Calibration of STRmix LRs following the method of Hannig *et al*.

John Buckleton[1,2], Maarten Kruijver[2], James Curran[1], and Jo-Anne Bright[2]

1. Department of Statistics, University of Auckland, New Zealand
2. Institute of Environmental Science and Research Limited, Auckland, New Zealand

Introduction

Calibration may be used to assess whether methods of LR assignment are reliable. Ramos and Gonzalez-Rodriguez [1] introduce the concept of calibration using weather forecasting as an example. Weather forecasters often give a probability of rain. Let us imagine that we wish to check whether these probabilities are being assigned sensibly. If we can assemble a number of days for which the prediction is, for example, around 50%, and of those days about half have precipitation, then this is evidence that this forecaster is operating sensibly - at least in this part of the probability range. This approach for assessing LR calibration is based on assessing the calibration of posterior probabilities for ground-truth known examples with varying prior probabilities. This is based on the LR being the multiplier that converts a prior probability into a posterior probability.

Two large calibration studies based on the work Ramos and Gonzalez-Rodriguez [1] of (N = 28,250,000 and 700,000,000) have been undertaken for STRmix™ [2, 3]. In the first, LRs for 2825 profiles generated using different multiplex kits from thirty-one (31) laboratories were analysed [4]. LRs were calculated using the Caucasian allele frequencies from the FBI expanded CODIS core set [5] and a theta (θ) of 0.01. In the second, 70 profiles generated using the Investigator® 24plex QS Kit (QIAGEN) dataset were analysed. Comparisons were undertaken using the allele probabilities from the FBI extended Caucasian allele frequencies and $\theta = 0$. The results of these two studies showed good calibration.

More recently Hannig et al. [6] have published a variant method for calibration using integration.

The Hannig et al. [6] method was applied without fiducial distribution fitting to the two datasets that have previously been described [2, 3]. For comparison, the previously published datasets were reanalysed using the method of Ramos and Gonzalez-Rodriguez [1] but recalculated to align with the Hannig et al. intervals.

Below we correct the typographical error in Hannig et al. [6] equation 1 applied in this research:

$$\log_{10}\left(G_b - G_a\right) - \log_{10}\left(bF_b - aF_a - \int_a^b F_r dr\right) = \delta$$

Where:

- G_x is the cumulative density function for the H_p true donors at $LR = x$
- F_x is the cumulative density function for the H_a true donors at $LR = x$
- a and b are two LR values

Hannig et al. make the claim: “Our approach is based on a direct assessment of how well the property that LR of LR is LR is satisfied and does not require consideration of prior or posterior probabilities.” Relatively simple algebra recovers the prior odds easily from their formula.

Hannig et al. describe the interpretation of δ "… the calibration plot suggests the following: Software A has a downward slope suggesting that as the reported LR values get increasingly larger than 1 they tend to increasingly overstate the weight of evidence in favor of H_P. In the case of Software B, the calibration plot suggests that it may be overstating the weight of evidence in favor of H_P by little less than a factor of 10."

Again this interpretation may not be sustainable. Our reasoning would include:

1. The values compared, G_x and F_x, are the sample values not the parameters,
2. The reported *LR* usually has deliberate conservativism included whereas often the tested *LR* does not (the 31 lab compilation has some but not all of the typical conservativism).

However, the two datasets examined above tend to show $-1 \le \delta \le 1$.

The results of the additional analyses are given in Tables 1 and 2 for each dataset.

Table 1. The results of the Hannig et al. (left) and Ramos and Gonzalez-Rodriguez (right) workup of 211 H_p true and 700,000,000 H_a true comparisons described in [2]. The n/a occurs because $G_b - G_a = 0$ or $bF_b - aF_a - \int_a^b F_r dr = 0$

$\log_{10}LR$		Count between a and b		δ	Posterior probability for $\log_{10}LR$		Observed probability of true donors	Comment regarding true donors
a	b	H_p true	H_a true		a	b		
0	1	2	8,476,455	-0.847	3×10^{-7}	3×10^{-6}	2.4×10^{-7}	Fewer
1	2	2	109,809	0.041	3×10^{-6}	3×10^{-5}	1.8×10^{-5}	Correct
2	3	0	20,055	n/a	3×10^{-5}	3×10^{-4}	0	Fewer
3	4	1	2,505	-0.618	3×10^{-4}	3×10^{-3}	4.0×10^{-4}	Correct
4	5	1	233	-0.587	3×10^{-3}	0.03	4.3×10^{-3}	Correct
5	6	11	30	0.345	0.03	0.23	0.27	More
6	7	9	6	-0.043	0.23	0.75	0.60	Correct
7	8	10	1	-0.220	0.75	0.97	0.91	Correct
8	9	11	0		0.97		1.00	

Table 2. The results of the Hannig et al. (left) and Ramos and Gonzalez-Rodriguez (right) workup of 10,101 H_p true and 28,250,000 H_a true comparisons described in Bright et al. [4]

$\log_{10}LR$		Count $\leq LR_a$		δ	Posterior probability for $\log_{10}LR$		Observed probability of true donors	Comment regarding true donors
a	b	H_p true	H_a true		a	b		
0	1	286	163,847	0.493	0.0004	0.004	0.0017	Correct
1	2	338	15,256	-0.052	0.004	0.035	0.022	Correct
2	3	366	1,440	0.052	0.035	0.263	0.203	Correct
3	4	380	121	0.111	0.263	0.781	0.758	Correct
4	5	445	16	0.203	0.781	0.973	0.965	Correct
5	6	509	4	0.151	0.973	0.997	0.992	Correct
6	7	536	0	n/a	0.997	0.99972	1.000	More
7	8	596	0	n/a	0.99972	0.99997	1.00000	More
8	9	606	0	n/a	0.99997			

Conclusion

We examined the rate of false inclusionary support using the concept of calibration. The thirty-one laboratory dataset [4] used $\theta = 0.01$. The Investigator® 24plex QS Kit (QIAGEN) dataset used $\theta = 0$. These would be expected to have a mild bias in favour of conservativism and a mild bias in favour of non-conservativism, respectively.

The method of Hannig et al. shows $-1 \le \delta \le 1$ for both datasets (Table 1 and 2). This means that the observed rate of false inclusionary support was within one order of magnitude of the expected rate.

The Ramos and Gonzalez-Rodriguez calibration also showed good alignment of observed and expected rates of false inclusionary support. Of the comparisons available 11 were scored "Correct" meaning that the observed rate was within the bounds of the expected rate, three were scored "More" meaning that the observed rate of true donors was more than expected, and two were scored "Fewer" meaning that observed rate of true donors was less than expected. but

Acknowledgements

We acknowledge the valuable assistance of Klaas Slooten and Amke Caliebe with correcting equation 1 in Hannig et al. [6]. This work was supported in part by grant 2017-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of their organizations or of the U.S. Department of Justice.

References

[1] Ramos D, Gonzalez-Rodriguez J. Reliable support: Measuring calibration of likelihood ratios. Forensic Science International. 2013;230:156-69.

[2] Bright J-A, Jones Dukes M, Pugh SN, Evett IW, Buckleton JS. Applying calibration to LRs produced by a DNA interpretation software. Australian Journal of Forensic Sciences. 2019:1-7.

[3] Buckleton JS, Bright J-A, Ciecko A, Kruijver M, Mallinder B, Magee A, et al. Response to: Commentary on: Bright et al. (2018) Internal validation of STRmix™ – A multi laboratory response to PCAST, Forensic Science International: Genetics, 34: 11–24. Forensic Science International: Genetics. 2020;44:102198.

[4] Bright J-A, Richards R, Kruijver M, Kelly H, McGovern C, Magee A, et al. Internal validation of STRmix™ – A multi laboratory response to PCAST. Forensic Science International: Genetics. 2018;34:11-24.

[5] Moretti TR, Moreno LI, Smerick JB, Pignone ML, Hizon R, Buckleton JS, et al. Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. Forensic Science International: Genetics.25:175-81.

[6] Hannig J, Riman S, Iyer H, Vallone PM. Are reported likelihood ratios well calibrated? Forensic Science International: Genetics Supplement Series. 2019;7:572-4.