Evaluating the Gap between Audit Standards and Practices: A Case Study on Probabilistic Genotyping Software Validation

Angela Jin
angelacjin@berkeley.edu
University of California, Berkeley
Berkeley, California, USA

Alexander Asemota
alexander.asemota@berkley.edu
University of California, Berkeley
Berkeley, California, USA

Dan E. Krane
Wright State University
Dayton, Ohio, USA

Nathaniel Adams
Forensic Bioinformatic Services
Fairborn, Ohio, USA

Rediet Abebe
ELLIS Institute, Max Planck Institute
for Intelligent Systems, &
Tübingen AI Center
Tübingen, Germany

Abstract

Efforts to make algorithm audits effective tools for accountability have highlighted the design of auditing standards as a crucial point of intervention. In this paper, we present ongoing work studying how a standard can fail to achieve its goals in practice through a case study of probabilistic genotyping software validation. We present findings on the gap between ASB 018, a technical standard for validating probabilistic genotyping software, and validation studies in practice. Examining three publicly available validation study reports, we find that all three validation studies fail to meet the guarantees the standard seeks to provide about these studies, but can still claim compliance with the formal requirements of the standard. We demonstrate how specific characteristics of the standard enable these compliant, yet unsatisfactory validation practices. To conclude, we discuss implications of our findings for the design of ASB 018 and for audit standards more broadly.

CCS Concepts

• **Social and professional topics → Government technology policy**; • **Software and its engineering → Software verification and validation**; • **Applied computing → Law**.

Keywords

audit standards, criminal legal system, accountability

ACM Reference Format:
Angela Jin, Alexander Asemota, Dan E. Krane, Nathaniel Adams, and Rediet Abebe. 2025. Evaluating the Gap between Audit Standards and Practices: A Case Study on Probabilistic Genotyping Software Validation. In *CHI '25 Workshop on Sociotechnical AI Governance (STAIG @ CHI '25), April 27, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

STAIG @ CHI '25, April 27, 2025, Yokohama, Japan

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recent work on auditing algorithmic systems has called for standardization of audits to ensure audit quality and consistency [8, 9, 12, 13]. Yet it is unclear how to design these standards to ensure audits are effective tools for accountability. Prior work has highlighted that efforts to standardize face a number of challenges, such as balancing the level of specificity in the standard's requirements to avoid promoting a check-list mentality [13] and assessing what amount and types of testing are sufficient to thoroughly investigate potential harms [12]. In this work, we explore audit standards and the auditing practices subject to those standards, with a specific focus on whether a standard ensures practices that align with its goals. We ask, *How might the design of an auditing standard allow for practices that fail to meet the standard's goals but nevertheless comply with its requirements?*

We adopt a case study approach, looking at this question in the context of the U.S. criminal legal system and specifically focusing on the ASB 018 standard [5] for performing internal validation (IV) studies of probabilistic genotyping software (PGS) commonly used to analyze challenging forensic DNA samples in criminal cases[1]. ASB 018 is the first standard for PGS validation to be developed through a process formally accredited against requirements for openness and due process, and we focus on its efforts to standardize internal validation (IV) studies – testing carried out by a forensic DNA lab after they acquire the PGS system but before they may use it in casework. We seek to understand *whether and how ASB 018 might allow for IV studies that comply with its requirements but fail to meet the guarantees ASB 018 seeks to provide.*

Assessing the potential gap between characteristics of IV studies that the standard envisions and promotes in its language and those that the standard enables in practice is critical, given the roles that IV studies and PGS validation standards play in courtroom debates over PGS reliability. When a forensic analyst presents a PGS output

[1]Figure 1 in the Appendix provides a simplified depiction of how PGS is used for forensic DNA testing. At a high level, a PGS system takes two DNA profiles as input: an evidence sample obtained from the crime scene and a reference profile from the person of interest (POI). The PGS system computes a likelihood ratio (LR) that compares the likelihood of observing the DNA evidence under two competing hypotheses: one that includes the POI and one that does not. For example, an LR may compare a hypothesis that the POI and two unknown others contributed to the DNA mixture against a hypothesis that three unknown individuals who are not the POI contributed to the DNA mixture.

as evidence, the analyst may proffer the lab's IV study to argue that the software performed reliably in this case and that the results should therefore be admitted as evidence [1]. Then, a key question that shapes the impact of the IV study is whether the study is of sufficient quality to support such an argument. Key decision-makers like judges and jurors are tasked with answering this question and may understand the IV study's purported compliance with ASB 018 to guarantee that the IV study is of sufficient quality to back such an argument. Our work demonstrates a need to be more guarded in this reliance on ASB 018 by demonstrating a gap between what the standard seeks to guarantee about compliant IV studies and what it actually ensures in practice.

We conducted our study in three parts, first conducting reflexive thematic analysis [6] of ASB 018 to understand the guarantees it seeks to establish for compliant IV studies. We interpret the standard's envisioned guarantees by drawing on the standard and the factsheet that ASB provides to accompany the standard. Next, we examine IV study practices through the lens of these guarantees, focusing on three publicly available IV study reports. Lastly, we inductively coded ASB 018's requirements and analyzed these codes alongside codes and high-level themes from our investigation of IV practices, to analyze how specific characteristics of the requirements enable observed patterns in IV study practices.

2 Results: Compliant validation practices frequently fall short of meeting the guarantees the standard seeks to provide about compliant studies.

In this section, we present each of the four guarantees ASB 018 seeks to provide about compliant IV studies and describe how IV practices fail to meet each guarantee. These findings suggest that the characteristics and quality of IV studies that comply with the standard may critically fall short of the qualities of IV studies that the standard envisions and advertises to those who rely on the standard to assess IV study quality and PGS reliability.

Guarantee 1. *Compliant IVs establish software limitations – boundaries around the use of software in casework.* The AAFS factsheet for the standard provides crucial context suggesting that limitations are boundaries on software use in casework. The factsheet elaborates that validation studies, including IV studies, of PGS *"define limitations for its use"* that *"establish the range of DNA profiles upon which the program may be used effectively"* (AAFS Factsheet). However, we find that **compliant studies rarely translated errors into actionable boundaries on software use.** The IV summaries we examined rarely mentioned, let alone established, boundaries, even when findings demonstrated or suggested potential for false positive or false negative LRs. For example, one lab's analysis of false positives revealed how studies may also dismiss errors as 'not the fault of the software' instead of recognizing that the software produced an LR supporting the false hypothesis. On the other hand, crucial boundaries that studies *did* acknowledge were not clearly communicated in a manner that would enable consistent application in casework. For example, studies frequently pointed to ambiguous plots as demonstrating limitations, though those limitations did not translate to objective restrictions on PGS use in casework (e.g., Figure 2 in the Appendix).

Guarantee 2. *Compliant IVs establish software limitations by evaluating software performance against falsifiable expectations.* We broadly define falsifiable expectations as expectations against which studies can assess observed software behavior and decisively state whether the observed behavior succeeds, or fails, in meeting that expectation. Importantly, in order for these expectations to support establishing limitations, it must be possible to identify when software behavior fails to meet the expectation. However, we find that compliant studies relied on **subjective and contradictory expectations make it difficult to reach mutual understandings on what expectations are and when expectations have been violated, a crucial step to establishing boundaries on use.** For example, all three reports relied on an expectation that the LR for true contributors should be 'high' when there is a 'high' template amount (i.e., DNA amount). However, 'high' is not defined in either case. Almost anything could be argued to meet these criteria, and this flexibility precludes any conclusion that the software fails to meet the expectation – conclusions critical to establishing when the software should or should not be used.

Guarantee 3. *Compliant IVs establish software limitations under conditions representative of casework conditions*, i.e., testing the software using the same equipment and procedures that will be used in casework. This representativeness not only ensures that these studies produce a *"realistic assessment of the readiness of the system for casework"* (ASB 018, p. 6), but also ensures that studies help inform lab-specific procedures for PGS use in casework. However, we find that **the testing environment and procedures used in IV studies may not represent that of casework.** Crucially, we find that software developers, and not users, may conduct significant portions of IV studies, and may fail to use lab-specific procedures and practices such as those used to prepare software inputs like the inferred number of contributors (inferred NoC).

Guarantee 4: *Compliant IVs conduct testing on sample types and sample sizes that are sufficient and appropriate for establishing limitations.* The number and types of samples tested should be *"**appropriate** [...] to demonstrate the potential limitations and reliability of the software"* (ASB 018, p. 3), and *"[w]hile specific requirements for the minimum number of studies and sample sets used for validation studies are not detailed in this standard, the laboratory shall perform **sufficient** studies to address the variability inherent to the various aspects of DNA testing, data generation, analysis and interpretation of data and user input parameters"* (ASB 018, p. 5). However, **we find no evidence of studies considering or assessing the sufficiency or appropriateness of the number or types of samples tested.** For example, all studies included an aggregate description of the samples tested, passively phrased almost exactly as follows: *"The profiles are of varying DNA quantity and mixture proportions."*

3 Results: ASB 018 seeks to enact change through isolated, vague, and incomplete requirements that enable compliant yet unsatisfactory internal validation studies.

Importantly, the studies we discussed in the previous section could still be interpreted as complying with ASB 018, despite failing to meet the guarantees ASB 018 seeks to establish for compliant IV

studies. In this section, we discuss how the standard enables such compliant yet unsatisfactory practices.

ASB 018 Design Choice 1: The standard emphasizes establishing limits, but fails to define what a limitation entails, and does not require studies to document limitations. The studies we examined did not clearly articulate boundaries, even when errors were identified. Instead, studies relied on plots for demonstrating limitations and argued that errors were not the failure of the software. These findings suggest a lack of clarity, in the standard, about what constitutes a software limitation. **Despite the repeated mention of studies establishing limitations, ASB 018 does not provide a clear definition of what a limitation is.** Details in the accompanying factsheet helped us infer what a 'limitation' may entail, but this detail is crucially lacking from the standard, the sole document that studies would be assessed against for compliance.

In addition to failing to specify what a limitation entails, ASB 018 does not include any requirement to document conclusions about software limitations. In other words, **the standard seeks to guarantee that studies establish limitations, but makes no requirement that studies actually do so**. For example, the standard only requires that studies *"address [...] accuracy, sensitivity, specificity, and precision"* (R4.1.3), without any requirements to establish limitations with respect to measurements of these characteristics of software performance.

Even if the standard included some requirement to establish limitations, ASB 018 fails to address a crucial ingredient for establishing limitations: specific, falsifiable expectations against which studies evaluate system behaviors. **While ASB 018 mentions assessing software performance against expectations, it fails to elaborate further on what those expectations should look like, and does not establish any requirements that set expectations or require the testing entity to set their own expectations.** Without such requirements, studies may define any expectations, including expectations that undermine the validity of conclusions that the software performs as expected, or none at all.

ASB 018 Design Choice 2: The standard relies on a checklist of disconnected requirements that fails to support establishing boundaries on software use. The standard **addresses the design of test scenarios across multiple, independent requirements about types of DNA profiles, types of artifacts, and types of input parameters, where some characteristics of mixture profiles are discussed alongside requirement to address performance characteristics** (R4.1.3)**, and others are presented as standalone requirements**, such as: *"Studies shall include evaluating user input parameters that vary run to run."* (R4.1.4). As a result, validation practices can address the input factors mentioned in these latter requirements in completely separate, standalone studies, with tests conducted on single samples. Such practices would still comply with the individual requirements, while failing to address factors that crucially impact system behavior and should therefore be included in any characterization of boundaries.

ASB 018 Design Choice 3: The standard fails to engage with how PGS testing and use is distributed across multiple actors. **ASB 018's requirements promote an understanding that the laboratory is the sole actor conducting an IV study, yet ambiguity in the requirements** – which stems from the standard's failure to engage with the reality of how IV studies may be split between laboratories and developers – **allows for a wide variety of configurations to be considered as compliant**, even when configurations fail to ensure that testing is conducted with technologies and procedures representative of how the software will be used in casework. The standard specifically requires that *"The laboratory shall validate a [PGS] system prior to its use for casework samples in the laboratory."* (R4.1) This language enables compliance by a laboratory that validates the PGS system, but with substantial involvement by the developer, which may importantly undermine the standard's goals.

ASB 018 Design Choice 4: The standard relies on subjective criteria, but does not hold actors accountable to any definition of these criteria. The standard seeks to shape study design by calling for studies to *"address"* (R4.1.3) certain performance characteristics, *"include"* certain types of inputs (R4.1.3), and *"consider"* certain types of input features (R4.1.5). To what extent should actors 'address', 'include', and 'consider'? The standard primarily establishes two criteria for sample types and sizes: that they are *'appropriate'* and *'sufficient'*. These criteria can be operationalized through a number of different study designs. While these criteria provide labs with flexibility to design studies specific to their local setting and intended use cases, the standard's lack of further elaboration on these criteria enables studies to test any types and number of samples (e.g., two samples) without justifying why their choice is sufficient or appropriate. In other words, **the standard does not hold actors accountable to any definition of these subjective criteria**.

4 Discussion

Overall, our work demonstrates a need for key decision-makers to be more guarded in their reliance on ASB 018 by demonstrating a gap between what the standard seeks to guarantee about compliant IV studies and what it actually ensures in practice. From our findings, we derive a set of recommendations for specific changes to ASB 018, which we share in the Appendix. Below, we outline two lessons that efforts on auditing standards in other domains can take away from our study. We additionally discuss lessons and open questions our work contributes to sociotechnical approaches to governance of algorithmic systems.

4.1 Implications for the design of audit standards

Explicitly naming the guarantees a standard seeks to establish can bolster efforts to hold standards developers accountable to the assurances the standard ostensibly promotes. In this study, we demonstrate how assessing practices against guarantees ASB 018 seeks to assure for compliant studies can demonstrate how the standard fails to achieve its goals in practice. However, we had to perform a significant amount of interpretation and draw on text beyond the standard – the accompanying factsheet – to develop our interpretations of the specific guarantees ASB 018 seeks to assure for compliant IV studies. This sort of ambiguity in the standard's communication of its own goals not only undermines the legitimacy of the standard as a tool for accountability, but can also block a key opportunity to scrutinize the standard's efficacy

and hold the standard's developers accountable for the assurances the standard promotes.

Clear definitions of auditing roles in audit standards, developed through engagement with real-world distributions of auditing responsibilities, are crucial to facilitate accountability. In Section 3, we discuss how the standard's language promotes an understanding that "the laboratory" is the sole actor conducting a validation study, yet is ambiguous enough to allow compliance by a validation study conducted collaboratively, by the developer and forensic lab. Examining this ambiguity in light of real-world configurations of IV study responsibilities that we observe in our analysis of validation studies, we find that studies may comply with the standard's requirements while failing to meet the standard's goals (e.g., that software limitations are established through testing under conditions representative of use).

We provide further evidence that the requirements of the standard may not actually support the high-level goals or guarantees the standard seeks to promote. Our study illustrates how the assurance a standard claims to provide can be critically misaligned with requirements used to determine compliance with the standard. This finding echoes prior work finding that companies could comply with FTC privacy assessments while still relying on problematic privacy practices [11], and also echoes recent work on LL144 finding that the structure and content of the law, amongst other factors, failed to support the law's goal to establish an effective third-party auditing regime that protects job-seekers [10, 14].

4.2 Mapping out ASB 018's sociotechnical gap

Our demonstration of the gap between the guarantees ASB 018 seeks to provide about compliant IV studies and what it actually ensures in practice parallels what Ackerman calls the *sociotechnical gap*, which Ackerman defines as the "divide between what we know we must support socially and what we can support technically" [3]. In the context of ASB 018, we conceptualize the sociotechnical gap as the difference between the standard's technical affordances (i.e., its requirements), and the social needs the standard seeks to support (i.e., the guarantees a judge or juror may hope to take away from a validation study's claim of compliance with the standard). Through this workshop and throughout our next steps, we plan to further explore ways the sociotechnical gap can be contextualized for standards such as ASB 018 and other AI governance tools.

Acknowledgments

We thank our anonymous reviewers for their feedback that helped improved this paper. We would also like to thank the following individuals for invaluable discussions and feedback that shaped this work: Jennifer Friedman, Tonya Nguyen, Niloufar Salehi, Richmond Wong. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] [n. d.]. Daubert Standard. *Wex, by the Legal Information Institute at Cornell University*. Accessed January 22, 2025. "https://www.law.cornell.edu/wex/daubert_standard"

[2] 2019. *Internal Validation of STRmix™ V2.4 for Fusion NYC OCME*. Internal Validation Study Summary. NYC Office of the Medical Examiner (OCME). https://www.nyc.gov/assets/ocme/downloads/pdf/STRmix-V2-4-Fusion-5C-Validation%20Summary.pdf

[3] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[4] Nathaniel Adams, Roger Koppl, Dan Krane, William Thompson, and Sandy Zabell. 2018. Appropriate Standards for Verification and Validation of Probabilistic Genotyping Systems. *Journal of Forensic Sciences* 63, 1 (2018).

[5] ANSI/ASB Standard 018, 1st Ed. 2020. *Standard for Validation of Probabilistic Genotyping Systems*. Standard. AAFS Standards Board, Colorado Springs, CO.

[6] Virginia Braun, Victoria Clarke, Nikki Hayfield, Louise Davey, and Elizabeth Jenkinson. 2023. Doing reflexive thematic analysis. In *Supporting research in counselling and psychotherapy: Qualitative, quantitative, and mixed methods research*. Springer, 19–38.

[7] Marc Canellas. 2021. Defending IEEE software standards in federal criminal court. *Computer* 54, 6 (2021), 14–23.

[8] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.

[9] Ellen P Goodman and Julia Trehu. 2022. Algorithmic Auditing: Chasing AI Accountability. *Santa Clara High Tech. LJ* 39 (2022), 289.

[10] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. 2024. Auditing work: Exploring the New York City algorithmic bias audit regime. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1107–1120.

[11] Chris Jay Hoofnagle. 2016. Assessing the Federal Trade Commission's Privacy Assessments. *IEEE Security & Privacy* 14, 2 (2016), 58–64.

[12] Khoa Lam, Benjamin Lange, Borhane Blili-Hamelin, Jovana Davidovic, Shea Brown, and Ali Hasan. 2024. A framework for assurance audits of algorithmic systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1078–1092.

[13] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 557–571.

[14] Lucas Wright, Roxana Mika Muenster, Briana Vecchione, Tianyao Qu, Pika Cai, Alan Smith, Comm 2450 Student Investigators, Jacob Metcalf, J Nathan Matias, et al. 2024. Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1701–1713.

A Recommendations for ASB 018

Based on our findings, we list some initial recommendations for the design of ASB 018 to support better alignment between the goals of the standard and its requirements.

Clearly state guarantees the standard seeks to assure about compliant validation studies. Understanding the guarantees ASB 018 seeks to provide plays a crucial role in our study, since on its own, the standard does not provide a clear picture of the guarantees it seeks to establish for compliant internal validation studies. For example, as we discussed earlier, the standard states that a key goal of internal validation studies is establishing limitations, but does not define what they mean by 'limitation'. Clearly stated goals are critical for evaluating whether the standard is achieving its goals. Echoing our discussion of falsifiability, if the guarantees the standard seeks to promote are not clearly stated, how do we know whether the standard succeeds or fails, in meeting the guarantees it seeks to establish?

Draw on well-established best practices for software verification and validation in software engineering, especially focusing on approaches to clearly specifying software requirements. Earlier, we reveal how studies can evaluate PGS behavior against expectations that make it difficult to establish mutual understandings of what boundaries to establish on software use. We join others [4, 7] in calling for PGS validation to draw from

software engineering standards such as IEEE 1012 address the importance of carefully-defined expectations for software behavior that would help hold validation efforts accountable to clear, falsifiable expectations.

Require clear documentation of who conducted the validation study, and how different actors were involved. During our analysis, we found that crucial details about how validation study responsibilities were split between developers and laboratories were unavailable in publicly available validation study reports. Instead, we found these details in hearing transcripts that the fourth author has access to. However, court transcripts are often not publicly available, typically cost hundreds of dollars per day-long transcript, as transcripts are owned and copyrighted by court reporters, and often span hundreds of pages. In light of these barriers, it is critical that these details are included within validation study reports.

B Figures

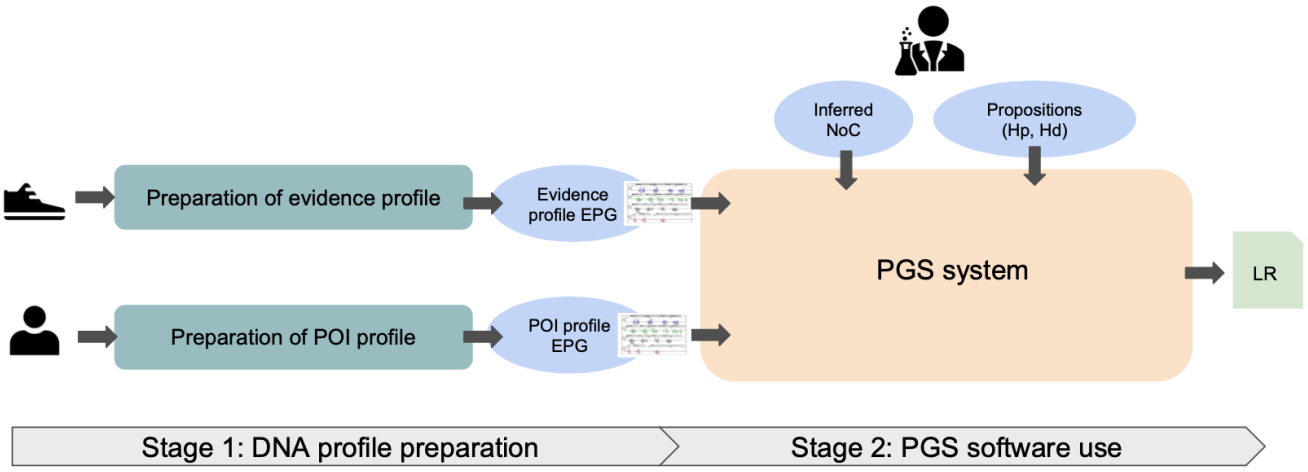

Figure 1: Simplified depiction of forensic DNA testing process using PGS, showing steps and key software inputs and outputs. A lab analyst typically begins their analysis by preparing physical DNA samples – the evidence sample and reference profile from the POI (the POI profile) – for input into the PGS system (Stage 1). Part of this process involves generating electropherograms (EPGs) for each of the profiles. After specifying additional inputs (e.g., the analyst's inference for the number of contributors to the mixture (Inferred NoC), the two hypotheses (Hp, Hd)), the analyst runs the software and produces a likelihood ratio for the POI. The likelihood ratio (LR) is then presented in court as evidence suggesting one of three possibilities: the POI contributed DNA to the evidence sample (inclusionary LR, or LR > 1), the POI did not contribute (exclusionary LR, or LR < 1), or the results are inconclusive (LR = 1). Throughout this pipeline, lab-specific equipment and procedures can all impact the LR output by the PGS system.

Figure 6b: Log(*LR*) versus APH per contributor (0 – 200 rfu x-axis logged scale) for four person mixtures amplified by the OCME laboratory.

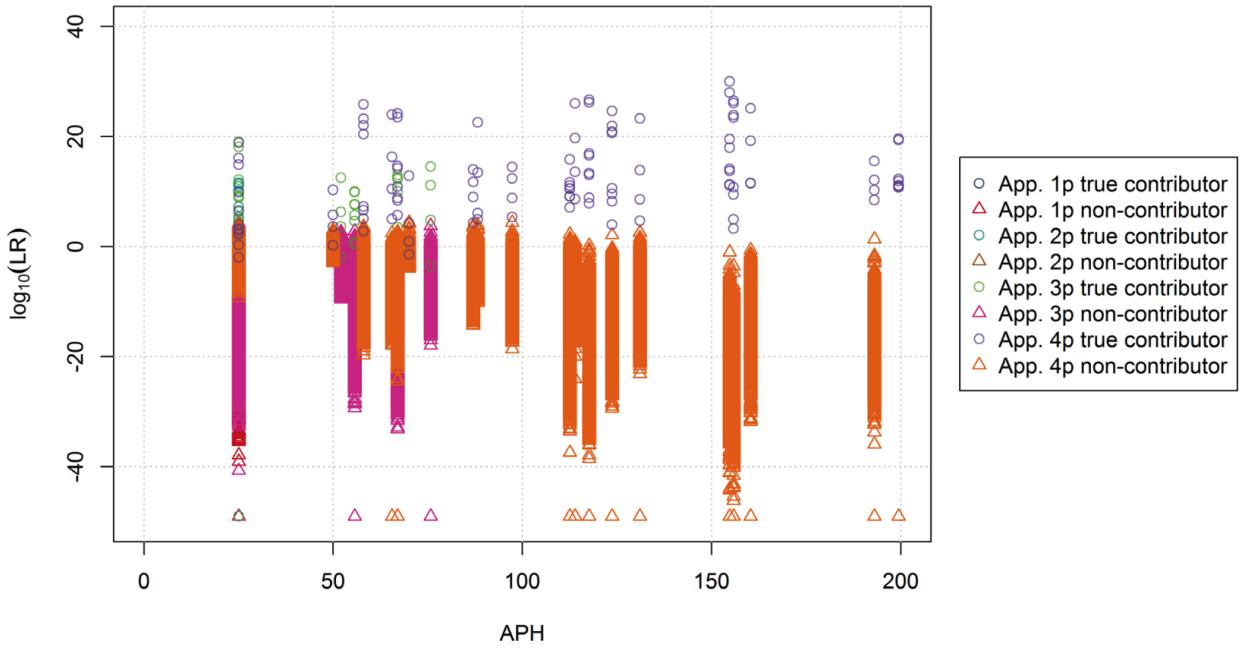

Figure 2: Screenshot of Figure 6b from the OCME report [2] The report claimed that this figure (and other similar plots) demonstrated "the lower limits of DNA" where false positives and false negatives may occur. (OCME, p. 11).