The Journey (or Adventure) Towards Evidence-informed Outcome Assessment
Paul Stratford - Enid Graham Memorial Lecture
Canadian Physiotherapy Association
2016 Annual Congress – May 2016
Victoria, British Columbia

I would like to thank CPA’s Awards Committee for this honour that bears the name of Enid Graham an extraordinary individual who had a profound impact on the inception and maturation of our profession. Here are a few things I learned about this remarkable person: (1) she studied massage and medical gymnastics in Heidelberg; (2) she was a member of the Voluntary Aid Detachment during World War I; (3) was a founding member of CPA; (4) established training schools in Montreal and Toronto; and (5) lobbied that physiotherapists should be given officer rank in the military which they received in 1940. She was truly a remarkable person.

So often what makes it possible to have the opportunity and privilege to receive an honour such as this, is that one’s colleagues took the time from their busy schedules to prepare a nomination. I would like to thank Jill Binkley, Vanina Del Bello Haas, Deborah Kennedy, Susan Harris, Dan Riddle, and Linda Woodhouse for taking time from their busy schedules to prepare a nomination.

Not only have I had the privilege to work with these folks, but I’ve also had the good fortune to work with and learn from many wonderful colleagues and students.

I would also like to thank my family for their support.

The title for the lecture is “The Journey (or Adventure) Towards Evidence-informed Outcome Assessment”.

One year ago this month, Dr. David Sackett—recognized as the father of Evidence-Based Medicine—died. His mission was to train the masses to transform high quality research findings into clinical actions that best served individual patients. It’s fitting that Evidence-based Practice is supported by a tripod that includes best research evidence, clinical expertise, and patient values. If one of the legs is missing Evidence-based Practice collapses. This triad also recognizes that evidence doesn’t make decisions, people do. Stated another way, best evidence isn’t intended to replace clinical expertise, but rather to augment it. In the fall of 1978 I had the good fortune of being a student in Dr. Sackett’s clinical trials course. He was gifted at taking the most complex concept and rotating it until it was understood by all. He was equally accomplished and considerate at providing feedback which often came in the form of what
would become known as the Sackett sandwich. Here’s an example: “What a novel idea, you might want to take a second look at your data, your slides look great.”

For me, the hallmark of his teaching and writing was the clinical vignette.

Here’s an illustration: “Suppose, for example, you detect 10 to 15 degrees of scoliosis in an otherwise healthy 12-year-old student who has come for her preschool examination. Do you tell her and her parents (and, if so, what do you say?), refer her to an ‘orthopod’ or what?”

These seemingly simple scenarios served as the starting point for the Sackett process of inquiry and learning. The process involved the following six steps: (1) structure a researchable question, (2) identify the most likely resources, (3) design an effective search strategy, (4) summarize and critique the evidence, (5) apply the evidence to the patient of interest, (6) evaluate the process and outcome.

Recognizing that his days were numbered, and I suspect growing weary of answering the same questions about his life, he put pen to paper, and in an interview format, delivered a 103 page accounting of his professional life and values. The volume ended with these thoughts: (1) be loyal to people rather than institutions, (2) serve the young, (3) you become what you pretend to be.

In the good old days (circa mid 70’s) assessing a patient’s outcome was much simpler than today. When a patient had 123° knee flexion on Friday and 120° on Monday he had lost 3°, no doubt about. And if you wanted to know about the activity level of patients you simply asked them. Sometimes responses were clear and could be charted with confidence in the medical record. However, all too often responses are ambiguous. The purpose of standardized outcome measures is not to replace conversations such as these, but rather to add clarity and confidence to their interpretation. Today I believe that assessing a patient’s outcome is a complex clinical skill that requires a conceptual framework and specific body of knowledge. Not complex in the sense that it’s difficult to learn or understand, but rather that the skill has multiple parts. Part of the framework consists of 5 important clinical questions, and the body of knowledge is the context specific answers to these questions. However, before I introduce these questions I want to share with you a bit of the history of how questions came into being.

In 1977 CPA became one of the first health care professions to provide direction for Quality Assurance. The initial report titled “Towards the Assessment of Quality of Care in Physiotherapy” stated that the model for measuring quality of care should be based on outcome measures. The second report gave an account of the development and testing of outcome measures relevant to physiotherapy care. Subsequently, two intrepid groups—the first led by Beverly Cole and the second by Elspeth Finch—produced the first and second editions of Physical Rehabilitation Outcome Measures—popularly known as the RED BOOK. More recently, CPA has provided educational webinars, resource links on its website, and an exciting initiative with Focus on Therapeutic Outcomes, which applies state of the art electronic
outcome assessment using computer-adapted testing.

CPA’s Quality of Care initiative was chaired by Helen Saarinen Rahikka. Helen was also CPA’s first physiotherapy consultant and the 1993 Enid Graham Recipient. Each year the Helen Saarinen Rahikka Student Leadership Award is given out at Congress. The introduction to this award reads as follows: “Helen Saarinen Rahikka devoted her professional life to the educational development of caring and capable clinicians in Canada and around the world. First with the Mohawk-McMaster Programme, and then within the School of Rehabilitation Science at McMaster University, her leadership, knowledge and compassion fostered two generations of students. Her ability to act as a mentor, recognize potential and facilitate excellence in others was unsurpassed within the physiotherapy profession.” When I think of Helen, she was never too busy to help colleagues and students meet their goals.

Not only has CPA as a professional body been active in encouraging and facilitating the use of standardized measures, but Canadian physiotherapists have also been at the forefront of measure development. Table 1 provides a few of the many measures developed by Canadian physiotherapists.

**Table 1. Measures Developed by Canadian Physiotherapists**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Developer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alberta Infant Motor Scale</td>
<td>Piper</td>
</tr>
<tr>
<td>Berg Balance Scale</td>
<td>Berg</td>
</tr>
<tr>
<td>Canadian Physiotherapy Assessment of Clinical</td>
<td>Mori</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
</tr>
<tr>
<td>Chedoke Arm and Hand Activity Inventory</td>
<td>Barreca</td>
</tr>
<tr>
<td>Chedoke-McMaster Stroke Assessment</td>
<td>Gowland</td>
</tr>
<tr>
<td>Continuing Care Activity Measure</td>
<td>Huijbregts</td>
</tr>
<tr>
<td>Clinical Outcome Variables Scale</td>
<td>Seaby</td>
</tr>
<tr>
<td>Harris Infant Neuromotor Test</td>
<td>Harris</td>
</tr>
<tr>
<td>Lower Extremity Functional Scale</td>
<td>Binkley</td>
</tr>
<tr>
<td>P4</td>
<td>Spadoni</td>
</tr>
<tr>
<td>Patient-Rated Elbow Evaluation</td>
<td>MacDermid</td>
</tr>
<tr>
<td>Patient Specific Functional Scale</td>
<td>Stratford</td>
</tr>
<tr>
<td>Reintegration to Normal Living Index</td>
<td>Wood-Dauphinee</td>
</tr>
<tr>
<td>Stroke Rehabilitation Assessment of Movement</td>
<td>Mayo</td>
</tr>
</tbody>
</table>
When it comes to the development and advocacy of standardized measures, CPA and its members have not followed where the path may lead, but rather, have gone where there was no path and blazed a trial.

Yet, despite these impressive efforts there has been resistance, both at home and internationally, to the systematic clinical application and added decision-making value of using standardized measures. For example, in a US survey Diane Jette found that only 48% of participants used standardized outcome measures. And here are a few quotes from participants in a study by McAuley et al: (1) “The use of outcome measures in this population is redundant. Why keep measuring when we know there will be improvement?” (2) “Not a focus in my patient caseload—a very fit group in general”, (3) “Most outcome measures are not useful in acute care”. To the first question I would ask: Do all patients improve, and if so to the same extent? How do you write measurable goals? The second and third responses share a common theme, that is an appropriate measure does not exist for these patient groups. I would ask these respondents if they have considered a class of measure known as “Patient Specific Measures”. Examples of these measures include the (1) Canadian Occupational Performance Measure, (2) Patient Generated Index, (3) MacTar, (4) Patient Specific Functional Scale (PSFS). As an example, the PSFS has been successful in assessing a spectrum of activities including: (1) marching with my tuba, (2) dredging for gold, (3) raising my arms to praise the lord, (4) delivering babies. Finally, a participant in a study done by McGlynn and Cott provided the following comment: “Well, sometimes I’d agree with them (research articles) and sometimes I disagree. It’s like reading a movie review; the reviewer may put down a movie, but you may love it.” Unlike opinion, standardized measures have formal descriptions of their conceptual framework, development, administration, scoring, and interpretation. And they can be defended to the extent that their measurement properties have been reported critiqued and substantiated in peer-reviewed forums.

Clearly, there are barriers to the successful clinical application of standardized outcome measures. Some are real and others perceived. Table 2 provides a few of the more frequently cited barriers.

Table 2. Frequently Cited Barriers

<table>
<thead>
<tr>
<th>Barrier</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>I’m really doing this for someone else, and that someone else is not my patient.</td>
<td>This is a real barrier and it likely stems from the top-down approach measures were imposed on clinicians by administrations and payers.</td>
</tr>
<tr>
<td>Barrier</td>
<td>Comment</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>I can infer activity limitations and participatory restrictions with a high degree of confidence from impairment measures that I currently apply.</td>
<td>This is a perceived barrier. At best there is a modest correlation between impairments and activities. Think Terry Fox.</td>
</tr>
<tr>
<td>I do not understand what a new measure’s score means.</td>
<td>This is a real barrier. Why would a busy clinician take the extra time and effort to administer another measure if it didn’t provide additional information and with greater confidence than is afforded by the current assessment methods?</td>
</tr>
<tr>
<td>I have insufficient time to administer and score the measure.</td>
<td>I would ask, How much time would you have if the new measure provided additional clinically useful information to enhance clinical decision-making?</td>
</tr>
</tbody>
</table>

To the aforementioned barriers I would add (1) the apparent failure to recognize measurement error, of which inherent variability in a truly stable patient is a major source, (2) authors have emphasized coefficient values rather their interpretation and application, (3) editors have been reluctant to allow the translation of research findings into obvious clinical applications. In contrast to the staunch views of some journals and editors, Physiotherapy Canada has taken knowledge translation very seriously. Each year, the Joan Cleather Silver Quill for Knowledge Translation is awarded. Joan was the editor of Physiotherapy Canada for 19 years and the 2008 Enid Graham recipient.

Paralleling these barriers was a climate, where for many, there was a chasm between clinicians and researchers. Perhaps the most vivid example of this for me was at the 1994 Joint CPA, APTA Congress in Toronto. A researcher had just finished his presentation concerning a randomized clinical trial involving a manual therapy technique, when a member of the audience rushed to the question microphone and began belittling the work. The audience member was quite aggressive in his criticism noting the dosage was incorrect—albeit—dosage studies did not exist.

What was truly fascinating about this diatribe was that the study had shown a therapeutic effect in favour of the intervention, however, this was little consequence because the study was done by a PT who wore the label “researcher”, a perceived outsider.

Now, I had a somewhat different experience, and although it was just as devastating, it wasn’t nearly as public. In the early days of the McMaster PT Program I sat on the admission’s committee and provided measurement support. My contribution was to figure out things like is it better to put three interviewers in the same room for 15-minutes or one interviewer in each
of 3-rooms for 5-minute interviews. Following several meetings in our department we had a meeting across campus with the higher ups. It was at this meeting that the assistant registrar introduced me as a statistician. I attempted to correct her stating I was a physiotherapist. However, she didn’t believe this until our chair said yes, Paul’s a PT. It was at this point that Laurel rocked back in her chair, sighed, rubbed her chin and uttered, “ah, a physiotherapist gone bad.” In retrospect she may have hit the nail on the head.

It was against this background that Jill Binkley conceived the North American Orthopaedic Rehabilitation Research Network. This group consisted of approximately 30 likeminded clinicians and researchers from 4 provinces and 7 states. Partnership with this group led not only to the development and testing of measures such as the LEFS, UEFI, and PSFS, but also to the conceptualization of a framework for assessing patient outcome and reporting results. In fact, with respect to the latter point, someone pointed out that group members spoke English, French, and Spanish, however, the results from measurement studies didn’t appear to be in any of languages. It was at this point that the journey became a quest for clarity and confidence.

**Five Questions to Consider When Assessing a Patient’s Outcome**

The framework I mentioned previously consists in part of five important clinical questions. The questions are (1) How confident can I be in a measured value? (2) To what extent can valid inferences be drawn from a measured value? (3) To what extent can valid inferences be drawn from a measure’s change score? (4) What is the target value for this patient? (5) What is the context specific ideal reassessment interval?

**How confident can I be in a measured value?**

This question considers the location of a measured value and not whether the measure is assessing what it is intended to measure. In clinical practice it’s customary to report a single value in the medical record. For example, we might record a patient’s Timed-up-and-go time as 12.5 seconds. However, the obtained value is just one of many values that could have been obtained on a truly stable patient. If the variation among potential responses is small, the recorded value is a reasonable representation of the patients true value, however, if the variation is large, we can’t be confident that the obtained value is a reasonable representation of the truth, and more importantly we can’t be confident in clinical decisions based on the measured value. The answer to the question “How confident can I be in a measured value?” comes from reliability studies. However, results are often presented in a form where their interpretation and direct application to clinical decision-making are not obvious. For example, it is typical for a researcher to report the results from a reliability study as follows: “The Type 2,1 intraclass correlation coefficient was 0.80 and the standard error of measurement was 1.3 seconds”. This requisite reporting could be augmented with the following: (1) 95% of patients with characteristics similar to the study patients will display random fluctuations of ± 2.5 seconds, (2) for a patient with a measured value of 12.5 seconds we would expect her true value to fall between 10.0 and 15.0 seconds, (3) if we were to base our assessment on the
average of 3-trials, rather than a single, trial we could be 95% certain that the patient’s true value would fall within \(\pm 1.5\) (e.g., 11.0 to 14.0 seconds).

**To what extent can valid inferences be drawn from a measured value?**

There are two aspects to this question. One addresses validity, the other considers interpretability. Validation studies have typically limited their reporting to various coefficients. Here’s an example for the Lower Extremity Functional Scale, a 20-item measure of lower extremity functional status: (1) the correlation between LEFS and WOMAC-PF scores is 0.78, (2) the difference in LEFS scores between patients requiring surgery and those who did not was statistically significant \((t_{98} = 4.52, p < 0.001)\), (3) the correlation between LEFS scores and WOMAC-PF were greater than between LEFS scores and pain. The problem with presenting coefficients only is that they don’t tell you what a measured value means.

Now, let’s consider interpretability. A patient has 121° of knee flexion. I suspect that without any effort a representation of this pops into your head. Now consider two patients with 121° of knee flexion, one is 12-weeks post total knee arthroplasty and the other 12-weeks post anterior cruciate ligament (ACL) reconstruction. I suspect that you assigned a different interpretation to this range for these patients. For the patient post-arthroplasty this range sounds right, however the patient post-ACL reconstruction there is consider range to be gained. Now consider a patient who has a LEFS score of 46/80.

To gain a sense of a person’s functional status from a multi-item measure it’s tempting to look at a patient’s responses to individual items, however, this isn’t appropriate for measures validated at the total score or sub-scale level: the item level error is simply too large to obtain a confident estimate. If fact, one must work backwards from the total score to gain an impression of what it means for a typical person with that score. For our patient with a LEFS score of 46, we would begin by applying what we learned from the first question. Rather than viewing this patient as having a LEFS score of exactly 46 we would consider the patients LEFS score to fall somewhere between 40 and 52 (i.e., the 90% confidence interval for a LEFS score is about 6). Referring to Figure 1 we see that persons with a LEFS score in this range would have moderate difficulty with walking 1-mile and heavy activities, and a little bit of difficulty walking 2-block, getting in/out of the bath, and light activities. Accordingly, to convey clinically useful information I believe it’s essential for researchers to provide resources such as user friendly item maps, in addition to the requisite validity coefficients.
The answers to the next two questions provide requisite information for forming measurable patient goals: the first considers a short-term change threshold value and the second a longer-term target value. These questions also illustrate how evidence builds on clinical expertise.
To what extent can valid inferences be drawn from a measure’s change score?

There are three considerations: (1) validity, (2) the change threshold value, and (3) the confidence in interpreting a change based on the threshold value. Although considerable attention has been afforded to the first two points little consideration has been devoted to the third point. This is more than a bit puzzling, because it’s the third point that reveals the added benefit, or not, of basing a decision on the advocated threshold value. To assist in fleshing out these points consider the following patient, a 62-year old man post total knee arthroplasty.

Following your usual assessment questions and test you believe there’s about a 50% chance (it’s a toss-up) that this patient’s functional status has improved an important amount. You then administer the LEFS and obtain a score of 44 which is 4-points greater than his previous value of 40. How do you interpret this 4-point difference? To assist with you decision you refer to a recent article which provided the following information: the area under a Receiver Operating Characteristic curve was 0.76, the threshold change value was 9-points, and the sensitivity and specificity for 9-points was 81% and 70% respectively. Although it’s obvious that 4-points is less than 9-points, it’s not clear the extent to which applying this value affects the confidence you have in your decision that the patient has not improved an important amount. Figure 2 provides a suggestion for how the typical reporting of information could be augmented.

Figure 2. Information gain curve
Step 1
Focusing on the x-axis, clinical expertise has got you to the toss-up region.

Step 2
This patient’s change was 4-point which is less than the threshold value of 9-points. Refer to the lower red line.

Step 3
Find the point on the red line that corresponds to an x-axis score of 50% and then identify the corresponding point on the y-axis (20%).

Interpretation: Given there’s only a 20% chance this patient has improved an important amount your consider it unlikely that he has improved.

Presenting a figure such as this also accommodates a range of pre-measure estimates of achieving a change.

**What is the target value for this patient?**

This question addresses making a prognosis; an undertaking physiotherapists perform with each patient. Often answers to this question have been based on expected norms. For example, we’ve used information from the contralateral limb to guide expectations concerning ROM and strength, and age and sex specific normal values when the goal is complete recovery. We’ve also applied test specific critical values such as 10 seconds for the TUG and 1.2 m/s for self-paced walking speed to guide decisions. However, these approaches do not provide time specific expected outcomes nor are they informative when the optimal outcome for a person is less that a full recovery. Increasingly these limitations are being addressed with evidence-based recovery curves. Figure 3 displays a set of evidence-based recovery curves based on the work of Alcock, Cupido, and Kennedy.
The upper left panel illustrates that the target value will be time dependent. Comparing the upper right to upper left panels shows that the target value is condition dependent and not measure dependent. The bottom panels show that applying prognostic indicators leads to refined estimates of target values. These graphs show the typical or average time specific target values. To avoid “locking in” to a single value represented by the average recovery curve, a more informative presentation might include a range of values. Figure 4 presents the average recovery bracketed by a 50% confidence interval.
What is the context specific ideal reassessment interval?

Two components of a measurable goal are the outcome value of interest and the expected timeframe for meeting the goal value. This question considers the confluence of the goal value of interest and the expected change trajectory for a patient. The idea that the choice of reassessment interval impacts on the chance of correctly labeling a patient as having achieved the goal value or not has virtually gone unnoticed. To illustrate this we will look at a patient who had a pre-total knee arthroplasty 6-minute walk distance (6MWD) of 344 m and form three goals: (1) at 4-weeks his 6MWD is 297 m, (2) at 12-weeks his 6MWD is 400 m, (3) and the final goal is the target or expected “recovery” 6MWD. An additional important piece of information is that a 60 m change in walk distance can be detect with a reasonable degree of confidence (i.e., MDC_90 = 60 m).
Figure 5. Setting measurable goals

Step 1
Identify the recovery curve closest to a pre-arthroplasty walk distance of 344 m. This is the middle or red line in Figure 5.

Set the goal applying the 4-week 6MWD: To increase the patient’s 6MWD by 60 m or more in 4-weeks.

Set the goal applying the 12-week 6MWD: To increase the patient’s 6MWD by 60 m or more in 10-weeks.

Set the target value: To increase the patient’s 6MWD to 450 m or more by 22-weeks post surgery.

The consequence of reassessing patients too early is to increase the chance of mislabeling them as improved when they have not. The opposite is true if patients are assessed beyond the ideal reassessment interval.

I realize that for some measures the answers to these questions lie just over the horizon. However, we are rapidly approaching that horizon.

Raymond Moriyama

About a decade ago, the renowned architect Raymond Moriyama delivered a convocation address at McMaster University. It was a brief account of his early life. He began by mentioning that he was born in Vancouver and as a lad was sent to an internment camp in the Slocan Valley. He didn’t say he knew hardship, but rather left it to the listener to infer.
He recounted several other life events along the way, some of them humorous. Each time he left the listener to infer the deeper message. He concluded with the day he graduated from Westdale high school in Hamilton. He noted that some of his classmates received cars as graduation gifts, and when his father handed him an envelope, Raymond hoped it would contain car keys or perhaps cash to help out with his upcoming university tuition. However, there was only a brief note that read: “Into temple of life drive a golden stake.” It was only at this point than Mr. Moriyama elaborated. He highlighted that his father’s note didn’t say he had to build the temple, just make a solid contribution.

We’re fortunate to be members of a profession that allows us to make contributions to the lives of others in many ways, whether it be to help patients take their next step or breath, or contribute to a person’s quality of life at home or half a world away. Our profession even has room for a fellow to embrace statistical and measurement concepts with the hope of enhancing clinical decision-making… just a little bit. I’ve been a physiotherapist for over 40 years and during that time I can count on my hand the number of days I’ve gone to work. I wish the same for each of you.

Thank you for this wonderful honour.