Organization on High: Expanding Use of Physician Examination Scores

Hilton P. Terrell, M.D., Ph.D.

ABSTRACT

As pressure mounts for physicians to take more frequent, costly examinations, some studies purport to show that examination scores correlate with performance measures: indices of preventive care and disease management in primary care, and complaints to licensure boards. The idea that physicians will be better at their tasks in proportion to the amount of knowledge they possess is unassailable on its face validity alone. The idea that government licensure examinations can measure this knowledge and use it to predict important practices of physicians is, however, ill founded. Examinations can predict some things, but not very important things. The data-driven approach, moreover, has great potential for misuse by those at nodes of central political and economic power. Political control, not science, appears to be the real agenda served by such studies.

In 2002, the American Medical Association (AMA) published a study, done in Canada, showing that scores on licensure examinations correlate with six measures supposed to represent physician competency. In 2007, another study done by many of the same authors showed that scores on a national skills examination predicted complaints to medical regulatory authorities.

In defense of such studies, it may be said that perfection is not to be expected, that competence is important, that incompetence abounds, and that the states have to do something to protect the public from medical harm. Medical educators need feedback in order to improve their methods. Nothing in the studies will *necessarily* be misused by those in positions of authority. Their methodology, however, has a magnetic attraction for centrists. It is an aviator's view, taking in a broad scope of landscape without the bother of fine detail. The practice of medicine needs broad epidemiologic views, but the actual delivery of care must be done with the feet on the ground, face to face with patients. The latter perspective has to rule in the end. The intention here is to offer a corrective, lest anyone in some organization on high try to misappropriate the findings of Dr. Robyn Tamblyn and her associates.

A Canadian study, published in 2002 by the AMA,[1] showed that licensure examination scores correlated with six measures of physician competency. A later study by many of the same authors, published in 2007,[2] showed that scores on a national skills examinations correlated with complaints to regulatory authorities. Articles such as these are likely to be cited as justifications for subjecting physicians to periodic, costly examinations as a condition of maintaining licensure.

Data Is Not King

It is tempting to join the argument at the level of data—answering a salvo of their favorite studies with a salvo of ours. Informational depth charges answer a spread of informational torpedoes. Sun Tzu, in his ancient classic *The Art of War* said, "The Skilful Warrior takes his stand on invulnerable ground; he lets no chance slip of defeating the enemy. The victorious army is victorious first and seeks battle later; the defeated army does battle first and seeks victory later."[3] The temptation to seek victory through battles over data must be resisted. To seek victory on this ground is to concede the war. The matter of what constitutes a competent physician is largely not a matter of such data.

While knowledge is required, there are physicians with great knowledge whom we would not allow to treat ourselves, and away from whom we direct patients. In addition to knowledge, successful medical care involves proportions of availability, congeniality, communication ability, compassion, judgment, and other characteristics. Licensure examinations, even oral examinations, cannot be expected to assess such things. At an examination, one is on one's best behavior, not tempted to shorten the test in order to get to the golf links, nor irritable from being on call, nor distracted by a declining retirement portfolio, nor overreacting to the last malpractice lawsuit. In a testing situation, one is communicating with people of very similar background, not across educational or cultural chasms. The issues presented in examinations are brief, having been largely cleaned of irrelevant material, quite unlike real life with its barnacles.

A Look at Data Nevertheless: Shots Across the Bow

The temptation to argue at the level of data is strong because the informational salvo offered in the targeted articles does not have the caliber to do the job.

In the 2002 article, the measures of physician competency are not comprehensive but supposedly representative. They are: mammography screening rate; an index of continuity of care; disease-specific, symptom-relief, and contraindicated prescribing rate as indicators of the quality of acute and chronic disease management; and referral rates.

That these measures are actually representative of good primary-care practice is not confirmed by an on-the-ground review. It is assumed. Should licensure examinations begin to be used for purposes other than pass/fail entry into medical practice, a possibility the authors mention, then the assumptions would become definitionally true. That is not the same as true; it is just "true because we say it is true." Data are from insurance claims, prescription drug databases, and hospital discharge diagnoses; they are thus derivatives of derivatives. No actual patients were interviewed. No episodes of patient care were observed. There was no review of physician reasoning in prescribing more symptomatic medicine than medicine with curative intent, or when prescribing "contraindicated" methocarbamol to a 70-year-old woman. No individual patient-centered outcomes were determined.

All of the measures were surrogates for the best question, which is, "Did these patients live longer, or better, or both?" Living longer is a rather straightforward and objective measure, if the patients of the physicians studied are in equal states of health. Even so, living longer may not reflect physician behavior. It can easily reflect physician selection of a certain class of patients. A physician who

practices in an impoverished urban center or isolated rural area may well have patients who do not live as long. One who caters to the Park Avenue set may appear to have better outcomes, irrespective of competency.

Living better is ineluctably subjective, and many patients will clearly trade away their estimate of longevity for their notion of quality. Therefore, living better and living longer are confounded with each other. To see whether a decision is good medicine, however, one has to join the patient in the arena of life, establish a line of communication, and listen. One has to risk getting vomit on one's shoes at 2 a.m. The study at issue was clean of such considerations; it has all the hygiene of computers dumping squeaky clean electrons into programs of analysis. Reviewing a physician's or patient's reasons is always messy. It is always arguable. It doesn't lend itself to counting. Centrists want to avoid this messiness. This research is an example of a common resort in medical research: things that don't count are counted, while the things that count are not counted because they are hard to count.

Another shot across the bow would be to remark on the unrepresentativeness of the substitute measures used to stand in for good physician behavior. As an example, the higher the mammography screening rate, the better the practice of primary-care medicine is supposed to be. Apparently, the researchers were unaware that the benefits of mammography are now and have always been quite controversial. There is still no definitive proof that mammographic techniques help women to live better or longer, all causes taken together. Gøtzsche and Olsen concluded from their review of mammography in Sweden that for every breast cancer death avoided by the technique, the total number of deaths increased by six, causing a firestorm of differing opinions.[4] Miller et al. found no reduced breast cancer mortality after 11 to 16 years of follow-up for four or five mammographic screenings, breast self-examination, and breast physical examination.[5]

One wouldn't have expected the promoters of expanded licensure examination score use to have known about the results of this Canadian study on breast cancer screening, since it was published in the same year as their own study. However, one would have expected them to recognize that some physicians may have had excellent reason not to herd all of their patients down the groupthink cattle chute. Some physicians with long experience in preventive medicine may detect flaws in the mammography fad. After discussing the issue with their physician, some patients may have decided they didn't want the considerable hassle of false positives in service to an uncertain benefit.[6] Perhaps it was the more astute physicians who dared to generalize from the failure of other popular "preventive" techniques to mammography.

It would take physicians astute enough to know the difference between all-cause and cause-specific mortality and morbidity, and the difference between absolute and relative risk reduction. It would take physicians who can take into account opportunity cost to the patient and who can help individual patients expend their attention, effort, and resources where there is apt to be the most benefit for them in particular, instead of applying a one-size-fits-all mentality. It would take a physician who realizes that any preventive study that enrolls a huge number of subjects—more than is needed for subgroup analysis or for erasing chance differences between the groups—is admitting on the front end that not much difference is expected. Gullible physicians may be taken in by a highly statistically significant result achieved by studying vast numbers, not realizing that the absolute and practical significance is nearly zero. One interpretation of the findings of Tamblyn et al. is that their licensure examinations were quantifying physician gullibility and inability to contextualize and apply research. The more gullible the physician, the higher their licensure scores and mammography rates. The test and the screening rate could be measuring the same thing from different angles.

The 2007 article by Tamblyn et al.[2] is a briar patch of statistics. The stated objective is "to assess whether patient-physician communication examination scores in the clinical skills examination predicted future complaints in medical practice." The outcome measure was the number of complaints that were filed against study physicians with regulatory authorities in Ontario or Quebec, and retained after investigation. I strongly suspect that the rate of 0.0491 complaints per physician practice-year is so very low as to be resistant to improvement. With the relative complaint rate differential that the test can allegedly identify no higher than 1.51, the *absolute* reduction in complaints available by any corrective method would seem to be close to nil. While no corrective method was presented in the paper, one suspects that the correction would be denial of licensure or relicensure pending successful completion of a rehabilitation program run by a profitable fascist private retraining facility. The identification of the less-apt communicators was defined by 2 standard deviations. One is again reminded of Garrison Keillor's Lake Wobegon where all the children are above average. If one "fixed" the problematic doctors, there would still be a group 2 standard deviations below the average, though the distribution would be more leptokurtic. Therein lies an endless supply for the rehabilitators.

Preoccupation with the standard deviation was also a feature of the 2002 article. Raising the passing score by "only 1 SD" in drug knowledge would have failed 16 more physicians over 4 years, reducing the risk of contraindicated prescriptions for elderly patients seen by these physicians by 42% (from 4.7% for these "low-scoring" physicians compared to 3.3% for an average physician).[1] Is this difference larger than the uncertainty in whether a prescription truly was contraindicated in a particular patient?

Evidence-Based, or GOBSAT?

Many "evidence-based guidelines" are best described as GOBSAT methodology: "good old boys (and girls) sat at a table" and decided what was best. That the experts considered scads of research is commendable. That they have impressive personal achievements means that their opinions should be considered. But in the end, which studies should be included in consideration, and how should they be weighted? A clear discussion of this methodology can be found in *Guide to Clinical Preventive Services*, 2nd ed.,[7] which includes a list of the good old boys and girls. The opinions in this commendable volume are appropriately hedged. As one example, they state:

> [I]n selected situations, even preventive services of proven efficacy may not be recommended due to concerns about feasibility and compliance. Benefits observed under carefully controlled experimental conditions may not be generalizable to normal medical practice. That is, the preventive service may have proven *efficacy* (effects under ideal circumstances), but lack *effectiveness* (effects under usual conditions of practice) [emphasis in original].[7, p liii]

There is a temptation for such guidelines to be written into requirements for insurance industry "quality" certification. The insurance company then produces score cards for physicians, and pressures them to comply.

Entanglements of such opinion with the coercive apparatus of a state medical licensing board endanger both the patients and the profession. Evidence, argument, and opinion seek to persuade. Licensure just uses the strong arm of the state. In 2002, Tamblyn et al. did not state a desire to see such entanglement, referencing only feedback from examination scores to medical education and a

"potential" for licensure regulation. If the authors did not intend to suggest use of licensure examination scores in any coercive way, it is nonetheless likely that someone will wish to do so. Misunderstanding and misuse of studies is common. The authors do it themselves in their introduction. They reference a study of the outcomes of acute myocardial infarctions according to the specialty of the admitting physician: "More knowledgeable physicians are more likely to adhere to evidence-based guidelines in the delivery of care and to achieve better outcomes."[8] If licensure examination scores partake liberally of guideline information, it is unsurprising that there would be a connection. The statement that there would be "better outcomes" conceals from the careless the fact that the outcomes are *not* case-by-case, individual evaluations of patients.

The outcomes to be evaluated are thus determined by the same mindset that sets up guidelines—the same mindset that creates some of the licensure examination questions. Finding a connection between any of these three features is announcing a decision, not a discovery. It takes the form: "We have decided to define good medical care as following our guidelines (on such items as not prescribing diphenhydramine to the elderly and making liberal referrals for specialty care). We have decided to write our licensure examination questions accordingly." It is only a short distance to: "We have decided to put your license on probation until you come to our reeducation camp for a two-week refresher on the benefits of prostate-specific antigen screening."

It is unsurprising that by 2007 Tamblyn et al. are connecting examination scores to enforcement processes by licensure agencies:

> Licensing examinations aim to assess a required level of proficiency, and thus minimum thresholds of communication ability may exist, below which the complaint rate is high and above which the rate is lower and relatively uniform.

Flawed Citations

The myocardial infarction study[8] cited by Tamblyn et al. has multiple flaws and will not carry the weight assigned to it. It did not identify board certification or actual training, only the self-identified specialty. Secondly, the study used only mortality as the measure of effectiveness. It is quite possible that generalist physicians were not attempting as many interventions, because of comorbidities, poor prognosis, or better communication on the constraints of therapy. Patients admitted by primary-care physicians had more severe disease, as shown by the lower proportion of inferior infarcts, higher rates of diabetes and cerebrovascular disease, and lower proportion of patients in Killip class 1. Statistical controls were used in an effort to manage the variation; experimental controls would have been superior. Moreover there was no accurate assessment of the splitting of management between the generalist and the cardiologist.

Since the study did not manipulate any variable experimentally, all the findings were associations, all ripe to be transmogrified into causations.

Many of these flaws were recognized and fairly stated by the authors, but such nuances are dropped in a citation when the pretence of understanding will suffice for the purposes of promoting groupthink. The most telling flaw was not noted by the authors: the study represents herd medicine, imagining that medical care is similar to managing a feedlot of cows. The information sets were abstracted from patient records, rendering the individualities of the patients a nullity. That *Sally Tucker* refused a catheterization with stent because her aunt had a stroke on the table during a similar procedure can't be allowed to matter. The generalist who has known her for years and believes that her fear of living with stroke damage outweighs even her fear of dying might honor her belief even if he judges it a misbelief. The specialist might pressure her into having the procedure. The physician who has the better cardiologic outcome is not necessarily the better physician.

In short, the study cited is like the study that cites it, obtaining mediocre answers to largely irrelevant questions.

Another study that cannot pull its load is one by Leape et al.,[9] which states that adverse drug events stemming from contraindicated prescribing accounts for a fifth of adverse drug events. This study looked at hospital systems for avoiding adverse drug events, using intensive real-time interviews. The authors found 334 potentially significant errors. There was no statement as to the denominator over which these errors sit, so no rates can be calculated. A medium-sized hospital might easily dispense half a million prescriptions in a year. A physician with a higher error rate, but a lower rate of prescribing, could have a lower absolute error rate, but the study is not designed to detect this. Tamblyn et al. were extracting information about physician knowledge from a study aimed at systems errors. Without knowing the physicians' denominator prescribing rates, they could not know how much absolute information the physicians had about drugs. Further, they chose to cite Beers et al.[10] for their criteria for contraindicated drug use in the elderly, a study already more than 10 years old. Beers et al. explicitly used GOBSAT methodology to arrive at their list. In 1997, Beers noted that the list was being misused in just the way that Tamblyn et al. have done: using criteria developed for the frailest elderly residing in nursing homes to evaluate prescribing for noninstitutionalized elderly populations. The licensure study was not limited to institutionalized elderly populations. Beers noted that "careful outcomes research" would ultimately be required to define the accuracy of such criteria.

The chances that such careful outcomes research will ever be done are not good. The history of medicine contains many instances of enthusiastic adoption of new reports, with rapid application of practices they suggest, and extension into areas marginal to the original report. The new practice acquires a following and financing, and becomes entrenched. It is never proven. It will disappear not through formal disproof, but only when an attractive new hierophant builds a new therapeutic shrine and issues a call to worship. Herein lies a danger of the association of licensure scores with reports from centralized databases pretending a valid view of primary physicians' practices.

In 2007, Tamblyn et al. admit that the communication score component of the skills examination has only "poor-to-moderate reliability." This "likely led to an underestimation of the strength of the relationship between communication and complaints." The researchers also admit that use of practice-years as a denominator would not take into account differences in frequency of patient contact, type of patients, and procedures performed, all of which may be associated with the risk of complaints. They conclude that "current examinations could be modified to test these attributes more efficiently and at earlier points in the training process"[11]—as though "efficiency" could some how remedy lack of reliability and relevancy.

Presuppositional Concerns: the Torpedoes

Enough of the warning shots! It's time to aim torpedoes below the water line. There is a deeper problem in the approach to medical care, education, and regulation than wrangling over data and design. Dr. Alfred O. Berg, one of the panel responsible for the *Guide to Preventive Medical Services* mentioned above, published an excellent essay on "the gap between evidence and practice."[12] He reviews the typical methods by which physicians today make clinical decisions, listing, for example, deferral to authority, reasoning from pathophysiology, personal clinical experience, and

so on. He describes succinctly how the U.S. Preventive Services Task Force should go about its task of producing evidence-based guidelines. He touches on some other considerations than evidence that deserve expansion, while omitting others that could have profitably been included.

On breast cancer screening, he states that numbers in an outcomes table are not directive: "[R]easonable physicians and patients might make different decisions based on the individual values they place on the risks and outcomes…. The patient's informed preferences are extremely important."[12] In other words, the data, however excellent it may be, is only a substrate. The enzymes are the patient and physician, taken together. These two individuals should be in control of the decisions.

With this insight, it is odd that Berg also listed as a problem the tremendous geographic variations in diagnostic and therapeutic practices. This moaning over variations in practice forms the bass line in centrist compositions on medical care. Yet, if patients and physicians were truly in control of decision-making, one would expect variations. Good medical care does not involve a standardized set of diseases smiting relatively standardized people, with a grudging indulgence here and there for personal foibles. At their best, diseases are fairly stable and useful explanatory constructs. They are not material things like a spoon or a spaniel. As useful as the disease model has been in Western medicine, it has lately captured too much of the imagination. Some patients tolerate what (to me) are horrendous skin lesions, refusing what (to me) are quite reasonably safe, inexpensive, and tolerable treatments. Other patients agonize over what (to me) are trivial concerns.

An epiphany came to me some years ago in the experience of a patient who wanted a bluish skin lesion removed from the back of her shoulder, although it was already known to be nonmalignant, simply because she didn't like it. The same day, I saw another patient with a patch that was nearly identical in size, color, position, and shape. It was a tattoo she had paid to have placed there.

The struggle over the proper practice of medicine is ultimately not over study design and data. It is definitional. What is a disease? How serious is it? How much effort should be expended over it? What is success? Who gets to decide the answer to all of these questions? If those of us who provide medical care to private persons cede the definitions to those in centrist positions, we may win a battle here and there, but we have already been defeated, and our patients with us. If the proper treatment of an individual—call him M. Rene Devereaux—is determined by brilliant experts as they circle in an airplane 15 km over his head in Thunder Bay, then both Rene and his physician are in serious jeopardy.

It is probably no accident that the studies by Tamblyn et al. originate from Canada. This nation has defined medical care as something that must be under central political control. The U.S. lumpenproletariat view of Canadian medicine is that you are required to help pay for everyone else's medical care but forbidden to pay for your own.

Who pays, who defines, and who decides? Centrists prefer to do it from 15 km up, as it were, where your humanity is merged with that of all others. The only significant feature is that you are a Canadian; that you are Rene Devereaux is no longer permitted notice.

It is not just medicine, but our view of life itself that is targeted by centrists. Tamblyn et al.[2] cite just two examples representative of "situations where communication is required for effective management": "discuss refusal of treatment for a terminal illness, counsel an adolescent about birth control." Examples of patient-physician communication that would receive a low score include "condescending, offensive, or judgmental behaviors," as well as ignoring patient responses. Not worthy of mention, apparently, was ignoring patients' faith or moral values. Will communication of politically incorrect messages be definitionally ineffective or offensive? Will physicians holding such views be screened out prelicensure to prevent potential future complaints?

The Nobel prize-winning economist F.A. Hayek succinctly stated the underlying issue in 1945. While he was speaking to economic issues, the same principles apply to medical questions: He stated:

> The knowledge of the circumstances of which we must make use never exists in concentrated or integrated form but solely as the dispersed bits of incomplete and recently contradictory knowledge which all the separate individuals [in society] possess. The economic problem of society is thus not merely a problem of how to allocate "given" resources—if "given" is taken to mean given to a single mind which deliberately solves the problem set by these "data." It is rather a problem of how to secure the best use of resources known to any of the members of society, for ends whose relative importance only these individuals know…. It is a problem of the utilization of knowledge which is not given to anyone in its totality.[13]

Conclusion

Anyone can know data. The particular selection and arrangement of data to solve a medical complaint, however, cannot be made available to a single mind. No Agency for Health Care Policy Research can do this. No sophisticated meta-analysis or book of algorithms can provide that mind. The economic resources for medical care are not properly a "given" resource for provincial authorities to reign over, nor are the informational ones. We need to stop playing with the idea that they are.

Hilton P. Terrell, M.D., Ph.D., practiced family medicine in South Carolina and was president-elect of AAPS. This article is published posthumously.

REFERENCES

1. Tamblyn R, Abrahamowicz M, Dauphinee WD, et al. Association between licensure examination scores and practice in primary care. *JAMA* 2002;288:3019-3026.
2. Tamblyn R, Abrahamowicz M, Dauphinee WD, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA* 2007;298:993-1001.
3. Sun Tzu. *The Art of War*. Minford J, trans. New York, N.Y.: Penguin Books; 2002.
4. Gøtzsche P, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet* 2000;355:129-134.
5. Miller AB, To T, Baines CJ, Wall C. The Canadian National Breast Screening Study—1: Breast cancer after 11 to 16 years of follow-up: a randomized screening trial of mammography in women age 40 to 49 years. *Ann Intern Med* 2002;137:305-312.
6. Schwade S. Warn women of false positives in breast tests. *Family Practice News*, May 15, 1998, p 40.
7. U.S. Preventive Services Task Force. *Guide to Clinical Preventive Services.* 2nd ed. Baltimore, Md.: Williams & Wilkins; 1996.
8. Jollis JG. Outcome of acute myocardial infarction according to the specialty of the admitting physician. *N Engl J Med* 1996;335:1880-1887.
9. Leape LL, Bates DV, Cullen DJ, et al. Systems analysis of adverse drug events: ADE Prevention Study Group. *JAMA* 1995;274:35-43.
10. Beers MH, Ouslander JG, Rollingher I, et al. Explicit criteria for determining inappropriate medication use in nursing home residents. *Arch Intern Med* 1991;515:1825-1832.
11. Beers MH. Explicit criteria for determining inappropriate medication use in nursing home residents: an update. *Arch Intern Med* 1997;157:1531.
12. Berg AO. Dimensions of evidence. *J Am Board Fam Pract* 1998;11:216-223.
13. Hayek FA. The use of knowledge in society. *American Economic Review* 1945;35(Sep):519-530.