Forensic Science International: Genetics 7 (2013) 52–54

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Decision-making in familial database searching: KI alone or not alone?

David J. Balding [a], Michael Krawczak [b], John S. Buckleton [c], James M. Curran [d,*]

[a] Institute of Genetics, University College London, Gower Street, London WC1E 6BT, UK
[b] Institute of Medical Informatics and Statistics, Christian-Albrechts University, Brunswiker Strasse 10, 24105 Kiel, Germany
[c] ESR Mt Albert Science Centre, Private Bag 92-021, Auckland, New Zealand
[d] Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 14 March 2012
Received in revised form 21 May 2012
Accepted 2 June 2012

Keywords:
Kinship
Familial searching
Likelihood ratio
Identity-by-descent
Neyman–Pearson Lemma
Simulation

ABSTRACT

We consider the comparison of hypotheses "parent–child" or "full siblings" against the alternative of "unrelated" for pairs of individuals for whom DNA profiles are available. This is a situation that occurs repeatedly in familial database searching. A decision rule that uses both the kinship index (*KI*), also known as the likelihood ratio, and the identity-by-state statistic (*IBS*) was advocated in a recent report as superior to the use of *KI* alone. Such proposal appears to conflict with the Neyman–Pearson Lemma of statistics, which states that the likelihood ratio alone provides the most powerful criterion for distinguishing between any two simple hypotheses. We therefore performed a simulation study that was two orders of magnitude larger than in the previous report, and our results corroborate the theoretical expectation that *KI* alone provides a better decision rule than *KI* combined with *IBS*.

1. Introduction

In a recent comparison of statistical methods for familial database searches [1], Ge et al. advocated decision-making based upon the combined use of *IBS* and *KI* as superior to the use of *KI* alone. Here, *IBS* ('identity-by-state') denotes the number of alleles shared by two given individuals, and *KI* ('kinship index') is the likelihood ratio in favor of a specified relationship over the alternative that they are unrelated. In familial database searching, the relationships of interest are "parent–child" and "full siblings", and we will write KI_{PC} and KI_{FS} respectively for the corresponding *KI*s. All humans are related to one another, though relationships may be remote and unknown. "Unrelated" usually refers to the simple, but artificial, model of independent sampling of alleles from a certain gene pool. A more realistic model would use population genetic parameter F_{ST} to account for the co-ancestry of "unrelated" individuals. We advocate using an F_{ST} adjustment in real applications, but for the purposes of this short note on the relative merits of different decision rules, we follow [1] and assume $F_{ST} = 0$.

The assertion that the pair of statistics (*IBS*,*KI*) gives a better decision rule than *KI* alone appears implausible because it conflicts with the Neyman–Pearson Lemma [2], which states that the likelihood ratio is the most powerful statistic for distinguishing between two simple hypotheses. Intuitively, *KI* includes all the information provided by the genotype data for distinguishing a specified relationship from unrelated, and its efficiency cannot be enhanced by the concurrent consideration of any other statistic. The fact that (*IBS*,*KI*) is a pair of numbers does not affect the logic of the Neyman–Pearson Lemma: in statistics, a 'statistic' is any function of the data, be it univariate, bivariate or multivariate.

There are superficial attractions to using *IBS* rather than *KI*. The former may be easier for non-experts to understand, and does not require knowledge of allele frequencies to compute. Moreover, use of *KI* entails deciding in advance the alternative hypothesis of interest (*e.g.* "parent–child" or "full siblings"), which *IBS* does not, and *KI* may fail to reject the null hypothesis of "unrelated" when the individuals are in fact related but the specified alternative hypothesis is wrong. These perceived advantages of *IBS* over *KI* are illusory however. Firstly, although allele frequencies are not required to compute *IBS*, they must be invoked to evaluate *IBS*. Secondly, to choose an appropriate *IBS* threshold, a power or similar analysis is required. This in turn requires specifying the alternative relationship as well. In any case, combining *IBS* with *KI* would lose any such seeming advantage of using *IBS* alone.

Ge et al. [1] provided simulation results that they interpreted as proving the superiority of (*IBS*,*KI*) over *KI* alone. For example, for "parent–child" *vs.* "unrelated", their false positive rate (*FPR*) of 0.0014 for $KI > 10{,}000$ drops to 0.0010 and 0.0007 when adding the requirement that $IBS > 15$ and $IBS > 16$, respectively. At the same

* Corresponding author.
E-mail addresses: d.balding@ucl.ac.uk (D.J. Balding), krawczak@medinfo.uni-kiel.de (M. Krawczak), John.Buckleton@esr.cri.nz (J.S. Buckleton), curran@stat.auckland.ac.nz (J.M. Curran).

1872-4973/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2012.06.001

time, the false negative rates (*FNR*) increases from 0.494 to 0.558 (incorrectly reported as 0.218) and 0.659. In their comments, the authors expressed the view that the gain in *FPR* would be worth the consequent loss in *FNR*. However, trading off *FPR* against *FNR* requires subjective judgments as to the relative harm of each type of error, and the conventional approach is to avoid these judgments by comparing the rates of one type of error when the other error rate is fixed. Here, we made these comparisons in the context of a much larger simulation study than was undertaken by Ge et al. [1].

2. Methods

One hundred million (10^8) each of full sibling and parent–child pairs, and one billion (10^9) unrelated pairs, were simulated according to Mendelian principles, using allele probabilities obtained from Caucasian population data for the 13 CODIS Short Tandem Repeat loci [3]. This is greatly in excess of the one million (10^6) simulated pairs previously employed [1], because estimating the difference in power between two methods requires larger sample sizes than estimating the power of a single method.

Following [1], we computed KI_{PC} for the parent–child and unrelated pairs, KI_{FS} for the full-sibling and unrelated pairs, and *IBS* for all pairs. To avoid specifying the relative costs/benefits of the two types of errors, we compared the false negative rate (*FNR*) when the false positive rate (*FPR*) was equalized for the two decision rules, and *vice versa*. For example, we compared the *FNR* of the decision rule that declares a pair to be full siblings when both $IBS \geq 15$ and $KI_{FS} \geq 1000$, with the *FNR* of declaring "full sibling" when $KI_{FS} \geq x$, where x is chosen so that both decision rules have the same *FPR*. Similarly, we compared the *FPR* of the two rules when x was chosen to equalize the two *FNR*s.

Additional computational effort was applied to the selection of x when $IBS \geq 16$ and $KI \geq 100{,}000$. Ge et al. report the false positive rates as 1 in a million for parent–child and 3 in a million for siblings. However, the sampling variation associated with the simulation of such small proportions in only one million pairs is going to be high. Therefore, we decided to invest more computational effort in firstly estimating the false positive rates more accurately, and secondly in determining appropriate thresholds for declaring kinship. The numbers of false positives in one billion unrelated pairs were 493 classified as parent–child and 478 classified as full siblings. To estimate the value of x giving equivalent *FPR*s by naively storing all the observed values would impose excessive memory requirements. To overcome this hurdle, we employed the method of Woodruff [4] which we explain by means of the following example:

Assume that we seek to accurately estimate the 90th percentile of a certain distribution. A straightforward way to do this would be to draw a sample of size one billion from this distribution, and use the 900,000,000th largest value as the sought-after estimate. However, one billion double precision numbers require about 7.5 GB of RAM to store, and disk-based sorting of the sample values, while using less RAM, would be extremely slow. The alternative [4] is to take a smaller sample from the distribution first and to use it to calculate a confidence interval for the desired percentile in the full sample. Subsequently, only values from the full sample that fall within the confidence limits need to be stored. For example, with an initial sample of size $n = 100$ drawn from a standard normal distribution, a 99.7% (± 3 standard deviations) confidence interval for the 90th percentile of the full sample would be obtained by first calculating a 99.7% confidence interval for binomial proportion $p = 0.9$ from the small sample. Using the normal approximation to the binomial, the required interval is

$$p \pm 3 \times \sqrt{\frac{p(1-p)}{n}}$$

which gives a confidence interval for the binomial proportion of [0.81, 0.99] if $p = 0.9$ and $n = 100$. In our example, the 0.81 and 0.99 sample quantiles of the small sample were 0.891 and 2.354, and these demarcated the confidence interval for the 90th percentile in the full sample. Now, when considering the full sample, only values between 0.891 and 2.354 had to be stored whereas values below this range were counted, but not stored. As a smaller illustration, we took a sample of size $n = 100{,}000$ (rather than one billion) and found 81,565 values below 0.891, and 17,492 values between 0.891 and 2.354. The sought-after 90,000th largest value of the sample was the 8435th ($= 90{,}000 - 81{,}565$) largest value among the stored values, which equaled 1.272. This compares favorably to

Table 1
Comparison of the percentage false positive rates (*FPR*) and false negative rates (*FNR*) of two decision rules for familial database searching. Columns 4 and 7 contain the *FPR* and *FNR* for the bivariate decision rule that declares a pair of individuals to be related if both $IBS > ibs_0$ and $KI > ki_0$, where the values of ibs_0 and ki_0 are specified in columns 1 and 2. Columns 3 and 6 give the values of ki_N and ki_P such that the univariate decision rule based upon $KI > ki_N$ has the same *FNR*, and that based upon $KI > ki_P$ has the same *FPR*, as the bivariate rule. Columns 5 and 8 give the *FPR* and *FNR* for the univariate decision rules using $KI > ki_N$ and $KI > ki_P$, respectively. The purpose of the table is to allow comparison of columns 4 and 5, and of columns 7 and 8, and since these columns contain error rates, smaller values indicate the better decision rule.

ibs_0	ki_0	ki_N	*FPR* (ibs_0,ki_0)	*FPR* (ki_N)	ki_P	*FNR* (ibs_0,ki_0)	*FPR* (ki_P)
Parent–child *vs.* unrelated							
14	100	291	4.77×10^{-4}	3.23×10^{-4}	126	0.048	0.018
14	1000	1228	1.36×10^{-4}	1.20×10^{-4}	1054	0.171	0.153
14	10,000	10,608	1.41×10^{-5}	1.35×10^{-5}	10,153	0.510	0.503
15	1000	2595	1.12×10^{-4}	6.22×10^{-5}	1319	0.275	0.180
15	10,000	14,109	1.25×10^{-5}	9.76×10^{-6}	10,994	0.558	0.516
16	1000	8610	7.22×10^{-5}	1.77×10^{-5}	2213	0.475	0.251
16	10,000	27,088	9.23×10^{-6}	4.18×10^{-6}	14,335	0.661	0.561
16	100,000	174,786	4.78×10^{-7}	2.75×10^{-7}	128,235	0.871	0.845
Full siblings *vs.* unrelated							
14	100	108	7.89×10^{-4}	7.63×10^{-4}	103	0.238	0.235
14	1000	1009	8.93×10^{-5}	8.79×10^{-5}	1004	0.426	0.426
14	10,000	10,011	7.56×10^{-6}	7.46×10^{-6}	9962	0.634	0.633
15	1000	1072	8.69×10^{-5}	8.25×10^{-5}	1034	0.432	0.428
15	10,000	10,113	7.52×10^{-6}	7.38×10^{-6}	10,089	0.634	0.634
16	1000	1410	7.72×10^{-5}	6.22×10^{-5}	1157	0.457	0.439
16	10,000	10,774	7.28×10^{-6}	6.80×10^{-6}	10,370	0.640	0.637
16	100,000	101,560	4.93×10^{-7}	4.73×10^{-7}	99,330	0.806	0.804
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)

the "true" value of 1.282, and we only had to store (and sort) fewer than one fifth of the simulated values.

3. Results

The results of our simulations are summarized in Table 1. Columns 4 and 7 of Table 1 correspond to Table 6(c) of Ge et al. [1]. The differences between the two tables reflect a higher precision resulting from the larger number of simulations performed in our study, and from the correction of one gross error in the earlier report, mentioned above. Although the values in the two tables are generally similar, they frequently differ by >10% and occasionally by >50%.

In no scenario considered was the univariate decision rule based upon *KI* alone inferior to the bivariate rule based upon (*IBS*,*KI*). Specifically, no entry of columns 4 and 7 of Table 1 is smaller than the corresponding entry of columns 5 or 8, respectively. In many settings the two estimated error rates are similar, but their ratio can range up to four in the scenarios considered.

In the final row of Table 1, the *FPR* estimates correspond respectively to 493 and 473 observed false positives in one billion trials. The standard deviations of these counts were both approximately equal to 22, and so their difference was not statistically significant. However, 10 of the 16 *FPR* comparisons and all of 16 *FNR* comparisons were significant at $\alpha = 0.05$ in favor of the univariate decision rule, and none yielded even nominal evidence against it.

The R package used to generate the data in Table 1 has been made available in the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/relSim/index.html).

4. Conclusion

In a simulation-based assessment of two decision rules for familial database searching, namely (*IBS*,*KI*) and *KI* alone, we found highly significant support for the latter. This result was to be expected because it is implied by the Neyman–Pearson Lemma [2]. The differences in error rates between the two approaches were small in many comparisons, but moderately large in others. Even if the gain in power of *KI* alone is small, because use of the compound decision rule adds complexity to the decision process and conveys no advantage, we recommend that *KI* alone is used to test any two competing hypotheses for the relationship between a pair of individuals, as occurs in familial database searching.

Acknowledgements

We gratefully acknowledge the comments of Lisa Melia, Michael Taylor, and two anonymous referees which greatly improved this paper. We would also like to acknowledge the assistance of Torben Tvedebrink whose help was invaluable in the validation and verification of the software. This work was supported in part by 136 grant 2011-DN-BX-K541 from the US National Institute of Justice.

References

[1] J. Ge, R. Chakraborty, A. Eisenberg, B. Budowle, Comparison of familial DNA database searching strategies, J. Forensic Sci. 56 (2011) 1448–1456.
[2] J. Neyman, E. Pearson, On the problem of the most efficient tests of statistical hypotheses, Philos. Trans. R. Soc. Lond. A 231 (1933) 289–337.
[3] B. Budowle, B. Shea, S.J. Niezgoda, R. Chakraborty, CODIS STR loci data from 41 sample populations, J. Forensic Sci. 46 (2001) 453–489.
[4] R.S. Woodruff, Confidence intervals for medians and other position measures, JASA 57 (1952) 622–627.