Published in final edited form as:
Curr Environ Health Rep. 2018 September ; 5(3): 338–350. doi:10.1007/s40572-018-0205-0.

Bayesian correction for exposure misclassification and evolution of evidence in two studies of the association between maternal occupational exposure to asthmagens and risk of autism spectrum disorder

Alison B. Singer, PhD[1,2], **M. Daniele Fallin, PhD**[3], and **Igor Burstyn, PhD**[4]

[1]Department of Epidemiology and Wendy Klag Center for Autism and Developmental Disabilities, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA

[2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[3]Department of Mental Health, Department of Epidemiology, and the Wendy Klag Center for Autism and Developmental Disabilities, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA

[4]Department of Environmental and Occupational Health, Department of Epidemiology and Biostatistics, and the A.J. Drexel Autism Institute, Drexel University Dornsife School of Public Health, Philadelphia, PA, USA

Abstract

Purpose of the article: Inference in epidemiologic studies is plagued by exposure misclassification. Several methods exist to correct for misclassification error. One approach is to use point estimates for the sensitivity (Sn) and specificity (Sp) of the tool used for exposure assessment. Unfortunately, we typically do not know the Sn and Sp with certainty. Bayesian methods for exposure misclassification correction allow us to model this uncertainty via distributions for Sn and Sp. These methods have been applied in epidemiologic literature, but are not considered a mainstream approach, especially in occupational epidemiology.

Recent findings: Here we illustrate an occupational epidemiology application of a Bayesian approach to correct for the differential misclassification error generated by estimating occupational exposures from job codes using a job exposure matrix (JEM).

Summary: We argue that analyses accounting for exposure misclassification should become more commonplace in the literature.

Keywords

exposure misclassification; occupational epidemiology; Bayesian; asthma; autism

Corresponding Author: Igor Burstyn, igor.burstyn@drexel.edu, Phone: (267)359.6062, Fax: (267)359.6006.

Conflict of Interest: The authors declare that they have no conflict of interest.

Human and Animal Rights: All reported studies with human or animal subjects performed by the authors have been previously published and complied with all applicable ethical standards.

Introduction

Heterogeneity of estimates of associations produced by epidemiologic analyses present a challenge for synthesis of evidence. One plausible cause of such heterogeneity is rooted in differences in the accuracy of exposure assessment since misclassification of a binary exposure may influence the location and uncertainty of the effect estimate. If non-differential misclassification of exposure is ignored, the association between exposure and disease is expected (for an infinite sample size) to be attenuated towards the null. Such non-differential misclassification can reduce power and lead to missed associations as well as false positives (1). In the case of differential exposure misclassification, effect estimates can be expected to be biased towards or away from the null, making interpretation particularly perilous. Differential misclassification with respect to outcome is of particular concern in studies where exposure assessment is collected following disease ascertainment. A typical approach to minimizing it involves separating the methods of data collection for the health outcome and determinants of exposure. However, differential exposure misclassification can also result from dichotomizing a continuous exposure that is measured with error. This type of misclassification has been called “differential due to dichotomization” (DDD) misclassification (2, 3). A detailed discussion of how differential exposure misclassification arises from non-differential measurement error is outlined in Appendix A, following the derivation given in Gustafson (2). Differential misclassification due to dichotomization can only occur when the exposure and outcome are associated, through either causality or confounding. Since we conduct studies with the idea that a relationship may exist between exposure and outcome, we should more carefully consider the possibility of differential misclassification in our analyses. Differential misclassification from this source cannot be remedied by common study design strategies, such as blinding the exposure assessment from the health status.

Many methods have been proposed to account for misclassification in the context of retrospective case-control studies where misclassification parameters (sensitivity, Sn, and specificity, Sp) are not known exactly. Such approaches are motivated by the observation that misclassification-corrected odds ratios obtained using a particular point estimate of Sn and Sp can be highly sensitive to small differences between the actual and guessed values (4, 5). More reliable estimation is obtained when we take into account the uncertainty regarding the actual values of the misclassification parameters. This can be accomplished by implementing a Bayesian approach where one posits prior distributions for the Sn and Sp (4). Intuitively, the procedure samples from these distributions and corrects for misclassification over many iterations. The procedure can distinguish between “good” and “poor” guesses of Sn and Sp by reconciling them with observed data and other parameters and models involved via evaluation of likelihood. Such methods obtain samples from the posterior distribution of the true (misclassification-corrected) effect estimates relating exposure to outcome. A detailed description of the approach, Bayesian Markov Chain Monte Carlo (MCMC) algorithms, can be found in Gustafson (2). Applications of these methods can be used to reconcile heterogeneity in effect estimates between different effect estimates obtained within a study using different exposure metrics (6), and this holds promise for resolving between-study heterogeneity as well. Luta et al. (7) presented and illustrated these methods in an accessible form that aids their dissemination to epidemiologists in general.

We apply Bayesian methods for correction of exposure misclassification in two population based case-control studies of the association between occupational asthmagen (agents that can trigger asthma) exposure and autism spectrum disorder (ASD): the Study to Explore Early Development (SEED) conducted in the United States (8) and a study nested within the Danish Registers (9). In both studies, we estimated maternal occupational asthmagen exposure job codes with an asthma-specific job exposure matrix (JEM) (10). We also illustrate the Bayesian evolution of evidence from a first hypothesis about exposure-ASD association and quality of exposure assessment postulated before the conduct of our first analysis in the smaller US sample (8), followed by integration of results from our first analysis, about both the exposure-ASD relationship and quality of exposure assessment, into the second, much larger, Danish study (9). We realize that knowledge derived from the U.S. regarding Sn and Sp of the JEM may not be entirely applicable to the study in Denmark due to differences in how occupational histories were collected, and address this by adjusting priors on the misclassification parameters to reflect this added uncertainty. We show the resulting effect size estimates and model of exposure misclassification that represent what we have learnt about the influence of maternal asthmagen exposure on risk of ASD and quality of exposure assessment tools.

Methods

Case-control studies

The first is the Study to Explore Early Development (SEED), a United States multi-site, case-control study designed to investigate risk factors, co-morbidities, and phenotypes of ASD (11). We focus on comparisons between the ASD cases (N=463) and population (POP) controls (N=710). Participants were required to be born and reside in one of six study catchment areas in California, Colorado, Georgia, Maryland, North Carolina, and Pennsylvania between September 1, 2003 and August 31, 2006. Children in the ASD were ascertained through service and educational providers for children with developmental disabilities, whereas POP children were identified through random sampling of vital records (11). ASD classification was based on results from the Autism Diagnostic Observation Study (ADOS) and the Autism Diagnostic Interview Revised (ADI-R) when children were 30 to 68 months of age. Maternal job histories were collected as part of computer assisted telephone interview shortly after enrollment in the study. Mothers reported jobs held for one month or more for at least ten hours per week from three months prior to the end of the pregnancy until the child was born or the mother stopped breastfeeding. As detailed in Singer, Windham (8), we coded jobs according to the International Labor Organization’s International Standard Classification of Occupations 1988 (ISCO-88) (12). Analyses were restricted to mothers reporting at least one job that overlapped with the pregnancy.

The second case-control study is nested in the Danish Registers (9). The sample consists of 29,359 controls and 6,706 ASD cases singleton births born in Denmark between January 1, 1993 and December 31, 2007 with an employed mother in the year representing the majority of the pregnancy. Details of the selection of the study sample are described elsewhere (9).

ASD cases were defined as having a reported ICD-10 diagnosis of autism spectrum disorder (ICD-10 codes: F84.0, F84.1, F84.5, F84.8 and F84.9) in the Danish Central Psychiatric Register (DPCR) (13) from January 1, 1995 into April 2013. We linked the ASD cases and controls to maternal occupation (Danish International Standard Classification of Occupation, DISCO-88) and industry codes (NACE codes) from the Employment Classification Module (AKM) (14) to determine maternal occupation and industry in the year that overlapped most with the pregnancy.

The specific details of the exposure assessment using asthma-specific JEM (10) for each study is described elsewhere (8, 9), but they broadly followed the procedures prescribed by Kennedy et al (2000). The JEM produces a binary indicator regarding presence of exposure to a compound or mixture known or strongly suspected of causing occupational asthma. Sensitivity and specificity of the JEM were previously evaluated in studies of occupational asthma under the assumption of non-differential exposure misclassification (15, 16), which revealed evidence in support of higher Sp compared to Sn, as intended by the creators of the JEM.

Correction for misclassification of exposure

Overview—The goal of this analysis is to correct for exposure misclassification generated by using an asthma JEM to classify exposure based on job codes into a binary exposure indicator. We begin with correction for exposure misclassification in an individual-level analysis that includes adjustment for covariates in the SEED study. For purely illustrative (not inferential) purposes, we contrast with a model where we assume near perfect exposure classification, as is typically done in occupational epidemiology studies (8), to those in which we assume exposure misclassification based on prior knowledge about the performance of the JEM in different settings. We allow the differential misclassification by case status but let data inform its extent by examining posterior distributions of Sn and Sp for cases and controls. Thus, we present two different models (Table 1): (1) assuming almost perfect classification of exposure, allowing for differential misclassification (model S_1), and (2) setting priors on misclassification parameters based on previous studies, allowing for differential exposure misclassification (model S_2). It is important to note that while we allow for differential misclassification of exposure, we do not force it to be differential, because the priors on Sn and Sp are the same for cases and controls.

In the second portion of the analysis, we apply misclassification correction using data from a two-by-two contingency table of maternal occupational asthmagen exposure by ASD case-control status from the Danish study, with Sn and Sp priors informed by the SEED study. (We do not conduct individual-level analysis in the Danish study because of logistical challenges in access to individual level data.) We set priors on Sn, Sp and the odds ratio based on posterior distributions from the SEED model S_2 (model D_1). We also acknowledge that performance of the JEM in SEED may have been different than in the Danish analysis and therefore also conduct an additional analysis with priors on Sn and Sp derived on the basis of SEED for model D_1 to have the same location but greater variance, leading to model D_2.

Since we use individual level data for the SEED and only a contingency table as data for the Denmark Bayesian analyses, the model specifications are different. However, all Bayesian analyses that account for exposure misclassification or measurement error share common features. We must specify three models: (a) an exposure model, (b) a measurement (misclassification) model and (c) an outcome model, following Gustafson (2). The details of model specification particular to our studies as well as details of sampling for the posterior distributions are given in Appendix B. Code is presented in Appendix C. We focus next on specifics of our analyses that deal with elucidation and specification of the prior distributions.

SEED: Specification of priors—We express the typical implicit assumption of almost perfect classification of exposure by setting prior distributions of *Beta* (1000,1) on Sn and Sp. These distributions have a mean of 0.999 and standard deviation of 0.001, and are constrained between 0 and 1 by the properties of the Beta distribution. These results should be roughly equivalent to typical uncorrected analyses.

We derived realistic priors on Sn and Sp of the JEM from previous literature. We detail here the origin of these priors, which were initially elucidated through expert opinion and then were updated with data from two analyses. Liu et al (15) asked experts (two occupational physicians and one occupational hygienist) to report the best guess, upper bound, and lower bound of true value for the Sn and Sp of the asthma-specific JEM. Liu et al (15) obtained posterior distributions for the Sn and Sp by updating these priors using data from workers' compensation claims and physician billing records from Alberta, Canada, to examine the association between asthmagen exposure and new adult onset asthma. Beach et al (16) further updated the Sn and Sp posteriors from the Liu et al (15) in a comparable analysis using similar data from British Columbia, Canada. We used the posterior distributions of sensitivity and specificity of the JEM from Beach et al (16) as a basis for our prior distributions in this analysis because they reflect the synthesis of evidence regarding exposure misclassification by the JEM prior to the SEED analysis.

Beach et al (16) reported Sn and Sp posterior distributions for 16 categories of asthmagen exposures. Since there was not much variability across the asthmagen categories for these misclassification parameters, we averaged the medians, 2.5th percentiles, and 97.5th percentiles to obtain a best guess of the mode, 2.5th percentile, and 97.5th percentile for Sn and Sp distributions for any asthmagen. The best guess of the 2.5th percentile, mode, and 97.5th percentile was 0.133, 0.381, and 0.728 for the Sn and 0.990, 0.992, and 0.994 for the Sp, respectively. We used the *betaExpert* function in R to determine parameters for beta distributions corresponding to the above percentiles and then used these *Beta* distributions as priors for Sn and Sp. Based on this, we set the prior distribution for Sn as *Beta*(3.6, 5.2) and the prior distribution for Sp as *Beta*(1000, 9.1); the priors on Sn and Sp were centered on means of 0.41 and 0.991, respectively.

We assumed a normally distributed uninformative prior with a mean of 0 and variance of 0.5 for all log-odds ratios in the outcome and measurement models, except for the intercept of the outcome model and the coefficient for sex. For these two coefficients, we set an uninformative normally distributed prior with a mean of 0 and a variance of 1. The $N(0, 0.5)$

distribution implies that the true odds ratio is roughly between 1/4 and 4 while a *N*(0, 1) distribution implies that the true odds ratio is roughly between 1/7 and 7. We judged these to be sufficiently vague and realistic while not informing the direction the associations. This was consistent with this being the first ever analysis of the association in question.

Denmark: Specification of priors—In the second series of Bayesian analyses, we update beliefs from the SEED study with a contingency table of data from a Denmark register study. Crude and adjusted odds ratios for the association between maternal asthmagen exposure and ASD in the Denmark register study are similar (9), so we are not concerned that confounders are not considered in these Bayesian analyses. Because we learnt about the magnitude of the odds ratio (θ=OR) relating the exposure and outcome from the SEED analysis, we set prior distributions on the prevalence of exposure among controls (r_0) and θ to induce a prior distribution on the true exposure prevalence among cases:

$$r_1 = \theta r_0 \Big/ (1 - r_0 + \theta r_0) .$$

In the first model (model D_1), we assume that exposure assessment was the same in the two studies, so we set priors on sensitivity as *Beta*(11.0, 38.5) for controls and *Beta*(8.3, 17.2) for cases; the prior on specificity for both cases and controls was *Beta*(718, 8.2). These distributions were derived from the posterior distribution from the SEED analysis. In the second model (model D_2), we inflate the variances of Sn and Sp to reflect the possibility that that differences in the collection of exposure history in the two studies impacted exposure assessment. We achieved this by setting the upper percentiles of priors on Sn to be 20% greater than that in model D_1, which roughly doubled the variance. Thus, for Model D_2, priors for Sn were *Beta*(5.9, 19.5) for controls and *Beta*(4.2, 8.1) for cases. We set a *Beta*(46.6, 3.4) prior for Sp, which assumes a prior mean of 0.93 and a standard deviation of 0.04. We set a uniform prior only on the true exposure prevalence among the controls, r_0, and set an informative Gaussian prior with a mean of −0.34 with a variance of 0.33 on the log-odds ratio based on the posterior from our analysis of SEED.

Results

The adjusted OR for maternal occupational asthmagen exposure comparing ASD cases to population controls in the SEED study was 1.39 (95% confidence interval (CI): 0.96 – 2.02), following a typical JEM exposure classification approach (8). When we assumed near perfect classification of exposure, which is essentially equivalent to not adjusting for exposure misclassification, we observed a posterior OR of 1.37 (95% CrI: 0.96 – 1.96) (model S_1) (Table 1). In the analysis allowing for differential exposure misclassification by the JEM (Model S_2), the median of the posterior of the adjusted OR was 0.71 (95% CrI: 0.23 – 2.42). The results allowing for differential exposure misclassification by the JEM suggest that the analyses not adjusted for exposure misclassification are positively biased, although the confidence intervals are overlapping.

The posterior distribution of sensitivity is concentrated among smaller values compared to its prior, but priors and posteriors were similar for the specificity (model S_2, Table 1). We also observe that the median of the posterior distribution for the Sn in cases was higher than the median of the Sn posterior distribution for the controls. Though the posterior Sn distributions produced for Sn for cases and controls from Model S_2 overlap, this suggests that measurable differential misclassification could be at play.

In the Danish case-control analysis, we found an inverse association between maternal asthmagen exposure and ASD using typical analytic approaches (crude OR: 0.92, 95% CI: 0.86–0.99) (9). The posterior distributions for Sn and Sp resulting from models D_1 are similar to the posteriors from SEED model S_2 (Table 1). In model D_1 were we set an informative prior on the odds ratio based on the posterior from SEED model S_2, we generate a posterior odds ratio with a median at 0.64 (95% CrI: 0.23 – 1.94). The posterior odds ratio is pulled towards the SEED result by inclusion of this prior. The posterior distribution for the Denmark model D_1 is similar to SEED model S_2 and only a little more concentrated around the median. Thus, despite the large sample size from this second study we learn little new about the effect estimate of interest. When we admit additional uncertainty about Sn and Sp in analysis of Danish study in model D_2, the posterior distribution for the odds ratio (median: 0.68, 95% CrI: 0.23 – 1.97) is similar to the posterior for model D_1 (Table 1). We illustrate how our knowledge of the asthmagen-ASD association and Sn changed from before analysis of SEED through to misclassification correction of the Danish study in Figure 1.

Discussion

In this paper, we illustrate a Bayesian method for correcting for exposure misclassification in the context of two studies examining the association between maternal occupational asthmagen exposure and ASD in the children. Inferentially, our models suggest that there is no measurable association between maternal asthmagen occupational exposure around the time of pregnancy and ASD. This conclusion was consistent with and without exposure misclassification adjustment, although the effect size estimates and confidence limits did fluctuate. We illustrate that it is hard to predict how misclassification of exposure affects every specific analysis. We also argue that it is more difficult than commonly realized to be sure that such misclassification is non-differential with respect to health outcome, even when exposure assessment is blind to outcome. We describe use of Bayesian tools that make correction for misclassification accessible to epidemiologists who collaborate with statisticians. We highlight the importance of setting defensible priors that capture knowledge that existed before the data in any given study is collected. This is particularly important when the data does not allow us to learn about all parameters of interest, as is typically the case with differential exposure misclassification. In doing so, we must take care to avoid showing over-confidence, as would arise from not considering a range of plausible priors. We consider these methodological matters in details below.

Our results nonetheless illustrate key points regarding exposure misclassification in epidemiologic studies. First, we demonstrate here that the odds ratio estimate can move in unexpected ways when we allow for misclassification by the JEM to be differential. In the epidemiologic literature, authors often assert the belief that exposure misclassification is non-differential and thus results are biased to the null, and argue that reported associations are likely not “spurious” under this rationale. However, in practice it may be difficult to know the true extent of differential misclassification. In the SEED study, as in any case-control study, we may suspect that differential recall for many self-reported variables because mothers who have a child with an ASD may recall the pregnancy differently than mothers of typically developing controls. However, mathematically, as discussed in Appendix A, the focus of this paper is the possibility for differential misclassification to occur through the process of dichotomizing a continuous exposure measured with error, regardless of the timing of exposure query (2, 3). The extent to which exposure misclassification is non-differential by dichotomizing a continuous exposure measure depends on the strength of exposure-outcome association (3). However, it can also be a product of uncontrolled confounding. Thus, we assert that it is important to consider the potential impact of differential exposure misclassification. If the true association does not exist and there is no uncontrolled confounding, then we do not expect exposure misclassification to deviate from non-differential. The fact that we observe little evidence for measurable differential misclassification in the Denmark study is concordant with the observation of no measurable association between exposure and outcome.

Second, we illustrate that even in situations of relatively large sample sizes, posterior effect estimates are affected by misclassification bias. In our example, we observed a precise protective effect estimate of occupational asthmagen exposure on ASD risk in the Denmark study, using typical analytic approaches. When we corrected for exposure misclassification based on prior evidence, we observed a more protective tendency in point estimate, but the credible intervals widened.

This illustrates the challenges faced in occupational epidemiology. The JEM we chose is among the best of its kind, yet the low sensitivity limits our ability to confidently identify new associations. The inability to improve our estimates and precision in the very large Danish study illustrates that without improving exposure measurements and assessment methods we have perhaps reached the limit of what we can discover with tools like JEMs when the associations are weak. Though we specifically focus on the JEM example here, this concern regarding exposure misclassification exists in any epidemiology study where we classify a continuous exposure measured with error.

One important caveat regarding these models is that we must be careful in setting prior distributions, especially given problems with non-identifiability. Recall that p_i=r_i*Sn + (1-r_i)*(1-Sp) where p_i is observed exposure prevalence and r_i is the true exposure prevalence; i=0 denotes controls and i=1 denotes cases. If we observe an exposure prevalence, p_i, of 0.21, and specificity, Sp, is approximately 1, then there are two possible solutions for (r_i, Sn): (0.3, 0.7) and (0.7, 0.3). The priors will determine the solution that is selected. Thus, if we place a prior on sensitivity that is concentrated on 0.3, the solution for the true exposure prevalence will converge upon 0.7. Since we have prior knowledge of the performance of the JEM and not true exposure prevalence in the selected samples, we place informative priors on the Sn and Sp instead of the true exposure prevalence. In our analysis, the particular identifiability issue only emerges when the specificity is close to one, but illustrates that use of these models should be guided by knowledge. Prior knowledge may also be complicated by the fact model parameters may not be directly transportable across different study populations, suggesting the importance of perhaps considering a few plausible priors.

Conclusion

In epidemiologic studies we often present results as if there is no measurement error despite the fact that it exists. We illustrate the sensitivity of effect size estimates to exposure misclassification in a limited series of studies and show how Bayesian procedures can be readily applied to address exposure misclassification. These methods accommodate our uncertainty regarding the amount of misclassification. We illustrated these methods through use of WinBUGS and R, although there are alternative packages that can be used for Bayesian inference, including OpenBUGS (an open-source version of BUGS software) (17), Just Another Gibbs Sampler (JAGS) (18), and STAN (19). We argue that analyses that account for this misclassification should become more commonplace within the epidemiologic literature in general. Though ultimately, these misclassification error methods cannot replace investing in the development of better exposure assessment tools and validating the quality of currently existing exposure assessment tools.

Acknowledgements:

We thank Professor Paul Gustafson for reviewing our explanation of how “differential due to dichotomization” misclassification arises.

Funding:

ABS was funded by an Autism Speaks Dennis Weatherstone Predoctoral Fellowship (#8576) and by a training grant from the National Institute of Environmental Health Sciences (T32ES007018). The Study to Explore Early Development was funded by six cooperative agreements from the Centers for Disease Control and Prevention: Cooperative Agreement Number U10DD000180, Colorado Department of Public Health; Cooperative Agreement Number U10DD000181, Kaiser Foundation Research Institute (CA); Cooperative Agreement Number U10DD000182, University of Pennsylvania; Cooperative Agreement Number U10DD000183, Johns Hopkins University; Cooperative Agreement Number U10DD000184, University of North Carolina at Chapel Hill; and Cooperative Agreement Number U01DD000498, Michigan State University.

Appendix A

We illustrate the manner in which non-differential measurement error produced differential exposure misclassification below, following Section 6.1 of Gustafson (2) but using heuristics that may be more widely appreciated by non-statisticians.

Imagine a continuous exposure C observed as C* is dichotomized at a constant w, such that into X*=0 when C*<w, and X*=1 when C*≥w. If C (true exposure) was dichotomized at a constant w instead of C* (observed exposure), then we would have X=0 when C<w, and X=1 when C≥w. We assume that C* is a decent surrogate of C and that two are correlated, such that when C* increases so does C and vice versa. This allows us to define:

$$\text{sensitivity as Sn} = Pr(\mathrm{X}^* = 1 \mid \mathrm{X} = 1) = Pr(\mathrm{C}^* \geq w \mid \mathrm{C} \geq w) \text{ and}$$
$$\text{specificity as Sp} = Pr(\mathrm{X}^* = 0 \mid \mathrm{X} = 0) = Pr(\mathrm{C}^* < w \mid \mathrm{C} < w)$$
$$\text{(analogous to expressions 6.3 and 6.4 of Gustafson (2))}.$$

(analogous to expressions 6.3 and 6.4 of Gustafson (2)).

For the moment, let us ignore the fact that we are interested in a binary outcome (such as case *vs.* control) but narrowly focus on exposures that are being segregated into two groups. If the value of *w* at which we divide exposure into "exposed" and "unexposed" reduces towards the lower end of the distribution of continuous exposure, the Sn increases. In other words, by lowering the threshold for defining subjects as exposed, we boost sensitivity, i.e. the chance that the expanded exposed group (X*=1, C*≥*w*) actually contains truly exposed subjects (X=1, C≥*w*). An analogous argument exists for specificity: as *w* decreases, Sp shrinks. The key point to observe is that the location of threshold within the distribution of observed exposures impacts misclassification in a predictable pattern. Consequently, if we have two groups with apparent exposure distributions centered on very different values and subject to the same measurement error, we can expect these two groups to have very different Sn and Sp when we apply the same threshold *w* to both groups. In the extreme case where observed exposure distributions are truncated at *w* (i.e. do not overlap), the classified X* for the more heavily exposed group will have perfect Sn and no Sp, whereas less exposed group will have Sn=0 and Sp=1.

If there is a relationship between Y and C (this does not have to be causal but can be due to uncontrolled confounding), then we can expect there also to exist a relationship between Y and C*. If this relationship is positive, this means that values of both C and C* will be greater among cases then controls, i.e. both the true and observed distributions of exposure among cases will be centered on larger values than that among controls if we condition on the part of the distribution where C is above the threshold, *w*. Please note that **we apply the same threshold *w* to both cases and controls**, but in this scenario, there will be higher chance of exceeding this threshold for cases. Recalling how the location of dichotomization within the distribution of apparent exposures affects Sn and Sp, we see that Sn and Sp must be different for cases and controls when there is an association between exposure and outcome, because the location of threshold within distributions of cases and controls will be different with respect to where true and observed exposures are concentrated. Since we demonstrate that Sn and Sp depend on the outcome without postulating that errors in C* depend on the outcome, we have illustrated how differential misclassification arises from non-differential measurement error.

It is important to note that this process of *differential due to dichotomization* (DDD) (2, p 142) arises only when exposure and outcome are associated, either due to a causal effect or uncontrolled confounding. Therefore, an important corollary of the above argument is that we must consider differential exposure misclassification when we suspect that the exposure is associated with the outcome. In other words, such a model is the most consistent one with a hypothesis (as in our papers on asthmagens and ASD) that exposure and outcome are related. Further reason to adopt a more flexible modeling approach that allows for differential misclassification is given by Gustafson (2, p 146) who warns that there is a risk of over-correction if non-differential misclassification is assumed when DDD is in fact present.

Appendix B

SEED: Model specification

For the SEED Bayesian misclassification correction, we specify three models: (a) an exposure model, (b) a measurement model and (c) an outcome model, following Gustafson (2). We express these three models below, where X indicates being “truly” exposed (X=1) or unexposed (X=0) to an occupational asthmagen; X* represents those classified (i.e. assigned by JEM) as exposed (X*=1) or unexposed (X*=0) to an occupational asthmagen; Sn_0 is the sensitivity of X* among controls; Sn_1 is the sensitivity of X* among ASD cases; Sp_0 is the specificity of X* among controls; Sp_1 is the specificity of X* among ASD cases; Y represents being an ASD case (Y=1) or a control (Y=0); **Z** represents a vector of confounders with coefficients **α** representing the log-odds of true asthmagen exposure, and coefficients **β** representing the log-odds of being an ASD case.

Exposure model:

$$\mathrm{logit}(\Pr(\mathrm{X}=1))=\alpha_0+\boldsymbol{\alpha}\mathbf{Z}$$

(b) Measurement model: differential exposure misclassification

$$\Pr(\mathrm{X}^*=1)=\begin{cases}\text{If Y}=0:\ \Pr(X=1)\mathrm{Sn}_0+(1-\Pr(X=1))(1-\mathrm{Sp}_0)\\ \text{If Y}=1:\ \Pr(X=1)\mathrm{Sn}_1+(1-\Pr(X=1))(1-\mathrm{Sp}_1)\end{cases}$$

(c) Outcome model:

$$\mathrm{logit}(\Pr(\mathrm{Y}=1))=\beta_0+\theta\mathrm{X}+\boldsymbol{\beta}\mathbf{Z}$$

The exposure model expresses the log odds of the true occupational asthmagen exposure conditional on the confounders in the model. The measurement model specifies the probability of observed (misclassified) occupational asthmagen exposure as a function of the probability of true occupational asthmagen exposure, the sensitivity, and the specificity, as well as ASD status. In the outcome model, we model the log odds of having a child with ASD as a function of the true maternal occupational asthmagen exposure status and the potential confounders. Confounders in our analysis included maternal age at child’s birth (continuous), parity (1, 2, >2), child’s sex, maternal race/ethnicity (white, black, Asian, Hispanic, multiracial or other), current maternal education (less than high school, high school, some college/trade school, bachelors, advanced degree), current total household income (<$30,000, $30,000–70,000, $70,000–110,000, >$110,000), maternal psychiatric condition history (yes, no), and active smoking during pregnancy (yes, no).

SEED: Model convergence and characterizing posteriors

We ran a complete case analysis with 463 ASD cases and 710 controls. Bayesian analysis was implemented in Winbugs 1.4 (20) through the R2WinBUGS (21) package in R version

3.1.2. We ran 3 chains for 20,000 iterations of the simulation, removing the initial 5,000 iterations to allow for a burn-in period. In order to reduce autocorrelation between neighboring iterations, we only sampled every 20th iteration of accepted samples from the posterior distribution. We reviewed trace plots, autocorrelation plots, density plots, and Gelman plots to check for convergence. The WinBugs code is included in Appendix C. The Bayesian approach generated posterior distributions for the adjusted odds ratio, consisting of a summary odds ratio (median of the posterior distribution) and a 95% credible interval (corresponding to the 2.5th and 97.5th percentiles of the posterior distribution). Posterior distributions were also obtained for the Sn and the Sp of the JEM, and the maternal occupational asthmagen exposure prevalence among controls.

Denmark: Model specification

In our contingency table, we have observed asthmagen exposure prevalences, X_0 and X_1, for controls and cases, respectively, for N_0 controls and N_1 cases. We assume that observed prevalences follow binomial distributions: $X_0 \sim \mathit{Bin}(p_0, N_0)$ and $X_1 \sim \mathit{Bin}(p_1, N_1)$.If r_0 and r_1 are the true exposure prevalences for cases and controls, respectively, then allowing for differential exposure misclassification we can calculate the true exposure prevalence among controls, $r_0 = (p_0+Sp_0-1)/(Sn_0+Sp_0-1)$, and the true prevalence among cases, $r_1 = (p_1+Sp_1-1)/(Sn_1+Sp_1-1)$. Over many MCMC iterations, we sample candidate values for Sn and Sp for cases and controls from prior distributions based on the SEED analyses and generate posterior distributions for corrected exposure prevalences, r_0 and r_1. Since we learned about the magnitude of the odds ratio (θ) relating the exposure and outcome from analysis of SEED, we set prior distributions on r_0 and log odds ratio to induce a prior distribution on the true exposure prevalence among cases, $r_1= ((\theta)(r_0))/((\theta)(r_0)+1-r_0)$. The distributions for r_0 and r_1 are then reconciled with the observed exposure prevalences. Selected candidate values for sensitivity, specificity, and θ are retained for the posterior distributions if they are deemed plausible based on the likelihood of the data given the model and priors.

Denmark: Model convergence and characterization of posteriors

The contingency table consisted of 5,876 exposed controls, 23,483 unexposed controls, 1,247 exposed cases, and 5,459 unexposed cases. We ran 200,000 iterations, removing the initial 10,000 iterations to allow for a burn-in period and selected every 100th iteration for inclusion in the posterior distribution in order to reduce auto-correlation. We generated posterior distributions for the odds ratio, sensitivity, specificity and exposure prevalence.

Appendix C

WinBugs code for Model S_1:

```
model {

for (i in 1:N) {

# Outcome model : includes the 'true' variable for asthmagen exposure
y[i] ~ dbern(pt[i])

logit(pt[i]) <- b0 + b1*astpregt[i] + b2*DR_AGEBIRTH_MX_C31[i] +
b5*DR_PSYALL_MX[i] +

b6*dr_parity_2[i] + b7*dr_parity_3[i] + b8*dr_sexm0f1[i] + b9*dr_mrace_bla[i] +
b10*dr_mrace_asi[i] + b11*dr_mrace_his[i] + b12*dr_mrace_oth[i] + b13*dr_medu_lhs[i]
+ b14*dr_medu_hs[i] + b15*dr_medu_sc[i] + b16*dr_medu_ad[i] + b17*dr_ti_1[i] +

b18*dr_ti_2[i] + b19*dr_ti_4[i] + b22*DR_ACTSMK_PREG[i]

# Measurement model

DR_ASTHMAGEN_PREG[i] ~ dbern(pm[i])

pm[i] <- SN0*(astpregt[i])*(1-y[i]) + (1-SP0)*(1-astpregt[i])*(1-y[i]) +
SN1*(astpregt[i])*(y[i]) + (1-SP1)*(1-astpregt[i])*(y[i])

# Exposure model

astpregt[i] ~ dbern(prop[i])

logit(prop[i]) <- g1 + g2*DR_AGEBIRTH_MX_C31[i] + g5*DR_PSYALL_MX[i] +

g6*dr_parity_2[i] + g7*dr_parity_3[i] + g8*dr_sexm0f1[i] + g9*dr_mrace_bla[i] +
g10*dr_mrace_asi[i] + g11*dr_mrace_his[i] +g12*dr_mrace_oth[i] + g13*dr_medu_lhs[i] +
g14*dr_medu_hs[i] + g15*dr_medu_sc[i] + g16*dr_medu_ad[i] + g17*dr_ti_1[i] +
g18*dr_ti_2[i] + g19*dr_ti_4[i] + g22*DR_ACTSMK_PREG[i]

}

# Calculate odds ratio

OR <- exp(b1)

# Calculate prevalence of exposure among unexposed

r0 <- (p0+SP0–1)/(SN0+SP0–1)

# PRIORS

b0 ~ dnorm(0,1)

b1 ~ dnorm(0,2)

b2 ~ dnorm(0,2)

b5 ~ dnorm(0,2)

b6 ~ dnorm(0,2)
```

```
b7 ~ dnorm(0,2)

b8 ~ dnorm(0,1)

b9 ~ dnorm(0,2)

b10 ~ dnorm(0,2)

b11 ~ dnorm(0,2)

b12 ~ dnorm(0,2)

b13 ~ dnorm(0,2)

b14 ~ dnorm(0,2)

b15 ~ dnorm(0,2)

b16 ~ dnorm(0,2)

b17 ~ dnorm(0,2)

b18 ~ dnorm(0,2)

b19 ~ dnorm(0,2)

b22 ~ dnorm(0,2)

g1 ~ dnorm(0,2)

g2 ~ dnorm(0,2)

g5 ~ dnorm(0,2)

g6 ~ dnorm(0,2)

g7 ~ dnorm(0,2)

g8 ~ dnorm(0,2)

g9 ~ dnorm(0,2)

g10 ~ dnorm(0,2)

g11 ~ dnorm(0,2)

g12 ~ dnorm(0,2)

g13 ~ dnorm(0,2)

g14 ~ dnorm(0,2)
```

```
g15 ~ dnorm(0,2)

g16 ~ dnorm(0,2)

g17 ~ dnorm(0,2)

g18 ~ dnorm(0,2)

g19 ~ dnorm(0,2)

g22 ~ dnorm(0,2)

SN0 ~ dbeta(1000,1)

SN1 ~ dbeta(1000,1)

SP0 ~ dbeta(1000,1)

SP1 ~ dbeta(1000,1)

}
```

WinBugs code for Model S_2:

```
model {

for (i in 1:N) {

# Outcome model : includes the ‘true’ variable for asthmagen exposure

y[i] ~ dbern(pt[i])

logit(pt[i]) <- b0 + b1*astpregt[i] + b2*DR_AGEBIRTH_MX_C31[i] +
b5*DR_PSYALL_MX[i] + b6*dr_parity_2[i] + b7*dr_parity_3[i] + b8*dr_sexm0f1[i] +
b9*dr_mrace_bla[i] + b10*dr_mrace_asi[i] + b11*dr_mrace_his[i] + b12*dr_mrace_oth[i] +
b13*dr_medu_lhs[i] + b14*dr_medu_hs[i] + b15*dr_medu_sc[i] + b16*dr_medu_ad[i] +
b17*dr_ti_1[i] + b18*dr_ti_2[i] + b19*dr_ti_4[i] + b22*DR_ACTSMK_PREG[i]

# Measurement model

DR_ASTHMAGEN_PREG[i] ~ dbern(pm[i])

pm[i] <- SN0*(astpregt[i])*(1-y[i]) + (1-SP0)*(1-astpregt[i])*(1-y[i]) +
SN1*(astpregt[i])*(y[i]) + (1-SP1)*(1-astpregt[i])*(y[i])

# Exposure model

astpregt[i] ~ dbern(prop[i])

logit(prop[i]) <- g1 + g2*DR_AGEBIRTH_MX_C31[i] + g5*DR_PSYALL_MX[i] +
g6*dr_parity_2[i] + g7*dr_parity_3[i] + g8*dr_sexm0f1[i] + g9*dr_mrace_bla[i] +
g10*dr_mrace_asi[i] + g11*dr_mrace_his[i] +g12*dr_mrace_oth[i] + g13*dr_medu_lhs[i] +
g14*dr_medu_hs[i] + g15*dr_medu_sc[i] + g16*dr_medu_ad[i] + g17*dr_ti_1[i] +
g18*dr_ti_2[i] + g19*dr_ti_4[i] + g22*DR_ACTSMK_PREG[i]

}

# Calculate odds ratio

OR <- exp(b1)

# Calculate prevalence of exposure among unexposed

r0 <- (p0+SP0–1)/(SN0+SP0–1)

# PRIORS

b0 ~ dnorm(0,1)

b1 ~ dnorm(0,2)

b2 ~ dnorm(0,2)

b5 ~ dnorm(0,2)

b6 ~ dnorm(0,2)

b7 ~ dnorm(0,2)

b8 ~ dnorm(0,1)

b9 ~ dnorm(0,2)

b10 ~ dnorm(0,2)

b11 ~ dnorm(0,2)

b12 ~ dnorm(0,2)

b13 ~ dnorm(0,2)

b14 ~ dnorm(0,2)

b15 ~ dnorm(0,2)

b16 ~ dnorm(0,2)

b17 ~ dnorm(0,2)

b18 ~ dnorm(0,2)

b19 ~ dnorm(0,2)
```

```
b22 ~ dnorm(0,2)
g1 ~ dnorm(0,2)
g2 ~ dnorm(0,2)
g5 ~ dnorm(0,2)
g6 ~ dnorm(0,2)
g7 ~ dnorm(0,2)
g8 ~ dnorm(0,2)
g9 ~ dnorm(0,2)
g10 ~ dnorm(0,2)
g11 ~ dnorm(0,2)
g12 ~ dnorm(0,2)
g13 ~ dnorm(0,2)
g14 ~ dnorm(0,2)
g15 ~ dnorm(0,2)
g16 ~ dnorm(0,2)
g17 ~ dnorm(0,2)
g18 ~ dnorm(0,2)
g19 ~ dnorm(0,2)
g22 ~ dnorm(0,2)
SN0 ~ dbeta(3.6,5.2)
SN1 ~ dbeta(3.6,5.2)
SP0 ~ dbeta(1000, 9.1)
SP1 ~ dbeta(1000, 9.1)
}
```

Priors for Model D_1:

```
data <- list(x0=5876, x1=1247, n0=29359, n1=6706, a.sn0=11.0, b.sn0=38.5,
a.sp0=718, b.sp0=8.2, a.sn1=8.3, b.sn1=17.2, a.sp1=718, b.sp1=8.2,

aa=1, bb=1, mu=−0.34, tau=3.0)
```

Priors for Model D_2:

```
data <- list(x0=5876, x1=1247, n0=29359, n1=6706, a.sn0=5.9, b.sn0=19.5,

a.sp0=46.6, b.sp0=3.4, a.sn1=4.2, b.sn1=8.1, a.sp1=46.6, b.sp1=3.4,

aa=1, bb=1, mu=−0.34, tau=3.0)
```

WinBugs code for Model D_1 and Model D_2:

```
model{

x0 ~ dbin(p0, n0)

x1 ~ dbin(p1, n1)

p0 <- r0*SN0 + (1-r0)*(1-SP0)

p1 <- r1*SN1 + (1-r1)*(1-SP1)

r0 ~ dbeta(aa,bb)

lor ~ dnorm(mu,tau)

SN0 ~ dbeta(a.sn0, b.sn0)

SN1 ~ dbeta(a.sn1, b.sn1)

SP0 ~ dbeta(a.sp0, b.sp0)

SP1 ~ dbeta(a.sp1, b.sp1)

OR <- exp(lor)

r1 <- (OR*r0)/(1-r0+OR*r0)

}
```

References

1. •Burstyn I, Yang Y, Schnatter AR. Effects of non-differential exposure misclassification on false conclusions in hypothesis-generating studies. Int J Environ Res Public Health. 2014;11(10):10951–66. [PubMed: 25337942] This paper demonstrates that exposure misclassification may result in false-positive conclusions.
2. Gustafson P Measurement error and misclassification in statistics and epidemiology : impacts and Bayesian adjustments. Boca Raton, Fla :: Chapman & Hall/CRC; 2004.
3. •Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 1991;134(10):1233–44. [PubMed: 1746532] This study is

the first ilustration in epidemiological literature of differential misclassification arising from non-differential measurement error.

4. Gustafson P, Le ND, Saskin R. Case-Control Analysis with Partial Knowledge of Exposure Misclassification Probabilities. Biometrics. 2001;57:598–609. [PubMed: 11414590]
5. Marshall JR. The use of dual or multiple reports in epidemiologic studies. Stat Med 1989;8(9): 1041–9; discussion 71–3. [PubMed: 2799130]
6. •Goldstein ND, Welles SL, Burstyn I. To Be or Not to Be: Bayesian Correction for Misclassification of Self-reported Sexual Behaviors Among Men Who Have Sex with Men. Epidemiology. 2015;26(5):637–44. [PubMed: 26214333] This study implements Bayesian methods for correction of exposure misclassification to resolve heterogenity of effect estimates.
7. •Luta G, Ford MB, Bondy M, Shields PG, Stamey JD. Bayesian sensitivity analysis methods to evaluate bias due to misclassification and missing data using informative priors and external validation data. Cancer Epidemiol 2013;37(2):121–6. [PubMed: 23290580] This study demonstrates exposure misclassifcation correction using Bayesian modeling in a manner that is accessible to epidemiologists in general.
8. Singer AB, Windham GC, Croen LA, Daniels JL, Lee BK, Qian Y, et al. Maternal Exposure to Occupational Asthmagens During Pregnancy and Autism Spectrum Disorder in the Study to Explore Early Development. J Autism Dev Disord. 2016;46(11):3458–68. [PubMed: 27511194]
9. Singer AB, Burstyn I, Thygesen M, Mortensen PB, Fallin MD, Schendel DE. Parental exposures to occupational asthmagens and risk of autism spectrum disorder in a Danish population-based case-control study. Environ Health. 2017;16(1):31. [PubMed: 28359263]
10. Kennedy SM, Le Moual N, Choudat D, Kauffmann F. Development of an asthma specific job exposure matrix and its application in the epidemiological study of genetics and environment in asthma (EGEA). Occup Environ Med 2000;57(9):635–41. [PubMed: 10935945]
11. Schendel DE, Diguiseppi C, Croen LA, Fallin MD, Reed PL, Schieve LA, et al. The Study to Explore Early Development (SEED): A Multisite Epidemiologic Study of Autism by the Centers for Autism and Developmental Disabilities Research and Epidemiology (CADDRE) Network J Autism Dev Disord. 2012.
12. International Labor Organization. International standard classifications of occupations (ISCO-88), 1988 ed. Geneva: International Labor Organization; 1991.
13. Mors O, Perto GP, Mortensen PB. The Danish Psychiatric Central Research Register. Scandinavian journal of public health. 2011;39(7 Suppl):54–7. [PubMed: 21775352]
14. Petersson F, Baadsgaard M, Thygesen LC. Danish registers on personal labour market affiliation. Scandinavian journal of public health. 2011;39(7 Suppl):95–8. [PubMed: 21775363]
15. Liu J, Gustafson P, Cherry N, Burstyn I. 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. [PubMed: 19691019]
16. Beach J, Burstyn I, Cherry N. Estimating the extent and distribution of new-onset adult asthma in British Columbia using frequentist and Bayesian approaches. The Annals of occupational hygiene. 2012;56(6):719–27. [PubMed: 22378842]
17. Thomas A, O'Hara BO, Ligges U, Sturtz S. Making BUGS Open. R News. 2006;6:12–7.
18. Plummer M A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. 2003.
19. Stan Development Team. Stan: A C++ Library for Probability and Sampling, Version 2.5.0. 2014.
20. Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing. 2000;10(4):325–37.
21. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A Package for Running WinBUGS from R. Journal of Statistical Software. 2005;12(3):1–16.

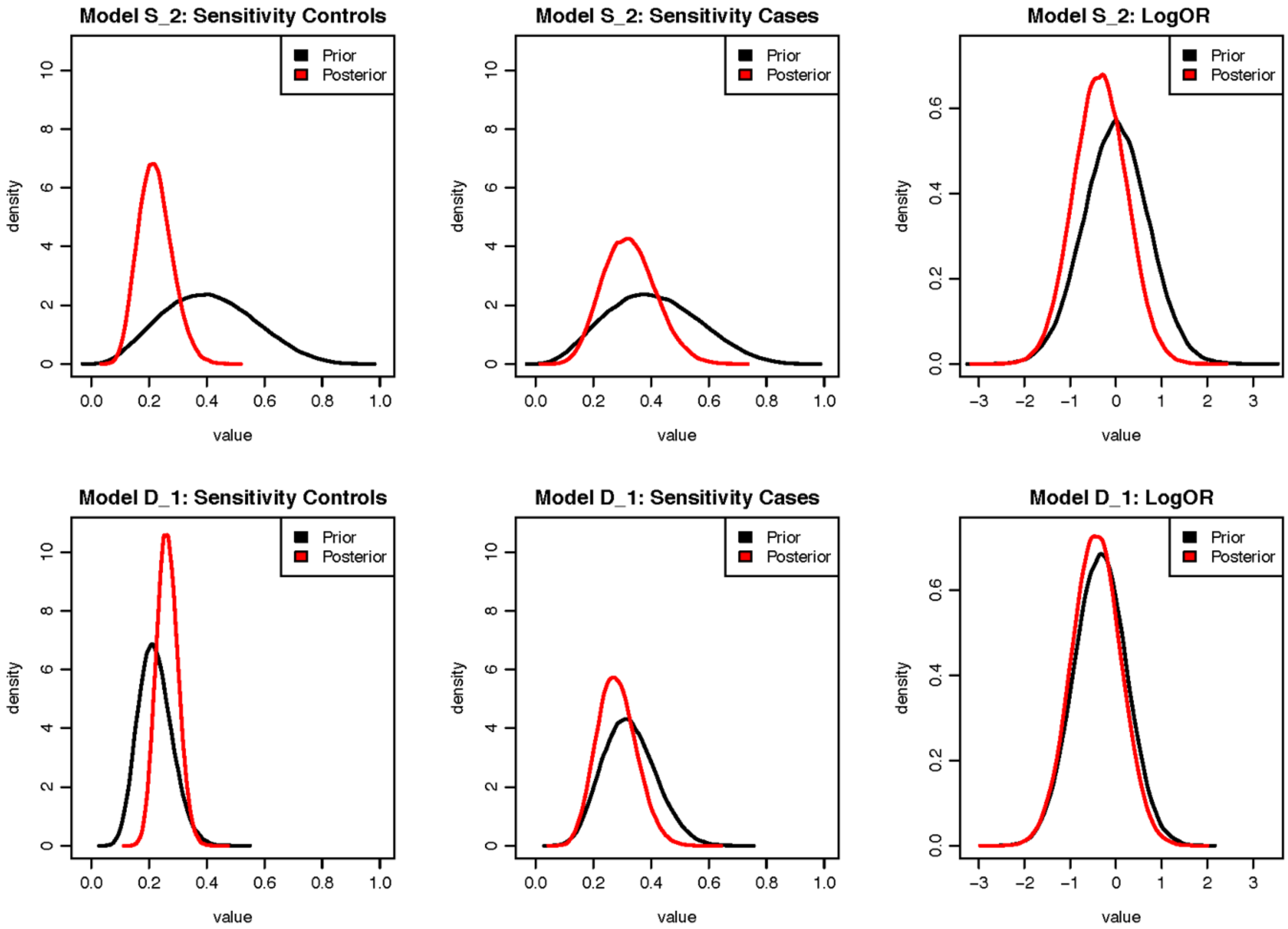

Figure 1:

Illustration of prior and posterior distributions for analysis of SEED (Model S_2) and study nested in Denmark (Model D_1); posterior distributions from Model S_2 were used to set priors for Model D_1

Table 1:

Posterior distributions, medians and 95% Credible Intervals after allowing for differential exposure misclassification and learning from study to study about association between exposure to asthmagens and autism spectrum disorders.

Study/country	SEED/USA		Denmark	
Prior on Odds Ratio	**Flat**		**Informed by posterior of SEED analysis S_2**	
Assumption about exposure classification (Model)	**Near-perfect exposure classification (S_1)**	**Realistic exposure misclassification (S_2)**	**Exposure misclassification the same as in SEED S_2 (D_1)[2]**	**Uncertain whether exposure misclassification is the same as in SEED S_2 (D_2)[3]**
Odds Ratio	1.37 (0.96 – 1.96) [1]	0.71 (0.23 – 2.42)[1]	0.64 (0.23 – 1.94)	0.68 (0.23 – 1.97)
Sensitivity				
Controls	0.999 (0.996 – 1.00)	0.21 (0.15 – 0.36)	0.26 (0.21 – 0.35)	0.26 (0.20 – 0.41)
Cases	0.999 (0.996 – 1.00)	0.31 (0.21 – 0.54)	0.27 (0.19 – 0.44)	0.27 (0.19 – 0.52)
Specificity				
Controls	0.999 (0.996 – 1.00)	0.99 (0.98 – 1.00)	0.99 (0.98 – 1.00)	0.93 (0.83 – 0.98)
Cases	0.999 (0.996 – 1.00)	0.99 (0.98 – 1.00)	0.99 (0.98 – 1.00)	0.93 (0.85 – 0.98)

[1]: adjusted for covariates as described in text for SEED

[2]: Priors on Sn, Sp and OR based posterior distribution from SEED

[3]: Priors on Sn based on inflating variance of posterior distributions from SEED (see text); prior on OR based on posterior distribution from SEED as in model D_1