Do We Now Know What Inappropriate Laboratory Utilization Is?

An Expanded Systematic Review of Laboratory Clinical Audits

Ronald G. Hauser, MD,1 and Brian H. Shirts, MD, PhD2

From 1Yale University School of Medicine, New Haven, CT; and 2University of Washington, Seattle.

Key Words: Utilization review; Guideline adherence; Health planning guidelines; Physicians’ practice patterns; Process assessment (health care); Delivery of health care; Quality assurance, health care; Cost savings; Efficiency, organizational; Health services misuse

ABSTRACT

Objectives: Many nonpathologists and some pathologists consider utilization review essential to laboratory quality improvement, but (1) confusion surrounding the definition of “appropriate” laboratory utilization, (2) the reliance on manual chart review, and (3) a lack of leadership have contributed to its unstandardized implementation. How the solutions to these barriers have evolved since the 1950s is described.

Methods: A systematic literature review is used.

Results: Current literature largely defines inappropriate laboratory utilization as any test order in violation of a guideline produced by a government or professional society. Audits performed without manual chart review (ie, database query) have dramatically increased since the mid-1990s. Most utilization audits do not involve any author with a pathology or laboratory medicine affiliation.

Conclusions: Literature consensus defining “inappropriate” utilization combined with the adoption of database technology has removed key obstacles to utilization reviews. Leadership is needed to unify and benchmark laboratory utilization.

The occurrence of misleading laboratory test results epitomizes the need for studying and improving laboratory utilization.1,2 Such results create needless surgical scars and unwarranted prescriptions while prolonging patient symptoms.3 Additional tests may follow, continuing this perilous journey through modern healthcare; the Ulysses syndrome has entrenched itself as “a disease of our time.”4,5

False-positive results arise from testing people without disease. Multiple sources from the United States, Canada, and the United Kingdom report an inexorable rise in absolute test volume from 1963 onward that, at a doubling rate every 5 to 10 years, outpaces the most obvious confounders such as population growth or a greater proportion of older patients.6-9 Literature indicates that, not only did test results become less predictive, but a new phenomenon dubbed “disease ambiguity” emerged. “Because there is ambiguity about what constitutes disease, more advanced or more frequent testing tends to produce more abnormalities and thus more diagnoses.”9

These scholarly reports reinforce the populist conception of overtreatment, demonstrated by the alignment of a consumer advocacy group, Consumer Affairs, with the American Board of Internal Medicine’s Choosing Wisely campaign to reduce unnecessary and potentially harmful tests.10 Books such as Overdiagnosed: Making People Sick in the Pursuit of Health and Overtreated: Why Too Much Medicine Is Making Us Sicker and Poorer also address this issue.11,12 While biological characterization of disease has improved with more accurate diagnostics, the now ubiquitous presence of laboratory testing may also drive diagnostic test use in healthy people.

Who should orchestrate the necessary change: patients,10 pathologists,13 or the clinician ordering the test?14 Over the last few years, a handful of case reports from unique academic
institutions propose a team approach,\textsuperscript{15-17} which may or may not include the pathologist. The team relies on utilization audits of the clinical laboratory, a method to assess the appropriate use of laboratory services, to determine usage patterns, plan interventions, and redirect ordering behavior in a cycle of continuous quality improvement. At the present time, utilization audits have not experienced widespread adoption, and no nationwide program exists to systematically compare requests of laboratory services in the United States.

We describe the ongoing evolution of laboratory test audits in reference to three historical barriers in this systematic review. First, confusion exists surrounding the definition used to classify provider test orders as appropriate or inappropriate. Our study updates the answer provided by a previous review from 1998 which asked, “Do we [even] know what inappropriate utilization is?”\textsuperscript{48} At that time, many studies had subjective or locally defined definitions of “appropriate.” We asked if this practice has changed in the ensuing 15 years. Second, the traditional audit of laboratory services involves manual patient chart review conducted by skilled abstractors. Structured clinical data repositories, such as electronic medical records, represent an alternative. We sought to determine the extent of their adoption to perform laboratory audits. Third, efforts to benchmark the use of laboratory services lack clear leadership. We sought to determine if pathologists have demonstrated an interest by reviewing their contribution to published audits.

Materials and Methods

Study Selection

To systematically identify published audits of laboratory services, we designed a search strategy Table\textsuperscript{1}. The MeSH terms used to perform the MEDLINE search included synonyms for “audit” and “test.” The intersection of these two ideas yielded the majority of the studies. The subheading utilization in combination with the MeSH terms clinical laboratory techniques, laboratories, and routine diagnostic tests identified additional studies. The controlled vocabulary of MeSH has changed over time, preventing an exact replication of the original systematic review.\textsuperscript{8} To compensate we used a broad search strategy designed to increase the number of screened articles compared with the original review’s MEDLINE search.

Articles included in the results met a basic definition of a laboratory audit. Our definition of a laboratory test excluded imaging, tissue stains performed by an anatomic pathologist, hearing tests, tuberculosis skin tests, and pulse oximetry. Non-discretionary screening tests (eg, cholesterol, universal human immunodeficiency virus screening) were excluded. Route preoperative testing has an existing systematic review, which we did not duplicate.\textsuperscript{18} Included studies contained a criterion for appropriateness of test ordering, applied the criterion to patient data, and reported criterion adherence. To clarify how we applied these three criteria, we provide an example: one study assessed postpartum glucose testing in women with gestational diabetes (criterion for appropriate testing) among a cohort of 36,251 women (applied to patient data) and found a 15.9% compliance rate (criterion adherence).\textsuperscript{19}

Audit Description

Author Affiliations and Study Design

To record the involvement in utilization audits by laboratorians, including the pathologist and others who oversee the laboratory, we noted the description of the author within their study. A study with at least one author with any (ie, partial or full) affiliation with a laboratory met this criterion. We could have required the lead author to have full laboratory affiliation, but instead we relaxed this criterion to create the highest sensitivity for involvement by the laboratory. We recorded the nationality of study authors as well.

Table 1

### MEDLINE Search Strategy

1. exp clinical laboratory techniques/
2. exp biopsy/
3. 1 not 2
4. exp chemistry techniques, analytical/
5. exp electrochemical techniques/
6. exp fluorescent antibody technique/or exp immunoassay/or exp immunologic tests/ or exp immunoprecipitation/ or exp interferon-gamma release tests/ or exp radioimmundetection/
7. exp genetic testing/or exp molecular diagnostic techniques/or exp nucleic acid amplification techniques/
8. exp biological factors/
9. exp laboratories/
10. exp health services/ec, ut
11. exp “delivery of health care”/ec, ut
12. exp quality assurance, health care/ec, ut
13. exp “cost savings”/
14. exp efficiency, organizational/
15. exp guideline adherence/
16. exp health planning guidelines/
17. exp health services misuse/
18. exp “process assessment (health care)”/
19. exp unnecessary procedures/
20. exp “utilization review”/
21. exp physician’s practice patterns/
22. 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21
23. 3 or 4 or 5 or 6 or 7 or 8 or 9
24. 23 and 24
25. exp clinical laboratory techniques/ut
26. laboratories/ut
27. exp diagnostic tests, routine/ut
28. 25 or 26 or 27
29. 28 or 29
30. limit 29 to (humans and yr="1997 - 2012")

ec, economics subheading; exp, the inclusion of all terms more specific than the key word; ut, utilization subheading.

To account for delays in MeSH indexing after publication, 1 year of overlap with the original study was included.
The authors designed their studies to address one of four aims. (1) Assessment studies sought only to measure the extent of adherence to the criteria. (2) Explanation studies went a step further to not only assess, but also explain the reasons for adherence. (3) Intervention studies assessed a baseline, formed a hypothesis for explaining adherence, and measured the performance of an intervention designed to improve adherence. (4) The last class of studies involved the creation and validation of new audit rules.

Each of these study designs allows for prospective or retrospective data collection. Studies that evaluated an intervention assessed adherence before and after the intervention. We recorded only the timing of their initial baseline assessment. In contrast, studies developing new audit criteria had to perform an initial validation of the proposed rule, followed by a second assessment of adherence to the newly developed rule. We designated these studies as prospective or retrospective depending on the method used in the final assessment. To highlight studies that examined not just adherence, but its broader influence on the patient and society, we recorded the presence of additional outcome measures.

Criteria for Test Appropriateness

The sources of the criteria for test appropriateness were grouped into five categories: (1) guidelines endorsed by an organization (ie, government, professional society), (2) primary literature, (3) local consensus without specific literature cited, (4) rules developed and validated within the article, and (5) individual opinion. A small number of studies used two categories, such as a governmental guideline and primary literature. We classified the articles with multiple sources according to the largest group consensus, so that government and professional society guidelines had preference over locally developed utilization or individual opinion. While this classification reflects consensus about the criteria, it does not strictly imply a level of evidence. For example, duplicate genetic testing for germline mutations is inappropriate at the highest level of evidence, common sense.20 It also reflects current knowledge because articles using local consensus or individual opinion generally echoed the sentiments of Pilon et al: “There were no clinical trials published...to validate or refute any possible indications.”21

Each audit criterion selects a cohort of patients. To describe the study cohort we selected setting of care (ie, inpatient, outpatient, and emergency) and the type of data required for patient selection (eg, diagnostics, medications, diagnostic codes) as summary measures. Audit results also depend on characteristics of the individuals placing the order. For example, risk-averse providers order more laboratory tests,22 and more experienced providers order fewer.23 We recorded any available information about health care providers who ordered the tests reviewed by the audit, including their training, years of experience, service affiliation, and the number of providers in the audit. Provider information creates an interpretive context when generalizing the results of an audit.24

Data Collection

Two studies with similar audit criteria may choose different methods for data collection. For example, audits that seek to assess follow-up testing for diabetes in postpartum mothers with a history of gestational diabetes may manually review patient charts or use a database query to identify patients with an International Classification of Diseases (ICD) code matching gestational diabetes. To determine the adoption of encoded data in laboratory audits, we recorded as a binary variable the presence or absence of manual intervention (eg, chart review) in either the selection of patients or the collection of audit data. We included studies from the prior review to trend this pattern across a longer period.

As a second data collection parameter, we documented the source of data for the audit. Example sources include electronic databases, patient charts, and test requisition forms. We tallied all data sources. Studies selecting a cohort of patients from a database followed by manual chart review were counted for both database and chart review. We also recorded the audit’s sample size to correlate with the method of data collection.

The review protocol has been registered in PROSPERO International Prospective Register of Systematic Reviews (crd.york.ac.uk/prospero/index.asp; Identifier: CRD42013004324).

Results

Study Selection

We identified 119 articles for inclusion in the study. A complete list of studies can be found in the Appendix.

Audit Description

Author Affiliations and Study Design

About one-fifth of all studies had an author affiliated with either pathology or laboratory medicine (23/118). One study from the United States did not list author affiliations.25 The majority of studies occurred in the United States (51/119), United Kingdom (22/119), Canada (9/119), and the Netherlands (6/119). The proportion of studies with an author affiliated with either pathology or laboratory medicine varied by country: United States (8/50), United Kingdom (11/22), and Canada (4/9).

For the study aim, the majority evaluated adherence only (73/119), followed by intervention (31/119), explanation...
and more comprehensive group (government > professional organization). Literature citations occurred less frequently (26/119), followed by individual opinion (10/119), local consensus (6/119), and within-study development (3/119). These findings differ from the original study, in which only three studies (3/44) based their criteria on published guidelines.

The source of audit criteria has changed dramatically over time. Since 1990, the use of organizational endorsement has become the predominant source of utilization criteria, accounting for over 75% of audits reports in recent years. On the other hand, author opinion, the dominant method in earlier studies, has become rare in recent utilization audit studies (see Figure 2).

Researchers selected patient cohorts in different ways. The most common method enrolled patients with a specific test order (50/119). For example, one study enrolled all patients having one of three genetic tests. Other studies used clinical data for patient selection (45/119). Cohorts defined by clinical data vs test data differ in their reporting of underutilization. Underutilization requires knowing who needed but did not receive the test, which was lacking in those cohorts that were defined by the tests ordered. Of studies with a clinical data cohort, fewer than half involved data encoded as a medication (5/45) or diagnostic code (14/45). One study with medication as an enrollment criterion looked for an alanine aminotransferase request after a patient began a statin drug. Another study found patients with an ICD-9 code for diabetes to monitor requisitions for hemoglobin A1c. The remainder of studies with clinical data relied on human chart review (26/45). Several studies combined clinical and test information (24/119).

The settings of patient care included outpatient (37/119), inpatient including emergency care (49/119), or both (30/119). Three studies that selected patients from the laboratory database did not specify the setting of care. Most of the

Figure 1: Results of the systematic search strategy and study selection process. To screen the abstracts and full text, we reviewed each study to determine if it contained a criterion to define “appropriate” testing (ie, a patient with diabetes should undergo a hemoglobin A1c test at least every 180 days), applied the criterion to a patient cohort (ie, Medicare diabetic patients in Louisiana), and reported adherence to the criterion (ie, 24%).

Figure 2: Change in the source of test utilization audit criteria from 1965 to 2012. The source of test utilization criteria is shown as absolute (A) and proportional (B) yearly totals. The curves fit the data with a locally weighted regression function.

Criteria for Test Appropriateness

The most common source of criteria came from an organizational endorsement (74/119). These organizations included governments (30/74) and professional organizations (44/74). For studies using both, we deferred to the larger (12/119), and rule development (3/119). A handful of the intervention studies (4/31) were randomized trials. Most studies gathered information retrospectively and others gathered prospective data (23/119).

Fewer than half of the studies documented any patient or societal outcome measure (50/119). Outcomes included cost savings from test deferment, patient complications from false-positive results incurred from overtesting, or cases initially missed by diagnostic underutilization.

© American Society for Clinical Pathology

Downloaded from https://academic.oup.com/ajcp/article-abstract/141/6/774/1766246/Do-We-Now-Know-What-Inappropriate-Laboratory by guest on 02 October 2017
studies took place exclusively in an academic care setting, as explicitly stated in the article (66/119). Others involved a managed care setting (7/119). Few studies documented information about the providers requesting the test (30/119).

**Data Collection**

The audits gathered data from multiple sources, including database queries, chart reviews, surveys and questionnaires, requisition forms, and direct observation. A study could use more than one source. Most used chart review (n = 71). Surveys and requisition forms were less popular (n = 15). One study directly observed test orders. Studies using a database (n = 38) for data collection often used administrative data such as diagnostic codes, medications, or the laboratory information system. A number of studies involved only database queries (36/119) to complete the entire audit. The remainder of the citations (83/119) required manual chart review. For comparison, one study (1/44) from the prior systematic review performed the audit with a database query. Over time, both the number and proportion of test utilization publications performed on encoded data using database queries have increased, with 40% of recent audits relying entirely on a database. Database audits had diverse data fields, including test results, order date, medications, and diagnosis codes. Unlike the studies involving chart review, the database audits tended to use simple criteria without complex inclusion or exclusion logic. For example, four studies monitored the time interval between hemoglobin A1c tests. Another five studies evaluated recommended laboratory monitoring after beginning a new medication (ie, statins and serum alanine aminotransferase). The study by Petitti et al is noteworthy for its complex database-selected cohort, because it identified patients with diabetes by linking pharmacy, laboratory, and clinical databases, whereas the other studies relied exclusively on ICD codes for diabetes. In general, studies performing the audit without manual review had a larger sample size.

**Discussion**

Since the previous appraisal of laboratory utilization in 1998, the field has evolved. Published guidelines have become increasingly popular as audit criteria (see Figure 2). This change appears to provide a consensus answer to the question of the original review, “Do we know what inappropriate laboratory utilization is?” because the majority of authors defined “inappropriate” as practicing medicine in opposition to an organizational guideline. We also observed an increase in the adoption of database queries as a method for data collection in laboratory utilization audits (see Figure 3). In addition, only 20% of studies had an author affiliated with pathology or laboratory medicine.

Organizational guidelines have multiple advantages over local consensus studies, which may have led to their adoption. In contrast to local consensus groups, governments and professional societies assemble panels of subject-specific experts to formulate their conclusions. As future research changes recommendations, networking experts will simplify maintenance of state-of-the-art guidelines. Referencing a recognized body of physicians also simplifies the process of local change management. The local definition of “appropriate” seems unlikely to compete with free, high-quality guidelines from a reputable organization, especially if the organization strives to remain unbiased despite financial conflicts of interest.

The proportion of studies analyzing exclusively digital data has increased to 40% which is likely because of the rapid, accurate, and low-cost transactions afforded by database technology. Database technology certainly has advantages over manual chart review, but the reliance of database technology

---

**Figure 3** Trends in test utilization audits carried out with a database query from 1965 to 2012. The number (A) and proportion (B) of yearly published articles conducting their entire utilization review with structured data using database queries are shown. These articles did not use manual chart review. The curves fit the data with a locally weighted regression function.
on structured data makes it impossible to conduct utilization audits that require interpretation of complex medical scenarios. As suggested by the National Quality Forum, any feasible national quality metric will require database audits.36

It remains unclear which medical domain, if any, will take ownership of laboratory utilization. Pathologists overseeing the laboratory do not universally view themselves as guardians of laboratory services, particularly in the United States. In our opinion, the directors of clinical laboratories should consider the further development of test utilization as a duty to their patients. Appropriate clinical referrals can decrease the rate of false-positive results from overutilization while encouraging necessary tests, as demonstrated by multiple studies in this review.28-30 It remains unclear if pathologists, who already expend a great effort on quality control, view test utilization as a quality control issue. Both quality control and test utilization increase test precision, protect against false results, and ensure the overall integrity of laboratory information.28-30,37

This study has a number of limitations. Although we screened more than twice the number of articles in the original study, this was not an exhaustive search. Many well-conducted audits are not indexed by PubMed or did not meet our search criteria. Also, audits evaluating criteria adherence do not represent the only means of managing test utilization. A common study design, without a criterion, correlates a decrease in laboratory services with no change in patient outcomes before and after an intervention.32

Suggestions for Future Work

Laboratory utilization studies, as outlined in this review, have considerable variability. The necessary complexity of this study’s descriptive parameters and their general lack of consensus support this assertion. The diversity of study methodology complicates the comparison of individual study outcomes. As a consequence of the heterogeneity, we avoided reporting an overall rate of inappropriate testing because we feel this would be misleading.

The future of utilization could greatly benefit from the creation of a standard for reporting clinical laboratory audits. Such a standard would contain required information for a laboratory audit such as patient inclusion criteria, the method of data collection, and appropriateness criteria used to judge adherence. A standard reporting method would allow for a transparent comparison of adherence to appropriateness guidelines.34

Our systematic review summarizes a half century of work by hundreds of individuals. In this time, laboratory diagnostics have transformed from essentially nonexistent to essential in the practice of medicine, yet a systematic effort to manage their use remains an elusive goal.

Address reprint requests to Dr Hauser: Dept of Laboratory Medicine, Yale University, 330 Cedar St, PO Box 208035, New Haven, CT, 06511; ronald.hauser@yale.edu.

Appendix

Complete List of Studies


References


7. Bulusu S. Rational, cost effective use of investigations: rising workload and costs in diagnostic departments must be contained. BMJ. 2002;325:222.

8. van Walraven C, Naylor CD. Do we now know what inappropriate laboratory utilization is? a systematic review


