What do measures of agreement (κ) tell us about quality of exposure assessment? Theoretical analysis and numerical simulation

Igor Burstyn, 1 Frank de Vocht, 2 Paul Gustafson 3

ABSTRACT

Background: The reliability of binary exposure classification methods is routinely reported in occupational health literature because it is viewed as an important component of evaluating the trustworthiness of the exposure assessment by experts. The Kappa statistics (κ) are typically employed to assess how well raters or classification systems agree in a variety of contexts, such as identifying exposed participants in a population-based epidemiological study of risks due to occupational exposures. However, the question we are really interested in is not so much the reliability of an exposure assessment method, although this holds value in itself, but the validity of the exposure estimates. The validity of binary classifiers can be expressed as a method’s sensitivity (SN) and specificity (SP), estimated from its agreement with the error-free classifier.

Methods and results: We describe a simulation-based method for deriving information on SN and SP that can be derived from κ and the prevalence of exposure, since an analytic solution is not possible without restrictive assumptions. This work is illustrated in the context of comparison of job-exposure matrices assessing occupational exposures to polycyclic aromatic hydrocarbons.

Discussion: Our approach allows the investigators to evaluate how good their exposure-assessment methods truly are, not just how well they agree with each other, and should lead to incorporation of information of validity of expert assessment methods into formal uncertainty analyses in epidemiology.

INTRODUCTION

The reliability of binary exposure classification methods is routinely reported in occupational health literature because it is viewed as an important component of evaluating the trustworthiness of the exposure assessment. The Kappa statistics (κ) are typically employed to assess how well the raters or classification systems agree in a variety of contexts, such as identifying exposed participants in a population-based epidemiological study of risks due to occupational exposures. Most recently, Offermans et al. 2 estimated agreement among various methods of assessing exposures in a cohort using various expert-based methods (job-exposure matrices and case-by-case evaluations). The authors reported κ coefficients for these methods that are not unlike those presented previously in a review by Teschke et al, 2 and that seems to suggest that κ values of about 0.6 or worse are a fair summary of what these methods generally yield in terms of inter-rater agreement in a typical study of occupational exposures. However, the question we are really interested in is not so much the reliability of a method to assess exposure, although this holds value in itself, but the validity of the exposure estimates.

The validity of binary classifiers can be expressed as a method’s sensitivity (SN) and specificity (SP), estimated from its agreement with the error-free classifier (also known as ‘gold standard’). 3 But how does one infer what κ tells us about the validity of exposure estimates (ie, SN and SP) when a true value (gold standard) is unavailable? Generally, reliability contains information on validity, 3 but in the case of κ, its relationship with SN and SP is also affected by prevalence of
exposure (Pr). An analytic solution in this case is not possible without restrictive assumptions about the actual prevalence and relationship between SN and SP. Therefore, we developed a simulation-based method for deriving information on SN and SP based on Pr. We illustrate this method in the context of a comparison of job-exposure matrices assessing occupational exposures to polycyclic aromatic hydrocarbons (PAHs). 1

METHOD

We propose a simulation-based method to calculate the values of SN and SP that are consistent with the observed κ and Pr. The relationship among κ, SN, SP and Pr can be described mathematically, if we assume two conditionally independent raters with the same validity, by:

\[
\kappa = \frac{Pr \times (SP - 1 + SN)^2}{(Pr - 1)/((Pr \times SN - SP - Pr + Pr \times SP))}
\]

(1)

\[
SN + SP > 1
\]

(2)

First, we define the distributions of the lower (κl) and upper (κu) bounds of κ by using uniform distributions (U) as κl ~ U(a1, a2) and κu ~ U(b1, b2). We further define the distribution of Pr as a Beta distribution—Pr ~ Beta(c, d). Information required to specify these distributions with reasonable credibility is available in reports evaluating inter-rater agreements, as in reference. 1 We can then calculate (multiple) the lower bounds (SNl and SP) for PAH by using U as κl ~ U(0.29, 0.31) and κu ~ U(0.59, 0.61). Some degree of judgements is involved in this but our formulation reflects the observation that in this case κ for PAHs lies between 0.3 and 0.6. We further define the distribution of Pr (mode of 5%, with 95% certainty that Pr does not exceed 10%) as Pr ~ Beta(6.2, 99.7). 5 The results of the rest of the calculations are summarised in figure 1, derived from 10 000 Monte Carlo samples for candidate values of SN and SP (step (b) above). They reveal that the mean SN for this example is about 0.78 (SD 0.15) and mean SP is about 0.96 (SD 0.09).

RESULTS

We apply our method to information provided in table 2 in the article by Offermans et al 7 for PAH exposure assessment. First, we define the distributions of the κl and κu for PAH by using U as κl ~ U(0.29, 0.31) and κu ~ U(0.59, 0.61). Some degree of judgements is involved in this but our formulation reflects the observation that in this case κ for PAHs lies between 0.3 and 0.6. We further define the distribution of Pr (mode of 5%, with 95% certainty that Pr does not exceed 10%) as Pr ~ Beta(6.2, 99.7). 3 The results of the rest of the calculations are summarised in figure 1, derived from 10 000 Monte Carlo samples for candidate values of SN and SP (step (b) above). They reveal that the mean SN for this example is about 0.78 (SD 0.15) and mean SP is about 0.96 (SD 0.09).

DISCUSSION

Our approach allows the investigators to evaluate how good their exposure-assessment methods truly are, not just how well they agree with each other, and should lead to incorporation of information of validity of expert assessment methods into formal uncertainty analyses in epidemiology. 6 Specifically, once we can represent knowledge about SN and SP by a joint distribution, we can use a number of existing techniques to evaluate the impact of exposure misclassification on the epidemiological results and to correct such results for known imperfections in exposure classification. Till now, knowledge of κ and exposure prevalence did not enable such analyses. It is noteworthy that Bayesian analyses that
appraised SN and SP of another job-exposure matrix produced a very similar appraisal for SP and lower value for average SN with a similarly wide distribution. This perhaps points to commonality of quality of expert assessment methods used in occupational epidemiology. It is important to note that simple comparison of measures of agreement across studies and instruments is not helpful because values of κ depend on the Pr, which may differ between applications even for the same SN and SP. Our method has a distinct advantage for such comparisons and assessment of validity. With knowledge about validity, even if it is uncertain, we can begin the work on incorporating this knowledge into routine epidemiological analyses.\(^7\)

**Contributors** All authors equally contributed to the writing of the manuscript. IB and PG jointly developed the algorithm. Theoretical derivations were performed by PG. Simulations were conducted by IB and verified by PG and FdV.

**Funding** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** An appendix is available online.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/

**REFERENCES**

1. Offermans NS, Vermeulen R, Burdorf A, et al. Comparison of expert and job-exposure matrix-based retrospective exposure assessment of occupational carcinogens in The Netherlands Cohort Study. *Occup Environ Med* 2012;69:745–51.
2. Teschke K, Olshan AF, Daniels JL, et al. Occupational exposure assessment in case-control studies: opportunities for improvement. *Occup Environ Med* 2002;59:575–93.
3. White E, Armstrong BK, Saracci R. *Principles of exposure measurement in epidemiology: collecting, evaluating and improving measures of disease risk factor*. Oxford University Press, 2008.
4. Feuerman M, Miller AR. Relationships between statistical measures of agreement: sensitivity, specificity and kappa. *J Eval Clin Pract* 2008;14:930–3.
5. Chun-Lung SU. Bayesian Epidemiologic Screening Techniques. 2013. http://www.epi.ucdavis.edu/diagnostictests/betabuster.html (accessed 21 Nov 2013).
6. MacLehose RF, Gustafson P. Is probabilistic bias analysis approximately Bayesian? *Epidemiology* 2012;23:151–8.
7. Liu J, Gustafson P, Cherry N, et al. Bayesian analysis of a matched case-control study with expert prior information on both the misclassification of exposure and the exposure-disease association. *Stat Med* 2009;28:3411–23.
8. Beach J, Burstyn I, Cherry N. Estimating the extent and distribution of new-onset adult asthma in British Columbia using frequentist and Bayesian approaches. *Ann Occup Hyg* 2012;56:719–27.
9. Gustafson P. *Measurement error and misclassification in statistics and epidemiology*. Chapman & Hall/CRC Press, 2004.

**Figure 1** Plausible pairs of sensitivity (SN) and specificity (SP) values for exposure-assessment methods for polycyclic aromatic hydrocarbons evaluated in ref. 1; hashed lines denote means.