Limited Usefulness of the Proportion of Tests with Normal Results in Review of Diagnostic Services Utilization

John M. Eisenberg and Sankey V. Williams

We tested the hypothesis that the proportion of normal results in tests ordered by a physician correlates directly with the proportion of tests that are disapproved by a utilization review audit for inappropriate test use. Auditors reviewed 3497 charts for 13 067 repeated determinations of 10 diagnostic tests ordered by 81 medical residents to determine if test use had been appropriate, according to explicit criteria. By the Pearson product-moment correlation coefficient, there was a significant \( p < 0.001 \) correlation between the proportion of normal results in tests ordered by residents and the residents' audit-disapproval rates, a previously undocumented relationship. Although statistically significant, the correlation coefficient of 0.25 means that the proportion of normal results predicted only 6% of the variance in the audit-disapproval rates. This relationship is not strong enough to use the proportion of normal results as a measure of a physician’s inappropriate test use, except to select physicians for utilization review.

Additional Keyphrases: economics of laboratory operation quality control

As increasing attention turns to the cost of diagnostic testing, there have been numerous efforts to alter physicians' use of diagnostic tests (1–6). However, few of these utilization review programs have included chart audit to determine whether the diagnostic testing was appropriate according to explicit criteria. Usually, the success of these utilization review programs has been measured by a decrease in the number of all tests rather than by a decrease in the number of inappropriate tests.

One reason for this reluctance to identify tests that are ordered inappropriately is the cost and time involved in chart audit. The difficulty of establishing criteria for the appropriate use of diagnostic tests is an additional obstacle. To avoid chart audit, some investigators have proposed that the proportion of normal test results be used as an indirect indicator of the inappropriateness of the tests (7). This study was designed to test the hypothesis that a physician’s proportion of normal test results is related to the level of inappropriate test use.

Materials and Methods

All 81 medical residents who rotated on the general medical services and the medical intensive-care unit at the Hospital of the University of Pennsylvania during 1978–1980 were included in this 24-month study.

The 10 diagnostic tests studied were among those most frequently ordered by physicians at the hospital: chest roentgenogram examination and measurement of serum albumin, alkaline phosphatase, amylase, total bilirubin, calcium, creatinine, lactate dehydrogenase, alanine aminotransferase, and thyroxin. None of these tests are ordered as part of a multiphasic screening panel at this hospital.

Auditors selected inpatient medical charts for review if multiple determinations of diagnostic studies were obtained within a short period of time, because inappropriate test use was expected to be high in these situations. Charts were audited if patients had two or more determinations of serum amylase or thyroxin, or if patients had three or more of one of the other diagnostic tests within a seven-day period.

Cases to be reviewed were identified by the clinical laboratory’s computerized reporting system or by the radiology department’s computer system (8). Every weekday, a computer program designed for this project listed all medical patients who had received multiple determinations of one of the tests being studied during a seven-day period. To prevent unnecessary duplication of audits, we recorded which tests were audited for each resident and the date the audit was performed.

Charts were audited by medical records specialists, using explicit criteria to determine whether over-utilization of the diagnostic test had taken place. The explicit criteria, designed by a panel of 12 physicians from the medical staff at this hospital (9), identified which situations should be present to justify multiple determinations of diagnostic procedures in a seven-day period. The criteria were designed to be applicable to all patients in the hospital who underwent repeated testing. As an example, Table 1 lists the criteria for appropriate determination of serum calcium. For each audit, the medical-records specialist summarized the patient’s clinical problems, listed the dates on which the test was performed, recorded the results, and determined wheth-

---

Table 1. Criteria for Appropriate Repeated Determinations of Serum Calcium within a Seven-Day Period

<table>
<thead>
<tr>
<th>Three determinations</th>
<th>Pancreatitis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapy for hypercalcemia (including furosemide and other diuretics, steroids, saline infusion, mithramycin)</td>
<td></td>
</tr>
<tr>
<td>Therapy for hypocalcemia (including dialysis, vitamin D, high calcium diet, intravenous calcium infusion)</td>
<td></td>
</tr>
<tr>
<td>Parathyroidectomy within the past two weeks</td>
<td></td>
</tr>
<tr>
<td>Thyroidectomy within the past two weeks</td>
<td></td>
</tr>
<tr>
<td>More than four blood transfusions in 48 h</td>
<td></td>
</tr>
<tr>
<td>Parenteral hyperalimentation</td>
<td></td>
</tr>
<tr>
<td>One normal and one abnormal value</td>
<td></td>
</tr>
</tbody>
</table>

More than three determinations

Pancreatitis with severe pain, hypotension, or decrease in hemoglobin

Therapy for symptomatic hypercalcemia (including lethargy, mental changes, coma)

Therapy for symptomatic hypocalcemia (including tetany, irritability, Trousseau's sign)

Exchange transfusion

Exception (including physician's judgement regarding a steadily increasing or decreasing result)
er the test's use met the explicit criteria. Each audit record was reviewed with one of the physicians directing the project (JME and SVW) to determine whether there was an exception to the criteria; we noted exceptions in about 5% of the audits. To confirm the accuracy of the audits, we periodically reviewed a sample of medical records on the hospital floor without knowledge of the results of the audit by the medical-records specialist. The audits were concordant in more than 90% of these reviews; when disagreement occurred, repeated review of the chart usually demonstrated the medical-records specialist to be correct. The audit procedure has been described in greater detail elsewhere (8, 9).

In addition to determining whether the diagnostic services had been used appropriately, the chart auditors recorded the number of abnormal and normal results for the test being audited. A test result was considered to be abnormal if its value was outside the normal range published by the hospital's laboratory (which performed all laboratory tests in the study). Criteria for normal chest roentgenograms were established by the investigators. Therefore, the ratio of abnormal to normal test results and the ratio of approved to disapproved audits were available for each resident.

We used the Pearson product-moment correlation coefficient to determine the relationship between the proportion of disapproved audits and the proportion of normal tests. Statistical significance was determined with the F-test. The physician was the unit of analysis.

Results

For the 81 medical residents, 3497 audits were performed, involving review of 13,067 determinations (mean, 3.7 determinations per audit). For each resident, a mean of 43.2 audits was performed and 161 test determinations reviewed. Of the test determinations reviewed, 7344 (56.2%) yielded abnormal results; a similar percentage of the audits (57.5%, or 2010 audits) revealed unnecessary use of the diagnostic service being reviewed.

There was a direct relationship between the proportion of tests that provided normal results and the proportion of charts that revealed inappropriate use for the same resident (Figure 1). The Pearson product-moment correlation coefficient for the relationship was 0.25 (p < 0.001). The regression equation comparing the proportion of tests with normal results (x) and the proportion of audit disapprovals (y) was y = 0.26 + 0.64x.

To determine whether the relationship between the proportion of normal results and the proportion of inappropriate testing for the entire group of tests was substantially influenced by one or several of the tests that we audited, we performed several iterations of the analysis in which one or more tests were excluded from consideration. For example, we excluded determinations of lactate dehydrogenase to evaluate the effect of testing patients admitted to the intensive care unit with chest pain (for which a normal value would have been considered to be appropriate in the process of "ruling out" myocardial infarction). In none of the analyses were the results substantially different from those reported for the entire set of diagnostic tests.

Discussion

Most published descriptions of interventions that were designed to reduce the over-utilization of diagnostic tests have reported their results in terms of reductions in the number of tests that are ordered by physicians (1–6). Bailet et al. (10), however, have questioned the assumption that less test utilization represents a more desirable level of care. Their study of a review system of dental utilization revealed no consistent differences in denial rates of claims (a measure of unnecessary care) from high vs moderate utilization practices. In our previous investigations we have used explicit criteria and chart audit to evaluate the appropriateness of repeated diagnostic tests (8, 9). However, chart audit is time-consuming and therefore an expensive method of utilization review. If a method of utilization review could be developed to estimate the appropriateness of diagnostic testing without the need for chart audit, the method could potentially improve the efficiency of programs to reduce unnecessary testing.

Dixon and Laszlo (7) have suggested that the number of normal and abnormal test results can be used to monitor the use of chemical laboratory tests by housestaff. After auditing 50 charts, they demonstrated a significant (p < 0.001) and powerful (correlation coefficient = 0.96) relationship between the number of abnormal results for each patient and the number of tests for which the results were used in the diagnosis or treatment of that patient. Our analysis differed from theirs in four important respects. First, our analysis concentrated on whether it was reasonable to order the test and not on whether the result was used in patient management. Second, we examined repeated determinations of tests, while Dixon and Laszlo examined all determinations of a set of tests that only partly matched ours. Third, in our analysis the physician, not the patient, was the unit of analysis. Fourth, we examined the proportion, not the number, of normal test results.

These differences in study design, especially the fourth, might explain the remarkable difference in results between the two studies. Dixon and Laszlo may have found a powerful correlation because they examined relationships involving the number, instead of the proportion, of test results. It is predictable that both the number of abnormal results and the number of results that are used in patient management will change in the same direction. For example, if the number of tests decreases, there should be a high correlation between the change in the number of abnormal tests and the change in the number of results that are used in patient management, simply because there will be fewer of both. However, because there are fewer tests, it is also likely that there will be fewer normal tests and fewer that are not used in patient management, although Dixon and Laszlo did not report these data. On the other hand, it is not

Fig. 1. Correlation between the proportion of tests with normal results and the proportion of audit disapprovals. Each point represents one of 81 physicians.
so clear that both the proportion of test results that are abnormal and the proportion of test results that are appropriate should change. For example, if the decision to order fewer tests did not involve the reduction of inappropriate tests more than appropriate tests, the proportion of test results that are appropriate need not change. Therefore, Dixon and Laszlo's results may have represented a tautology, because it would be surprising if there were not a greater number of tests that changed management when the number of tests performed was greater, even if there were no change in the proportion of appropriate tests.

Our study shows a direct relationship between the proportion of tests with normal results and the proportion of tests that reflect inappropriate use of diagnostic services after chart review with explicit criteria. Although the relationship between abnormal test results and appropriate test use is statistically significant ($p < 0.001$), the correlation coefficient is sufficiently low (0.25) to preclude use of the proportion of tests with normal results to predict if an individual physician is using diagnostic tests appropriately. Moreover, there was substantial variation in these proportions by physician (see Figure 1).

The correlation coefficient of 0.25 means that the model can explain only 6% of the variance in physicians' inappropriate use of tests. In addition, because one of the criteria for appropriate testing of several of the tests was the presence of one previously abnormal determination, this criterion may, by definition, increase the correlation between the proportion of tests that are abnormal and the proportion of tests that are defined as appropriate. Therefore, even the 6% figure may be an overstatement. At best, the ratio of normal to abnormal results in diagnostic tests ordered by physicians may provide an initial screen to select physicians who are more likely to have ordered unnecessary tests. Once selected, the charts of these physicians could be reviewed before any intervention was instituted. Our results do not support the replacement of chart audit with determination of the proportion of normal test results as a means of detecting inappropriate test use.

These results do not imply that 57.5% of laboratory test use in our hospital was inappropriate. In our utilization review program, an audit indicated inappropriate test use if any of the multiple determinations did not meet the explicit criteria, even if all the others did. In addition, it is likely that the patients who qualified for this study by undergoing multiple test determinations were those most likely to have undergone unnecessary diagnostic testing. The high proportion of reviewed cases that revealed unnecessary test use does suggest, however, that the strategy of monitoring the repetitive ordering of tests was efficient in selecting cases for audit. The results of these audits were used in a program designed to change physicians' use of diagnostic tests, the results of which are now being reviewed for analysis.

This research was supported in part by a grant from Blue Cross of Greater Philadelphia and in part by the National Health Care Management Center at the Leonard Davis Institute of Health Economics through Grant HS02577, National Center for Health Services Research, OASH, DHHS. SVW is a Henry J. Kaiser Family Foundation Faculty Scholar in General Internal Medicine. JME is Sol Katz Associate Professor of General Medicine at the University of Pennsylvania.

References