THE STATA JOURNAL

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the web page

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- Science Citation Index Expanded (also known as SciSearch®)
- CompuMath Citation Index®

A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies

Nicola Orsini
Division of Nutritional Epidemiology
Institute of Environmental Medicine
Karolinska Institutet
Stockholm, Sweden
nicola.orsini@ki.se

Rino Bellocco
Department of Statistics
University of Milano-Bicocca
Milano, Italy

Matteo Bottai
Department of Epidemiology and Biostatistics
Arnold School of Public Health
University of South Carolina
Columbia, SC

Alicja Wolk
Division of Nutritional Epidemiology
Institute of Environmental Medicine
Karolinska Institutet
Stockholm, Sweden

Sander Greenland
Departments of Epidemiology and Statistics
University of California, Los Angeles
Los Angeles, CA

Abstract. Classification errors, selection bias, and uncontrolled confounders are likely to be present in most epidemiologic studies, but the uncertainty introduced by these types of biases is seldom quantified. The authors present a simple yet easy-to-use Stata command to adjust the relative risk for exposure misclassification, selection bias, and an unmeasured confounder. This command implements both deterministic and probabilistic sensitivity analysis. It allows the user to specify a variety of probability distributions for the bias parameters, which are used to simulate distributions for the bias-adjusted exposure–disease relative risk. We illustrate the command by applying it to a case–control study of occupational resin exposure and lung-cancer deaths. By using plausible probability distributions for the bias parameters, investigators can report results that incorporate their uncertainties regarding systematic errors and thus avoid overstating their certainty about the effect under study. These results can supplement conventional results and can help pinpoint major sources of conflict in study interpretations.

Keywords: st0138, episens, episensi, sensitivity analysis, unmeasured confounder, misclassification, bias, epidemiology

1 Introduction

Conventional statistical methods to estimate exposure–disease associations from observational studies are based on several assumptions, such as no measurement error and no selection bias (i.e., selection, participation, and retention of subjects are purely

random). If the associations are interpreted as causal effects, another assumption of random-exposure assignment within levels of controlled covariates is also implicitly made. When such assumptions are not met, tests and estimates for the association between exposure and disease are likely to be biased and may fail to capture most of the uncertainty about the estimated parameter (Greenland 2005).

There are many proposed methods to adjust uncertainty assessments for unmeasured sources of bias or systematic error (Chu et al. 2006; Eddy, Hasselblad, and Shachter 1992; Fox, Lash, and Greenland 2005; Greenland 2001; Greenland 2003b; Greenland 2005; Greenland and Lash 2008; Hoffman and Hammonds 1994; Lash and Fink 2003; Phillips 2003; and Steenland and Greenland 2004). Nonetheless, few published papers in epidemiologic journals use quantitative methods to investigate the role of potential bias in the observed findings (Jurek et al. 2006). To facilitate the use of both deterministic and probabilistic sensitivity analysis, we present a flexible and easy-to-use tool to assess the uncertainty of exposure–disease associations due to misclassification of the exposure, selection bias, and unmeasured confounding. The proposed tool is implemented as a one-line Stata command. Here we illustrate the use of the tool by analyzing a published medical study reporting a positive association between occupational resin exposure and lung-cancer deaths in a case–control study.

2 Methods

We consider the simplest situation in which there are only two factors: the disease and the exposure. Each factor is considered as being either present or absent, and so the data can be summarized in a 2×2 table. The term *relative risk* (RR) will be used as a generic term for the risk ratio (ratio of proportions getting disease), rate ratio (ratio of person–time incidence rates), and odds ratio (ratio of odds, most often used in case–control data). The formulas implemented for correction of the observed RR due to misclassification of the exposure, selection bias, and a binary unmeasured or uncontrolled confounder are described in detail elsewhere (Greenland 1996; Greenland and Lash 2008).

Deterministic (ordinary or classical) sensitivity analysis provides an external adjustment of the observed RR upon specification of a list of hypothetical values for the bias parameters. The main limitation of this approach is related to the lack of explicit accounting for uncertainty about the bias parameters (Greenland 1998). To account for this uncertainty, probabilistic sensitivity analysis allows the user to specify a variety of probability densities for the bias parameters and use these densities to obtain simulation limits for the bias-adjusted exposure–disease relative risk. The accompanying Stata tool allows the user to specify a variety of probability density functions for the bias parameters (table 1). Probabilistic sensitivity analysis through Monte Carlo (random-number–based) simulations involves two iterated steps: 1) draw a random sample (one set of bias parameters) from the user-specified probability density functions of the bias parameters, and 2) back-calculate a bias-adjusted ("corrected) RR from the drawn parameters. These two steps are repeated several times to obtain a distribution of bias-adjusted RR.

Table 1. Probability distributions for the bias parameters used in adjustment for misclassification of the exposure, selection bias, and unmeasured or uncontrolled confounding

Type of systematic error and bias parameters	Description	Probability density functions (pdf)
Misclassification of the exposure		
`dseca()`	Sensitivity cases	`constant(`#`)`
`dspca()`	Specificity cases	`uniform(`a b`)`
`dsenc()`	Sensitivity noncases	`triangular(`a b c`)`
`dspnc()`	Specificity noncases	`trapezoidal(`a b c d`)`
		`logit-logistic(`m s $\lceil lb\ ub \rceil$`)`
		`logit-normal(`m s $\lceil lb\ ub \rceil$`)`
Selection bias		
`dpscex()`	Pr selection cases exposed	`constant(`#`)`
`dpscun()`	Pr selection cases unexposed	`uniform(`a b`)`
`dpsnex()`	Pr selection noncases exposed	`triangular(`a b c`)`
`dpsnun()`	Pr selection noncases unexposed	`trapezoidal(`a b c d`)`
		`logit-logistic(`m s $\lceil lb\ ub \rceil$`)`
		`logit-normal(`m s $\lceil lb\ ub \rceil$`)`
`dsbfactor()`	Selection bias factor	`constant(`#`)`
		`log-logistic(`m s`)`
		`log-normal(`m s`)`
Unmeasured confounding		
`dpexp()`	Pr confounder exposed	`constant(`#`)`
`dpunexp()`	Pr confounder unexposed	`uniform(`a b`)`
		`triangular(`a b c`)`
		`trapezoidal(`a b c d`)`
		`logit-logistic(`m s $\lceil lb\ ub \rceil$`)`
		`logit-normal(`m s $\lceil lb\ ub \rceil$`)`
`drrcd()`	RR counfounder–disease	`constant(`#`)`
`dorce()`	OR confounder–exposure	`log-logistic(`m s`)`
		`log-normal(`m s`)`

The results of the simulation can be summarized by descriptions of the distribution of bias-adjusted RR. For example, the median (50th percentile) and the 2.5th and 97.5th percentiles can serve as analogues of the point and interval estimate for the bias-adjusted RR. To take into account uncertainty due to random error, we subtract from the distribution of the bias-adjusted ln(RR) a random draw from a normal distribution with zero mean and standard deviation equal to the standard error of the conventional ln(RR) estimate.

In a situation where more than one systematic error occurred during the study (uncontrolled confounding, selection bias, misclassification of the exposure) and these errors can be treated as independent, we can perform a multiple probabilistic bias analysis with adjustment made in the reverse order of their occurrence. Suppose that the order of events is as follows: confounded associations arise in the population used as the source of study subjects; subjects are selected; and finally, subjects are classified by exposure (with no misclassification of disease). Then, at each iteration of the simulation, adjustment of the observed exposure–disease RR follows this order: adjustment for misclassification of the exposure, then adjustment for selection bias, and finally, adjustment for uncontrolled confounders.

The rest of the article is organized as follows: section 3 presents the syntax of the command `episens` and its immediate form `episensi`; section 4 provides some examples in which the command is applied to published data; and section 5 contains a discussion of strengths and limitations of sensitivity analysis.

3   The episens command

3.1   Syntax

`episens` *var_case* *var_exposed* $\left[\,var\_time\,\right]$ $\left[\,if\,\right]$ $\left[\,in\,\right]$ $\left[\,weight\,\right]$ $\left[\,,\ options\,\right]$

`episensi` #a #b #c #d $\left[\,,\ options\,\right]$

3.2   Description

`episens` performs deterministic and probabilistic sensitivity analysis of the exposure–disease relative risk for misclassification of the exposure, selection bias, and unmeasured or uncontrolled confounding.

`episensi` is the immediate form of `episens`.

3.3   Options

The probability distribution function (pdf) of each bias parameter is specified as an argument of an option. The list of probability distributions is presented in *pdf for the bias parameter (pdf_options)* below, as well as in table 1 organized by type of systematic error.

Misclassification of the exposure

<u>dsec</u>a(*pdf_options*)	define the sensitivity among the cases
<u>dspc</u>a(*pdf_options*)	define the specificity among the cases
<u>dsen</u>c(*pdf_options*)	define the sensitivity among the noncases
<u>dspn</u>c(*pdf_options*)	define the specificity among the noncases
<u>corrsens</u>(#)	set the correlation between case and noncase sensitivities to #
<u>corrspec</u>(#)	set the correlation between case and noncase specificities to #

Selection bias

<u>dpscex</u>(*pdf_options*)	define the selection probability among cases exposed
<u>dpscun</u>(*pdf_options*)	define the selection probability among cases unexposed
<u>dpsnex</u>(*pdf_options*)	define the selection probability among noncases exposed
<u>dpsnun</u>(*pdf_options*)	define the selection probability among noncases unexposed
<u>dsbf</u>actor(*pdf_options*)	define the selection-bias factor

Uncontrolled confounder

<u>dpexp</u>(*pdf_options*)	define the prevalence of the confounder among the exposed
<u>dpunexp</u>(*pdf_options*)	define the prevalence of the confounder among the unexposed
<u>drrc</u>d(*pdf_options*)	define the confounder-disease relative risk
<u>dorc</u>e(*pdf_options*)	define the confounder-exposure odds ratio
<u>corrprev</u>(#)	set the correlation between exposure-specific confounder prevalences to #

pdf for the bias parameter (*pdf_options***)**

<u>constant</u>(#)	constant value equal to #
<u>uniform</u>(a b)	uniform between min $= a$ and max $= b$
<u>triangular</u>(a b c)	triangular with min $= a$, mode $= b$, and max $= c$
<u>trapezoidal</u>(a b c d)	trapezoidal with min $= a$, modes between b and c, and max $= d$
<u>logit-logistic</u>(m s $\begin{bmatrix} lb & ub \end{bmatrix}$)	logit–logistic with mean $= m$ and scale $= s$, shifted between $\begin{bmatrix} lb & ub \end{bmatrix}$
<u>logit-normal</u>(m s $\begin{bmatrix} lb & ub \end{bmatrix}$)	logit-normal with mean $= m$ and scale $= s$, shifted between $\begin{bmatrix} lb & ub \end{bmatrix}$
<u>log-logistic</u>(m s)	loglogistic with mean $= m$ and scale $= s$
<u>log-normal</u>(m s)	lognormal with mean $= m$ and scale $= s$

Simulations

<u>reps</u>(#)	specify the number of replications to be performed
<u>nodots</u>	suppress the replication dots
<u>seed</u>(#)	set the random-number seed to #
<u>ndraw</u>(#)	number of observations drawn at each replication
<u>saving</u>(*filename*)	save results to *filename*
<u>grprior</u>	histogram of the priors
<u>grarrsys</u>	histogram of the adjusted relative risk (systematic error)
<u>grarrtot</u>	histogram of the adjusted relative risk (systematic error plus random error)

Study design, format, combined analysis

<u>study</u>(cc | cs | ir)	specify the type of study
<u>f</u>ormat(%*fmt*)	set the display format for numbers
<u>combined</u>	specify combined analyses of multiple biases

3.4   Saved results

`episens` saves the following in `r()`:

Scalars

Deterministic sensitivity analysis

`r(bias_mie)`	percentage of bias due to misclassification of the exposure
`r(rrdx_mie)`	exposure–disease relative risk adjusted for misclassification of the exposure
`r(bias_sel)`	percentage of bias due to selection bias
`r(rrdx_sel)`	exposure–disease relative risk adjusted for selection bias
`r(bias_unc)`	percentage of bias due to unmeasured confounding
`r(rrdx_unc)`	exposure–disease relative risk adjusted for unmeasured confounding

Probabilistic sensitivity analysis

`r(rrdx_mie_pm)`	median of the distribution of exposure–disease relative risks adjusted for misclassification of the exposure
`r(rrdx_mie_plb)`	2.5th percentile of the distribution of exposure–disease risks adjusted for misclassification of the exposure
`r(rrdx_mie_pub)`	97.5th percentile of the distribution of exposure–disease risks adjusted for misclassification of the exposure
`r(rrdx_sel_pm)`	median of the distribution of exposure–disease relative risks adjusted for selection bias
`r(rrdx_sel_plb)`	2.5th percentile of the distribution of exposure–disease risks adjusted for selection bias
`r(rrdx_sel_pub)`	97.5th percentile of the distribution of exposure–disease risks adjusted for selection bias
`r(rrdx_unc_pm)`	median of the distribution of exposure–disease relative risks adjusted for unmeasured confounding
`r(rrdx_unc_plb)`	2.5th percentile of the distribution of exposure–disease risks adjusted for unmeasured confounding
`r(rrdx_unc_pub)`	97.5th percentile of the distribution of exposure–disease risks adjusted for unmeasured confounding
`r(rrdx_all_pm)`	median of the distribution of exposure–disease relative risks adjusted for all user-specified biases
`r(rrdx_all_plb)`	2.5th percentile of the distribution of exposure–disease risks adjusted for all user-specified biases
`r(rrdx_all_pub)`	97.5th percentile of the distribution of exposure–disease risks adjusted for all user-specified biases

4   Examples

4.1   Deterministic sensitivity analysis

To illustrate how to perform a sensitivity analysis using the command `episens`, we used the crude data from a case–control study comparing cases of lung-cancer deaths with controls based on occupational exposure to resins (Greenland et al. 1994).

```
. cci 45 94 257 945, woolf

                                                            Proportion
                     |   Exposed    Unexposed  |     Total     Exposed
        -------------+------------------------ +----------------------
                Cases|        45           94  |       139      0.3237
             Controls|       257          945  |      1202      0.2138
        -------------+------------------------ +----------------------
                Total|       302         1039  |      1341      0.2252

                     |   Point estimate        |   [95% Conf. Interval]
        -------------+------------------------ +----------------------
            Odds ratio|         1.760286        |   1.202457    2.576898 (Woolf)
        Attr. frac. ex.|          .4319106        |    .1683693     .6119365 (Woolf)
        Attr. frac. pop|          .1398272        |
        -------------+------------------------ +----------------------
                            chi2(1) =      8.63  Pr>chi2 = 0.0033
```

The authors found a positive association between occupational exposure and lung-cancer deaths (OR=1.76, 95% CI, 1.20–2.58). Further adjustment for age or year did not substantially change this association. Nonetheless, the measured exposure to resins must be misclassified to some extent.

Exposure misclassification

The sensitivities and specificities of classification among the cases and noncases would allow us to adjust the observed data for classification error (Greenland 1996; Greenland and Lash 2008). We can perform a deterministic sensitivity analysis assuming nondifferential misclassification of the exposure and assigning a specific (fixed) value to the sensitivity and specificity among cases and noncases, say, 0.9.

```
. episensi 45 94 257 945, st(cc) dseca(c(.9)) dspca(c(.9)) dsenc(c(.9))
> dspnc(c(.9))
Se|Cases   : Constant(.9)
Sp|Cases   : Constant(.9)
Se|No-Cases: Constant(.9)
Sp|No-Cases: Constant(.9)
Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]
Deterministic sensitivity analysis for misclassification of the exposure
   External adjusted Odds Ratio = 2.34
   Percent bias = -25%
```

The odds ratio (OR adjusted for misclassification of the exposure is 2.34, with a percentage of bias of $(1.76 - 2.34)/2.34 * 100 = -25\%$. However, under the assumption that the sensitivity among the cases (0.9) is higher than the sensitivity among the

noncases (0.8) with specificities at 0.8, the OR adjusted for misclassification of the exposure would be 9.11.

```
. episensi 45 94 257 945, st(cc) dseca(c(.9)) dspca(c(.8)) dsenc(c(.8))
> dspnc(c(.8))
Se|Cases   : Constant(.9)
Sp|Cases   : Constant(.8)
Se|No-Cases: Constant(.8)
Sp|No-Cases: Constant(.8)
Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]
Deterministic sensitivity analysis for misclassification of the exposure
   External adjusted Odds Ratio = 9.11
   Percent bias = -81%
```

One can repeat this procedure for various likely combinations of sensitivities and specificities among the cases and noncases and present the adjusted ORs in a table (Greenland 1996; Greenland and Lash 2008).

Selection bias

Because of lack of adequate job records for exposure reconstruction, some data available from this study indicate that the probabilities of selecting a case and a noncase are 0.7 and 0.6, respectively.

If selection is associated with both exposure to resins and lung-cancer death, considerable selection bias could result. The selection-bias factor dsbfactor is given by the exposed versus unexposed selection probabilities comparing cases (dpscex/dpscun) and noncases (dpsnex/dpsnun). If the selection probabilities among cases and noncases do not differ across exposure status, there is no bias [dsbfactor = (dpscex/dpscun)/(dpsnex/dpsnun) = 1].

```
. episensi 45 94 257 945, st(cc) dpscex(c(.7)) dpscun(c(.7)) dpsnex(c(.6))
> dpsnun(c(.6))
Pr Case Selection Exposed: Constant(.7)
Pr Case Selection No-Exposed: Constant(.7)
Pr No-Case Selection Exposed: Constant(.6)
Pr No-Case Selection No-Exposed: Constant(.6)
Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]
Deterministic sensitivity analysis for selection bias
   External adjusted Odds Ratio = 1.76
   Percent bias =   0%
```

However, if the probabilities of selecting a case and a noncase are different with respect to the exposure status, the selection-bias factor will be greater than 1 if (dpsecx/dpscun) > (dpsnex/dpsnun) and lower than 1 if (dpsecx/dpscun) < (dpsnex/dpsnun). For instance, lets suppose that the probability of selecting a case exposed is 0.9, a case unexposed is 0.5, a noncase exposed is 0.5, and a noncase unexposed is 0.7. The selection-bias factor is equal to $(.9/.5)/(.5/.7) = 2.5$.

```
. episensi 45 94 257 945, st(cc) dpscex(c(.9)) dpscun(c(.5)) dpsnex(c(.5))
> dpsnun(c(.7))

Pr Case Selection Exposed: Constant(.9)
Pr Case Selection No-Exposed: Constant(.5)
Pr No-Case Selection Exposed: Constant(.5)
Pr No-Case Selection No-Exposed: Constant(.7)

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for selection bias
   External adjusted Odds Ratio = 0.70
   Percent bias = 152%
```

The selection-bias adjusted OR is equal to $1.76/2.5 = 0.7$. In an opposite scenario, the probability of selecting a case exposed is 0.5, a case unexposed is 0.9, a noncase exposed is 0.7, and a noncase unexposed is 0.5. The selection-bias factor is equal to $(.5/.9)/(.7/.5) = 0.4$.

```
. episensi  45 94 257 945, st(cc) dpscex(c(.5)) dpscun(c(.9)) dpsnex(c(.7))
> dpsnun(c(.5))

Pr Case Selection Exposed: Constant(.5)
Pr Case Selection No-Exposed: Constant(.9)
Pr No-Case Selection Exposed: Constant(.7)
Pr No-Case Selection No-Exposed: Constant(.5)

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for selection bias
   External adjusted Odds Ratio = 4.44
   Percent bias = -60%
```

The selection-bias adjusted OR is equal to $1.76/0.4 = 4.4$. These two extreme scenarios, however, do not take into account that there is no reason to expect big differences comparing the case and noncase selection probabilities with respect to the exposure; that is, dpscex should be similar to dpscun, and dpsnex should be similar to dpsnun.

Uncontrolled confounders

In the case–control study of occupational exposure to resins and lung-cancer mortality, the authors had no data on smoking. Therefore, we want to quantify the potential bias introduced by ignoring smoking in the published analysis. To back-calculate the smoking adjusted OR, we assume that the RR relating smoking to lung-cancer death is 5, and the smoking prevalences among the resins exposed and unexposed are 0.7 and 0.5, respectively.

```
. episensi 45 94 257 945, dpexp(c(.7)) dpunexp(c(.5)) drrcd(c(5))

Pr(c=1|e=1): Constant(.7)
Pr(c=1|e=0): Constant(.5)
RR_cd      : Constant(5)

Observed Odds Ratio [95% Conf. Interval]= 1.76 [1.20, 2.58]

Deterministic sensitivity analysis for unmeasured confounding
   External adjusted Odds Ratio = 1.39
   Percent bias =  27%
```

The resin lung-cancer death OR adjusted for smoking is 1.39, which is lower than the observed OR because we assumed positive associations between the confounder and the outcome ($5 > 1$) as well as between the confounder and the exposure ($0.7 > 0.5$). For sensitivity analysis, one can repeat the above command using other plausible values for the resins-specific smoking prevalences and the smoking lung-cancer OR (Greenland 1996; Greenland and Lash 2008).

4.2 Probabilistic sensitivity analysis

The main limitation of deterministic sensitivity analyses is that they treat the bias parameters as if they were known or as if they can assume only certain fixed values. It also fails to discriminate among the different scenarios in terms of their likelihood, and it is not straightforward to summarize all the bias-adjusted RR calculated under a variety of possible values for the bias parameters. Therefore, we next assume that we can specify prior probability distributions for the bias parameters that capture our uncertainty about those parameters and then use these distributions in a probabilistic sensitivity analysis.

Exposure misclassification

We first assume nondifferential misclassification of the exposure with probability density functions for sensitivities and specificities among cases and noncases equal to trapezoidal distributions with a minimum of 0.75 and a maximum of 1, and an interval of equally probable values between 0.85 and 0.95.

A technical issue is that the formulas used to back-calculate the relative risk can yield negative adjusted counts, which are impossible. To avoid negative adjusted counts, the prior distributions for sensitivity and specificity must be bounded by `dsenc()` $\geq$ (number of noncases classified exposed / total number of noncases) and `dspnc()` $\geq$ (number of noncases classified unexposed / total number of noncases) among noncases and by `dseca()` $\geq$ (number of cases classified exposed / total number of cases) and `dspca()` $\geq$ (number of cases classified unexposed / total number cases) among cases. The command `episens` automatically discards draws of sensitivities and specificities from user-specified distributions falling into the region of negative adjustment. It is the user's decision whether to check that the resulting truncated distribution still appears to be reasonable.

Here negative adjustments would occur whenever `dsenc()` $< (257/1202 = 0.214)$ and `dspnc()` $< (945/1202 = 0.786)$ among noncases, and `dseca()` $< (45/139 = 0.324)$ and `dspca()` $< (94/139 = 0.676)$ among cases. Among these four bounds, however, only one is of interest (`dspnc()` < 0.786) because we specified trapezoidal distributions between 0.75 and 1.

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots
> dseca(trap(.75 .85 .95 1)) dspca(trap(.75 .85 .95 1))
> dsenc(trap(.75 .85 .95 1)) dspnc(trap(.75 .85 .95 1)) seed(123)

Se|Cases   : Trapezoidal(.75,.85,.95,1)
Sp|Cases   : Trapezoidal(.75,.85,.95,1)
Se|No-Cases: Trapezoidal(.75,.85,.95,1)
Sp|No-Cases: Trapezoidal(.75,.85,.95,1)

Probabilistic sensitivity analysis for misclassification of the exposure

                              Percentiles          Ratio
                            2.5     50       97.5    97.5/2.5
                            ----------------------------------------
Conventional                1.20    1.76     2.58    2.14
Systematic error            1.87    2.47     14.71   7.86
Systematic and random error 1.49    2.57     15.00   10.07
```

The 2.5th and 97.5th percentiles of the simulated distribution of bias-adjusted OR are 1.9 and 14.7, and the median estimate is 2.5. Including random error in the distribution of bias-adjusted OR, the 2.5th and 97.5th percentiles of the simulated distribution become 1.5 and 15. Unsurprisingly, given the high uncertainty about the bias parameters, the ratio of the bias-adjusted simulation limits (15/1.5) is about 5 times the ratio of the conventional limits (2.6/1.2). The option `grprior` helps to visualize the assumed prior probability distributions by showing histograms of the draws of the bias parameters from those distributions (figure 1). The specificity distribution among noncases is truncated at 0.786 because the command `episens` discards draws leading to negative adjustments. The option `saving(`*filename*`)` can be useful to inspect the sampled distributions of the bias parameters and the bias-adjusted odds ratios and to control various aspects of the graphs.

We can allow for differential misclassification by drawing the sensitivities and specificities from different trapezoidal distributions for cases and controls. Because the sensitivities/specificities among the cases are not independent of the sensitivities/specificities among the noncases, we should specify a high correlation between sensitivities and specificities respectively, say, 0.8. The options `corrsens()` and `corrspec()` can help to control the degree of differentiality. Assuming the same priors for cases and noncases, a correlation of 1 means no difference between sensitivities/specificities among cases and noncases (nondifferential misclassification).

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dseca(trap(.75 .85 .95 1))
> dspca(trap(.75 .85 .95 1)) dsenc(trap(.7 .8 .9 .95))
> dspnc(trap(.7 .8 .9 .95)) corrsens(.8) corrspec(.8) seed(123) grprior
Se|Cases    : Trapezoidal(.75,.85,.95,1)
Sp|Cases    : Trapezoidal(.75,.85,.95,1)
Se|No-Cases: Trapezoidal(.7,.8,.9,.95)
Sp|No-Cases: Trapezoidal(.7,.8,.9,.95)
Corr Se|Cases and Se|No-Cases : .8
Corr Sp|Cases and Sp|No-Cases : .8
Probabilistic sensitivity analysis for misclassification of the exposure

                                     Percentiles        Ratio
                             2.5     50      97.5       97.5/2.5
                            ----------------------------------------
Conventional                 1.20    1.76    2.58       2.14
Systematic error             1.81    3.48    48.19      26.57
Systematic and random error  1.61    3.60    48.92      30.47
```
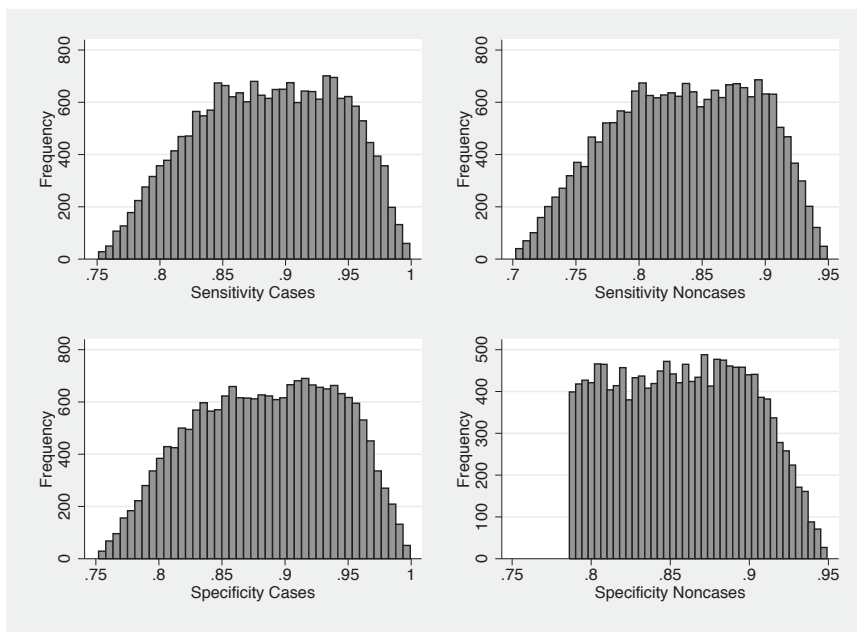

Figure 1. Histograms of 20,000 draws from trapezoidal prior distributions ($a = 0.75$, $b = 0.85$, $c = 0.95$, $d = 1$) for the sensitivity and specificity among cases and noncases.

The 95% simulation limits including systematic and random error were 1.6 and 49, with a median estimate of 3.6.

Selection bias

Although *Selection bias* of section 4.1 shows how sensitive the resins lung-cancer death OR is to different scenarios of selection bias, these scenarios are of no help because only a small association (if any) between lack of records and lung-cancer death is expected (`dsbfactor()` = 1). Instead of assigning a distribution to each selection probability (`dpscex()`, `dpscun()`, `dpsnex()`, `dpsnun()`), we can directly assign a prior distribution to the selection-bias factor (figure 2). Particularly, we assume a lognormal distribution with mean 0 and standard deviation 0.21, which yields 95% prior probability of the bias factor falling between $\exp(0 - 1.96 * 0.21) = 0.7$ and $\exp(0 + 1.96 * 0.21) = 1.5$.

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dsbfactor(log-n(0 0.21))
> seed(123) grprior

selection bias factor: Log-Normal(0.00,0.21)

Probabilistic sensitivity analysis for selection bias
                                 Percentiles          Ratio
                          2.5      50      97.5      97.5/2.5
                        ----------------------------------------
Conventional             1.20     1.76     2.58      2.14
Systematic error         1.16     1.76     2.66      2.29
Systematic and random error 1.01  1.75     3.08      3.05
```
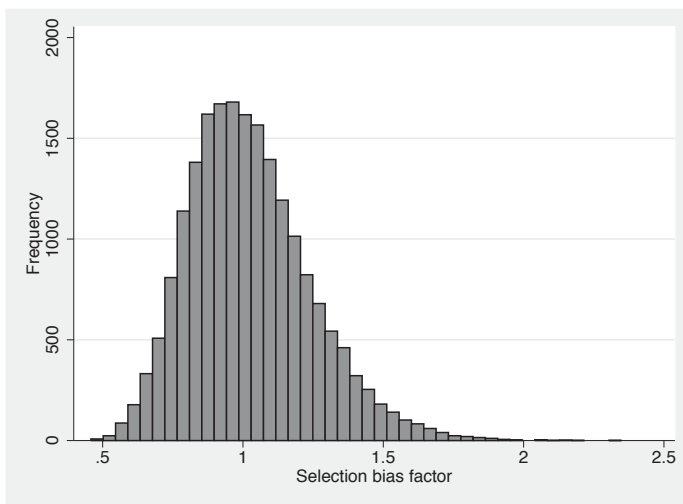

Figure 2. Histogram of 20,000 draws from a lognormal distribution for the selection-bias factor ($m = 0, s = 0.21$).

As expected, the median estimate of the selection-bias adjusted OR 1.75 is not practically different from the conventional 1.76, but the ratio of 95% simulation limits including systematic and random error (3.05) is 43% higher than the conventional one (2.14).

Uncontrolled confounder

As a starting example, we specify two uniform independent distributions for the smoking prevalences among exposed and unexposed between 0.4 and 0.7. We also independently specify a prior probability distribution for the smoking lung-cancer mortality RR that is lognormal with 95% confidence limits of $\ln(5)$ and $\ln(15)$. These limits imply that the mean of this prior RR distribution is $\{\ln(15)+\ln(5)\}/2 = 2.159$ with standard deviation $\{\ln(15)-\ln(5)\}/(2*1.96) = 0.280$. Figure 3 shows the draws from these prior probability distributions for the bias parameters (option `grprior`).

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dpexp(uni(.4 .7))
> dpunexp(uni(.4 .7)) drrcd(log-n(2.159 .280)) seed(123)
> grarrtot grprior
Pr(c=1|e=1): Uniform(.4,.7)
Pr(c=1|e=0): Uniform(.4,.7)
RR_cd      : Log-Normal(2.16,0.28)
Probabilistic sensitivity analysis for unmeasured confounding
```

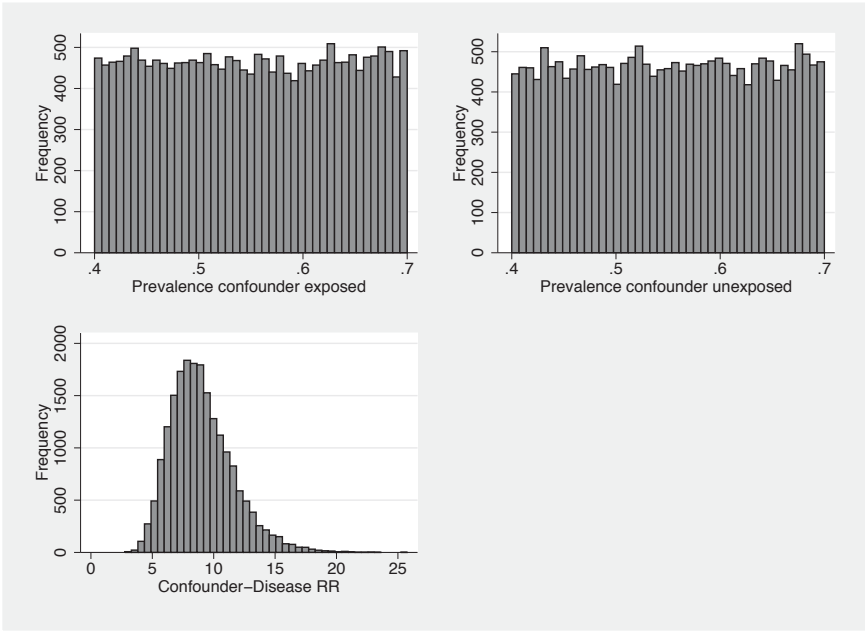|	2.5	Percentiles 50	97.5	Ratio 97.5/2.5
Conventional	1.20	1.76	2.58	2.14
Systematic error	1.25	1.76	2.49	2.00
Systematic and random error	1.05	1.76	2.96	2.83

Figure 3. Histograms of 20,000 draws each from prior distributions for the smoking-exposure specific prevalences and the confounder-disease RR.

From the 20,000 draws for each bias parameter, the median smoking-adjusted resins lung-cancer OR is 1.76 with 2.5th and 97.5th percentiles of 1.05 and 2.96. As expected, the ratio of the smoking-adjusted simulation limits (2.83) is 32% higher than the ratio of the conventional limits (2.14). The distribution of the bias-adjusted OR, including both systematic and random error is shown in figure 4 (option `grarrtot`).
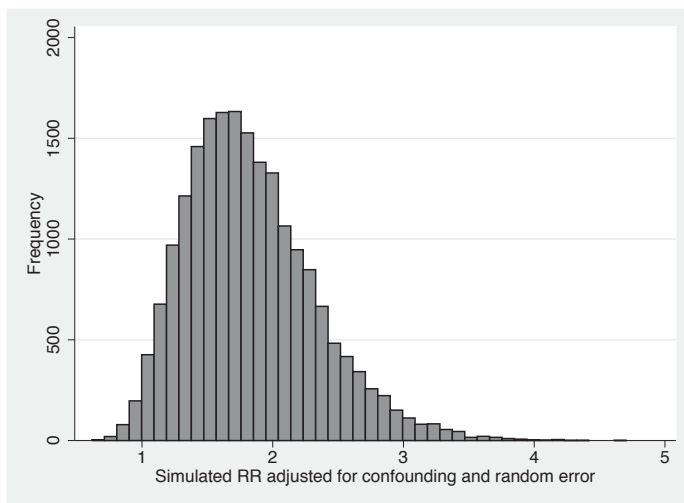

Figure 4. Distribution of the 20,000 smoking-adjusted resins lung-cancer odds ratios derived from the data and the prior distributions in figure 3.

Given that there is no reason to expect great differences in the prevalence of smoking among resins exposed and unexposed, small differences are more likely than large ones. Therefore, it is unrealistic to assume two independent priors for the two prevalences of smoking `dpexp()` and `dpunexp()`. A way to incorporate this consideration in the probabilistic sensitivity analysis is to specify a probability distribution for the confounder-exposure OR (option `dorce()`) instead of the prevalence of the confounder among the exposed (option `dpexp()`). Using independent priors for the confounder-exposure OR and the prevalence of the confounder among the unexposed is more reasonable and easier to specify realistically than using independent priors for the confounder prevalences among the exposed and unexposed.

Suppose that we assign to the confounder-exposure OR a lognormal distribution with mean 0 (that is, `dpexp()` is expected to be similar to `dpunexp()`) and 95% prior limits equal to $\{(1 - .7) * .4\}/\{0.7 * (1 - 0.4)\} = 0.286$ and $\{.7 * (1 - .4)\}/\{(1 - .7) * 0.4\} = 3.5$. These limits are derived calculating a confounder-exposure OR at the extreme values of 0.4 and 0.7 for `dpexp()` and `dpunexp()`. The standard deviation for the lognormal distribution is equal to the standard error calculated from the prior limits $\{\ln(3.5) - \ln(0.286)\}/(1.96 * 2) = 0.639$.

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dpunexp(uni(.4 .7))
> dorce(log-n(0 0.639)) drrcd(log-n(2.159 .280)) seed(123) grprior

Pr(c=1|e=0): Uniform(.4,.7)
RR_cd      : Log-Normal(2.16,0.28)
OR_ce      : Log-Normal(0.00,0.64)

Probabilistic sensitivity analysis for unmeasured confounding
                              Percentiles          Ratio
                         2.5    50      97.5       97.5/2.5
                        ----------------------------------------
Conventional             1.20   1.76    2.58       2.14
Systematic error         1.25   1.76    3.02       2.42
Systematic and random error 1.04 1.79   3.36       3.23
```

Given the priors graphically presented in figure 5, the median bias-adjusted OR is equal to 1.79 with 95% simulation limits 1.04 and 3.36, which have a ratio 3.2 or 14% higher than the earlier ratio of 2.8 based on unrealistic independent priors of the smoking prevalences.
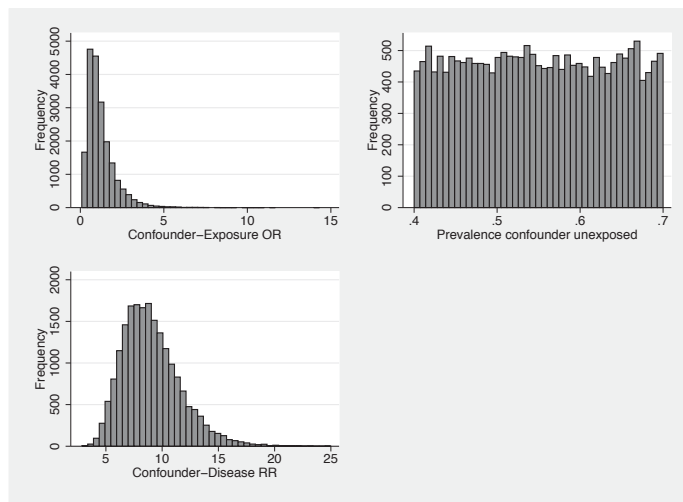

Figure 5. Histograms of 20,000 draws each from prior distributions for the confounder-exposure odds ratio, the prevalence of smoking among the unexposed to resins, and the confounder-disease RR.

4.3 Combined analysis of biases

Adjustment for multiple biases can be done by specifying the option `combined`. To illustrate, we will adjust the observed OR for differential misclassification of the resins exposure and the selection bias and for uncontrolled confounding by smoking using the probability density functions for the bias parameters specified in the above sections.

```
. episensi 45 94 257 945, st(cc) reps(20000) nodots dseca(trap(.75 .85 .95 1))
> dspca(trap(.75 .85 .95 1)) dsenc(trap(.7 .8 .9 .95)) dspnc(trap(.7 .8 .9 .95))
> corrsens(.8) corrspec(.8) dsbfactor(log-n(0 0.21)) dpunexp(uni(.4 .7))
> dorce(log-n(0 0.639)) drrcd(log-n(2.159 .280)) seed(123) combined
Se|Cases   : Trapezoidal(.75,.85,.95,1)
Sp|Cases   : Trapezoidal(.75,.85,.95,1)
Se|No-Cases: Trapezoidal(.7,.8,.9,.95)
Sp|No-Cases: Trapezoidal(.7,.8,.9,.95)
Corr Se|Cases and Se|No-Cases : .8
Corr Sp|Cases and Sp|No-Cases : .8
selection bias factor: Log-Normal(0.00,0.21)
Pr(c=1|e=0): Uniform(.4,.7)
RR_cd      : Log-Normal(2.16,0.28)
OR_ce      : Log-Normal(0.00,0.64)
Probabilistic sensitivity analysis - Combined corrections
    Misclassification of the exposure
    Selection bias
    Unmeasured confounding

                                     Percentiles        Ratio
                            2.5      50       97.5      97.5/2.5
                            ----------------------------------------
Conventional                1.20     1.76     2.58      2.14
Systematic error            1.47     3.87     56.17     38.14
Systematic and random error 1.34     3.92     57.56     42.98
```

A comparison of the combined analysis with the single-bias analyses of the previous sections shows that, under the given priors confounding by smoking and selection bias have little impact on the observed resins lung-cancer OR, and that the greatest source of uncertainty is misclassification of the exposure.

5   Discussion

We have presented a new Stata command, `episens`, to perform both deterministic and probabilistic sensitivity analysis to assess the potential impact of systematic errors on observed exposure–disease associations. To illustrate, we applied `episens` to a case–control study regarding occupational resin exposure and lung-cancer deaths.

The advantages of a probabilistic sensitivity analysis have been discussed previously (Greenland 2001; Greenland 2003a; Greenland 2005; Greenland and Lash 2008; Lash and Fink 2003; Phillips 2003; Phillips and LaPole 2003; and Steenland and Greenland 2004). Briefly, a probabilistic sensitivity analysis requires the investigator to make explicit this uncertainty about bias parameters. This explication is done by using prior distributions for the parameters, which reflect background information and judgments of the investigator about sources of systematic error. The resulting distribution of bias-adjusted estimates captures the uncertainty about bias that is ignored by conventional statistics (such as confidence intervals). Under certain common conditions, this distribution can be viewed as an approximation to the more computationally demanding posterior distribution of Bayesian analysis (Greenland 2005).

Concerns have been raised about the arbitrariness in the particular distributions assumed for the bias parameters. The important point however is that changing the

prior distributions can result in different distributions for the bias-adjusted exposure–disease RR. This relation corresponds to the fact that if different investigators have different opinions about sources of bias, it should be no surprise if their conclusions differ.

Differences of opinion about bias sources may be represented by different prior distributions. The different bias-adjusted RR distributions that result then reflect the differences in conclusions (final opinions) we should expect when prior opinions differ and decisive data are lacking (as is usually the case in epidemiology). Thus, by varying the input prior distributions for probabilistic sensitivity analyses, we can illustrate the extent to which differences in prior opinions about various sources of bias may contribute to conflicting interpretations of the study. The possibility of conflicting outputs may encourage analysts to provide the best available evidence or arguments to support their own choices for prior distributions. As with earlier, more specialized SAS macros (Fox, Lash, and Greenland 2005), the Stata command presented in this paper greatly eases such variation by automating the transformation of the input priors to the output bias-adjusted distributions.

In conclusion, we have provided a user-friendly command suitable for both deterministic and probabilistic sensitivity analysis to evaluate bias due to misclassification of a binary exposure variable, selection bias, and bias due to an uncontrolled confounder. We hope that future refinements will provide extensions to variables with multiple levels, and allow for misclassification of multiple variables.

6 References

Chu, H., Z. Wang, S. R. Cole, and S. Greenland. 2006. Sensitivity analysis of misclassification: A graphical and a Bayesian approach. *Annals of Epidemiology* 16: 834–841.

Eddy, D. M., V. Hasselblad, and R. D. Shachter. 1992. *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*. Boston: Academic Press.

Fox, M. P., T. L. Lash, and S. Greenland. 2005. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *International Journal of Epidemiology* 34: 1370–1376.

Greenland, S. 1996. Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology* 25: 1107–1116.

———. 1998. The sensitivity of a sensitivity analysis (invited paper). In *1997 Proceedings of the Biometrics Section*, 19–21. Alexandria, VA: American Statistical Association.

———. 2001. Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Analysis* 21: 579–583.

———. 2003a. Generalized conjugate priors for Bayesian analysis of risk and survival regressions. *Biometrics* 59: 92–99.

———. 2003b. The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields to childhood leukemia. *Journal of the American Statistical Association* 98: 47–54.

———. 2005. Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A* 168: 267–306.

Greenland, S., and T. L. Lash. 2008. *Bias analysis*. In *Modern Epidemiology*, ed. K. J. Rothman, S. Greenland, and T. L. Lash, 3rd ed., in press. Philadelphia: Lippincott–Raven.

Greenland, S., A. Salvan, D. H. Wegman, M. F. Hallock, and T. J. Smith. 1994. A case–control study of cancer mortality at a transformer-assembly facility. *Interntional Archives of Occupational and Environmental Health* 66: 49–54.

Hoffman, F. O., and J. S. Hammonds. 1994. Propagation of uncertainty in risk assessments: The need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk Analysis* 14: 707–712.

Jurek, A. M., G. Maldonado, S. Greenland, and T. R. Church. 2006. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *European Journal of Epidemiology* 21: 871–876.

Lash, T. L., and A. K. Fink. 2003. Semiautomated sensitivity analysis to assess systematic errors in observational data. *Epidemiology* 14: 451–458.

Phillips, C. V. 2003. Quantifying and reporting uncertainty from systematic errors. *Epidemiology* 14: 459–466.

Phillips, C. V., and L. M. LaPole. 2003. Quantifying errors without random sampling. *BMC Medical Research Methodology* 3: 9.

Steenland, K., and S. Greenland. 2004. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology* 160: 384–392.

About the authors

Nicola Orsini is a Ph.D. student, Division of Nutritional Epidemiology, the National Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.

Rino Bellocco is associate professor of biostatistics, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, and Associate Professor of Biostatistics, Department of Statistics, University of Milano Bicocca, Milano, Italy.

Matteo Bottai is assistant professor of biostatistics, Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina.

Alicja Wolk is professor of nutritional epidemiology, the National Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.

Sander Greenland is professor of epidemiology, UCLA School of Public Health, and professor of statistics, UCLA College of Letters and Science, Los Angeles, CA.