Errors, mistakes, blunders, outliers, or unacceptable results: how many?

David L. Witte,1* Sue Ann VanNess,1 Debbie S. Angstadt,1 and Beverly J. Pennell2

We have studied 219,353 individual clinical chemistry results obtained in methods comparison studies. Each result was prospectively compared with its replicate, comparative, or repeat value to identify differences from expected values. Unacceptable results were defined as differing from the expected values by ≤7 SDs or CVs. We believe these differences represent special-cause variation and should be expressed as unacceptable rates per million results (ppm). We observed 447 ppm unacceptables: 196 ppm in control samples and 251 ppm in patients’ samples. Results judged likely to alter patient care occurred at a rate of 41 ppm. To better understand the magnitude of these rates, we compared these results with reports of error rates in HIV testing and the airline industry. The measurements reported were made for the purpose of quality improvement, not judgment or discovery. The significance of these findings for laboratorians, manufacturers, and regulators is discussed.

INDEXING TERMS: accuracy • bias • precision • quality assurance • quality control • statistics • special-cause variation • statistical process control

Managing quality has always been a primary goal in the practice of clinical chemistry. Laboratory has developed systems for quality control, quality assurance, and continuous quality improvement. Quality system design continues to evolve. These efforts are aimed at creating results such that each and every result is believable and thought to have been reported without errors.

Others have studied the sources of errors throughout the testing cycle [1]. Ross and Boone have shown that 93% of the errors are either in the preanalytical or postanalytical portions of the process [2]. However, the 7% of errors occurring in the analytical portion are entirely the laboratory’s responsibility and are the focus of this paper.

Burnett states that any meaningful measure of analytical quality must include three statements specifying precision (CV or SD), accuracy (bias), and frequency of outliers [3]. The increased complexity of the analytical process leads to opportunities for increased variation. Variation may be of the common-cause variety, which is measured by CVs, or may be special-cause variation (i.e., mistakes), as described by Hinkley and Barkan [4, 5].

Few data have been published regarding the frequency of outliers in the clinical chemistry laboratory. In a study of nearly 1 million results, Lapworth and Teal reported 120 errors, of which 38 were believed to have occurred in the analytical phase [6]. Among these 38, most were attributed to sample mix-ups rather than instrument error. Kazmierczak and Catrou reported the results of replicate creatinine analyses in a study of random analytical errors [7]. Their rather strict definition of analytical error yielded a rate of 8.9%. Their data revealed that 2 of 438 results were >6 SD from the expected result. How should a significant analytical error be defined? How should this error rate be communicated among laboratorians? How should the error rate be communicated to other stakeholders? Gambino suggests that error rates could be reported in parts per million (ppm) [8], which would allow comparison with other error (catastrophe) rates in our society. Gambino reports there are 2.5 deaths per million anesthesia cases, 164 deaths per million automobile crashes, and 0.18 deaths per million passenger enplanements. The Lapworth and Teal data included ~38 analytical errors per million results. In the Kazmierczak and Catrou data, errors >6 SD occurred at a rate of 4566 per million results.

The Lapworth and Teal estimate of error rate utilized retrospective review of results by either laboratory check-out systems or clinician comparison with other patients’ data to detect an error. The Kazmierczak data involved replicate analyses of specimens and measured the analytical error rate. The former study could be viewed as retrospective, the latter as prospective. We report a summary of the frequency of widely discrepant analytical results observed in multiple prospective method evaluation studies over 18 years and including >200,000 results.
Materials and Methods

Methods comparison studies involved technology already in the market or immediately moving to wide marketing; i.e., these are the results of methods comparison studies on mature systems with well-characterized reagents. Studies using less-well-characterized technology were excluded from our review.

The studies reviewed were completed between March 1977 and September 1995 in five locations and included at least 14 operators and 16 discrete analyzers. The results were divided into routine discrete cuvette chemistries, discrete analyzer electrode measurements, and immunoanalyses, whether performed in a dedicated immunoanalyzer or in a discrete chemistry analyzer.

Every individual result had at least two associated results with which it could be evaluated, including the replicate analysis result for the same sample, the comparison method result, or the results for the quality-control specimens.

We obtained Institutional Review Board approval for use of excess sample for analytical evaluations. All results reported here were for analytical evaluation only. No results were reported to patients' records.

We tabulated the unacceptable results that were highly likely to have been reported. Results with method process flags indicating internal instrument problems were not tabulated. Unacceptable results attributable to operator sample mix-up were not tabulated. Rather, the results tabulated would have been believed to be accurate and reportable without the contradictory evidence of the replicate result.

The true result for a control sample was specified as the mean result. The true result for a patient's sample was specified as the duplicate result, a repeat analysis result, a comparison method result, or a combination of these results. The difference from “truth” for controls was calculated by using the SD for the control; i.e., the SD interval (SDI) was the difference divided by the SD. The difference from “truth” for patients’ samples was calculated from the CVs for controls with a similar concentration; i.e., CVI equals difference divided by CV times “truth.”

Results

Table 1 summarizes the results of 219,353 separate analyses, of which 98 were shown to be >7 SD different from the expected value. An additional 79 results differed from the expected value by 4.0 to 6.9 SD.

All of the 98 results differing by >7 SD were reviewed in detail. Controls accounted for 43 of the 98 results, patients’ samples for the remaining 55. No patient harm occurred because these results were for methods comparison studies and no result was reported to a patient chart. Nonetheless, we categorized the errors for the patients’ samples as to the likelihood of causing a patient management problem (see Table 2). Many of these large differences for patients’ samples did not cross typical decision points, were not independent tests for decision-making, or were frankly not believable results (e.g., cholesterol at 50 mg/L). Some results for patients’ samples did cross decision lines but were judged not likely to alter patient management because the abnormalities were relatively small. Table 3 illustrates the 9 results that were judged to potentially cause errors in patient management. These discrepant results could have suggested an erroneous change in a therapeutic drug concentration, an erroneous hypoglycemia, 3 erroneously normal results, and an erroneous hyperferremia.

Discussion

We have chosen to refer to these widely discrepant analytical results as unacceptable results. Healy chose outliers and avoided the pejorative word error [9]. Lapworth and Teal [6] chose blunder. However, as Healy points out, not all blunders necessarily yield outlying results. Burnett [3] also chose to refer to these discrepant results as outliers. Burnett observed that outliers can be identified only if the method has sufficient precision such that a widely discrepant result can be shown to be different from the usual variation. Burnett chose to define outliers as >3.83 SD from the mean when the number of samples in the study was 400.

Lapworth and Teal attributed the majority of their analytical errors to sample mix-ups. The differences we observed were prospectively investigated at the time of the study and were felt not to be explainable by a patient or control sample mix-up. We conclude that the unacceptable rate reported here is more likely attributable to malfunction of automated analytical equipment. We recognize, however, that other explanations may be considered.

<table>
<thead>
<tr>
<th>No. of results</th>
<th>Cuvettes</th>
<th>Electrodes</th>
<th>Immunoassays</th>
<th>Total</th>
<th>Total ppm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference &gt;7 SD</td>
<td>180 261</td>
<td>22 978</td>
<td>16 114</td>
<td>219 353</td>
<td>NA</td>
</tr>
<tr>
<td>Controls</td>
<td>85</td>
<td>1</td>
<td>12</td>
<td>98</td>
<td>447</td>
</tr>
<tr>
<td>Patients</td>
<td>39</td>
<td>0</td>
<td>4</td>
<td>43</td>
<td>196</td>
</tr>
<tr>
<td>Difference 4–6 SD</td>
<td>69</td>
<td>0</td>
<td>10</td>
<td>79</td>
<td>360</td>
</tr>
<tr>
<td>Controls</td>
<td>46</td>
<td>0</td>
<td>1</td>
<td>47</td>
<td>214</td>
</tr>
<tr>
<td>Patients</td>
<td>23</td>
<td>0</td>
<td>9</td>
<td>32</td>
<td>146</td>
</tr>
<tr>
<td>Totals</td>
<td>154</td>
<td>1</td>
<td>22</td>
<td>177</td>
<td>807</td>
</tr>
</tbody>
</table>
The potential laboratory outcome of unacceptable quality-control specimens would most likely be a repeated analytical run. However, the decision to repeat a run varies between laboratories, depending on the quality-control rules used in each laboratory.

Not all patients’ samples with unacceptable results are equally likely to alter patients’ outcomes. The 9 results judged likely to alter patient management translate to 41 ppm, which can be compared with other error rates. Certainly the severity of these patient management problems does not compare with a death in an aviation accident (0.18 deaths per million enplanements); nonetheless, they are judged important. Boone [2] reported 5 cases of risky tests being performed on patients as a result of unacceptable laboratory results. Converting Boone’s data to parts per million nomenclature reveals that 375 patients per million would have undergone tests associated with risk.

In our study, 14 results were judged to be a nuisance or to cause confusion in patient management; this translates to 64 ppm. For comparison, let us consider another kind of confusing results: lost baggage. About 5000 cases of lost baggage are reported per million passengers [10]. Does this suggest that laboratories cause less confusion than aviation baggage handlers?

Many of the unacceptable results for patients’ samples would have caused no patient management problem. Physiologically unlikely results such as that already mentioned for cholesterol would not be reported by any responsible laboratorian and should be detected by all quality-assurance systems. Therefore, these types of results were not judged to be important in patient management; these discrepant results do, however, alter the productivity of a laboratory.

The gaussian distribution predicts that results with a difference >3 SD would occur at a rate of 2700 ppm, values >4 SD would occur at 63 ppm, values >5 SD at 0.6 ppm, >6 SD at 0.002 ppm, and >7 SD at 3 per 10^12 (3 × 10^-6 ppm) [11]. We chose to tabulate differences >7 SD or CV from the expected. Differences this great should be rare in a gaussian distribution. Our observation of 447 ppm differences >7 SD suggests that the gaussian distribution is inadequate to explain these differences and suggests the differences are not of the common-cause variety. Perhaps another distribution or chaos theory and Cantor sets would be more descriptive [9, 12]. We believe the 447 ppm rate is consistent with the quality-improvement literature that suggests the presence of common-cause variations and special-cause variations. These unacceptable results are most probably special-cause variations and not part of an easily described statistical distribution.

Our report of 447 ppm unacceptable results with differences >7 SD should be compared with other reports. The consumer electronics industry frequently aims for error rates as low as 2–3 ppm (the 6 sigma goal? [11]). The laboratory study of Boone and Ross cited by Boone [2] was retrospective and their report of 375 per million patients undergoing additional diagnostic procedures and incurring risk is not attributed only to the analytical process. Another laboratory study in a practice surveillance network [13] determined 112 laboratory errors in physician office laboratories per million patient visits; in addition, 37 reference laboratory results were shown to be in error for these million patient visits. Among these 149 (112 plus 37) errors, 50 were thought to alter patient management (50 ppm). These error rates are considerably higher than the 1 ppm error rate reported by Hutchinson [12] for HIV testing in the military.

Has the unacceptable result rate improved over time? A review of the initial proficiency data of Belk and Sunderland [14] suggests larger errors in the past. Converting the data from that 1947 study to parts per million gives an unacceptable result rate in that study of 162 116 per million laboratory tests. Example errors reported in that study were glucose >1000 mg/L when the true value was 600 mg/L, and blood urea nitrogen <200 mg/L when the true value was 450 mg/L. A group of Australian laboratories were recently reported to have error rates from 2% (20 000 ppm) to 30% (300 000 ppm) on quality-assurance samples [15] (the error criteria were not stated in SDI units, and many were transcription errors). Both of these proficiency testing error rates are much higher than the 12 904 ppm total unacceptables (7955/616 467) reported by the College of American Pathologists [16]. In the latter data, 1108 unacceptables were attributed to instrument error, or 1797 ppm.

The unacceptable result rate we report here needs to be contrasted with the common-cause variation and error rate discussed by others [17-19]. Klee [17] has discussed the errors in patient management that occur when there is a shift in the analytical process such that a greater percentage of sample results will be judged to be above or below a given clinical decision concentration. This bias is

<table>
<thead>
<tr>
<th>Test</th>
<th>“Error”</th>
<th>“Truth”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iron</td>
<td>237</td>
<td>64</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>176</td>
<td>363</td>
</tr>
<tr>
<td>LD-1</td>
<td>24.6</td>
<td>111.0</td>
</tr>
<tr>
<td>Uric acid</td>
<td>7.9</td>
<td>11.2</td>
</tr>
<tr>
<td>Glucose</td>
<td>28</td>
<td>90</td>
</tr>
<tr>
<td>Theophylline</td>
<td>3.3</td>
<td>19.8</td>
</tr>
<tr>
<td>Theophylline</td>
<td>6.9</td>
<td>16.2</td>
</tr>
<tr>
<td>Tobramycin</td>
<td>14.8</td>
<td>5.7</td>
</tr>
<tr>
<td>Phenytoin</td>
<td>23.2</td>
<td>15.8</td>
</tr>
</tbody>
</table>

Table 2. Patient differences >7 SD.

<table>
<thead>
<tr>
<th>Cuvettes</th>
<th>Electrodes</th>
<th>Immuno</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td></td>
<td>Total ppm</td>
</tr>
<tr>
<td>Unlikely problem</td>
<td>146</td>
<td></td>
</tr>
<tr>
<td>Possible confusion</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>Potential error</td>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Potential patient management error.
different from the kinds of unacceptable values we discussed. Likewise, Westgard et al. [18] and Hatjimihail [19] discussed quality-control parameters used to detect shifts or random common-cause errors in the analytical process. Instead, we are reporting special-cause variation, which is a different source of error than the bias or random variation discussed by Klee and Westgard.

We believe that the decreasing analytical error rate in clinical chemistry laboratories of today makes it more difficult for an individual laboratory to estimate its unacceptable result rate. Similarly, Busch and Alter have discussed the problems with identifying errors in processes where errors are very few; e.g., it has become very difficult to estimate the likelihood of transmission of HIV via transfusion because the transmission rate is so low [20]. Likewise, it is difficult to devise studies to identify error rates as low as those that probably currently exist in the clinical chemistry laboratory. The measurement of low error rates will require large databases.

At present, debate is growing regarding quality-control systems necessary to assure acceptable results in clinical chemistry laboratories. Much of the debate centers around the use of unit test devices. The laboratory community has great concern regarding the robustness of results from unit test devices. However, how much does the laboratory community know about the robustness of the “standard” technology currently in use? We believe this study, showing error rates in the 50–500 ppm range, is representative of many current devices and may help in judging future analytical devices. For comparison a manufacturer of a new unit test device could use the American National Standard for Inspection/American Society for Quality Control sampling procedures (ANSI/ASQC 21 4–1993, Table X-N-1) to evaluate manufacturing batch sizes of 35 000 to 150 000 units. This standard states that if 500 units are tested and no more than 1 is defective, the batch meets the average quality limit of 250 ppm unacceptables. Another way of expressing this is that if the evaluated devices have 211 ppm unacceptables, then 90% of groups of 500 samples will have 1 or 0 unacceptables. Manufacturers, laboratorians, and regulators need to find commercially viable means to generate and share unacceptable rate data for both existing and new devices.

What unacceptable rate is tolerable? Affordable? Attainable? A worthy goal? What will reduce poor patient outcomes? The error rate in clinical chemistry analyses is probably higher than the rate of transfusion-related transmission of viral diseases. Contrast the hundreds of unacceptable results per million with the likelihood of transmitting any infectious viral disease by transfusion (i.e., 29 per million transfusions) [21]. How much more effort is needed to reduce the unacceptables rate in clinical chemistry laboratories? Again relating to the transfusion medicine example, the US has recently added testing for HIV antigen in donated blood in an attempt to prevent from 5 to 16 HIV transmissions in 12 million transfusions. That means that the policy makers for blood safety wish to have HIV transmissions <1 per million transfusions. It further indicates that the goal of reducing the HIV transmission rate from 2.2 ppm to somewhere between 0.4 and 1.3 ppm is worth the considerable cost of analyzing every donor unit for HIV antigen.

Again taking from the operational examples of the aviation industry, Leape [22] has suggested that our understanding of errors in medicine needs to be improved. Leape points out that the aviation industry accepts that errors are unavoidable and uses feedback and redundancies in the process to “absorb” the errors. Absorption of errors means that many errors occur but they are identified and corrected before any bad outcomes could occur. The aviation industry also has highly standardized systems for decision-making in the presence of an error. This should be contrasted with medicine, which is highly individualistic. Further, aviation vigorously inspects its near misses. But, perhaps the most important lesson from the aviation industry is that many redundancies and feedback systems can prevent bad outcomes in the presence of a measurable error rate. Laboratory processes could benefit from this mistake-proofing.

Laboratorians have shown benefit from mistake-proofing through using information systems to check patients’ results against predetermined limits or against previous results on the same patient, i.e., limit and delta checks. In addition, laboratorians currently “balance electrolytes” before reporting. We call this “intercheck” and extend it to other relationships between laboratory results such as albumin and globulin or aspartate and alanine aminotransferases. Are these the rudimentary examples of the feedback systems we need?

Acceptance of risk by many industries may thwart the search for systems with error rates in the 1–2 ppm range. Risk homeostasis theory states that great care to remove one risk of error may cause less care to be undertaken for different risks [23]. Laboratorians must be careful that clinical chemistry laboratories do not succumb to this risk homeostasis. Also, laboratorians must not lower standards just because it was acceptable for us to have an error rate of a certain size in the past: Continuous improvement is the goal.

We believe the data presented here will be useful when considering statements regarding regulatory controls on quality. Analytical quality is part of the total quality process. Research measurements can be made to discover new information, judge a current situation, or seek improvement opportunities. The present report is an example of seeking improvement opportunities [24]. Our data reaffirm that analytical errors are probably a very rare contributor to patient management errors but improvement should be possible. The partnership between laboratorians, industry, and regulators can continue to improve the credibility of laboratory results.
References