Toward Optimal Laboratory Use

Do We Know What Inappropriate Laboratory Utilization Is?
A Systematic Review of Laboratory Clinical Audits

Carl van Walraven, MD, MSc, FRCPC; C. David Naylor, MD, DPhil, FRCPC

Objective.—Laboratory utilization has steadily increased, and some studies suggest inappropriate utilization. Therefore, we wished to assess studies that measure inappropriate laboratory use in light of methodological criteria.

Design.—Systematic review of published studies.

Data Sources.—MEDLINE, HEALTHSTAR, and EMBASE databases were searched from 1966 to September 1997 using a broad and inclusive strategy with no language restriction. In addition, the references of all retrieved studies and 3 textbooks on diagnostic testing were hand-searched.

Study Selection.—All studies that provided and applied criteria for inappropriate laboratory use.

Data Extraction.—Studies were categorized based on whether the criteria were implicit (objective criteria for inappropriate utilization not provided or very broad) or explicit. Guidelines for evaluation were applied to each study by a single reviewer.

Data Synthesis.—Forty-four eligible studies were identified. Eleven studies used implicit criteria for inappropriate laboratory utilization and contained small numbers of patients or physicians. Most did not adequately assess the reliability of the implicit criteria. Thirty-three studies used explicit criteria based on the appropriateness of test choice, frequency, and timing, as well as the probability of a positive result. There were large variations in the estimates of inappropriate laboratory use (4.5%-95%). Evidence supporting the explicit criteria was frequently weak by the standards suggested for therapeutic maneuvers, but was nonetheless compelling based on principles of physiology, pharmacology, and probability.

Conclusions.—Many studies identify inappropriate laboratory use based on implicit or explicit criteria that do not meet methodological standards suggested for audits of therapeutic maneuvers. Researchers should develop alternative evidentiary standards for measuring appropriateness of laboratory test use.

THE APPROPRIATE utilization of laboratory tests is necessary for optimal patient care. The utilization of diagnostic laboratories has increased over the last several decades in many medical jurisdictions around the world. In Ontario, Canada’s largest province, the total number of tests performed annually has increased by more than 130% between 1976 and 1992. The cost of increased laboratory utilization has been considerable, rising from an annual expenditure (in inflation-adjusted dollars) of Can $33.29 per person in 1979 to Can $103.36 in 1992. This increased laboratory use is appropriate if it allows accurate diagnoses to be made, ideal therapy to be identified and monitored, and accurate prognoses to be established.

Perhaps because of this increased use, laboratory utilization has been analyzed extensively with studies focused primarily on physician-ordering practices. Many studies have suggested that inappropriate test ordering is a primary reason for increased laboratory use. If inappropriate laboratory utilization does exist, it must be corrected for several reasons. First, inappropriate testing not only causes unnecessary patient discomfort, it also increases the likelihood of generating false positive results, which cause unnecessary worry and further investigation. Second, laboratory tests are examples of “little-ticket” health care technologies, which, because they are commonly used, cost the health care system large amounts of money. Finally, inappropriate laboratory utilization may be associated with other inefficiencies in health care delivery. Identifying inadequacies in the use of laboratory services may disclose problems in other areas of health care.

For editorial comment see p 565.

These considerations highlight the importance of determining when inappropriate laboratory utilization occurs and raise the question, “Do we know what inappropriate laboratory utilization is?” Evidence-based users’ guides to assessing the medical literature have focused on laboratory test performance characteristics, but have not directly addressed test utilization. With this systematic review, we set out to identify published studies that measured inappropriate laboratory utilization. We used guidelines for appraising laboratory utilization reviews to determine the validity of each study’s criteria for assessing appropriateness and the application of the criteria. However, since those methodological guidelines are directed primarily toward therapeutic utilization reviews, we determined the studies’ compliance with the guidelines as well as the applicability of the guidelines to diagnostic utilization reviews. Finally, we have identified directions for future research in this area.

METHODS

The MEDLINE database was searched from 1966 to September 1997 by crossing several subject headings (explode laboratories, diagnostic services, diagnostic services—routine, explode diagnosis–laboratory, or explode quality assurance–health care) with several topic headings (attitude of health personnel, physicians’ practice pat-
tions, explode guidelines, or utilization review) or text words (unnecessary, duplication, efficiency, inappropriate, overutilization, underutilization, quality control, quality assurance, guidelines, utilization, or utilization review). Citations were limited to human subjects but no language restriction was used. A very inclusive search strategy was used because laboratory utilization is a broad topic applying to almost all areas of medicine. Each citation’s title and abstract was reviewed and all studies potentially concerned with efficient laboratory use or methods to decrease laboratory use were retrieved.

Studies were included if they specified criteria for inappropriate laboratory utilization and actually used them in an audit. Studies were excluded if they assessed radiological or pathological tests, dealt exclusively with laboratory quality control issues (eg, culturing urine from leaky specimen containers), or assessed the appropriateness of the test’s urgency (eg, was a “stat” order appropriate?). Studies that developed criteria for inappropriate laboratory utilization were excluded if the criteria were not applied to a second group of patients. Finally, we excluded studies that determined the appropriateness of screening tests (eg, preoperative screening batteries, preoperative bleeding times, blood cross-matching). We reasoned that, instead of determining the appropriateness of a test for a particular patient in a particular situation, studies addressing the appropriateness of screening tests determine whether a particular test is appropriate for a group of patients. Determining the appropriateness of screening tests has been addressed in several publications.

These inclusion and exclusion criteria were applied to the references of all articles as well as citations retrieved from EMBASE and HEALTHSTAR databases with a search strategy similar to that used in MEDLINE. Finally, hand-searching was used to screen all articles in the Toward Optimal Laboratory Use text and retrieved. Seventy-one of these studies were determined.

Four articles were excluded because they dealt exclusively with serum sampling time without assessing the appropriateness of ordering the test itself.

Forty-four articles meeting our inclusion criteria were identified. One study presented explicit criteria for both serum calcium and lactate dehydrogenase, which are discussed separately. Criteria for inappropriate use were classified as implicit for 11 studies and explicit for 33 studies. Studies were published between 1965 and 1995. Most studies (70%) originated from the United States, with the remainder from the United Kingdom, the Netherlands, Australia, Canada, and 1 each from Egypt and Thailand.

Studies Using Implicit Appropriateness Criteria

The 11 studies that used implicit criteria for appropriateness are listed in Table 1. All of these studies were conducted in teaching institutions and, with 2 exceptions, had appropriate laboratory use in the study title or abstract without providing a definition of these terms in the manuscript. Twelve articles from the MEDLINE search were appropriate for study inclusion. Thirty-two other articles were identified by searching EMBASE and HEALTHSTAR databases, reviewing references of all retrieved articles, and hand-searching journal series and textbooks on laboratory testing. Thirty of these were in the MEDLINE database but were not captured by our initial search strategy or were overlooked during the initial citation screening. Four articles were excluded because they dealt exclusively with serum sampling time without assessing the appropriateness of ordering the test itself.

Figure 1.—Result of search strategy. The asterisk indicates that 1 study provided criteria for 2 separate tests.

![Figure 1](https://example.com/figure1.png)
tions, studied laboratory utilization for patients in hospitals. The number of patients assessed in the studies ranged from 25 to 555. Four studies cited the number of physicians ordering tests, which ranged from 3 to 18. Investigations that were assessed in the studies included basic biochemistry tests alone, with hematology tests, or with hematology and microbiology tests. It was unclear from Dickinson’s article what investigations were assessed.

Validity of the Criteria

Each study’s criteria were based partly on whether the test helped diagnose or treat patients. The criteria were subjective and relied on interpretation by the reviewer. Several studies added qualifiers to the criteria, apparently to increase the reliability of the reviews; however, the qualifiers themselves left latitude for interpretation. Since the criteria were broad, it is not surprising that none were based on specific published evidence. Two studies used a consensus conference to develop the broad criteria for implicit review, but an explicit group process was not reported. None of the studies correlated compliance with the criteria and favorable patient outcomes. However, this methodological standard had limited applicability because the studies were assessing the ordering of basic laboratory tests in isolation.

Criteria Application

Most studies used medical record review or a summary of the medical record review to apply the criteria. All audit criteria required the chart reviewer to identify each patient’s symptoms or diagnoses to determine test appropriateness. Most studies used a single reviewer for medical record abstraction. Two studies clarified uncertainties with the opinion of a second clinician or interviews with the requesting physician. In the study by Bold and Corrin, the senior surgical residents from each team determined appropriateness of all tests, presumably based on their memory of each case.

Only 2 studies directly compared the classification of laboratory appropriateness made by different reviewers. In 1 study, 42.8% of 339 laboratory tests were classified as inappropriate by a panel of 3 physicians compared with only 26.5% by a single pathologist. Bloomgarden and Sidel showed that agreement between 2 physician groups was significantly beyond chance (\(\kappa = 0.63, P < .001\)). Dickinson determined that intrarater agreement was 97% when the same reviewer reabstracted the charts after 3 months. One study allowed test appropriateness to be classified as uncertain but no sensitivity analysis (ie, the effect of reclassifying uncertain categorizations on the study’s conclusion) was reported on these data.

The proportion of inappropriate tests varied among studies. Dixon and Laszlo and Sandler had the highest estimates, citing more than 90% of tests as inappropriate. These studies arguably also had the most stringent assessment criteria (Table 1). Data provided by Young did not allow for an estimate of test inappropriateness by individual test type. The rest of the studies concluded that between 10% and 50% of tests were inappropriate. None of the studies reported confidence intervals for the prevalence of inappropriate tests.

Studies Using Explicit Appropriateness Criteria

These studies are presented in Tables 2, 3, and 4 and are categorized according to the laboratory test studied. In contrast with the implicit group, these studies concentrated on single laboratory...
tests. With several exceptions, most studies were conducted on patients hospitalized within teaching centers. The number of patients varied among studies (range, 15-2897) but they generally contained more than the implicit minimum of 15. Studies (range, 15-2897) but they generally contained more than the implicit minimum of 15. In most studies, the sampling frames of these studies were sometimes very narrow and often included patients only if they had testing at a particular frequency, with a completed form, or for a particular indication.1

Validity of the Criteria

Although the explicit criteria for inappropriate use vary (Tables 2, 3, and 4), several themes are common to them. Criteria were based on testing frequency, with a completed form, or for a particular indication.1

Most criteria were based on published literature or local consensus. Several studies* justified the criteria used by citing a varying number of published articles on the topic. Schoenenberger et al109 documented the methods used for conducting a systematic review to derive their criteria. Three studies based criteria on guidelines published by the American College of Physicians108,115 and other specialty societies.111 Local expert opinion was specified for criteria derivation in some studies106,107,110,112 but the methods used to elicit these opinions were not delineated. Rhyme and Gehlbach109 based their criteria for thy-

<table>
<thead>
<tr>
<th>Study</th>
<th>Setting</th>
<th>Inappropriateness Criteria</th>
<th>Inappropriate Tests, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eisenberg et al107 1977</td>
<td>15 teaching-hospital patients having ≥3 calcium tests per week</td>
<td>&gt;3 calcium tests without pancreatitis, therapy for hypercalcemia or hypocalcemia, parathyroidectomy or thyroidec- tomy, &gt;4 transfusions in 2 d, hyperalimentation, or at least 1 abnormal test result</td>
<td>9/15 (60)</td>
</tr>
<tr>
<td>Branger et al110 1995</td>
<td>1500 patients from ambulatory care clinic</td>
<td>Biochemical or hematological test repeated within 5 d of previous test</td>
<td>6748/41 655 (16)</td>
</tr>
<tr>
<td>Bartlett et al110 1994</td>
<td>214 teaching-hospital patients having urine culture requested using special form</td>
<td>In patients without UTI symptoms: urine culture without abnormal urinalysis, fever with no source, active urinary sediment, or follow-up for UTI treatment</td>
<td>11/214 (5)</td>
</tr>
<tr>
<td>Albright et al111 1991</td>
<td>552 patients having VDRL testing on cerebrospinal fluid</td>
<td>Cerebrospinal fluid for VDRL test in patient with negative (or unknown) serologic tests for syphilis (serum rapid plasma reagin or fluorescent treponemal antibody absorption)</td>
<td>522/552 (94.5)</td>
</tr>
<tr>
<td>Gross et al112 1988</td>
<td>315 patients having blood culture in teaching hospital</td>
<td>Only 1 or &gt;4 blood cultures for a single patient§</td>
<td>151/315 (47.9)</td>
</tr>
<tr>
<td>Bowman et al113 1992</td>
<td>505 ambulatory or hospitalized patients having stool culture</td>
<td>Culture for enteric pathogens or C difficile culture without stool positive for white blood cells, watery stool, specific hospitalization length</td>
<td>or previous antibiotic use¶</td>
</tr>
<tr>
<td>Morris et al114 1992</td>
<td>All ambulatory and hospitalized patients having stool culture</td>
<td>&gt;1 stool for ova and parasites (unless first sample was negative and clinical suspicion was high) or any stool for ova and parasites for patients developing diarrhea &gt;3 d following admission</td>
<td>265/811 (32.7)</td>
</tr>
</tbody>
</table>
A laboratory test is inappropriate if it is ordered on the wrong patient, for the wrong indication, at the wrong time, or without the patient’s knowledge. To determine the frequency of inappropriate laboratory utilization, a primary step is to develop explicit criteria that state in precise terms which tests and test situations are inappropriate. A secondary step is to apply the criteria to laboratory data collected in a way that is both feasible and precise. Finally, the test results are analyzed to determine the proportion of test orders that are inappropriate.

### Criteria Application

Since explicit criteria were used, these studies had more options for data collection necessary for criteria application. Many used data from laboratory databases or from comprehensive surveys of patients investigated at their hospital. Finally, fundamentals of physiology and pharmacology were used for criteria in the remaining studies.

### Prevalence of Inappropriate Laboratory Utilization

Estimates for the proportion of inappropriate tests varied extensively (range, 4.5%-95%). Tests with the highest estimates for inappropriate utilization included prothrombin time, calcium, cerebrospinal fluid analysis for VDRL test, and antiepileptic drug monitoring. The study by Saxena et al probably overestimates inappropriate laboratory use because the criteria used for timing of repeat testing were stringent (i.e., a meaningful difference between a serial creatine kinase level measured at 10 hours vs one measured at 12 hours). Estimates for inappropriate laboratory utilization ranged from 5% to 50% in the other studies. It is plausible that criteria relying on data from requisition forms would have higher inappropriateness rates because physician documentation on the requisition could be very incomplete. However, studies using this method did not find a higher prevalence of inappropriateness.

### COMMENT

Appropriate laboratory utilization is a cornerstone of optimal medical practice. Patients benefit from accurate diagnoses, proper therapeutic monitoring, and precise prognostications, all of which result from the use of sensible diagnostic technologies. Inappropriate laboratory utilization can not only harm patients, it is also expensive. The authors of all studies cited in this review deserve commendation for trying to measure its prevalence. A broad range of criteria were used for different laboratory tests in distinct clinical situations. The validity of the criteria and the reliability of their application varied among the studies. The prevalence of inappropriate laboratory use also varied extensively in its magnitude.

Some studies used implicit as opposed to explicit criteria. Implicit criteria are versatile because a single definition can be applied to a broad range of tests. They allow more comprehensive evidence for inappropriate laboratory utilization because the reviewer can consider multiple components of a patient’s situation when determining the appropriateness of a laboratory test. However, there are many drawbacks to using implicit appropriateness criteria, including interviewer variability in their interpretation and application. Use of explicit criteria usually demands detailed review of medical records. If physicians do not completely document their clinical reasoning or the patient’s findings, laboratory tests may be judged inappropriate. The review could then become more of an assessment of clinical documentation than of laboratory utilization. Also, medical record review is time-consuming and expensive. This will decrease the number of diagnostic episodes that can be assessed and therefore decrease the scope and precision of any assessment of laboratory utilization. With explicit criteria, it is frequently possible to collect the necessary data from databases or data requisition forms, thus avoiding many of the problems associated with retrospective medical record review. Finally, explicit criteria make it possible to determine appropriateness criteria using prospectively collected data (e.g., from a data requisition form). Prospectively collected data help avoid bias due to omission of physician documentation.

Two studies classified tests as inappropriate if therapy did not change as a result of test ordering. This criterion is problematic because the reviewer determining test appropriateness could not know whether or how therapy could have changed if the test result was different. Also, it is unreasonable to expect a laboratory test to single-handedly alter therapy because test results are only one of several parameters used in medical
### Table 4.—Studies Using Explicit Criteria for Inappropriate Laboratory Use of Therapeutic Drug Monitoring

<table>
<thead>
<tr>
<th>Study</th>
<th>Setting</th>
<th>Inappropriate Criteria</th>
<th>Inappropriate Tests, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goldberg and Abbott, 1974†</td>
<td>27 teaching-hospital patients</td>
<td>Digoxin level without arrhythmias, nausea or vomiting, altered mental status, worsening congestive heart failure despite digoxin use, congestive heart failure with chronic renal failure, question of digoxin use</td>
<td>8/27 (29.6)</td>
</tr>
<tr>
<td>Clague et al, 1983</td>
<td>166 teaching-hospital patients having digoxin assay</td>
<td>Digoxin level without possible noncompliance, toxic effects, or interaction with other drugs; inadequate response to digoxin; changing metabolic factors; digoxin used for prevention of supraventricular tachycardia</td>
<td>165/200 (82.5); 11/200 (5.5) had inappropriate timing*</td>
</tr>
<tr>
<td>Slaughter et al, 1978</td>
<td>145 teaching-hospital patients having digoxin level measured</td>
<td>Digoxin level without possible noncompliance or toxic effects or worsening congestive heart failure on digoxin</td>
<td>71/145 (49); 6/144 (4) had inappropriate timing*</td>
</tr>
<tr>
<td>Greenlaw et al, 1980</td>
<td>61 non–teaching-hospital patients having digoxin level measured</td>
<td>Digoxin level without possible noncompliance or toxic effects or worsening congestive heart failure on digoxin</td>
<td>22/120 (18.3)</td>
</tr>
<tr>
<td>Ives et al, 1984</td>
<td>20 teaching-clinic patients having digoxin level measured</td>
<td>Digoxin level without possible noncompliance or toxic effects or worsening congestive heart failure while receiving digoxin excepting baseline levels, levels done when interacting drugs started, or worsening renal function</td>
<td>8/55 (14.5)</td>
</tr>
<tr>
<td>Pearce and Day, 1990</td>
<td>152 teaching-hospital patients having drug level measured</td>
<td>Drug level &lt;5 d following previous measurement</td>
<td>24/153 (15.7)</td>
</tr>
<tr>
<td>Bussey and Hoffman, 1983</td>
<td>244 non–teaching-hospital patients having drug level measured</td>
<td>Drug level &lt;4 half-lives after previous level or change in dose (without toxic symptoms, decreased renal function, or new drugs)</td>
<td>88/244 (36.1)</td>
</tr>
<tr>
<td>Pengis and Martin, 1984</td>
<td>187 teaching-hospital patients</td>
<td>Drug level without possible toxic effects or drug interaction, treatment failure, or compliance assessment</td>
<td>3/64 (4.7)</td>
</tr>
<tr>
<td>Levin et al, 1981†</td>
<td>384 teaching-hospital patients</td>
<td>Digoxin, theophylline, gentamicin, or phenytoin level without subtherapeutic response or suspected toxic effects‡</td>
<td>89/153 (58.1)</td>
</tr>
<tr>
<td>Schoenengerber et al, 1995</td>
<td>330 teaching-hospital patients having antiepileptic drug levels measured</td>
<td>Antiepileptic drug level without seizure &lt;6 h previously, possible toxic effects, recent start of antiepileptic drug, change in dose, addition of second antiepileptic drug, or change in liver or gastrointestinal tract function§</td>
<td>624/855 (73), 140/231 (61) had inappropriate timing†</td>
</tr>
<tr>
<td>Wing and Duff, 1989†</td>
<td>125 teaching-hospital patients having dilantin level measured</td>
<td>Nonbaseline phenytoin level without subtherapeutic response or suspected drowsiness</td>
<td>27/125 (22), 68/98 (69) had inappropriate timing</td>
</tr>
<tr>
<td>Levine et al, 1988</td>
<td>80 teaching-hospital patients who received phenytoin during hospitalization</td>
<td>Similar to Schoenengerber et al, except that patient criteria were also used to identify patients who inappropriately did not have phenytoin level measured</td>
<td>22/80 patients (27.5) met criteria at least once, 83/113 (73.5) phenytoin levels were inappropriate</td>
</tr>
<tr>
<td>Sargent et al, 1985</td>
<td>33 teaching-hospital patients having theophylline level measured</td>
<td>Theophylline level without subtherapeutic response, suspected noncompliance or toxic effects, or theophylline steady state¶</td>
<td>48/102 (47)</td>
</tr>
<tr>
<td>Guernsey et al, 1984</td>
<td>121 teaching-hospital patients having theophylline level measured</td>
<td>Above criteria applied to nonbaseline levels and included follow-up of previously toxic levels¶</td>
<td>79/352 (23), 136/491 (27) had inappropriate timing</td>
</tr>
<tr>
<td>Greelelaw et al, 1979</td>
<td>53 community-hospital patients having gentamicin level measured</td>
<td>Gentamicin level without possible toxic effects, severe infection or endocarditis, renal impairment, history of aminoglycoside toxic effects, no response to treatment, or questionable pharmacodynamics</td>
<td>3/67 (4.5)</td>
</tr>
</tbody>
</table>

---

*Level drawn less than 6 to 8 hours after last dose.
†Intervention study; only data prior to intervention provided.
‡Drug sampling had to be at steady state (specified for each antiepileptic drug).
§Less than 4 half-lives following change in drug dose.
¶Subcriteria were defined for each criterion.

Many studies that used explicit criteria did not comply with guidelines for establishing the validity of utilization review criteria. Poor study design or poor reporting could explain noncompliance with these guidelines. However, it is also arguable that these guidelines are not as readily applicable to diagnostic as to therapeutic utilization reviews. Diagnostic technologies are often not tested with the same evaluative rigor as many therapeutic technologies. This is understandable in light of the lower cost of many diagnostic tests, their use as only 1 element in clinical decision making, and the attendant uncertainties in their link to patients’ outcomes. As a result, there is a paucity of high-level evidence (such as randomized controlled trials) assessing diagnostic technology on which audit criteria can be based. This could explain why less than half of the studies cited evidence on which their audit criteria were based. Morris et al suggested for this by conducting a cohort study at their own institution on which their appropriateness criteria were based.
Compliance with utilization review guidelines is problematic for audits of diagnostic technologies for other reasons. For criteria development, expert opinion should be tapped using an explicit, systematic, and reliable process when published evidence is lacking or needs supplementation.27 This could be more difficult in the diagnostic arena because test appropriateness is highly dependent on the pretest probability of the condition being sought. The number of clinical situations in which pretest probabilities have been rigorously determined is relatively small. Also, the subjective determination of these probabilities varies widely among physicians. These subjective issues may explain why only a few studies tapped expert opinion to derive criteria and why none used systematic methods in doing so.125

Finally, the validity of the criteria is enhanced when their compliance is associated with improved patient outcomes.27 For example, Kravitz et al126 showed that patients meeting panel criteria for revascularization who did not undergo such procedures fared significantly worse than those who did. In contrast, it is difficult for diagnostic technologies to conform with this criterion because tests are often used to determine whether the primary prognosticator (ie, the disease) is present. Test use by itself may therefore not be associated with patient outcomes. To compensate for this, several studies tried to validate their criteria by associating criteria compliance with surrogate outcomes such as costs or the prevalence of abnormal test results.82,85 Other investigators dealt with the difficulty of complying with these accepted methods of establishing valid audit criteria by basing them on physiological principles and pharmacological tenets. In our view, these are defensible alternatives, since pharmacokinetic and pharmacological criteria would be based on known pharmacodynamic tenets, such as those by Pearce and Day84 and Bussey and Hoffman.103 Clinical probability-based criteria would identify scenarios where a test may be useless if its result is significantly worse than the pretest probability. This would occur when a test's operating characteristics (eg, positive or negative likelihood ratios) are too weak to be useful. This would happen when the operating characteristics of a test are insufficient to alter the probability of a patient's disease state. For example, Reid et al26 and would compare tests based on their operating characteristics such as sensitivity, specificity, and likelihood ratios. The latter criterion, determining a test's appropriateness without another test for comparison, could be based on physiological, pharmacological, or probabilistic data. Physiological criteria would be based on our knowledge of disease processes, such as the criteria of Albright et al for VDRL testing in cerebrospinal fluid.98 Pharmacological criteria would be based on known pharmacodynamic tenets, such as those by Pearce and Day84 and Bussey and Hoffman.103 Clinical probability-based criteria would identify scenarios in which test results would almost never be useful. This would occur when a test's operating characteristics (eg, positive or negative likelihood ratios) are too weak to significantly alter posttest probabilities. Finally, a test may be useless if its result is highly predictable. Cohort studies could be used to identify these situations.† We hope that our model will serve as a framework for the development of appropriateness criteria for future studies and that these studies will concentrate first on establishing valid criteria for determining test appropriateness before it is measured.
Where should we go from here? Identifying inappropriate laboratory use is important and requires valid methods. Promising criteria for inappropriate laboratory use need to be validated and then applied to divergent populations of patients and physicians using methodologically sound studies. Further research is necessary to measure test operating characteristics and methods of determining pretest probabilities accurately. If inappropriate laboratory utilization is identified, methods to rectify the problem must be developed, tested, and implemented. Although appropriate laboratory testing does not necessarily translate into appropriate utilization of test results, it is an important step toward optimal care of the patient and better performance of the health care system.

Dr van Walraven is an R. Samuel McLaughlin Foundation research fellow at the Institute for Clinical Evaluative Sciences, North York, Ontario.

References


53. Tierney WM, Miller ME, McDonald CJ. The effect of test ordering of informing physicians of the


