Journal of Forensic Toxicology & Pharmacology

Research Article A SCITECHNOL JOURNAL

Prevalence Dependent Accuracies Misstate the Diagnostic Power of the Standardized Field Sobriety Test

Greg Kane[1*] and Elizabeth Kane[2]

[1]Department of BAs math and physics, Englewood, Colorado, United States

[2]Department of BA applied mathematics, Englewood, Colorado, United States

***Corresponding Author:** Greg Kane, Department of BAs math and physics, Englewood, Colorado, United States, Tel: 303-919-1041; Fax: 303-741-4622; Email: gregkanemd@gmail.com

Received date: February 11, 2021; **Accepted date:** February 26, 2021; **Published date:** March 05, 2021

Abstract

In the 1990s as the legal blood alcohol limit for driving changed, validation studies reported the Standardized Field Sobriety Test (SFST) to be accurate at discriminating between Blood Alcohol Concentrations (BAC) above or below several legal limits: 0.10%, 0.08%, 0.05% and 0.04%. This effectively made the implications of the test depend on the legal jurisdiction involved and on whether the driverÃ¢â,¬â„¢s license was passenger or commercial class. We investigated the contribution of the validation studiesÃ¢â,¬â„¢ choice of accuracy statistic to the SFSTÃ¢â,¬â„¢s reported accuracy.

Methods: Using the data set from a commonly cited SFST validation study, we calculated the arrest accuracy and overall accuracy of the SFST at identifying BACs above or below 31 different target BACs from 0.00% to 0.30%. We organized the results in tables; we observe and explain trends.

Results: At target BAC 0.30% the arrest accuracy of the SFST is 1%; at BAC 0.15%, 34%; at BAC 0.00%, 100%. The statistics arrest accuracy and overall accuracy describe the SFST, a test designed to identify changes caused by alcohol, as less accurate when the changes are severe, more accurate when changes are mild, and as 100% (arrest) and 93% (overall) accurate when there are no changes at all.

Conclusion: The statistic arrest accuracy identifies the SFST as currently used by US law enforcement as 78% accurate. This number is an artifact of the prevalence dependence of arrest accuracy. Calculations independent of prevalence show that the SFST actually has no meaningful power to discriminate between drivers with high and low BACs. The statistics overall accuracy and arrest accuracy to not quantify the probability that impaired driving defendants who failed the SFST had an elevated BAC or were impaired.

Keywords: Sobriety test; Blood alcohol concentration

Introduction

This paper is about statistical methods widely understood to validate the law enforcement Standardized Field Sobriety Test; it is not about the sobriety test itself. The SFST is a physical examination, most often done at the roadside by a trained law enforcement officer in the course of an impaired driving investigation. Developed beginning in the 1970s [1-3], and field tested in projects sponsored by the National Highway Traffic Safety Administration (NHTSA) in the 1990s [4-6]. The procedure has been well described [7-13], and its statistical properties considered [14]. Initially investigated for its ability to identify elevated Blood Alcohol Concentration (BAC), the SFST is also used [15] and investigated [16-23] as a method to identify impairment and the presence of various drugs.

In court SFSTs are presented as evidence in impaired driving prosecutions, typically through the testimony of the police officer who administered the test to the defendant. Officers are taught [24] that the SFST was validated in NHTSA's lab and field tests, and that if possible their testimony should reference the San Diego Field Validation Study [6], commonly cited as Stuster and Burns 1998 (here Stuster and Burns).

Stuster and Burns reported that when officers in the study used the SFST to estimate BAC as above or below the then new legal limit of 0.08%, the overall accuracy of their estimates was 91%. In current SFST training police officers are taught that this statistic shows that the SFST is valid – 91% accurate. Officers are taught that Stuster and Burns also show that the SFST subtests are accurate. Horizontal Gaze Nystagmus (HGN) 88% accurate; Walk and Turn (WAT) 79% accurate; One Leg Stand (OLS) 83% accurate – in each case “accurate” referring to overall accuracy. Stuster and Burns further reports that officers arrest decisions were 94% accurate (positive predictive value, PPV) at identifying BACs below 0.04% or above 0.04 but below 0.08%.

Officers learn that the earlier Colorado Field Validation Study [4] reported that, in assessing drivers as being above or below Colorado’s lower legal limit of 0.05%, arrest decisions were 86% (overall) and 93% (arrest) accurate. Officers also learn about Anderson et al [3], which reported that, in assessing drivers as being above or below the then legal limit BAC of 0.10%, officer’s overall accuracy was 80%.

Based on this use of the statistics overall accuracy and PPV, police officers are taught that the SFST is accurate at identifying BACs above or below 0.10%, above or below 0.08%, above or below 0.05% and above or below 0.04%. This allows officers to testify that the SFST is accurate at identifying BACs above or below various legal limits commonly used in the USA.

Quantifying the performance of diagnostic tests

In quantifying the performance of a diagnostic test, the overall accuracy and PPV are indeed properties of the test, but not of just the test. Overall accuracy and PPV are prevalence dependent [25-27]. They depend on, and vary with, the prevalence of the condition tested for in the group from which test subjects are selected. Difficulties arising from the use of overall accuracy have been described [28-29], as has the mathematics of quantifying diagnostic accuracy [30-35]. Vecchio [36] illustrates difficulties with the PPV with the example of a test whose sensitivity and specificity are both 95%. When this test indicates Yes, condition present, the answers are correct 98% of the time –but only when the people tested are taken from a group 75% of whom have the condition tested for, *i.e.* prevalence 75%. When exactly the same test is administered to people from a group with a prevalence of 1%, the Yes answers are correct only 16% of the time.

The diagnostic performance of a test can be quantified with statistics that are not prevalence dependent. Sensitivity and specificity are widely used. They may be combined to form the Likelihood Ratio (LR), which describes the performance of a diagnostic test with a single number, making it possible to compare different tests and to get a general sense of how a test performs [37-39]. The LR level at which a test becomes useful depends on the application-specific meaning of "useful." There are general guidelines. The American Medical Association teaches medical practitioners that tests with LR greater than 10 generate large changes in probability; LR 5-10 moderate changes; LR 2-5 small changes; LR 1-2 small changes that are rarely important [30].

The practical implications of the actual numbers quantifying a test's sensitivity, specificity, and LR are often not easy to see. This is not surprising; the implication of a test depends on a factor, prevalence that has been removed from those statistics. One fix is to return prevalence to the calculation, to use the empirically discovered sensitivity and specificity of a test to calculate the PPV (the probability a test's answer Yes, condition present is correct) several times, at each of a range of pre-test probabilities. The resulting table, pairing pre-test and post-test probabilities, gives an accessible and intuitive picture of how "accurate" a test is. In Vecchio's example a very accurate test, 95% sensitive and specific, changed a pre-test probability of 1% only to 16% –a surprising result, though well described [32].

The SFST is reported to be validated by accuracy statistics that are prevalence dependent. As per se impaired driving laws have changed [12], the test has been reported to be accurate at identifying BACs above or below several legal limits: 0.10%, 0.08%, 0.05% and 0.04%. This paper investigates the contribution of prevalence to the SFST accuracies reported by Stuster and Burns.

Methods

Using the Stuster and Burns data set, we calculated the arrest accuracy and overall accuracy of the SFST at identifying BAC as above or below 31 target BACs from 0.00 to 0.30%. We arranged the answers in a table and observed trends. We constructed similar tables for the SFST subtests HGN, WAT, OLS, and for two artificial data sets, one random, and the other with every test result set to Yes, BAC high. We compare and explain results.

We calculated the LR of Stuster and Burns' SFST at each target BAC. We constructed a table of the PPV of the SFST at several target BACs and pre-test probabilities.

Data Sets

Stuster and burns: The data set for Stuster and Burns was obtained from Dr. Jack Stuster, author of the study. A second copy was shared by researcher Dr. Mike, who originally got the file from NHTSA *via* a FOIA request [14]. The two Excel files were identical. Stuster and Burns data included HGN, WAT, and OLS "clue" counts paired with evidentiary BAC levels. The data set included other information (encounter date, time, driver gender, officer's BAC estimate, etc.) not considered here. Stuster and Burns gave criteria for interpreting individual subtests but not for interpreting the three subtests taken together. The data set does not include entries for an overall SFST result.

Randomized sobriety test

The first artificial data set constitutes what this paper calls, for convenience, the Randomized Sobriety Test. Data consisted of 261 randomly generated SFST results paired with randomly generated BACs.

San Diego all fail sobriety test

The second artificial data set consists of the Stuster and Burns evidentiary BAC results, but with each of those real BACs paired with a sobriety test "result" set to BAC high. No one passes this test; everyone fails.

Index and reference tests

We calculated the overall accuracy and arrest accuracy (=PPV) with which index tests predicted reference test results. In general, the index test was the sobriety test result interpreted as indicating a BAC above or below the target level; the reference test was the BAC.

For the stuster and burns data set the index text was the combined SFST test, defined as follows. SFST results (index) were counted as indicating BAC high when any one of three subtests (HGN, WAT, OLS) indicated BAC high according to the criteria in Stuster and Burns. Those criteria are based on "clue" counts. Tests indicate BAC high when: HGN ≥ 4; OLS ≥ 2; WAT ≥ 2, and BAC low otherwise. SFSTs for drivers who failed one subtest and passed the other two were counted as BAC high. Only SFST results for divers who passed all three subtests were counted as indicating BAC low. The data set included drivers with results reported for one or two subtests, but not all three. Their SFSTs were not included in the calculations here.

The reference test was the evidentiary BAC.

For the randomized sobriety test, the index text consisted of sobriety test "results" for each driver–BAC high or BAC low–generated at random, each with a 50% probability. The reference test consisted of random BACs from 0.00 to 0.30%. Values were randomized by the Excel function rand.

For the San Diego all fail sobriety test the index test was the sobriety test "result", which was in every case chosen to be BAC high. The reference test was Stuster and Burns' evidentiary BAC.

Target values

Target BAC values were chosen from 0.00 to 0.30%, in increments of 0.01%.

Illustrative chart

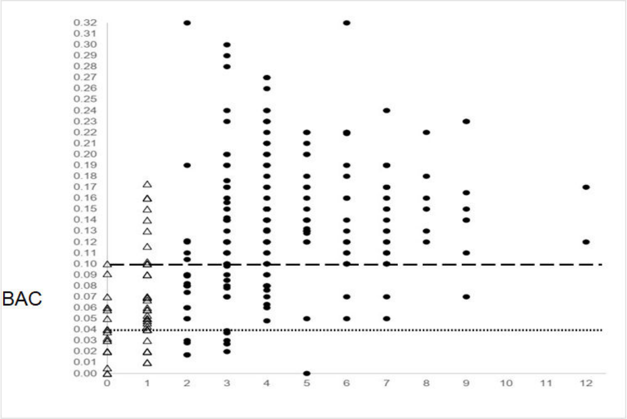

Figure 1: BAC *vs.* One Leg Stand "clue" count for each driver in the Stuster and Burns data set. Circles represent drivers who failed the SFST; triangles represent drivers who passed.

In Figure 1, OLS "clue" count is plotted against evidentiary BAC, for each driver with reported OLS results in the Stuster and Burns data set.

Calculations

The interpretation criteria in Stuster and Burns were used to dichotomize drivers' HGN, WAT, OLS results as indicating BAC high or BAC low. Drivers' evidentiary BACs were dichotomized as high or low compared to a particular target BAC. The sobriety test interpretations high or low (index) were compared with the evidentiary BACs high or low (reference), and counts were made of the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) results at that target BAC. From these empiric quantifications of SFST performance, calculations were done for the statistics: PPV TP/(TP+FP); overall accuracy (TP+TN)/(TP+TN+FP+FN); sensitivity TP/(TP+FN); specificity TN/(TN+FP); and likelihood ratio sensitivity/(1-specificity). These statistics were calculated at each target BAC.

For each subtest (HGN, WAT, and OLS) arrest accuracy and overall accuracy were calculated in the same manner, except that drivers with incomplete records were included when possible. For example, HGN scores were counted for any driver with a recorded HGN result, regardless of whether the data set reported that driver's OLS and WAT scores. These statistics were calculated for each target BAC.

For the Randomized Sobriety Test and San Diego All Fail Sobriety Test, arrest accuracy and overall accuracy were calculated in the same manner.

Using the empirical sensitivities and specificity of Stuster and Burns' SFST, and a convenient formulation of PPV [34]:

$$PPV = \frac{sensitivity \times prevalence}{(sensitivity \times prevalence) + (1 - specificity)(1 - prevalence)}$$

Prevalence-dependent Arrest Accuracies (PPV) were calculated at four target BACs, 0.10, 0.08, 0.05, 0.04%, at three pre-test probabilities of elevated BAC, 10, 50, and 90%.

Results

The Stuster and Burns data set included records for 297 drivers. Results including all three subtests, HGN, WAT, OLS, were reported for 261 drivers. Of these, 242 failed (had at least one subtest indicating BAC high) and 19 passed (no subtest indicating BAC high).

The arrest accuracy (PPV) and overall accuracy of the Stuster and Burns SFST, of the Randomized Sobriety Test, and of the San Diego All Fail Sobriety Test, at each of the target BACs, is presented in Table 1.

	Standardized field sobriety test		Randomized sobriety test		San Diego all fail sobriety test	
Target BAC	Arrest accuracy	Overall accuracy	Arrest accuracy	Overall accuracy	Arrest accuracy	Overall accuracy
0	100%	93%	100%	58%	100%	100%
0.01	99%	93%	96%	56%	98%	98%
0.02	98%	92%	92%	55%	97%	97%
0.03	95%	89%	89%	55%	93%	93%
0.04	93%	89%	81%	52%	90%	90%
0.05	90%	89%	75%	50%	86%	86%
0.06	86%	86%	73%	50%	80%	80%
0.07	83%	84%	68%	49%	77%	77%
0.08	78%	79%	63%	48%	72%	72%
0.09	73%	75%	60%	49%	68%	68%
0.1	69%	72%	59%	50%	64%	64%
0.11	62%	65%	58%	50%	57%	57%
0.12	56%	59%	53%	48%	52%	52%
0.13	48%	51%	50%	48%	44%	44%
0.14	43%	47%	46%	49%	39%	39%
0.15	34%	39%	43%	48%	31%	31%
0.16	29%	34%	40%	47%	26%	26%
0.17	24%	30%	38%	48%	22%	22%
0.18	19%	25%	34%	47%	18%	18%
0.19	18%	24%	32%	46%	16%	16%
0.2	14%	20%	30%	47%	13%	13%
0.21	11%	17%	29%	48%	10%	10%
0.22	10%	16%	28%	48%	9%	9%
0.23	7%	14%	27%	49%	7%	7%
0.24	5%	12%	26%	49%	5%	5%
0.25	3%	10%	22%	48%	3%	3%
0.26	3%	10%	16%	45%	3%	3%

0.27	3%	10%	11%	43%	3%	3%
0.28	2%	9%	9%	43%	2%	2%
0.29	2%	9%	6%	42%	2%	2%
0.3	1%	8%	3%	43%	1%	1%

Table 1: The arrest accuracy and overall accuracy with which three sobriety tests identify BAC as above or below target levels from 0.00 to 0.30%.

For the Stuster and Burns HGN, WAT, OLS tests, similar statistics are presented in Table 2.

	HGN		**WAT**		**OLSOLS**	
Target BAC	**Arrest accuracy**	**Overall accuracy**	**Arrest accuracy**	**Overall accuracy**	**Arrest accuracy**	**Overall accuracy**
0	100%	81%	100%	81%	100%	78%
0.01	100%	82%	99%	81%	100%	79%
0.02	99%	82%	98%	80%	99%	79%
0.03	98%	84%	95%	79%	97%	80%
0.04	98%	88%	93%	80%	95%	80%
0.05	97%	90%	91%	81%	95%	84%
0.06	93%	89%	89%	83%	92%	85%
0.07	91%	89%	87%	83%	90%	85%
0.08	87%	88%	82%	79%	85%	83%
0.09	82%	84%	78%	77%	80%	78%
0.1	77%	80%	74%	75%	77%	77%
0.11	71%	76%	67%	70%	71%	74%
0.12	64%	70%	61%	66%	65%	70%
0.13	54%	62%	52%	59%	54%	61%
0.14	48%	58%	47%	55%	48%	57%
0.15	39%	51%	37%	47%	38%	50%
0.16	33%	46%	32%	44%	32%	45%
0.17	29%	42%	27%	41%	28%	44%
0.18	24%	38%	22%	37%	23%	40%
0.19	22%	37%	20%	35%	21%	38%
0.2	18%	34%	16%	31%	16%	35%
0.21	15%	31%	12%	28%	13%	32%
0.22	13%	30%	11%	27%	11%	31%
0.23	10%	27%	8%	25%	8%	29%
0.24	6%	24%	5%	23%	6%	26%
0.25	5%	23%	4%	22%	4%	25%
0.26	4%	23%	3%	21%	4%	25%
0.27	3%	22%	3%	21%	3%	25%
0.28	3%	21%	2%	20%	2%	24%
0.29	2%	21%	1%	20%	2%	23%
0.3	2%	21%	1%	20%	1%	23%

Table 2: The arrest accuracy and overall accuracy with which each SFST subtest identifies BAC as above or below target levels from 0.00 to 0.30%.

For the Stuster and Burns SFST, further statistics are presented in Table 3, and prevalence-dependent arrest accuracies are presented in Table 4.

Target BAC	**Arrest accuracy**	**Overall accuracy**	**LR**	**Prevalence**	**Sensitivity**	**Specificity**	**TP**	**FP**	**TN**	**FN**
0	100%	93%	-	1	0.93	-	242	0	0	19
0.01	99%	93%	1.87	0.98	0.93	0.5	240	2	2	17
0.02	98%	92%	1.5	0.97	0.94	0.38	237	5	3	16
0.03	95%	89%	1.23	0.93	0.94	0.24	229	13	4	15
0.04	93%	89%	1.32	0.9	0.95	0.28	224	18	7	12
0.05	90%	89%	1.5	0.86	0.97	0.35	218	24	13	6
0.06	86%	86%	1.49	0.8	0.99	0.33	208	34	17	2
0.07	83%	84%	1.42	0.77	1	0.3	200	42	18	1
0.08	78%	79%	1.35	0.72	1	0.26	188	54	19	0
0.09	73%	75%	1.29	0.68	1	0.23	177	65	19	0
0.1	69%	72%	1.26	0.64	1	0.2	168	74	19	0
0.11	62%	65%	1.21	0.57	1	0.17	150	92	19	0
0.12	56%	59%	1.18	0.52	1	0.15	136	106	19	0
0.13	48%	51%	1.15	0.44	1	0.13	115	127	19	0
0.14	43%	47%	1.14	0.39	1	0.12	103	139	19	0

0.15	34%	39%	1.12	0.31	1	0.11	82	160	19	0
0.16	29%	34%	1.11	0.26	1	0.1	69	173	19	0
0.17	24%	30%	1.1	0.22	1	0.09	58	184	19	0
0.18	19%	25%	1.1	0.18	1	0.09	46	196	19	0
0.19	18%	24%	1.1	0.16	1	0.09	43	199	19	0
0.2	14%	20%	1.09	0.13	1	0.08	33	209	19	0
0.21	11%	17%	1.09	0.1	1	0.08	26	216	19	0
0.22	10%	16%	1.09	0.09	1	0.08	23	219	19	0
0.23	7%	14%	1.08	0.07	1	0.08	18	224	19	0
0.24	5%	12%	1.08	0.05	1	0.08	12	230	19	0
0.25	3%	10%	1.08	0.03	1	0.08	8	234	19	0
0.26	3%	10%	1.08	0.03	1	0.08	8	234	19	0
0.27	3%	10%	1.08	0.03	1	0.07	7	235	19	0
0.28	2%	9%	1.08	0.02	1	0.07	5	237	19	0
0.29	2%	9%	1.08	0.02	1	0.07	4	238	19	0
0.3	1%	8%	1.08	0.01	1	0.07	3	239	19	0

Table 3: Statistics quantifying the diagnostic performance of Stuster and Burns' SFST at identifying BAC as above or below target levels from 0.00 to 0.30%. LR: Likelihood Ratio. Sensitivity and specificity in the table are rounded to two significant digits. LRs in the table were calculated with unrounded values.

		Pre-test probability of high BAC		
		10%	**50%**	**90%**
Target BAC	0.04	13%	57%	92%
	0.05	14%	60%	93%
	0.08	13%	57%	92%
	0.1	12%	56%	92%

Table 4: The post-test probability that drivers who fail the Stuster and Burns SFST actually have an elevated BAC, as a function of target BAC and pre-test probability. When a group, 10% of whom have BACs above 0.08%, is tested with the SFST and just those people who fail the SFST are considered, 13% will have a BAC above 0.08%.

A plot of Stuster and Burns' OLS "clue" count vs. evidentiary BAC is presented in Figure 1. Linear regression yielded r2=0.20.

DISCUSSION

Arrest accuracy and overall accuracy are higher at each lower BAC

Table 1 reveals that the lower the target BAC "identified" by the SFST, the higher the arrest accuracy. At higher BACs the SFST is less accurate. At BAC 0.24% –three times the legal limit in most US jurisdictions– the SFST is only 5% accurate.

This trend is not an artifact of some unusual feature of the Stuster and Burns data set. It holds for the Random Sobriety Test, for the San Diego All Fail Sobriety Test, and for each of the SFST's component tests, HGN, WAT, and OLS.

The fact that a test designed to identify changes caused by alcohol is less accurate when the changes are severe and more accurate when changes are mild is surprising and counterintuitive. It is also surprising that a sobriety test that simply identifies every driver as having a high BAC, and a random sobriety test equivalent to a coin toss, are both also accurate at identifying drivers with ever lower BACs.

Explanation

This phenomenon arises from the choice of prevalence dependent statistics to assess diagnostic power.

Because Stuster and Burns evaluated the accuracy of an already defined SFST, the clue-count interpretations do not vary. What does vary, here and in Stuster and Burns, is the target BAC. Each time the target BAC is lowered, BAC high is redefined to include BACs above the new, lower target. More False Positives are redefined as True Positives, and the prevalence dependent overall and PPV accuracies go up.

Figure 1 illustrates the phenomenon. Stuster and Burns' interpretation criteria dichotomize OLS "clue" counts of 2 or more as indicating BAC high, and counts of 0 or 1 as indicating BAC low. In Figure 1, drivers with an OLS result indicating BAC high (index) are represented by circles. Thus, TP and FP counts involve just the circles. Now consider the dashed horizontal line at BAC 0.10%. Circles above the line are defined as True Positives. Circles below the line are defined as False Positives. The arrest accuracy of the OLS test is the percentage of circles that are above the line.

This is the case for target BAC 0.10%. Then, each time the target BAC is changed, true and false are redefined to mean above or below the new horizontal line at the new target BAC (reference). For target BAC 0.04% the arrest accuracy is the percentage of circles that are above the horizontal line at 0.04%.

There are more circles above the line at 0.04% than there are above the line at 0.10%. The arrest accuracy of the Stuster and Burns SFST is greater at 0.04% than at 0.10%. In general, as lower target BACs redefine ever lower results as True, the arrest accuracy increases. This relationship holds for all target BACs down to 0.00, at which point all results have been redefined as True and the arrest accuracy is 100%. The arrest accuracy changes with target BAC because changing the target BAC changes –redefines– the prevalence of the condition BAC high.

Drivers judged by Stuster and Burns' OLS test to have BAC low are represented in Figure 1 as triangles. TN and FN counts involve just the triangles. For these drivers, BACs above the line are False; below the line are true.

The overall accuracy is the percentage of all drivers whose results are true; that mean circles above the line and triangles below the line. As target BACs get smaller the overall accuracy trends towards its final value, the percentage of all drivers whom the sobriety test judges to have a high BAC. At target BAC 0.00, all circles are defined as true; all triangles are defined as False.

Table 3 quantifies this phenomenon for the overall SFST. The columns TP and FP represent the 242 drivers who, according to the interpretation prescribed in Stuster and Burns, failed the SFST. The arrest accuracy is the percentage of this number who are classified as TP. At high target BACs the number of TPs is low–few drivers had BACs above, say, 0.24%. As the correctness of the SFST's interpretation is redefined at each lower target BAC, the number of TPs increases and the number of FPs decrease. The mechanics of the test do not change, but the definitions of True and False do. Finally, at target BAC 0.00, everyone who failed the test has been redefined as a TP and the arrest accuracy is 100%. As changes to the target BAC effectively redefine the target condition, the sensitivity and specificity also change.

No clinically meaningful power to discriminate between BAC high and BAC low

The arrest accuracy and overall accuracy statistics overstate the diagnostic power of the SFST. Table 3 reveals that the 78% arrest accuracy of Stuster and Burns' SFST at target BAC 0.08% arises not from the diagnostic power of the test, but from the 72% prevalence in Stuster and Burns' study group of BACs above that level, to which is added a small contribution –6%–from the test itself.

When a test's sensitivity and specificity are known, calculations of PPV can quantify the degree to which knowledge of a test result changes the probability of the condition tested for. Table 4 shows the results of this calculation for Stuster and Burns' SFST, using the sensitivities and specificities in Table 3, at target BACs 0.10%, 0.08%, 0.05% and 0.04%.

When a group, 10% of whom have BACs above 0.08%, is tested with the SFST and just those people who fail the SFST are considered, 13% will have a BAC above 0.08%. That's a change in probability of just 3%. At 50% prevalence the change is 7%. In practice the changes in certainty created by a failed SFST are smaller than the uncertainty in the original guesstimate of pretest probability.

This finding is consistent with the likelihood ratios reported in Table 3. All are in a range, 1-2, that the American Medical Association teaches medical practitioners quantify the performance of tests as being so weak that they change probabilities to degree that is rarely clinically important.

Impairment

In their forensic use SFSTs are said to identify not just BAC but also "impairment." Stuster and Burns report only the accuracy of the SFST at identifying BAC. The study explicitly does not report the accuracy of the test at identifying "impairment." Any reference to the Stuster and Burns SFST's accuracy at identifying impairment derives from the test's reported accuracy at identifying BAC, and inherits the limitations we report here.

Not addressed

Officer's opinions: This paper addresses the statistics arrest accuracy and overall accuracy as applied to the HGN, WAT, OLS (index) and evidentiary BAC (reference) data in the Stuster and Burns data set. Stuster and Burns also reported the 91% overall accuracy of the dichotomized "officers' estimated BACs" (index) at identifying the dichotomized evidentiary BAC (reference). In making their estimations officers did not, as Stuster and Burns' interpretation criteria prescribe, predict BACs above or below the 0.08% target. Instead, they estimated BAC to two significant digits, recording 27 different predicted values from 0.00% to 0.26%. The method they used to do this is not reported. Officers not only did SFSTs, they also used a portable breath testing device and measured each driver's breath alcohol concentration. Officers knew the actual BAC of each driver before they handed in the data sheet recording their own unstandardized BAC estimate. Stuster and Burns report that the officers' BAC estimates were "within the margin of error of sophisticated evidentiary testing equipment" [6]. There being no reproducible connection between officers' BAC estimates and drivers' HGN, WAT, and OLS performance, the estimates are not considered here.

Conclusion

The statistic arrest accuracy identifies the Stuster and Burns SFST, as currently used by law enforcement, as 78% accurate. This number is an artifact of the prevalence dependence of arrest accuracy. Calculations independent of prevalence show that Stuster and Burns' SFST actually has no meaningful power to discriminate between drivers with high and low BACs. The statistics overall accuracy and arrest accuracy to not quantify the probability that impaired driving defendants who failed the SFST had an elevated BAC or were impaired.

References

1. Burns M, Moskowitz H (1977) Psychophysical tests for DWI arrest. Southern California Research Institute.
2. Tharp V, Burns M, Moskowitz H (1981) Development and field test of psychophysical tests for DWI arrest. Southern California Research Institute.
3. Anderson TE, Schweitz RM, Snyder MB (1983) Field evaluation of a behavioral test battery for DWI. U.S. Department of Transportation, National Highway Traffic Safety Administration.
4. Burns M, Anderson EW (1995) A Colorado validation study of the standardized field sobriety test (SFST) battery. U.S. Department of Transportation, National Highway Traffic Safety Administration.
5. Burns M, Dioquino T (1997) A florida validation study of the standardized field sobriety test (S.F.S.T.) battery. U.S. Department of Transportation, National Highway Traffic Safety Administration.
6. Stuster JW, Burns M (1998) Validation of the standardized field sobriety test battery at BACS below 0.10 percent. U.S.

Department of Transportation, National Highway Traffic Safety Administration.

7. DWI detection and Standardized Field Sobriety Testing (SFST) Participant manual (2018) US Department of Transportation, National Highway Traffic Safety Administration.
8. Rubenzer SJ (2008) The standardized field sobriety tests: A review of scientific and legal issues. Law Hum Behav 32: 293-313.
9. Downey LA, Hayley AC, Porath-Waller AJ, Boorman M, Stough C (2016) The Standardized Field Sobriety Tests (SFST) and measures of cognitive functioning. Accid Anal Prev 86: 90-98.
10. Yoshizuka K, Perry PJ, Upton G, Lopes I, Ip EJ (2014) Standardized field sobriety test: False positive test rate among sober subjects. J Forensic Toxicol Pharmacol 3.
11. Burns M (2007) The robustness of the horizontal gaze nystagmus test. U.S. Department of Transportation, National Highway Traffic Safety Administration.
12. Burns M (2003) An overview of field sobriety test research. Perceptual and Motor Skills 97: 1187-1199.
13. Cole S, Nowaczyk RH (1994) Field sobriety tests: Are they designed for failure? Perceptual and Motor Skills 79: 99-104.
14. Hlastala MP, Polissar NL, Oberman S (2005) Statistical evaluation of standardized field sobriety tests. J Forensic Sci 50.
15. Drug recognition expert course participant manual (2018) US department of transportation, National Highway Traffic Safety Administration.
16. Bosker WM, Theunissen EL, Conen S, Kuypers KP, Jeffery WK, et al. (2012) A placebo-controlled study to assess standardized field sobriety tests performance during alcohol and cannabis intoxication in heavy cannabis users and accuracy of point of collection testing devices for detecting thc in oral fluid. Psychopharmacology 223: 439-446.
17. Stough C, Boorman M, Ogden E, Papafotiou K (2006) Detection of impaired driving. In: An evaluation of the standardized field sobriety tests for the detection of impairment associated with cannabis with and without alcohol. National Drug Law Enforcement Research Fund monograph no. 17. Canberra. Australian Institute of Criminology.
18. Perry PJ, Fredriksen K, Chew S, Ip EJ, Lopes I, et al. (2015) The effects of dextromethorphan on driving performance and the standardized field sobriety test. J Forensic Sci 60: 1258-1262.
19. Downey LA, King R, Papafotiou K, Swann P, Ogden E, et al. (2012) Examining the effect of dl-3,4-methylenedioxymethamphetamine (MDMA) and methamphetamine on the standardized field sobriety tests. Forensic Sci Int 220: e33-36.
20. Downey LA, King R, Papafotiou K, Swann P, Ogden E, et al. (2012) Detecting impairment associated with cannabis with and without alcohol on the standardized field sobriety tests. Psychopharmacology 224: 581-589.
21. Declues K, Perez S, Figueroa A (2016) A 2-year study of d 9-tetrahydrocannabinol concentrations in drivers: examining driving and field sobriety test performance. J Forensic Sci 61: 1664-1670.
22. Ip EJ, Bui QV, Barnett MJ, Kazani A, Wright R, et al. (2013) The effect of trazodone on standardized field sobriety tests. Pharmacotherapy 33: 369-374.
23. Silber BY, Papafotiou K, Croft RJ, Stough CK (2005) An evaluation of the sensitivity of the standardised field sobriety tests to detect the presence of amphetamine. Psychopharmacology 182: 153-159.
24. DWI detection and Standardized Field Sobriety Testing (SFST) Instructor Guide (2018). US department of transportation, National Highway Traffic Safety Administration.
25. Mandrekar JN (2010) Simple statistical measures for diagnostic accuracy assessment. J Thorac Oncol 5: 763-764.
26. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, et al. (2016) STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. BJM 6.
27. Scott IA, Greenberg PB, Poole PJ (2007) Cautionary tales in the clinical interpretation of studies of diagnostic tests. Intern Med J 38: 120-129.
28. Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M (2004) The use of “overall accuracy” to evaluate the validity of screening or diagnostic tests. J Gen Intern Med 19: 460-465.
29. Grunau G, Linn S (2018) Detection and diagnostic overall accuracy measures of medical tests. Rambam Maimonides Med J, 9: e0027.
30. Furukawa TA, Straus SE, Bucher HC, Agoritsas T, Guyat G (2015) Diagnostic Tests. In: Guyatt G, Meade M, Rennie D, Cook, D., (Eds.) Users’ guides to the medical literature essentials of evidence-based clinical practice (3rd ed.). New York. McGraw Hill Education.
31. Carvajal DN, Rowe PC (2010) Sensitivity, specificity, predictive values, and likelihood ratios. Pediatrics in Review 31: 511-513.
32. Streiner DL (2003) Diagnosing tests: Using and misusing diagnostic and screening tests. J Pers Assess 81: 209-219.
33. Eusebi P (2013) Diagnostic accuracy measures. Cerebrovasc Dis 36: 267-272.
34. Chu K (1999) An introduction to sensitivity, specificity, predictive values and likelihood ratios. Emergency Medicine 11: 175-181.
35. Espallardo NL (2003) Decisions on diagnosis in family practice: Use of sensitivity, specificity, predictive values and likelihood ratios. Asia Pac Fam Med 2: 229-232.
36. Vecchio TJ (1966) Predictive value of a single diagnostic test in unselected populations. NEJM 274: 1171-3.
37. Akobeng AK (2007) Understanding diagnostic tests 2: Likelihood ratios, pre- and post-test probabilities and their use in clinical practice. Acta Pædiatrica 96: 487-491.
38. Florkowski CM (2008) Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: Communicating the performance of diagnostic tests. Clin Biochem Rev 29: S83-S87.
39. Attia J (2003) Moving beyond sensitivity and specificity: Using likelihood ratios to help interpret diagnostic tests. Australian Prescriber 26: 111-113.