related quality of life (HRQoL) by comparing it to other global health measures and multiattribute instruments, miss the point; you cannot compare ordinal multiattribute scales [4]. But more to the point is the willingness of those who support the EQ-5D-5L to push to one side any notion of the contribution and limitations of fundamental measurement. Certainly, we can point to correlations, but they are as meaningful as using a rain gauge as a latent index of depression severity in a community.

The current thinking for the PC-CIS is to focus on disease or target patient population specific 'impacts', including health outcomes, *but also all other meaningful concepts patients might report*. Indeed, the wish list is considerable including *a variety of life impacts, such as symptoms, function, survival, biomarkers, out of pocket costs, family stresses and much more;* expanding upon the existence of core value sets in many disease states. This is a tall order, and probably overblown, but must be premised on a commitment to the standards of normal science and fundamental measurement for each PC-CIS attribute and its value claim; otherwise PC-CIS will be a wasted effort The purpose of this brief commentary is to detail: (i) the required standards for core outcome or impact claims, whether they are patient centric or not; (ii) to emphasize the critical role played by measurement theory in developing patient-centric measure to support core claims; and (iii) to make clear the distinction between core clinical measures and core patient-centric measures; which are all too often collapsed into a single measure or where the clinical parameters dominate the development of measures such as multiattribute health related quality of life (HRQoL). The take home message is that we have few patient-centric or patient reported outcome (PRO) measures that meet the required standards for normal science and fundamental measurement [5] [6] [7] [8] [9].

NORMAL SCIENCE AND FUNDAMENTAL MEASUREMENT

For those coming from a background and commitment to what is defined as pharmacoeconomics, the application of assumption driven simulations to create imaginary non-evaluable claims for cost-effectiveness, the case that this endeavor, which has lasted for some 30 years, is an analytical dead end may come as a surprise. It is not the role of core impact measures to support lifetime modelled claims. This is not mainstream economics, neither is it science. Rather, it fails the demarcation test, the appeal to superior evidence, between science and non-science of pseudoscience. The failure of this belief system or meme has been made clear. We can only understand the importance of PC-CIS core value sets if we make clear that such a value set must, to meet the standards of normal science, recognize criteria for credibility of the claim or construct, the potential for empirical evaluation and a commitment to replication of the measure across target patient populations in the PC-CIS focused disease state. This is the first gateway that must be passed if the PC-CIS is to be developed, accepted and applied given the intended audience for the various PC-CIS measures. Indeed, the audience must be defined; is it to track patients in a registry, is it to identify unmet medical and social need or is it to support value claims in formulary submission? Irrespective of the audience, these standards must apply. This holds, for example, for the FDA CDER Pilot Grant Program for Standard Core Clinical Outcome Assessments (COAs) and related endpoints program and the International Consortium for Health Outcomes Management (ICHOM) program for standard data sets. In neither case is there any apparent effort being devoted to ensuring that instruments and measures meet the standards of normal science and fundamental measurement.

If experience to date with disease specific patient-centric or PRO measures is any guide, there will be considerable wasted effort in creating ordinal composite scores. A recent example of wasted effort is the international consensus on outcome measures for child and youth anxiety, depression, obsessive compulsive disorder and post-traumatic stress disorder. This commits the expected *faux pas* of relying on classical statistical analysis to conform the various instruments but overlooks the constraints imposed by fundamental measurement and the importance of Rasch or modern measurement[10] . An example is the recently released ICHOM international consensus standard set of outcome measures for child and youth anxiety, depression, obsessive compulsive disorder and post-traumatic stress disorder [11]. As an example of the misplaced choice of instruments, is the self-reported (ages 8 – 18) Revised Children's Anxiety and Depression Scale (RCADS-P) which comprises 47 items (questions) and 12 subscales with each item response on a 4-point Likert scale (never, sometimes, often and always) with the responses scored from 0 to 3 [12]. The item integer values are aggregated for an overall score and various subscale scores. As noted below, the limitations of Likert-based instrumentation render the RCADS-P redundant; a wasted effort if the objective is to assign a status category and evaluate response to therapy as the scale and subscales are ordinal and lack construct validity. The redundancy of Likert-based instruments is discussed below.

The second gateway is that the PC-CIS instrument must meet the standards for fundamental measurement, notably the application of conjoint simultaneous measurement with the application of Rasch or modern measurement theory. This requirement follows from the classification of scales of measurement: nominal, ordinal. Interval and ratio. Each scale has one or more of the following properties: identity where each value has a unique meaning (nominal); magnitude where values on a scale have an ordered relationship with each other but the distance between each is unknown (nominal); invariance of comparison where scale units are equal in an ordered relationship with an arbitrary zero (interval scale); and a true zero (or a universal constant) where no value on the scale can take negative values (ratio); the ratio scale has the interval property where the scale supports claims for both absolute and relative differences. To these should be added the major contribution of conjoint simultaneous measurement to ensure that measurement in the social science for PROs matches the standards of the physical sciences [13] [14]. Applying RMT for PRO latent constricts or attributes creates, if feasible, an instrument or measure that combines the difficulty of an item with the likelihood of a respondent completing that item [1]. Under certain circumstances this invariant interval scale can be transformed to a bounded ratio scale, the ideal measure for PRO value claims in therapy response in terms of relative differences [15].

If an instrument is to be valid it must meet ratio or interval measurement properties, this allows the application of classical statistical techniques and meaningful claims for response to therapy. This means we have to recognize the systemic error in instrument development that has characterized disease specific patient-centric or PRO measures over the past 30 years: the misapplication of Likert scale integer responses to support instrument scores and claims for therapy response. As will be detailed below, these measures are just composite ordinal scores and lack totally scientific merit.

PREMISES FOR PC-CIS VALUE CLAIMS

These premises it must be emphasized apply across the board to all value claims in disease and therapy areas: clinical claims, PRO claims, drug utilization claims and resource utilization claims. These are all potential value claims, PRO or PC-CIS value claims are of specific interest because of the measurement constraints on their development which, unfortunately, are usually ignored or unappreciated.

An understanding of the standards of fundamental measurement sets the stage for the two premises that support value claims for pharmaceutical products and devices. These are critical not only, in the case for PRO response claims, but also as a necessary basis for clinical and resource utilization value claims:

- All value claims for a product or therapeutic intervention must refer to a single attribute that meets the demarcation standards for normal science: all value claims must be credible, evaluable and replicable
- All value claims must be consistent with the limitations imposed by the axioms of fundamental measurement: they must be unidimensional and meet interval or ratio measurement standards

These premises apply to value claims that are disease or target patient population specific, where every claim is supported by a reporting and assessment protocol. Unfortunately, few PRO value claims meet these standards. Note, however, that the difficulties associated with PRO claims are not shared with other value claims which can support claims for clinical benefit. The formulary committee or health system is in the box seat to determine the relevance of claims for a target patient population and the process for factoring these into pricing and access recommendations. The key point is that claims assessment is an ongoing process where each claim is judged by its credibility, ability to be empirically evaluated and replicated across different treating environments. If not, then that value claim should be rejected. All value claims must be supported by a protocol detailing how that claim is to be evaluated and reported in a meaningful time frame.

THE LIKERT FALLACY

As noted in the case of the RCADS-P, a common feature in disease specific outcome measures is the development of instruments that are built around items for patient (or caregiver) response with responses presented as Likert scales. While these are popular, comprising the overwhelming majority of disease specific PROs, they are fundamentally flawed. Each Likert scale presents responses as a ranked ordinal scale where the distance between integer values assigned to the response level are unknown; we might as well apply letters instead of numbers. Assigning numbers to the various thresholds in the scale and then aggregating these across the various Likert response items is disallowed because each Likert response is ordinal. The aggregate score (and subscale scores) is just a composite, based on the number of Likert items, ordinal scale. It can tell us nothing about response to therapy or support any arithmetic operation. At best we can apply non-parametric statistics and focus on medians and modes as measures of the dispersion of the ranked ordinal scores.

The RCADS-P is a classic example of these fundamental errors. If we are the place reliance on the aggregate Likert score as a measure of response to therapy, then four conditions have to be me: (i) that the Likert items and the proposed scale refer to a coherent and meaningful single attribute or latent construct; (ii) that all of the Likert items (or statements) are, from the prospective respondents perspective, of equal difficulty; (iii) that the thresholds between integer steps for each Likert item are of equal value or equal distance and (iv) that each Likert item has the same number of integer responses or thresholds [1]. If these assumptions cannot be demonstrated then the '*add em'up*' procedure for the integer values yields only a multiattribute ordinal scale. Failing to meet these conditions ensures that Likert-based multiattribute PRO instruments with a single overall integer-based response score are clearly meaningless, and possibly misleading, as the basis for therapy response claims. The RCADS-P clearly fails on each of the first three criteria: the four-level Likert responses are ordinal as the distance between them is unknown, there is no attempt to assess item difficulty for the patient, where patients are likely to differ in the ability to respond and there is no appreciation of the need to conceptualize measure single latent constructs or attributes. The result is a dimensionally homogeneous ordinal score that lacks construct validity.

Unfortunately, unless there is a concerted effort on the part of those promoting PC-CIS and the search for patient-centric measures of the limitations imposed by fundamental evidence the same mistakes will be repeated; as they will be for the FDA and ICHOM recommendations for a core value set. Indeed, if part of this effort is to determine their relevance for capturing response to therapy, then each will have to be evaluated. Fortunately, a check list for evaluating PRO measures is available [5]. Unfortunately, in the less robust social sciences we assume, without proof, that the scales we use are interval measures; whereas they are uniformly ordinal. Unless PC-CIS recognizes this, core and supplemental instrument claims will be redundant. What is important in the Rasch model is to combine two elements to capture the probability of a positive (binary) response to a question: the difficulty of the item and the ability of the patient to realize that item (to respond positively). In a questionnaire, for example the that has identified needs for a target group, the needs will be ranked by the Rasch model in terms of their difficulty with a sample of respondents ranked in terms of their ability to respond. Interval level measures are derived when the levels of one attribute increase with the increase in the values of the other attributes. The Rasch model infers latent trait interval measures from raw counts of ranked item responses.

BEST PRACTICES

The claim is made that there is a lack of standardized approaches to construct PC-CIS; the plea is for best practices for *guiding and maintaining PC-CIS with a primary focus on patient centricity, and flexibility for innovation and evolution of the set(s).* This is patently untrue; we have the standards and techniques that are required, once we accept the critical role played by Rasch or modern measurement theory. Rasch measurement, the application of conjoint simultaneous measurement, to instrument development has been recognized since the 1960s. The primary focus of Rasch measurement, as described by Bond and Fox, is to develop units of measurement which art first may be arbitrary (ordinal) but can be iterated along a scale of interest (interval) so the unit values remain the same (relative difference), with the ultimate (and difficult) objective of the Holy Grail of a ratio measure [9].

It is instructive to contrast the Rasch framework for constructing fundamental measures to that typically found in instrument development where data have primacy with the results describing those data; the resulting instrument or model is exploratory of all data elements that are observed and describe the data [9]. This stands in contrast to the confirmatory Rasch model where the data elements are selected to fit a predictive model. To achieve the standards of the Rasch model to achieve fundamental evidence rules have to be applied to select the required data elements to ensure they fit the model. This applies the principles of probabilistic conjoint simultaneous measurement to focus on the size and structure of residuals for fitting. If this is achieved, and there is no guarantee, then it can be claimed that the results can be applied as a measurement scale for the attribute or latent construct that has invariant interval properties. This does not deny the application of statistical analysis, just that Rasch measurement precedes it in establishing a claim for interval measurement. The fundamental question, therefore, that has to be applied to have confidence in best practice for instrument development is whether or not it yields interval level properties for evaluating, among other applications, response to therapy. As far back as 1946, Stevens made the observation that *…most of the scales used widely and effectively by psychologists are ordinal scales* [9 16].

PC-CIS NEEDS-FULFILLMENT

This is only the starting point. If the focus is on therapy interventions for target patient groups, then we need to report on the value claims and their impact on outcomes defined in both clinical and patient-centric terms over the lifetime of the product and the patient; a basis for ongoing disease area and therapeutic class reviews. In this framework the patient (and caregiver) should have potentially equal billing in therapy assessment. This can be captured with the long-standing concept of disease or target patient population specific needs-fulfillment with Rasch measurement modelling; the subjectively assessd needs of the patient and/or caregiver and to the extent to which we might infer that their needs being met. Purely clinical parameters (e.g., functional status) may only go part way, if at all, to meeting needs. This can be considered in quality of life terms, where life takes its value from needs being fulfilled. The more needs that are fulfilled, the greater the quality of life. The important point is that over the past 25 years needs fulfillment measures across some 30 disease states have been developed that meet both the standards of normal science and Rasch or modern measurement theory. This sets the minimum standards bar for the potential identification and development for other PC-CIS patient centric measures as inferences of latent traits and as complements to purely clinical measures.

The challenge for PC-CIS is to recommend for specific target populations patient-centric measures that meet the required evidence standards. This is not easily accomplished as it requires not only subjective assessments based on patient (and caregiver) interviews, possibly to meet the somewhat ambitious (and probably unrealizable) objective of *gathering information from patients, carers and families about what is important, and deriving from them what is most important to them,* but to ensure that the latent constructs or attributes are credible and measurable. This is why the needs-fulfillment Rasch framework is not only critical but essential. At the same time, more recently a transformation algorithm has been developed to translate the Rasch interval measure to a bounded ratio scale; that is a scale defined in the range 0 (a true zero) and capped at unity [17] . This allows are direct measure of the extent to which needs are fulfilled in the target population and also the possibility of assessing the factors most closely associated with the level of needs fulfillment. If the RACDS-P is to be rejected, then the answer is to develop a needs-fulfillment quality of life instrument, following the Rasch model, for each of the sub-domains identified as potentially measurable latent-constructs.

CONCLUSIONS

If the proposed PC-CIS core and supplemental disease specific tool is to meet its objectives in patient-centric outcome assessment then attention has to be given to the standards of normal science and fundamental measurement. Notably, in the latter case, the application of conjoint simultaneous measurement and Rasch or modern measurement framework in instrument development. The objective must be to accept only those measures that yield interval or ratio measurement standards. Ordinal data are not measurement. Fortunately, with the wide application of Rasch measurement over the past 60 years, we have the techniques for ensuring that we achieve measures of latent traits, commonplace in patient-centric investigations, that allow meaningful measures of response to therapy. This is seen in the development over the past 25 years of Rasch needs-fulfillment instruments across some 30 disease states [18].

At the same time a review is required to assess core instruments that have already been proposed to support health technology assessment in specific disease states. Many will, no doubt, fail to meet the required measurement standards. These need to be weeded out. rather than accepting a core list of required instruments to capture, accurately or otherwise, the impact or outcomes of interventions in disease states, we should start with the required instrumentation standards; notably in respect of patient-centric or patient reported impacts, to agree on a minimum set that are meaningful. Only then is it appropriate to case a wider net to capture ancillary or complementary measures, but ones that must meet the required standards. In this framework (and agenda), PC-CIS is premature. Until a sound evidentiary base is agreed it seem pointless to speculate and develop additional measures.

REFERENCES

[1] Langley P. Nothing to Cheer About: Endorsing Imaginary Economic Evaluations and Value Claims with CHEERS 22 [version 1; peer review: 2 approved]. *F1000Research* 2022, **11**:248 (https://doi.org/10.12688/f1000research.109389.1)

[2] Perfetto E, Oehtlein F, Love T et al. Patient-centered Core Impact Sets: What they are and why we need them. *The Patient – Patient-Centered Outcomes Research*. 2022.

[3] Langley P. The Great I-QALY Disaster. *InovPharm*. 2020; 11(3): No 7

[4] Feng Y-S, Kohlmann T, Janssen M et al. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. Qual Life Res. 2021;30:647-73

[5] Langley PC and McKenna SP. Measurement, modeling and QALYs [version 1; peer review: 2 approved] *F1000Research* 2020, 9:1048 https://doi.org/10.12688/f1000research.25039.1

[6] McKenna S, Heaney A, Langley P. Fundamental Outcome Measurement: Selecting Patient Reported Outcome Instruments and Interpreting the Data they Produce. *InovPharm*. 2021; 12(2): No. 17

[7] McKenna S, Heaney A. Setting and maintaining standards for patient-reported outcome measures: can we rely on the COSMIN checklists? *J Med Econ*. 2021;24(1):502-511.

[8] Mokkink L, Terwee C, Bouter et al. Reply to the concerns raised by McKenna and Heaney about COSMIN..*J Med Econ*. 2021;24(1):857-859.

[9] McKenna SP, Heaney A COSMIN reviews: the need to consider measurement theory, modern measurement and a prospective rather than retrospective approach to evaluating patient-based measures. *J Med Econ*. 2021;24(1):860-861.

[10] Bond T, Yen Z, Heene M. Applying the Rasch Model: Fundamental measurement in the human sciences, 4th Ed. New York: Routledge, 2021

[11] Krause K, Chung S, Adewuya A et al. International consensus on a standard set of outcome measures for child and youth anxiety, depression, obsessive compulsive disorder, and post-traumatic stress disorder. *Lancet Psychiatry*. 2021;8(1):76-86

[12] Chorpita B, Moffitt C, Gray J. Psychometric properties of the Revised Child Anxiety and Depression Scale in a clinical sample. *Behavioral Res Ther*, 2005;432:309-22

[13] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske, Institut, 1960

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

[15] Langley P. McKenna S. Fundamental Measurement: The Need Fulfilment Quality of Life (N-QOL) Measure. *InovPharm*.2021;12(2): No. 6

[16] Stevens s. On the theory of the scales of measurement. Science. 1946;103:677-80

[17] Langley P. McKenna S. Fundamental Measurement: The Need Fulfilment Quality of Life (N-QOL) Measure. *InovPharm*.2021;12(2):No. 6

[18] Galen Research, Manchester UK. www.galen-research.com