Use of Computer Adaptive Testing in the Development of Machine Learning Algorithms

In this issue, Tighe and colleagues present a novel approach for “classifying large amounts of data into useful information.” Their study demonstrates the use of machine learning (ML) algorithms for clinical predictions and highlights the promise of such advances in mining secondary databases. Recently, a new approach to ML, Computer Adaptive Testing (CAT), has been applied to the collection of patient-reported outcomes (PROs). CAT relies on computer algorithms that select the most informative subset of items for each individual from a larger “bank” of self-report items. The algorithm improves efficiency by administering fewer items and reducing response burden. As the choice of items to be administered is tailored to the individual, measurement precision is not sacrificed. The ML algorithms applied by Tighe and colleagues were used to identify accurate and efficient formulas for predicting subsequent femoral nerve blocks as are the CAT algorithms. In CAT, levels of self-reported health outcomes are estimated (predicted) by administering the most efficient subset of items for any given individual. The methodological advances of ML and CAT have the potential to meaningfully impact both research and clinical practice. This editorial describes ML as it relates to the collection of PROs and explores recent applications of this computer-based measurement strategy. In addition, we discuss how the existence of CAT-assessed PROs might improve the prediction of ML algorithms such as those applied by Tighe and colleagues.

How CATs Work

Underlying every CAT is a bank of items “calibrated” to an IRT model. An IRT model is a mathematical equation that, based on a prior sample’s responses to items, predicts how new individuals with various levels of the trait being measured are likely to respond to each item. In this respect, IRT is similar to the ML-generated prediction equations described in the work by Tighe et al. Because IRT is a probability model, scores can be estimated based on an individual’s answer to any subset of the entire bank. Individuals may respond to different subsets of items, but their scores will be expressed on the full bank’s trait continuum. Thus, scores of different persons taking different subsets of items are directly comparable.

In a CAT-based measure, an initial item from the bank is administered. An individual’s response to that item provides a gross estimate of his or her level of the trait being measured. Consider the example of a CAT-based measure of pain interference [2]. The individual’s response to an initial item that has five response categories would provide substantial information about the range of pain interference the individual is experiencing, as the five choices represents five ranges of the continuum. With subsequent questions, the range of probable levels of pain interference is progressively narrowed. For example, a patient who initially chooses the response, “Very Much,” when asked, “How much did pain interfere with your day to day activities?” likely has a high level of pain interference. The items subsequently administered would be those known to be effective in discriminating among people experiencing high pain. The individual in this example might next be given an item about pain interference in social activities. Social items have been found, in IRT calibrations, to be good at discriminating among individuals who are experiencing high levels of pain interference. Conversely, an individual who responded “Never” to the question about pain interference in daily activities would be asked succeeding questions known to be good at discriminating among people with low levels of pain interference. Items about the impact of pain on recreational activities and enjoyment of life, in turn, have been found to be good at discriminating among those experiencing little pain interference.

Patients can typically complete a PRO type CAT in an average of five questions with reliability similar to a typical 15–25-item survey measure. This extreme reduction in test length can be leveraged to present 5–10 complete PRO measures in only 5–10 minutes. With this low level of overall burden, it becomes much easier to regularly
administer more comprehensive assessments to patients. In the case of pain assessment, this can result in added time to measure the emotional covariates of pain and not just basic pain constructs.

CAT test specifications can be modified to insure content coverage, meet test length restrictions, and vary the stopping rule to insure that a desired level of reliability is obtained (e.g., the level of reliability required to find a group “mean” as part of a clinical trial is much less than would be required to insure accurate assessment in clinical treatment). It should further be noted that scores produced by all IRT-based instruments created from a common item bank, including CATs and fixed-length short forms, are fully comparable. This enables easy future comparisons of scores, both locally, and across other clinical and research settings [3].

CATs in Action—PRO Measurement Information Systems

CAT-based assessments of PROs (Patient Reported Outcomes) have become much more accessible through the creation of the PRO Measurement Information System (PROMIS), an initiative funded by the National Institutes of Health (NIH) Common Fund (formerly known as the NIH Roadmap initiative), to improve assessment of self-reported symptoms and other health-related quality of life domains across many chronic diseases. A primary goal for PROMIS was the creation of a CAT system to enable efficient, psychometrically robust assessment of PROs for a wide range of chronic disease outcome research [4]. To achieve this goal, customized banks of items for measuring PROs have been developed and tested at 12 primary research and data collection sites. During the item banking process, the PROMIS network conducted focus groups, patient interviews, and The Lexile® Framework for Reading (reading level) analyses to refine the meaning, clarity, and literacy demands of all items. The item banks were administered to over 20,000 respondents and calibrated using IRT models. In total, PROMIS has developed and validated 21 adult item banks across the broad domains of physical, mental, and social health (see Table 1). Parallel item banks have been developed for pediatric and proxy use. Additional banks, developed with specific disease-focused emphases, are in the process of being validated. For each bank, CAT algorithms were created. All of the instruments have been tested in Spanish and additional translations are at various stages of completion.

The PROMIS network has conducted multiple/ongoing studies to evaluate and validate the item banks and examine the value of CAT-based administrations. It should be noted that many PROMIS instruments were developed by scientists who had previously authored the most well-known PROs. For example, the developers of the Short Form-36, the Functional Assessment of Chronic Illness Therapy instruments, and the Health Assessment Questionnaire, all had lead roles in insuring that the psychometric properties of the PROMIS instruments exceed their “parent” assessments. In practical application, the PROMIS banks have generally shown to have greater reliability, and fewer ceiling and floor effects than the “gold standard” instruments in the same area. Consequently, PROMIS instruments are also more likely to be responsive to change—better able to identify when people experience clinical benefit or decline.

The use of CAT in the assessment of health has grown exponentially, particularly due to the rapid use of PROMIS. To date, over 2,600 investigators in 45 countries have registered to use PROMIS software (which can administer both CATs and short-form versions of PROMIS instruments, as well as, other custom instruments provided by the registered researcher) [5]. Additionally, the NIH has funded research to provide score lookups that enable scores obtained using PROMIS measures to be directly compared with legacy measures. More information can be found at http://www.nihpromis.org.

PROMIS CATs have been successfully delivered online, on numerous computer platforms (including iPads), and via telephone using integrated voice response, whereby patients respond to verbal prompts by using the keypad on their phones. The instruments also have been piloted for use within electronic health records system (such as EPIC) Epic Care Ambulatory Core EMR. Using PROMIS with a system like Epic would provide a single point of patient contact to be maintained and results made available within the electronic workflow that supports doctors’ management of patient care. We also have piloted the

<table>
<thead>
<tr>
<th>Physical Health</th>
<th>Mental Health</th>
<th>Social Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain behavior</td>
<td>Anxiety</td>
<td>Ability to participate in roles and activities</td>
</tr>
<tr>
<td>Pain interference</td>
<td>Depression</td>
<td>Satisfaction with roles and activities</td>
</tr>
<tr>
<td>Fatigue</td>
<td>Anger</td>
<td>Companionship</td>
</tr>
<tr>
<td>Physical function</td>
<td>Illness impact negative</td>
<td>Emotional support</td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td>Illness impact positive</td>
<td>Informational support</td>
</tr>
<tr>
<td>Sleep-related impairment</td>
<td>Applied cognition—concerns</td>
<td>Instrumental support</td>
</tr>
<tr>
<td>Sexual function</td>
<td>Applied cognition—abilities</td>
<td>Social isolation</td>
</tr>
</tbody>
</table>

Table 1  Patient-Reported Outcome Measurement Information System domain framework

1451
regular use of CAT-administered PROs to follow patients at home on a weekly and even daily basis. Severe changes in outcomes status can generate automatic alerts to charge nurses, social workers, or dieticians. Medical societies are turning to CATs to serve as the basis for patient registries that can be used to evaluate treatments administered across the practices of thousands of member doctors and surgeons.

The PROMIS of CATs in the Treatment of Pain

The Institute of Medicine’s (IOM) report, Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research, highlights the importance of the regular assessment of self-reported pain intensity, pain impact, emotional state, and physical functioning throughout the course of treatment [6]. Traditional measures of these variables exact significant burden for both patients and providers. Consequently, a complete assessment of the breadth recommended by the IOM usually is limited to the initiation of a course of pain treatment (if then), and rarely administered thereafter. Yet, comprehensive reassessments could provide immediate information for refining treatment, improving outcomes, and enhancing the patient–clinician experience. CAT-based assessments decrease the burden of regular and comprehensive pain assessment. The increased availability and use of CATs makes it more likely that IOM’s goal of continuing assessment of pain and its correlates throughout the course of treatment is realized.

Having begun this editorial pointing out parallels between CAT and the ML methods applied by Tighe et al., it is appropriate to end by noting the synergy between the two methods. Tighe et al. obtained impressive prediction using only the set of variables typically collected prior to surgery. The database included a 0–10 pain rating scale and information regarding prior pain medication use. One has to wonder how much the prediction models they developed could improve with inclusion of a more comprehensive set of PRO measures.

The impact of a patient’s pain intensity and prior medicine use become far more meaningful when combined with the patients detailed self-reporting of their personal experience and impact of that pain. Collecting information on patients’ perceptions of pain impact, emotional state, and physical functioning, as recommended by the IOM, would provide clinicians and researchers new opportunities for exploring the interaction among pain and pain correlates, and improving the quality of clinician patient interactions. Such collection could take place in the waiting room, or between visits, on a weekly, or even a daily diary basis, to fully capture the true pain experience—and not just a 1-month retrospective. Perhaps in the future, the medical enterprise will be able to entertain real-time treatment in response to real-time patient reporting of patient pain and its corresponding emotional and physical components. Using ML and CAT can make the inclusion of these additional variables in the treatment of pain a reality.

RICHARD C. GERSHON, PhD, and KARON COOK, PhD
Northwestern University

References


