RISK: Health, Safety & Environment (1990-2002)

Volume 11 | Number 4 Article 5

September 2000

# Performance Indicator Analysis of Proficiency Criteria in the Drug-Testing-Laboratory Certification Process of the DHHS

John M. Gleason

Darold T. Barnum

# Performance Indicator Analysis of Proficiency Criteria in the Drug-Testing-Laboratory Certification Process of the DHHS

John M. Gleason & Darold T. Barnum*

## Introduction

Approximately 8.3 million adult drug users were employed in the United States in 1997.[1] These employees cost U.S. employers $100 billion annually due to higher costs related to turnover, absenteeism, accidents, decreased productivity, and health care.[2] Consequently, in recent years, testing of job applicants and employees for the presence of illegal substances has increased. Pre-employment drug screening is conducted by 95% of Fortune 500 companies, and more than six million workers are subject to mandatory testing because of governmental requirements.[3] Moreover, 81% of the corporations responding to an American Management Association survey used some form of drug testing.[4]

To be acceptable for detecting the use of unlawful substances, drug tests must accurately discriminate between those who use drugs and those who do not.[5] False positive errors in the testing process are

* Dr. Gleason is Professor of Decision Sciences in the Department of Information Systems and Technology and U.S. West Technology Fellow, Creighton University. He received B.S. (mathematics) and M.B.A. degrees from the University of Missouri at Kansas City, and a D.B.A. degree from Indiana University. *E-mail: jgleason@creighton.edu.*

Dr. Barnum is Professor of Managerial Studies at the University of Illinois, Chicago. He received a B.B.A. from the University of Texas at Austin, and M.B.A. and Ph.D. degrees from the Wharton School, University of Pennsylvania.

1 *See* U.S. Dep't of Health and Human Services, National Household Survey on Drug Abuse (1998).

2 *See* Jane Easter Bahls, *Drugs in the Workplace*, 43(2) H.R. Mag. 80 (1998); Steve Bates, *House Passes Bill to Curb Workplace Substance Abuse*, 86(8) Nation's Bus. 8 (1998); David Warner, *The War on Drugs Wants You*, 84(2) Nation's Bus. 54 (1996); *Workplace Substance Abuse: A Snapshot*, 17(6) Behav. Health Mgmt. 14 (1997).

3 *See* Bahls, *supra* note 2; Kevin B. Zeese, Drug Testing Legal Manual (1995).

4 *See* American Mgmt. Ass'n, 1996 AMA Survey on Workplace Drug Testing and Drug Abuse Policies (1996).

5 *See* Darold T. Barnum & John M. Gleason, *Analyzing Proficiency Criteria of Health Technology Systems: The Case of Drug Testing*, 46 IEEE Trans. Engin. Mgmt. 359 (1999);

especially serious, given the systems of jurisprudence in the United States and other industrialized nations which require individuals be presumed innocent until proven guilty.[6] Consequently, empirical studies are routinely conducted to estimate the proficiency of drug-testing laboratories. Many studies of drug-testing laboratories in the United States, and more recently in various European countries, have been published in major biomedical journals in the last two decades.[7] Moreover, laboratory proficiency studies are routinely conducted by government agencies as part of laboratory certification processes.

Darold T. Barnum & John M. Gleason, *The Credibility of Drug Tests: A Multi-stage Bayesian Analysis,* 47 Indus. Lab. Rel. Rev. 610 (1994); John M. Gleason & Darold T. Barnum, *A Probabilistic Analysis of Multiple-Drug Testing Procedures in Sports Doping Control,* 1 Int'l Trans. Operat. Research, 395 (1994); John M. Gleason & Darold T. Barnum, *Estimating Actual Rates of Drug Use,* 27 Socio-Economic Plan. Sci. 199 (1993); John M. Gleason & Darold T. Barnum, *Predictive Probabilities in Employee Drug Testing,* 2 RISK 3 (1991); Lennart E. Henriksson, *Drug Testing and Grievance Rates,* 23 J. Collect. Negot. Pub. Sector 211 (1994).

6 *See* Robert P. De Cresce et al., Drug Testing in the Workplace (1989); Tia S. Denenberg & Richard V. Denenberg, Alcohol and Other Drugs: Issues in Arbitration (1991); Helen Elkiss & Joseph Yaney, *Recent Trends in Arbitration of Substance Abuse Grievances,* 42 Lab. L.J. 556 (1991); Frank Elkouri & Edna Asper Elkouri, How Arbitration Works (1985); Marvin F. Hill & Anthony V. Sinicropi, Evidence in Arbitration (1987); Bureau of National Affairs, The Developing Labor Law (Charles J. Morris ed., 1983); Kenneth W. Thornicroft, *Arbitrators and Substance Abuse Discharge Grievances: An Empirical Assessment,* 14(4) Lab. Stud. J. 40 (1989). *See* Zeese, *supra* note 3.

7 *See, e.g.,* Roser Badia et al., *Survey on Drugs-of-Abuse Testing in the European Union,* 27 J. Analyt. Tox. 117 (1998); Joe D. Boone et al., *Laboratory Evaluation and Assistance Efforts: Mailed, On-site and Blind Proficiency Testing Surveys Conducted by the Centers for Disease Control,* 72 Am. J. Pub. Health 1364 (1982); D. Burnett et al., *A Survey of Drugs of Abuse Testing by Clinical Laboratories in the United Kingdom,* 27 Annals Clin. Biochem. 213 (1990); Kenneth H. Davis, Richard L. Hawks & Robert V. Blanke, *Assessment of Laboratory Quality in Urine Drug Testing,* 260 JAMA 1749 (1988); Christopher S. Frings, Robert M. White & Danielle J. Battaglia, *Status of Drugs-of-Abuse Testing in Urine: An AACC Study,* 33 Clin. Chem. 1683 (1987); Christopher S. Frings, Danielle J. Battaglia & Robert M. White, *Status of Drugs-of-Abuse Testing in Urine Under Blind Conditions: An AACC Study,* 35 Clin. Chem. 891 (1989); Edward Gottheil, Glenn R. Caddy & Deborah L. Austin, *Fallibility of Urine Drug Screens in Monitoring Methadone Programs,* 236 JAMA 1035 (1976); Hugh J. Hansen, Samuel P. Caudill & Joe D. Boone, *Crisis in Drug Testing: Results of CDC Blind Study,* 253 JAMA 2382 (1985); Susan J. Knight et al., *Industrial Employee Drug Screening: A Blind Study of Laboratory Performance Using Commercially Prepared Controls,* 32 J. Occup. Med. 715 (1990); P. Lafolie & O. Beck, *Deficient Performance of Drugs of Abuse Testing in Sweden: An External Control Study,* 54 Scandinavian J. Clin. Lab. Invest. 251 (1994); Louis C. LaMotte et al., *Comparison of Laboratory Performance with Blind and Mail-distributed Proficiency Testing Samples,* 92 Pub. Health Rep. 554 (1977); J. Segura et al., *Proficiency Testing on Drugs of Abuse: One Year's Experience in Spain,* 35 Clin. Chem. 879 (1989); J. F. Wilson et al., *External Quality Assessment of Techniques for the Detection of Drugs of Abuse in Urine,* 31 Annals Clin. Biochem. 335 (1994); J.F. Wilson et al., *Performance of Techniques Used to Detect Drugs of Abuse in Urine: Study Based on External Quality Assessment,* 37 Clin. Chem. 442 (1991).

The focus of the proficiency studies is to determine the risk of false positive and false negative errors in the laboratories' testing processes. This article reports the results of a Performance Indicator Analysis[8] of proficiency criteria in the drug-testing-laboratory certification process mandated by the U.S. Department of Health and Human Services (DHHS). We examine the problems inherent in the use of laboratories as the unit of analysis, and identify conditions under which these problems may result in misleading conclusions, incorrect decisions, and inappropriate demands for corrective actions.

## The DHHS Guidelines

The importance of accurate drug testing in clinical laboratories in the United States is reflected in federal policy through the Clinical Laboratory Improvement Amendments of 1988 (and the implementation of the regulations in 1993), in policies and procedures required by the U.S. Department of Transportation that mandate testing of safety sensitive employees in the transportation industry, and in the mandatory guidelines promulgated by the DHHS for federal workplace drug-testing programs.[9]

The DHHS guidelines specify scientific and technical requirements for federal agencies' workplace drug-testing programs. These guidelines also establish a certification process that applies to laboratories that perform drug testing for federal agencies. The guidelines require that random drug-testing programs test for marijuana and cocaine. The programs also may test for opiates, amphetamines, and phencyclidine, and for other drugs when tests are conducted on the basis of reasonable suspicion, accident, or safe practice.[10]

The importance of accuracy of the testing process is recognized in the DHHS guidelines: "Reliable discrimination between the presence, or absence, of specific drugs ... is critical, not only to achieve the goals of the testing program but to protect the rights of the Federal

8 *See* Darold T. Barnum & John M. Gleason, *Analyzing Proficiency Criteria of Health Technology Systems: The Case of Drug Testing,* 46 IEEE Trans. Engin. Mgmt. 359 (1999).

9 *See* P. Bachner, *Is It Time to Turn the Page on CLIA 1988?,* 279 JAMA 473 (1998); Procedures for Transportation Workplace Drug Testing Programs, Quality Assurance, and Quality Control, 49 C.F.R. § 40.31 (1997); Mandatory Guidelines for Federal Workplace Drug Testing Programs, 59 Fed. Reg. 29908, 29918, Subpart B, 2.1(a) (June 9, 1994).

10 *See* 59 Fed. Reg. 29908, Subpart B, 2.1(a).

employees being tested."[11] Furthermore, the guidelines indicate that drug testing should be considered a special application of forensic toxicology, because "of the impact of a positive test result on an individual's livelihood or rights, together with the possibility of a legal challenge of the results."[12] Accordingly, the DHHS laboratory certification program provides stringent proficiency standards that must be satisfied by laboratories involved in drug testing for federal agencies.

However, criteria used in proficiency-testing processes are often flawed. For example, we developed a process we refer to as Performance Indicator Analysis (PIA) to analyze drug-testing accuracy measures that were used in laboratory proficiency studies published in peer-reviewed medical and scientific journals from 1976 through 1998. We found that most indicators of proficiency are faulty, and that flawed indicators are present in every study.[13]

In this paper, we use PIA to identify problems with the DHHS proficiency-testing criteria. The PIA identifies a variety of proficiency-criteria problems that result when laboratories are considered as the units of analysis. Specifically, there are conditions under which laboratories with equal degrees of accuracy will have different false-positive reporting probabilities. Moreover, false-negative reporting probabilities may differ between laboratories, resulting in cases in which a superior-performing laboratory yields results which suggest lower proficiency than an inferior laboratory.

## False Positives and the DHHS Guidelines

The effect of a false positive depends on the certification status of the laboratory. During the initial DHHS certification process, any false positive that occurs during proficiency testing disqualifies a laboratory from further consideration. On the other hand, the DHHS proficiency-testing requirements specify that if any false positives result from blind challenges sent to a *certified* laboratory, then that laboratory will be identified as out of compliance and corrective actions must be taken.[14]

---

11 *Id.* at 29925, Subpart C, 3.2(b).

12 *Id.*, Subpart C, 3.2(c).

13 *See* Barnum & Gleason, *supra* note 8.

The guidelines require that laboratories processing more specimens receive more blind challenges.[15] Consequently, the PIA indicates that the probability that a laboratory reports a false positive is higher for the higher-volume laboratories, even though these laboratories identify a negative challenge at the same rate as lower-volume laboratories.

For example, assume that 20 blank proficiency test specimens (a blank specimen is one that does not contain drugs) are being tested for five drugs each, and that each drug test conducted by the laboratory has an individual specificity of 0.9999 (where specificity is the probability of a negative test result, given that the drug is not in the specimen); that is, there is only one chance in 10,000 that a negative drug challenge will yield a positive test result. It follows then that the probability of an individual *specimen* being correctly declared negative (that is, the probability of no false positives) is $0.9999^5$, which equals 0.9995 (assuming independency of tests). However, while there are five negative challenges per specimen, there are 100 negative challenges for the laboratory (20 specimens with five challenges per specimen), so the probability of a laboratory declaring all challenges correctly negative is $0.9999^{100}$, which equals 0.9900. That is, the probability of one or more false positives is 0.0100 (1.0 – 0.9900).

Now, assume that another (higher volume) laboratory receives 100 blank proficiency test specimens that are to be tested for five drugs each, and that each drug test conducted by the laboratory has an individual specificity of 0.9999 (that is, the laboratory has a specificity rate for each drug identical to that of the first laboratory). Again, the probability that an individual *specimen* is correctly declared negative (that is, the probability of no false positives) is $0.9999^5$, which equals 0.9995. However, while there are five negative challenges per specimen, there are 500 negative challenges for the laboratory (100 specimens with five challenges per specimen), so the probability of the laboratory declaring all challenges correctly negative is $0.9999^{500}$, which equals 0.9512. That is, the probability of one or more false positives is 0.0488 (1.0 – 0.9512).

---

14 *See* 59 Fed. Reg. at 29927, Subpart C, 3.19.

15 *See id.* at 29924, Subpart B, 2.5(d).

Thus, the probability that the higher-volume laboratory will report one or more false positives (0.0488) is nearly five times as high as the probability for the lower-volume laboratory (0.0100). Therefore, the higher-volume laboratory will have a much greater probability (than the lower-volume laboratory) of being disqualified for reporting a false positive during the initial certification process, and, subsequent to certification, it will have a much greater probability of reporting a false positive during routine proficiency testing.

Moreover, because the goal of drug testing is to correctly classify *individuals* according to the presence or absence of drugs in their specimens, the false positive rate for a laboratory does not provide information relative to the probability of correctly classifying an individual's specimen. Both hypothetical laboratories have equal probabilities (0.9995) of correctly classifying a blank specimen when testing for five drugs, meaning that both laboratories have equal probabilities (0.0005) of one or more false positives for a given *specimen*. However, the probabilities of one or more false positives for the lower-volume *laboratory* (0.0100) and for the higher-volume *laboratory* (0.0488) are significantly greater than the probability for a given *specimen* (0.0005).

This problem becomes even more significant for lower specificity rates. For example, if the specificity in the above example were 0.9990 rather than 0.9999, the lower-volume laboratory would have a probability of 0.0952 of reporting one or more false positives, versus 0.3936 for the higher-volume laboratory.

### False Negatives and the DHHS Guidelines

The DHHS guidelines also require that a laboratory seeking initial certification "correctly identify and confirm 90 percent of the total drug challenges" over the aggregate three cycles of proficiency testing required for *initial* certification.[16] Should a certified laboratory involved in ongoing proficiency testing fail to satisfy a similar 90% requirement over the span of two consecutive cycles during the required four cycles of proficiency testing in a given year, suspension or revocation of certification may result.[17]

16 *Id.* at 29927, Subpart C, 3.19(a)(2).

17 *See id.* at 29928, Subpart C, 3.19(b)(2).

The specific wording of these guidelines, requiring the identification *and* confirmation of 90% of the total drug challenges, indicates that negative challenges are not included. The DHHS guidelines indicate that a screening test should be conducted "to eliminate 'negative' urine specimens from further consideration and to identify the presumptively positive specimens that require confirmation or further testing."[18] Since negative test results are not confirmed, the only challenges subject to confirmation are those that yield positive results on the screening test: that is, either true positives or false positives. Because false positives are considered separately under the DHHS guidelines (as discussed in the previous section), only true positives (that is, positive results yielded by positive challenges) could actually be confirmed. Therefore, the DHHS guidelines are equivalent to a requirement that the laboratory must identify and confirm 90% of the *positive* drug challenges; that is, the DHHS term "total drug challenges" does not include negative challenges. This requirement is important because mixing positive and negative challenges among "total challenges" leads to more significant problems.[19]

Unfortunately, the more narrow (and more appropriate) DHHS focus on positive challenges also creates problems because the 90% requirement is implicitly based on a *ratio of sums*. We will consider the implications of ratios of sums relative to single-laboratory and multiple-laboratory test results.

*Implications Relative to a Single Laboratory*

Assume we are testing for the presence of two drugs (recall that, at a minimum, random tests must test for marijuana and cocaine). Further, suppose that the true sensitivity of the laboratory in detecting the first drug is 0.95, and the true sensitivity in detecting the second drug is 0.85 (where sensitivity is the probability of a positive test result, given that the drug is in the specimen). For the first drug, assume there are twenty positive challenges, and, based on a sensitivity rate of 0.95, nineteen are identified as positives; that is, there are nineteen (20 x 0.95) true positives. Similarly, for the second drug, assume there are 40 positive challenges and 34 (40 x 0.85) are identified as true positives. If

18 *Id.* at 29917, Subpart A, 1.2.

19 *See generally* Barnum & Gleason, *supra* note 8.

we use the ratio of sums method to calculate the ability of the laboratory to detect positive challenges, we have (19+34)/(20+40) = 0.88; that is, 88% of the positive challenges are identified. Recall that the DHHS guidelines require that at least 90% of the challenges be identified, which in this example requires the identification of at least 54 (60 x 0.9) challenges, rather than the 53 that are identified.

Instead, assume we have 40 positive challenges for the first drug and obtain 38 (40 x 0.95) true positives, and 20 positive challenges for the second drug and obtain 17 (20 x 0.85) true positives. Note that the true sensitivity of the laboratory in detecting each drug is the same as in the previous example, and that there is the same total number of positive challenges (60) as in the previous example. If we use the ratio of sums method to calculate the ability of the laboratory to detect positive challenges, we have (38+17)/(40+20) = 0.92; that is, 92% of the positive challenges are identified. In this case, the DHHS 90% guideline is satisfied.

Thus, the drug identification rate changes from 88% to 92% solely because of the change in the relative number of challenges for each drug. Therefore, the ability of a laboratory to satisfy the 90% DHHS requirement is dependent upon the relative mix of challenges for each drug.

*Implications Relative to Multiple Laboratories*

Consider two laboratories, A and B, each of which is testing for the presence of two drugs. Suppose that the true sensitivity of Laboratory A in detecting the first drug is 0.96, and the true sensitivity in detecting the second drug is 0.87. For the first drug, suppose there are 100 positive challenges which yield 96 (100 x 0.96) true positives; for the second drug, suppose there are 200 positive challenges which yield 174 (200 x 0.87) true positives. If we use the ratio of sums method to calculate the ability of the laboratory to detect positive challenges, we have (96+174)/(100+200) = 0.90.

Now suppose the true sensitivity of Laboratory B in detecting the first drug is 0.95, and the true sensitivity for the second drug is 0.85. For the first drug, suppose there are 200 positive challenges which yield 190 (200 x 0.95) true positives; for the second drug, suppose there are 100 positive challenges which yield 85 (100 x 0.85) true positives. Note

that the total number of positive challenges (300) is identical to that for Laboratory A. If we use the ratio of sums method to calculate the ability of Laboratory B to detect positive challenges, we have (190+85)/(200+100) = 0.92. Thus, the performance of Laboratory B appears to be better than Laboratory A, even though the ability of Laboratory A exceeds the ability of Laboratory B to detect the presence of each drug (0.96 vs. 0.95, and 0.87 vs. 0.85). This result occurs simply because of the relative mix of the 300 positive challenges for the two different drugs.

A single change in this example leads to a more troublesome result. Suppose the true sensitivity of Laboratory A in testing for the second drug deteriorates from 0.87 to 0.86. Again, the ability of Laboratory A to detect the presence of each drug exceeds the ability of Laboratory B. However, while Laboratory B satisfies the DHHS 90% requirement, Laboratory A does not satisfy the requirement (it will have identified only 89% of the positive drug challenges).

## Conclusions and Recommendations

Drug-testing programs involve a variety of potential risks: the personal and economic costs associated with denying or terminating employment on the basis of faulty test results, legal pitfalls ranging from the need to satisfy federally-mandated scientific standards regarding random testing, and issues related to tort liability. To reduce the risk of inaccurate testing processes, the U.S. government has taken significant steps to ensure that drug-testing laboratories maintain extremely high levels of accuracy and also has set rigorous standards to ensure exemplary laboratory performance. The analysis herein focuses on only one question: Whether indicators used to measure accuracy are flawed in ways that result in some drug-testing laboratories being held to higher standards than others.

The primary conclusion of the Performance Indicator Analysis is that the proficiency criteria mandated in the DHHS guidelines can result in differing treatment of laboratories in proficiency tests for both false positives and false negatives. Laboratories with identical levels of performance in avoiding false positives (that is, laboratories with identical specificity levels) can yield different false-positive results

under the DHHS proficiency-testing process. Similarly, a laboratory with a superior level of performance in avoiding false negatives (that is, a laboratory that has a superior sensitivity level in comparison to another laboratory) can yield poorer false-negative performance under the DHHS proficiency-testing process. These aberrations are a consequence of the *laboratory* focus of the DHHS guidelines. The laboratory-focused guidelines fail to appropriately measure the true issue of concern in proficiency testing: the ability of the laboratory to correctly identify an individual *specimen.*

These results have several implications. A higher-volume laboratory may have a greater probability of being disqualified for reporting a false positive than a lower-volume laboratory during the initial DHHS certification process. Moreover, subsequent to certification, a higher-volume laboratory may have a greater probability of reporting a false positive during routine proficiency testing. Thus, such a laboratory may have to undertake corrective actions that would not be required of a lower-volume laboratory with an identical level of proficiency in avoiding false positives.

On the other hand, with respect to properly identifying true positives in tests for multiple drugs, the ability of a laboratory to satisfy the 90% DHHS requirement is dependent upon the relative mix of challenges for each drug because of the implication of ratios of sums. An extension of this complication, in tests of more than one laboratory, is that a superior-performing laboratory can yield a poorer proficiency result than an inferior-performing laboratory. In such cases, these misleading outcomes may result in demands for corrective actions that are not warranted.

The problems with the laboratory-focused DHHS guidelines are further exacerbated by the regulations developed by various federal agencies in their attempt to implement the guidelines. As a result, implementation processes yield even more complications relative to the true purpose of proficiency testing: to ensure that individual specimens are correctly identified.

Based on our findings, we recommend that all drug-testing accuracy indicators used in proficiency-testing processes required by the DHHS, or by other federal agencies, be subjected to PIA. Measures

that incorporate potential biases should be replaced with measures that are free of such errors. Indicators that do not use the individual specimen as the unit of analysis are especially likely to reflect biases, and they should be the first to be examined. But even indicators based on individual specimens can contain criterion errors of the types we have identified; therefore, they should not be exempt from scrutiny.