

# Biased by Design

## How Poorly Designed Experiments And Surveys Create And Contribute To Urban Legends



Cliff Spiegelman  
Maria Cuellar, Lucas  
Mentch, David Sheets, &  
Bill Tobin

# Co-Authors



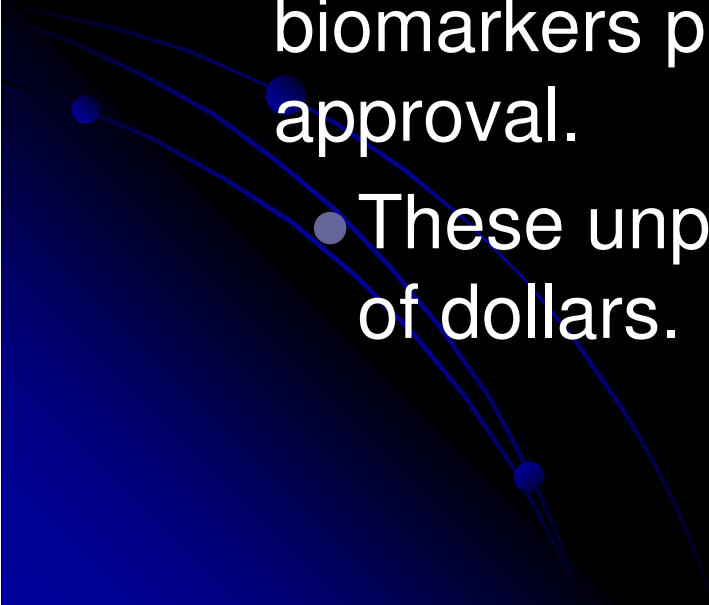
## Two focal points of the talk

- Flawed design examples using biomarkers and forensic markers
- Flawed design examples including sample size flaws using firearm/toolmarks

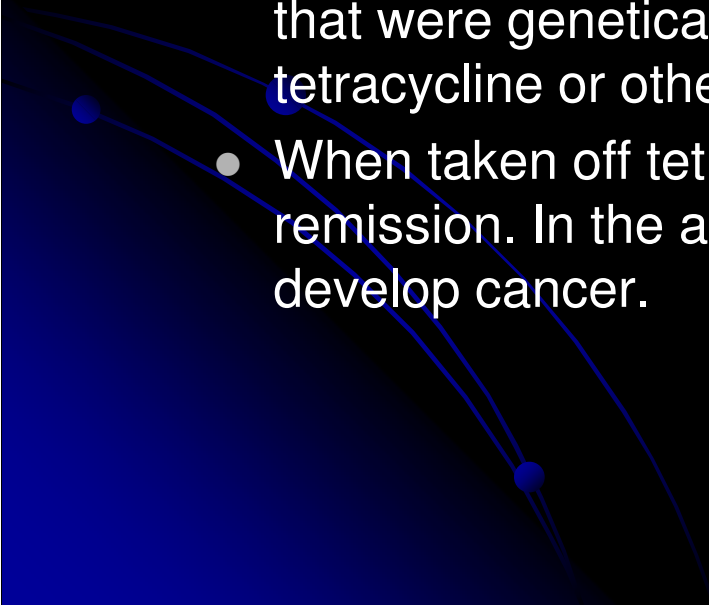
# How poor study designs lead to supporting the 'urban legends' of science

- Examples include:
  - Cancer biomarker searches
  - Arson forensic markers
  - Abusive head trauma markers (shaken baby syndrome)
  - Near zero error rates for matches of guns to crime scene bullets
    - And a near endless list of other forensic findings

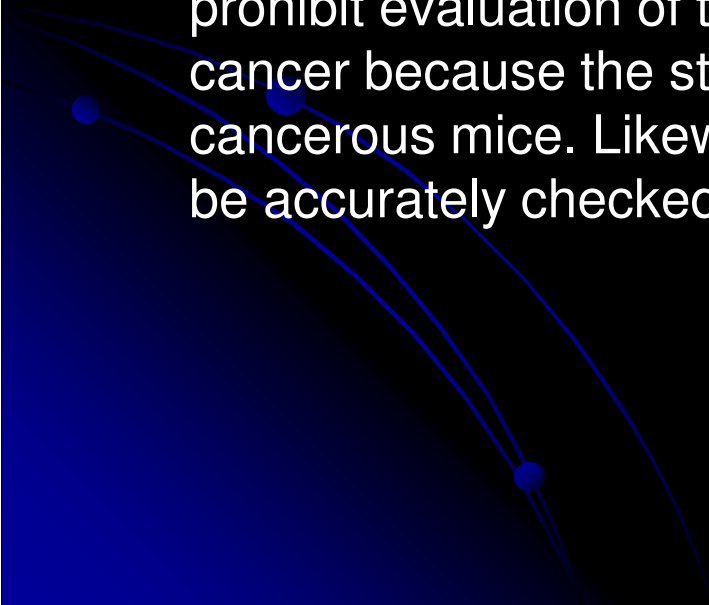
# Flawed Biomarker and Forensic Marker Searches

- The search for biomarkers for early stage cancer and other diseases is a major scientific and medical undertaking.
    - There are 1000s of cancer biomarker papers published each year, and only about 2 cancer biomarkers per year are granted FDA approval.
    - These unproductive searches cost billions of dollars.
- 

## We examine some of the poor study designs used in most non-FDA approved biomarker studies

- A common study designs for cancer biomarkers uses genetically altered mice (see Kelly- Spratt et al, 2008, Zhang et al, 2015) and are commonly referred to as mouse models.
    - A common study designs for cancer biomarkers uses genetically altered mice (see Kelly- Spratt et al, 2008, Zhang et al, 2015) and are commonly referred to as mouse models.
    - The design uses pairs of litter mates (1-treatment & 1-control) that were genetically altered to have a specified cancer when fed tetracycline or other antibiotics.
    - When taken off tetracycline, the cancer temporarily goes into remission. In the absence of tetracycline, the mice do not develop cancer.
- 

# Looking more deeply into this design

- It is apparent that neither the cancer groups nor the non-cancer groups are representative of their respective target populations. The cancerous mice do not have colds, medicines, broken bones, bruises, sore throats, and the cancer-free “normal” mice are all too normal.
  - None of the cancer mice have other forms of cancer, tuberculosis, flu, bruises, moles, etc. As a result, these confounding conditions prohibit evaluation of the sensitivity of the biomarkers to the target cancer because the study design has no confounding factors in the cancerous mice. Likewise, the selectivity for cancerous mice cannot be accurately checked because the normal mice are too normal.
- 

# Forensic Markers for AHT

- For certain studies about Abusive Head Trauma (AHT) (Maguire et al. 2009, Maguire et al. 2011, and Cowley et al. 2015) the designs have possible bias as well. The first evident source of bias is that the population in these studies is nonrandom, since it was selected from children's hospitals with physicians and child abuse specialists who were aware of and interested in AHT.
  - For example, one would not look for markers of battered women by using women only at battered women's shelters.
  - Another source of bias is introduced by the responses used in the studies.



## AHT continued

- Specifically, the authors of these studies determine that a child was abused if the child fulfills either of these two criteria: “Abuse confirmed at case conference or civil, ...“Abuse confirmed by stated criteria including multi-disciplinary assessment.” Thus, it is possible and even likely that some children recorded as not having been abused, were in fact abused, but this could not be confirmed according to these criteria.

# One of many sad AHT false conviction stories

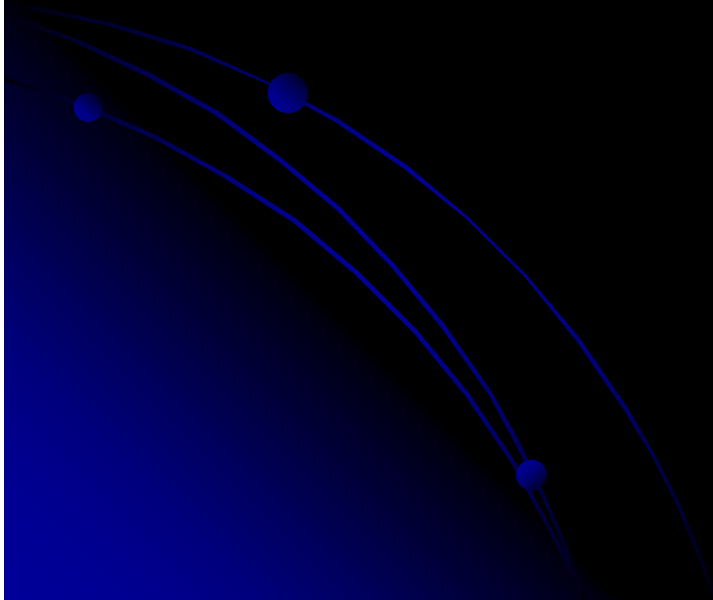
- Court proceedings have been reversed due to wrongful convictions. See the case of Julie Baumer <https://www.law.umich.edu/clinical/innocenceclinic/acceptedcases/Pages/baumer.aspx>, for example, who was taking care of her six-week-old nephew in 2003. The infant had a specific type of retinal bleeding (a common biomarker for AHT), and the physician attributed the infant's brain and eye injuries to AHT. Baumer was convicted of first-degree child abuse by a jury in 2005 and spent four years in prison. But new medical opinion arose that venous sinus thrombosis could cause the same bleeding symptoms as AHT. So, in 2010 at her second trial, the jury found Baumer not guilty of child abuse. So among other events, the non-specificity of a biomarker caused Baumer to be wrongly convicted.

# Arson Biomarkers

- The last several decades have seen drastic changes in the way fire scenes are processed and analyzed for arson indicators. A number of indicators such as large alligatoring (deep patterns of severe scorching) and crazed glass (unusual fracture patterns in glass surfaces) once thought to definitely indicate the presence of accelerants – many even supported at the time by the National Bureau of Standards (NBS – now the National Institute of Standards and Technology (NIST)) – are now largely recognized as no more than myth (1980). For a more thorough history, including once approved arson indicators such as spalling and V-pattern angles, we refer the reader to Lentini (2013) and Tobin (1990)

# Choosing Hypotheses, Factors & Sample Sizes Carelessly

- Firearm/Toolmarks is generally regarded by forensic examiners as just behind fingerprints in development as a pattern evidence science.
- The following are typical of 'proof of low error rates'



**Law, Probability and Risk Advance Access published October 1, 2012**

*Law, Probability and Risk* (2012) 0, 1–19

doi:10.1093/lpr/mgs028

**Analysis of experiments in forensic firearms/toolmarks practice offered as support for low rates of practice error and claims of inferential certainty**

CLIFFORD SPIEGELMAN

*Distinguished Professor of Statistics, Department of Statistics, Texas A&M University,  
College Station, TX, USA*

AND

WILLIAM A. TOBIN<sup>†</sup>

*Forensic Engineering International, 2708 Little Gunstock Road, Bumpass, VA, 23024-8882, USA*

This article critically evaluates experiments used to justify inferences of specific source attribution ('individualization') to '100% certainty' and 'near-zero' rates of error claimed by firearm toolmark examiners in court testimonies, and suggests approaches for establishing statistical foundations for firearm toolmarks practice that two recent National Academy of Science reports confirm do not currently exist. Issues that should be considered in the earliest stages of statistical foundational development for firearm toolmarks are discussed.

*Keywords:* firearms identification; forensic certainty; validation studies; ballistics; forensic error rate; forensic testimony.

# Firearm/toolmarks

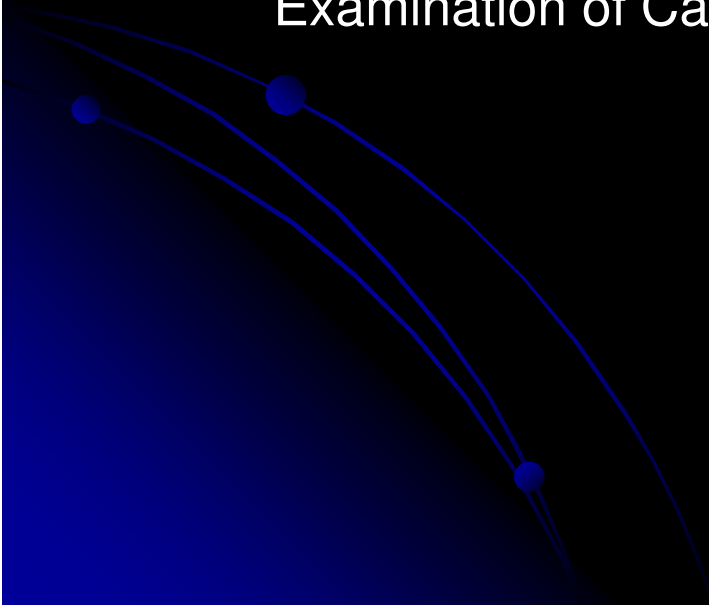
- Firearm toolmarks examinations and comparisons are often used in investigations of homicides involving a firearm, spent bullets and/or cartridge cases that are recovered from crime scenes. Most frequently, one or more firearms are recovered during investigation of a shooting incident and typically submitted for forensic comparisons with bullets and/or cartridge cases recovered from the scene. **The forensic practice used to associate or eliminate a particular firearm as the murder weapon is based on comparisons of characteristics imparted to bullets and cartridge cases during cycling of a cartridge through the gun, and is known as firearm/toolmarks examination.**

# Typical testimony

If it is concluded that the submitted weapon “matches” the crime scene bullets, the firearm/toolmarks examiner typically testifies at trial that the crime scene bullets were fired from the gun to the exclusion of all other possible weapons, although sometimes “to a practical certainty.”

# Poorly Formulated Hypotheses

- The stated hypotheses are, “1) that marks imparted to cartridge cases from different guns rarely if ever display sufficient agreement to lead qualified firearms examiners to conclude the specimens were fired from the same gun and 2) that marks imparted to cartridge cases from the same gun will rarely if ever lead a qualified firearms examiner to conclude the specimens were fired from different guns.”
  - From: “A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases” by Bunch and Murphy (2003).





# Restated Hypotheses

$H_{01}$ :  $P\{\text{Marks imparted to cartridge cases from different guns display sufficient agreement to lead qualified firearms examiners to conclude specimens were fired from same gun}\} \geq \delta$ .

$H_{a1}$ :  $P\{\text{Marks imparted to cartridge cases from different guns display sufficient agreement to lead qualified firearms examiners to conclude specimens were fired from same gun}\} < \delta$ .

And,

$H_{02}$ :  $P\{\text{Marks imparted to cartridge cases from same gun will lead a qualified firearms examiner to conclude specimens were fired from different guns}\} \geq \delta$ .

$H_{a2}$ :  $P\{\text{Marks imparted to cartridge cases from same gun will lead a qualified firearms examiner to conclude specimens were fired from different guns}\} < \delta$ .

# Factors

- Factors:
- Ammunition brand and perhaps ammunition lot
  - Charge
  - Shape
  - Undersized or not
- Firearm brand and perhaps lot
  - Caliber
  - Rifling

# Factors Continued

- Examiners' labs
  - Methods CMS or not CMS
- Experimental representativeness
- Blinding?
- Contextual Information?
  - ..... and so on

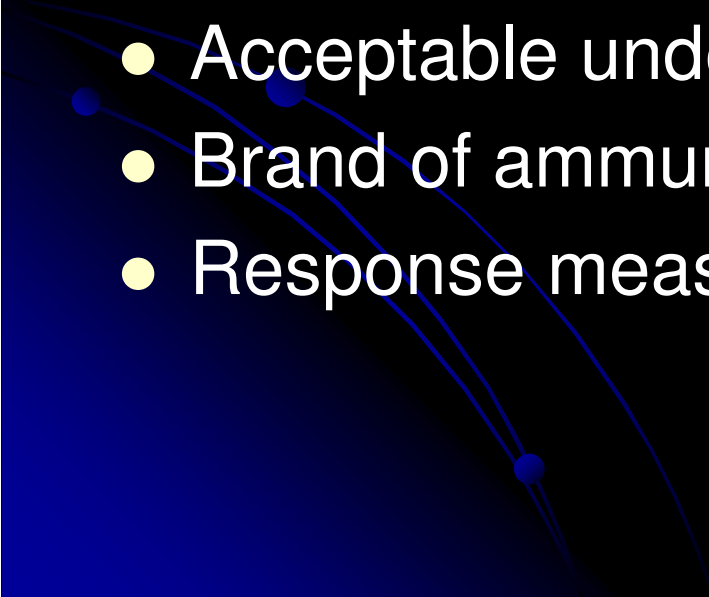
# Factors Chosen by Experimenters

- 10 casings given to each examiner which were a mix of Glock and other cartridge casings. Examiners were to treat the challenge as casework.
- Factors and sample sizes chosen by experimenters were:
- Ten Glock Pistols (9 MM) with consecutively manufactured *breechfaces*
- One Beretta model 92F (Luger 9mm)
- One SigSauer model P226 (Luger 9mm)
- Eight **FBI** firearm/toolmarks examiners
- 42 test fires from consecutively manufactured Glocks
- 318 other cartridges were used; it is difficult to know from what weapons the 318 other cartridges were fired

# Factors or Variables **Missing** from the 'Validity Study'

- Factors not considered or discussed, and sample size issues:
- Ammunition type
- Ammunition charge
- Cartridge case hardness
- Primer cup hardness
- Breechface hardness
- Firing pin hardness
- Different batches of Glocks

## Missing factors continued

- Non-Glock firearms “feeds and speeds” of production (or, alternatively, economic conditions of the manufacturing environment)
  - Batch and sample not chosen randomly;
  - Two 9 mm Lugers were not chosen randomly
  - Acceptable undersized ammunition
  - Brand of ammunition
  - Response measures
- 

## Factors or Variables **Missing** from the 'Validity Study' (**Continued**)

- Ids from firing pins
- Ids from ejector marks
- Ids from breechface marks
- Ids from combination of ejector, firing pin, and breechface marks
- Weapon cleanliness
- Participants' experience as toolmarks examiners
- Fabrication tooling hardness
- Type of workpiece
- Alloy used for workpiece

## Missing Factors continued

- Participants asked to handle test as casework, but no measures of effectiveness for this instruction reported
- Break-in period for pistols
- Lubrication regime, present or not
- Condition of lubrication system, nested within lubrication regime
- Fabrication tooling materials, if any



# So A Reformulation Of The Hypotheses?

- Should hypotheses be reformulated condition on the factors used. For example, 'clear marks'

$H_{01} : P\{\text{Marks imparted to cartridge cases from different guns display sufficient agreement to lead qualified firearms examiners to conclude specimens were fired from same gun} \mid \text{clear marks}\} \geq \delta .$

$H_{a1} : P\{\text{Marks imparted to cartridge cases from different guns display sufficient agreement to lead qualified firearms examiners to conclude specimens were fired from same gun} \mid \text{clear marks}\} < \delta .$

## Stated Conclusions

- There were 70 true IDs (positives) that examiners could make, and 100 percent were made. Of 290 true exclusions, examiners made 118, the remainder was declared inconclusive. The two propositions tested by this experiment were confirmed, that is, the two null hypotheses were rejected. [No significance level is reported.]

# Comments on the experiment and stated conclusions

- There was no meaningful SOP or detailed criteria of how matches were made or not. Factors used in the experiment were too few: only three types of weapon with unbalanced numbers of each were used, the type(s) of ammunition used is unknown, and automated toolmarks equipment has shown that ammunition brand matters for automated identification, Bachrach (2006), all examiners from one organization, one manufacturing batch used for the ten Glocks, and unknown number of break-in test fires.

# Comments continued

- The fact that **examiners were not chosen at random** but rather from what is considered to be an elite unit in the FBI Crime lab limits generality. **Even if we pretend that the eight examiners are typical, an exact 95% confidence interval for the percentage of examiners that would also be perfect on a similar exam is approximately 63 to 100 percent.**

## Comments continued

- Thus, a fair assessment would be that at least 63 percent of examiners would do as well on this experiment. The typical sample sizes are small, one group of examiners, three types of weapons (two with sample size 1). Thus, the experiment provides little support for the conclusions presented. In addition, considering the fact that the level of most factors was not recorded (e.g., cartridge or breechface hardness, or chamber pressure), it is difficult to see how different studies, without properly measured factors, can be combined to assess cartridge IDs.

## Comments continued

- In addition, examiners knew they were being tested and, according to well-accepted principles, this creates a challenge to applying results of this study to general toolmarks community actual casework.

See the first two chapters of Cook and Campbell (1979), and Shadish, Cook, and Campbell (2002) for factors affecting the validity of field experiments. Examiners could estimate how similar markings on test fires were for the different (consecutively manufactured) weapons provided. In essence, they observed what statisticians call 'between-variation', or an intuitive feel for the 'between sums of squares'. This is quite unusual in actual forensic casework.

# References

- **References**
- Boris Freidlin, Lisa M. McShane, Edward L. Korn, Randomized Clinical Trials With Biomarkers: Design Issues, *JNCI J Natl Cancer Inst* (2010) doi: 10.1093/jnci/djp477 .477.
- Karen S. Kelly-Spratt, A. Erik Kasarda, Mark Igra, and Christopher J. Kemp, A Mouse Model Repository for Cancer Biomarker Discovery, *J Proteome Res.* (2008); 7(8): 3613– 3618. doi:10.1021/pr800210b.
- Kim, Kelly Y. and McShane, Lisa M. and Conley, Barbara A., Designing biomarker studies for head and neck cancer, *Head & Neck*, 36,7, 1069-1075.
- L. E. Cowley, C.B. Morris, S.A. Maguire, D.M. Farewell, A.M. Kemp. *Validation of a Prediction Tool for Abusive Head Trauma.* *Pediatrics*, Vol. 136, No. 2, 2015.
- Mei-Yin C. Polley, Boris Freidlin, Edward L. Korn, Barbara A. Conley, Jeffrey S. Abrams, and Lisa M McShane, Statistical and Practical Considerations for Clinical Evaluation of Predictive Biomarkers, *JNCI J Natl Cancer Inst* (2013) 105 (22): 1677- 1683 doi:10.1093/jnci/djt282.
- NFPA. (1992). *NFPA 921 - Guide for Fire and Explosion Investigations* (1992 ed.). Quincy, MA: National Fire Protection Association.
- Lentini, John J., and FABC CFEI. "The mythology of arson investigation. "*Scientific Fire Analysis*: <http://www.firescientist.com>.(accessed 16th August 2013).

# References (continued)

- Polley, M.-Y. C., Freidlin, B., Korn, E. L., Conley, B. A., Abrams, J. S., & McShane, L. M. (2013). Statistical and Practical Considerations for Clinical Evaluation of Predictive Biomarkers. *JNCI Journal of the National Cancer Institute*, 105(22), 1677–1683. <http://doi.org/10.1093/jnci/djt282>.
- S. Maguire, A.M. Kemp, R.C. Lumb and D.M. Farewell. *Estimating the Probability of Abusive Head Trauma: A Pooled Analysis*. *Pediatrics* Vol. 128, No. 3, 2011.
- S. Maguire, N. Pickerd, D. Farewell, M. Mann, V. Tempest, A.M. Kemp. *Which clinical features distinguish inflicted from non-inflicted brain injury? A systematic review*. *Archives of Disease in Childhood*. Vol. 94, No. 11, pp. 860–867, 2009.
- S.M. Kassin, K.L. Kiechel. "The Social Psychology of False Confessions: Compliance, Internalization, and Confabulation." *Psychological Science*, Vol. 7 No. 3 125-128, 1996.
- Zhang, H., Cao, J., Li, L., Liu, Y., Zhao, H., Li, N., ... Chen, L. (2015). Identification of urine protein biomarkers with the potential for early detection of lung cancer. *Scientific Reports*, 5, 11805. <http://doi.org/10.1038/srep11805>.
- US Dept. of Commerce, and National Bureau of Standards. "Fire Investigation



# References (continued)

- Collapsed Springs in Arson Investigation: A Critical Metallurgical Evaluation, W.A. Tobin & K.L. Monson, Fire Technology, Volume 25, Number 4 (November 1989), National Fire Protection Association.
- Arson Investigations, W.A. Tobin, Law Enforcement Bulletin ('Focus' feature), February 1990, Federal Bureau of Investigation.
- What Collapsed Springs Really Tell Arson Investigators, W.A. Tobin, Fire Journal, Volume 84, No. 2 (March/April 1990), National Fire Protection Association.
- What Collapsed Springs Really Tell Arson Investigators, W.A. Tobin; course instructional material, Fire/Arson Investigation Resident Course, October 1994, U.S. Fire Administration, National Fire Academy; requested and reprinted with permission for inclusion in NFPA 921 Guide to Arson Investigations.