BMJ Open

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]

To cite: Burstyn I, de Vocht F, Gustafson P. What do measures of agreement (κ) tell us about quality of exposure assessment? Theoretical analysis and numerical simulation. *BMJ Open* 2013;**3**:e003952. doi:10.1136/bmjopen-2013-003952

► Prepublication history and additional material for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2013-003952).

Received 5 September 2013
Revised 25 October 2013
Accepted 1 November 2013

[1]Department of Environmental and Occupational Health, School of Public Health, Drexel University, Philadelphia, Pennsylvania, USA
[2]Centre for Occupational and Environmental Health, Centre for Epidemiology, Institute of Population Health, Manchester Academic Health Sciences Centre, The University of Manchester, Manchester, UK
[3]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence to
Dr Igor Burstyn;
igor.burstyn@drexel.edu

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.

Strengths and limitations of this study

- The main strength of our approach is that it is flexible and easy to implement.
- Our methodology accounts for realistic uncertainties that an epidemiologist faces in evaluating the plausible extent of exposure misclassification.
- The main limitation of our work is that it does not yet account for correlated errors in exposure estimates that are common in the field, and the importance of this limitation remains to be understood.

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*[1] 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.[4] Therefore, we developed a simulation-based method for deriving information on SN and SP based on κ and the 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:

$$\begin{aligned}\kappa =&(\Pr\times(\mathrm{SP}-1+\mathrm{SN})^2)\\ &\times(\Pr-1)/((\Pr\times\mathrm{SN}-\mathrm{SP}-\Pr+\Pr\times\mathrm{SP})\\ &\times(\Pr\times\mathrm{SN}+1-\mathrm{SP}-\Pr+\Pr\times\mathrm{SP}))\end{aligned}\quad(1)$$

We assume that exposure classification by experts is better than chance, as expressed by:

$$\mathrm{SN}+\mathrm{SP}>1\quad(2)$$

First, we define the distributions of the lower (κ_l) and upper (κ_h) bounds of κ by using uniform distributions (U) as $\kappa_l \sim U(a_1, a_2)$ and $\kappa_h \sim U(b_1, b_2)$. 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 of SN and SP (SN_l and SP_l) that are consistent with these distributions, following:

$$\mathrm{SN}_l=\kappa_l/((1-\Pr)+\kappa_l\times\Pr),\quad(3)$$

and

$$\mathrm{SP}_l=\kappa_l/(\Pr+\kappa_l\times(1-\Pr))\quad(4)$$

The upper theoretical bounds on SN and SP are known (ie, these are 1) and, even though no other information is available, this enables us to sample plausible SN and SP values from the uniform distribution constrained by the lower bounds (SN_l and SP_l, respectively) and the upper bounds of 1. Using Monte Carlo sampling, this procedure is repeated multiple times to generate sets of possible (SN and SP) combinations.

The proposed procedure is a hierarchical process that starts with (a) selecting a set of (κ_l, Pr) values from specified distributions to calculate (SN_l, SP_l; Eqs. (3) and (4)), and is followed by (b) selecting candidate set (SN and SP) from values uniformly distributed between the lower bounds (SN_l and SP_l) and the upper theoretical maximum of 1, and completed by (c) imposing constraints on the candidate set of (SN and SP) that are implied by Eqs. (1) and (2) (see next paragraph for details of the last step). The purpose of step (a) in the procedure is to calculate SN_l and SP_l. The purpose of step (b) is to sample candidate values of SN and SP that lie between their respective theoretical lower and upper boundaries. The purpose of step (c) is to limit the sets of values of SN and SP selected in step (b) to only those that, first, are congruent with the theoretical model that relates validity to reliability (Eq. 1), and, second, satisfy the assumption that classification of exposure is better than random (Eq. 2).

By chance, some values of Pr, SN and SP selected in this way will correspond to values of κ, implied by Eq. (1), that lie outside of bounds on κ that we have specified by choosing specific values of κ_l and κ_h from corresponding distributions. Furthermore, some combinations of SN and SP will not be consistent with Eq. (2) (ie, imply that exposure classification was worse than chance). Consequently, the candidate sets of values of SN and SP that are not in agreement with our starting assumptions are eliminated from the sample used to estimate the distributions of SN and SP. The resulting combinations are consistent with our knowledge of agreement between different exposure assessment methods and foretell how valid these exposure assessment methods can be expected to be in general.

Calculation can be implemented in R, and is available in Appendix 1 (available online) with input values specific to the illustrative example described below.

RESULTS

We apply our method to information provided in table 2 in the article by Offermans *et al*[1] for PAH exposure assessment. First, we define the distributions of the κ_l and κ_h for PAH by using U as $\kappa_l \sim U(0.29, 0.31)$ and $\kappa_h \sim U(0.59, 0.61)$. 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 0.03).

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

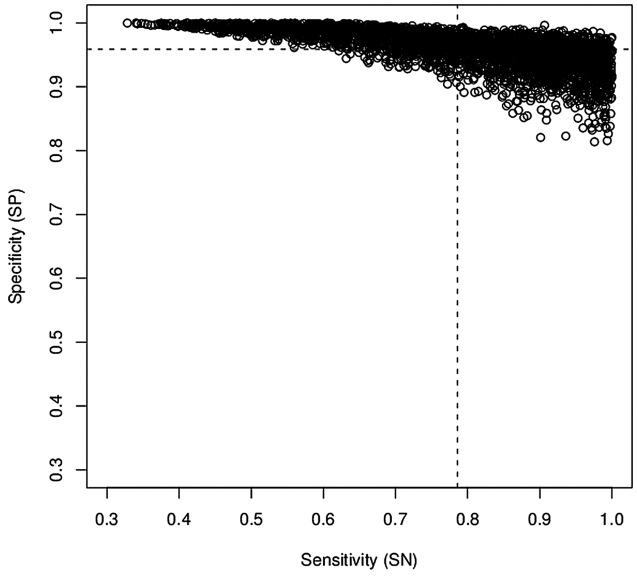

Figure 1 Plausible pairs of sensitivity (SN) and specificity (SP) values for exposure-assessment methods for polycyclic aromatic hydrocarbons evaluated in ref. 1; hashed lines denote means.

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.[7 8] 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.[9]

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.

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.