MAIMON WORKING PAPERS www.maimonresearh.net

MEASUREMENT MALPRACTICE: COMPOSITE PATIENT REPORTED OUTCOMES

WORKING PAPER No. 19 August 2020

Paul C Langley, Ph.D. Adjunct Professor, College of Pharmacy, University of Minnesota

ABSTRACT

A common characteristic of patient reported outcomes (PRO) instruments is their composite nature. That is, they combine variables capturing different dimensions of health experience. This applies to both generic health related quality of life (HRQoL) as well as to the majority of disease specific PORO instruments. While composite PROs may be defended by their ability to capture and summarize in a single score a range of health outcomes they are, by the axioms of fundamental measurement, meaningless. They are ordinal constructs that lack dimensional homogeneity and hence construct validity. The implications of this, as a recent paper by McKenna and Heaney points out, are profound. As measures such as the EQ-5D-3L are ordinal composite scores; they cannot be used to create QALYs. Although widely used over the past 30 years, the QALY is an impossible mathematical construct. Value claims based on QALY models are meaningless; a situation that the Institute for Clinical and Economic Review (ICER) and other organizations in health care decision making have yet to come to terms with in their pricing and access recommendations.

INTRODUCTION

The hallmark of normal science is the testing of hypotheses to discover new facts. This can only occur if the instruments employed to evaluate claims are consistent with the axioms of fundamental measurement and dimensional homogeneity in applying a coherent theoretical model to combine the various item components. This has been recognized since the 17th century with considerable effort dedicated to ensuring the ability of the instrument to measure the desired attribute (e.g., temperature). Accurate measurement of an attribute is critical both to support decisions but also to ensure that claims are credible, evaluable and replicable. The key feature in measurement is unidimensionality. The instruments must be designed to measure a single attribute. To achieve this, the instrument can be designed so that the scores created have either interval or ratio properties. This is a critical distinction. In the former case the instrument is designed to provide invariant comparisons; differences between scores are the same over the entire scale. A response from one point on the scale is captured with the same metric as the response from any other point (e.g., 12 inch ruler). The scale supports addition and subtraction. In the latter case the scale has a true zero (e.g., weight); there can be no observations below zero. This scale, which includes invariance of comparisons, supports the further operations of multiplication and division.

Unfortunately, when we consider the development and application of patient reported outcomes measures (PROs) in medicine and its offshoot health technology assessment, notions of meeting

measurement standards for unidimensionality, invariance of comparisons and a true zero is, to all intents and purposes, ignored (or simply not recognized). The implications of this have been detailed in a recent paper by McKenna and Heaney that challenges the willingness of practitioners to construct and apply composite measures [1]. Put simply: a failure to recognize the irrelevance of composite measures in health system decision making. A willingness, often through ignorance rather than design, to develop PROs which fail to meet the required axioms of fundamental measurement; instruments which are nothing more than ordinal or ordered measures. That is, the scores produced by the instrument can be ranked but we have no idea of the distance between the tabulated scores. We might put the scores on an interval scale to give the impression that we 'know' the difference but this is an illusion. If an instrument is not designed to have an interval scale then it won't have one (e.g., an elastic tape measure).

COMPOSITE MALPRACTICE

McKenna and Heaney point to three areas where composite instruments are employed as performance or outcome measures: (i) health system performance; (ii) clinical trials; and (iii) patient's assessment of their health status or treatment impact. In each case the composite measures employed (and the same composite instrument may be used in more than one area) is intended to create a single score for a multidimensional concept, where the concept is defined in terms of two or more variable or attributes (e.g., depression and pain) that are considered to related with a concept such as health related quality of life (HRQoL) or just statistically. These variables, which may have latent or ordinal characteristics, are added together as simple additions, the typical case, or as weighted observations within a composite total score algorithm.

There are some obvious limitations on composite measures. What is driving a change in the aggregate score? Do we need to disaggregate to obtain any meaning? How transparent is the score? How are score changes to be interpreted by clinicians? How were the items comprising the score selected and reported on by patients? Are the items weighted? On what basis were weights employed? Will different weights produce different outcomes? Are the items comprising the composite measure interrelated? How useful are composite measures in decision making? Is the composite measure effectively redundant? While these issues are relevant to all composite measures, the problems do not arise when we accept the role of single items or attributes and construct a unidimensional measure.

But there is a more fundamental problem: limitations imposed by the axioms of fundamental measurement. Unless designed to meet required response properties, composite measures will create ordinal rankings. They will not capture response to therapy; their role in clinical decision making is extremely limited and may not even be meaningful. All we can say is that score A > score B. Not by how much; just a ranking with an unknown difference.

APPLICATIONS OF MALPRACTICE

There is any number of inappropriate applications of composite scores. McKenna and Heaney provide an example of a composite measure to assess the quality of hospital surgery. Variable measures included mortality, hospital volume and patient covariates (e.g., comorbidities, socioeconomic status). Adding these variables breaks the rule of dimensional homogeneity where variables can only be combined if they have the same dimension. Otherwise they lack construct validity. *In consequence, composite measures should not be employed in the case of PRO measures.*

Perhaps the most egregious example of measurement malpractice is to be found with the QALY measure where time spent is a disease state is multiplied by a 'utility' score on a presumed range of 0=death and 1=perfect health to yield a measure of years of perfect health. Unfortunately, as has been pointed out by McKenna and Heaney and others, the QALY is a mathematically impossible construct. As the utility scale is ordinal, you cannot multiply time spent by an ordinal score. The EQ-5D-3L has neither latent nor ratio properties. Indeed, the EQ-5D-3L, in common with other generic HRQoL measures can take negative values (its range is from 1 to -0.59). This raises the intriguing prospect of impossible negative QALYs.

The fact that the QALY is an impossible construct has profound consequences. It has been center stage for 30 years as the primary cost-effectiveness measure creating incremental cost-per-QALY claims. It is central to the business case of the Institute for Clinical and Economic Review (ICER). Without the QALY the ICER business model collapses. Recommendations for pricing and product access are meaningless, as are its recent QALY-based models and proposals for COVID-19 pricing. It is unlikely ICER will acknowledge this situation; it is up to health care decision makers to reject is recommendations; for ICER to recant five years of modeled evidence reports would be a bombshell [2].

ABANDONING COMPOSITE MEASURES

Given the importance of accurate and meaningful measures in health care, there is no option but to emphasize that the failure to achieve standards for instrument development that characterize the physical sciences is unacceptable. This applies both to generic HRQoL instruments such as the EQ-5D-3L and HUI Mk 3 as well as to the majority of disease specific outcomes measures. The task of informing health care decision makers and physicians in clinical practice will be a challenging task. A recent search of the PubMed data base with the keywords "Cost and QALY" yielded some 16,500 hits. Even so, a possible starting point would be through the leading health technology assessment and medical journals, and professional associations such as the National Pharmaceutical Council (NPC) and the Pharmaceutical Research and Manufacturers of America (PhRMA). A recent paper has already criticized the NPC proposals for QALY-based lifetime value assessment [3]. To this might be added the introduction of 'measurement malpractice' flags for instrument data bases such as MAPI (Adelphi) values and the Tufts Center for Evaluation of Value and Risk in Health (CEVR).

If we are to, at least, minimize inappropriate decisions for pharmaceutical products and devices, particularly for high risk and rare diseases, then there is no option but to acknowledge past malpractice and embrace the axioms of fundamental measurement. As McKenna and Heaney conclude: *Valid outcome measurement must be unidimensional but composite indices by definition do not have this*

quality ….inappropriate composite measurement is 'counterproductive, undermines the professionalism of dedicated clinicians and erodes patient trust'.

REFERENCES

[1] McKenna S, Heaney A. Composite outcome measurement in clinical research; the triumph of illusion over reality? *J Med Econ*. 2020doi:10.1080/13696998.2020.1797755

[2] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-2023 value assessment framework for constructing imaginary worlds. *Inov Pharm*. 2020;11(1):No. 12

[3] Langley P. The National Pharmaceutical Council: Endorsing the construction of imaginary worlds in health technology assessment. *Pharmacy*. 2020;8(119);doi:10.3390/pharmacy8030119