Estimating Error Rates in Binary Decisions with Inconclusive Outcomes

Karen Kafadar

with Sydney Campbell and Jordan Rodu

Department of Statistics

University of Virginia

kkafadar@virginia.edu

http://statistics.as.virginia.edu/faculty-staff/profile/kk3ab

Acknowledgements: CSAFE (NIST, via Iowa State)

Center for Statistical Applications in Forensic Evidence (CSAFE)
NIST-Funded Center of Excellence

1. Iowa State Univ: Alician Carriquiry (PI)
2. Univ of California-Irvine: Hal Stern
3. Univ of Virgunia: Karen Kafadar
4. Carnegie Mellon Univ: Stephen Fienberg $\rightarrow$ Robin Mejia
5. West Virginia Univ (Criminal Justice): Keith Morris
6. Duke Univ (School of Law): Brandon Garrett

Motivation: **How accurate are Forensic Exam Processes?**

Given evidence from Crime Scene, Potential Suspect:

- **(A)** Truly from same source: True Match **(TM)**
- **(B)** Truly different sources: Non-match **(NM)**

But the examiner may offer *three* assessments:

- "Identification" (Right if **A**, wrongful conviction if **B**)
- "Exclusion" (Right if **B**, free perpetrator if **A**)
- "Inconclusive" (i.e., "DK" = "Don't know")

If only two answers were possible, calculating error rates would be straightforward. But with three answers?

True State	Examiner's call		
	"ID"	"Exclusion"	"Inconclusive"
A: True Match	Correct Conviction	Person free More crimes	Person may be free (commit more crimes)
B: True Non-Match	Wrongful Conviction	Correct exclusion	Person may be free (justifiably)

In (A): "Inconclusive" may be closer to "Wrong" answer:
Guilty person may commit more crimes (unless other evidence)

In (B): "Inconclusive" may be closer to "Right" answer

How do we account for "Inconclusives"?

OUTLINE

1. Problem: **Big differences** in reported error rates, depending on how "*Inconclusive*" decisions are counted
2. Literature review: Previous approaches
3. Proposed approach: standardization based on "difficulty"
4. Implementation
5. Further considerations

The Problem in Latent Print Studies

Ulery et al. 2011 *PNAS*, Appendix p.13:

Accuracy and reliability of forensic latent print decisions

- 169 examiners made decisions on 17,121 presentations of 744 image pairs: 11,578 "mates" (TM); 5,543 "non-mates" (NM)
- 3,389 TM (29.3%), 558 NM (10.1%): deemed "NV" (no value)
- Reported FPR = 6/5,543 = 0.11%, FNR = 611/11,578 = 5.3%
- Exclude "NV": FPR = 6/4,985 = 0.12%, FNR = 611/8,189 = 7.5%
- "Inconclusive" on 3,875 of 8,189 Mates of Value: 47.3%
- "Inconclusive" on 1,032 of 4,985 Non-Mates of Value: 20.7%
- **Should we count Inconclusive decisions as correct?**

Why it matters: Suppose $n = 200$ tests = 100 (TM) + 100 (NM)

Decisions: "M"/"NM"/"Inc", Ignore "Inconclusives"

1. TM: 80 / 20 / 0 (FNR=20%); NM: 30 / 70 / 0 (FPR=30%) $\Rightarrow$ 25%

2. TM: 60 / 0 / 40 (FNR= 0%); NM: 0 / 60 / 40 (FPR= 0%) $\Rightarrow$ 0%

Big difference! Not an academic problem:

- Ulery et al. 2011 (next slide)
- Baldwin et al. 2014: Cartridge case comparisons:
  746 "inconclusives": "$\sim$1%" if counted as "correct";
  22.8% if counted as "incorrect"

Ulery et al. 2011: "FPR = 6/5,543 = 0.1%, FNR = 611/11,589 = 5.3%"
(included 3,389 TMs and 558 NMs deemed "No Value" in denominator)

Include "*Inc*":	FPR	FNR
Yes, as "Correct"	6/4,985 = 0.12%	611/8,189 = 7.5 %
No (Ignored)	6/3,953 = 0.15%	611/4,314 = 14.2%
Yes, as "Incorrect"	1,038/4,985 = 20.8%	4,486/8,189 = 54.8%

Pacheco et al. 2014, *Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy and Precision In Latent Fingerprint Examinations* (NIJ report):

- 109 participants, 80 pairs: each saw 20 prints
  10 "mate" (TM) + 10 "non-mate" (NM) pairs
- Of 5,963 assessments: Excluded 1,411 deemed "NV"
  $\Rightarrow$ "3,138 Mates + 1,398 Non-mates = 4,536" (lost 16)
- Reported FPR & FNR, With/Without "*Inconclusives*"

Pacheco et al. 2014: Reported Error Rates

Include "*Inc*":	FPR	FNR
Yes, as "Correct"	42/1,398 = 3.0%	235/3,138 = 7.5%
No (Ignored)	42/995 = 4.2%	235/2,692 = 8.7%
Yes, as "Incorrect"	445/1,398 = 31.8%	681/3,138 = 21.7%

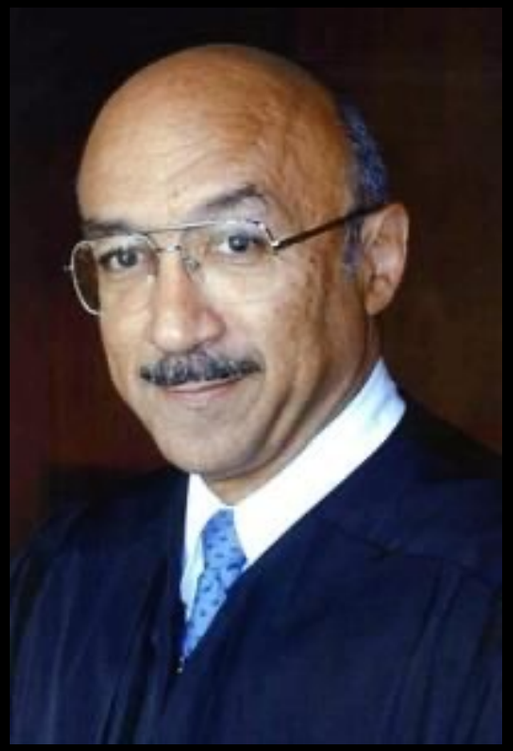
Pacheco et al. (2014) also give information on "difficulty":

Assessed Level	Mate (56 pairs)	Non-Mate (24 pairs)
Insufficient to Difficult	14 pairs 25%	6 pairs 25%
Difficult to Moderate	21 pairs 37.5%	9 pairs 37.5%
Moderate to Easy	21 pairs 37.5%	9 pairs 37.5%

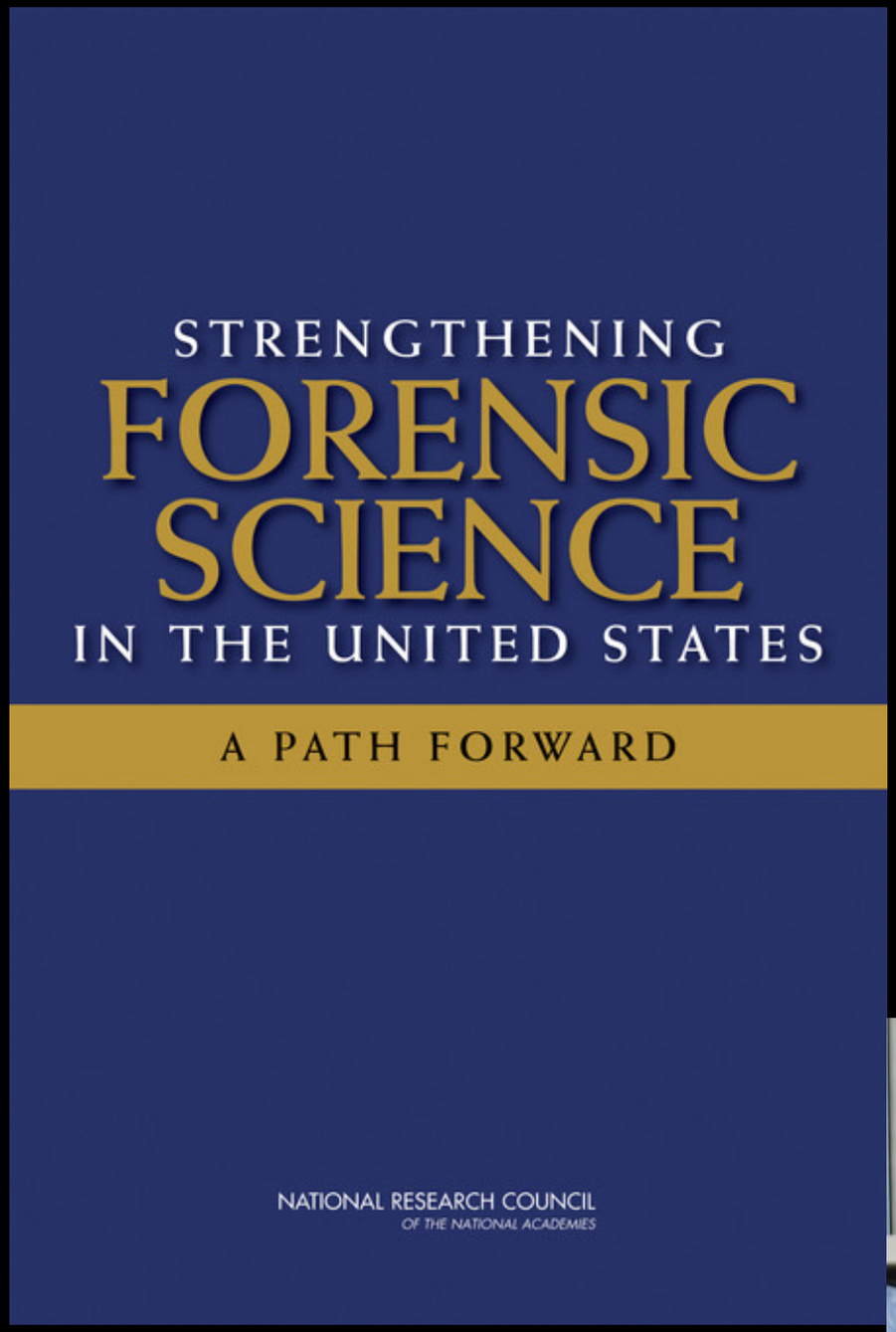
Lessons from these examples:

- Treatment of "Inconclusive" decisions greatly influences reported error rates
- Inconclusives occur in studies of accuracy in forensic decisions in many disciplines: latent prints, ballistics, hair analyses,...
- We need *consistent* treatment of "inconclusive" decisions when calculating error rates
- Studies of accuracy in diagnostic imaging also report "inconclusives"; in real life, doctor will request new image, additional tests, ... so "inconclusive" decisions may be less common/troublesome

- No standard guidelines on when to report "inconclusive"
- Some recommendations; mostly, labs have their own policies
- NRC (2009), *Strengthening Forensic Science in the United States: A Path Forward*: "*If neither an identification nor an exclusion can be reached, the result of the comparison is inconclusive*" (p138)
- *Organization of Scientific Area Committees* (OSAC) *Best Practices* (2020) provides a quality scale to help examiners evaluate prints in initial stage of latent fingerprint ID process
- Information on "Level of difficulty" (cf. Pacheco et al. 2014) would enable more fair comparisons across studies.



Hon. Harry T. Edwards



Prof Constantine Gatsonis, Brown

"With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source."

Source: National Research Council. 2009. *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: The National Academies Press. (p 7)

2. Literature: Current approaches

Three views on "inconclusives" in forensic exams:

1. Dror & Scurich (2020) *For. Sci. Int.*:

   - "Inconclusive" *can* be a "correct" decision
   - If so, then report error rate in usual way
   - How to decide if "Inconclusive" was correct?
   - Dror & Scurich: "Ask experts and take majority vote"

2. Hofmann, Carriquiry, Vanderplas (2021) *LPR*:

- No opinion on "*Inconclusive*" as "correct" or "incorrect"
- Report error rates when "inconclusives" are counted as (i) "correct" and (ii) "incorrect" and (iii) "ignored"
- What will jury do with three error rates?
- Layperson has trouble understanding even *one*
- Temptation to choose the rate of most convenience
  Prosecution: Cite lowest rate; Defense: Cite highest rate

3. Arkes & Koehler (2021) *LPR*:

- Inconclusives are not "always correct" nor "always incorrect"
- So ignore them in calculating error rates
- "Ignore them" $\approx$ "treat them as if they don't exist"?
- **Big** ranges in %inconclusives:
  Ulery et al.: 47% of 8,189 mates, 21% of 4,985 non-mates (37%);
  Pacheco et al.: 24% overall

Diagnostic imaging proficiency tests face similar issue
but not in real life: 'inconclusive' $\Rightarrow$ re-take image
We cannot "re-take" forensic evidence

3. Proposed Approach: Standardization

When might an examiner report "inconclusive"?

- Forensic evidence: Subjective, based on examiner's expertise
- Presumably, P{"*inconclusive*"} depends on *image quality*: *Higher* [*Lower*] probability if image quality is *Poor* [*Clear*]
- Can also depend on examiner's "risk level": *Higher* [*Lower*] probability if examiner is *More* [*Less*] risk averse, more cautious – even if image is clear
- Studies will differ in proportions of good/poor quality evidence
- OSAC *Best Practices* suggests examiners assess print quality using 6-point scale (0=poor, 5=excellent)

We know how to standardize rates when data can be stratified into categories having different proportions of an important factor

- Simpson's Paradox: Misleading conclusions can arise when we collapse data over relevant categories (e.g., quality)
- Analogy: we standardize incidence/mortality rates by proportions of Standard Population in 5-year age groups
- Almost impossible to assess examiner's "risk level"
- But we *can* objectively measure "latent print quality"

`forensicstats.org/quality-metric-algorithms-for-fingerprint-images`

Standardization:

- facilitates comparisons of reported error rates across studies *within* a discipline
- treats "inconclusive" decisions consistently across all studies, allowing us to assess whether "latent print exams" really are more "accurate" than "ballistics" or "handwriting"
- Requires knowing proportions of (standard) population that fall into categories of the relevant variable

Define a *standard distribution* of OSAC quality levels (0, ..., 5)

Pacheco et al. (2014) give us a good first start:

- "Insufficient to Difficult": 25%
- "Difficult to Moderate": 37.5%
- "Moderate to Easy": 37.5%

Existing databases of latent print images (e.g. NIST 302a) will enable translation of Quality Metric (QM) Score (0-100) to 6-point scale (0-5) $\Rightarrow$ proportions of prints having QM $= j$ and typical proportion of "inconclusive" decisions for QM_j-level prints.

Ex: Two studies, 1000 prints each, 3 outcomes:
Treat "Inconclusive" as 'Error":

Quality	#Prints	Correct	Incorrect	Inconcl	Error rate
Low	900	543	57	300	357/900 = 40%
High	100	97	1	2	3/100 = 3%
Total	1000	640	58	302	360/1000 = 36%

Quality	#Prints	Correct	Incorrect	Inconcl	Error rate
Low	200	100	80	20	100/200 = 50%
High	800	600	100	100	200/800 = 25%
Total	1000	700	180	120	300/1000 = 30%

If Inconclusive is treated as "correct":

Study 1: 57/900 = 6.3%, 1/100 = 1.0%, 5.8% overall;

Study 2: 80/200 = 40.0%, 100/800 = 12.5%, 18.0% overall

If Inconclusive is ignored:

Study 1: 57/700 = 8.1%, 1/98 = 1.0%, 8.3% overall;

Study 2: 80/180 = 44.4%, 100/700 = 14.3%, 20.4% overall;

Study 1 had more "Inconclusives" on "hard" cases (33.3% vs 10.0%), but fewer "Inconclusives" on "easy" cases (2.0% vs 12.5%).

Put studies on equal footing:

Apply *Stratum-Specific Rates* to *Standard Population* dist'n

Ex: 50-50 split in Low-High quality prints in *both* studies:

Study 1: 40% (SE 1.2%); Study 2: 50% (SE 2.2%)

Treatment of "inconclusive" decisions:

- "Correct"
- "Incorrect"
- "Ignored"
- "half-incorrect if Low quality";
  "full-incorrect if High quality"
  (don't penalize entirely if poor-quality evidence)

Extend: weight accuracies within six quality categories:
Quality=0: Weight = 1.0; Quality=1: Weight = 0.8 ...
Quality=4: Weight = 0.2; Quality=5: Weight = 0.0

Count half [all] of "Inconclusives" in Low [High] categories as "incorrect" and use Standard Pop distribution (50-50):

Quality	Prints	Correct	Incorrect	Inconcl	Std rate
Low	900	543	57	300	115/500 = 23%
High	100	97	1	2	15/500 = 3%
Total	1000	640	58	302	130/1000 = 13%

Quality	Prints	Correct	Incorrect	Inconcl	Std rate
Low	200	100	80	20	225/500 = 50%
High	800	600	100	100	94/500 = 19%
Total	1000	700	180	120	319/1000 = 32%

4. Implementation

How to proceed?

- Standardization will require reasonable assessment of *proportions of inconclusives by quality category*
- Presumably, $p_j = \mathrm{P}\{\text{“}\textit{Inconclusive}\text{”} \mid \text{Quality level } j\}$ decreases as j increases: higher quality $\Rightarrow$ fewer inconclusives:

$$p_0 > p_1 > p_2 > p_3 > p_4 > p_5$$

- This “standard distribution” will penalize labs for deciding “inconclusive” out of an abundance of caution (these studies are *not* blind, much less double-blind)
- Houston Forensic Science Center (one of very few US crime labs *not* housed in Police Dept) will help (need others)

5. Conclusions & Further Considerations

Standardization allows:

- Consistent treatment of "*inconclusive*" decisions when estimating error rates
- P{"*Inconclusive*"} sensibly depends on latent print quality
- Facilitates honest comparisons of error rates across studies *within* a forensic discipline *as well as* across disciplines; e.g., are error rates for Latent Prints $<$ those for bite marks?
- Applies to other scientific disciplines where "*Inconclusive*" is an acceptable answer (e.g. diagnostic imaging)
- Changing the forensic culture will not be easy.

Thank you!

Some references

Arkes HR, Koehler JJ (2021), *Law, Probability & Risk* 20:153-168

Dror IE, Scurich N (2020), *Forensic Sci Int'l Synergy* 2:333-338

Fiumara GP et al (2019), *NIST Soecial Database 302*

Hofmann H, Carriquiry A, Vanderplas S (2021), *Law, Probability & Risk* 19:317-365

National Research Council (2009), *Strengthening Forensic Science in the United States: A Path Forward*, `www.nap.edu`

Pacheco I, Cerchiai B, Stoiloff S (2014), Final Technical Report, Award Number 2010-DN-BX-K268

Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2011), *PNAS* 108.19:7733-38.