STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:2911–2933 (DOI: 10.1002/sim.1879)

TUTORIAL IN BIOSTATISTICS
Regression analysis of multiple source and multiple informant data from complex survey samples

Nicholas J. Horton[1,*,†] and Garrett M. Fitzmaurice[2]

[1] *Department of Mathematics, Smith College, College Lane, Northampton, MA 01063, U.S.A.*
[2] *Department of Biostatistics, Harvard School of Public Health, Boston, MA, U.S.A.*

SUMMARY

In this tutorial, we describe regression-based methods for analysing multiple source data arising from complex sample survey designs. We use the term '*multiple-source*' data to encompass all cases where data are simultaneously obtained from multiple informants, or raters (e.g. self-reports, family members, health care providers, administrators) or via different/parallel instruments, indicators or methods (e.g. symptom rating scales, standardized diagnostic interviews, or clinical diagnoses). We review regression models for analysing multiple source risk factors or multiple source outcomes and show that they can be considered special cases of generalized linear models, albeit with correlated outcomes. We show how these methods can be extended to handle the common survey features of stratification, clustering, and sampling weights. We describe how to fit regression models with multiple source reports derived from complex sample surveys using general purpose statistical software. Finally, the methods are illustrated using data from two studies: the *Stirling County Study* and the *Eastern Connecticut Child Survey*. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: clustering; missing data; multiple indicators; multiple informants; sampling weights; stratification

1. INTRODUCTION

In many epidemiologic studies, information about health outcomes, risk factors, and service use is obtained from multiple sources (or informants). As an example, in psychiatric epidemiologic studies of childhood psychopathology, the child's parent is routinely used as a proxy data source; other informants (e.g. self-report, peers, teachers, clinicians, or trained observers) may also be employed, depending on the child's age and the nature of psychopathology under

[*] Correspondence to: Nicholas J. Horton, Department of Mathematics, Smith College, College Lane, Northampton, MA 01063, U.S.A.
[†] E-mail: nhorton@smith.edu

Contract/grant sponsor: NIMH: contract/grant number: RO1/MH54693

Received October 2003
Accepted March 2004

study. Contemporary psychiatric epidemiologic studies of children commonly use two or three informants as data sources, since assessment of psychopathology is inherently difficult [1] and there is often lack of reproducibility.

While multiple sources are almost routinely used in population- and community-based assessments of children's mental health and behaviour, multiple source data commonly arise in a wide variety of different fields of study. Family history studies, where many relatives are interviewed about the status of the proband and other family members generate similar types of data [2, 3]. In studies of breast cancer survivors, multiple sources can be used to provide information on medical comorbidity [4]. For investigations in nutritional epidemiology, multiple dietary instruments (e.g. food frequency questionnaires, 24-h recalls, food diaries) are used to assess nutrient intake [5, 6]. In studies that assess the quality of health care it has become routine to obtain data from both health care service providers and service users, thereby providing multiple perceptions (i.e. the professionals' and patients' perceptions) of the particular aspect of health care under study [7]. Leape *et al.* [8] used two physician reviewers to identify and evaluate adverse medical events. Service utilization studies also feature multiple informants, where both user and provider(s) are asked to report types of services obtained/provided (e.g. Reference [9]) or when multiple information sources, such as self-report and administrative database, are queried [10]. Finally, multiple sources have become increasingly common in hospital-based and outpatient-based assessments of treatments for mental illness [11–13].

Our use of the general term '*multiple-source*' data is intended to encompass all cases where data are simultaneously obtained from multiple informants or raters (e.g. self-reports, family members, health care providers, administrators) or via different/parallel instruments, indicators or methods (e.g. symptom rating scales, standardized diagnostic interviews, or clinical diagnoses). Note, however, that we restrict our use of this term to multiple source data that are commensurate. That is, multiple source data are thought to provide multiple measures of the same underlying variable and are measured on a similar scale.

Multiple sources can provide information on either risk factors or outcomes. An example of the former was reported in a follow-up study of children into adolescence and adulthood, where a child's self-report and a parent's report on anxiety at ages 5 and 9 were used as multiple source predictors of a diagnosis of depression at age 18 [14]. In an example of the latter, taken from two parallel community-based surveys of child psychopathology in Connecticut [15, 16], multiple source data on children's mental health outcomes were primarily assessed by a parent's and a teacher's report via parallel forms of a standardized symptom checklist.

Irrespective of whether they arise as risk factors or outcomes, one of the key methodological challenges in analysing multiple source data concerns how they should best be represented in statistical models. Many of the traditional methods for analysing multiple-source data have not been completely satisfactory. An approach commonly used in the past is a 'pooling' strategy, where the information from the multiple sources is somehow combined into a single number summary for each subject. The 'best estimate' diagnoses [17, 18], in which clinicians review all available data and arrive at a diagnosis on a case-by-case basis, is one such pooling method. Alternatively, a variety of strategies and algorithms for pooling and combining multiple source data have been introduced. Two examples of some contemporary pooling strategies include the 'or' and the 'and' algorithms. In the so-called 'or' algorithm binary source data are considered to be positive if *any* of the source data are positive, and negative otherwise [19]. In the 'and' algorithm binary source data are considered to be positive if *all* of the source data are positive, and negative otherwise [20]. Another strategy for producing a single number summary is to take the arithmetic average of the multiple source data, a strategy that is somewhat more appealing when the source data are quantitative. There are many reasons why the pooling of data from multiple sources is not very desirable. These include: (1) the optimal algorithm for combining multiple source data depends upon the type of measurement error present; (2) pooling does not permit the examination of differences in risk-factor effects across sources; and (3) many pooling algorithms are not clearly defined in the presence of missing data from one or more sources.

The main alternative to pooling has been to conduct separate analyses for each source and report the results separately. This approach has its own drawbacks, however: (1) separate analyses yield multiple (and often differing) sets of results for the different sources, which may be difficult for the consumer of the research findings to interpret; (2) separate analyses provide no formal means of evaluating how similar or different the results are across the various sources (or to summarize them in a single set of results, if they are sufficiently similar); and (3) separate analyses may be based on different subsets of the data, if some subjects are missing data from one source and others are missing data from another source.

In response to the shortcomings of existing analytic methods for multiple-source data, Fitzmaurice and colleagues [21, 22] proposed regression methodology for simultaneously analysing binary multiple source outcomes, while Daskalakis *et al.* [23] generalized these methods to categorical multiple source outcomes. Kuo *et al.* [24] and Goldwasser and Fitzmaurice [25] independently proposed extensions of these methods to continuous multiple source outcomes. In related work, Kraemer *et al.* [26] addressed methods to develop consensus methods for informant reports, in terms of the context and perspective of these sources.

Horton *et al.* [27] considered the shortcomings of existing methods for analysing multiple source data when they are used as risk factors or predictor variables. They developed regression methods for the case where both the multiple source risk factor and the outcome are binary. These regression-based methods treat the multiple sources as providing either conceptually different information or the same information measured with error. They also make full use of all available information, even from subjects who have missing data from one or more sources. A notable feature of these regression-based methodologies for simultaneously analysing multiple source risk factors and outcomes is that they can be considered special cases of generalized linear models, albeit with potentially correlated outcomes.

An additional complication for the analysis of multiple source data from many epidemiologic studies is the use of complex survey designs. Administrative, pragmatic, as well as scientific factors may lead researchers to divide data collection into separate geographic districts, or to oversample particular groups that are of main interest. In this tutorial, we consider the analysis of multiple source data arising from complex survey designs. In particular, we demonstrate how existing regression-based methods for simultaneously analysing multiple source data are related to generalized linear models for correlated data and indicate how these methods can be extended to handle complex survey designs, and describe their application in two examples.

The remainder of this tutorial is organized as follows. In Section 2, we review regression-based methods for simultaneously analysing multiple source data. We consider regression models for multiple source outcomes and multiple source predictors separately. We emphasize how these models can be expressed as generalized linear models, albeit with potentially correlated dependent variables. In Section 3, we consider multiple source data arising from complex survey samples. We review some basic concepts from the survey sampling literature and describe how the regression model parameters can be estimated using an approximate likelihood that takes account of common survey features of stratification, clustering, and sampling weights. Some close links between this approach and generalized estimating equations (GEE) are noted. In Section 4, the methods are illustrated using data from two studies: the *Stirling County Study* and the *Eastern Connecticut Child Survey*. We discuss software to fit these models in Section 5, and conclude with a discussion of other aspects of the analysis of multiple source reports in this setting.

2. GENERALIZED LINEAR MODELS FOR MULTIPLE SOURCE DATA

2.1. Regression methods for multiple source outcomes

In previous research on methods for analysing outcomes measured by multiple sources, Fitzmaurice and colleagues [21, 22] and Daskalakis *et al.* [23] described regression models for binary, categorical, and ordinal multiple source outcome data. They proposed a multivariate regression model, where the measures obtained from all sources are analysed simultaneously. There is an extensive literature relating to the analysis of paired or other multivariate outcomes (e.g. References [28–30]). Many of these methods for categorical multiple source outcomes can be readily extended and generalized to the case of continuous or count data. In this section, we review these methods, make the connections with generalized linear models more explicit and show how these methods can handle diverse types of multiple source outcomes.

We first establish notation, describe the data typically collected in these settings, and propose a general modelling strategy. We assume that there are N independent subjects, each with an outcome obtained from J different sources. The outcome can be binary, ordinal, discrete, continuous or count data. Let Y_{ij} represent the outcome obtained for the ith subject from the jth source (with $i=1,\ldots,N;\ j=1,\ldots,J$). In addition, let $\boldsymbol{X}_{ij}$ be a $p \times 1$ vector of covariates, associated with the outcome obtained for the ith subject from the jth source. Note that each $\boldsymbol{X}_{ij}$ will, in general, contain both source variables (or indicators for the different sources) and subject-specific risk factors. That is, $\boldsymbol{X}_{ij}$ may include a constant (for the intercept term), a set of $(J-1)$ indicator variables indicating the source, various subject-specific risk factors, and possibly source by risk factor product terms. Finally, we let $\boldsymbol{Y}_i=(Y_{i1},\ldots,Y_{iJ})'$ be the $J \times 1$ outcome vector for the ith subject, and $\boldsymbol{X}_i$ be the associated $J \times p$ matrix of covariates.

Following the approach described in Fitzmaurice *et al.* [21], regression models relating the mean of the outcome measured by each source to the vector of covariates or risk factors can be developed. Specifically, we consider multivariate regression models for the mean of $\boldsymbol{Y}_i$, conditional on both source and risk factors, that are of the following general form:

$$g(E[Y_{ij} \mid \boldsymbol{X}_{ij}]) = \boldsymbol{X}_{ij}'\boldsymbol{\beta} \tag{1}$$

where $g(\cdot)$ is a known link function. Thus, the conditional mean of Y_{ij} may depend on the source, any other covariates of interest (e.g. risk factors and potential confounders), and possibly their interactions. When the outcome is continuous, the identity link function is a natural choice for $g(\cdot)$; with binary or count data, the logit or log link functions are natural choices, respectively. However, in principle, any suitable link function can be selected.

A feature of the regression model shown in equation (1) is that it can be used to specify hypotheses about the effects of the sources and of risk factors on the outcome (with the latter usually being the effects of primary scientific interest), as well as possible source by risk factor interactions. Note that the source by risk factor interactions represent contrasts of within-subject effects. A test of a source by risk factor interaction is equivalent to a test of the differences among the source-specific regression coefficients for the corresponding risk factor. Because the multiple source reports are expected to be positively correlated, these regression coefficients are also positively correlated and we have more power than we usually would have for testing interactions.

A simple example can help illustrate the potential application of the model shown in equation (1). Consider the Eastern Connecticut Child Survey, an epidemiologic study of children that assesses psychopathology using ratings from parents and teachers. Suppose that we are interested in assessing the association of a binary risk factor, e.g. family stressors, with child psychopathology. The following bivariate regression allows a single regression model to be fit to the data from both sources,

$$g(E[Y_{ij}|X_{ij}]) = \beta_0 + \beta_1\,\mathrm{SOURCE}_{ij} + \beta_2\,\mathrm{STRESS}_i + \beta_3\,(\mathrm{SOURCE}_{ij} \times \mathrm{STRESS}_i) \qquad (2)$$

where SOURCE $(0=\text{parent},\ 1=\text{teacher})$ and STRESS $(0=\text{no},\ 1=\text{yes})$ are indicator variables. This model assumes that the effect of family stress on the (appropriately transformed) mean rating may vary by source (parent or teacher). In general, source-related differences in the effect of family stress can be evaluated via tests of β_1 and/or β_3 equaling zero. For example, the simplified bivariate regression model,

$$g(E[Y_{ij}|X_{ij}]) = \beta_0 + \beta_1\,\mathrm{SOURCE}_{ij} + \beta_2\,\mathrm{STRESS}_i \qquad (3)$$

assumes that the effect of family stress on the mean rating does not vary by source, but overall the mean rating reported by parents and teachers may differ (i.e. $\beta_1 \neq 0$).

Note that, so far, no distributional assumptions have been made other than the conditional mean of the Y_{ij}'s is given by equation (1). However, this model is very flexible and actually represents a broad class of regression models that include as special cases linear regression models for continuous multiple source data, logistic regression models for binary multiple source data, and log-linear regression models for multiple source count data. Later, we make some additional assumptions about the Y_{ij}'s and the association among the Y_{ij}'s.

2.2. Regression methods for multiple source predictors

Multiple source reports are also commonly used as predictors of an outcome. Horton *et al.* [27] catalogued a variety of methods that have been suggested in the literature, and discussed the advantages and disadvantages of each type. These methods can be grouped into two categories: approaches that include all source reports in a single regression, and approaches that simultaneously estimate separate regression equations, one for each source report. Of note, both approaches fit within the generalized linear model framework described earlier and can be expressed in the form of (1).

To illustrate these two general approaches, we consider an example from the Stirling County Study, where multiple source reports of psychiatric diagnosis (from self-report [SELFDIAG] or physician-report [GPDIAG]) were used to predict mortality. A regression model can be fit that includes both source reports:

$$g(E[Y_i|X_i]) = \beta_0 + \beta_1\,\mathrm{SELFDIAG}_i + \beta_2\,\mathrm{GPDIAG}_i + \beta_3\,(\mathrm{SELFDIAG}_i \times \mathrm{GPDIAG}_i) \qquad (4)$$

where Y_i is an indicator of mortality. Horton *et al.* [27] describe the relationship between this model and regression models with a single predictor derived from *ad hoc* combination rules (such as the 'or' and 'and' rules). Note that estimation of the parameters in model (4) is straightforward, since there is only a single observation per subject. While this model may be particularly attractive if the primary goal is prediction of the outcome, the regression parameters do not have useful interpretations in terms of the effect of the risk factor on the outcome. Indeed, the association between the risk factor and outcome is likely to be attenuated in model (4) due to the conditioning on all source reports. For example, in the above model, the regression parameters have interpretation in terms of the effect of a positive report from one source, conditional on the report of the other source. However, in many settings, the marginal association of each source report with the outcome may be of greater scientific interest. The potential attenuation of covariate effects in the case of linear models (assuming an identity link) can be readily seen by considering the following simple illustration. Suppose that X_{i1} and X_{i2} are two source reports, having equal variances σ_x^2, and with common marginal associations with Y_i. The common association with Y_i can be expressed in terms of the correlation, $\text{Corr}(Y_i, X_{i1}) = \text{Corr}(Y_i, X_{i2}) = \rho_{yx}$. In addition, the two source reports are assumed to be positively correlated, $\text{Corr}(X_{i1}, X_{i2}) = \rho_{x_1x_2} > 0$. Then the regression coefficient for X_{i1} (or X_{i2}) in the linear regression of Y_i on X_{i1} (or X_{i2}) alone is given by $(\sigma_y/\sigma_x)\rho_{xy}$, where $\sigma_y^2 = \text{Var}(Y_i)$. However, note that the regression coefficient for X_{i1} (or X_{i2}) in the linear regression of Y_i on both X_{i1} and X_{i2} is given by $(\sigma_y/\sigma_x)\rho_{xy}/(1+\rho_{x_1x_2})$, and is always smaller than $(\sigma_y/\sigma_x)\rho_{xy}$ when $\rho_{x_1x_2} > 0$. Furthermore, the degree of attenuation is related to the magnitude of the (positive) correlation among source reports. Similar attenuation arises also in regression models with non-linear link functions.

In light of these concerns, an alternate approach, described independently by Horton *et al.* [27] and Pepe *et al.* [31], involves simultaneously fitting separate regression equations, one for each source. In the Stirling County Study example, this could be represented by the following model:

$$\begin{aligned} g(E[Y_i|X_{i1}]) &= \beta_0 + \beta_1\,\text{SELFDIAG}_i \\ g(E[Y_i|X_{i2}]) &= (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1)\,\text{GPDIAG}_i \end{aligned} \tag{5}$$

where the same outcome appears in each equation, but with different source-dependent predictors. Other predictors may also be included in this model. The $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ parameters denote the source-related differences in the effect of psychiatric diagnosis on mortality. For example, a test of $\alpha_1 = 0$ can be used to determine if the association between diagnosis and mortality differs for the two source reports. When there are no significant source differences in (5), a simplified model may be fit that pools all of the information

$$\begin{aligned} g(E[Y_i|X_{i1}]) &= \beta_0 + \beta_1\,\text{SELFDIAG}_i \\ g(E[Y_i|X_{i2}]) &= \beta_0 + \beta_1\,\text{GPDIAG}_i \end{aligned} \tag{6}$$

In model (6) the β_1 parameter is interpreted as the association between diagnosis and the outcome, averaged over source reports.

2.3. Estimation

So far, no distributional assumptions have been made other than the conditional mean of each Y_{ij} can be expressed in terms of a generalized linear model [32] given by (1). Next, we assume that the marginal distribution of each Y_{ij} belongs to the exponential family,

$$f(Y_{ij}) = \exp[\{Y_{ij}\theta_{ij} - a(\theta_{ij})\}/\phi + b(Y_{ij}, \phi)]$$

where θ_{ij} is a 'location' parameter, often referred to as the 'canonical' parameter, and ϕ is a 'scale' parameter (where ϕ is sometimes known). Note that $a(\cdot)$ and $b(\cdot)$ are simply specific functions that distinguish distributions belonging to the exponential family. The exponential family of distributions include the normal, Bernoulli, and Poisson distributions. Given that each Y_{ij} is assumed to have a distribution from the exponential family, the marginal expectations of the Y_{ij}'s, $E(Y_{ij}) = \mu_{ij}$, is then modelled as a function of covariates by equation (1). In addition, the marginal variance of Y_{ij} depends on the marginal mean according to, $\mathrm{Var}(Y_{ij}) = v_{ij} = v(\mu_{ij})\phi$, where $v(\mu_{ij})$ is a known 'variance function' (i.e. a known function of μ_{ij}) and ϕ is a scale parameter that may or may not need to be estimated.

Note, however, that multiple source data are usually positively correlated. This correlation must be accounted for when analysing either multiple sources outcomes or predictors. To account for the correlation, there are three broad approaches. The first is to completely specify the joint distribution of $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$. The second is to specify a model for the correlation among the Y_{ij}'s; note that the latter does not specify the joint distribution of $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$, except in the special case where $\boldsymbol{Y}_i$ has a multivariate normal distribution. The third approach is to assume that the Y_{ij}'s are independent (and hence uncorrelated) for the purposes of estimation, but make a suitable adjustment to the standard errors for the correlation among the Y_{ij}'s. Note that the magnitude of the true correlation among the Y_{ij}'s does not alter in any way the interpretation of $\boldsymbol{\beta}$. While there are merits to each of these approaches, only the third approach provides a relatively straightforward extension when multiple source data arise from complex survey samples.

Estimation of the regression parameters proceeds similarly for models with multiple source outcomes or multiple source predictors. Assuming that the Y_{ij}'s are independent, the maximum likelihood estimate of $\boldsymbol{\beta}$ is obtained by taking the derivative of the log likelihood with respect to $\boldsymbol{\beta}$, and then finding the values of $\boldsymbol{\beta}$ that make those derivatives equal to 0. Given

$$\ln L = \sum_{i=1}^{N}\sum_{j=1}^{J} (\{Y_{ij}\theta_{ij} - a(\theta_{ij})\}/\phi + b(Y_{ij}, \phi)) \tag{7}$$

the derivative of the log likelihood with respect to $\boldsymbol{\beta}$ is,

$$\partial \ln L/\partial\boldsymbol{\beta} = \sum_{i=1}^{N}\sum_{j=1}^{J} (\partial\theta_{ij}/\partial\boldsymbol{\beta})\{Y_{ij} - \mu_{ij}\}/\phi$$

In cases where a 'canonical' link function, $\theta_{ij} = \boldsymbol{X}_{ij}'\boldsymbol{\beta}$, has been assumed,

$$\partial \ln L/\partial\boldsymbol{\beta} = \sum_{i=1}^{N}\sum_{j=1}^{J} \boldsymbol{X}_{ij}'\{Y_{ij} - \mu_{ij}\}/\phi$$

Solving the set of simultaneous equations, $\sum_{i=1}^{N}\sum_{j=1}^{J}\boldsymbol{X}_{ij}'\{Y_{ij} - \mu_{ij}\} = 0$, yields an estimate of $\boldsymbol{\beta}$.

The analysis can proceed by naively assuming that the multiple source outcomes for any given subject, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$, are independent observations. While this 'naive' approach yields estimates of $\boldsymbol{\beta}$ that are valid, their nominal standard errors (under the independence assumption) are not. However, valid standard errors can be obtained using the well-known empirical variance estimator, first proposed by Huber [33]. We defer any further discussion of estimation of standard errors, with suitable adjustments for the correlation among the Y_{ij}'s, to Section 3.

In summary, the analysis of multiple source outcomes or predictors can proceed in two stages. In the first stage, the correlation among the multiple source data is simply ignored and standard generalized linear regression is used to obtain estimates of $\boldsymbol{\beta}$ (and possibly of ϕ). In the second stage, valid standard errors are obtained using an alternative, but widely implemented, variance estimator that properly accounts for the correlation among multiple source outcomes. The chief advantage of this approach is that it can be readily implemented using standard, widely available, statistical software for generalized linear models [34]. Finally, we note that this approach of using standard regression models intended for a single outcome to analyse a multivariate outcome is a special case of the generalized estimating equations approach [35]. Next, we consider how this general approach can be extended to handle multiple source data arising from complex survey samples.

3. GLMs FOR MULTIPLE SOURCE DATA FROM COMPLEX SURVEY SAMPLES

As mentioned earlier, an additional complication in many epidemiologic studies is that the multiple source data arise from complex survey samples. For example, the Eastern Connecticut Child Survey used a complex survey design with stratification, multi-stage clustering and unequal sampling weights. Because of the complex survey designs used, traditional methods of analysis that assume simple random sampling cannot be applied. In this section, we consider extensions of the generalized linear models for multiple source data presented in Section 2 to handle complex survey designs. Before doing so, we review the three main features that need to be accounted for in the analysis of multiple source data from complex surveys: (i) stratification, (ii) clustering, and (iii) sampling weights. A more detailed description of modern model-based methods for analysing survey data can be found in Särndal *et al.* [36] while the special issue of *Statistical Methods in Medical Research* featured a number of accessible review papers [37–42]. Dunn [43] provides a gentle introduction to use of these methods in psychiatry in the context of two examples in psychiatric morbidity. Pickles *et al.* [44] and Vázquez-Barquero *et al.* [45] have also reported results using this methodology.

It is very common in sample surveys to divide the population of sampling units into distinct subpopulations, referred to as *strata* (e.g. geographic regions or administrative units). A distinctive feature of these strata is that each sampling unit can occur in one, and only one, stratum. Within each of the strata, a separate sample is selected from among all the sampling units that comprise that stratum. Of note, the selection of samples is carried out separately and independently within each stratum. From a purely statistical perspective, stratification is often used to reduce the variances of the sample estimates, with the variance decreasing as a function of the degree to which any stratum-specific statistics diverge and the sampling units within stratum are homogeneous. More generally, the variance of sample estimates is reduced to the extent that the variability among sampling units within the strata is less than their variance in the entire population. Stratification is very common in sample surveys, and the main motivation for its use is often for logistical rather than statistical reasons. Failure to account for stratification in the analysis has little impact on estimates of population parameters but will, in general, result in an overestimation of the variability (i.e. standard errors are overestimated and confidence intervals are too wide).

The second feature that must be accounted for in the analysis of sample survey data is clustering. Epidemiologic surveys commonly use cluster sampling, where the sampling unit (or unit of selection), contains more than one population element. That is, clusters are sampling units containing several elements. For example, the cluster could be a classroom of students. Alternatively, the cluster could be a subject, with the elements of the cluster being the multiple source data on that subject. Because the clusters are selected first, they are generally referred to as the *primary sampling units* or PSUs. Note that there can be more than a single level of clustering in survey data. In multistage clustering, the clusters selected at the first stage are the PSUs; in the second and later stages, further sample selection occurs within the PSUs and so on. The final or ultimate sample obtained from the selected PSUs are often referred to as 'ultimate' or 'primary' clusters. The ultimate clusters represent the aggregation of all units (or observations) included in the sample from a PSU. In general, failure to account for clustering in the data has little impact on estimates of population parameters but will result in underestimation of the variability (i.e. standard errors are underestimated and confidence intervals are too narrow). In the statistical analysis of multistage samples, a commonly used approximation in the survey literature involves the specification of the first stage strata and PSU identifiers at the highest level only [36, 42]. The lower level clustering is subsumed within the PSU and the analysis proceeds as if there were a single levels of clustering at the level of the 'ultimate' or 'primary' cluster. Essentially this treats the primary clusters as i.i.d. draws from some superpopulation, where the lower stages are subsumed into the i.i.d. process.

Finally, the sample selection in many epidemiologic surveys commonly involves unequal selection probabilities. That is, each PSU does not have an equal probability of selection. As a result, the data analysis must take account of the sampling weights. The intuition for why an adjustment is required is that the sampling weights can be considered a measure of how many units in the population the sampled PSU represents. That is, if the sampled PSU's probability of selection was small, say π, then the analysis must inflate that PSU's contribution by a factor of $1/\pi$ so that the PSU represents itself and those that were not selected. In general, failure to account for the weights in the analysis will yield estimates of population parameters that are biased and can result in underestimation of the variability (i.e. standard errors are underestimated and confidence intervals are too narrow).

Thus, when multiple source data arise from a complex survey, stratification, clustering, and unequal selection probabilities must all be taken into account in the analysis in order to avoid misleading inferences concerning the population parameters of interest. Next, we consider estimation of the regression model parameters in (1) when the multiple source data arise from a complex sample survey design. To do so, we need to modify the notation used in Section 2. Here, we assume that the population can be divided into S distinct subpopulations or strata ($s=1,\ldots,S$). We assume that our sample is comprised of N_s PSUs from strata s ($i=1,\ldots,N_s$) and that the ultimate cluster is comprised of n_{si} units ($j=1,\ldots,n_{si}$). The response and covariates for the sijth unit are denoted by Y_{sij} and $\boldsymbol{X}_{sij}$, respectively. Finally, the (known) sampling weight for the sijth unit is w_{sij}.

The regression parameters in (1) can be estimated using the following approximate log likelihood in place of (7),

$$\sum_{s=1}^{S}\sum_{i=1}^{N_s}\sum_{j=1}^{n_{si}} w_{sij}(\{Y_{sij}\theta_{sij} - a(\theta_{sij})\}/\phi + b(Y_{sij}, \phi)) \tag{8}$$

In cases where a ‘canonical’ link function, $\theta_{sij} = \boldsymbol{X}'_{sij}\boldsymbol{\beta}$, has been assumed this involves solving the following set of estimating equations:

$$\mathscr{L}(\boldsymbol{\beta}) = \sum_{s=1}^{S}\sum_{i=1}^{N_s}\sum_{j=1}^{n_{si}} \mathscr{L}_{sij} = \sum_{s=1}^{S}\sum_{i=1}^{N_s}\sum_{j=1}^{n_{si}} \boldsymbol{X}'_{sij} w_{sij}\{Y_{sij} - \mu_{sij}\} = 0 \tag{9}$$

Note that this is simply a weighted version of the usual score equations for a generalized linear model and an estimate of $\boldsymbol{\beta}$ can be obtained using any of the widely available software for generalized linear models that allow for the inclusion of weights in the analysis.

While the solution to (9) provides a valid estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, for inferences about $\boldsymbol{\beta}$ we need an estimator of the variance of $\hat{\boldsymbol{\beta}}$ that takes accounts of the stratification, clustering, and unequal selection probabilities. A valid estimator is provided by

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{V}^{-1}\hat{K}\hat{V}^{-1} \tag{10}$$

where

$$V = \sum_{s=1}^{S}\sum_{i=1}^{N_s}\sum_{j=1}^{n_{si}} w_{sij} v_{sij}(\boldsymbol{X}'_{sij}\boldsymbol{X}_{sij})$$

$$K = \sum_{s=1}^{S} K_s, \quad K_s = \frac{N_s}{N_s - 1}\sum_{i=1}^{N_s}(\mathscr{L}_{si} - \bar{\mathscr{L}}_s)(\mathscr{L}_{si} - \bar{\mathscr{L}}_s)'$$

$$\mathscr{L}_{si} = \sum_{j=1}^{n_{si}} \mathscr{L}_{sij}, \quad \bar{\mathscr{L}}_s = \frac{1}{N_s}\sum_{i=1}^{N_s} \mathscr{L}_{si}$$

and v_{sij} is the known variance function. Note that (10) has exactly the same form as the empirical variance estimator proposed by Huber [33] and also advocated by Liang and Zeger [35] for the special case of the ‘independence’ estimating equations approach, except that (10) accounts for unequal selection probabilities and uses K, a pooled within-stratum estimator of the covariance of $\mathscr{L}(\boldsymbol{\beta})$. An interesting discussion of the relation between these two variance estimators can be found in Williams [46]. Finally, for non-canonical link functions, there are very similar expressions for $\mathscr{L}(\boldsymbol{\beta})$ and V.

4. APPLICATIONS

4.1. Eastern Connecticut Child Survey

The Eastern Connecticut Child Survey (ECCS) [16], was designed to estimate the prevalence of mental health problems in children, in a three-county non-metropolitan planning region. The study sample was drawn from class enrolment data from public, private and institutional schools nested within differing geographic areas: small cities, suburban areas, and rural townships. We will use the 7 strata (3 small cities, 3 suburban, and 1 rural) in our analysis. The mean of the sampling weights was 24.8, and the weights ranged from 4.4 to 79.5. Here the ultimate cluster was schools, with children nested within schools, and multiple source reports nested within children.

Researchers solicited multiple source reports of psychopathology from the parents and teachers of this sample of school children. In particular, each child's parent (or primary caregiver) and teacher completed parallel versions of a standardized psychiatric symptom checklist, namely the Child Behavioral Checklist [47] and the Teacher's Report Form [48], which were completed by parents (Y_1) and teachers (Y_2) respectively. In the study, 44 per cent of teacher ratings on children were unobserved. Missingness of this magnitude is not uncommon: a similar rate was reported in their Ontario Child Health Study [49]. There were a variety of causes of missingness for the teacher reports, including school district non-participation, parental refusal to give consent, and teacher nonresponse. Fitzmaurice *et al.* [22] considered the question of whether the missingness in this data set is related to the unobserved teacher's rating, and found little evidence for this hypothesis. The raw scores were dichotomized at the cutpoint for borderline/clinical internalizing problems. A score of 1 indicates internalizing problems, and a score of 0 indicates normal range.

One of the primary research questions concerned estimation of prevalence of internalizing problems, measured from parent and teacher reports, and how prevalence estimates differed according to characteristics of the child. To illustrate the methods, we consider the model described by Fitzmaurice *et al.* [21], for a sample of $n = 1688$ subjects with a total of 2636 multiple source reports of internalizing problems. Predictors of psychopathology in this model include: gender of the child (BOY: $1 = $boy, $0 = $otherwise), area of residence (rural, suburban, or small city), social class (low, middle or high), single parent (MOMSING: $1 = $yes, $0 = $no), maternal stress and dissatisfaction with family life (MATSTRS: $1 = $yes, $0 = $no), child's health (HLTHPRO: $1 = $yes, $0 = $no), grade repetition (GRADEREP: $1 = $the child had repeated a grade, $0 = $otherwise), and family stress (e.g. divorce or death, FAMSTRS: $1 = $yes, $0 = $no). The model fit by Fitzmaurice *et al.* [21], using data from both the Eastern Connecticut Child Survey and the New Haven Child Survey, tested for potential interactions of gender with selected risk factors and source-risk factor interactions. Their final model, which was fit using maximum likelihood methods [22], incorporated source-specific effects for maternal stress, child's health and family stress, as well as an interaction between family stress and gender. We replicated this general model, accounting for the survey design using the `svylogit` command in Stata [50]. For this type of model, an interaction between a predictor (such as child's health) and source is equivalent to the model given by (2) for that term. For a predictor (such as grade repetition) where no source interaction is included, the model is equivalent to that given by (3). In preliminary analyses, the interaction between maternal distress and source was found to be not significant, and the interaction term was dropped.

The results from the final model, displayed in Table I, are interpreted as log odds ratios. There was a significant interaction between source and health problems ($p = 0.03$), indicating that the association between health problems and internalizing behaviour was larger when the outcome was reported by the parent. There was weak evidence that the relationship between internalizing behaviour and family stress varied by source ($p > 0.10$). There was little association between internalizing problems and area of residence ($F_{2,61} = 1.16$, $p = 0.32$) or single parent status ($p > 0.70$). There was a borderline significant association between social class

Table I. Regression parameter estimates (representing log OR for internalizing problems) from a logistic regression model with source effects.

	Estimate* (Standard error)		
Parameter	Teacher		Parent
Intercept	−2.28 (0.28)		−3.08 (0.26)
Area of residence			
Suburban		0.01 (0.18)	
Small city		0.15 (0.11)	
Social class			
Middle		0.19 (0.14)	
Low		0.56 (0.31)	
Momsing		−0.08 (0.25)	
Matstrs		0.80 (0.13)	
Hlthpro	0.06 (0.31)		0.87 (0.16)
Graderep		0.40 (0.16)	
Boy		0.78 (0.24)	
Famstrs	0.38 (0.29)		0.86 (0.25)
Boy×famstrs		−0.73 (0.27)	
Girls, family stress	0.38 (0.29)		0.86 (0.25)
Boys, family stress	−0.36 (0.23)		0.13 (0.24)

*Single estimates indicate common effects for teachers and parents; separate estimates are shown for effects that vary by source.

and the outcome ($F_{2,61} = 2.78$, $p = 0.07$), with low SES subjects more likely to have internalizing problems. Maternal satisfaction, health problems (using parent report), grade repetition, and family stress (using parent report) were all associated with more internalizing problems. Finally, the association between family stress and the outcome differed by sex ($p = 0.01$).

4.2. *Stirling County Study*

The Stirling County Study is a long-term investigation in psychiatric epidemiology conducted among adult residents of an area in Atlantic Canada. The fictitious name of ‘Stirling’ is used to protect identity [51]. The study began in 1952 and involves repeated cross-sectional surveys as well as cohort follow-up investigations extending to 1992. We consider the 953 subjects with complete data from the 1952 sample.

The study collected multiple source reports (self- and physician-report) about psychiatric disorders. Here, we are interested in using these reports as predictors of mortality over a 16-year follow-up period (approximately one quarter of the sample died during this time). These multiple reports of psychiatric disorders have been used in conjunction with information on follow-up to understand the association between psychiatric disorders and mortality in a community population.

In earlier reports of the Stirling County Study, psychiatric cases were identified by having psychiatrists review the materials assembled from both sources [52]. Later work focusing on the common theme of the relationships between psychiatric disorders and mortality analysed the sources separately [53, 54]. Horton *et al.* [55] jointly analysed these data using multiple source methods and found little evidence that relationships varied according to the sources. They concluded that psychiatric disorders were significantly associated with mortality; in particular, subjects who died before reaching the age of 50 were especially likely to have had a psychiatric disorder. However, their analysis did not account for the survey design.

The county was divided into 9 strata for purposes of sampling. Samples from each of the strata included between 33 and 255 subjects, with varying sampling weights. Overall, the sampling weights ranged from 1 to 19.7, with a mean of 8.2. Several districts were oversampled to incorporate additional information about economically advantaged and disadvantaged communities: the mean weight per district ranged from 1.8 (district 1) to 16.6 (district 9). Here, subjects are the PSUs and districts are the strata.

To examine the relationship between mortality and psychiatric diagnosis we utilize discrete time survival models [56] to assess the magnitude and significance of the relationship between independent variables and mortality. These regression methods specify a piecewise exponential survival distribution and approximate the proportional hazards model but have the distinct advantage that they fit within the framework of model (1).

More formally, we partition the 16-year follow-up period in the study into four mutually exclusive, exhaustive intervals $\Omega_1, \ldots, \Omega_4$, with a constant hazard function within each interval. We observe (Y_{ik}, T_{ik}) for $i = 1, \ldots, N$ and $k = 1, \ldots, 4$, where Y_{ik} denotes whether the ith subject died during the kth period, and T_{ik} denotes the time at risk for the ith subject during the kth period. Here $Y_{ik} = 1$ for some i and k implies that $T_{ik'} = 0$ for $k' = k + 1, \ldots, 4$. The model for the mortality rate (expected number of events per year) is given by the Poisson regression model

$$\log E[Y_{ik} | \mathbf{T}_i, \mathbf{X}_i] = \log(T_{ik}) + \mathbf{X}_i \boldsymbol{\beta}$$

Table II displays part of the observed data for two hypothetical subjects A and B. Because subject A was observed for all 16 years, the analysis data set includes 8 records, one for each source report (self and physician) for each of the 4 time intervals. The same outcome is repeated for each source report. Subject B died after 6.5 years, so this subject contributes 4 years to interval 1 and 2.5 years to interval 2, for each of the self- and physician-reports.

In addition to using the case assessments from the two sources (SELFDIAG and GPDIAG), other predictors included in the regression model were gender, age and time interval. Age in 1952 was divided into three categories: <50, $50-69$, and $70+$. The last time interval was used as the reference group. We fit a model for mortality rate that included main effects of age, diagnosis, gender and interval, along with the interaction between age (2 df) and diagnosis. Other predictors were assumed to be constant over time (i.e. no interaction with interval).

The piecewise exponential model was fit using `svypoisson` in Stata and the parameter estimates (log annual mortality rate ratios) are displayed in Table III. In a preliminary analysis, we fit a model that allowed the association between risk factor and outcome to vary by source (main effects of source and risk factor plus their interactions). There was no evidence of any significant interaction between source and the risk factors (all p-values >0.10; omnibus or overall test, $F_{5,1066} = 0.96$, $p = 0.44$), so these terms were dropped, yielding a model similar to (6). Dropping the interaction with source implies that the association between each risk factor and mortality did not differ by source, and yields a simple model that combines information from the sources. The final model is similar to that described by previous reports [55], though this analysis accounts for the survey design. Overall, the force of mortality tends

Table II. Data set framework for Stirling County Study Poisson regression example.

id		died	years	selfdiag	gpdiag	gender
Original (*one observation per subject*) *data set*						
A		0	16	0	1	F
B		1	6.5	0	0	M
Analysis (*transformed*) *data set*						
id	Interval	Died (Y)	Years (T)	Informant	Diag	Gender
A	1	0	4	Self	0	F
A	2	0	4	Self	0	F
A	3	0	4	Self	0	F
A	4	0	4	Self	0	F
A	1	0	4	gp	1	F
A	2	0	4	gp	1	F
A	3	0	4	gp	1	F
A	4	0	4	gp	1	F
B	1	0	4	Self	0	M
B	2	1	2.5	Self	0	M
B	1	0	4	gp	0	M
B	2	1	2.5	gp	0	M

Table III. Regression parameter estimates (representing log annual mortality rate ratios) from a piecewise exponential survival model with no source effects.

Parameter	Estimate	(SE)
Intercept	−5.58	(0.29)
Interval (0–4)	−0.96	(0.21)
Interval (5–8)	−0.57	(0.19)
Interval (9–12)	−0.36	(0.20)
Interval (13–16)	—	
Gender (F)	−0.13	(0.15)
Gender (M)	—	
Age (<50)	—	
Age (50–69)	2.48	(0.28)
Age (⩾70)	3.53	(0.29)
Diag	1.62	(0.33)
Diag×age (<50)	—	
Diag×age (50–69)	−1.35	(0.38)
Diag×age (⩾70)	−1.31	(0.46)

to increase over time ($F_{3,1068} = 7.65$, $p < 0.0001$). Older subjects and those with a psychiatric diagnosis have a significantly higher rate of mortality, but the association of psychiatric diagnosis report and mortality is significantly larger for younger subjects (Incidence rate ratio

[IRR] = exp(1.62) = 5.1, 95 per cent CI = 2.7–9.6) than for middle-aged subjects (IRR = 1.3, 95 per cent CI = 0.9–1.9) or older subjects (IRR = 1.4, 95 per cent CI = 0.7–2.6).

5. SOFTWARE FOR IMPLEMENTING REGRESSION MODELS

It is straightforward to fit regression models for multiple source data in Stata 8.0 while accounting for complex survey designs using the `svy` commands. We illustrate the commands required to produce the analysis of the ECCS data presented in Section 4.1. The analysis can be divided into several parts:

1. specifying the complex survey design,
2. fitting the regression, and
3. calculating contrasts of the regression parameters and testing specific hypotheses of interest.

Figure 1 displays the Stata syntax and output to specify the complex survey design for the example considered in Section 4.1. The `list` command can be used to display some or all of the observations; the values for three selected subjects are shown. The data set consists of two records per subject, corresponding to observations associated with the two sources (parent and teacher). For subjects 8 and 15, the source reports from parents and teachers were identical, however for subject 16, the reports were discordant.

The `svyset` command allows the specification of the survey design variables. Stata can incorporate a variety of designs through this mechanism. For example, if the data are weighted, but not clustered or stratified, the `strata` and `psu` statements can be left out, and the analysis can proceed. Multistage sampling designs are also supported. The `svydes` command provides a summary of the design including the number of PSU's per stratum, and distribution of observations within PSU's.

Stata supports a number of complex survey estimation commands (a comprehensive list is provided in Table IV). Figure 2 displays the syntax and output to fit the final regression model for this example (after non-significant source interactions were dropped). The `xi` command allows the specification of interactions in a flexible manner, though the names generated by Stata for the interaction terms are somewhat inelegant. The `svylogit` model is by default overparametrized, and redundant terms are dropped before the model is estimated.

To calculate the values reported in Table I, or to determine if non-significant source interactions can be dropped, some post-processing of the regression results is necessary.

Figure 3 displays the Stata code to calculate contrasts of the regression parameters. To calculate the intercept term for TEACHER in Table I based on the results from the regression model reported in Figure 2, the _CONS and TEACHER terms must be added. The `svylc` (or `lincom`) command can be used to calculate such linear combinations, along with the associated standard error for this function. In addition to this calculation, examples are given for the calculation of the log OR and OR for the HLTHPRO predictor for teachers (0.055 and 1.057, respectively).

In addition, there may be interest in performing tests of specific hypotheses regarding the parameters in the model.

Figure 4 displays the Stata code to carry out such tests. The AREA main effect was fit using two indicator variables (one for SUBURB, and one for CITY). To carry out multiple df

```
. list strata school childid intern teacher hlthpro

        strata    school    childid    intern   teacher   hlthpro
  1.       M01      1201         08         0         1         0
  2.       M01      1201         08         0         0         0
  3.       M01      1201         15         0         1         0
  4.       M01      1201         15         0         0         0
  5.       M01      1201         16         0         1         1
  6.       M01      1201         16         1         0         1
  (other output not displayed)

. svyset, psu (school)
. svyset [pweight=weight]
. svyset, strata (strata)
. svydes

pweight:  weight
Strata:   strata
PSU:      school
                                         #Obs per PSU
 Strata                           ----------------------------
strata     #PSUs     #Obs        min      mean       max
--------  --------  --------  --------  --------  --------
     M01         3       216        63      72.0        84
     M02         2       112        46      56.0        66
     M03         4       282        38      70.5       101
       R        47      1348         1      28.7       103
     S01         4       266        12      66.5       105
     S02         6       315         6      52.5        94
     S03         3        97        17      32.3        45
--------  --------  --------  --------  --------  --------
       7        69      2636         1      38.2       105
```

Figure 1. Stata commands and output to specify the complex survey design for the Eastern Connecticut Child Survey.

tests of an overall AREA effect, the `svytest` command with the `accumulate` option is used. As reported earlier, there was no significant AREA effect (df $=2$, $p=0.32$). For the models fit to the Stirling County Study example in Section 4.2, similar invocations of the `svytest` command were used to conduct multiple df tests of source effects, and to estimate IRR's.

6. DISCUSSION

The methods reviewed in this paper describe a principled approach to the incorporation of (often discordant) multiple source reports when fitting regression models using data from a complex sample survey. These methods have advantages over more *ad hoc* approaches that combine the reports, and allow formal assessment of whether covariate (e.g. risk factor) effects

Table IV. Stata commands for analysing survey data.

Command	Function
`svymean`	Estimation of population (and subpopulation) means
`svyprop`	Estimation of population proportions
`svyratio`	Estimation of population ratios
`svytotal`	Estimation of population totals
`svyregress`	Linear regression (for survey data)
`svyivreg`	Instrumental variables regression
`svyintreg`	Interval and censored regression
`svylogit`	Logistic regression
`svymlogit`	Multinomial logistic regression
`svyologit`	Ordered logistic regression
`svyprobit`	Probit models
`svyoprobit`	Ordered probit models
`svypoisson`	Poisson regression
`svynbreg`	Negative binomial regression
`svygnbreg`	Generalized negative binomial regression
`svyheckman`	Heckman selection model
`svyheckprob`	Probit estimation with selection
`svytab`	Two-way tables for survey data
`svylc`	Calculate estimates of parameters
`svytest`	Test hypotheses regarding parameters
`_robust`	Programmer's command (survey variance estimator)

vary according to the source. In addition, they allow the incorporation of individuals with possibly missing source reports in the joint analysis. An appealing feature of the proposed methods is that they can be implemented using existing, general purpose, statistical software. Although we have focussed on the application of these methods to child and adult psychopathology, surveys that include multiple source or informant data are commonly conducted in studies of the elderly and in health services research. Also, in both of the applications considered in this paper, there were only two sources. The methods described in Sections 2 and 3 can be extended in a natural and straightforward way to handle three or more source reports. As an example, Lash *et al.* [4] analysed five multiple source reports of comorbidity as predictors of tamoxifen usage in a group of women with breast cancer. With more than two source reports there is the potential for a proliferation of regression parameters if all possible source effects need to be incorporated. As a result, the power to detect source-related differences in covariate effects may be somewhat low when the number of sources exceed three or four. However, in principle, the methodology can be applied in the same fashion as we have described.

Of note, the proposed methods allow for the pooling of information from different sources when appropriate. For example, in the analysis of the Stirling County Study data, there were no significant source effects, so a single model that pooled information from physician and self-report was fit to these data. This joint analysis of both source reports resulted in smaller standard errors than those obtained from separate analyses of each source report. In the Eastern Connecticut Child Study, some of the covariate effects differed for the two sources and it

```
. xi:  svylogit intern teacher suburb city middle low momsing momstrs
hlthpro i.hlthpro*teacher graderep famstrs i.famstrs*boy boy
i.famstrs*teacher

i.hlthpro         _Ihlthpro_0-1       (naturally coded; _Ihlthpro_0 omitted)
i.hlth~o*teac~r   _IhltXteach_#       (coded as above)
i.famstrs         _Ifamstrs_0-1       (naturally coded; _Ifamstrs_0 omitted)
i.famstrs*boy     _IfamXboy_#         (coded as above)
i.fams~s*teac~r   _IfamXteach_#       (coded as above)

Survey logistic regression

pweight:  weight                                Number of obs    =      2636
Strata:   strata                                Number of strata =         7
PSU:      school                                Number of PSUs   =        69
                                                Population size  = 65353.985
                                                F(  14,     49)  =      9.45
                                                Prob > F         =    0.0000
------------------------------------------------------------------------------
      intern |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     teacher |    .798415   .2614177     3.05   0.003     .2758485    1.320982
      suburb |   .0146484   .1812493     0.08   0.936    -.3476639    .3769606
        city |   .1511786   .1071006     1.41   0.163    -.0629124    .3652697
      middle |   .1930796   .1435674     1.34   0.184    -.0939075    .4800667
         low |   .5563123   .3081134     1.81   0.076    -.0595975    1.172222
     momsing |  -.0800196   .2482151    -0.32   0.748    -.5761945    .4161553
     momstrs |   .7952221    .134978     5.89   0.000      .525405    1.065039
     hlthpro |   .8692232   .1593975     5.45   0.000     .5505922    1.187854
 _IhltXteac~1 |  -.8141739   .3649264    -2.23   0.029    -1.543651   -.0846965
    graderep |   .3974364   .1611943     2.47   0.016     .0752135    .7196593
     famstrs |   .8618834   .2510834     3.43   0.001     .3599749    1.363792
         boy |   .7765334   .2364629     3.28   0.002     .3038509    1.249216
  _IfamXboy_1 |  -.7343468   .2685819    -2.73   0.008    -1.271234   -.1974592
 _IfamXteac~1 |  -.4838623   .3138062    -1.54   0.128    -1.111152    .1434273
       _cons |  -3.078798   .2614573   -11.78   0.000    -3.601444   -2.556152
------------------------------------------------------------------------------
```

Figure 2. Stata commands and output to fit multiple source regression models for the Eastern Connecticut Child Survey.

was possible to quantify the magnitude of these differences. In the joint analysis of both source reports, the model fit to these data pooled information from both source reports to estimate certain covariate effects, while allowing estimation of source-specific effects for other covariates.

An alternate approach to the proposed regression methods is to combine the information across the different source reports prior to the analysis. For example, the arithmetic average

```
. svylc _cons + teacher

 ( 1)  teacher + _cons = 0.0
------------------------------------------------------------------------------
      intern |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -2.280383   .2805399    -8.13   0.000    -2.841174   -1.719592
------------------------------------------------------------------------------

. svylc hlthpro + _IhltXteach_1

 ( 1)  hlthpro + _IhltXteach_1 = 0.0
------------------------------------------------------------------------------
      intern |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0550494   .3143999     0.18   0.862    -.5734271    .6835259
------------------------------------------------------------------------------

. svylc hlthpro + _IhltXteach_1, or

 ( 1)  hlthpro + _IhltXteach_1 = 0.0
------------------------------------------------------------------------------
      intern | Odds Ratio   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.056593   .3321927     0.18   0.862     .5635906     1.98085
------------------------------------------------------------------------------
```

Figure 3. Stata commands and output to calculate contrasts of regression parameters for the Eastern Connecticut Child Survey.

```
. svytest suburb=0

Adjusted Wald test
 ( 1)  suburb = 0.0
       F(  1,    62) =    0.01
            Prob > F =    0.9358

. svytest city=0, accumulate

Adjusted Wald test
 ( 1)  suburb = 0.0
 ( 2)  city = 0.0
       F(  2,    61) =    1.16
            Prob > F =    0.3195
```

Figure 4. Stata commands and output to calculate tests of specific hypotheses for the Eastern Connecticut Child Survey.

of the multiple source reports ($Y_i^* = (Y_{i1} + Y_{i2})/2$ for multiple source outcomes or $X_i^* = (X_{i1} + X_{i2})/2$ for multiple source predictors) provides a single number summary that can be used for subsequent analysis. This strategy will be somewhat more appealing when the source data are quantitative. However, we note that regression models for the combined reports (Y_i^* or X_i^*) can be shown to be special cases of the regression models for the joint reports presented in Section 2 [25]. However, a potential disadvantage of this arithmetic mean pooling approach is that it must often make a number of *a priori* assumptions. When the corresponding analysis is expressed in terms of the regression models for the joint source reports, these assumptions are testable from the data at hand. In addition, when there are missing source reports the arithmetic mean pooling approach uses a form of 'mean imputation' for the missing source reports and can yield biased estimates of the regression parameters and standard errors when the data are missing at random [25]. In contrast, the regression models for the joint source reports make full use of all available information and likelihood-based methods for multivariate normal outcomes yield valid estimates of the regression parameters when missing reports are missing at random. As a result, when the arithmetic mean pooling approach is deemed to be appropriate we recommend that it should be implemented using the regression models for the joint source reports presented in Section 2.

We have reviewed one approach for the analysis of multiple source data using marginal regression models. Alternative approaches that may be considered include the use of multilevel random effects (e.g. Rabe-Hesketh *et al.* [57], or Longford [58]), latent variable [59–61], or measurement error models [62]. In addition, structural equation models could be utilized to address estimation in this setting [63, 64]. Of note, in many of these alternative approaches the target of inferences is somewhat different than in the marginal regression models considered in this article. While each of the alternative methods has merits, it is beyond the scope of this tutorial article to provide a detailed comparison among alternative approaches.

One practical difficulty in analysing multiple source reports is that there is often a substantial amount of missing data. Multiple source reports are commonly missing since, by definition, they are collected from sources other than the primary subject of the study. Multiple stages of informed consent and willingness to participate often lead to a large proportion of the subjects having missing data from one or more sources. Neither pooling strategies nor separate analyses have addressed the potential bias resulting from missing data in this setting. Missingness can induce bias as well as loss of inferential efficiency [65, 66]. The methods discussed in this tutorial can incorporate incomplete observations when missingness is due to a process that is 'completely at random' [65]. It is straightforward to incorporate weights into the analysis that account for missingness that is 'at random' (related to observed quantities) [55, 67–69], a less restrictive assumption. In particular, when missingness is by design (e.g. for two-stage studies where incomplete observations are due to design decisions in the study) these methods are particularly attractive. A number of researchers have considered estimation in such settings [44, 45, 70].

Finally, we note that the methods described in Section 3 could be implemented using the generalized estimating equations approach [35], where the 'cluster' is the PSU and the known sampling weights are incorporated in the estimating equations. Use of the generalized estimating equations approach, with a working independence correlation structure, will produce identical estimates of the regression parameters. However, use of the empirical variance estimator, ignoring the stratification, will result in a less efficient estimator of the variances [70]. This increased variability, particularly when the stratification is done to reduce the variances of the sample estimates, may adversely affect the coverage probability of confidence intervals constructed from the empirical variance estimates.

ACKNOWLEDGEMENTS

We thank Gwen Zahner for use of data from the Eastern Connecticut Child Survey, which was conducted under contract to the Connecticut Department of Children and Youth Services. We also thank Jane Murphy for use of data from the Stirling County Study. We are grateful for the support provided by NIMH grant R01-MH54693.

REFERENCES

1. Verhulst FC. Recent developments in the assessment and diagnosis of child psychopathology. *European Journal of Psychological Assessment* 1995; **11**(3):203–212.
2. Kendler KS, Silberg JL, Neale MC, Kessler RC, Heath AC, Eaves LJ. The family history method: whose psychiatric history is measured? *American Journal of Psychiatry* 1991; **148**:1501–1504.
3. Silverman JM, Breitner JCS, Mohs RC, Davis KL. Reliability of the family history method in genetic studies of Alzheimer's disease and related dementia. *American Journal of Psychiatry* 1986; **143**(10):1279–1282.
4. Lash TL, Thwin SS, Horton NJ, Guadagnoli E, Silliman RA. Multiple informants: a new method to assess breast cancer patients' comorbidity. *American Journal of Epidemiology* 2003; **157**(3):249–257.
5. Rosner B, Gore R. Measurement error correction in nutritional epidemiology based on individual foods, with application to the relation of diet to breast cancer. *American Journal of Epidemiology* 2001; **154**(9):827–835.
6. Espeland MA, Kumanyika S, Wilson AC, Reboussin DM, Easter L, Robertson J, Brown WM, McFarlane M. TONE cooperative research group. Statistical issues in analyzing 24-hour dietary recall and 24-hour urine collection data for sodium and potassium intakes. *American Journal of Epidemiology* 2001; **154**(10): 996–1006.
7. Hundley V, Penney G, Fitzmaurice A, van Teijlingen E, Graham W. A comparison of data obtained from service providers and service users to assess the quality of maternity care. *Midwifery* 2002; **18**(2):126–135.
8. Leape LL, Brennan TA, Laird NM, Lawthers AG, Localio AR, Barnes BA, Hebert L, Newhouse JP, Weiler PC, Hiatt H. The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. *New England Journal of Medicine* 1991; **324**(6):377–384.
9. Zahner GEP, Daskalakis C. Factors associated with mental health, general health and school-based service use for psychopathology. *American Journal of Public Health* 1997; **87**(9):1440–1448.
10. Horton NJ, Saitz R, Laird NM, Samet JH. A method for modeling utilization data from multiple sources: application in a study of linkage to primary care. *Health Services and Outcomes Research Methodology* 2002; **3**:211–223.
11. Young AS, Sullivan G, Burnam MA, Brook RH. Measuring the quality of outpatient treatment for schizophrenia. *Archives of General Psychiatry* 1998; **55**(7):611–617.
12. Rosenheck RA, Desal R, Steinwachs D, Lehman A. Benchmarking treatment of schizophrenia: a comparison of service delivery by the national government and by state and local providers. *Journal of Nervous and Mental Disease* 2000; **188**(4):209–216.
13. Wang PS, Demler O, Kessler RC. Adequacy of treatment for serious mental illness in the United States. *American Journal of Public Health* 2002; **92**(1):92–98.
14. Reinherz HZ, Giaconia RM, Pakiz B, Silverman AB, Frost Ak, Lefkowitz ES. Psychosocial risks for major depression in late adolescence: a longitudinal community study. *Journal of the American Academy of Child Adolescent Psychiatry* 1993; **32**(6):1155–1163.
15. Zahner GEP, Pawelkiewicz W, DeFrancesco JJ, Adnopoz J. Children's mental health service needs and utilization patterns in an urban community. *Journal of the American Academy of Child Adolescent Psychiatry* 1992; **31**:951–960.
16. Zahner GEP, Jacobs JH, Freeman DH, Trainor K. Rural-urban child psychopathology in a northeastern U.S. state: 1986–1989. *Journal of the American Academy of Child Adolescent Psychiatry* 1993; **32**:378–387.
17. Leckman JF, Sholomskas D, Thompson WD, Belanger A, Weissman MM. Best estimate of lifetime psychiatric diagnosis. *Archives of Generali Psychiatry* 1982; **39**:879–883.
18. Bird HR, Canino G, Rubio-Stipec M, Gould MS, Ribera J, Sesman M, Woodbury M, Huertas-Goldman S, Pagan A, Sanchez-Lacay A, Moscoso M. Estimates of the prevalence of childhood maladjustment in a community survey in Puerto Rico. *Archives of General Psychiatry* 1988; **45**:1120–1126.
19. Piacentini JC, Cohen P, Cohen J. Combining discrepant diagnostic information from multiple sources: are complex algorithms better than simple ones? *Journal of Abnormal Child Psychology* 1992; **20**(1):51–63.
20. Boyle MH, Offord DR, Hofmann HG, Catlin GP, Byles JA, Cadman DT, Crawford JW, Links PS, Rae-Grant NI, Szatmari P. Ontario child health study. I. Methodology. *Archives of General Psychiatry* 1987; **44**:826–831.
21. Fitzmaurice GM, Laird NM, Zahner GEP, Daskalakis C. Bivariate logistic regression analysis of childhood psychopathology ratings using multiple informants. *American Journal of Epidemiology* 1995; **142**(11): 1194–1203.

22. Fitzmaurice GM, Laird NM, Zahner GEP. Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association* 1996; **91**(433):99–108.
23. Daskalakis C, Laird NM, Lipsitz SR. Simultaneous risk factor and agreement analyses for multiple-source categorical outcomes, submitted.
24. Kuo M, Mohler B, Raudenbush SL, Earls FJ. Assessing exposure to violence using multiple informants: application of hierarchical linear model. *Journal of Child Psychology and Psychiatry* 2000; **41**(8):1049–1056.
25. Goldwasser MA, Fitzmaurice GM. Multivariate linear regression of childhood psychopathology using multiple informant data. *International Journal of Methods in Psychiatric Research* 2001; **10**(1):1–10.
26. Kraemer HC, Measelle JR, Ablow JC, Essex MJ, Boyce WT, Kupfer DJ. A new approach to integrating data from multiple informants in psychiatric assessment and research: mixing and matching contexts and perspectives. *American Journal of Psychiatry* 2003; **160**(9):1566–1577.
27. Horton NJ, Laird NM, Zahner GEP. Use of multiple informant data as a predictor in psychiatric epidemiology. *International Journal of Methods in Psychiatric Research* 1999; **8**:6–18.
28. Glynn RJ, Rosner B. Comparison of alternative regression models for paired binary data. *Statistics in Medicine* 1994; **13**:1023–1036.
29. Graubard BI, Korn EL. Regression analysis with clustered data. *Statistics in Medicine* 1994; **13**:509–522.
30. Ananth CV, Preisser JS. Bivariate logistic regression: modelling the association of small for gestational age births in twin gestations. *Statistics in Medicine* 1999; **18**:2011–2023.
31. Pepe MS, Whitaker RC, Seidel K. Estimating and comparing univariate associations with application to the prediction of adult obesity. *Statistics in Medicine* 1999; **18**:163–173.
32. McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman & Hall: London, 1989.
33. Huber PJ. The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967; 221–233.
34. Horton NJ, Lipsitz SR. Review of software to fit generalized estimating equation (GEE) regression models. *The American Statistician* 1999; **53**:160–169.
35. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
36. Särndal C-E, Swensson B, Wretman J. *Model Assisted Survey Sampling*. Springer: New York, 1992.
37. Dunn G. Complex surveys. *Statistical Methods in Medical Research* 1996; **5**:213.
38. Brick JM, Kalton G. Handling missing data in survey research. *Statistical Methods in Medical Research* 1996; **5**:215–238.
39. Pfeffermann D. The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research* 1996; **5**:239–261.
40. Graubard BI, Korn EL. Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research* 1996; **5**:263–281.
41. Rust KF, Rao JNK. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* 1996; **5**:283–310.
42. LaVange LM, Stearns SC, Lafata JE, Koch GG, Shah BV. Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research* 1996; **5**:311–329.
43. Dunn G. *Statistics in Psychiatry*. Edward Arnold Publishers Ltd: London, 2000.
44. Pickles A, Dunn G, Vázquez-Barquero JL. Screening for stratification in two-phase ('two-stage') epidemiological surveys. *Statistical Methods in Medical Research* 1995; **4**:73–89.
45. Vázquez-Barquero JL, Garcia J, Artal Simón J, Iglesias C, Montejo J, Herrán A, Dunn G. Mental health in primary care: an epidemiological study of morbidity and use of health resources. *British Journal of Psychiatry* 1997; **170**:529–535.
46. Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics* 2000; **56**:645–646.
47. Achenbach TM. *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Department of Psychiatry, University of Vermont, 1991.
48. Achenbach TM. *Manual for the Teacher's Report Form and 1991 Profile*. Department of Psychiatry, University of Vermont, 1991.
49. Boyle MH, Offord DR, Racine YA, Fleming JE, Szatmari P, Links PS. Predicting substance use in early adolescence based on parent and teacher assessments of childhood psychiatric disorder: results from the Ontario child health study follow-up. *Journal of Child Psychology and Psychiatry* 1993; **34**(4):535–544.
50. StataCorp. *Stata Statistical Software*: *Release 8.0*. Stata Corporation: College Station, TX, 2003.
51. Hughes CC, Tremblay MA, Rapoport RN, Leighton AH. *People of Cove and Woodlot*: *the Stirling County Study of Psychiatric Disorder and Sociocultural Environment*, vol. II. Basic Books, Inc.: New York, 1960.
52. Leighton DC, Harding JS, Macklin DB, Leighton AM. *The Character of Danger*: *The Stirling County Study of Psychiatric Disorder and Sociocultural Environment*, vol. III. Basic Books, Inc.: New York, 1963.
53. Murphy JM, Monson RR, Olivier DC, Sobol AM, Leighton AH. Affective disorders and mortality: a general population study. *Archives of General Psychiatry* 1987; **44**:473–480.
54. Murphy JM, Monson RR, Olivier DC, Sobol AM, Pratt LA, Leighton AH. Mortality risk and psychiatric disorders. *Social Psychiatry and Psychiatric Epidemiology* 1989; **24**:134–142.

55. Horton NJ, Laird NM, Murphy JM, Monson RR, Sobol AM, Leighton AH. Multiple informants: mortality associated with psychiatric disorders in the Stirling County Study. *American Journal of Epidemiology* 2001; **154**(7):649–656.
56. Laird NM, Olivier D. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* 1981; **76**:231–240.
57. Rabe-Hesketh S, Yang S, Pickles A. Multilevel models for censored and latent responses. *Statistical Methods in Medical Research* 2001; **10**(6):409–427.
58. Longford NT. Multilevel analysis with messy data. *Statistical Methods in Medical Research* 2001; **10**(6): 429–444.
59. Dunn G, Everitt B, Pickles A. *Modelling covariances and latent variables using EQS*. Chapman & Hall Ltd: London, 1993.
60. S-Y Lee, J-Q Shi. Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data. *Biometrics* 2001; **57**(3):787–794.
61. Landrum MB, Normand S-LT, Rosenheck R. Selection of related multivariate means: monitoring psychiatric care in the Department of Veterans Affairs. *Journal of the American Statistical Association* 2003; **98**:7–16.
62. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models*. Chapman & Hall: London, 1995.
63. Bollen KA. *Structural Equations with Latent Variables*. Wiley: New York, 1989.
64. Hoyle R. *Structural Equation Modeling*: *Concepts*, *Issues and Applications*. Sage Publications: Thousand Oaks, CA, 1995.
65. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
66. Zhao LP, Lipsitz SR, Lew D. Regression analysis with missing covariate data using estimating equations. *Biometrics* 1996; **52**:1165–1182.
67. Xie F, Paik MC. Generalized estimating equation model for binary outcomes with missing covariates. *Biometrics* 1997; **53**:1458–1466.
68. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846–866.
69. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
70. Williams P, Ryan L. Design of multiple binary outcome studies with intentionally missing data. *Biometrics* 1996; **52**:1498–1514.
71. LaVange LM, Keyes LL, Koch GG, Margolis PA. Application of sample survey methods for modelling ratios to incidence densities. *Statistics in Medicine* 1994; **13**:343–355.