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 explicit (objective criteria for inappropriate utilization not provided or very broad) or implicit. Guidelines for evaluation were applied to each study by a single reviewer.

Data Synthesis.—Forty-four eligible studies were identified. Eleven studies used explicit 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.¹ ² In Ontario, Canada’s largest province, the total number of tests performed annually has increased by more than 130% between 1976 and 1992.³ 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 1992.³ 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.⁶ ¹¹ Many studies have suggested that inappropriate test ordering is a primary reason for increased laboratory use.¹² ¹⁸ 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.²¹ Second, laboratory tests are examples of “little-ticket” health care technologies,²² which, because they are commonly used, cost the health care system large amounts of money.²² ²³ 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,²² 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 utilization reviews²⁷ 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-
In the past, inappropriate laboratory utilization has been addressed in several publications. Determining the appropriateness of screening tests has been addressed in several 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 series from JAMA and all references in 3 commonly cited textbooks on laboratory test use.

Descriptive data that were abstracted from each study included country of publication, study setting, laboratory tests assessed, and the number of patients and physicians included in the study. All articles were categorized based on whether the criteria for inappropriate laboratory use were implicit (objective criteria for inappropriate use not provided or criteria very broad, leaving much latitude for interpretation by the reviewer) or explicit. Guidelines for the evaluation of utilization reviews were applied to each study by a single reviewer. These guidelines help determine whether the criteria are valid and whether they are applied appropriately. Validity of the criteria was assessed based on 3 components: the quality of the evidence on which the criteria are based; the methods used to tap expert opinion; and whether compliance with the criteria is associated with better patient outcomes. The strength of criteria application was based on 2 components: the reliability of criteria use and whether the impact of uncertainty in their application was measured. Estimates for the prevalence of inappropriate laboratory utilization were documented. When interventions aimed at altering laboratory use were studied, only preintervention measurements of inappropriate laboratory use were determined.

RESULTS

Figure 1 illustrates results of the search strategy. Eighty-three articles potentially providing criteria for laboratory inappropriateness were identified and retrieved. Seventy-one articles were excluded because they provided no appropriateness criteria (n = 24); were review articles (n = 19); or studied screening tests (n = 10); laboratory quality control (n = 6), the appropriateness of laboratory test urgency (n = 3), or other topics (n = 9). Of the 24 studies without a definition of inappropriate utilization, 4 derived criteria but did not apply them to a separate group of patients, whereas the other studies measured laboratory use without an appropriateness assessment. Several of these studies referred to “inappropriate” and “needless” 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 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.

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 excep-

©1998 American Medical Association. All rights reserved.
Table 1.—Studies Using Implicit Criteria for Inappropriate Laboratory Use

<table>
<thead>
<tr>
<th>Study</th>
<th>No. of Patients</th>
<th>Inappropriateness Criteria</th>
<th>Inappropriate Tests, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rafieh and El-Tobgui, 94 1995</td>
<td>66 admissions to single public hospital</td>
<td>Test not relevant to patient's symptoms and provisional diagnosis</td>
<td>83/264 (31.4)</td>
</tr>
<tr>
<td>McConnell et al, 89 1982</td>
<td>25 nonrandomly chosen patients admitted to teaching hospital</td>
<td>Test was unnecessary for suspected medical purpose for which it was ordered*</td>
<td>554/1651 (33.6)</td>
</tr>
<tr>
<td>Ruangkanchanase et al, 93*</td>
<td>All patients seen in pediatric clinic during 5 mo (total number not given)</td>
<td>Staff rated appropriateness of each test from 1 to 10; tests with mean scores &lt;5 were inappropriate</td>
<td>Hematology, 29%† Chemistry, 33%‡ Microbiology, 50%‡</td>
</tr>
<tr>
<td>Young, 80 1980*</td>
<td>287 patients admitted to medicine ward in teaching hospital</td>
<td>Test not justified by patient's condition or results of other tests</td>
<td>103/155 (66) patients had at least 1 inappropriate test</td>
</tr>
<tr>
<td>Bolling and Sidel, 87 1980</td>
<td>476 patients nonrandomly sampled from all visits to teaching-hospital emergency department</td>
<td>3 reviewers in 2 groups classified each test as necessary for decisions in emergency relevant to patient's problem, required for medicolegal reasons, or inappropriate</td>
<td>Hematology, 46/177 (26.0) Biochemistry, 114/222 (51.4) Microbiology, 13/28 (46.4)</td>
</tr>
<tr>
<td>Gortmaker et al, 88* 1985</td>
<td>100 patients randomly sampled from physicians ordering extreme number of tests in hospital</td>
<td>Test was not admission or preparative test; was not follow-up to abnormal test result; did not add or exclude diagnosis, change treatment, or monitor treatment;‡ was not associated with specific note describing impact</td>
<td>34% tests for high laboratory-use physicians, 25% for low laboratory-use physicians</td>
</tr>
<tr>
<td>Dixon and Laszlo, 74 1974</td>
<td>25 inpatients</td>
<td>Test did not generate order for treatment or need for other care; was not considered in planning for subsequent evaluation was not considered in constructive assessment in progress note was not appropriately repeated test (not defined); (if normal) did not generate comment that it adequately ruled out diagnostic consideration§</td>
<td>95% of tests</td>
</tr>
<tr>
<td>Baigelman et al, 85 1985</td>
<td>145 patients admitted to intensive care unit</td>
<td>Ordering more than 1 set of electrolytes within 24 h when it was “not necessary”</td>
<td>58/140 (41.7) of repeated electrolytes</td>
</tr>
<tr>
<td>Dickinson, 87 1987*</td>
<td>30 patients with abdominal pain in teaching-hospital emergency department</td>
<td>Testing was “excessive”</td>
<td>59/162 (36.4)</td>
</tr>
<tr>
<td>Sandler, 1984</td>
<td>555 medical patients in teaching-hospital emergency department</td>
<td>Test result was not abnormal and did not lead to diagnosis not suspected from initial evaluation or change in therapy</td>
<td>2133/2372 (90)</td>
</tr>
<tr>
<td>Bold and Corrin, 85 1965</td>
<td>344 laboratory tests on surgical patients in teaching hospital</td>
<td>Test was not diagnostic or supportive of clinical diagnosis, a useful negative, or was not used in monitoring or treatment of disease progress</td>
<td>37/344 (10.8)</td>
</tr>
</tbody>
</table>

*Intervention study; only data prior to intervention provided. †Data are abstracted from graphical representation. ‡Frequency limits given in article. §These criteria were also used by Myers et al.¶ Whole numbers calculated from data in article.

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, 30 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 diagnosis or treat patients (Table 1). As is the norm with implicit review, 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 839 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 ($\chi = 0.63, P<.001$). Dickinson determined that intrarater agreement was 97% when the same reviewer reread 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 inappropriate 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
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 al 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 Physicians, and other specialty societies. Local expert opinion was specified for criteria derivation in some studies, but the methods used to elicit these opinions were not delineated. Rhino and Gehlbach based their criteria for thy-

Table 2.—Studies Using Explicit Criteria for Inappropriate Laboratory Use for General and Microbiological Tests

<table>
<thead>
<tr>
<th>Study</th>
<th>Setting</th>
<th>Inappropriateness Criteria</th>
<th>Inappropriate Tests, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valenstein et al, 1988</td>
<td>198 teaching-hospital patients having &gt;1 biochemical profile per admission</td>
<td>&gt;1 biochemical profile per day</td>
<td>254/936 (27)</td>
</tr>
<tr>
<td>Branger et al, 1995</td>
<td>1500 patients from ambulatory care clinic</td>
<td>Biochemical or hematomatological test repeated within 5 d of previous test</td>
<td>674/41 655 (16)</td>
</tr>
<tr>
<td>Studnicki et al, 1993</td>
<td>2897 teaching-hospital patients</td>
<td>More than prespecified maximum number of tests per day and week; prespecified maximum number was greater if any test result was abnormal</td>
<td>2133/20 124 (10.6%†)</td>
</tr>
<tr>
<td>Erban et al, 1989</td>
<td>375 teaching-hospital patients</td>
<td>Prothrombin time or partial thromboplastin time without liver disease, malabsorption or malnutrition, active bleeding, history of abnormal bleeding, use of anticoagulants, possible deep vein thrombosis, pulmonary embolism, or disseminated intravascular coagulation</td>
<td>210/300 (70)</td>
</tr>
<tr>
<td>Eisenberg et al, 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 thyroidectomy, &gt;4 transfusions in 2 d, hypercalciuria, or at least 1 abnormal test result</td>
<td>9/15 (60)</td>
</tr>
<tr>
<td>Kroenke et al, 1987</td>
<td>All admissions to medical specialties in teaching hospital</td>
<td>Urinalysis without UTI symptoms, increased urea or creatinine, fever or sepsis, decreased albumin or edema, abdominal or back pain, volume depletion, abnormal sodium, metabolic acidosis, pregnancy or preoperative assessment, chronic illness, or nephrototoxic drugs</td>
<td>Urinalysis: 669/1341 (49.9)</td>
</tr>
<tr>
<td>Bartlett et al, 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, or fever within 2 d of surgery. In patients with UTI symptoms: urine culture with previous positive culture &lt;3 d previously or &lt;2 d after treatment started or completed</td>
<td>11/214 (5)</td>
</tr>
<tr>
<td>Albright et al, 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>Schiffman et al, 1991</td>
<td>544 patients having blood culture in 3 teaching hospitals</td>
<td>Only 1 blood culture for a single patient</td>
<td>158/544 (29)</td>
</tr>
<tr>
<td>Gross et al, 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 al, 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, or previous antibiotic use¶</td>
<td>84/505 (16.3%) for C difficile, 251/505 (49.7%) for enteric pathogens#</td>
</tr>
<tr>
<td>Morris et al, 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>

*Intervention study; only data prior to intervention provided.
†Data are abstracted from graphical representation.
‡UTI indicates urinary tract infection.
§Other criteria for appropriateness given but not measured.
¶Hospitalization length of <4 days for enteric pathogens, >4 days for Clostridium difficile.
#Clostridium difficile culture criteria only.
†Note that laboratory protocol dictated the performance of these laboratory tests on all specimens.

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 group. Finally, 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.

Validity of the Criteria

Although the explicit criteria for inappropriateness appear diverse (Tables 2, 3, and 4), several themes are common to them. Criteria were based on testing frequency, timing of the test in relation to previous medications and tests, test choice compared with possible alternatives, clinical indications for the test, or probability that a test result was abnormal. Two studies defined a single blood culture as inappropriate because distinguishing between a true positive blood culture and a contaminant can be difficult when only 1 culture is done. Flynn et al 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 Physicians and other specialty societies. Local expert opinion was specified for criteria derivation in some studies, but the methods used to elicit these opinions were not delineated. Rhino and Gehlbach based their criteria for thy-
Table 3.—Studies Using Explicit Criteria for Inappropriate Laboratory Use of Cardiac Enzymes and Thyroid Tests

<table>
<thead>
<tr>
<th>Study</th>
<th>Setting</th>
<th>Cardiac Enzymes</th>
<th>Thyroid Function</th>
<th>Inappropriate Tests, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eisenberg et al14, 77 1977</td>
<td>69 teaching-hospital patients having ≥3 lactate dehydrogenase measurements in a week</td>
<td>&gt;1 lactate dehydrogenase measurement per week without possible myocardial infarction or hemolysis or abnormal value; &gt;3 lactate dehydrogenase measurements per week without recurrent chest pain</td>
<td>Thyrotropin measured for symptoms having a total score of 0</td>
<td>35/69 (51)</td>
</tr>
<tr>
<td>Saxena et al, 106 1993</td>
<td>774 hospitalized patients having creatine kinase–MB measured</td>
<td>Creatine kinase–MB measured &lt;3 times and at times other than admission, 12 h later, and 24 h later</td>
<td>T3 resin uptake and total thyroxine tests ordered for symptoms having “low indication” score</td>
<td>No creatine kinase measured at 12 h: 451/480 (94)</td>
</tr>
<tr>
<td>Saxena et al, 104 1993</td>
<td>774 hospitalized patients having creatine kinase–MB measured</td>
<td>Creatine kinase–MB measured &lt;3 times and at times other than admission, 12 h later, and 24 h later</td>
<td>T3 resin uptake and total thyroxine tests ordered for symptoms having “low indication” score</td>
<td>No creatine kinase measured at 24 h: 281/293 (96)</td>
</tr>
<tr>
<td>Rhyne and Gehlhab, 100 1979</td>
<td>90 patients having triiodothyronine resin uptake and total thyroxine measured at ambulatory clinic</td>
<td>Any test other than thyrotropin alone for thyroid function testing</td>
<td>50/90 (55)</td>
<td></td>
</tr>
<tr>
<td>Schoedel et al, 105 1991†</td>
<td>1425 ambulatory patients with thyroid function testing</td>
<td>Any test other than those specified for thyroid screening, detection of hypothyroidism and hyperthyroidism, and monitoring of therapy</td>
<td>67/181 (37)</td>
<td></td>
</tr>
<tr>
<td>Finn et al.114 1988</td>
<td>181 teaching-hospital patients having thyroid testing</td>
<td>Any test other than those specified for thyroid screening, detection of hypothyroidism and hyperthyroidism, and monitoring of therapy</td>
<td>67/181 (37)</td>
<td></td>
</tr>
<tr>
<td>Boon-Falleur et al.13 1995</td>
<td>794 ambulatory patients with thyrotropin measured</td>
<td>Thyrotropin measured for symptoms having a total score of 0</td>
<td>138/794 (17.4)</td>
<td></td>
</tr>
</tbody>
</table>

*Intervention study: only data prior to intervention provided.

Laboratory testing on a survey involving physicians in training. Boon-Falleur et al13 used a probability analysis associating symptoms with thyroid dysfunction to create criteria. Morris et al20 based their criteria on detailed data from a cohort of patients investigated at their hospital. Finally, fundamentals of physiology36 and pharmacology were used for criteria in the remaining studies.

**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 requisition forms completed by the ordering physician.100,106,108,112,113 The remainder of the studies required medical record review. The data source for 2 studies was uncertain.103,105 Several studies108,109,112,115 determined abstraction reliability with a second medical record reviewer but only Schoenenger et al105 measured agreement beyond chance (κ = 0.61, P < .001). Two studies clarified data from chart review by interviewing the physician who ordered the questionable tests106,114 None of the studies associated compliance with the criteria with positive patient outcomes. Bowman et al14 and Morris et al,20 however, showed that tests that complied with the criteria were significantly more likely to have abnormal results.

**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,115 calcium,133 cerebrospinal fluid analysis for VDRL test,136 and antiepileptic drug monitoring.58,108 The study by Saxena et al106 probably overestimates inappropriate laboratory use because the criteria used for timing of repeat testing were stringent (ie, it assumes 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 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 explicit as opposed to implicit criteria. Implicit criteria are versatile because a single definition can be applied to a broad range of tests. They allow a more complete review of 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 implicit 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 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 often possible to collect the necessary data from databases or data requisitions, 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 (eg, from a data requisition form). Prospectively collected data help avoid bias due to omission of physician documentation. Two studies58,106 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 1 of several parameters used in medical decisions.
decision making. Finally, even tests that do not change therapy may provide useful information to physicians and patients.

Table 4.—Studies Using Explicit Criteria for Inappropriate Laboratory Use of Therapeutic Drug Monitoring

<table>
<thead>
<tr>
<th>Study</th>
<th>Setting</th>
<th>Inappropriateness Criteria</th>
<th>Inappropriate Tests, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goldberg and Abbott,131 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,52 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)</td>
</tr>
<tr>
<td>Slaughter et al,131 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,131 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,13 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 or levels done when interacting drugs started, or worsening renal function</td>
<td>8/55 (14.5)</td>
</tr>
<tr>
<td>Pearce and Day,84 1990</td>
<td>152 teaching-hospital patients having drug level measured</td>
<td>Drug level = 5 d following previous measurement</td>
<td>24/153 (15.7)</td>
</tr>
<tr>
<td>Bussey and Hoffman,131 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,52 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,131 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,131 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,52 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)</td>
</tr>
<tr>
<td>Levine et al,52 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,131 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,131 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 level¶</td>
<td>79/352 (23), 136/491 (27) had inappropriate timing</td>
</tr>
<tr>
<td>Flynn et al,131 1978</td>
<td>61 teaching-hospital patients receiving gentamicin</td>
<td>Single gentamicin level</td>
<td>5/45 (11)</td>
</tr>
<tr>
<td>Greelaw et al,131 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.22 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.23 This could explain why less than half of the studies cited evidence on which their audit criteria were based. Morris et al52 anticipated 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. 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 test's appropriateness125 is more difficult in the diagnostic arena because the test is infrequently performed at a prespecified interval in relation to some therapeutic intervention or physiological change in the patient’s status.

For these reasons, it is understandable that audits of laboratory test utilization infrequently comply with methodological guidelines for validation of appropriateness criteria. What, then, of compliance with guidelines for application of the criteria? As discussed herein, using chart review in a “reliable and unbiased” fashion is difficult given the limitations of physician documentation. Several studies avoided medical record review by using clinical data documented on laboratory requisitions,81,84,106,114,127 although none of these studies validated these data. Despite the potential problems with medical record review, only 4 studies used 2 or more reviewers and compared their abstraction results. None of the studies determined the effect of uncertainty on study results with sensitivity analysis. These issues are less of a concern when objective criteria are used.

Are the results of these utilization reviews generalizable? Almost all the studies took place in teaching centers on hospitalized patients and analyzed the ordering practices of a small number of physicians-in-training. This represents a very select patient and physician group. Hospitalized patients are less healthy than the ambulatory population and laboratory-requesting patterns for them may not represent those of the ambulatory physician group. Physicians in training have less experience in diagnostic strategies and may order tests less efficiently than established practitioners. Since laboratory use is dependent on both patient and physician factors, it is necessary to research other clinical settings to determine the extent of inappropriate laboratory utilization. This could be done by using the criteria that are applicable to data potentially found within laboratory and administrative databases. Certainly, allusions to extensive inappropriate laboratory use128,129 should not be made without appropriate qualifiers. This review has several limitations that should be mentioned. First, the study deliberately excluded all radiology, pathology, and screening tests. Determining how well inappropriate utilization is measured in these areas would merit future reviews. Second, quality of study design and quality of reporting can be distinct. It is possible that the reliability of some studies was compromised when criteria for definitions or comparisons among raters were edited from the manuscripts. Finally, the diverse nature of this topic makes it possible that some eligible studies were not identified with our search strategy. The observation that less than half of the studies were identified from the initial MEDLINE search increases this concern.

The studies in this review suggest a paradigm for classifying criteria that determine the appropriateness of laboratory use (Figure 2). Appropriateness criteria could be based on a direct comparison among laboratory tests or a single test in isolation. The former could be based on studies using methods suggested by Reid et al19 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.

Pharmacological criteria would be based on known pharmacodynamic tenets, such as those by Pearce and Day24 and Bussey and Hoffman. 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.27 Further research is necessary to measure test operating characteristics and methods of determining pretest probabilities accurately.13,138 When inappropriate laboratory utilization is identified, methods to rectify the problem must be developed, tested, and implemented.13,134 Although appropriate laboratory testing does not necessarily translate into appropriate utilization of test results,13 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
75. Kuangkanchanaset S. Laboratory investigations utilizing a computerized departmental test order entry system. Ramathibodi Hospital, J Med Assoe Thai. 1990;76(suppl 1):194-208.