What do measures of agreement (κ) tell us about quality of exposure assessment?

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-003952
Article Type:	Research
Date Submitted by the Author:	05-Sep-2013
Complete List of Authors:	Burstyn, Igor; Drexel University, Department of Environmental and Occupational Health de Vocht, Frank; The University of Manchester, Occupational and Environmental Health Research Gustafson, Paul; University of British Columbia, Statistics
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Epidemiology, Occupational and environmental medicine
Keywords:	Epidemiology < TROPICAL MEDICINE, OCCUPATIONAL & INDUSTRIAL MEDICINE, STATISTICS & RESEARCH METHODS

BMJ Open

What do measures of agreement (κ) tell us about quality of exposure assessment?

Igor Burstyn[1*], Frank de Vocht[2], Paul Gustafson[3]

1: Department of Environmental and Occupational Health, School of Public Health, Drexel University, Philadelphia, PA, 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, Canada.

*: corresponding authors: Tel: 215.762.2909 | Fax: 215.762.8846 | email: igor.burstyn@drexel.edu

Abstract:

Reliability of binary exposure classification methods is routinely reported in occupational health literature because it is viewed as an important component of evaluating trustworthiness of the exposure assessment by experts. Kappa statistics (κ) are typically employed to assess how well raters or classification systems agree in a variety of contexts, such as identifying exposed subjects 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. 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. Our approach allows 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.

Article summary

(1) Article Focus

- Although evaluation of reliability of exposure classification is routine in occupational epidemiology, little is known about how to use this information to access validity of exposure classification
- We developed procedure for inferring sensitivity and specificity from evaluation of inter-rater agreement that is suitable for Bayesian analysis of data.

(2) Key Messages

- Information about reliability of exposure classifiers contains information about validity of exposure estimator.
- Our method is essential step before epidemiological studies that use misclassified binary exposure estimates can correct for exposure misclassification when only reliability of classification is known.

(3) Strengths and Limitations.

- 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 plausible extent of exposure misclassification.
- The main limitation of our work is that does not yet account for correlated errors in exposure estimates that are common on 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 important component of evaluating trustworthiness of the exposure assessment. Kappa statistics (κ) are typically employed to assess how well raters or classification systems agree in a variety of contexts, such as identifying exposed subjects in a population-based epidemiological study of risks due to occupational exposures. Most recently in this journal, 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 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 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 (aka gold standard) [3]. But how does one infer what κ tells us about validity of exposure estimates (i.e. *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 prevalence of exposure. We illustrate this method in the context of a comparison of job-exposure matrices assessing occupational exposures to polycyclic aromatic hydrocarbons (PAH) [1].

Method

We propose a simulation-based method to calculate 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 rates with the same validity, by:

$$\kappa = (Pr\times(SP\text{-}1+SN)^2)\times(Pr\text{-}1)/((Pr\times SN\text{-}SP\text{-}Pr+Pr\times SP)\times(Pr\times SN+1\text{-}SP\text{-}Pr+Pr\times SP)) \quad \text{[Eq. 1.]}$$

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

$$SN+SP>1 \quad \text{[Eq.2]}$$

We first 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 Beta distribution: $Pr \sim Beta(c, d)$. Information required to specify these distributions with reasonable credibility is available in reports evaluating inter-rater agreements, as in [1]. We can then calculate (multiple) lower bounds of *SN* and *SP* (SN_l and SP_l) that are consistent with these distributions, following:

$$SN_l = \kappa_l/((1\text{-}Pr) + \kappa_l\times Pr), \text{ and} \quad \text{[Eq.3]}$$

$$SP_l = \kappa_l/(Pr + \kappa_l\times(1\text{-}Pr)) \quad \text{[Eq.4]}$$

The upper theoretical bounds on *SN* and *SP* are known (i.e. 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 bound of 1. Using Monte Carlo sampling this procedure is repeated multiple times to generate sets of possible (*SN*, *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) (Eq. 3 and 4), and is followed by [b] selecting candidate set (*SN*, *SP*) from values uniformly distributed between lower bounds, (SN_l, SP_l), and completed by [c] imposing constraints on the candidate set of (*SN*, *SP*) that are implied by Eq. 1 and 2 (see next paragraph for details of the last step).

By chance, some values of *Pr*, *SN* and *SP* selected in this way will correspond to values of κ, implied by by Eq. 1, that lie outside of bounds on κ that we have specified by choosing specific values of κ_l and upper κ_h from corresponding distributions. Furthermore, some combinations of *SN* and *SP* will not be consistent with Eq. 2 (i.e. 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 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 ***eAppendix*** with input values specific to the illustrative example described below. There is no additional data to share.

Because this research did not involve human subjects, ethics clearance was not required.

This research was author-initiated and unfunded.

Results

We apply out 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 lower (κ_l) and upper (κ_h) bounds of κ for PAH by using uniform distributions (*U*) as $\kappa_l \sim U(0.29, 0.31)$ and $\kappa_h \sim U(0.59, 0.61)$. Some degree of judgments 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 $Pr \sim Beta(6.2, 99.7)$[5]. The results of the rest of the calculations are summarized in the Figure, 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 (standard deviation (sd) 0.15) and mean *SP* is about 0.96 (sd 0.03).

Discussion

Our approach allows 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 (e.g. [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 impact of exposure misclassification on epidemiologic 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 JEM, produced 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 prevalence of exposure, 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 in epidemiological analyses [9].

Figure: Plausible pairs of SN and SP values for PAH exposure assessment methods evaluated in [1]; hashed lined denote means

Funding

None

Competing Interests

None

Contributorship

All authors equally contributed to writing 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.

Data sharing

No additional data available.

Reference List

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** (10):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** (9):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** (5):930-3.

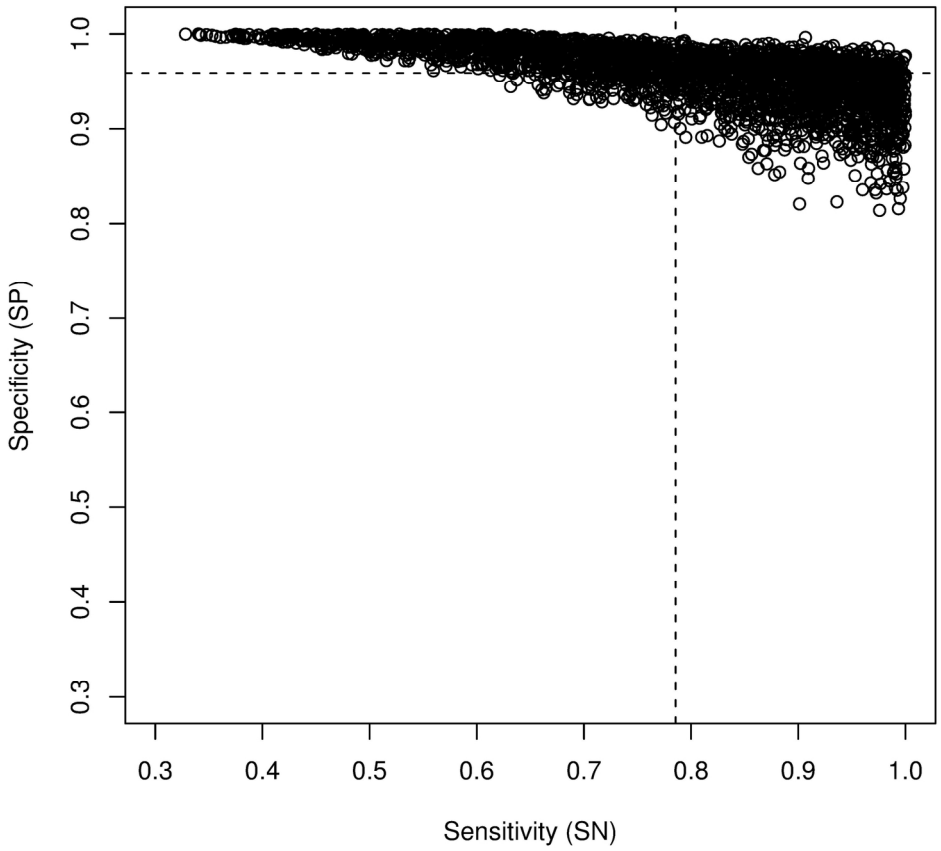
5 Chun-Lung Su. Bayesian Epidemiologic Screening Techniques. 2013.

6 MacLehose RF, Gustafson P. Is probabilistic bias analysis approximately Bayesian? *Epidemiology* 2012;**23** (1):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** (27):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** (6):719-27.

9 Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall/CRC Press 2004.



Plausible pairs of SN and SP values for PAH exposure assessment methods evaluated in [1]; hashed lined denote means
165x165mm (300 x 300 DPI)

```
################################################################################
#APPENDIX: What do measures of agreement (κ) tell us about quality of exposure assessment?
#script that is to be implemented in R software
################################################################################
#START

##INPUTS
k<-10000 #size of simulation
#informed by DATA from PAH from http://oem.bmj.com/content/69/10/745.full
KP.LO<-runif(k, 0.29, 0.31) #UNIFORM DISTN of lower bound on kappa
KP.HI<-runif(k, 0.59, 0.61) # UNIFORM DISTN of high bound of kappa
PREV.CLBRT<-rbeta(k, 6.1946, 99.6983) #BETA distribution of exposure prevalence

##CALCULATIONS
#lower bound on SN and SP
SN.LO<-KP.LO/((1-PREV.CLBRT) + KP.LO*PREV.CLBRT)
SP.LO<-KP.LO/(PREV.CLBRT + KP.LO*(1-PREV.CLBRT))

#unconstrained priors on SN and SP
SN<-runif(k, SN.LO,1)
SP<-runif(k, SP.LO,1)

#apply constraints
p<-PREV.CLBRT
kappa.naive<-(p*(SP-1+SN)^2)*(p-1)/((p*SN-SP-p+p*SP)*(p*SN+1-SP-p+p*SP))
lo<-rep(0, k)
hi<-rep(0, k)
for (i in 1:k) {if(kappa.naive[i] < KP.LO[i] ) lo[i] <- 1}
sum(lo)
for (i in 1:k) {if(kappa.naive[i] > KP.HI[i] ) hi[i] <- 1}
sum(hi)
random<-rep(0, k)
add<-SN+SP
for (i in 1:k) {if(add[i]<1) random[i] <- 1}
sum(random)

#prior after constraints
pq1<-cbind(SN, SP, lo, hi, random)
pq2<-data.frame(pq1)
prior_ <- subset(pq2, lo == 0 & hi == 0 & random==0)

##PRESENT RESULTS IN A FIGURE
plot(prior_$SN, prior_$SP, xlab="Sensitivity (SN)", ylab="Specificity (SP)", xlim=c(0.3, 1), ylim=c(0.3, 1))
length(prior_$SN)
abline(v=mean(prior_$SN), lty=2)
abline(h=mean(prior_$SP), lty=2)

# END
```

BMJ open

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

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-003952.R1
Article Type:	Research
Date Submitted by the Author:	25-Oct-2013
Complete List of Authors:	Burstyn, Igor; Drexel University, Department of Environmental and Occupational Health de Vocht, Frank; The University of Manchester, Occupational and Environmental Health Research Gustafson, Paul; University of British Columbia, Statistics
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Epidemiology, Occupational and environmental medicine
Keywords:	Epidemiology < TROPICAL MEDICINE, OCCUPATIONAL & INDUSTRIAL MEDICINE, STATISTICS & RESEARCH METHODS

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]

1: Department of Environmental and Occupational Health, School of Public Health, Drexel University, Philadelphia, PA, 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, Canada.

*: corresponding authors: Tel: 215.762.2909 | Fax: 215.762.8846 | email: igor.burstyn@drexel.edu

Abstract:

Background: Reliability of binary exposure classification methods is routinely reported in occupational health literature because it is viewed as an important component of evaluating trustworthiness of the exposure assessment by experts. Kappa statistics (κ) are typically employed to assess how well raters or classification systems agree in a variety of contexts, such as identifying exposed subjects 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 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.

Article summary

(1) Article Focus

- Although evaluation of reliability of exposure classification is routine in occupational epidemiology, little is known about how to use this information to access validity of exposure classification.
- We developed procedure for inferring sensitivity and specificity from evaluation of inter-rater agreement that is suitable for Bayesian analysis of data.

(2) Key Messages

- Information about reliability of exposure classifiers contains information about validity of exposure estimator.
- Our method is essential step before epidemiological studies that use misclassified binary exposure estimates can correct for exposure misclassification when only reliability of classification is known.

(3) Strengths and Limitations.

- 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 plausible extent of exposure misclassification.
- The main limitation of our work is that 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 trustworthiness of the exposure assessment. Kappa statistics (κ) are typically employed to assess how well raters or classification systems agree in a variety of contexts, such as identifying exposed subjects 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 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 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 (also known as "gold standard") [3]. But how does one infer what κ tells us about the validity of exposure estimates (i.e. *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 prevalence of exposure. We illustrate this method in the context of a comparison of job-exposure matrices assessing occupational exposures to polycyclic aromatic hydrocarbons (PAH) [1].

Method

We propose a simulation-based method to calculate 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:

$$\kappa = (Pr\times(SP\text{-}1+SN)^2)\times(Pr\text{-}1)/((Pr\times SN\text{-}SP\text{-}Pr+Pr\times SP)\times(Pr\times SN+1\text{-}SP\text{-}Pr+Pr\times SP)) \quad \text{[Eq. 1.]}$$

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

$$SN+SP>1 \quad \text{[Eq.2]}$$

We first 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 Beta distribution: $Pr \sim Beta(c, d)$. Information required to specify these distributions with reasonable credibility is available in reports evaluating inter-rater agreements, as in [1]. We can then calculate (multiple) lower bounds of *SN* and *SP* (SN_l and SP_l) that are consistent with these distributions, following:

$$SN_l = \kappa_l/((1\text{-}Pr) + \kappa_l\times Pr), \text{ and} \quad \text{[Eq.3]}$$

$$SP_l = \kappa_l/(Pr + \kappa_l\times(1\text{-}Pr)) \quad \text{[Eq.4]}$$

The upper theoretical bounds on *SN* and *SP* are known (i.e. 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 bound of 1. Using Monte Carlo sampling this procedure is repeated multiple times to generate sets of possible (*SN*, *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) (Eq. 3 and 4), and is followed by [b] selecting candidate set (*SN*, *SP*) from values uniformly distributed between lower bounds, (SN_l, SP_l), and upper theoretical maximum of 1, and completed by [c] imposing constraints on the candidate set of (*SN*, *SP*) that are implied by Eq. 1 and 2 (see next paragraph for details of the last step). The purpose of step [a] in the procedure is to calculate lower bounds on sensitivity and specificity. The purpose of step [b] is to sample candidate values of sensitivity and specificity that lie between their respective theoretical lower and upper boundaries. The purpose of step [c] is to limit the sets of values of sensitivity and specificity 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 by Eq. 1, that lie outside of bounds on κ that we have specified by choosing specific values of κ_l and upper κ_h from corresponding distributions. Furthermore, some combinations of *SN* and *SP* will not be consistent with Eq. 2 (i.e. 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 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 ***eAppendix*** with input values specific to the illustrative example described below. There is no additional data to share.

Because this research did not involve human subjects, ethics clearance was not required.

This research was author-initiated and unfunded.

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 lower (κ_l) and upper (κ_h) bounds of κ for PAH by using uniform distributions (*U*) as $\kappa_l \sim U(0.29, 0.31)$ and $\kappa_h \sim U(0.59, 0.61)$. Some degree of judgments 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 summarized in the Figure, 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 (standard deviation (sd) 0.15) and mean *SP* is about 0.96 (sd 0.03).

Discussion

Our approach allows 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 (e.g. [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 impact of exposure misclassification on epidemiologic 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 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 prevalence of exposure, 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].

Figure legend:

Figure : Plausible pairs of SN and SP values for exposure assessment methods for polycyclic aromatic hydrocarbons evaluated in [1]; hashed lined denote means

Data sharing

No additional data available.

Funding

None

Competing Interests

None

Contributorship

All authors equally contributed to writing 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.

Reference List

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** (10):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** (9):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** (5):930-3.

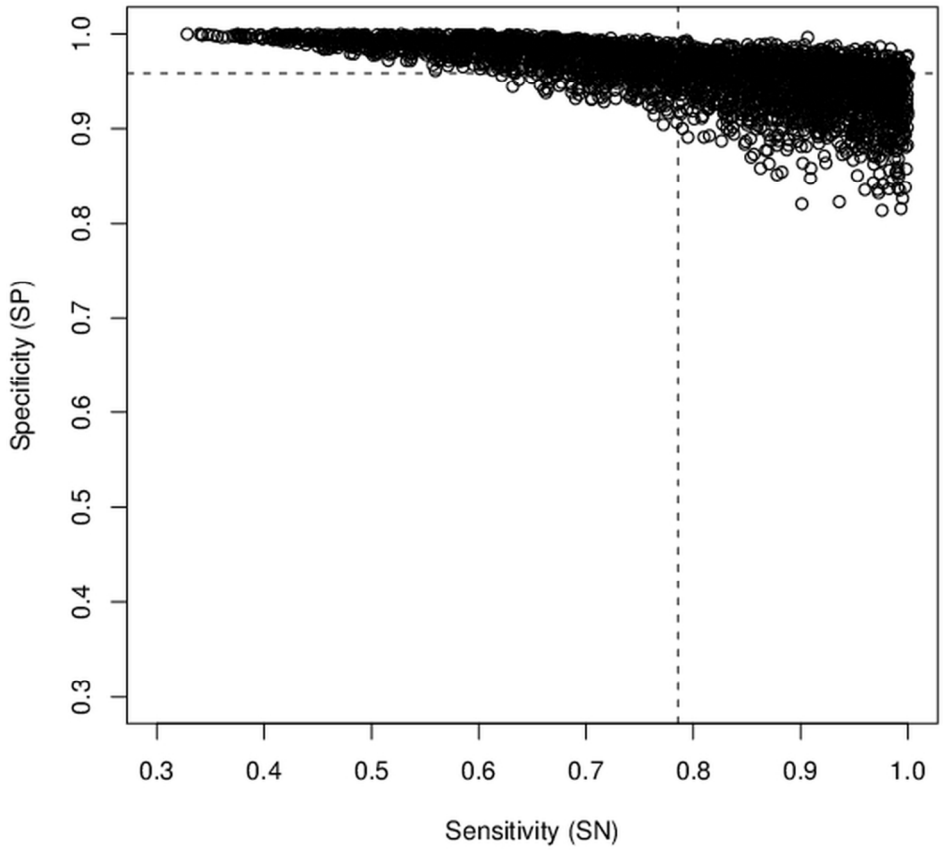
5 Chun-Lung Su. Bayesian Epidemiologic Screening Techniques. 2013.

6 MacLehose RF, Gustafson P. Is probabilistic bias analysis approximately Bayesian? *Epidemiology* 2012;**23** (1):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** (27):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** (6):719-27.

9 Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall/CRC Press 2004.



Plausible pairs of SN and SP values for exposure assessment methods for polycyclic aromatic hydrocarbons evaluated in [1]; hashed lined denote means
90x90mm (300 x 300 DPI)

```
################################################################################
#APPENDIX: What do measures of agreement (κ) tell us about quality of exposure assessment?
#script that is to be implemented in R software
################################################################################
#START

##INPUTS
k<-10000 #size of simulation
#informed by DATA from PAH from http://oem.bmj.com/content/69/10/745.full
KP.LO<-runif(k, 0.29, 0.31) #UNIFORM DISTN of lower bound on kappa
KP.HI<-runif(k, 0.59, 0.61) # UNIFORM DISTN of high bound of kappa
PREV.CLBRT<-rbeta(k, 6.1946, 99.6983) #BETA distribution of exposure prevalence

##CALCULATIONS
#lower bound on SN and SP
SN.LO<-KP.LO/((1-PREV.CLBRT) + KP.LO*PREV.CLBRT)
SP.LO<-KP.LO/(PREV.CLBRT + KP.LO*(1-PREV.CLBRT))

#unconstrained priors on SN and SP
SN<-runif(k, SN.LO,1)
SP<-runif(k, SP.LO,1)

#apply constraints
p<-PREV.CLBRT
kappa.naive<-(p*(SP-1+SN)^2)*(p-1)/((p*SN-SP-p+p*SP)*(p*SN+1-SP-p+p*SP))
lo<-rep(0, k)
hi<-rep(0, k)
for (i in 1:k) {if(kappa.naive[i] < KP.LO[i] ) lo[i] <- 1}
sum(lo)
for (i in 1:k) {if(kappa.naive[i] > KP.HI[i] ) hi[i] <- 1}
sum(hi)
random<-rep(0, k)
add<-SN+SP
for (i in 1:k) {if(add[i]<1) random[i] <- 1}
sum(random)

#prior after constraints
pq1<-cbind(SN, SP, lo, hi, random)
pq2<-data.frame(pq1)
prior_ <- subset(pq2, lo == 0 & hi == 0 & random==0)

##PRESENT RESULTS IN A FIGURE
plot(prior_$SN, prior_$SP, xlab="Sensitivity (SN)", ylab="Specificity (SP)", xlim=c(0.3, 1), ylim=c(0.3, 1))
length(prior_$SN)
abline(v=mean(prior_$SN), lty=2)
abline(h=mean(prior_$SP), lty=2)

# END
```

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]

1: Department of Environmental and Occupational Health, School of Public Health, Drexel University, Philadelphia, PA, 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, Canada.

*: corresponding authors: Tel: 215.762.2909 | Fax: 215.762.8846 | email: igor.burstyn@drexel.edu

Abstract:

Background: Reliability of binary exposure classification methods is routinely reported in occupational health literature because it is viewed as an important component of evaluating trustworthiness of the exposure assessment by experts. Kappa statistics (κ) are typically employed to assess how well raters or classification systems agree in a variety of contexts, such as identifying exposed subjects 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 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.

Article summary

(1) Article Focus

- Although evaluation of reliability of exposure classification is routine in occupational epidemiology, little is known about how to use this information to access validity of exposure classification.
- We developed procedure for inferring sensitivity and specificity from evaluation of inter-rater agreement that is suitable for Bayesian analysis of data.

(2) Key Messages

- Information about reliability of exposure classifiers contains information about validity of exposure estimator.
- Our method is essential step before epidemiological studies that use misclassified binary exposure estimates can correct for exposure misclassification when only reliability of classification is known.

(3) Strengths and Limitations.

- 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 plausible extent of exposure misclassification.
- The main limitation of our work is that does not yet account for correlated errors in exposure estimates that are common ~~on~~ 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 trustworthiness of the exposure assessment. Kappa statistics (κ) are typically employed to assess how well raters or classification systems agree in a variety of contexts, such as identifying exposed subjects in a population-based epidemiological study of risks due to occupational exposures. Most recently ~~in this journal~~, 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 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 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 (~~aka~~ also known as "gold standard") [3]. But how does one infer what κ tells us about the validity of exposure estimates (i.e. *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 prevalence of exposure. We illustrate this method in the context of a comparison of job-exposure matrices assessing occupational exposures to polycyclic aromatic hydrocarbons (PAH) [1].

Method

We propose a simulation-based method to calculate 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:

$$\kappa = (Pr\times(SP\text{-}1+SN)^2)\times(Pr\text{-}1)/((Pr\times SN\text{-}SP\text{-}Pr+Pr\times SP)\times(Pr\times SN+1\text{-}SP\text{-}Pr+Pr\times SP)) \quad \text{[Eq. 1.]}$$

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

$$SN+SP>1 \quad \text{[Eq.2]}$$

We first 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 Beta distribution: $Pr \sim Beta(c, d)$. Information required to specify these distributions with reasonable credibility is available in reports evaluating inter-rater agreements, as in [1]. We can then calculate (multiple) lower bounds of *SN* and *SP* (SN_l and SP_l) that are consistent with these distributions, following:

$$SN_l = \kappa_l/((1\text{-}Pr) + \kappa_l\times Pr), \text{ and} \quad \text{[Eq.3]}$$

$$SP_l = \kappa_l/(Pr + \kappa_l\times(1\text{-}Pr)) \quad \text{[Eq.4]}$$

The upper theoretical bounds on *SN* and *SP* are known (i.e. 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 bound of 1. Using Monte Carlo sampling this procedure is repeated multiple times to generate sets of possible (*SN*, *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) (Eq. 3 and 4), and is followed by [b] selecting candidate set (*SN*, *SP*) from values uniformly distributed between lower bounds, (SN_l, SP_l), and upper theoretical maximum of 1, and completed by [c] imposing constraints on the candidate set of (*SN*, *SP*) that are implied by Eq. 1 and 2 (see next paragraph for details of the last step). The purpose of step [a] in the procedure is to calculate lower bounds on sensitivity and specificity. The purpose of step [b] is to sample candidate values of sensitivity and specificity that lie between their respective theoretical lower and upper boundaries. The purpose of step [c] is to limit the sets of values of sensitivity and specificity 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 by Eq. 1, that lie outside of bounds on κ that we have specified by choosing specific values of κ_l and upper κ_h from corresponding distributions. Furthermore, some combinations of *SN* and *SP* will not be consistent with Eq. 2 (i.e. 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 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 ***eAppendix*** with input values specific to the illustrative example described below. There is no additional data to share.

Because this research did not involve human subjects, ethics clearance was not required.

This research was author-initiated and unfunded.

Results

We apply our~~t~~ 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 lower (κ_l) and upper (κ_h) bounds of κ for PAH by using uniform distributions (*U*) as $\kappa_l \sim U(0.29, 0.31)$ and $\kappa_h \sim U(0.59, 0.61)$. Some degree of judgments 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 $Pr \sim Beta(6.2, 99.7)$[5]. The results of the rest of the calculations are summarized in the Figure, 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 (standard deviation (sd) 0.15) and mean *SP* is about 0.96 (sd 0.03).

Discussion

Our approach allows 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 (e.g. [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 impact of exposure misclassification on epidemiologic 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 ~~JEM~~job-exposure matrix~~,~~ produced 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 prevalence of exposure, 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].

Figure: Plausible pairs of SN and SP values for ~~PAH~~ exposure assessment methods for polycyclic aromatic hydrocarbons evaluated in [1]; hashed lined denote means

Reference List

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** (10):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** (9):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** (5):930-3.

5 Chun-Lung Su. Bayesian Epidemiologic Screening Techniques. 2013.

6 MacLehose RF, Gustafson P. Is probabilistic bias analysis approximately Bayesian? *Epidemiology* 2012;**23** (1):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** (27):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** (6):719-27.

9 Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall/CRC Press 2004.