Review: PNAS 2024-01326

This is a great, short piece that addresses a concern that could apply to a variety of forensic examination techniques but uses one method (toolmark comparison) as an example. The authors communicate a simple but important message clearly and mostly without domain-specific jargon. I also love their stance on transparency and documenting all of the "*n*"s associated with any analyses that involve multiple comparisons (numbers of samples and numbers of comparisons).

Most of my suggestions/comments are minor except for the first two, which concern existing alternative solutions to the problem of multiple comparisons and the existing psychological literature that speaks to this kind of comparison task. I think that, with some fairly minor edits that address the concerns I have included below, this piece would be ready for publication.

Suggestions/Comments:

1. **Existing literature with strategies that could help with these types of comparison tasks:** I am not sure if this technique has been applied outside of trace analysis/physical fit analysis, but there is a method that has been empirically tested for improving decision making in physical fit examinations (e.g., determining whether the ends of two pieces of duct tape were once joined together and, thus, from the same source). Regardless, the concept is similar—without the approach in the paper that I will cite here, the analysts are making a holistic, perceptual judgment that involves summarizing multiple judgments made about a long surface. I think that this paper would benefit from discussing this approach, even if it is not perfect, as it is very practical and provides a way of quantifying analyst's individual judgments as well as comparing two analysts' judgments when verifying analyses using numbers rather than an overall, holistic judgment about whether the two samples are from the same source/align. The study is Prusinowski et al. (2023): https://doi.org/10.1016/j.forc.2023.100487

2. **Psychological concepts that shed light on this type of visual comparison process and decision-making context:** Another way to think about what analysts are doing in this context is to consider how people make holistic judgments ("*These two surfaces show mostly the same features."*) versus forcing someone to zoom in and analyze small parts of the materials prior to evaluating the sum of these analyses together as a whole *("80% of the striations line up exactly, 10% are ambiguous such that I am not comfortable stating either way if they are different or the same, and the remaining 10% appear to be inconsistent"*). This is a similar concept to that discussed in the paper cited in point 1, but it is important because this is one way that biases and heuristics can creep into our judgments. When we make holistic assessments in situations where we are technically making multiple comparisons in quick succession, we are likely to discount evidence that does not align with our expectations or other information available and weigh evidence that is consistency with our expectations/other information more heavily. Thus, many

visual comparison tasks in sequence like this is one situation where that biasing effects can have a significant impact on the final conclusion if the decision-maker isn't forced to be deliberative in their analysis of the individual components of the data. I would be good to see some discussion of the cognitive processes that mimic this statistical technique/method/concept given that most of this is still done by people, not technology/actual measurements.

3. **Clarity/precision/minor issues**: Here are some places where the explanations/descriptions could be clearer.
    - Lines 27-28: "This often ignored issue increases the false positive rate, and can contribute to the erosion of public trust in the justice system." Make it clear what issue you are trying to address here, primarily—multiple comparisons, close non-matches, large databases, or all of the above. Also, I think the gap between higher false positive rates and erosion of public trust in the justice system may need to be closed a little more here – it felt like a big leap to go form "false positives" straight to "public trust has been eroded". How? Why?
    - Lines 30 to 34: Discussion of the Mayfield case – The issue is this case was that the database was large AND the fingerprint was a partial latent, so there was less information to use in the comparison to begin with, thus increasing the chance that another fingerprint with a section of friction ridge skins like this might be located in the very, very large number of prints in the database.
    - Line 36: Describe and/or define a cross-correlation function here for those who do not know what it is.
    - Lines 38-40: "…and continues to be used in many pattern searching algorithms to find the best alignment between two images and to quantify their overall similarity." Under what circumstances is this approach used? How often? By who? It currently reads as though this is extremely common, though that may or may not be true – authors should clarify or cite relevant data.
    - Lines 156 – 158: What "*l"* is seems kind of unclear to me – I had to read it a bunch of times and inspect the Figure to figure it out.
    - Lines 226 to 230: "As the number of comparisons increases, the probability of encountering a coincidental match increase." This and the lead up are a little hard to follow – I think this needs to be stated very simply. Something like…"The chance of error compounds/adds up as the number of comparisons performed increases, which means that multiple comparisons increase the chance that a random comparison will yield a result that suggests the samples were caused by the same tool when they are not." Honestly, maybe even simpler that that – I struggle to explain this concept to undergraduate students in research methods, so it is not easy to follow even when you are very motivated to make it so.

- Lines 241 to 242: "…subjective, individual rules" – I am not sure this is categorically true. It might be true in some cases that the individual analysts is entirely responsible for determining the criteria that samples must meet to say "the same tool cut both of these wires", but I don't think that examiners are always testifying to "subjective individual rules" necessarily. Happy to be proven wrong though!
- Lines 266-267: How was the data pooled? Weighted by sample size etc?