Original Article

Questionnaire validation may not validate- a critical analysis

SILVIA PINANGO-LUNA[1], PETER PETROS[2,3]

[1] *Hospital "Dr. Miguel Pérez Carreño" Caracas, Venezuela*
[2] *University of NSW, Professorial Dept of Surgery, St Vincent's Hospital, Sydney*
[3] *University of Western Australia, Crawley, Perth WA*

Abstract: The genesis of this work came from analysis of a single patient QOL graph mainly for pain. It gave rise to 6 questions concerning the validity of the validation process using the ICIQ questionnaire as an example. The questions raised against 'validation' were: 1. The assessment was almost entirely subjective. 2. The test-re-test time frame of 2 weeks could lead to major errors. 3. The questions tested the collective memory, not variation. 4. Replacement of the physician's interaction and the considerable benefits thereof. 5. The questionnaires are reductionist, seemingly oblivious of the holistic anatomical control mechanism. 6. Validations add another layer of complexity and do not add to what can be obtained using the simple language of a questionnaire. In conclusion there seems no benefit in 'validating' what are really simple questions based on plain English. As long as the authors define what they are talking about in the methods that should suffice.

Keywords: Questionnaire validation; Art of Medicine.

BACKGROUND

From Hippocrates onwards, symptoms have been an essential element in medical diagnosis. Normally the physician elicits symptoms by listening to the patient's story or by direct questioning. Because this was said to introduce bias, patient administered questionnaires are being increasingly used to remove the bias from the physician.

The original aim of this work was to test the reproducibility of the pain symptom within an individual patient over a 3 month time period. Analysis of the graphs and the 'validated' questions themselves brought the whole concept of questionnaire validation into question.

INTRODUCTION

In urinary incontinence, wide variation in symptoms within an individual patient caused symptoms to be viewed as unreliable, thereby promoting the use of 'objective' urodynamics[1].

A later development questioned the validity of the questions themselves, leading to psychometric systems for validation of questionnaires. Psychometric validation of questionnaires involved a complex mainly subjective system involving several parameters[2], including: Face validity * Content validity* Construct validity*, Criterion validity* Test-retest for reproducibility**, statistical inner consistency (Cronbach's alpha coefficient), responsiveness*.

* physician subjective

** patient subjective

A critical analysis of the process involved in validation gave rise to 6 questions. These are detailed below.

The first question concerning 'validation'

A simple analysis of these parameters inevitably concludes that nearly all are subjective. This subjectivity raises the first question: how is a subjective 'validated questionnaire' more valid than a 'subjective' history from an experienced perceptive clinician?

A simple methodology

A 68 year old woman, parity 3, mainly with chronic pelvic pain, some nocturia and some non-stress non-urge urine loss agreed to keep a daily diary over a 3 month period to monitor her chronic pelvic pain using a 1-10 visual analogue scale (VAS), entry made immediately before retiring for the night.

Three pain charts

Three charts with no gaps are presented (Figure 1). Pain intensity was recorded on 81 consecutive days. In 12/81 days, the pain was severe, VAS 7 or above (Figure 2). The month of May had the most severe episodes, 8 with VAS>7. The outstanding feature of these graphs is the massive variations even in the space of a few days. The pain varied from VAS 8 to VAS 1, between day 16 to day 26 (May), from VAS 7 to VAS 1, between day 7 to day 10 (June), from VAS 8 and VAS 2, between day 10 to day 19 (July).

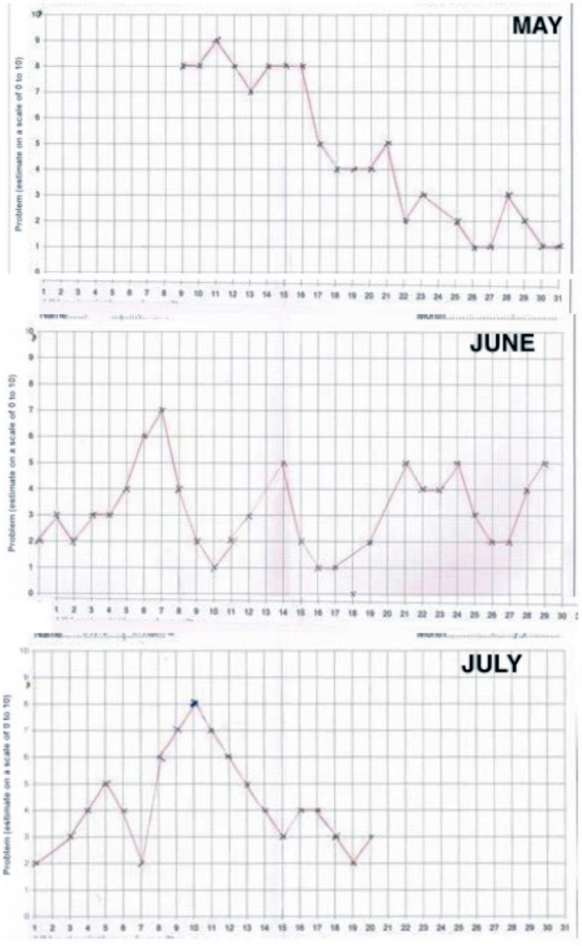

Figure 1. – Graphic display of VAS QOL scores, chronic pelvic pain.

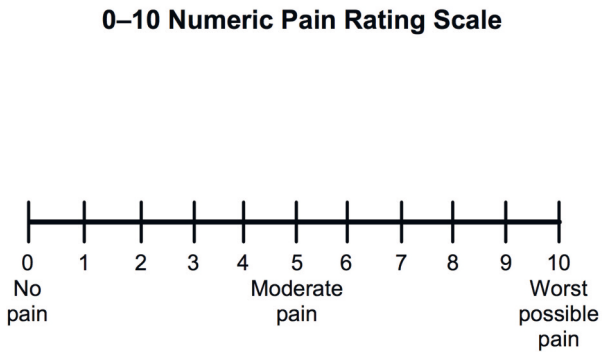

Figure 2. – Visual analogue scale 0-10 Numeric Scale.

The second question concerning 'validation'

The time for the pain to fall from severe to mild varied widely (Figure 1): 10 days in May, 3 days in June and 7 days in July. Fig1 raises the 2nd question, how valid is any 'validated questionnaire'* when it relies on a test-retest analysis which can vary so widely within 2-3 days?

* The 'scientific' basis for the recommended 7-14 day test-re-test interval is yet another subjective but 'authoritative' proclamation[3] which does not fit in any way with the graphs, (Figure 1). *From Coyne K, Kelleher C,Patient Reported Outcomes:* The ICIQ and the State of the Art[3]. *"Test-retest reliability, or reproducibility, indicates how well results can be reproduced with repeated testing. To assess test-retest reliability, the same patient completes the questionnaire more than once, at baseline and again after a period of time during which the impact of symptoms is unlikely to change (e.g., a few days or weeks). It is important to keep the test-retest period a reasonably short period of time - such as 7-14 days"*[3].

The 3rd question concerning 'validation'

'Reproducibility' may not be the parameter which the "test-retest" methodology of questionnaire 'validation' is actually testing. For example, the questions from the ICIQ questionnaire are in the present tense. They test the patient's *collective memory* of the pain. A typical question is *"To what extent does your urinary problem affect your household tasks (e.g. cleaning, shopping, etc.)?"* Another is "*Does your urinary problem affect your job, or your normal daily activities outside the home?*" and so on. These questions request a global average answer based on collective memory of individual events, recorded and averaged over a time period by the cortex. The mental process behind this question is similar to "What sort of food does your mother cook, bad, average, good?" If it is mainly 'good food', the occasional bad or average meal is discounted by the cortex. This process certainly would not vary in a two week interval. So what is really being tested by a test-retest questionnaire for this ICIQ *is the reliability of the patient's memory of the symptom, not the reproducibility of the symptom itself.*

All of which begs the 3rd question, why bother to validate ICIQ when it does not assess reproducibility?"

The 4th question concerning 'validation'

There is less provision for the doctor-patient interaction in such constructs which are essentially reductionist. What generally happens is that the physician takes the patient administered questionnaire, explains the result and prescribes some treatment or other. Inevitably, there are inroads into the "Art of Medicine" [4]. Like the function of the human body, the "Art of Medicine" is non-linear. Highly experienced clinicians read a patient's body language, sense barriers such as shyness, personality disorders, inhibitions, ask further questions and use their art to penetrate further into the cause of the problems.

The 5th question

The questions in questionnaires such as the ICIQ are reductionist. Nowhere is there any space for variation. Yet, as the VAS graph demonstrates, symptoms vary widely. How can this be? Given the pelvic structures and exponential nature of the control mechanisms for muscles, fascia, organs, ligaments, even a minor variation of anatomy, if additive, could cause sufficient laxity in the uterosacral ligaments (Figure 3), to set off a cascade of widely varied events ([4-6]). An extreme analogy is the fluttering of a butterfly causing a cyclone on the other side of the globe[6]. This gives rise to question no 5, "How can the originators of such 'validations', be so authoritatively reductionist, when the control mechanisms are holistic, exponentially determined and holistically controlled?"

The 6th question

It examines whether new constructs such as 'questionnaire validation" really add anything beyond the same 'unvalidated' questions, when both use simple language.

Karl Popper described such constructs as 'new languages'. He considered them artificial and unnecessary if they could be described in more simple terms[7].

Popper's viewpoint is that a Theory (or concept) can never be entirely validated. It can, however, be easily invalidated: one validated exception invalidates the whole concept[7]. For example if one states that all swans are white, the production of one black swan invalidates that concept. The concept of good test-retest correlations within two week intervals lie at the core of questionnaire 'validation'[2,3]. In the same way as the black swan, the VAS graph invalidates the whole test retest concept and with it, the whole process of validation.

Popper described constructs such as 'questionnaire validation' as "artificial model languages". He stated that contradictions arise when an'artificial model language' is created.

In 1980 Popper stated *"Thus the method of constructing artificial model languages is incapable of tackling the problems of the growth of our knowledge; and it is even less able*

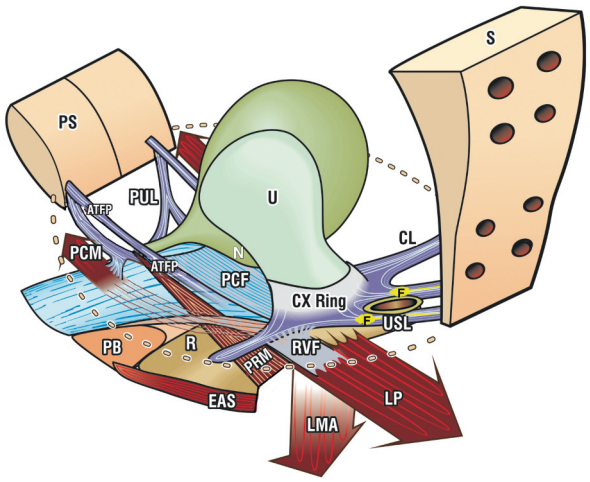

Figure 3. – Opposite directional forces (arrows) stretch CL & USL ligaments and fascia to support Frankenhauser (F) and sacral nerve plexuses . The posterior directional forces (LP/LMA) act against the cardinal (CL) and uterosacral ligaments (USL). PS = pubic symphysis; S = sacrum; PUL = pubourethral ligament; ATFP = arcus tendineus fascia pelvis; USL = uterosacral ligament; CL= cardinal ligament; PCM = pubococcygeus muscle; LP = levator plate; LMA = longitudinal muscle of the anus; PRM = puborectalis muscle; PCF = pubocervical fascia;RVF = rectovaginal fascia; PB = perineal body; EAS = external anal sphincter
Note Nerve plexuses 'F' (yellow) at the base of USLs.

to do so than the method of analysing ordinary languages, simply because these model languages are poorer than ordinary languages. It is a result of their poverty that they yield only the most crude and the most misleading model of the growth of knowledge - the model of an accumulating heap of observation statements".

Is the validated ICIQ questionnaire poorer than using simple language? Questionnaires such as ICIQ combine symptoms and give results as 'scores'. For example, the statement to an ordinary person " Your ICIQ score improved from 'x' before surgery to 'y' after surgery" is totally meaningless. A patient complains of a symptom, not a number. Furthermore, there may be differential improvement in say their urgency but not nocturia. Surely it is more informative to list each symptom and say whether it is improved (or not). This example confirms Popper's statement "*these model languages are poorer than ordinary languages. It is a result of their poverty that they yield only the most crude and the most misleading model of the growth of knowledge - the model of an accumulating heap of observation statements"*.

CONCLUSIONS

There seems no benefit in 'validating' what are really simple questions based on plain English. As long as the authors define what they are talking about in the methods that should suffice.

Declarations

Permission was obtained from the patient to anonymously publish the graphs and clinical details. No conflicts financial or otherwise.

Participation

Petros: conceptualization, analysis, diagrams writing. Pinango-Luna: analysis, writing.

REFERENCES

1. Bates P, Bradley WE, Glen E, Hansjorg M, Rowan D, Sterling A, Hald T. International Continence Society First Report on the Standardisation of Terminology of Lower Urinary Tract Function (1975).
2. Barber MD. Questionnaires for women with pelvic floor disorders. Int Urogynecol J. 2007; 18: 461+465. DOI 10.1007/s00192-006-0252-1.
3. Coyne K, Con Kelleher C Patient Reported Outcomes: The ICIQ and the State of the Art , Neurourology and Urodynamics 2010, 29, 645-651.
4. Petros PE The Art and Science of Medicine. Lancet, 2001, 358, 818-1819.
5. Petros PEP. Ch 6.1.2 The Chaos Theory Framework – its impact on the understanding of bladder control and urodynamic charting in "The Female Pelvic Floor, 3rd Ed Springer Heidelberg, pp 239-246.
6. Gleick J. "Inner Rhythms" in Chaos – Making a New Science Cardinal, Penguin, England; 1987, 275-300.
7. Popper KR. Theories. Falsifiability. The Logic of Scientific Discovery. Unwin, Hyman, London, 1980, 27-146.

Correspondence to:
Peter Petros
E-mail: pp@kvinno.com

Editorial commentary

The study of medicine, especially characterization of disease symptoms, evaluation of treatment outcomes, quality of life, and even patient satisfaction, constantly lies in the tension that exists between deduction and induction. We try to learn from the single case in order to infer the rule for many as we also try to discard information gathered from many for the single patient. While in medicine we can refer to the past decade or two as the era of evidenced-based medicine, we can witness today a transition to an era of personalized or tailored medicine. Personalized medicine is not a new concept in medicine, but rather a return to the past. We are all familiar with the typecast of the old rural doctor with the holistic approach that treats a person as a whole of body and soul, characterized by special physician-patient relations, familiarity with the patient, their family and their environment and recognition of all the physical mental and emotional factors that can promote their health.

In this article by Pinango-Luna and Petros the authors refer to this gap between deduction and induction and list and discuss the disadvantages of deduction. I completely agree with the authors that make some strong arguments in favor of the personal and direct doctor-patient contact. We must never abandon the skills of history taking and questioning so that we can obtain the whole picture of our patients' condition. In addition, the authors also argue strongly against the use of questionnaires and their validation. I don't think that validated questionnaires need to be abandoned all together, they do however need to be considered with caution.

Although studies using validated questionnaires are preferred to attain objectively obtained reproducible data, we need to be cautious in interpreting these data.

One example is the use of condition-specific questionnaires. Condition-specific questionnaires are validated to discriminate between women with and without a certain condition (i.e. pelvic pain, sexual function, quality of life, etc.), within the group of patients suffering from a broader condition (for example pelvic pain among women treated for pelvic floor dysfunction). Indeed, in some studies, questionnaires have shown responsiveness to change after surgery. After surgery, new aspects such as dyspareunia, worries about damaging the operative results, onset of new symptoms, unsatisfactory surgical results, or development of complications become relevant because of the treatment. Hence, a state following pelvic floor surgery should be regarded as a new clinical condition, necessitating a new condition-specific validated questionnaire. It may be that these questionnaires even those that are condition specific for one condition are not optimal to detect their goal after surgery because these new aspects are not represented in the questionnaire. By neglecting the negative impact that pelvic floor surgery may have on its own accord, evaluation following surgery might be too positive[1].

I believe that although interpreted with caution, validated questionnaires should be used for the study of medicine, however in no instance should they replace the doctor-patient interaction.

REFERENCE

1. 1. Weintraub AY. A validated tool would greatly enhance future research on the impact of surgery on sexual function. J Womens Health (Larchmt). 2016, 25, 327-8.

ADI Y. WEINTRAUB
Editor Pelviperineology
Soroka University Medical Center, Beer Sheva - Department of Obstetrics and Gynecology
adiyehud@bgu.ac.il