Medically Relevant Laboratory Performance Goals: An Overview

David L. Witte, M.D., Ph.D.
Laboratory Control, Ltd.
Ottumwa, Iowa

Abstract: Medically relevant laboratory performance goals are incompletely addressed by mathematical logic built upon biologic variability, analytic variability, analytic bias, and Bayesian reasoning. Relevant goals need to also include an understanding of the perceptions and preferences of patients, their partners, providers, payers and the population served.

Human thought is subject to predictable errors. Formal study of medical cognitive processes is new. We should expect medical decisions may be equally at risk as other decisions are to cognitive flaws. Common flaws include overconfidence, inadequate feedback from previous decisions, too close an attachment to one's first idea, attachment to the status quo, extrapolating the representative case to the general, inaccurate probability estimates and framing the wrong questions. Relevant laboratory testing processes will recognize cognitive traps and attempt to minimize their impact on medical decisions.

Relevant goals are more than analytical goals and also include doing the RIGHT test at the RIGHT point in the care process and facilitating the RIGHT intervention decisions. Designing relevant testing strategies includes avoidance of cognitive errors, accurate probability (Bayesian) calculations and pathophysiologically sound reasoning.

Building relevant goals will require using the mathematical constructs of the past and adding new insights into the preferences and perceptions which define the desired outcomes. How can we design programs to identify the relevant requirements for each part of the total testing process?

"Relevant" is defined as related to the matter at hand, pertinent or important. To be relevant, pertinent or important implies the item under consideration contributes to satisfying some need.[1] We have come to specify these needs as the desired outcomes from the health care process, which include among others the reduction of a complaint or dysfunction, reassurance of health and satisfaction with the care and the cost. It is unusual for the laboratory to independently satisfy one of these needs. The laboratory contributes to the process of care by facilitating good choices in uncertain situations.

Laboratory effort should always be focused on the desired outcomes of care (Figure 1). Relevance in laboratory testing means making a positive difference in the outcome.[2] The laboratory should continuously build upon its allocentric (outside the lab) focus while maintaining excellent inside the laboratory processes. The process requires framing the RIGHT clinical questions and selecting the RIGHT tests to perform in that situation. The tests must be performed at the RIGHT time after the RIGHT patient preparation. The analytical process must have the RIGHT precision, RIGHT accuracy (analytical

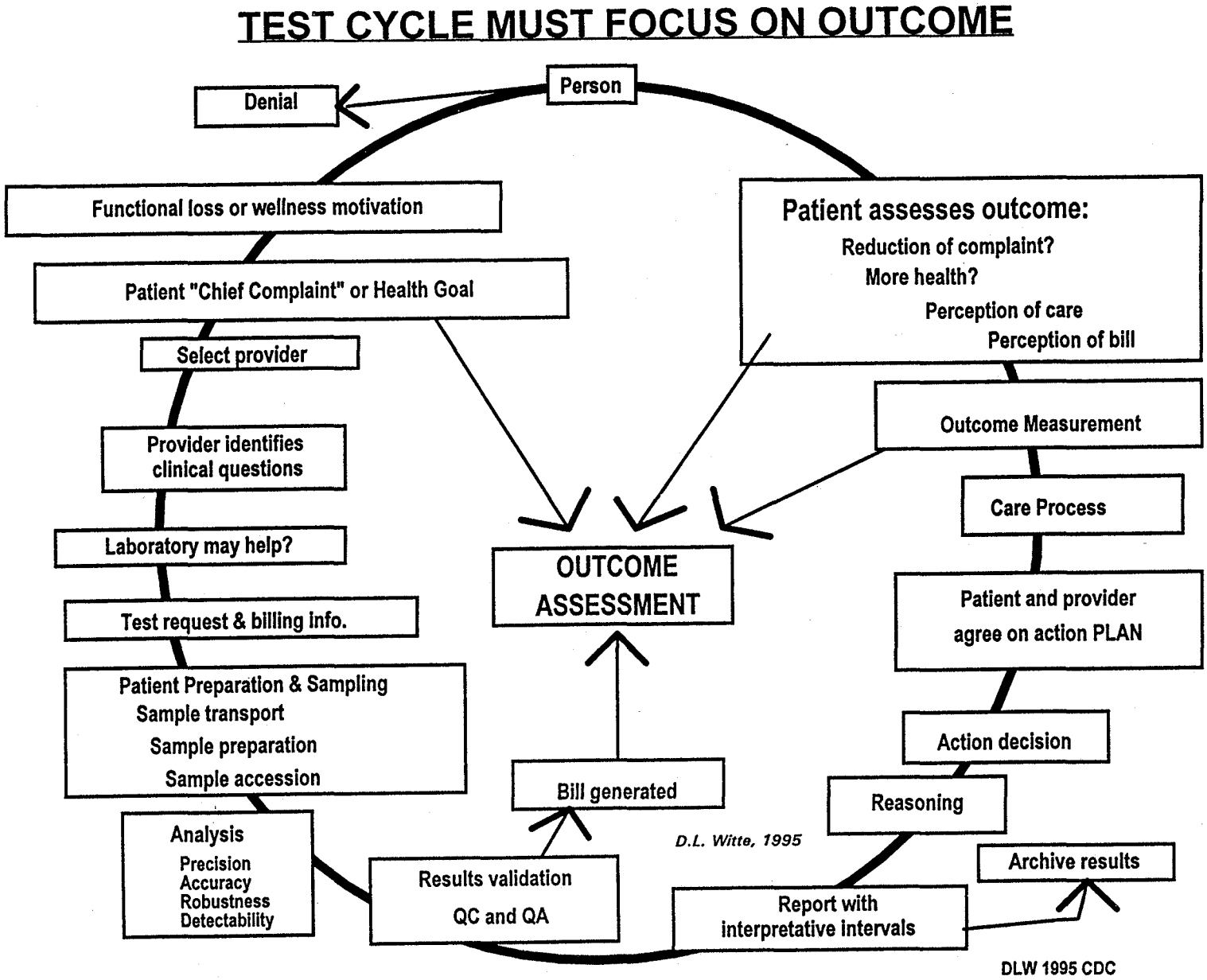

Figure 1. Laboratory efforts and outcomes of care.

specificity and calibration) and the RIGHT detectability (analytical sensitivity). The result must be reported with the RIGHT interpretive aids to facilitate the RIGHT decision process and the RIGHT choice of intervention. The RIGHT outcomes must be monitored with the RIGHT evaluative and accounting measures over the RIGHT time frame.

The analytical portion of the test cycle is the focus of Workshop 5. Much has been written about the contribution of random variation (common cause variation) and calibration bias relative to the medical relevance and analytical quality control. Much less is known about special cause variation ("blunders") in laboratories.[3,4] Elsewhere in this Institute Dr. Reed indicates a laboratory error rate of 1.1 per 1000 tests (or 1100 parts per million, PPM), and Dr. Hearn reported 27 errors in about 13,500 HIV tests or about 2000 PPM. We have studied methods comparison data where every sample had a duplicate result on the test method and singlicate from the reference method. Table 1 shows the frequency of common cause (arbitrarily chosen as 4-10 S.D. differences) and special cause (greater than 10 S.D. differences) errors in the test method results. If the Gaussian distribution approximates common cause random variation, then approximately one in 10,000 (or 100 PPM) errors are predicted to exceed about 4 standard deviations from the mean observed in a stable process. The data in Table 1 suggest we have measurable special cause variation in the analytical phase. Others have discussed the frequency of errors in the pre- and post-analytical phases.[3] Relevance includes careful consideration of common cause variation and bias, but the perception should be expanded to include pre- and post-analytical phases as well as special cause variation.

The broader focus of Workshop 8 includes pre- and post-analytical relevance (Table 2). Relevance means facilitating the right decisions. Decision making is thinking, the science of cognition. The papers in Workshop 8 illustrate intuitive, probabilistic, pathophysiologic and rule-based thinking processes. Each type of thinking is subject to bias and/or errors.[5-8] Cognition (how we think) is an important new area for study as part of the process of medical care and health promotion. Medical decisions and diagnoses are subject to the same errors in thinking as any other decision process when conditions are uncertain.[6,9-11] Neal Dawson discusses processes to avoid cognitive errors. Laboratorians have important opportunities to minimize errors by implementing effective reporting schema with decision aids and reference ranges.

Laboratorians are familiar with Bayesian probabilistic reasoning. Errors in judgment, however, can be caused by biased estimates of a test's clinical sensitivity and specificity. George Bergus outlines some of these errors which must be avoided in future test evaluation research.

Tests based on known pathophysiologic relationships illustrate causal reasoning. Joseph Keffer presents excellent examples. Gordon Schectman presents examples of rule-based, decision-making aids that improve the intermediate outcome of serum cholesterol.

Relevance is frequently measured in dollars. Unfortunately the majority of research has confused cost and charge.[12] Future research must identify the differences. The definitions of cost and charge will be difficult. Until we discontinue the common error of assuming charge is an appropriate proxy for cost, however, we will not make

Studies (n)	Tests	Common? 4-10 S.D.	Special? >10 S.D.	Total PPM
Routine cuvette chem (136)	146,393	61	43	710
Electrodes (7)	21,208	0	0	0
Immunoassays (23)	10,320	6	5	1,066

Table 1. Differences Between Replicates

- Framing the clinical questions
- Pre-analytical variation
- Analytical variation
  - Common cause (Cva)
  - Special cause (outliers)
  - Statistical process control
- Analytical bias
  - Calibration
  - “Robustness”
  - “Detectability”
- Biological normal variation
- Pathological variation
- Probability and prediction (Bayes)
- Pathophysiologic interrelationships
- Preferences of those served
- Minimization of cognitive errors
- Meeting outcome expectations
- Appropriate financial accounting

Table 2. Elements of Medically Relevant Laboratory Performance Goals

good societal health decisions. Society will benefit when we develop good understanding of the direct fixed, stepped and variable costs for providing a given health benefit. The indirect cost must also be considered, but identified appropriately, i.e., the RIGHT accounting schema.

Relevance is a perception and therefore only has meaning from a specific viewpoint.[1,13] There are many stakeholders in the processes of both disease care and health promotion. Research must consider the viewpoints of the person-patient, provider, payor and the population (4 P's). Too many research efforts focus narrowly on the viewpoints of one or two stakeholders, i.e., 2 of the P's consider policy without input from the other 2 P's. Research must take an enterprise wide or society wide viewpoint and consider expectations of all 4 P's.

References

1. DeBono E. I am right -- you are wrong. Penguin Books, 1991.

2. Witte DL. Measuring Outcomes: Why Now? *Clin Chem.* 1995;41(5):775-780.

3. Lapworth R, Teal TK. Laboratory blunders revisited. *Ann Clin Biochem.* 1994;31:78-84.

4. Gambino R. Laboratory error rates should be reported in parts per million (PPM) rather than percent - moreover, proficiency tests do not measure true blunder rates. *Lab Report.* 1994;16(3):22-23.

5. Kassirer JP. Diagnostic Reasoning. Ann Intern Med 1989;110:893-900.

6. Dawson NV. Physician Judgment in Clinical Settings: Methodological Influences and Cognitive Performance. *Clin Chem.* 1993;39(7):1468-1480.

7. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. *Science.* 1974;185:1124-1131.

8. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science.* 1981;211:453-458.

9. Kassirer JP, Kopelman RI. Cognitive errors in diagnosis: instantiation, classification, and consequences. *Am J Med.* 1989;86:433-441.

10. Schiff GD. Commentary: Diagnosis tracking and health reform. *Am J Med Qual.* 1994;9(4):149-152.

11. Williamson J. Editorial: Quality of current diagnostic performance -- Our most serious health care quality problem? *Am J Med Qual.* 1994;9(4):145-147.

12. Finkler SA. The distinction between cost and charges. *Ann Intern Med.* 1982;96:102-9.

13. Witte DL. Medically relevant laboratory-performance goals: A listing of the complexities and a call for action. *Clin Chem.* 1993;39(7):1530.

Cognitive Limitations and Methods for Improving Judgments: Implications for Establishing Medically Relevant Performance Goals

Neal V. Dawson, M.D.
Associate Professor of Medicine
Case Western Reserve University
MetroHealth Medical Center
Cleveland, Ohio

Abstract: During the past two decades, increasing amounts of information have become available about the performance of physicians for the common cognitive tasks of making diagnoses and estimating prognoses. These studies have demonstrated both excellent and poor performance by physicians. Among the reasons for less than optimal performance are cognitive limitations associated with the use of heuristics (rules of thumb) and the occurrence of cognitive biases. Obstacles to accurate probability estimation include three heuristics: availability (using the ease with which instances come to mind as a proxy for likelihood of occurrence), representativeness (pattern recognition), anchoring and adjustment (updating an initial estimate after additional information becomes available) and three cognitive biases: ego (self-serving probability estimates), hindsight (knowledge that the event occurred inflating the estimate that it would have occurred), and anticipated regret (allowing the undesirability of a diagnosis or outcome to alter the estimate of its likelihood of occurrence). Impediments to optimal information synthesis include confirmatory bias (tendency only to seek information that will confirm, rather than disconfirm, hypotheses), ignoring negative evidence (using abnormal but not normal findings to make a diagnosis or estimate prognosis), and framing (different ways to present the same information). Knowledge of cognitive limitations and development of techniques to assess components of judgmental accuracy (e.g., lens model analysis) have allowed specific methods for improving judgments to be identified and investigated. Such knowledge also should influence the research agendas of those who wish to design and test methods of providing interpretive guidance or influencing decision making based on laboratory test results.

In most clinical settings, much medical decision making occurs at an informal or intuitive level and includes such tasks as synthesizing information and estimating the likelihood of current unknowns (e.g., diagnoses) or future events (e.g., prognoses). Systematic errors in judgments (cognitive bias) have been documented in nonmedical and medical settings.[1,2,3] Cognitive limitations provide both challenges for current methods of decision making and opportunities for a systematic research effort into methods to enhance both medical judgments and outcomes.

In the sections that follow, I provide a) brief definitions and examples of specific impediments and obstacles to intuitive decision making and b) outline methods that have been used (or might be examined) to avoid or minimize the associated cognitive limitations.

Impediments to optimal information synthesis

Confirmatory bias is the tendency to seek evidence that can be used to confirm (but not disconfirm) hypotheses.[4,5] One can view such evidence as contributing to predictive value positive, rather than predictive value negative. Eddy[4] cites an article from the surgical literature where the author discusses how a "positive" mammogram can increase the likelihood of breast cancer (predictive value positive). The fact that a "negative" mammogram decreases the likelihood of breast cancer was not considered (predictive value negative). Confirmatory bias may also affect the interpretation of data. Walston[6] demonstrated that both medical students and practicing physicians used low-relevance information to support their own diagnoses.

Ignoring negative evidence is a phenomenon related to confirmatory bias. It represents the tendency to use abnormal but not normal findings in making judgments. Although both abnormal and normal findings should be used to make diagnoses efficiently, a study of practicing physicians demonstrates how they used abnormal, but not normal, findings in diagnosing pneumonia in outpatients.[7]

Framing, i.e., alternative ways of presenting the same information, can influence or even reverse medical decisions. McNeil et al[8] demonstrated how physicians' preferences for lung cancer treatment shifted between surgery and radiation therapy when data were presented as the probability of living as opposed to the probability of dying, when the treatments were specifically identified versus not identified and when life expectancy was provided rather than cumulative probability. Another medically relevant example of data presentation relates to the willingness of physicians to initiate therapy when study results are presented as relative risks as opposed to differences in absolute risks.[9-11] Mathematically equivalent rates may not be cognitively equivalent, a fact that has not been lost on pharmaceutical companies or research investigators.

Obstacles to accurate likelihood (probability) estimation

Heuristics: Familiar "rules of thumb" or other intuitive shortcuts may help simplify complex decision tasks but can lead to systematic errors in judgments.

The availability heuristic occurs when a physician uses the ease with which diagnoses or outcomes are recalled as a proxy for the likelihood they will occur. Although common events may come easily to mind, other cases and occurrences may be easily remembered because of their rarity, uniqueness, or personal meaningfulness due to a physician's research interests, personal experiences or recency of their occurrence. Not all easily recalled instances are, in fact, common. Detmer and colleagues[12] asked surgeons to estimate the surgical mortality rate for the entire Surgical service. Surgeons from high mortality specialties (cardiovascular, neurosurgery, general surgery) estimated the overall mortality rate to be more than double that of the estimated rate by surgeons from low mortality services (plastic surgery, orthopedics, urology). A surgeon's own experiences would be expected to be more available than the experiences of others and seemed to exert a disproportionate effect on judgments about the mortality rate for the entire Surgery service.

Representativeness, or pattern recognition, is a method which uses resemblance as a quick way of assessing likelihood, i.e., the probability that "A"

belongs to class "B" is directly related to the degree that "A" resembles "B." Pattern recognition is taught and commonly used in medicine but is not influenced by several factors that are known to affect actual likelihood: the prior probability of disease (or outcome), the fact that data from a small sample may be an unreliable estimator of the underlying (population) characteristic, the degree to which the event may be predictable, the likelihood of the event occurring by chance alone, and regression to the mean. For example, a single blood pressure reading may not be representative of a person's average blood pressure, and most patients with obesity, glucose intolerance and hypertension do not have Cushing's disease. Investigators need to obtain empirical evidence regarding the effects of this heuristic in medical decisions.

The anchoring and adjustment heuristic may be used by physicians in circumstances where an initial probability estimate is re-evaluated as new information becomes available. This describes the manner in which much of the diagnostic and prognostic information becomes available in medical settings, e.g., an initial impression is based on the history and physical examination, which is updated as routine laboratory and more specialized test results become available. Studies in nonmedical settings suggest that people tend to be too conservative as they adjust their initial estimate upward or downward, as if they were"anchored" to their initial estimate.[13]

Cognitive bias: These impediments to the accurate assessment of likelihood are not related to the use of cognitive short cuts or heuristics.

Ego bias occurs when estimates of probability are altered in a self-serving manner. Psychological research indicates that we tend to attribute our successes to skill and our failures to chance, i.e., "bad luck." In a study of estimated surgical mortality, Detmer et al[12] found that most surgeons estimated the mortality rate for their own patients to be lower than the mortality rate for the entire service. (This can be seen to be similar to the Lake Wobegon phenomenon where all children are above average.) Ego bias also may affect the confidence with which estimates are made.

Hindsight bias: knowledge that an event has occurred tends to inflate estimates that it would have occurred (compared with true a priori estimates). Hindsight bias has been shown to occur in clinicopathologic conferences.[14] A related phenomenon may affect judgments of physicians in quality improvement and malpractice reviews.[15]

Value induced bias is generally manifested as anticipated regret. This phenomenon can distort probability estimates when two steps in the judgment process are combined, i.e, when the likelihood estimate is influenced by the (un)desirability of the diagnosis or outcome. Inflation of probability estimates in medical settings related to value induced bias have been shown by Wallsten[6] and Poses.[16]

In the past few years, several research groups have been critical of the potential generalizability of the "heuristics and biases program" of Kahneman and Tversky.[17] For example, Lopes and Oden[18] have noted a) that the difference between "right" and "wrong" answers may be numerically rather small and b) that in certain instances heuristics seem to be properties of a general process of pattern recognition rather than individual "rules" that people apply to solve problems. Gigerenzer and colleagues[19,20] assert that internal problem representation, rather than general heuristics, drives

probability assessments. They further note that presenting problems as frequencies, rather than probabilities, leads to a smaller proportion of respondents who seem to use heuristics.

Many of the original heuristics and bias studies as well as those of their critical counterparts have been performed as pencil and paper experiments on college students. In contrast, many of the studies cited above have been performed in naturalistic medical settings. The importance of the use of heuristics and the occurrence of cognitive biases in medical settings is an empirical question about which we currently know too little. Although the medical examples cited above offer plausible arguments regarding the importance of heuristics and biases in medical settings, the frequency of occurrence and the magnitude of their effects are largely still unknown and will require ongoing research efforts.

Beyond the effect of heuristics and cognitive biases lie another set of challenges to clinical judgments. Methodologic considerations and inadequate feedback about prior judgments can greatly limit the opportunities to decipher the "true" predictive value of the data acquired in usual clinical practice. Spectrum and several forms of test related bias (verification bias, diagnostic review bias, test review bias and incorporation bias) can obscure the actual predictive characteristics of clinical data for the clinical observer.[2,21,22] In addition, ethical, cost, and pragmatic concerns often inhibit clinicians from performing the appropriate gold standard test. In many such circumstances, this can prevent feedback which could be used to recognize incorrect judgments and to appropriately alter perceptions of predictive information. An example of this phenomenon was recognized among physicians who were trying to distinguish outpatients with pneumonia from those who had other causes of acute cough.[23]

The accuracy of judgments can be partitioned into three general components: 1) prevalence (base rate), 2) discrimination (the ability to discern occasions when the event of interest will or will not occur), and 3) calibration (the ability to provide realistic probability estimates).

Over the past decade, an increasing number of common medical judgments have been scrutinized to determine their accuracy. Poses and colleagues[16] demonstrated that experienced physicians have difficulty predicting streptococcal pharyngitis in adult patients. These physicians had modest discrimination (receiver operating curve [ROC] area = .67) and a consistent tendency to overestimate the likelihood of strep (average estimate = 62%, actual prevalence = 8%). A similar tendency to overestimate occurrence rates has been demonstrated for physicians predicting pneumonia in outpatients.[24] When practicing physicians estimated the likelihood of pneumonia to be 90%, it was present in 20% of cases. In a similar study, Dawson and Speroff[25] also documented poor calibration and modest discrimination (ROC area = .73) by physicians predicting outpatient pneumonia. Tape and colleagues[26] demonstrated variability across three study sites in both accuracy and apparent physician use of clinical information for predicting outpatient pneumonia.

Tierney and co-workers[27] studied physicians' judgments of probability of myocardial infarction among emergency room patients with chest pain. They demonstrated very good physician discrimination (ROC area = .87) and generally good calibration, except for

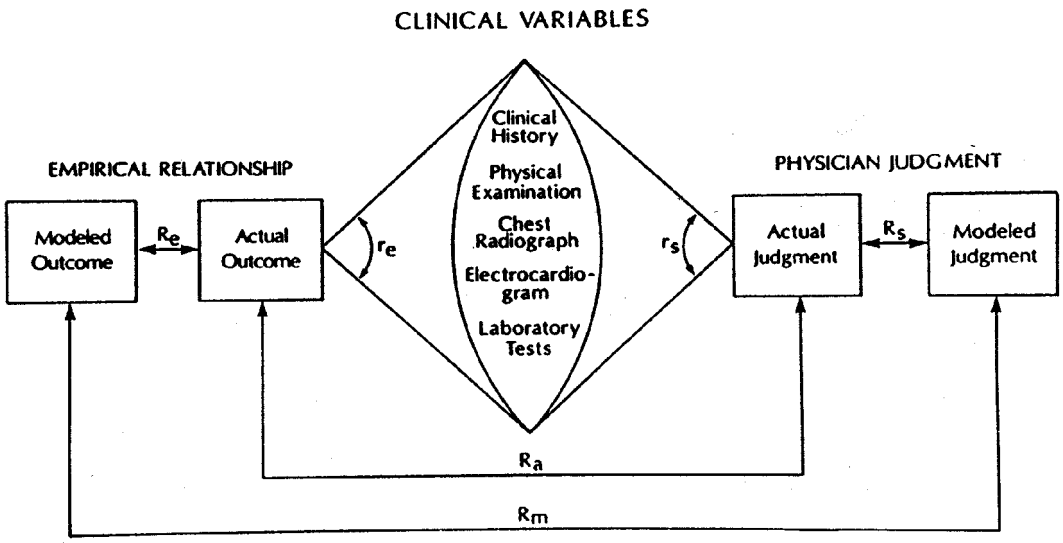

Figure 1. Lens model analysis of clinical variables.

estimates in the mid range of probabilities (between 30 and 70%) where physicians tended to overestimate the likelihood of myocardial infarction. Very good discrimination (ROC areas = .83 - .90) and variable calibration have been demonstrated for physicians' judgments of in-hospital mortality for intensive care unit patients.[28,29] Good discrimination (ROC area = .78) and a general tendency to overestimate mortality also have been documented for physicians' judgments about longer term outcomes (2 and 6 month survival) for seriously ill hospitalized adults.[30,31] These patients, physician accuracy increased as physician confidence in their predictions increased.[32] This finding is in contrast to many nonmedical and medical studies which tend

to show a lack of relationship between confidence levels and accuracy.[33]

A technique called lens model analysis (see Figure 1) has recently been introduced into medical studies and can help dissect out separate components of the judgment process.[34-37] It is designed to compare the relationships among clinical predictors (cues), and the outcome of interest (r_e), as well as cues and judgments made by physicians or others (r_s). The adequacy of the models of judgment (R_s) and outcome (R_e) can be assessed. In addition, the relationship between the outcome and judgment (R_a) and between the two models (R_m) can be compared. Speroff and coworkers[34] used it to examine why physicians have difficulty predicting hemodynamic status with noninvasive measures. They discovered that physicians seem to underutilize some important cues (30% of the explained variance came from data from the laboratory, chest radiograph and electrocardiogram whereas these data accounted for only 7% of the variance in physicians' judgments). In addition, physicians seem to place too much emphasis on other important cues (physicians placed too much weight on the clinical impression of the presence of congestive heart failure).

Hammond[38] has offered a theoretical model of decision making in situations where assessments may change over time. His theory of "dynamic tasks" asserts that the output from a task system will tend to stimulate a form of cognitive activity that lies on a continuum from calculation to intuition. The form of cognitive activity that is induced may (or may not) be compatible with the task system. He further argues that judgmental accuracy should be highest when the induced cognitive activity matches that part of the continuum that is appropriate for the task system (calculation, intuition or a combination). Hammond's formulation provides a specific structure against which future research in dynamic decision making can be tested.

Stewart and Lusk[39] recently have developed a method to assess judgmental accuracy for continuous outcome measures. Their decomposition of accuracy is explicitly linked to lens model analysis and allows specific methods for improving judgments to be identified and investigated (see Table 1). The first five components of prediction (rows 1-5) relate to judgment discrimination. The last two components (rows 6, 7) relate to calibration. Rows 3 through 7 are at least partially under the control of the judge. Columns A through N denote potential methods for improving judgments. Letters in individual cells indicate literature cited by the authors (n = nonmedical, b = both medical and non medical) that investigated a particular component of prediction. They also note areas that should be investigated (X).

In combination, the theoretical and analytic structures provided by Hammond[38] and Stewart and Lusk[39] provide a powerful construct within which the systematic development and evaluation of information systems and subsequent judgments by physicians can be evaluated.

References

1. Dawson NV, Arkes HR. Systematic errors in medical decision making: judgment limitations. Journal of General Internal Medicine 1987;2:183-7.

	C. Adapted from Stuart and Lusk (1994) **Method for Improving Judgements**																	
Component of Prediciton	**A***	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**
1. Inherent (environmentally) predictability	**n****																	
2. Fidelity of information system		**n**																
3. Match between environment and judge			**n**	**n**	**b**	**b**												
4. Reliability of acquiring information							**X**	**X**	**X**									
5. Reliability of processing information										**n**	**b**	**n**	**b**	**b**				
6. Regression bias				**b**								**n**			**X**	**n**	**n**	**b**
7. Base rate bias				**b**											**X**	**n**		**b**

**n = nonmedical studies cited, usually weather forecasting and psychology

b = both medical and nonmedical studies cited

x = no studies found specifically designed to improve judgements in these areas (although reliability of acquiring information has been shown to be a problem
in medical and nonmedical studies)

*A) Research to find new predictors
B) Develop better measures of true predictors
C) Train judge about environmental system
D) Experience with specific judgement problem
E) Cognitive feedback biases
F) Train judge to ignore non-predicitve cues
G) Develop clear definitions of cues
H) Training to improve cue judgements
I) Improve information displays
J) Replace judge with a model
K) Combine several independent judgements
L) Require justification for judgements
M) Decompose the judgment task
N) Mechanical combination of cues
O) Statistical training
P) Feedback about judgment
Q) Search for discrepant information
R) Statistical correction for bias

Table 1. Method to assess judgmental accuracy for continuous outcome measures
(Adapted from Stewart and Lusk)

2. Dawson NV. Physician judgment in clinical settings: methodological influences and cognitive performance. *Clin Chem.* 1993;39:1468-80.

3. Hershberger PJ, Park HM, Markert RJ, Cohen SM, Finger WW. Development of a test of cognitive bias in medical decision making. *Academic Medicine.* 1994;69:839-42.

4. Eddy DM. Probabilistic reasoning in clinical medicine: problems and opportunities. In: Kahneman D, Slovic P, Tversky A, eds. Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press, 1982:249-67.

5. Wolf FM, Gruppen LD, Billi JE. Differential diagnosis and the competing hypothesis heuristic: a practical approach to judgment under uncertainty and Baysian probability. *JAMA.*1985;253:2858-62.

6. Wallsten TS. Physician and medical student bias in evaluating diagnostic information. *Medical Decision Making* 1981;1:145-64.

7. Christensen-Szalanski JJ, Bushyhead JB. Physicians' misunderstanding of normal findings. *Medical Decision Making*. 1983;3:169-75.

8. McNeil BJ, Pauker SG, Sox HC, Tversky A. On the elicitation of preferences for alternative therapies. *N Engl J of Med.* 1982;306:1259-62.

9. Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. *Amer J of Med.* 1992;92:121-4.

10. Naylor CD, Chen E, Strauss B. Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Ann of Intern Med.* 1992;117:916-21.

11. Bucher HC, Weinbacher M, Gyr K. Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration. *Brit Med J.* 1994;309:761-4.

12. Detmer DE, Fryback DG, Gassner K. Heuristics and biases in medical decision making. *J of Med Educ.* 1978;53:682-3.

13. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science.* 1974;185:1124-31.

14. Dawson NV, Arkes HR, Siciliano C, Blinkhorn R, Lakshmanan M, Petrelli M. Hindsight bias: an impediment to accurate probability estimation in clinicopathologic conferences. *Medical Decision Making*. 1988;8:259-64.

15. Caplan RA, Posner KL, Cheyney FW. Effect of outcome on physician judgments of appropriateness of care. *JAMA* 1991;265:1957-60.

16. Poses RM, Cebul RD, Collins M, Fager SS. The accuracy of experienced physicians' probability estimates for patients with sore throats. *JAMA.* 1985;254:925-9.

17. Kahneman D, Slovic P, Tversky A (eds). Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press; 1982.

18. Lopes LL, Oden GC. The rationality of intelligence. Poznan Studies in the Philosophy of the Sciences and the Humanities 1994;21:199-223.

19. Gigerenzer G, Hell W, Blank H. Presentation and content: the use of base rates as a continuous variable. *J of Exper Psych.: Human Perception and Performance.* 1988;14:513-25.

20. Gigerenzer G. Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In Wright G, Ayton P, eds. Subjective Probability. New York: John Wiley and Sons; 1994:129-161.

21. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J of Med.* 1978;299:926-30.

22. Begg CB. Biases in the assessment of diagnostic tests. *Stat in Med.* 1987;6:411-23.

23. Bushyhead JB, Christensen-Szalanski JJ. Feedback and the illusion of validity in a medical clinic. *Medical Decision Making*. 1981;1:115-23.

24. Christensen-Szalanski JJ, Bushyhead JB. Physicians' use of probabilistic information in a real clinical setting. *J of Exper Psych.* [Human Perceptions] 1981;4:928-35.

25. Dawson NV, Speroff T. Validated decision aid for outpatient pneumonia [Abstract]. *Clin Res.*1989;37:774.

26. Tape TG, Heckerling PS, Ornato JP, Wigton RS. Use of clinical judgment analysis to explain regional variations in physicians' accuracies in diagnosing pneumonia. *Medical Decision Making*.1991;11:189-97.

27. Tierney WM, Fitzgerald J, McHenry R, Roth BJ, Pstay B, Stump DL, Anderson FK. Physicians' estimates of the probability of myocardial infarction in emergency room patients with chest pain. *Medical Decision Making*.1987;6:12-7.

28. Poses RM, Bekes C, Copare FJ, Scott WE. The answer to "What are my chances, Doctor?" depends on who is asked: prognostic agreement and inaccuracy for critically ill patients. *Crit Care Med.* 1989;17:827-33.

29. McClish DK, Powell SH. How well can physicians estimate mortality in a medical intensive care unit? *Medical Decision Making*. 1989;25-32.

30. Knaus WA, Harrell FE, Lynn J, Goldman L, Phillips RS, Connors AF, Dawson NV, Fulkerson WJ, Califf RM, Desbeins N, Layde P, Oye RK, Bellamy PE, Wagner DP for the SUPPORT Investigators. The SUPPORT prognostic model: prediction of survival for seriously ill hospitalized adults. *Ann of Intern Med*. 1995;122:190-202.

31. Arkes HR, Dawson NV, Speroff T, Harrell FE, Alzola C, Phillips RS, Goldman L, Fulkerson W, Califf R, Desibiens N, Oye RK, Knaus W, Connors AF, and the SUPPORT Investigators. The covariance decomposition of the probability score and its use in evaluating prognostic estimates. *Medical Decision Making*. 1995;15:120-31.

32. Connors AF, Dawson NV, Speroff T, Arkes H, Knaus WA, Harrell FE, Lynn J, Teno J, Goldman L, Califf R, Fulkerson W, Oye R, Bellamy P, Desbiens N and the SUPPORT Investigators. Physicians' confidence in their estimates of the probability of survival: relationship to accuracy [Abstract]. *Medical Decision Making*. 1992;12:336.

33. Dawson NV, Connors AF, Speroff T, Kemka A, Shaw P, Arkes HR. Hemodynamic assessment in the critically ill: "Is physician confidence warranted?" *Medical Decision Making*. 1993;13:258-66.

34. Speroff T, Connors AF, Dawson NV. Lens model analysis of hemodynamic status in the critically ill. *Medical Decision Making*. 1989;9:243-52.

35. Wigton RS. Use of linear models to analyze physicians' decisions. *Medical Decision Making*. 1988;8:241-52.

36. Schwartz S, Griffin T. Medical Thinking: The Psychology of Medical Judgment and Decision Making. New York: Springer Verlag, 1986.

37. Hursch C, Hammond KR, Hursch J. Some methodological considerations in multiple-cue probability studies [Review]. *Psych Rev*. 1964;71:42-60.

38. Hammond KR. Judgment and decision making in dynamic tasks. *Information and Decision Technologies*. 1988;14:3-14.

39. Stewart TR, Lusk CM. Seven components of judgmental forecasting skill: implications for research and the improvement of forecasts. *J of Forecasting* 1994;13:579-99.

Probabilistic Aspects of Medical Testing: Test Results, Test Performance and Medical Decision Making

George R. Bergus, M.D.
Department of Family Practice
University of Iowa
Iowa City, Iowa

Abstract: Clinicians order tests for the diagnostic information contained in test results. Laboratorians focus on the analytical performance characteristics of their tests, but the performance characteristics of concern to clinicians are test sensitivity and specificity. Test results do not directly provide the diagnostic information that clinicians seek, but Bayes' Theorem allows clinicians to use the results to make diagnostic assessments. This probabilistic approach requires appraisal of a patient's pre-test probability of disease and knowledge of a test's likelihood ratio. When the test is interpreted as a dichotomous outcome, the likelihood ratio is calculated from the test's sensitivity and specificity. A further refinement is to use the full information available in a test result by using result-specific likelihood ratios to revise probability assessments.

Estimating test sensitivity and specificity can be biased by methodologic problems which include spectrum bias, test referral bias, reference test bias, and sampling variability. These biases need to be recognized and avoided, although occasionally researchers either ignore or cannot avoid these problems.

The information contained in a test result cannot be appropriately used if clinicians disregard Bayes' Theorem or researchers use biased methodologies to assess a test's performance characteristics. Because of these factors, improved analytic performance in the laboratory might not result in the clinician having greater knowledge about the health state of a patient.

Probability Revision

Because of the inherent error in most clinical tests, using a test result in clinical medicine is a complex procedure. The clinician might attempt to use a test result alone to determine whether the patient is diseased or healthy, but this simplistic approach can lead to incorrect and dangerous conclusions. Instead, the test result should be used to revise the probability of disease that the physician had before testing by the use of Bayes' theorem. The post-test probability of disease is determined by the test result, the probability of disease before testing, and the performance characteristics of the test.[1]

It is possible to have a negative test result despite the presence of disease. Imagine that a clinician estimates a patient has a 90% pretest probability of disease and decides to confirm his/her impression with a test. If the test comes back "negative," the clinician could decide that either there is a laboratory error, or the patient does not have the disease. Instead, the clinician needs to appropriately interpret the test result by asking about the probability of disease given the negative test. This probability is easily calculated once one has an estimate of the test performance characteristics; we will assume that the sensitivity of this test is 90% and the specificity is 80%. Bayes' theorem

indicates that the probability of disease despite the negative test is 53% (appendix 1, calculation 1). Therefore, despite a negative test, the patient has a slightly better than even chance of having the disease. Similarly, a clinician can end up with a non-diseased patient with a positive test. While incongruent results can result with the clinician demanding a "better" test than the one the laboratorian is providing, the clinician should consider a better method of using the test information.

Bayes' theorem permits a new piece of information to be interpreted within the context of prior knowledge. This approach requires that the new piece of information be given an explicit weight known as a likelihood ratio (LR). The LR is the probability of a certain finding in individuals with “disease X” divided by the probability of the same finding in individuals without the disease. Although Bayes' theorem has been available for over 2 centuries, it has not become the standard method by which clinicians interpret a piece of laboratory data. Currently, for clinicians to use Bayes' theorem in their work, they have to take the report from the laboratory, look up the LR for the test result in a textbook or journal article and then calculate the post test probability. This multi-step procedure does not invite probability revision. It is possible that clinicians could be encouraged to use Bayes' theorem if the laboratorian provided on the lab report both the numerical result and its associated LR. Additionally, to ease the computational burden that comes with probability revision, the lab report could incorporate simple Bayesian nomograms.[2,3]

A test result is measured on a continuous scale but frequently used for Bayesian probability revision as a dichotomous (i.e., positive or negative) outcome. The dichotomized results are given one LR if positive (the LR+) and another if negative (the LR-). These LRs can be easily calculated because they are directly derived from the sensitivity and specificity of a test; the LR+ is the sensitivity of the test divided by [1 - specificity of the test] and the LR- is [1 - sensitivity] divided by the specificity. While dichotomizing the test result makes it easier for the clinician to use Bayes' theorem, it also degrades the available information from the test because, regardless of how extreme, there is only one LR for all "positive" results and a single LR for "negative" results.

A simple laboratory test, urine microscopy, can serve as an illustration of how information can be lost.[4] Table 1 is the 2 by 2 table for the urinalysis when 5 or greater WBC per hpf is considered a "positive" result. The LR+ for urine pyuria is 4.0 (appendix 1, calculation 1). Because of the dichotomizing, 5 WBC/hpf, has the same Bayesian weight as 10 WBC/hpf which is both intuitively objectionable and conceptually unsound. A refinement is to increase the number of categories a result can be placed into, so that unique LRs are assigned to narrower ranges of test results. As the number of categories is increased, the data from which the LRs are calculated become increasing sparse. Table 2 contains LRs calculated directly from the data set which has been partitioned into 6 levels. As can be noted, by using these additional levels, 5-9 WBC/hpf now has a different LR than 10-14 WBC/hpf although, because of sampling variability, the LRs do not monotonically increase with increasing number of WBC/hpf. When a stratified analysis is used on this small data set, one could conclude that 5 WBC/hpf is more supportive of urinary tract infection (UTI)

	UTI Present	UTI Absent
WBC ≥ 5	171	32
WBC < 5	67	148
Total People	238	180

Table 1. Microscopic pyuria data from Ferry et al [4] placed into a 2 by 2 table.

Patients with UTI	Leukocyte Count	Patients without UTI	Likelihood Ratio
124	≥ 15 WBC/hpf	14	6.70
14	10-14 WBC/hpf	11	0.96
33	5-9 WBC/hpf	7	3.57
22	3-4 WBC/hpf	22	0.76
21	1-2 WBC/hpf	49	0.32
24	0	77	0.24
238	Total Patients	180	

Table 2. Microscopic pyuria data adapted from Ferry et al. [4] and placed into 6 test-result intervals.

Leukocytes on Micro UA	Calculated LR
15 WBC/hpf	2.23
10 WBC/hpf	1.59
5 WBC/hpf	0.98
3 WBC/hpf	0.62
1 WBC/hpf	0.33

Table 3. Using the microscopic pyuria data found in Table 2, the LRs have been calculated using a MLE algorithm[5] and ROC curve analysis. [6]

than is 10 WBC/hpf!

If the modeling approach is pushed further, an unique LR can be assigned for each and every level of a test result. To deal with the problem of sparse data, the LRs can be determined using statistical techniques and modeling. Table 3 contains result-specific LRs which were calculated from the original data set at 5 different levels of WBC/hpf using a MLE estimator[5] and ROC curve analysis.[6] Alternately, logistic modeling can be used to calculate an LR at any level of WBC/hpf.[7,8] Calculating result-specific LR is beyond the skills of most clinicians but could be provided by a sophisticated laboratory information system and then attached to the lab report sent to the clinician.

Biased assessment of test performance

A second major challenge to the use of Bayes' theorem in clinical medicine is the need for unbiased estimates of a test result's LR. Because post-test probability of disease is directly related to the estimates of test performance, precise and accurate assessment of test performance is essential. Diagnostic test performance is assessed by identifying two groups of diseased and nondiseased patients and then observing how the test classifies these people. Biased estimates of test performance result in biased estimates of the post-test probability. Common biases affecting the assessment of test performance can be divided into two broad categories.[9] The first category pertains to how subjects are selected for assessing the test. The second category of biases are methodologic in origin. Before looking at these biases in greater detail, we first need to focus on two basic definitions: The first is the gold standard test, which defines the truth, as well as we can know it, about a patient's condition. The second is the index test, which we typically use in practice for information about a patient's condition because gold standard test is too expensive, too dangerous or not available.

It has been widely believed that sensitivity and specificity are qualities of a test invariant to the population selected.[10] While this immutability is attractive, it is also a misconception. Severe disease is generally easier to detect than mild disease, and therefore, the sensitivity of a test will, in part, be determined by the severity of disease in the diseased subjects being tested. This bias, known as spectrum bias, is common because many tests are developed in academic medical centers where the spectrum of disease can be very different than in a community hospital.[11] Spectrum bias can also distort the measured sensitivity of a test if researchers attempt to avoid misclassifications by using only patients they are highly certain of having the disease. This approach also gathers very extreme cases of disease.

An example of spectrum bias can be found in research focusing on the sensitivity of the urine dipstick to diagnose UTI. The range reported in the literature is wide, estimated from 66% to 100%,[12] suggesting that some of the variation in the estimates arise from the patients selected to define the sensitivity of the test. In an interesting study, Lachs calculated sensitivity of the dipstick in subgroups of patients stratified by pretest probability of UTI.[13] The dipstick had excellent sensitivity, 92%, in patients with extreme symptoms and a high clinical probability of infection. In contrast, the sensitivity in patients with few symptoms and a low probability of infection was only 56%. Whether the dipstick is a sensitive test for UTI depends on the spectrum of disease

being tested.

Spectrum bias can also impact test specificity because this measure is related to selecting controls. Naturally, if healthy medical students are used as controls, the test usually correctly identifies them as nondiseased and therefore demonstrates a very high specificity. Of greater importance is whether the test correctly identifies nondiseased patients who have signs and symptoms easily confused with the disease. Returning to the example of the urine dipstick, the literature contains a wide range of estimates of specificity for this test, from 60% to 98.4%. The study by Lachs also confirms that the specificity is greatly dependent on the patients in the nondiseased group. In noninfected patients, clinically with a low probability of UTI, the specificity was 78%; but in patients with a high probability of UTI, the specificity was much lower at 42%.

The solution to the problem of spectrum bias is to have the developers of a test clearly define the spectrum of disease in their population and to use reasonable controls for determining the specificity. To help clinicians use the appropriate test characteristics in their clinical populations, test developers could report LRs for well identified subgroups of patients.

A second type of bias, work up/verification bias, arises from researchers using the index test to decide which patients will also undergo the gold standard test. The size of this bias is directly related to how tightly the index test result is used as a selection criterion. If only patients with positive index tests are sent for gold standard tests, the index test will appear to have a sensitivity of 100% because all individuals with a positive gold standard test also have a positive index test. In this situation, the specificity of the index test will appear to be 0% because all individuals with a negative gold standard test also have a positive index test. In reality, work up/verification bias is rarely this extreme or as obvious.

A more subtle case of this bias arises if 100% of persons with a positive index test but only 20% of people with a negative index test were sent for a dangerous biopsy, the gold standard test in this example. Imagine 100 persons with positive index tests, all of whom are sent for biopsy. Eighty of the biopsies return positive. In contrast, another 100 persons have negative index tests, but only 20 are sent for biopsy. These 20 biopsies yield 10 positive results. The sensitivity of the index test appears to be 88.8% because of the 90 patients with positive biopsies 80 had positive index tests. In truth, the sensitivity of the test is much lower because many individuals with negative index tests were not included in the calculation. After realizing that only one fifth of the individuals with negative index tests ended up with biopsy, the clinician should estimate the sensitivity to be 61.5% (appendix 1, calculation 3).

This source of bias in the calculated performance of a test might seem obvious and easy to control. All patients with an index test need to undergo the gold standard test; however, because many gold standard tests are dangerous or very expensive, this control is not always instituted. Alternately, as illustrated in the above calculation, not all index test negative patients need to undergo the gold standard if a random subgroup of these patients need to be referred for this test.

A third common source of bias arises from the gold standard and is known as reference test bias. Although the performance characteristics of an index test

are quoted relative to the true disease state of a patient, they are actually calculated relative to another fallible test, the gold standard. If we assume that the gold standard determines the true health state, then we will ignore any classification errors made by the gold standard. When the index test is discordant with the gold standard, we assume that the index test is imperfect. The misclassification, however, could be on the part of the gold standard.

The relationship between an index test's true performance and its observed performance is predictably related to the prevalence of disease when the index and gold standard tests are conditionally independent (appendix 1, equation 1).[14] The observed sensitivity of the index test approaches its true value when the prevalence of disease approaches 100%; if all subjects in a population are diseased, the gold standard can no longer misclassify nondiseased individuals as diseased. At lower disease prevalence we will observe a lower sensitivity for the index test. The observed specificity of an index test will approach its true value when the prevalence of disease approaches 0% for a similar reason.

In truth, the index and gold standard tests are generally conditionally dependent, causing the relationship between index test characteristics and disease prevalence to be variable (appendix 1, equation 2). The observed sensitivity of an index test can increase, decrease, or remain unchanged with a rise in disease prevalence.[15]

This source of bias is particularly worrisome. While we need to use the best possible gold standard test, it is all too easy to fall into a circular argument about true disease state and gold standard test result. The effect of this bias is predictable and correctable if the index and gold standard tests are conditionally independent.[16] In the face of conditional dependencies, however, the size and direction of the bias cannot be predicted unless the true performance of the index and gold standard tests are known.[9]

The final source of bias to be discussed in this paper arises from sampling variability impacting the sensitivity, specificity and LR reported for a test.[17] These estimates of test performance can be numerically unstable if too few patients have been evaluated. In general, the larger the study the more stable the estimate of performance. For example, a test might have a sensitivity that has been reported to be 80%. When this estimate is based on 50 subjects, the 95% CI of this estimate is quite wide and ranges from 68.9% to 91.1%. When the same estimate of sensitivity is made based on 250 subjects, the 95% CI is narrower, 75.0% to 85.0%. When the estimate is based on 10,000 subjects, the 95% CI is 79.2% to 80.8% and the estimate is quite precise. It is obvious that this source of bias can be controlled by using large sample sizes for quantifying the performance of a test, but this is not always done because of expense, time, or the scarcity of patients with a certain disease. The clinician needs to be aware that although the analytic process behind a test might be very precise, the data on the sensitivity and specificity of the test might be unstable.

Summary

In this paper, we focused on probabilistic aspects of medical testing and presented ways for laboratorian to encourage clinicians to use this approach. The laboratory could attach result-specific LRs to the lab report and integrate computation tools into the lab report in the form of nomograms. We have also detailed four

common sources of bias that impact the LR calculated for a test result and suggested means for limiting their impact.

The clinician uses testing to obtain information about an individual, often in hopes of clarifying a clinical situation. Since tests have inherent error, their results need to be interpreted within the clinical context that the physician is hoping to clarify. It is possible that, by using Bayes' theorem with unbiased result-specific LRs, a clinician could obtain more information from a test result than is currently available.

References

1. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. 2nd ed. Boston/Toronto/London:Little, Brown and Company, 198S:23.

2. Fagan TJ. Nomogram for Bayes's theorem. *N Engl J Med.* 1975;293:2S7.

3. Glasziou PP. Probability Revision. *Pri Care.* 199S;22:235-24S.

4. Ferry S, Andersson S, Burman LG, Westman G. Optimized urinary microscopy for assessment of bacteriuria in primary care. *J fam.. Pact.* 1990;3l:lS3-161.

5. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. *J Math Psychol.* 1969;6:487-96.

6. Bergus GR. When is a test positive? The use of decision analysis to optimize test interpretation. *Fam Med.* 1993;S:6S660.

7. Albert A. On the use and computation of likelihood ratios in clinical chemistry. *Clin Chem*. 1982;28:1113-1119.

8. Knottnerus JA. Application of logistic regression to the analysis of diagnostic data: Exact modeling of a probability tree of multiple binary variables. *Med Decision Making.* 1992;12:93-108.

9. Begg CB. Biases in the assessment of diagnostic tests. *Stat in Med.* 1987;6:411-423.

10. Diamond GA. Clinical epistemology of sensitivity and specificity. *J Clin Epidemiol.* 1992;4S:9-13

11. Salive ME. Referral bias in tertiary care: The utility of clinical epidemiology. *Mayo Clin Proc.* 1994;69:808-809.

12. Kellogg JA, Manzella JP, Shaffer SN, Schwartz BB. Clinical relevance of culture versus screens for the detection of microbial pathogens in urine specimens. *Am J Med.* 1987;83:739-4S.

13. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: Lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med.* 1992;117:135-140.

14. Boyko EJ, Alderman BW, Baron AE. Perspectives. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. *J Gen Intern Med*. 1988;3:476-481.

15. Bergus GB, Witte DL. Predicting the impact of reference test bias on the observed test characteristics of an index test. *Med Decision Making*. 1994;14:425.

16. Diamond GA, Rozanski A, Forrester JS, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. *J Chronic Dis*. 1986;39(5):343-S5.

17. Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? *JAMA*. 1990;263(2):27S-278.

Appendix 1

Calculation 1

$$0.53 = \frac{0.90 * 0.10}{(0.10 * 0.80) + (0.90 * 0.10)}$$

$$\text{post-test probability} = \frac{\text{pre-test probability} * \text{sensitivity}}{(\text{1-pre-test probability}) * (\text{1-specificity}) + (\text{pre-test probability} * \text{sensitivity})}$$

Calculation 2

$$\frac{171 / 238}{32 / 180} = 4.0 = LR+$$

Calculation 3

$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = \text{Sensitivity}$$

$$\frac{80}{80 + (10 * 5)} = 61.5\%$$

Equation 1

$$S_{observed} = \frac{[p\, S_{index\ test} S_{imperfect\ reference} + (1 - p)(1 - SP_{imperfect\ reference})(1 - SP_{index\ test})]}{[p\, S_{imperfect\ reference} + (1 - p)(1 - SP_{imperfect\ reference})}$$

S is the sensitivity of a test, SP is the specificity of a test and p is the prevalence of disease.

Equation 2

$$S_{observed} = \frac{[p\, \beta_1\, S_{imperfect\ reference} + (1 - p)(1 - SP_{imperfect\ reference} - SP_{index\ test} + \beta_2 S_{imperfect\ reference})]}{[p\, S_{imperfect\ reference} + (1 - p)(1 - SP_{imperfect\ reference})]}$$

β_1 is defined as the percentage of actual disease that is positive by the imperfect reference and also positive on the index test, and β_2 is the fraction of people without disease and negative by imperfect reference who are is negative on the index test. These Betas are measures of the dependence between the imperfect reference and the index in diseased and non-diseased populations.

A Success Story for the Biomedical Model: Improved Understanding of Pathophysiologic Processes Coupled with Improved Analytical Procedures

Joseph H. Keffer, M.D.
Professor and Director
Clinical Pathology
University of Texas Southwestern Medical Center
Dallas, Texas

Abstract: As we attempt to address the question, "How can we improve laboratory testing," it is appropriate to divide the issue and to refine the question. In this writing, I suggest that "testing" from the analytical aspect is quite excellent. Those who are concerned with improvement are aware that test selection and utilization of the data are less optimal. As we consider the subject, it is appropriate to avoid generalizations about "all laboratory testing." That testing which is pathophysiologically based leads to better physician utilization than less clearly defined statistically validated testing. Examples are presented contrasting uric acid and cholesterol with thyroid function testing and measures of myocardial ischemia. Future directions for improving the contribution of laboratory testing are dependent upon continuing advances of pathophysiological understanding. This understanding contributes to the expanding the "evidence-based medicine" database, applying this to the individual longitudinal electronic medical record, and incorporating of artificial intelligence systems to empower physicians. Supporting research should augment the already existing direction of these efforts which are clearly established.

Introduction

The traditional view of the biomedical model holds that if we perform research so that we understand the normal physiology of the human body, and the disturbances associated with pathologic disease states, and we can measure those disturbances, then we can define disease. Along the lines of this model, we have achieved substantial success. Many disease states are definable in terms of precise and accurate analytical measurements performed in the clinical laboratory. This combination should lead to appropriate and successful application of the biomedical sciences to human disease. It is widely held that clinical laboratory testing is often inappropriately utilized and that the data are misinterpreted or ignored. Generalizations, however, may not be appropriate. Testing that reflects integration into pathophysiologic insights may be less of a problem than testing requiring complex statistical validation.

This conference is convened to address clinical laboratory testing and the utilization of the resultant data. The goal is clearly to determine directions for research study to improve upon the current state of affairs. There is a sense that there is too much of the wrong testing and too little of the right testing with inappropriate response to the data. In short, the current state of affairs is unsatisfactory. This response will be on two

levels: the first contrasts two categories of testing in terms of usefulness, clinical applications, success and failure, and proposes consideration which may reflect on how and why tests are used and abused.

On another level, I shall address the limits of physician practice more globally and cite the growing force of "evidence-based medicine" in conjunction with the approaching development of the electronic medical record. Ultimately, we look toward these advances in conjunction with the development of expert systems. When these are linked, we anticipate routine, transparent incorporation of evidence-based medicine to better empower physicians currently overwhelmed by the enormous body of medical knowledge.

Testing Categories

I propose to reclassify clinical laboratory testing into two categories for this discussion, *"Markers"* and *"Integral Component Elements" (ICE)*. This distinction is based on a contrast and comparison of characteristics of the two groups. (Table 1) The former is represented by cholesterol and its association with atherosclerosis, and uric acid with its association with gout. In this sense, it is recognized as contributory, not definitive testing. The *"ICE"* category will be represented in this discussion by measuring of thyrotropin (TSH) and free thyroxine (FT_4) in assessment of thyroid function and creatine kinase-MB isoenzyme (CK-MB) and cardiac troponin I (cTnI) in assessment of ischemic myocardial injury. It is proposed that much of the discussion about the uncertainty of medical relevance goals for analytical testing, reference range debates, issues of predictive value theory applied to clinical medicine, and uncertainty with regard to appropriate utilization relate to the category of "markers." In contrast, there are a growing number of clinical laboratory tests which fit the category of "ICE.". My intent is to use this term as reflecting tests which produce definitive diagnostic evidence. In common parlance, they "ICE" the diagnosis. With these, there is a basis for consensus and appropriate application of medical laboratory testing including assessment of outcomes.

"Markers" are levels of analytes which are associated with disease states based on values observed in populations characterized as diseased or non-diseased. Often, the elements used in the separation of diseased from well individuals are poorly defined because we lack the understanding of the fundamental pathophysiology of the disease or because of the heterogeneity or complexity of those processes. Common properties of this group are the focus on a solitary analytical value, the value is addressed in isolation, little focus on the individual's own reference range in contrast with a population based determination of reference range, and poor linkage between the analyte and the disease state in terms of pathophysiologic understanding. Often this is a labored association. Clinicians find the association vague, producing weak compliance with testing norms for these states. In this group, the application of decision support and predictive value theory is widely applied and essential. The association is statistical.

In contrast, the *"Integral Component Elements" (ICE)* are thoroughly understood in their essential relation to the pathophysiologic state which is addressed. They are commonly interpreted in light of serial trending values rather than in solitary determinations; they are assessed in combination with analytes related by integration into the pathophysiologic understanding; and the individual's own reference range, if defined, plays a critical role in assessment of the differentiation of health versus evolving disease. As a result, they are fundamental to assessing the outcome

associated with the disease process. Physicians intuitively find these analytes appealing because of the relation to the overall process producing better compliance with good medical practice. Bayesian rules, advocated for widespread use by some, do not apply when test results define the very entity being sought. Unfortunately, as reported by Kassirer "[1]... in many instances, Bayesian reasoning and the fundamentals of decision analysis have been incorporated into the curriculum. Yet these methods have limited scope. Bayes' rule does not explain, for example, which diagnoses or how many should be considered or discarded in a given situation, which of many possible tests is likely to have the greatest diagnostic value, or how to incorporate notions of causality into the diagnostic process. And although the principles of decision analysis are worth learning, teaching students how to apply this formal approach has been difficult. In addition, the technique is simply too cumbersome for routine clinical use."[1] While appealing, there is often insufficient information to apply Bayes' rule. We generally do not know the prior probability.[2]

The examples of uric acid in association with gout and cholesterol in association with atherosclerosis are considered in relation to the "MARKER" category. Measurement of these values in isolation provides a weak association with related diseases. Physicians often apply the tests inappropriately and interpret them inappropriately.

In the case of uric acid, the definition of hyperuricemia most commonly applied is a statistical one based on a mean and two standard deviations, reflecting the findings in a randomly chosen population of normal, healthy individuals.[3] However, "the factor(s) responsible for the formation of monosodium crystals in any individual are simply not known ..."[3] Physicians know that "the risk of developing gout increases with increasing hyperuricaemia, but the rise is not proportional and there is no point at which gout is inevitable."[4] It is no wonder that physicians seem to show a lack of respect for this type of laboratory data.

In the case of hypercholesterolemia, the statistical distribution of the "normal range" prevailed for many years, with the result that inappropriately high levels were ignored. In spite of aggressive attempts to now re-educate the physician community, compliance among physicians with regard to dealing with elevated cholesterol measurements is disappointing.[5] This may possibly be explained by continuing debates in the literature which indicate the large differences in absolute mortality from coronary heart disease at a given cholesterol level. It is acknowledged that diet, among other factors, significantly alters outcomes associated with the impact of a given cholesterol level. Physicians recognize that there is a multifactorial process involved and that the underlying pathophysiology remains a subject of debate.[6] Physicians recognize that hypercholesterolemia is a factor; however, given the continuing debate in the literature with regard to the various lipid analytes, it may well be that physicians do not respond as uniformly as the experts desire.[7,8] These observations are not presented to attempt to contradict the significance of lipid abnormality, but rather to identify the confusion which is created in the minds of practicing physicians. It is unlikely that the widespread application of predictive value theory on individual laboratory reports will change this behavior. However, the incorporation of repeated electronic reminders to physicians with specific suggested interventions suggested may be effective in achieving the stated response.[9]

By contrast, as understanding of the pathophysiology of thyroid function and the

acute ischemic coronary process evolves, the application of newer analytes with improved analytical methods permits optimal differentiation of health and disease including prognosis. We now know that the remarkable and successful adoption of the recommendations[10] for thyroid function testing which include thyrotropin (TSH) and free thyroxine (FT_4) measurement is based on the inherent stability of the physiologic relationships of these two analytes.[11] We understand the physiology of thyroid function and the pathophysiologic states which distort this. The analytical measurements are remarkably precise and reproducible, permitting their application in clinical medicine. The reference range for the individual patient is defined by their own "set point." With serial sampling, the data corroborate each other, establishing either the presence of intact physiology or the pathophysiologic abnormalities which characterize a disease state. Indeed, the application of predictive value theory is inappropriate in this setting in that the analyte levels determine the definition of the disease. For example, it is inappropriate to refer to the sensitivity of the TSH measurement for primary hypothyroidism since the TSH must be elevated to make this diagnosis. Indeed, the growing understanding of the pathophysiologic states permits the prediction of outcomes and the measurement of the analyte, TSH, represents a surrogate test which can predict the outcome of atrial fibrillation in the elderly, if untreated.[12] This is a remarkable contrast to the previous use of the laboratory with analytes, such as the uric acid or cholesterol.

A further example of the "ICE" category is now available with the myocardial markers of ischemic cardiac events.[13-15] Progressively, in recent years, we have learned to measure serial samples of CK-MB by ever more precise assays[13] followed by the cardiac troponin T,[14] and now, the completely cardiospecific marker, cardiac troponin I.[15,16] These permit the definition of myocardial injury as characterized by serial elevation and fall of these markers in association with clinically observed events which permit the foreknowledge of prognosis based on the finding of serial elevation of these analytes. They not only predict short term prognosis associated with an acute myocardial event, but in addition, subsequent cardiac mortality over a two-year period.[13]

The relevance of these distinctions to analytical goals for assay methods is self-evident. The methods must be precise and truly define the elements of the pathophysiologic process required. In turn, we have achieved such goals for both thyroid and cardiac testing applications, and these are being rapidly adopted by physicians with appropriate systematic incorporation into the practice of medicine.[10,17] Indeed, failure of physicians to apply analytical testing along these lines will predictably result in increased exposure to malpractice suits because they represent quality standards in medicine.

Improving Laboratory Test Utilization

Frequently, discussions are held relating to the inadequacy of physician utilization of laboratory testing, and the lack of understanding of physicians with regard to the sensitivity and specificity of testing. In short, there is concern with the standards of testing. This leads to assessment of the analytical performance of the laboratory, and further anxieties relating to the analytical process. Rather, the effort to improve "laboratory testing" must relate to the prior steps involved in the sequence and selection of laboratory probes and the appropriate follow-up response to the analytical data, including the interpretation and physician response. The

thesis of this paper holds that physicians will respond to meaningful laboratory testing where the association with disease is conforms to the optimal goal of the biomedical model, that is, the understanding of disease with subsequent intervention.

In the daily routine of medicine, no physician, regardless of level of expertise or training, can call upon and truly master all of the relevant knowledge appropriate to daily problem solving. This must be incorporated into the extensive variability associated with individual patient care including such considerations as age, sex, concurrent medical conditions, medications, and other variables. Consequently, laboratory data reporting reference ranges for well populations are inherently limited and not sufficient.

Evidence-Based Medicine and Expert Systems: Laboratory Implications

Advocates of evidence-based medicine appropriately argue that individual patient evaluation and medical decision making should be based on evidence tailored to the individual. The term "evidence-based medicine" was coined at McMaster Medical School in Canada in the 1980s to label this clinical learning strategy, which people at the school had been developing for over a decade.[18] Four steps are described in evidence-based medicine: 1) Setting the question, 2) Finding the evidence, 3) Appraising the evidence, and 4) Acting on the evidence. This is a growing area which will impact the practice of medicine extensively. The databases are expanding with electronically accessible avenues. The case is persuasive and further research expansion of the concept is needed.[19] Indeed, a joint publishing venture between the British Medical Journal and the American College of Physicians will be launching a new journal based on this concept.[20] Fundamentally, in order to take "evidence-based medicine" beyond the anecdotal report in the literature, we must link this database to the real-time longitudinal electronic medical record with sentinel events triggered by the entry of key laboratory data, pharmacy orders, lists of clinical problems incorporated in the electronic medical record, ongoing addition of the working diagnosis and other components.[21] These sentinels must activate inquiry into the evidence-based electronic database and use sophisticated expert systems.[22] They then will selectively present considerations to the physician for definitive response.

Future Needs

In attempting to cope with the increase in medical knowledge, the profession has explored specialization and sub-specialization. This strategy is successful in one sense, but a failure in the larger sense. As we return to an emphasis on the generalist in medicine, we have no choice but to empower the physician in a new way. Three elements are required to achieve the enhancement described: First, evidence-based medicine must be strengthened and the useable database expanded. This includes further development of pathophysiological understanding of disease. Second, the electronic medical record, a true longitudinal history of the individual, must become a reality with real-time current update of acute episodes. Third, truly sophisticated expert systems must integrate the first two so as to present relative selective considerations to an empowered physician. The technology exists and is not a limiting factor. We, as a society, can achieve this and we have no alternative if long sought goals are to be attained, and in an affordable manner.

Our research direction is clear. The analytical product of the laboratory is satisfactory; in fact, in most cases, it is

exemplary in both precision and accuracy. Now, there is need to improve physician utilization. Augmentation of existing research directions is warranted. If we do this, we will advance the goal of curing or caring.

References

1. Kassirer JP. Teaching problem-solving: how are we doing? [Editorial] 1995;332:1507-9.

2. Browne RH. Bayesian Analysis and the GUSTO trial. *JAMA.* 1995;274:873.

3. McCarty JD. Gout without hyperuricemia. *JAMA.* 1994;271:302-3

4. Smith ML. Gout, hyperuricaemia, and crystal arthritis. *Brit Med J.* 1995;310:521-4.

5. Grover SA, Coupal L, Hu XP. Identifying adults at increased risk of coronary disease. How well do the current cholesterol guidelines work? *JAMA.* 1995;274:801-6.

6. Verschuren WMM, Jacobs DR, Bloemberg BPM, Kromhout D, Menotti A, Aravanis C, et al. Serum total cholesterol and long-term coronary heart disease mortality in different cultures. Twenty-five-year follow-up of the seven countries study. *JAMA.* 1995;274:131-6.

7. Denke MA, Winker MA. Cholesterol and coronary heart disease in older adults. No easy answers [Editorial]. *JAMA.* 1995;274:575-7.

8. Levine GN, Keaney JF, Vita JA. Cholesterol reduction in cardiovascular disease. Clinical benefits and possible mechanisms. *N Engl J Med.* 1995;332:512-21.

9. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician behavior. A systematic review of the effect of continuing medical education strategies. *JAMA.* 1995;274:700-5.

10. Becker DV, Bigos ST, Gaitan E, Morris JC, Rallinson ML, Spencer CA, et al. Optimal use of blood tests for assessment of thyroid function [Letter]. *JAMA.* 1993;269:2736.

11. Keffer JH. Pre-analytical considerations in thyroid function testing. *Clin Chem.* In press.

12. Sawin CT, Geller A, Wolf PA, Belanger AJ, Baker E, Bacharach P, et al. Low serum thyrotropin concentrations as a risk factor for atrial fibrillation in older persons. *N Engl J Med.* 1994;331:1249-52.

13. Ravkilde J, Nissen H, Horder M, Thygesen K. Independent prognostic value of serum creatine kinase isoenzyme MB mass, cardiac troponin T and myosin light chain levels in suspected acute myocardial infarction. *J Am Coll Cardiol.* 1995;25:574-81.

14. Hamm CW, Ravkilde J, Gerhardt W, Jorgensen P, Peheim E, Ljungdahl L, et al. The prognostic value of serum troponin T in unstable angina. *N Engl J Med.* 1992;327:146-50.

15. Guest TM, Ramanathan AV, Tuteur PG, Schechtman KB, Ladenson JH, Jaffe AS. Myocardial injury in critically ill patients. A frequently unrecognized complication. *JAMA.* 1995;273:1945-9.

16. Adams JE, Bodor GS, Davila-Roman VG, Delmez JA, Apple FS, Ladenson JH, et al. Cardiac troponin I: a marker with high specificity for cardiac injury. *Circulation.* 1993;88:101-6.

17. Gibler WB, Runyon JP, Levy RC, Sayre MR, Kacich R, Hattemer CR, et al. A rapid diagnosis and treatment center for patients with chest pain in the emergency department. *Ann Emerg Med.* 1995;25:1-8.

18. Rosenberg W, Donald A. Evidence based medicine: an approach to clinical problem-solving. *Brit Med J.* 1995;310:1122-6.

19. Davidoff F, Haynes B, Sackett D, et al. Evidence based medicine. *Brit Med J.* 1995;310:1085-6.

20. Davidoff F, Case K, Fried PW. Evidence-based medicine: why all the fuss? *Ann Intern Med.* 1995;122:727.

21. Tierney WM, Overhage JM, McDonald CJ. Toward electronic medical records that improve care [Editorial]. *Ann Intern Med.* 1995;122:725-6.

22. Connelly D, Bennett ST. Expert systems and the clinical laboratory information system. *Clin Lab Med.* 1991;11:135-51.

A New Arena for Clinically Related Performance Goals: The Case of Cholesterol Management... or Is the Laboratory Responsible for Ensuring Quality Patient Care?

Gordon Schectman, M.D.[+]*
Edward Sasse, Ph.D.
Department of Medicine,
Division of General Internal Medicine[+], and
Department of Pathology,
Medical College of Wisconsin,
Milwaukee, Wisconsin

***Presenting Author**

Abstract: The treatment of hypercholesterolemia relies heavily upon laboratory data for proper case selection and management. Recently, the precision and accuracy of lipid testing has markedly improved, and further advances in this direction are likely to be dwarfed by the large biologic variability inherent in lipid measurements. Despite these improvements in laboratory testing, however, most individuals with hypercholesterolemia are not receiving proper therapy according to current guidelines. Barriers identified for poor physician adherence to recommended guidelines for hypercholesterolemia management include I) limited physician awareness of current recommendations; ii) lack of physician knowledge concerning proper use of drug therapy; and iii) the absence of health care delivery systems which facilitate lipid disorder management.

To overcome these barriers, more medically relevant performance goals may be sought to extend the influence of the laboratory into the clinical setting. Using existing computer technology, specific tasks for the laboratory to improve patient care may include I) sending laboratory-generated reminders to the clinician and/or patient to encourage cholesterol screening when appropriate; ii) reporting, along with cholesterol levels, the recommended LDL cholesterol goals appropriate for that specific patient, with a comment regarding whether drug therapy should be considered; iii) suggestions of specific therapeutic options for the clinician if the lipid profile had not reached optimal levels; and iv) close collaboration with health care delivery teams in the managed care setting to improve the turnaround time (speed) and costs of laboratory testing.

By assuming a more prominent role in the clinical setting, the laboratory may help to overcome existing barriers to the implementation of lipid-lowering therapy, thereby directly improving patient care.

Within the past two decades, knowledge that low density lipoprotein (LDL) cholesterol lowering correlates closely with reduced coronary heart disease (CHD) morbidity and mortality heralded an era where accurate lipid measurements suddenly became necessary to identify and treat individuals with lipid abnormalities.[1] Randomized studies

documenting that interventions to reduce LDL cholesterol significantly reduced coronary heart disease events confirmed initial epidemiologic associations,[2] and encouraged the formation of the National Cholesterol Education Program (NCEP) to develop national guidelines.[3,4] These practice guidelines recommended cholesterol screening for all adults, and suggested management algorithms to assure that patients were appropriately diagnosed and treated to achieve specific LDL cholesterol goals. Target LDL cholesterol goals vary for each patient, depending upon the number of cardiovascular risk factors present and the overall heart disease risk.

For meeting the performance goals outlined by the NCEP, accurate and precise laboratory tests are necessary to reduce the potential for incorrect classification of hypercholesterolemia.[5] In particular, accurate LDL cholesterol measurements are essential, as successful therapy hinges upon the ability of the patient to reduce LDL cholesterol below a specific level.[4] Because LDL cholesterol calculations depend upon total cholesterol, triglyceride and high density lipoprotein (HDL) cholesterol assays,[6,7] accurate and reliable measurements of all these lipid measurements are necessary.[8] Several publications have highlighted the importance of accurate measurements and pointed out the consequences of poor test precision and accuracy.[5,8-11] As a result, the Adult Treatment Program Laboratory Standardization Panel concluded that total cholesterol accuracy and precision should be reduced to less than 3%.[12]

With rapid technical improvements in commercially available autoanalyzers, accuracy and precision standards mandated by the Laboratory Standardization Panel appear to have been met. Recent papers report precision data well within 3% for total cholesterol, and also less than 3% for triglycerides and HDL cholesterol.[13,14] As a result of these improvements, the LDL cholesterol calculation also has improved accuracy and precision.[14] Because the large biologic variability inherent in most lipid measurements remains unchanged, total test variability will not be appreciably improved from further refinements in cholesterol, triglyceride and HDL cholesterol assays.[8]

Despite technical improvements in lipid testing, achievement of LDL cholesterol goals through appropriate treatment is currently substandard, suggesting that clinicians may not be using these tests properly.[15] Although at least 50% of patients with coronary heart disease will benefit from cholesterol lowering medications, surveys show that only between 8 and 30% receive it.[16-18] Therefore, modern advances in laboratory testing to improve test precision and accuracy have not correlated with the ability of the clinician to correctly use this laboratory information to implement NCEP guidelines.

For satisfactorily implementing hypercholesterolemia management guidelines, the question arises as to whether the laboratory should directly assist the clinician to properly use results of cholesterol testing. In other words, does the domain of the laboratory extend beyond ensuring adequate test accuracy and precision, particularly when the test is being incorrectly used by the clinician? Should laboratory performance standards include the responsibility to ensure the presence of a dialogue between laboratory and clinician to prompt the clinician to use laboratory information wisely? Should laboratory personnel provide guidance to the clinician to increase the likelihood that cholesterol testing is utilized correctly, leading to cardiovascular risk reduction and improved patient care? Currently, most laboratories only ascertain that each test is performed with appropriate

accuracy and precision, and report values along with appropriate normal/abnormal values for a specific reference population. Some laboratories have also included a table reviewing NCEP recommendations for total cholesterol, HDL cholesterol, triglycerides, and LDL cholesterol.

A ready familiarity with computerized processing of laboratory data and automated test reporting enables the laboratory to consider novel approaches to influence the clinician. Computerized information retrieval and display systems, like reminder systems, have been shown to have an impact on physician behavior. For example, introduction of a clinician's workstation to facilitate data retrieval resulted in a 32% reduction in laboratory testing charges in two bone marrow transplant units.[19] Similarly, physician test-ordering behavior can be improved through concurrently providing displays of past test results,[20] probability estimates of obtaining an abnormal result,[21] or test charges.[22] The potential for the computer to influence physician behavior has been recently reviewed.[23] These studies indicate that creative uses of computer technology can enhance clinician interpretation and implementation of laboratory data.

As the computerized medical record and comprehensive clinical databases become increasingly utilized, information systems are being refined which fully integrate all clinical data, including that obtained from the clinical examination and laboratory. With this technology, the potential of the laboratory to provide powerful decision support for the clinician becomes very realistic. For example, incorporating into the clinical database the patient's disease profile, risk factor status, and drug regimen allows an assessment of whether LDL cholesterol values have reached goal levels, and makes possible automated suggestions regarding further therapy. Such information can be used either by the clinician or by allied health professionals to re-evaluate and modify therapy until goal lipid values are finally achieved. This level of feedback has been demonstrated to be helpful in improving cholesterol management.[24]

Computerized reminders and/or feedback to improve hypercholesterolemia management and/or screening could include:

1) Prompts for cholesterol screening if the patient has not had a cholesterol measurement performed within the past 3 years. These reminders, generated for the clinician and/or the patient, are likely to improve screening rates and increase the number of patients receiving adequate treatment.

2) Interpretation of the triglyceride, HDL and LDL cholesterol values within the context of the NCEP guidelines, suggesting whether diet and/or drug therapy should be considered for the patient. The report could evaluate the specific risk factor status of the individual and advise the clinician whether goal levels have been achieved. This type of report would allow the clinician to apply appropriate treatment guidelines to his patient without memorizing all aspects of the guidelines.

3) Treatment recommendations including whether diet or drug therapy is appropriate, and specifying which drug or drugs would be reasonable considering the clinical setting. To implement this approach, clinical patient information and simple treatment algorithms could be

SMITH, ALCUS ID # 318 Age 61.9 Drug Trial # 41

Hypolipidemic Drug Therapy (Daily dose in grams, mg, capsules, or tablespoons)		
Colestipol:	Niacin: 1500	Gemfibrozil:
Psyllium:	Levastatin: 40	Fish Oil:

Major Illness		Other Risk Factors	
CHD:	No	Gender?	Yes
Diabetes:	No	Obese:	No
Htn:	Yes	Smoker?	No
PVD:		Fhx CHD:	Yes
CVA:		BOG LVH:	

Risk Factor Totals: 3
Age Adjusted CHD Risk/10 yrs:
No Risk Factors: 10.6%

	Date of Last Lipid Profile	Number of Profiles	TC	TG	HDLC	LDLC	LDL/HDL	CHD Risk/10 yr
Baseline		2	286	126	35.5	225	6.33	26.8%
Drug Rx	12/8/92	3	241	139	36.0	177	4.92	22.8%
Change from Baseline			-16%	10%	1%	-21%	-22%	-15%

Absolute Risk Change: -4.0%

QA Corner

Yearly Drug Costs: $906 Rx too costly; Above 80th% ($444)

Cost per Unit Change in LDL/HDL Ratio: $643 Rx NOT Cost-Effective! Above 80th% ($462)

Keep Trying! Therapy has NOT achieved LDL goals or reduced LDL/HDL ratio by >30%.

Figure 1. Example of patient report comparing baseline results with results obtained during treatment.

programmed into the computer to provide this information to the clinician.

At the Medical College of Wisconsin and the Milwaukee Department of Veteran Affairs Medical Center, a computerized database integrating laboratory and pharmacy data with information derived from the clinical examination has been in existence since 1988 for use in the Lipid Disorder Treatment Program, and allows a comprehensive computerized assessment of patient progress. The database formats a report comparing baseline lipid profiles with those obtained during treatment and prepares a report available to the clinician as the patient is seen at the clinic visit (see Figure 1). The tabular printout allows the clinician to determine effectiveness of current drug therapy by comparing the mean of recent lipid values obtained on the current regimen with baseline values. A summary of risk factors is compiled to allow the clinician to quickly assess CHD risk, which is also computed according to risk estimates from Framingham.[25] Brief summary statements are provided to the clinician assessing whether NCEP goals have been achieved for that particular patient, describing whether therapy has been effective, and whether the response for that particular patient justifies the cost of therapy, in comparison with cost-effectiveness data from patients of similar risk status in the clinic. A similar report is prepared for the patient, describing his/her progress in simple terms.

Developing this system serves several goals. First, it enhances the efficiency of the clinic visit, allowing the clinician to spend more time discussing current patient concerns, rather than spending time locating important data scattered

in different places in the chart. Second, it automatically provides for storing clinical data which can be used to detemine the effectiveness of therapy administered within the clinic. This clinical data may also serve important quality monitoring functions. Third, it provides a structure to assist physician extenders in taking a more active role in clinical management of disease by using the computerized decision support as an initial basis for clinical decision-making. Fourth, it enhances communication with the patient through a computer generated personalized report specific for the patient discussing his/her progress.

At our own site, this computerized system has been effectively used in some of these areas. We have evaluated the effectiveness of cholesterol-lowering drug therapy administered in the clinic setting,[26] assessed our own ability to achieve defined lipid goals among our patients treated with cholesterol-lowering drugs,[27] evaluated the ability of allied health professionals to use this system effectively to implement cholesterol-lowering therapy, thereby serving as cost-effective "physician extenders",[28] and used this computerized infrastructure to test alternative approaches to improve administration of cholesterol-lowering drug therapy.[29,30]

Additional support that the laboratory could provide to improve clinician performance includes rapid performance of laboratory tests (within minutes or hours) so that the clinician can review results with the patient at the same visit, rather than scheduling a second clinic visit to discuss results and consider therapeutic changes. In addition, if screening total cholesterol and HDL cholesterol values are not normal, then the laboratory could consider performing other lipid tests automatically (perhaps using a direct LDL cholesterol assay if the patient wasn't fasting), thereby providing more information to the clinician and prompting further action if levels are undesirable.

In conclusion, by assuming a more prominent role in the clinical setting, the laboratory may help to overcome existing barriers to implementing lipid-lowering therapy, thereby directly improving patient care. The use of computer technology offers an ideal avenue for this process to proceed. In addition, this approach may have applicability to other areas in clinical medicine which rely heavily on laboratory support for therapeutic decision-making.

References

1. Staniler J, Wentworth D and Neaton JD. Is relationship between serum cholesterol and risk of premature death from coronary heart disease continuous and graded? Findings in 356,222 primary screenees of the Multiple Risk Factor Intervention Trial (MRFIT). *JAMA*. 1986;256:2823-2828.

2. Brown BG, Zhao XQ, Sacco DE and Albers JJ. Lipid lowering and plaque regression. New insights into prevention of plaque disruption and clinical events in coronary disease. *Circulation*. 1993;87:1781-1791.

3. The Expert Panel. Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. *Arch Intem Med*. 1988;148:36-69.

4. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol Levels. Summary of the second report of the National

Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel 11). *JAMA*. 1993;269:3015-3021.

5. Weissfeld JL and Holloway JJ. Precision of blood cholesterol measurement and high blood cholesterol case-finding and treatment. *J Clin Epidemiol*. 1992;45:971-984.

6. Friedewald WT, Levy RI and Fredrickson DS. Estimation of the concentration of low density lipoprotein cholesterol in plasma, without the use of the preparative centrifuge. *Clin Chem*. 1972;18:499-502.

7. McNamara JR, Cohn JS, Wilson PWF and Schaefer EJ. Calculated values for low density lipoprotein cholesterol in the assessment of lipid abnormalities and coronary disease risk. *Clin Chem*. 1990;16:36-42.

8. Schectman G and Sasse E. Variability of lipid measurements: Relevance for the clinician. *Clin Chem*. 1993;39:1495-1503.

9. Cooper GR, Myers GL, Smith SJ and Sampson EJ. Standardization of lipid, lipoprotein, and apolipoprotein measurements. *Clin Chem*. 1988;34:B95-B 105.

10. Superko HR, Bachorik PS and Wood PD. High-density lipoprotein cholesterol measurements. *JAMA*. 1986;256:2714-2717.

11. Cooper GR, Myers GL, Smith SJ and Schlant RC. Blood lipid measurements. Variations and practical utility. *JAMA*. 1992;267:1652-60.

12. National Cholesterol Education Program. Current status of blood cholesterol measurement in clinical laboratories in the United States: Report from the Laboratory Standardization Panel. *Clin Chem*. 1988;34:193-201.

13. Smith SJ, Cooper GR, Myers GL and Sampson EJ. Biological variability in concentrations of serum lipids: sources of variation among results from published studies and composite predicted values. *Clin Chem*. 1993;39:1012-22.

14. Bookstein L, Gidding SS, Donovan M and Smith FA. Day-to-day variability of serum cholesterol, triglyceride, and high-density lipoprotein cholesterol levels. Impact on the assessment of risk according to the National Cholesterol Education Program guidelines. *Arch Intern Med*. 1990;1 50:1653-7.

15. McBride PE, Plane MB and Underbakke G. Hypercholesterolemia: The current educational needs of physicians. *Am Heart J*. 1992; 123:817-824.

16. Cohen MV, Byme MJ, Levine B, Gutowski T and Adelson R. Low rate of treatment of hypercholesterolemia by cardiologists in patients with suspected and proven coronary artery disease.

Circulation. 1991;83:1294-304.

17. Northridge DB, Shandall A, Rees A, and Buchalter MB. Inadequate management of hyperlipidemia after coronary bypass surgery shown by medical audit. *Brit Heart J*. 1994;72:466-67.

18. The Clinical Quality Improvement Network (CQIN) Investigators. Low incidence of assessment and modification of risk factors in acute care patients at high risk for cardiovascular events, particularly among females and the elderly. *Am J Cardiol*. 1995;76:570-73.

19. Connelly DP, Sielaff BH and Willard KE. A clinician's workstation for improving laboratory use: Integrated display of laboratory results. *Am J Clin Pathol*. in press. 1995;

20. Tierney WM, McDonald CJ, Martin DK and Rogers MP. Computerized display of past test results. Effect on outpatient testing. *Ann Intem Med*. 1987;107:569-74.

21. Tierney WM, McDonald CJ, Hui SL and Martin DK. Computer predictions of abnormal test results. Effects on outpatient testing. *JAMA*. 1988;259:1194-8.

22. Tierney WM, Miller ME and McDonald CJ. The effect on test ordering of informing physicians of the charges for outpatient diagnostic tests. *N Engl J Med*. 1990;322:1499504.

23. Elson RB, Connelly DP. Computerized patient records in primary care. Their role in mediating guideline-driven physician behavior change. *Arch Fam Med*. 1995;4:698-705. (Review)

24. Headrick LF, Speroff T, Pelecanos HI and Cebul RD. Efforts to improve compliance with the National Cholesterol Education Program Guidelines: Results of a randomized controlled trial. *Arch Int Med*. 1992; 151:2490-2496.

25. Anderson KM, Wilson PWF, Odell PM and Kannel *YVB*. An updated coronary risk profile. A statement for health professionals. *Circulation*. 1991;83:356-362.

26. Schectman G., Hiatt J. and Hartz A. Evaluation of the effectiveness of lipid-lowering therapy (bile acid sequestrants, niacin, psyllium and lovastatin) for treating hypercholesterolemia in veterans. *Amer Jl of Cardiol*. 1993;71:759-65.

27. Schectman G and Hiatt J. Drug therapy for hypercholesterolemia in patients with cardiovascular disease: Factors limiting achievement of lipid goals. *Amer J Med*. 1995; In press.

28. Schectman G, Wolff N, Byrd JC, Hiatt JG and Hartz A. Physician extenders for cost-effective hypercholesterolemia management. *J Gen Intem Med*. 1995; In press:

29. Heudebert GR, Ruiswyk J Van, Hiatt J and Schectman G. Combination drug therapy for hypercholesterolemia: The trade-off between cost and simplicity.

Arch Int Med. 1993;153:1828-1837.

30. Schectman G, Hiatt J and Hartz A. Telephone contacts do not improve compliance to niacin or bile acid sequestrants. *Ann of Pharmacotherapy*. 1994;28:29-34.

Summary of Workshop 8: Establishing Medically Relevant Performance Goals for the Laboratory

Facilitator: David L. Witte, M.D., Ph.D.
Laboratory Control, Ltd.
Ottumwa, Iowa

CDC Liaisons: Tina Stull, M.D. and Mark White

Key Questions:

1) How are clinically related performance goals established and evaluated?
2) How can clinically related performance goals be translated to medically relevant performance goals?

The Presentations

Relevance means making a difference. Making a positive difference in care processes and care outcomes requires good decision making. Good decisions are required in all phases of health care: the pre-analytical and pre-clinical, the laboratory analytical and clinical, the post-analytical and post-clinical phases. This workshop included five presentations and a vigorous discussion of current knowledge and desired future improvements in clinical decision making utilizing laboratory data. Developing clinically related, medically relevant performance goals requires a clear and quantitative understanding of how a change in the precision and accuracy of a laboratory result may change the decision-making process and therefore may change a health care outcome.

Medical thinking or cognition involves an interplay of at least four different thinking strategies: intuition, probabilistic reasoning, pathophysiologic or causal reasoning and the use of rules or heuristics.[1] Each paper presents details on these cognitive processes.

Dawson discusses the common thought processes used by clinical decision makers. Clinical decisions are at risk for all the potential errors and biases known to occur in other types of decision making. Understanding these errors will facilitate developing analytical goals. More importantly, laboratory reports can be formatted with more appropriate decision aids to prevent the common errors in the decision process.

Dawson points out that clinical decision makers frequently overestimate the likelihood of disease in a given patient. Two cognitive biases contribute to this phenomenon: If the negative consequences of an error of omission (e.g., missing a streptococcal throat infection in patient with previous rheumatic symptoms) far outweigh the consequences of the obverse error, the anticipated regret causes one to overestimate the likelihood of streptococcal throat infection in these patients. Similarly, the availability bias causes one to overestimate the probability of the most easily recalled possibility. All laboratories have seen a change in test utilization after a conference or presentation of a problem patient.

Understanding these predictable biases in decision making can guide efforts to define the precision, accuracy and supporting interpretive information necessary to facilitate the desired decision. Will the decision process and

ultimate clinical outcome be improved if the decision maker knows the thyroid stimulating hormone (TSH) result has an analytical uncertainty of 10%? Will the decision process be improved if the decision maker knows this method for glycosylated hemoglobin is predictably 10% higher than the method used in the Diabetes Control and Complications Trial? Will the decision process change if the blunder rate[2-4] is known to be 1 in 800? Will the decision process be improved if the decision maker knows the frequency of positive streptococcal throat culture in children of this age has been approximately 20%? Or that 99% of people previously tested had a percent transferrin saturation less than this patient? Or 99% of clinic patients had an alanine aminotransferase (ALT) less than this patient? Or that the likelihood percentage for an abnormality of this magnitude is 100?

Bergus discusses the possible errors in evaluating the Bayesian predictive properties of a laboratory test. Adequate interpretive data cannot be provided with inadequate test evaluation. The precision and accuracy of the test strongly influence the predictive value and choice of decision levels. Probabilistic reasoning with laboratory data is a cornerstone for the relevance of laboratory testing. Specificity and sensitivity are not fundamental properties of laboratory tests but rather observations of the interaction of tests and tested populations. Will a change in precision or accuracy change the ability of a test to facilitate a correct decision? How can we determine and assure adequate precision and accuracy and demonstrate these properties to the decision makers?

Keffer outlines the biochemical model of disease and the use of well characterized laboratory tests to identify specific pathophysiologic processes. This is causal reasoning at its strongest. We need to strive for more complete understanding of both health and disease to identify more biochemically defined tests. In causal reasoning is usually found a strong correlation between analytical precision and accuracy and the ability to make an accurate clinical decision.

Schectman shows the positive outcomes associated with decision making by predetermined rules. Displaying drug doses and lipid concentrations together facilitates decisions that produce lower blood lipids. Combining the biochemical model and decision rules can be beneficial.

The Discussion

As laboratorians seek to define relevant goals, we must take a broad view. We must facilitate the four different reasoning strategies. Relevance requires traversing the boundaries between pre-analytical and post-analytical factors. Non-laboratorians rightly expect that laboratory quality will be high. We must continue to provide and improve that quality. The workshop discussants believe the major opportunities for quality improvement lie across the boundaries that traditionally enclose the laboratory.

Medical relevance means attaching laboratory results to other data and interpretive information and integrating the data into the care processes. Medical relevance is providing equal quality results in multiple locations and care settings. Two adjectives were prominent in the group discussion: delightful and informative. The laboratory report must be informative enough to prevent judgment errors and delightful to use. Delightful reports allow easy visual interpretation of both the result and the reference information. Delightful reports will integrate laboratory data with other data such as drug doses and prevalence of specific findings. The delightful report format will improve the intellectual quality of decision

making by leading in the desired direction.

Medically relevant means a positive impact on outcomes. We must know both the expectations for outcomes and the outcomes being achieved. Outcomes are quantitatively measurable. Satisfaction with care, cost of care, days lost from work, and days with impaired activities are a few of the relevant outcomes. These outcomes are not easily measured but we must increase our efforts. Only by knowing if a change in laboratory performance is associated with a change in outcomes will we be able to define relevant goals. Do laboratory data plus sound reasoning reduce later care costs? We need to avoid some of the predictable errors. Charge for care is rarely an acceptable quantitative proxy for cost of care. It is a well known cognitive bias that we tend to under-value the outcomes of preventive care. Are we challenged to evaluate the outcome when nothing bad has happened?

Medically relevant goals must be defined through collaboration of multiple stakeholders. Each stakeholder must also be aided and coached to avoid the cognitive errors discussed. The stakeholders' list is long. One stakeholder has frequently argued the non-relevance of many laboratory procedures through Bayesian logic using one laboratory result at a time and concentrating on the value of the positive results. The discussants believed that multivariate approaches with a more appropriate understanding of the value of the negative or normal result would yield an analysis that more accurately reflected clinical decision making.

Research Agenda

The discussants defined four general areas for fruitful future research: First, outcomes measurement and the attribution of outcomes to laboratory information must be better defined. The gaps between expected outcomes and observed outcomes provide a major opportunity to identify relevant new laboratory practices. Second, the cognitive use of laboratory data offers significant opportunities for improvement. Understanding the impact of results on decisions is largely unknown. Will reports with decision aids impact the decisions and outcomes? Can we devise multivariate predictive schemes to evaluate test impact? What is the decision making value in the normal result? Third, many test evaluations are subject to predictable biases. How can we identify these biases and prevent errors in decision making? Fourth, can improved test request systems providing interpretive information in the pre-analytical phase improve test utilization, other resource utilization and outcomes? Progress in these four research areas will move us toward defining medically relevant analytical performance goals. The discussants encourage taking an enterprise-wide or care system-wide view of the relevance of laboratory tests and discover the impact of changes in laboratory performance on the decision making process and outcomes of the care process.

References:

1. Kassirer JP. Diagnostic Reasoning. *Ann Intern Med.* 1989;110:893-900.

2. Lapworth R, Teal TK. Laboratory blunders revisited. *Ann Clin Biochem.* 1994;31:78-84.

3. Gambino R. Laboratory error rates should be reported in parts per million (PPM) rather than percent - moreover, proficiency tests do not measure true blunder rates. *Lab Report.* 1994;16(3):22-23.

4. Witte DL. Medically Relevant Laboratory Performance Goals. This symposium.