Accuracy in Transit Drug Testing: A Probabilistic Analysis

DAROLD T. BARNUM AND JOHN M. GLEASON

The accuracy of drug tests for transit and other transportation workers is an important subject. Concepts of accuracy that are most relevant are applied to laboratory proficiency and transportation worker drug-usage data, and accuracy levels that could occur in transit drug testing programs are estimated. Seemingly accurate tests for abused drugs can sometimes be inaccurate to a disturbingly high degree. Therefore, a methodology by which decision makers can correctly set desired accuracy standards for their organizations is suggested to avoid potential inaccuracies.

Substance abuse by transportation workers has become a major concern to public officials, transportation agencies, and unions. As a result, there has been a rapid increase in the testing of workers for the presence of illegal substances. Recent federal requirements for the testing of transportation employees has accelerated the trend (1–5).

The use of drug tests is based on the premise that they correctly identify both the presence and the absence of certain drugs. If the tests do not identify most drug users, then they are not an effective method for apprehending abusers or encouraging abstinence. Even worse, if nonusers test positive for drugs, they may be marked for life for an offense they did not commit. Our legal system, as well as our system of workplace jurisprudence, requires that a person be assumed innocent until proven guilty with compelling evidence. Thus, drug testing should be used only when the tests can correctly categorize users and nonusers, with the categorization of nonusers being especially important.

The U.S. Department of Transportation (DOT) has recognized the crucial importance of correct test results and requires that rigorous accuracy standards be met by drug testing laboratories, with special emphasis on false positive errors. A laboratory’s certification is subject to review and revocation unless the accuracy requirements are met (4).

The accuracy of drug tests for transit and other transportation workers is therefore an important subject. In the following sections, concepts of accuracy that are most relevant are reviewed and applied to laboratory proficiency and transportation drug use data, and accuracy levels that could occur in transit drug testing programs are estimated. Seemingly accurate tests for abused drugs may sometimes be inaccurate to a disturbingly high degree. Therefore, a methodology is suggested by which decision makers can correctly set desired accuracy standards for their organizations, in order to avoid potential inaccuracies.

TERMINOLOGY

To clarify the terminology and concepts used in this paper, the following discussion is offered. Excellent discussions from somewhat differing viewpoints may be found in Quality Assurance in Drug-Use Testing (6) and Drug Testing in the Workplace (7).

When a specimen is tested for drugs, one of four outcomes must occur:

- True positive—specimen with drugs tests positive for drugs,
- True negative—specimen with no drugs tests negative for drugs,
- False positive—specimen with no drugs tests positive for drugs, or
- False negative—specimen with drugs tests negative for drugs.

Given that a specimen contains drugs, it must test either (true) positive or (false) negative. That is, the probability of a positive test result (given drugs are present) plus the probability of a negative test result (given drugs are present) must equal 1.0. This equation may be written

\[ P(+) + P(-) = 1.0 \]  \hspace{1cm} (1)

In other words, when drugs are present in a specimen, the probabilities of a true positive and a false negative must total 1.0.

Similarly, a specimen without drugs must test either (false) positive or (true) negative. This equation may be written

\[ P(+) + P(-) = 1.0 \]  \hspace{1cm} (2)

Thus, the probability of a positive test result (given no drugs are present) plus the probability of a negative test result (given no drugs are present) equals 1.0.

Three measures of drug test accuracy are used in the health-related professions: sensitivity, specificity, and predictive value. Sensitivity is the probability that a specimen with drugs will test positive. Thus, it equals the probability of a true positive, i.e.,

\[ \text{Sensitivity} = P(+) \]  \hspace{1cm} (3)

A second concept used in evaluating the accuracy of laboratory tests is specificity. Specificity is the probability that a drug-free specimen will test negative, i.e., the probability of a true negative.

\[ \text{Specificity} = P(-) \]  \hspace{1cm} (4)
Thus, sensitivity measures the ability of the test to correctly report the presence of drugs, whereas specificity measures the ability of the test to correctly report the absence of drugs. For an ideal test, sensitivity and specificity would both be equal to 1.0, meaning that every drugged specimen tests positive and every nondrugged specimen tests negative.

In drug testing, the most important concern is incorrect results, that is, occurrence of false positives and false negatives. Sensitivity and specificity are indirect indicators of the false result rates. $P(+|\text{no drugs})$, the probability of obtaining a false positive, is equal to $(1 - \text{specificity})$. $P(-|\text{drugs})$, the probability of obtaining a false negative, is equal to $(1 - \text{sensitivity})$. Thus, the higher the sensitivity, the lower the false positive rate. The higher the specificity, the lower the false negative rate. What laboratories should do, therefore, is maximize sensitivity and specificity, because this will minimize the false negative and false positive rates.

An important concept, although not widely used in studies measuring laboratory proficiency, is the positive predictive value (PPV) of a test. For drug tests, the PPV is the probability that the drug is present in a specimen, given that the test yielded a positive result. That is,

$$\text{PPV} = P(\text{drugs}|+)$$

For example, if 90 out of every 100 people testing positive for drugs have truly taken drugs, then the PPV of the test is 0.90. The probability that a person with a positive test result truly has not taken drugs is 0.10. That is, if a drug test has a PPV of $X$, then the probability is $(1 - X)$ that a person testing positive is drug free.

Thus, maximizing the PPV of a test also minimizes the probability that specimens testing positive are truly drug free. The latter probability is the false conviction rate. That is,

$$\text{False conviction rate} = P(\text{no drugs}|+) = 1 - \text{PPV}$$

This concept is important in determining whether a positive result on a drug test provides sufficient evidence of drug usage. If, for example, positive results on a test are known to be untrue in 1 out of every 10 cases, i.e., the false conviction rate = 0.1, then a positive test would not be considered sufficient evidence to convict a person of drug use. More important, in protecting the innocent from false accusation the PPV (or, equivalently, the false conviction rate) of the test is of prime concern.

Specificity, sensitivity, and predictive value may be stated as either probabilities or percentages (6–8). Both are used here, as the context dictates.

**DRUG ABUSE BY TRANSPORTATION WORKERS**

Not surprisingly, estimates of drug usage vary widely. Because using drugs of abuse is usually illegal but often considered to be in style, self-reports may not be very reliable. Similarly, those making pronouncements about the extent of drug abuse often have strong incentives for claiming that drug usage is either very high or very low. Finally, the characteristics of the group about which the estimates are made can cause wide variations, because drug usage varies on the basis of age, geographic location, and other variables. Considering these caveats, several estimates of the extent of drug abuse are provided.

In late 1989, the National Institute on Drug Abuse (NIDA) estimated, on the basis of self reports, that 7.1 percent of all Americans over 18 used some illegal substance at least once during the month before their 1988 household survey, although usage dropped to 2.1 percent for people over 35 years old (9,10).

A 1989 U.S. Department of Labor report (11) based its estimates on the results of employer drug tests. For all of the industries surveyed in 1988, 8.8 percent of the employees tested positive, with the workforces of the smallest firms tending to have the lowest positive rates. For the transportation industry, 5.6 percent of the employees tested positive. (Rates in other industries ranged from 20.2 percent in wholesale trades to 3.1 percent in services; transportation's rate was the second lowest of the industries reported.)

In 1986 and 1987, tests of railroad workers involved in train accidents revealed that 29 of 759 people, or 3.8 percent, tested positive for drugs of abuse (12).

Estimates are also available for one of the nation's largest transit properties. In 1988, 2.74 percent of its employees tested for drugs tested positive, as did 2.58 percent of those tested in early 1989 (13).

These last three estimates were based not on random drug tests primarily, but on tests often administered to employees under conditions where one would expect drug usage to be higher than usual. Thus, they may overestimate the extent of drug usage for the three workforce groups involved.

In a final estimate, DOT randomly tested more than 16,000 of its own employees between July 1, 1988, and June 30, 1989. Of these, 99 tested positive, for a rate of 0.619 percent (14).

On the basis of these estimates, transit employee drug usage may be in the range of 0.6 to 5.6 percent. Typically, transit systems in the larger cities and with younger workers would have higher rates, whereas systems in smaller towns and with older workforces would have lower ones. Allowing that some systems will have higher or lower rates, for the nation as a whole the transit average is probably in the 0.6- to 5.6-percent range. Consequently, rates from this range are used in the analyses in this paper. (When this statistical methodology is used for a specific system, then the estimated drug abuse rate for that particular system should be used.)

**ACCURACY OF DRUG TESTING**

A topic no less subject to disagreement than the proportion of workers using drugs is the accuracy of the tests for such substances. Here also, estimates of accuracy often seem to be based more on the self-interest of the claimants than on sound empirical evidence. Additionally, a variety of factors, such as characteristics of the laboratories involved, the testing methods used, the testing protocols actually followed, whether proficiency tests are open or blind, concentrations of the drugs in spiked challenges, and other influences can cause large variations in accuracy.

Because of the concerns over whether this testing correctly identifies the presence or absence of drugs, a number of lab-
ors of the 1988 JAMA study (16) did conduct confirmation testing of specimens that initially screened positive, and the most frequently used confirmation method was GC/MS.

More specifically, the 1985 study (15) did not report whether its laboratories conducted confirmation tests on specimens reported positive by the initial screening test. The 1988 study said that during the open phase, specimens screening positive were subjected to confirmation testing. The most frequently used confirmation method was GC/MS. GC/MS was always used for confirmations by about half of the laboratories and sometimes used for confirmations by about another one-sixth.

The remaining confirmations used either high-performance thin-layer chromatography or high-performance liquid chromatography. The authors of the 1988 article (16) were uncertain whether confirmations took place for all specimens screening positive in the blind portion of the study. However, there is evidence to suggest that such confirmations did occur.

The false positive rates for the open and blind phases were almost identical, each rounding to a rate of 0.002 (the rate during the open phase being slightly higher). Because the false positive rates were virtually the same for both the open and blind phases, it is likely that the same confirmation procedures were used for each. Because it is certain that confirmations took place during the open phase, when the laboratories knew precisely which specimens were being used for the test, the rate of 0.002 appears to be the applicable false positive rate when confirmation is required.

GC/MS is generally considered to be the confirmation method that can theoretically produce the most accurate results. However, the fact that it was not used in every case in the 1988 JAMA blind study (16) does not necessarily mean that the resulting specificity estimates were lower (or, equivalently, that the false positive rates were higher) than they would have been if GC/MS had been the sole confirmation method used. GC/MS requires a higher level of skill and care than some of the other methods. These requirements may not always be met in the routine testing that blind proficiency studies attempt to duplicate, so the theoretical advantage of GC/MS may be lost. Although confirmation with GC/MS theoretically makes misidentification less likely, mistakes are possible (7). In short, no empirical evidence proves that blind proficiency studies of laboratories using only GC/MS for confirmations would yield higher specificity (that is, lower false positive rates) than did the laboratories in the 1988 JAMA study (16).

Moreover, there have been no independently conducted blind proficiency studies of the performance of laboratories certified by NIDA under actual transit agency testing conditions. When such data are available and applicable, the resulting sensitivity and specificity rates should be used. The mere fact that UMTA-required standards might in theory result in more accurate tests is not sufficient to prove that they do so in practice. The Navy also had high standards, yet there was substantial inaccuracy, because the standards were not initially enforced adequately in practice (2,19). Instances of high theoretical standards and lax enforcement in transportation drug testing have already occurred (14,20,21).

Because it seems that the two JAMA articles (15,16) contain the most relevant published averages to date, they are used here for illustrative purposes. The 1988 article (16) is more representative of recent experience than the 1985 article.
(15), and it is closer to what can be expected of NIDA-certified laboratories. However, many laboratories today are similar to those reported on in the 1985 study, and there may be laboratories that are even less accurate (22).

If the accuracy of a laboratory used by a particular transit agency is claimed to be higher or lower than those in the JAMA studies, the burden of proof should be on the claimant to provide supporting empirical evidence. That is, because laboratories differ in their sensitivity and specificity rates, individual transit properties should obtain blind study estimates for the laboratories they use in order to estimate the true accuracy of their own results. This should be done whether or not theoretically uniform transit standards exist.

Based on calculations from data presented for the blind phases of the JAMA proficiency studies, the false positive rate (false positives/negative challenges) was 0.014 in the 1985 article (15) and 0.002 in the 1988 article (16), representing findings on the proportions of drugless samples where drugs were incorrectly reported to be present. (A negative challenge is a specimen that does not contain a drug being tested for by the study. These statistics are, of course, estimates of $P(+|\neg\text{drugs})$ and are equivalent to specificity levels of 98.6 and 99.8 percent, respectively.

The authors of the 1988 article (16) calculated the false positive rates by a different method, thus reporting false positive rates of 0.013 for the blind phase and 0.016 for the open phase of their study, as compared to the blind and open rates of 0.0018 and 0.0019 calculated here. The false positive rates reported in the 1988 study (16) are about 10 times higher than those resulting from these calculations.

The false negative rates (false negatives/positive challenges) were 0.618 in the 1985 study (15) and 0.311 in the 1988 study (16). These statistics estimate the value of $P(\neg|\text{drugs})$ and reflect sensitivity levels of 38.2 and 68.9 percent, respectively. (A positive challenge is a specimen that does contain a drug being tested for by the study.)

The results of open tests, those involving urine samples that laboratories knew were being used for quality checks, are not used herein. Results of blind tests, those in which the laboratories did not know they were being tested, are much more representative of what one could typically expect when actual urine samples are submitted. Therefore, only blind test results were used for the research presented here.

**BAYESIAN ANALYSIS**

Most employers do not want to accuse workers of drug usage if there is reasonable doubt about their guilt. However, if 99 percent of drug-free employees test negative, then many employers would conclude that reasonable doubt does not exist when an employee tests positive. Because the percentages of the drug-free specimens testing negative in the JAMA studies were 98.6 and 99.8 percent, it would appear that drug tests are accurate enough that reasonable doubt could not be established.

However, the results are not what they seem, as Bayesian analysis of AIDS test and drug test results have exhibited (23–25). The first case is described in Table 1.

As shown for Case 1 in Columns 2 and 3 of the table, a urine specimen must either contain drugs ($S_+)$ or contain no drugs ($S_-$). For this case, it is further assumed that the probability is 0.056 that the urine specimens truly contain drugs, requiring that the probability is 0.944 that they do not, as shown in Column 4. These probabilities imply that 5.6 percent of the target population uses drugs; recall that this is the rate estimated for all transportation workers in the U.S. Department of Labor survey. The next column, Column 5, identifies the probability of the urine specimen's testing positive for drugs when there truly are drugs present (0.382), and when there truly are no drugs in the sample (0.014). That is, $P(+|\text{drugs}) = 0.014,$ and $P(+|\neg\text{drugs}) = 0.382.$ These probabilities are taken from the 1985 JAMA study (15).

The numbers in Column 6 are the products of the numbers in the two previous columns. That is, for the population being tested, the probability that a person truly is on drugs and tests positive for drugs is 0.021392, whereas the probability that a person truly is not on drugs and tests positive for drugs is 0.013216. The sum of these two probabilities, denoted by $P(+)$ and equal to 0.034608, is the probability of a positive test result.

Dividing each of the numbers in Column 6 by $P(+)$ yields the numbers in Column 7, which are the probabilities of being in the particular states, given a positive test result. Thus, the probability that specimens that test positive will contain drugs is 0.618, meaning the test has a PPV of 61.8 percent. The probability that specimens testing positive will truly contain no drugs is 0.382, meaning the test has a false conviction rate of 38.2 percent. That is, $P(\text{drugs}+|+) = 0.618,$ and $P(\neg\text{drugs}+|+) = 0.382,$ with the two probabilities totaling 1.

These same results can be developed more intuitively by considering a group of 1,000 workers who are tested for drug usage. If 5.6 percent of the group are truly taking drugs, then 56 workers will provide urine specimens that contain drugs, and the remaining 944 will provide specimens that are drug free. Of the specimens containing drugs, 56(0.382) = 21 will test positive for drugs, and the remaining 35 will test negative. Similarly, of the specimens not containing drugs, 944(0.014) = 13 will test positive for drugs, and the remaining 931 will test negative. Thus, $21 + 13 = 34$ specimens will test positive for drug usage, although 13 of these 34 do not actually contain drugs. That is, $13/34$ (or 38.2 percent) of those testing positive for drug usage will truly be drug free.

Thus, almost two out of every five workers testing positive will truly be drug free. With probabilities such as these, it is highly unlikely that a positive drug test would provide a preponderance of evidence that an individual was taking drugs, let alone meet higher levels of proof, such as clear and convincing evidence or evidence beyond a reasonable doubt. Not only would employers lose arbitration or court cases with such meager evidence, it would seem illogical, from the standpoint of good personnel practice, to dismiss or discipline employees with such unreliable evidence.

Of course, the actual probability that a person who tests positive is not on drugs will differ under different assumptions about (a) the percentage of the target population that is actually taking drugs, (b) test specificity, and (c) test sensitivity. Different assumptions are presented in Cases 2 through 8.

In Cases 2 through 4, the estimates of the proportion of the target population on drugs were varied. Whereas Case 1 used the rate (0.056) estimated for all transportation workers, Case 2 used the rate (0.038) estimated for railroad workers.
TABLE 1  APPLICATION OF BAYESIAN ANALYSIS TO TRANSPORTATION DATA

| CASE | STATE | S<sub>j</sub> | P(S<sub>j</sub>) | P(+) | P(+|S<sub>j</sub>)P(S<sub>j</sub>) | P(S<sub>j</sub>|+) |
|------|-------|---------------|----------------|------|-------------------------------|----------------|
| 1    | Drugs | S<sub>1</sub> | 0.056          | 0.382 | 0.021392                      | 0.618          |
|      | No Drugs | S<sub>2</sub> | 0.944          | 0.014 | 0.013216                      | 0.382          |
|      | Total |               | 1.000          |       | P(+)=0.034608                 | 1.000          |
| 2    | Drugs | S<sub>1</sub> | 0.038          | 0.382 | 0.014516                      | 0.519          |
|      | No Drugs | S<sub>2</sub> | 0.962          | 0.014 | 0.013468                      | 0.481          |
|      | Total |               | 1.000          |       | P(+)=0.027984                 | 1.000          |
| 3    | Drugs | S<sub>1</sub> | 0.026          | 0.382 | 0.009932                      | 0.421          |
|      | No Drugs | S<sub>2</sub> | 0.974          | 0.014 | 0.013636                      | 0.579          |
|      | Total |               | 1.000          |       | P(+)=0.023568                 | 1.000          |
| 4    | Drugs | S<sub>1</sub> | 0.006          | 0.382 | 0.002292                      | 0.141          |
|      | No Drugs | S<sub>2</sub> | 0.994          | 0.014 | 0.013916                      | 0.859          |
|      | Total |               | 1.000          |       | P(+)=0.016208                 | 1.000          |
| 5    | Drugs | S<sub>1</sub> | 0.056          | 0.689 | 0.038584                      | 0.953          |
|      | No Drugs | S<sub>2</sub> | 0.944          | 0.002 | 0.001888                      | 0.047          |
|      | Total |               | 1.000          |       | P(+)=0.040472                 | 1.000          |
| 6    | Drugs | S<sub>1</sub> | 0.038          | 0.689 | 0.026182                      | 0.932          |
|      | No Drugs | S<sub>2</sub> | 0.962          | 0.002 | 0.001924                      | 0.068          |
|      | Total |               | 1.000          |       | P(+)=0.028106                 | 1.000          |
| 7    | Drugs | S<sub>1</sub> | 0.026          | 0.689 | 0.017914                      | 0.902          |
|      | No Drugs | S<sub>2</sub> | 0.974          | 0.002 | 0.001948                      | 0.098          |
|      | Total |               | 1.000          |       | P(+)=0.019862                 | 1.000          |
| 8    | Drugs | S<sub>1</sub> | 0.006          | 0.689 | 0.004134                      | 0.675          |
|      | No Drugs | S<sub>2</sub> | 0.994          | 0.002 | 0.001988                      | 0.325          |
|      | Total |               | 1.000          |       | P(+)=0.006122                 | 1.000          |

Involving in accidents, Case 3 used the most recent rate (0.026) found at the large urban transit system, and Case 4 used the rate (0.006) found as a result of the random tests of its workers by DOT. This variation in assumed rate provides an indication of the sensitivity of the final outcome to changes in the rate of drug usage in the target population.

In Cases 5 through 8, the same four rates of drug usage were repeated, but the 1988 JAMA article (16) estimates of sensitivity and specificity were substituted. These estimates were 68.9 and 99.8 percent (which represent a false negative rate of 0.311 and a false positive rate of 0.002, respectively).

Thus, for Cases 1 through 4, about two out of every five positives will represent those who have not taken drugs when all transportation workers are considered, one out of every two positives will represent drugless employees for railroad workers in accidents, three out of every five positives will represent drugless employees for workers at the large transit system, and six out of every seven positives will represent drugless employees for the population of transportation employees being tested by DOT. In other words, for these four cases the false conviction rates are 38, 48, 58, and 86 percent, respectively. Under any of these circumstances, such test results have no value in proving drug use.

The error rates are lower for Cases 5 through 8, where sensitivity and specificity rates were drawn from the 1988 JAMA article (16). In these four scenarios, approximately 1 out of every 20 transportation employees who tested positive would be falsely accused, as would approximately 1 out of every 15 railroad employees, 1 out of every 10 of the large transit system employees, and 1 out of every 3 DOT employees. These represent false conviction rates of 5, 7, 10, and 33 percent, respectively. Although these results are better than those for the first four cases, it seems that convicting even 5 percent falsely, with such serious consequences, is an extremely high error rate.

Very importantly, these false conviction rates are based on our estimated average rates for drug usage, sensitivity, and specificity and are not necessarily applicable to any particular transit agency. But all of our estimated rates are ones that could occur in some circumstances. Because of the extremely serious consequences of being convicted of drug use, an employer would be wise to determine that these estimated rates, or similar rates, do not apply before acting on positive drug test results.

The situation would be improved if sensitivity and specificity levels could be increased. Another alternative is to test
only those groups with high rates of drug use, and to test only in those situations where sufficient accuracy can be obtained. One way to do this is examined next.

**METHODOLOGY FOR DEVELOPING ACCEPTABLE PREDICTIVE PROBABILITIES**

Once a decision maker has estimated the percentage of the target population taking drugs and has determined the acceptable percentage of those testing positive but not on drugs, that is $P(\text{no drugs}|+)$, then he or she can identify various combinations of specificity and sensitivity rates that will yield the acceptable percentage of those testing positive but not on drugs. Then, by conducting blind tests of the laboratories being used, it would be possible to determine whether the sensitivity and specificity requirements are actually being met. This methodology is discussed in this section.

First, assume that out of every 100 samples that test positive for drugs, a maximum of 1 should be truly drug free, $P(\text{no drugs}|+) \leq 0.01$. That is, the minimum PPV of the test would be 99 percent. These standards require that out of every 100 workers testing positive for drug usage, at least 99 of them will indeed have drugs in their systems, but 1 at the most will not. Thus, the false conviction rate will be no more than 1 employee out of every 100 testing positive, or 1 percent, because the ratio $(\text{drug users})/(\text{positive test results}) \geq 99/100$.

Some might argue that this probability is too high, and others may feel that it is too low. Whatever level is chosen could be determined by collective bargaining in unionized situations, unilaterally by top decision makers, or perhaps set by DOT decree. However, it is the PPV, not the sensitivity or specificity rates, that must be set correctly. Any PPV desired could be substituted for the one used here.

Second, it is necessary to estimate the proportion of the target population that is using drugs. For purposes of this example, 3.0 percent of the employees are assumed to have drugs in their systems. This parameter is slightly above the usage rate found at the large transit agency described earlier, but it could be replaced by any appropriate percentage.

Once the drug usage rate and the desired PPV have been established, it is possible to calculate the combinations of sensitivity and specificity levels that will be acceptable. For this example, it is estimated that $P(\text{drugs}) = 0.03$, and it is required that $P(\text{no drugs}|+) \leq 0.01$. Using these values in the appropriate Bayesian formula and simplifying yields

$$P(-|\text{no drugs}) = 1 - [0.0003124\cdot P(+|\text{drugs})]$$

Now, the different combinations of specificity $P(-|\text{no drugs})$ and sensitivity $P(+|\text{drugs})$ that satisfy the equation can be calculated.

As the formula indicates, the lower the sensitivity, the higher must be the specificity. The acceptable sensitivity and specificity combinations are shown in Figure 1. Any combination on the line will exactly satisfy the requirements, any combination above the line will more than satisfy them, and any combination below the line (in the shaded area) will not satisfy them.

When the sensitivity in Figure 1 is zero, meaning none of the true drug users are identified, the required specificity is at its highest level, 1.00 (i.e., 100 percent). That is, if the probability of a false negative given drug usage is 1.00, then there will be no true positives, and so sensitivity is zero. Hence, there must be no false positives in order to meet the standard that no more than 1 out of 100 total positives is false.

When sensitivity is 1.0, meaning that all tested drug users are identified as such, then specificity must be at least 0.9997 (or 99.97 percent). This figure represents a false positive rate of $(1 - 0.9997) = 0.0003$, meaning that even with perfect sensitivity, no more than 3 out of every 10,000 negative specimens should result in positive test results. Such precise results are unlikely to occur in practice, given the current state of technology. Even under the extremely favorable conditions used in the AACC study (18), 7 out of every 10,000 negative challenges would result in positive test results, a ratio more than twice the acceptable level.

It is also instructive to consider the case at DOT, where 0.006 of the more than 16,000 specimens taken from July 1988 through June 1989 tested positive. Again assuming a desired

![FIGURE 1 Sensitivity-specificity levels (drug use = 3 percent, false conviction rate = 1 percent).](image)
false conviction rate of no more than 1 percent, and substituting the parameters into the appropriate Bayesian equation, the following calculation is performed:

\[ P(\neg \text{drugs}) = 1 - [0.0000609 - P(\text{drugs})] \]  

(8)

Assume a sensitivity rate of 90 percent, that is \( P(\text{drugs}) = 0.90 \), the minimum required by DOT regulations. Substituting this into the equation yields a specificity \( P(\neg \text{drugs}) \) of 0.999945127. The maximum false positive rate, therefore, is \( P(\neg \text{drugs}) = 1 - 0.999945127 = 0.000054873 \). This result means that there can be no more than 1 reported positive out of every 18,224 truly negative specimens. As this relates to quality control testing, out of every 18,224 blind negative challenges submitted, the laboratories could report no more than 1 positive without violating the desired standards. On the basis of the laboratory accuracy studies observed to date, it would be impossible for most laboratories to be this accurate. Also, given that for the July 1988 through June 1989 period DOT submitted only 79 blind negative challenges at the most (14), there is little evidence from DOT’s experience that the 1 out of 18,224 ratio has been satisfied.

On the basis of currently published data, therefore, if 3 percent or less of the target population uses drugs, it would be difficult to maintain false negative and false positive rates at levels low enough to ensure that no more than 1 out of every 100 people testing positive is falsely accused.

These results do not mean, however, that drug testing should never be used if one wishes a false conviction rate of no more than 1 out of every 100 positive tests. When the proportion of the population using drugs is higher than 3 percent, then the required sensitivity-specificity combinations will be lower than those shown in Figure 1 (except when sensitivity equals zero).

One likely situation in which substantially more than 3 percent of the target population may be using drugs is when testing is done only for reasonable cause, rather than on a random or universal basis. If supervisors and others are well trained in the signs of drug use, then workers that they refer for testing should be much more likely to be on drugs than those in the workforce as a whole (17). Testing for reasonable cause has the added benefit that a direct link between the drug use and performance can be established, providing a much stronger case for discipline than when drug usage is found but no decline in performance can be shown.

Assume, for example, that 50 percent of those individuals tested for cause are actually taking drugs, which is the percentage of DOT employees tested for cause between September 1987 and August 1989 that tested positive (26). Figure 2 shows the various sensitivity and specificity combinations that meet the requirements for this situation.

As before, any point on the line in Figure 2 meets the requirements exactly, whereas any point above the line exceeds the requirements. Although the requirements are still high, they are much lower than before and are such that they could be met by laboratories such as those represented in the 1988 JAMA study (16). For example, given that 50 percent of those tested are truly on drugs, let us assume a required sensitivity of 90 percent, the level required by DOT regulations (4). When this sensitivity is achieved, the specificity would only have to be 99.091 percent. This level is within the capacity of the 1988 JAMA study (16) laboratories, which achieved an average specificity of 99.8 percent. Thus, when testing is done for cause, using first-rate laboratories and a rigorous blind proficiency evaluation procedure, drug tests can indeed provide reasonable protection from the false conviction of the innocent.

Assuming laboratory sensitivity and specificity levels at the rates identified in the 1988 JAMA article (16), the desired protection against false convictions could be obtained in universal or random testing by use of a two-stage testing procedure. Under such a procedure, no action would be taken against someone testing positive for the first time, except that the individual would be tested again. If the results of the second test were also positive for the same drugs, then it could be assumed that the person is truly taking the drugs, and normal actions for drug usage could be taken.

The rationale is as follows. Assume that the initial target population has a drug usage rate between 0.006 and 0.056, and the sensitivity and specificity on the initial tests are 68.9
and 99.8 percent, which are the assumptions used in Cases 5 through 8 in Table 1. Thus, from Table 1, the percentage of those testing positive who truly are on drugs ranges from 67.5 to 95.3 percent. The second-stage test is a universal retesting of all those who initially tested positive, so the second-stage target population has a drug usage rate of between 0.675 and 0.953. To be extremely conservative, the usage rate will be set at 0.50. Then, this second-stage target population has the same usage rate as that illustrated for a target population of individuals tested for cause, and similar results apply. (Note that the two-stage procedure assumes that the initial false positives are not the result of some underlying condition that will cause the incorrect results to recur on the second test.)

This result means that current technology is sufficient to obtain no more than 1 false positive out of every 100 positive tests if the target group has a usage rate of at least 0.50. This condition may hold when testing is done for cause or if second tests of those initially testing positive are conducted.

### CONCLUSIONS AND POLICY IMPLICATIONS OF THE STUDY

The initial conclusion of this analysis is that allegedly accurate tests for abused drugs can be inaccurate to a disturbingly high degree, under circumstances that may sometimes occur in transit. For example, consider the laboratory accuracy rates from the most favorable published blind study to date, and current drug usage rates of railroad, transit, and DOT employees. Using these averages, 1 out of every 15 positive tests of railroad workers would be false, as would be 1 out of 10 positive tests of employees at one transit agency, and 1 out of 3 positive tests of DOT employees.

Such potential inaccuracies should serve to encourage both governmental regulators and transit operators to proceed with the utmost care in implementing drug testing programs. Further, the findings demonstrate the need to alert transit operators, employees, and their unions that seemingly reliable drug tests can be inaccurate and to educate all groups concerning methods for assessing accuracy.

Moreover, it is important for policy makers to be aware of the potential problems and to consider the implications in their decision-making processes. Failure to do so creates the potential for lost lawsuits and arbitrations and could cost many innocent workers their jobs and many unwitting agencies good employees. Further, failure to do so may violate the principles of our legal and workplace jurisprudence systems by allowing convictions with insufficient evidence; this could lead to political pressure sufficient to outlaw drug testing even in justifiable cases.

It is important that the proportion of workers falsely accused decreases when a higher percentage of the population being tested is truly on drugs. If, for example, 50 percent or more of the target group is truly taking drugs, then it is possible to ensure a rate of no more than 1 false conviction out of every 100 people tested positive. Thus, if testing is limited to high usage groups, the results can provide an acceptable degree of accuracy.

One common case in which drug usage in a sample may be high would be for workers tested for reasonable cause. In DOT's testing experience for reasonable cause, for example, 50 percent of those tested have tested positive.

If random or universal testing must be used for a group with a low drug usage rate, acceptable accuracy can be obtained through the use of a two-stage process. No action would be taken concerning those who test positive for the first time, other than to retest them. Positive results for the same drugs on the second test would often provide acceptable accuracy, because the retested group would often have a drug-use rate exceeding 50 percent.

Conclusions concerning for-cause and two-stage testing assume that laboratory sensitivity and specificity rates are at least as high as the averages from the 1988 JAMA study (16), and that target populations have drug usage rates of at least 0.50. The two-stage process also assumes that errors are random and not the result of conditions that will cause them to recur for the same individuals.

Whether it is decided to test randomly, universally, or on the basis of probable cause, a Bayesian methodology should be used. Such a framework will ensure that the test results are sufficiently specific and sensitive to achieve the desired PPV. If groups with different abuse rates are being tested, then the procedure should be applied to each group separately.

DOT has not directly addressed the most important accuracy standard, that of the minimum acceptable PPV. The PPV is the percentage of all positive tests in which drugs are truly present. Thus, if DOT wishes to regulate the accuracy of test results with the objective of protecting the innocent as well as identifying the guilty, then it needs to set explicit standards for the PPV or, equivalently, for the false conviction rate, $(1 - \text{PPV})$. The false conviction rate is the percentage of all positive tests in which drugs are absent.

This work has used accuracy data under conditions when chain of custody was not a major problem and when no non-drug substances could produce positive test results, because all challenges contained only the drugs for which tests were being conducted. When such factors are a significant consideration, as they would be during actual drug tests of real workers, the reliability of the test results will usually be worse than those from laboratory proficiency studies.

Because some will quarrel with our estimates of drug use, sensitivity, and specificity rates, the primary emphasis of this paper may be obscured. The average rates believed to be the most applicable to transit drug testing were used. These rates were based on empirical evidence, from studies published in reputable sources. When more relevant average rates become available, they should be used in place of the ones used herein.

Further, regardless of whose average estimates are used, conclusions about the extent of error that are based on averages are not applicable to a particular situation. False conviction rates for a specific organization can only be estimated if the underlying Bayesian methodology is combined with data specific to that particular case.

Our main message is that appropriate rates should be used in a Bayesian framework to develop estimates of false conviction rates, and that the maximum acceptable false conviction rate should be used in setting the required combinations of sensitivity and specificity. That is, the rates for PPV, sensitivity, specificity, and drug usage form a system. The key rate is false conviction rate $(1 - \text{PPV})$, which is affected by the levels of the other three rates. Thus, using a Bayesian frame-
work one can either determine the actual false conviction rate from the other three rates or determine the acceptable combinations of sensitivity and specificity given an actual drug usage rate and the desired maximum false conviction rate.

The use of abused drugs by the nation's transit workforce is indeed a serious problem that must be addressed on many fronts. In proper circumstances, urine testing is a valuable weapon for decreasing drug use, and its use for safety-sensitive employees is recommended. However, although testing should be available, it should be used sparingly and with the utmost care. The recent rush toward routine universal drug testing may result in personal and economic costs far in excess of any potential benefit.

REFERENCES


DISCUSSION

MELISSA J. ALLEN
Office of the Secretary, U.S. Department of Transportation, 400 7th Street, S.W., Washington, D.C. 20590.

Barnum and Gleason draw two general conclusions. First, they state that available data indicate that drug tests will result in significant false positives and false negatives. Second, they note that, using a Bayesian analysis and available data, drug testing will result in very high percentage of those identified as using drugs being falsely identified. For example, the paper states that under the DOT program for testing its own employees, as many as one-third of those identified as positive may be falsely identified. DOT personnel believe that the paper relies on data that is invalid for purposes of analyzing either its internal testing program or the recently initiated testing in the transportation industry resulting from DOT regulations. We also believe that the Bayesian analysis is an inappropriate technique for analyzing testing programs whose primary purpose is deterrence.

The basis for the authors' analysis is data taken from a 1988 JAMA article on the results of blind and open proficiency tests of laboratories conducted by NIDA in 1986 and 1987. The proficiency tests conducted at these laboratories in 1986 and 1987 are not the same as proficiency tests conducted at NIDA-certified drug-testing laboratories; as the authors acknowledge, confirmatory analysis may not have occurred and any confirmatory analysis that did occur is not significant.

The results of the NIDA-certified drug testing laboratories as the authors acknowledge, confirmatory analysis may have occurred and confirmatory analysis that did occur is not significant.
and to implement a laboratory certification program in April 1988. The NIDA guidelines are used in the DOT employee testing program and were adopted with minor modifications (reviewed by NIDA) for DOT's industry drug testing rules. Among other things, the following requirements must be met:

1. Drug testing laboratories must be HHS/NIDA certified. NIDA certification requires thorough documentation of laboratory personnel, procedures, and facilities; open and blind proficiency testing programs; onsite inspection by NIDA inspectors; and a comprehensive quality assurance program for all aspects of specimen processing.

2. GC/MS confirmatory testing is required for all samples that screen positive on immunoassay. This virtually error-free test is an extremely important element of any reliable testing program. Over half of the laboratories in the JAMA study did not use it.

3. Drug testing is authorized only for five drugs [cocaine, marijuana, opiates, phencyclidine (PCP), and amphetamines], and specific cutoff levels for a positive result are mandated. The NIDA-approved drug testing protocols are used to ensure standardization of positive test results. Cutoff levels are established for both immunoassay and GC/MS analysis.

4. Reanalysis (retest) of positive specimens is authorized. Employees may request reanalysis of a positive specimen. The reanalysis can be conducted at another NIDA-certified laboratory or at the NIDA laboratory that conducted the original analysis. Reanalysis of over 100 positive specimens in the DOT employee drug testing program has resulted in affirmation of the positive result in all cases.

5. A Blind Proficiency Testing Program is required of all employers. Blind quality control specimens (both specimens containing no drugs and specimens containing drug metabolites) are submitted to the laboratories by employers. The quality control specimens are indistinguishable from employee specimens. Results of the laboratory's performance on the quality control specimens are monitored by the employer, and unsatisfactory laboratory performance must be reported to DOT.

6. Medical Review Officer (MRO) verification of laboratory positives is mandated. Every laboratory positive must be reviewed by an MRO to ensure that there are no documentation or other errors in the testing process. The MRO verification process also permits employees to present documentation of an authorized medical use of a controlled substance. This verification process is an additional safeguard for employees whose specimen tests positive at the laboratory.

The testing procedures reflected in the JAMA studies are not comparable to, and are substantially less stringent than, the testing procedures used by DOT for its own employees and in its testing program for private sector transportation employees.

The DOT federal employee drug testing program was implemented in September 1987. The laboratory used for drug testing was certified by the Department of Defense (DOD). NIDA required the use of DOD-certified laboratories until the NIDA laboratory certification program was implemented. Extensive data are available on the blind proficiency testing program conducted by the Armed Forces Institute of Pathology (AFIP) for all DOD-certified drug testing. The DOD certified both U.S. military-operated laboratories and commercial laboratories used by the Department of the Army and the U.S. Coast Guard for drug testing. A summary of the AFIP blind proficiency test data for January 1988 through June 1989 is presented in Table 2.

DOT has also conducted a blind proficiency testing program of the NIDA-certified laboratory used for DOT employee testing. The results of this blind proficiency testing program for July 1989 through February 1990 are summarized in the following table.

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>Number of Samples</th>
<th>Number Correct</th>
<th>False Negative Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCA</td>
<td>2,476</td>
<td>2,476</td>
<td>0.00</td>
</tr>
<tr>
<td>EHRT</td>
<td>2,454</td>
<td>2,454</td>
<td>0.00</td>
</tr>
<tr>
<td>CL</td>
<td>1,725</td>
<td>1,725</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**TABLE 2:** AFIP BLIND PROFICIENCY TEST DATA FOR JULY 1989 THROUGH FEBRUARY 1990

Note: NCA = Northwest Toxicology Associates. EHRT = Environmental Health, Research, and Testing. CL = Compuchem Laboratories.

*Blind spikes for amphetamines, opiates, and PCP were not tested at Army contract laboratories because the Army program only tests for cocaine and THC.*
20 false negative rates obtained at the NIDA laboratories in the certified laboratories. These data validate the accuracy of drug testing at NIDA-certified laboratories.

There are many other problems or errors with the paper that are worth brief mention here:

1. The JAMA article study covered testing for drugs not included in the DOT program or regulations. The error rates for those other drugs appeared to be higher. It appears, therefore, that the error rate relied on by the authors for analysis of DOT testing should be lower.

2. As noted above, the DOD has conducted blind proficiency testing in its military drug testing laboratories using immunoassay and GC/MS methodologies for the past 6 years. The Navy tests cited by the authors that were declared "scientifically unsupportable" in 1982 were not conducted using GC/MS confirmatory analysis.

3. The data used in the paper for estimating potential drug use in the transportation industry are not a valid basis for the conclusions reached. For example, some of the estimates appear to be based on preemployment drug testing that, as a scheduled test, is not a reliable indicator of overall drug use. It will not identify users who can stop long enough to get a negative on a test. Recent data from the 1988 NIDA Household Survey on Drug Abuse indicate that approximately 12.5 percent of adults employed in transportation occupations have used illegal drugs.

Finally, it seems highly inappropriate to use a Bayesian analysis on a drug testing program that stresses deterrence. Under the Bayesian approach, even though the actual number of errors essentially stays constant, the number of errors becomes an increasingly higher percentage of the total number of positives as the actual number of real positives decreases. In other words, the Bayesian analysis penalizes a program that effectively deters usage—the ultimate goal—by making it appear that the best program has the worst rate for falsely identifying drug users. An appropriate analysis would simply indicate the actual number or rate of false positives. More important, because the data presented indicates that the paper should have used a predictive value of virtually 100 percent—if not 100 percent—under the Bayes' theorem, the conclusion of the paper should have been: transit industry testing will have virtually no false positives.

AUTHORS' CLOSURE

We have carefully analyzed Allen's discussion of our paper. Our most important conclusions are these. First, the discussant provided no valid evidence about laboratory accuracy in transit, and we are still convinced that drug tests sometimes may be very inaccurate in current transit agency testing. Second, Allen missed the primary intent of our paper. Our purpose was not to estimate average drug test accuracy in the transit industry but to present a methodology for ensuring acceptable accuracy at individual transit properties. Third, we were able to identify only one of the sources that the discussant utilized. In this case, we discovered that she had badly misinterpreted the data. Furthermore, Allen misrepresented a number of the points we made in our paper. Consequently, we are concerned about the accuracy of all of the information that she presented. Our discussion of these three conclusions, and a number of other points of disagreement, are covered in this closure.

In her discussion of our paper, Allen's response begins with a totally incorrect statement about our conclusions. In describing the conclusions that we supposedly drew, she states:

First, they state that available data indicate that drug tests will result in significant false positives and false negatives. Second, they note that... drug testing will result in a very high percentage of those identified as using drugs being falsely identified.

Neither of these conclusions appears in our paper. First, we never said that drug tests would result in significant false positives and false negatives. In fact, we showed that inaccurate results could occur in situations where the false positive and false negative rates were both very low. Second, we were very careful not to claim that drug testing will result in high percentages of those testing positive being falsely identified. Indeed, we explicitly stated:

Very importantly, these false conviction rates are based on our estimated average rates for drug usage, sensitivity, and specificity and are not necessarily applicable to any particular transit agency. But all of our estimated rates are ones that could occur in some circumstances. Because of the extremely serious consequences of being convicted of drug use, an employer would be wise to determine that these estimated rates, or similar rates, do not apply before acting on positive drug test results.

Allen also states that our paper should have concluded that "transit industry testing will have virtually no false positives." We disagree. As discussed next, we are even more strongly convinced than before that seemingly accurate tests for abused drugs may sometimes be inaccurate to a disturbingly high degree in current transit industry testing. We feel that it is even more appropriate to use the methodology we suggested in the original paper, by which transit decision makers can correctly set desired accuracy standards for their own organizations, thereby easily avoiding these potential inaccuracies.

The discussant's statement attempted to show the accuracy of NIDA-certified laboratories. However, it neglected to mention that, since January 1990, the DOT (through UMTA regulations) use laboratories certified by NIDA for all of their drug testing. These laboratories had to confirm all positive test results, using GC/MS for all confirmations. On January 19, 1990, after less than a month of drug testing under these regulations, UMTA's authority to impose requirements was revoked by the courts (I). Thus, at the time of this writing, April 9, 1990, federal regulations do not prevent transit system drug testing from
being conducted by anyone, and from using any confirmation method desired. For example, one management company that operates a large number of transit systems has told its resident managers that they may use any laboratory in which they have confidence, whether or not it is NIDA certified. Transit drug testing is regulated by widely varying local laws, some very strict and others very lenient. Testing at many systems is not regulated by government in any way. Thus, the average accuracy of the transit industry’s drug testing might be comparable to the 1985 JAMA study, the 1988 JAMA study, or some other level. Consequently, accuracy levels used in our paper, which are based on the 1985 and 1988 JAMA articles, might be higher or lower than the average levels actually present in the industry at this point in time.

Moreover, we are concerned about whether the DOD-certified laboratory accuracy averages that Allen presented would be applicable even if transit were required to use NIDA-certified laboratories. These data show a quantum leap in accuracy when compared to all other blind laboratory proficiency studies. We would be more reassured about the validity of the data if they had come from an article in a well-respected refereed journal, or from another source that had been subjected to rigorous refereeing, and if the data had been gathered by a disinterested party. The military results may or may not be applicable to DOT-required testing at NIDA-certified laboratories, but there appear to be important differences between the DOD and DOT laboratory procedures, at least on the surface (2,3). For example, although two tests are required under the DOT procedures (3), Irving (2) states that the Navy requires three tests before a specimen is considered positive. The extra test should significantly lower the false positive rate. Moreover, DOD cutoff concentrations for some drugs appear to be higher than DOT cutoff concentrations (2,3), which would also make it less likely that the DOD laboratories would have false positives when compared to the laboratories used by DOT. There are other differences as well, but these two should serve to illustrate the fact that DOD results may well be more accurate than NIDA results. It is completely invalid to estimate one from the other without a much more detailed examination of the procedures, refereed by experts in the field. Therefore, in our opinion, the information presented does not demonstrate the accuracy of DOT drug testing procedures at NIDA-certified laboratories, because too many questions remain unanswered.

If, at some future time, transit drug tests are indeed proved to have no false positives, then Bayesian analysis would not be necessary, because the false conviction rate would be zero. However, there must be zero false positives, not, as Allen states, “virtually no false positives,” especially where drug usage rates are low. For example, if DOT testing of its own employees reduces their drug usage by the same percentage that occurred in the military, DOT’s usage rate will drop from 0.6 percent to less than 0.1 percent. Incidentally, 0.1 percent is the rate found during U.S. Customs Service random testing of its current employees, so having only one drug user out of every 1,000 employees is attainable (4). Using this 0.1-percent usage rate and the concepts we discussed in our paper, there could be no more than 1 false positive out of every 98,901 drug-free specimens, if we wanted to falsely convict no more than 1 out of every 100 people testing positive. Or, assume it is required that there be no more than 1 false positive out of every 1,000 positive test results, a level the unions might find more reasonable than only 1 out of 100. Using the 2.6-percent drug usage rate recently found at a large transit system and DOT’s maximum allowable false negative rate of 10 percent (3), then there could be no more than 1 false positive out of every 41,582 drug-free specimens. In either case, even the military’s alleged 40,000 tests of drug-free specimens without a false positive would not be enough to attain acceptable accuracy, even if the military results were applicable. Thus, Bayesian analyses should be performed unless there are no false positives at all.

Most important, however, the discussant’s statement did not address the primary intent of our paper. Our main purpose was not to estimate drug test accuracy in the transit industry but to present a methodology for ensuring acceptable accuracy. Because this point is so important, we repeat here the critical paragraphs.

Because some will quarrel with our estimates of drug use, sensitivity, and specificity rates, the primary emphasis of this paper may be obscured. The average rates believed to be the most applicable to transit drug testing were used. These rates were based on empirical evidence, from studies published in reputable sources. When more relevant average rates become available, they should be used in place of the ones used herein. Further, regardless of whose average estimates are used, conclusions about the extent of error that are based on averages are not applicable to a particular situation. False conviction rates for a specific organization can only be estimated if the underlying Bayesian methodology is combined with data specific to that particular case.

Our main message is that appropriate rates should be used in a Bayesian framework to develop estimates of false conviction rates, and that the maximum acceptable false conviction rate should be used in setting the required combinations of sensitivity and specificity.

Our article presented an easy solution to the problem of insufficient accuracy, that Allen apparently missed in her concern about the various laboratory proficiency studies. The solution is introduced in the second and most important part of our study, the section entitled “Methodology for Developing Acceptable Predictive Probabilities.” It is easy to lower the false conviction rate by retesting a second time. That is, a positive screening test is already automatically followed by a confirmation test. To lower the false conviction rate still further, all that is necessary is to follow a positive confirmation test with a second confirmation test. For cases in which there is extremely low drug usage, or in which extremely low rates of false convictions are required, a positive result on the second confirmation could be followed by a third confirmation. To decide on the number of confirmations needed, it would be necessary to do a Bayesian analysis with the desired parameters. The multiple testing must not be optional, nor should it be conducted on a request by the involved employee, but it must automatically occur when the screen and first confirmation are positive. This suggestion is a variation on the two-stage procedure that we discussed at length in the paper; it accounts only for laboratory errors, not for errors in the chain of custody or contaminated vessels or other problems occurring outside the laboratory itself. The procedure suggested in the article accounts for these external problems as well.

We also wish to comment on the discussant’s belief that “Bayesian analysis is an inappropriate technique for analyzing
testing programs whose primary purpose is deterrence." We disagree. As we noted in the paper, both the American legal system and its system of workplace jurisprudence are based on the principle that a person is assumed innocent until proven guilty with compelling evidence. Under our legal and industrial jurisprudence systems, for example, it is unlikely that a defendant would be convicted if there were only a 50-percent chance that the prosecution's evidence was true, but it is likely that a defendant would be convicted if there were a 99.9-percent chance that the evidence was true. Bayesian analysis identifies the correct probability that the evidence provided by a positive drug test is true, and we ignore it to the peril of our system of justice. It is highly appropriate that it be used.

Allen is correct in stating that as drug usage declines, the number of false positives remains essentially constant although the false conviction rate increases. But this fact is irrelevant. It assumes that we must choose between a situation with high drug use and low false conviction rates and one with low drug use and high false conviction rates. We can easily have both low use and low false conviction rates by requiring multistage testing in appropriate situations. Bayesian analyses tell us, for given rates of usage and desired ceilings on rates of false convictions, when multistage testing is necessary. Thus, although we hope DOT's drug testing program will lower drug abuse rates, the simultaneous application of Bayesian techniques will ensure that employees testing positive are indeed likely to have taken drugs, by adjusting the retesting requirements for the drug abuse rate of the population being tested.

The discussant misrepresented several other statements in our paper. She states that we acknowledge that confirmation tests may not have occurred in the 1988 JAMA study. The opposite is true; we stated that "It appears that the laboratories in the 1988 JAMA study did conduct confirmation testing of specimens that initially screened positive, and the most frequently used confirmation method was GC/MS." Also, Allen says that some of our estimates of drug use appear to be based on preemployment tests. To the contrary, all of the estimates we used in the analyses were based solely on drug tests of current employees.

Moreover, while we don't want to debate numbers, Allen makes the following statement about drug use: "Recent data from the 1988 NIDA Household Survey on Drug Abuse indicate that approximately 12.5 percent of adults employed in transportation occupations have used illegal drugs." This is not true. The actual results from NIDA indicate that 12.5 percent of the surveyed males between 18 and 40 in the transportation industry said that they had used some illicit substance in the past month, with the illicit substances including the nonmedical use of stimulants, sedatives, tranquilizers, or analgesics (5). Figures for all adults in transportation are not even available. For all occupations, however, males were almost twice as likely as females to say they had used some illicit substance, and workers between 18 and 35 were almost six times more likely than workers over 35 to say they had used some illicit substance (5). The usage rate would be even lower if only illicit drugs, which is all that DOT is testing for, are considered. In short, not only has the discussant incorrectly identified the reference group and the illicit substances to which the 12.5-percent figure applies, but also, by saying it represents the drug use rate of all adult transportation workers, has drastically overestimated drug use by transportation employees. We assume that this error was unintentional and not a conscious attempt to inflate the drug-use figures. But, if Allen could make so basic an error in so simple a data situation, it calls into question the validity of all of her data, much of which is very complex and requires expert interpretation. Because the discussant did not provide complete references for any of her sources, we were not able to check the accuracy of the information, except for the one case just cited. However, the error that we did discover reemphasizes our earlier point about the laboratory data presented: the information should not be accepted as valid until subjected to a rigorous and complete review by neutral experts in the field of laboratory proficiency studies.

In closing, we are pleased that DOT, through Allen, chose to offer a discussion of our paper. It introduced new data that we were unaware of and provided us with the opportunity to correct misunderstandings that our original paper inadvertently generated. We feel that DOT personnel are working hard to make their drug testing procedures accurate, and we hope that they and others will use our suggestions to further improve the certainty of error-free tests.

REFERENCES

5. A. Kopstein and J. Giroeret. Drug Use Patterns and Demographics of Employed Drug Users: Data from the 1988 National Household Survey on Drug Abuse. Division of Epidemiology and Prevention Research, National Institute on Drug Abuse, Rockville, Md.

Publication of this paper sponsored by Committee on Transit Management and Performance.