Westat®

Improving lives through research™

Literature Search on Combining Survey and Administrative Records

Task Order 2, BLS BPA 1625DC-17-A-0001

Authors

Lou Rizzo
J. Michael Brick

December 7, 2017

Prepared for:
Bureau of Labor Statistics
2 Massachusetts Ave. NE
Washington, DC 20212

Prepared by:
Westat
An Employee-Owned Research Corporation®
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Table of Contents

Chapter			Page
1	Introduction: Lohr and Raghunathan (2017)		1-1
2	Overview of Macro Approaches		2-1
	2.1	Calibration Approaches	2-1
	2.2	Multiple Frame Approaches	2-2
	2.3	Small-Area Estimation Basic Area Models	2-3
	2.4	Hierarchical Models for Combining Data Sources	2-4
3	Overview of Micro Approaches		3-1
	3.1	Linking Information from Individual Records	3-1
		3.1.1 PRL: Methods and Models	3-2
		3.1.2 PRL: Imperfect Links and Potential Biases	3-3
	3.2	Data Fusion	3-4
	3.3	Imputation	3-5
4	Conclusions and Review in the Context of the OSHA/SOII Application		4-1
	4.1	Review of Macro Approaches in the Context of the OSHA/SOII Application	4-1
	4.2	Review of Micro Approaches in the Context of the OSHA/SOII Application	4-3
	4.3	Overall Review of Approaches	4-4
R	References		R-1

Appendixes

A	Detailed Review of Papers on Calibration and Composite Estimation		A-1
B	Detailed Review of Papers on Multiple Frame Methods		B-1
	B.1	Combining Cellphone and Landline Frames	B-1
	B.2	National Crime Victimization Survey	B-3

Appendixes (continued) **Page**

C Detailed Review of Papers on Small-Area Area-Level Estimation C-1

C.1 NHIS and BRFSS: Reweighting Approach C-1
C.2 NHIS and BRFSS: Small-Area Estimation Approach C-2
C.3 Kim, Park, and Kim: Small-Area Estimation Combining Surveys with Fixed Bias C-4
C.4 Small-Area Estimation Theory: General C-5

D Detailed Review of Papers on Hierarchical Models for Combining Surveys D-1

D.1 Bayesian Models at the Estimate Level: No Covariates D-1
D.2 Bayesian Models at the Estimate Level: Corn Yields from the National Agricultural Statistical Services. D-3

E Detailed Review of Papers on Data Linkage E-1

E.1 Data Linkage:Traditional Methods and Mixture Models E-1
E.2 Data Linkage: Bayesian Methods E-3
E.3 Adjusting for Imperfect Data Linkage E-6

F Detailed Review of Papers on Data Fusion F-1

F.1 Data Fusion: Moriarity and Scheuren (2001) F-1
F.2 Data Fusion: Rässler (2002) F-2

G Detailed Review of Papers on Imputation G-1

G.1 Imputation: Self-Reports and Clinical Data from NHIS and NHANES. G-1
G.2 Imputation: Incomplete Data on Hospice-Use from the CanCORS. G-2
G.3 Imputation: Combining Cancer Registry Data with Followup Survey Data G-3
G.4 Mass Imputation: Combining Complex Surveys by Imputing to the Full Population G-4

Contents (continued)

Table		Page
G-1.	Table 1 from He et al. (2014): 3,027 CanCORS lung and colorectal cancer patients who died within 15 months of diagnosis.	G-2

1 Introduction: Lohr and Raghunathan (2017)

Lohr and Raghunathan (2017) is an excellent recent review paper on the subject of combining survey data with other sources, which corresponds exactly with the issue at hand (combining SOII and OSHA). Lohr and Raghunathan outline several general approaches for combining survey from multiple studies in their Sections 2 through 8. These can be divided into 'macro' approaches (approaches which combine or adjust estimates at a high level), 'micro' approaches (approaches which link individual records in some way, and 'amalgamate' the sources), and combinations of macro and micro. These basic categories will inform the literature review. Section 2 provides an overview of the macro approaches and Section 3 the micro approaches. The Appendices provide details of the individual articles relevant to each topical subsection. Section 4 provides a discussion of all of this material in the context of BLS's needs for the OSHA/SOII initiative.

2 Overview of Macro Approaches

The macro approaches, also called 'basic area level approaches' in Rao (2003), are as follows:

- Calibration;
- Multiple frame methods;
- Small-area estimation basic area level models; and,
- Hierarchical models.

These are discussed in Sections 2.1 through 2.4 respectively.

2.1 Calibration Approaches

Calibration is a methodology in which one study provides 'control totals' for particular population domains (at a high level of quality: very low or zero mean squared error), which are used to calibrate estimates from the other study. This is ideal for a situation in which one study (the 'gold standard study') is of the highest quality, but only provides estimates for a limited set of characteristics (in the classic case, this would be population totals for particular domains). The other study (the 'main study') provides a much more extensive amount of data, but has higher sampling variance and possibly other sources of survey error (biases). Calibration can reduce the variance for any characteristics from the main study which are correlated to variables available in the gold standard study, and can also successfully reduce biases in the estimates from the main study as well. Calibration is used extensively in survey practice, and there is an extensive literature regarding its properties. Section 2 of Lohr and Raghunathan (2017) provide an overview of calibration and a set of references.

Appendix A provides details of two references from Merkouris (2004, 2010). These are directly relevant to the OSHA/SOII application, as they apply to multiple surveys of the same population that need to be reconciled. In his first paper, Merkouris provides a composite estimator of the two surveys, with calibration of each to independent control totals, with the additional constraint that

they 'calibrate to each other': they are required to match estimates for totals for particular common items. In the OSHA/SOII context, the OSHA and SOII estimates could be calibrated to each other. His second paper then expands to the case of domain estimation, comparing schemes which utilize different levels of domain-specific information. A strong point of this approach is that it is easy to implement. A weak point is that it doesn't explicitly deal with bias: it only corrects biases that can be fixed by calibrating to control totals. The OSHA data is likely to have more serious biases that cannot be so easily corrected.

2.2 Multiple Frame Approaches

Multiple frame approaches are similar to the composite estimation and calibration approach described in the previous section, but it does not assume that both frames cover the full population. Generally, the union of the frames is assumed to cover the full population.

The classic multiple frame method has two independent samples from two frames A and B respectively, with the population divided conceptually into three mutually exclusive, exhaustive domains a, b, and ab (ab is the intersection of A and B, a contains elements only in A, and b contains elements only in B). Domain a is covered by the A sample, domain b by the B sample, and a simple linear combination between the two samples provides an optimal estimator for domain ab. The weights allocated to each sample in this linear combination may be based on relative precision, or on other factors. In some cases, both samples may be viewed as unbiased estimators of the domain ab, in which case relative precision is based on sampling variance alone. In other cases, the two sample estimators have differing expectations due to mode effects, differential nonresponse, etc., which adds bias considerations to the relative precision. In the OSHA/SOII context, SOII would cover the whole universe with the OSHA component only covering a portion of the universe (so that B is actually a subset of A, and ab is equal to B). We need also to view the OSHA component as having a bias due to mode-type effects. Section 5 of Lohr and Raghunathan (2017) give an overview of multiple frame methods and a set of references.

Appendix B provides a detailed review of papers by Lohr (2011), Brick et al. (2011), and Lohr and Brick (2012, 2014). The Brick et al. (2011) papers covers the application of combining landline and cellphone samples in telephone surveys. Lohr and Brick (2012) present an application from the National Crime Victimization Survey. These papers go beyond the Merkouris papers in that they allow for the modelling of bias directly. The biases in the Brick et al. (2011) paper are generated

from nonresponse differentials. The Lohr and Brick (2012) paper relating to the National Crime Victimization Survey allow for mode biases as well as biases from nonresponse differentials. Biases are difficult to estimate without auxiliary information of some kind: in the Lohr and Brick (2012) paper the leverage to estimate biases is gained by shrinking domain bias estimates to national-level bias estimates, assuming that mode effects are equal (or close to equal) across domains. Lohr and Brick (2012) rely on a non-Bayesian random effects model, generating Empirical Bayes type shrinkage estimates.

2.3 Small-Area Estimation Basic Area Models

Small-area estimation basic area models (Type A models in the terminology of Rao (2003)) are similar to multiple frame methods in that for any particular domain they may be based on two estimators which are linearly combined using weights which are based on relative precision. One of these estimators may be an unbiased estimator, but with high variance, and the other may be biased but with much lower variance. Small-area estimation differs from classic multiple frame methods in its reliance on an explicit model regarding the estimated values. The low-variance estimator is model-design unbiased: unbiased if the model is true. Model misspecification will lead to bias. The original paper on direct small-area estimation is Fay and Herriot (1979). The first part of Section 6 of Lohr and Raghunathan (2017) provide an overview of small area methods. If credible models can be developed linking OSHA estimates to SOII estimates, then a small-area estimation direct approach could be an option.

Appendix C has a detailed reviews of two papers regarding county-level estimates of cancer risk prevalence values based on a low-bias national survey (the National Health Interview Survey: NHIS) and a high-bias survey with sufficient sample sizes at the county level (the Behavior Risk Factor Surveillance System: BRFSS): Elliot and Davis (2005) and Raghunathan (2007). The source of bias in BRFSS originates in the population imbalance induced from relying on landline-only telephone samples and due to low response rates.

Elliot and Davis (2005) deal with BRFSS's coverage biases in an innovative way: weighting factors are attached to the BRFSS sample weights which ultimately adjust the BRFSS estimates to NHIS estimates at the regional level. An important aspect of this is again being able to apply regionally-based adjustments at the county level (particular critical conditional probabilities are assumed to be equal across counties). This approach works to adjust the biases, but at the cost of variance inflation

from the weighting adjustments. This variance inflation is large enough to force Elliot and Davis to develop hybrid estimators to minimize mean-squared-error (not relying entirely on the bias-adjusted, but high-variance, weight-adjustment estimates).

Raghunathan et al. (2007) addresses the same problem with a more standard small-area estimation framework. The basic building block is a county-level vector of NHIS telephone household prevalence, NHIS non-telephone household prevalence, and BRFSS telephone household prevalence. The Raghunathan researchers differ from Elliot and Davis in having access to county-level identification information for NHIS, allowing the construction of county-level estimates (Elliot and Davis worked with the NHIS public-use file, which allows for only construction of regional estimates). The two NHIS estimates have high sampling variability at the county level (and sometimes samples do not exist at all), but no assumed bias. The bias in the BRFSS estimates is addressed directly as a parameter at the county level and its variability is partially explained via county-level information. This allows for credible bias estimation. The Raghunathan approach is a full-scale Bayesian approach (with hierarchical priors).

Appendix C also provides some information about a theoretical paper Ybarra and Lohr (2008) which deal with the issue of measurement error in the covariates used in the small-area estimation model (a generally overlooked issue).

Kim, Park, and Kim (2015) provide a similar small-area estimation approach for the case in which one has a 'gold-standard' measurement from a small survey nested within a larger survey which has flawed measurement for which bias has to be allowed for. Estimation is done for a large set of small-area domains. They develop a non-Bayesian theory which is comprehensive, and leads to Empirical Bayes type estimators. The paper makes a fairly strong assumption that the bias from the measurement error is constant across domains. If this is judged to be a reasonable assumption finally in the OSHA/SOII application, then this theory will be very helpful. Otherwise the more complicated models that allow for differential biases will be needed.

2.4 Hierarchical Models for Combining Data Sources

Hierarchical models for combining data sources arises from the meta-analysis paradigm. Under this paradigm, many studies are essentially estimating the same quantity. The models developed for bringing these estimates together are similar to multiple frame models and small-area basic area level

models, but have a more explicit Bayesian approach. The regression-type models defining the means from the various studies are simpler than many small-area basic area models, but with a random-effects assigned for the means. It also is a Bayesian hierarchical model with defined prior distributions. Section 7 of Lohr and Raghunathan (2017) discuss recent work in this area.

Appendix D presents details about Manzi et al. (2011), which provides a Bayesian model for combining small area smoking prevalence estimates in 48 Local Areas in Eastern England. All of the various estimators which are combined together are assumed to have biases associated with them. The primary 'leverage' in getting at biases in this application is that the 48 Local Areas should aggregate to a UK General Household Survey prevalence estimate for Eastern England which is assumed to be unbiased. Manzi et al. also develop a non-Bayesian two-way ANOVA type approach leading to Empirical Bayes which is similar to the Lohr and Brick (2012) model, except that Lohr and Brick assume that one of their component estimates is unbiased. In the Manzi et al. application, the unbiased benchmark is the General Household Survey overall regional prevalence estimate.

Appendix D also has a similar application from the National Agricultural Statistical Services (NASS) for US Corn Yields. Three different programs estimate corn yields: (1) a probability sample of farms in which corn is measured by NASS personnel (an 'objective' survey), (2) a monthly interview sample of corn farmers with some coverage exclusion, and (3) a December national interview survey of corn farmers with no coverage exclusion. The three of these are put together using a Bayesian model. The December national interview survey is assumed to be unbiased. The objective survey appears to suffer from measurement-error bias and the monthly interview survey of corn farmers excludes large farms for burden reasons. Wang et al. (2011) provide a Bayesian approach for bringing the three surveys together for a national estimate. Leverage in estimating the biases in the two surveys assumed to have biases is gained by specifying a substantive process model which models true corn yield as a function of exogenous variables such as weather and amount of corn planted in the spring. Nandram et al. (2014) extend the Bayesian model to cover state-level domain estimates.

3 Overview of Micro Approaches

The micro approaches, also called 'basic unit level approaches' in Rao (2003), are as follows:

- Linking information from individual records;
  - Deterministic record linkage (DRL);
  - Probabilistic record linkage (PRL);
  - Data fusion;
- Imputation;

These are discussed in Sections 3.1 and 3.2 respectively.

3.1 Linking Information from Individual Records

Linking information from individual records between OSHA and SOII would be the basis of any credible unit-level approach, as unit-level data from the two surveys cannot be well-combined unless they are linked (identical records are recognized as being such). There is a burgeoning literature in this area. DRL is an exact link between records from two files based on a comprehensive set of matches. DRL may be possible in the OSHA/SOII context especially if the Employment Identification Number (EIN) was included in the OSHA data collection. PRL allows for uncertainty in the link. A similarity score is assigned that quantifies the degree to which two records may be the same record. There are also Bayesian versions of PRL. Section 3 of Lohr and Raghunathan (2017) discusses recent work in this area.

This section (and Appendix E) will be divided into two subsections, 3.1.1 and 3.1.2, which correspond respectively to portions of the recent literature. The first subsection will cover methods and models for PRL. The second subsection will deal with the related issue of evaluating the effect of PRL on analyses of the imperfectly linked data sets and developing methods for avoiding bias in analyses from inaccurate linkages.

3.1.1 PRL: Methods and Models

Appendix E.1 gives an overview of three references for data linkage: Christen (2012), Bohensky et al. (2010), and Winkler (2014). Christen is a book which provides a useful summary of current methods in actually accomplishing data linkage, including theoretical overviews, and discussion of currently available software. Winkler provides a further recent overview of current methods in data linkage. Winkler's paper overlaps much of the other material discussed in this literature search, so it is not discussed, but it is a good further summary. Bohensky et al. (2010) is a meta-review paper of data linkage as it is practiced in the medical research literature.

Winglee et al. (2005) provide a case study from the Medical Expenditure Panel Survey (MEPS). MEPS collects information about 'medical events' such as hospital stays annually from both household respondents (self-reports of medical events in the past period), and the medical providers for those same events. The linkage between these files is imperfect and using identifying fields such as event dates and other information about the medical events. From the Winglee et al. paper it appears that MEPS uses traditional probabilistic linkage methods (Fellegi and Sunter 1969) based on setting cutpoints for potential pairs based on the degree of matching using the linkage fields. Their paper is a discussion of research that they did to refine the definition of these cutpoints using methodology from Belin and Rubin (1995), and also simulations.

There are a number of recent papers that develop Bayesian models for linking data files in a probabilistic way. Appendix E.2 provides the details. Goldstein et al. (2012) presents theory for linking a 'File of Interest' (FOI) with a 'Linked data File' (LDF). The FOI is the primary file for which analysis is done, and the LDF provides auxiliary information. Each relevant record on the FOI is linked to multiple records on the LDF, with a vector of probabilities defining the posited probability that a particular LDF record links with the FOI record (the probabilities in the probability vector add to 1 for each FOI record). Goldstein et al. mixes this probability vector with a second factor which defines a multiple-imputation type distribution for the FOI record. In this sense, this Goldstein paper may also belong in Section 3.3 (or may exclusively belong there) because it mixes linking and multiple imputation methods to fill out the FOI file records.

Steorts et al. (2016) is a full-scale Bayesian approach to the process of linking LDF records to FOI records. They define latent random variables of 'true persons' and link the 'true persons' to records on the LDF and FOI files, with corresponding probabilities. They also define a latent random variable that determine if a particular field on a particular record is distorted. Distorted fields will

obscure links which are valid. Their complete Bayesian formulation of this whole structure generates posterior probabilities of linkages between records that can be used to make final linkages.

Gutman et al. (2013) provides another Bayesian model. Their application is for linking persons who have died in the United States, with cause of death on one file, and Medicare expenditures on the other file. They have strong blocking variables based on demographics and geography (creating blocks with small cells), but not strong matching within the blocks. They define as a random variable a linking vector within each block. The sample universe of these vectors define all possible pairs between the files (within the blocks). By fitting a Bayesian model with these vectors, the field values on each file, and a set of appropriate parameters, they implicitly are allowing for a wide range of pairings with implicit PRL probabilities assigned to each possible pair. The models then are fit directly.

Tancredi and Liseo (2015) and Gutman et al. (2015) present further related Bayesian models: the former paper for a regression model and the latter paper a survival analysis type model. In both cases, they draw the matching process and the model more fully together, so that the model itself informs the assignment of matching probabilities (those match pairs that agree more closely with the regression model are marginally favored). This may be less useful in the context of OSHA/SOII where many different analyses may be done with the final data set.

3.1.2 PRL: Imperfect Links and Potential Biases

There are a number of recent papers reported in Appendix E.3 where there is a (non-Bayesian) analysis of the effects of imperfect data linkage on analysis, and on ways of adjusting for this effectively. Lahiri and Larsen (2005) and Kim and Chambers (2012) present results for linear regression. Chipperfield et al. (2011) presents results for contingency tables and logistic regression. In all cases, the approach is to estimate a probability for the link that is used, and include this probability into the estimating equations for the analysis. Doing this is asserted as eliminating any biases from imperfect linkages. Kim and Chambers differs from the other two papers in designating an overall probability of mislinking based on exchangeability arguments, whereas the other two papers define distinct probabilities for each possible pair (though in estimating these, the probabilities may be the same for particular propensity estimation cells). Hof and Zwinderman (2015) present a more general likelihood model that defines in its likelihood all possible pairs between two files to be matched. The probabilities of linkage are incorporated in this way into the likelihood estimating equations. To estimate linkage probabilities Lahiri and Larsen use a mixture

model derived from an earlier paper by Larsen and Rubin (2001). Details on Larsen and Rubin's paper are in Appendix E.1. In all these cases, clerical review on a subset of pairs is the main leverage for estimating linkage probabilities.

3.2 Data Fusion

Unlike data linkage, data fusion proceeds without a direct link (deterministic or probability) between two pairs of records on the two files. Under data fusion, sets of records are matched between two data sources, allowing for the exploration of correlational relationships.

Appendix F presents a paper by Moriarity and Scheuren (2001) which provides an overview of data fusion (statistical matching). Their starting point is a paper by Kadane from 1978 which is reprinted with their 2001 paper. Kadane (2001) sets out a theoretical approach for the basic scenario of having one file with records with an **X** vector and a **Y** vector, and a second file with records with an **X** vector and a **Z** vector, and wanting to do analysis of the joint distribution of (**X**, **Y**, **Z**). If one is willing to make the strong assumption of conditional independence of **Y** and **Z** given **X**, it is easy enough to match a record ($\mathbf{x_1}$, $\mathbf{y_1}$), ($\mathbf{x_1}$, $\mathbf{z_1}$) on $\mathbf{x_1}$and create a synthetic record ($\mathbf{x_1}$, $\mathbf{y_1}$, $\mathbf{z_1}$). If there is a posited correlation between Y and Z conditional on X then creating a synthetic file for analysis is much more difficult. One needs to posit the correlation up front, and then synthetic files can be generated with this correlation built in. Moriarity and Scheuren criticize the Kadane approach as not succeeding in producing synthetic data sets that actually match the posited (Y, Z) correlations. Their approach produces random residuals for Y on the one file and for Z on the other file, and then matches on (X, Y, Z) using this augmented file. Simulation studies find that this matching approach provides a synthetic file that has all of the correct distributions. This 'right' matching approach doesn't seem too much different however than a mass imputation approach.

Rässler (2002) presents a further overview of statistical matching. In her monograph, she presents a solution to the statistical matching problem through Bayesian multiple imputation, treating the missing **Y** data in the Y-missing file and the missing **Z** data in the Z-missing file as missing data, assuming the missing blocks in the two files are data which are 'missing at random' conditional on **X**[1]. She assumes multivariate normality, and draws the equivalent of the **Y** and the **Z** vector for filling in the data using standard multivariate normal multiple imputation. The key difference with

[1] In Rässler's presentation, the data known on both files is **Z** (rather than **X**), and the missing data is **X** and **Y** (not **Y** and **Z**). To avoid confusion in the presentation, we've used the Moriarity and Scheuren notation, but note this if you wish to follow up by referencing Rässler's monograph.

standard multiple imputation with a single data set (and missing blocks of **Y** and **Z**) is that the conditional covariance between **Y** and **Z** is completely unknown, and has to be specified arbitrarily or drawn from a prior distribution for that parameter (in standard multiple imputation with one data set, the data set would provide data-based estimates of the conditional covariance which would feed into the computation of posterior distributions). Appendix E provides further details. Rässler's preferred multiple imputation approach differs very little from those described in Section 3.2 below, and her work on this approach could easily be included in that section.

3.3 Imputation

Imputation is primarily a tool for dealing with item-level missing data in a single survey, but can be applied in this context. The two surveys are viewed as a large single survey, with items on one survey but not the other considered as a species of item-level missingness. The large literature on imputation can then be utilized to provide models to allow for a 'filling-in of the item nonresponse', accomplishing in this way a fusion between the data sets. In cases where there is a solid linking between sources (e.g., DRL), and no mode differences, imputation can be a powerful and flexible instrument for achieving source fusion. The methodology is also applicable where PRL or Data Fusion is possible, but is less useful in the presence of mode differences between the two data sources, where the item outcomes must be viewed as being from differing population distributions. Section 4 of Lohr and Raghunathan (2017) discusses recent work in this area.

Appendix G details a number of papers that apply imputation methods in this context of combining surveys. Raghunathan (2006) and Schenker et al. (2010) present an application which is similar to the OSHA/SOII context using cancer prevalence items (smoking, obesity, diabetes). The National Health Interview Survey (NHIS) asks individuals for self-reports on these topics and those are assumed to have measurement error. The NHIS is a large nationally representative sample. The National Health and Nutrition Examination Survey (NHANES) has clinical measurements which are assumed to be without measurement error, but its sample is much smaller. Schenker et al. link the two studies by generating a model based on linking the NHIS self-report items to the clinical information on NHANES (the self-report items are also asked on the NHANES questionnaire). They apply this model to generate mass multiple imputations on the NHIS records (self-report to imputed clinical report). In this way, the bias from self-reporting on NHIS is estimated and can be eliminated (if the model is valid).

He et al. (2014) provide a second example regarding hospice use for lung and colorectal cancer patients from the Cancer Care Outcomes Research and Surveillance Consortium (CanCORS) sponsored by the National Cancer Institute (NCI) in 2003-05. There are two different sources for hospice use information: Medicare claims and medical records. Medicare claims are assumed accurate but only apply for age 65+ persons. Medical records can sometimes be incomplete about hospice use in particular. He et al. construct a Bayesian latent variable multiple imputation type approach to yield an overall estimate. They make the key assumption that hospice use is implied if either Medicare or medical records claim it, but missing records or 'no' in both does not necessarily mean no hospice use. They assume overreporting of hospice use by these sources is not possible: only underreporting. Appendix Section F.2 provides details. In the OSHA/SOII context, it may be possible to assume that overreporting of occupational injuries or sicknesses cannot occur, only underreporting: a positive report from either OSHA or SOII can be taken at face value. As He et al. argue in their context, this assumption facilitates identifiability.

Yucel and Zaslavsky (2005) present a similar example regarding the effect of chemotherapy in the two-year survival rate for colorectal cancer. The primay data source is the cancer registry for the state of California, which in principle is a census of all persons in California who are diagnosed with colorectal cancer (similar to the plan for OSHA for occupational injuries and illnesses). This registry suffers from underreporting of certain aspects of patients' records, in particular whether the patient received chemotherapy or not (as chemotherapy is done sometimes by other doctors and clinics than those responsible for the primary care of the cancer, and the report of it falls through the cracks, so to speak). The Quality of Cancer Care project was a validation study that drew a sample of patients in the California registry for followup with surveys of the presiding physicians. The physicians could provide more accurate information about the true course of therapy for the patient including whether or not the patient received chemotherapy.

The Yucel and Zaslavsky framework is thus similar to the OSHA/SOII framework in that there is a larger, and assumedly more complete database (OSHA/ state registry), with most of the population covered, but with inaccurate information, and a smaller study (SOII/Quality of Cancer Care project) with better data. In this case, there is no issue with linking: the Quality of Cancer Care sample units are drawn directly from the Cancer Registry. The validation study is fairly small (1,956 patients; 1,422 respondents).

Appendix G shows that the Yucel and Zaslavsky approach is a mass imputation approach: a model is developed using the sample data linked with the registry data, and this model is used to generate

mass multiple imputations for the registry data. He and Zaslavsky (2009) extend this approach to a multivariate vector of therapies (rather than a single therapy indicator).

Finally, Appendix G also gives a short description of Dong et al. (2014a, 2014b) which provides a theory of mass imputation to create a population from a complex survey data set. This may be of value in 'unraveling' a complex survey sample to compare to a population-based data set from the same frame. Normally the unit-level records from a complex survey sample shouldn't be used out of the context of the complex sample design from which they were derived. This approach allows one to pry away that context and get back to a population-type data set that has no sample design that needs to be respected and considered (at the cost of considerable effort).

4 Conclusions and Review in the Context of the OSHA/SOII Application

The literature provides two major branches in bringing two surveys such as OSHA and SOII together: a macro and a micro approach. Sections 4.1 and 4.2 will summarize these two approaches respectively, and Section 4.3 will provide further conclusions.

4.1 Review of Macro Approaches in the Context of the OSHA/SOII Application

The macro approach keeps the two surveys and their estimates separate, and builds a composite estimator from the separate survey estimates. This will generally entail a partitioning of the population universe into discrete strata. In some cases, one or the other surveys may not cover part of the population universe, in which case the estimator for that stratum will be from one of the surveys only. In other parts of the population universe where there is overlap, both surveys contribute to a composite estimator.

If both surveys are believed to be unbiased estimators, then the weights for each survey estimator in the composite estimator can be computed based on relative precision alone. The composite weights are proportional to the estimated precision of each estimator. Calibration methods can be used to benchmark to auxiliary sources of data to improve the precision of the composite estimator. For small domains, small-area estimation type models can be utilized to further bolster the estimator by borrowing strength across domains (using auxiliary information). The work in Ybarra and Lohr (2008) on adjusting for auxiliary information measurement error should be considered when the auxiliary information has measurement error on the same order of magnitude as the primary estimates, to provide accurate evaluations of the level of mean squared error.

If one of the two surveys is viewed to have bias, then simple precision-based composition is not possible. The OSHA/SOII application certainly falls into this category, as the OSHA estimates have to be suspected as having bias from measurement error (companys' failure to fully and accurately report on their occupational safety and health) and low response rates. In most cases in the literature (with a few notable exceptions such as Manzi et al. (2011)), one of the survey estimators (the

'primary survey' below) is viewed as being unbiased, with the other survey estimators (the 'secondary survey(s)' below) assumed to be potentially biased. This bias can arise from nonresponse, lack of coverage, or mode differences. In the OSHA/SOII application, the SOII estimate would be viewed as unbiased (not because it is exactly, but it is relatively unbiased after all adjustments are made for effects of SOII survey nonresponse, and is accepted then as such in composite estimation).

This bias is not easy to evaluate. A naïve estimator of it is the simple difference between the two survey estimates. But then all the data from the secondary survey really only goes into estimating the bias, and in reality the primary survey is carrying the load of estimating the population value. The bias has to be deconstructed in a way that allows the secondary survey to provide information to the estimation of the population value.

Lohr and Brick (2012) provide a random effects type approach. The primary survey estimator is assumed to be unbiased, and the secondary survey estimator has a bias which is estimated directly at the national level. Within subdomains, the biases are allowed to be different across subdomains, but the subdomain biases are shrunk to a national level bias estimate by random-effects Empirical Bayes type methods. These methods could be applied in the OSHA/SOII context. Raghunathan et al. (2007) provide a similar but more explicitly Bayesian approach, and provide a model which implicitly breaks down the secondary survey biases at the county level using demographic characteristics: a substantive explanatory model of the relevant bias levels. Wang et al. (2011) also provide a substantive explanatory model which allows for leverage in the measurement of bias levels in an explicitly Bayesian model. The general theme in the literature is that good models have to be developed to allow the necessary leverage to measure biases in secondary surveys. A Bayesian approach allows for the full application of this model-based approach, but random-effects Empirical Bayes is also used in current applications.

For the simpler methods such as the multiple frame methods, it may be possible to evaluate precision with no reference to the estimation variable. In this case, the composite estimation method can be implemented by assigning a single set of weights to all units in the multiple files comprising the composite estimation method. Most of the methods described in Appendices A and B can be done this way. Otherwise, if the parameters of the composite estimator depend on the estimation variable, then each estimation variable may have a separate composite estimator. There cannot be one set of universal weights. The methods described in Appendices C and D fall into this category. The form of the composite estimator varies across the estimation variables, so each estimation variable has to be run through the composite estimation system individually.

4.2 Review of Micro Approaches in the Context of the OSHA/SOII Application

The micro-level approaches relevant for the OSHA/SOII application can be divided into probabilistic linkage approaches, data fusion, and imputation approaches. Data fusion is an approach where data sets with partially overlapping field sets are merged (one data set has one set of fields, the other data sets another set of fields, with some overlapping fields that can be used in linking). Data fusion is suboptimal compared to probabilistic linkage, as one is forced to make strong prior assumptions about the correlation structure between the nonoverlapping fields sets. The relevant question items in OSHA and SOII have considerable overlap in this case, so data fusion with its deficiencies doesn't likely need to be considered, except possibly for limited items (or restricted strata with more limited reporting requirements).

Probabilistic linkage is a complex methodology for linking the records from two surveys covering the same population when deterministic linkage (linkage based on a gold standard common identifying field) is not available. There are a variety of different approaches to this in the literature. In the 'classical approach' going back to Fellegi and Sunter (1969), one estimates the probability of a link to a particular File Of Interest (FOI) record and then chooses the field from the Linked Data File (LDF) record with the highest probability of linking to the FOI record. There is a large literature (and software available) for carrying this out.

There are a number of references discussed in Section E.3 which accept a set of 'best' single probabilistic links between FOI records and LDF records, but attempt to account in regression analyses for the uncertainty inherent in the probabilistic link, by incorporating estimated linkage probabilities into the regression estimating equations (in some form).

A second, more theoretically ambitious set of references go beyond a single link and work directly with a vector of LDF records linking to each FOI record, with a vector of corresponding linkage probabilities. These can result in a multiple imputation type approach in which multiple imputations incorporate into each augmented FOI record a set of linked field information from multiple donors from the LDF file. Several references provide a fully Bayesian analysis which does not create a final set of multiple imputation files but completes a full analysis from a Bayesian standpoint.

Finally, the imputation methods discussed in Section 3.3 describe further more explicit imputation methods. The Schenker et al. (2010) application deals with measurement error: there are a small number of records from one survey that is a subsample from a larger survey with both gold-standard reports with no assumed measurement error and flawed reports with measurement error, and a larger number of records only with the flawed reports with measurement error. The gold standard in this case is a clinical report of medical status, and the flawed report is a self-report of the same medical status from the individual. The strong relationship between gold-standard reports and flawed reports in the smaller survey drive an imputation system for the larger survey that in effect fills in the gold-standard report for all records on the larger survey. This paradigm is not likely to be relevant for OSHA/SOII directly, as the presumption is that OSHA injury and illness reports will not be less accurate than SOII ones if they are in fact reported (the parallel between self-reports and clinical reports in the medical context is not really on target).

The other two papers He et al. (2014) and Yucel and Zaslavsky (2005) are more likely to be relevant as the matched surveys in these cases are both measuring the same incidence of the same event with the error in both cases being from missing values: a failure to report the event. In both cases, the researchers make the assumption that error can be in only one direction: if there is no event, then neither source will report the event, but if there is an event, then one or the other source (or both sources) will report it. This particular paradigm is likely to be directly relevant to the OSHA/SOII application. Both applications rely on Bayesian modeling to tease out the correct conditional probability distributions, backing out mass imputations to represent the right posterior distributions given the source information and the specified models.

4.3 Overall Review of Approaches

The micro approach will provide a better result than the macro approach if the data linkage is strong between the surveys, as it captures more information in linking data at the individual record level. The macro approach completely severs the two surveys and cannot measure the information provided by the correlations within these linked records.

The micro approach is certainly more difficult and costly to implement as it requires a careful combination of the two surveys. The macro approach can remain within the framework of each individual survey, and put together the estimates as a last step. Thus there is a cost/benefit analysis that is needed in deciding between micro and macro.

The micro approach's cost and benefit both depend on the quality of the link. A direct deterministic link will be easiest to implement and will also give the best results. The quality and complexity of probabilistic methods depend implicitly on how effective the linking is. As the quality of the links goes down, the benefit of the micro approach decreases, and the cost and complexity of the necessary modeling increases. It seems on the surface of things that the OSHA/SOII link has a very good chance of being solid enough to yield strong estimates.

References

Bohensky, M.A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D.V., Scott, I., and Brand, C.A. (2010). Data linkage: A powerful research tool with potential problems. *BMC Health Services Research,* 10, 346.

Brick, J.M., Flores Cervantes, I., Lee, S., and Norman, G. (2011). Nonsampling errors in dual frame telephone surveys. *Survey Methodology,* 37(1), 1-12.

Chipperfield, J.O., Bishop, G.R., and Campbell, P. (2011). Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data". *Survey Methodology* 37(1), 13-24.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection.* Berlin-Heidelberg: Springer.

Dong, Q., Elliott, M.R., and Raghunathan, T.E. (2014a). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology,* 40(1), 29-46.

Dong, Q., Elliott, M.R., and Raghunathan, T.E. (2014b). Combining information from multiple surveys. *Survey Methodology,* 40(2), 347-354.

Elliot, M.R., and Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys. *Applied Statistics,* 54(3), 595-609.

Fay, R.E. III, and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association,* 74, 269-277.

Fellegi, I., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association,* 64, 1183-1210.

Goldstein, H., Harron, K., and Wade, A. (2012). The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine,* 31, 3481-3493.

Gutman, R., Afendulis, C.C., and Zaslavsky, A.M. (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association,* 108, 34-47.

Gutman, R., Sammartion, C.J., Green, T.C., and Montague, B.T. (2015). Error adjustments for file linking methods using encrypted unique client identifier (eUCI) with application to recently released prisoners who are HIV+. *Statistics in Medicine,* 35, 115-129.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

He, Y., Landrum, M.B., and Zaslavsky, A.M. (2014). Combining information from two data sources with misreporting and incompleteness to assess hospice-use among cancer patients: A multiple imputation approach. *Statistics in Medicine,* 33, 3710-3724.

He, Y., and Zaslavsky, A.M. (2009). Combinng information from cancer registry and medical records data to improve analyses of adjuvant cancer therapies. *Biometrics,* 65, 946-952.

Hof, M.H.P., and Zwinderman, A.H. (2015). A mixture model for the analysis of data derived from record linkage. *Statistics in Medicine,* 34, 74-92.

Kadane, J.B. (2001). Some statistical problems in merging data files. *Journal of Official Statistics,* 17(3), 423-433. (Reprint of 1978 Department of Treasury report).

Kim, G., and Chambers, R. (2012). Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica,* 66, 64-79.

Kim, J.-K., Park, S., and Kim, S.-Y. (2015). Small area estimation combining information from several sources. *Survey Methodology,* 41(1), 21-36.

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association,* 100, 222-230.

Larsen, M.D., and Rubin, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association,* 96, 32-41.

Lohr, S.L. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology,* 37(2), 197-213.

Lohr, S.L., and Brick, J.M. (2014). Allocation for dual frame telephone surveys with nonresponse. *Journal of Survey Statistics and Methodology,* 2, 388-409.

Lohr, S.L., and Brick, J.M. (2012). Blending domain estimates from two victimization surveys with possible bias. *The Canadian Journal of Statistics,* 40(4), 679-696.

Lohr, S.L., and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science,* 32(2), 293-312.

Manzi, G., Spiegelhalter, D.J., Turner, R.M., Flowers, J., and Thompson, S.G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society A*, 174(1), 31-50.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association,* 99, 1131-1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society B,* 72(1), 27-48.

Moriarity, C., and Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics,* 17(3), 407-422.

Nandram, B., Berg, E., and Barboza, W. (2014). A hierarchical Bayesian model for forecasting state-level corn yield. *Environmental and Ecological Statistics,* 21, 507-530.

Raghunathan, T.E. (2006). Combining information from multiple surveys for assessing health disparities. *Allgemeines Statistisches Archiv,* 90, 515-526.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.L., David, W.W., Dodd, K.W., and Feuer, E.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association,* 102, 474-486.

Rao, J.N.K. (2003). *Small area estimation.* Hoboken, NJ: John Wiley & Sons.

Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. Lecture Notes in Statistics 168. New York: Springer.

Schenker, N., Raghunathan, T.E., and Bondarenko, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine,* 29, 533-545.

Steorts, R.B., Hall, R., and Fienberg, S.E. (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association,* 111*,* 1660-1672.

Tancredi, A., and Liseo, B. (2015). Regression analysis with linked data: Problems and possible solutions. *Statistica,* 57(1), 19-35.

Wang, J.C., Holan, S.H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2011). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics,* 17(1), 84-106.

Winglee, M., Valliant, R., and Scheuren, F. (2005). A case study in record linkage. *Survey Methodology,* 31(1) ,3-11.

Winkler, W.E. (2014). Matching and record linkage. *WIREs Computer Statistics,* 6, 313-325, doi: 10.1002/sics.1317.

Ybarra, L.M.R., and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika,* 95(4), 919-931.

Yucel, R.M., and Zaslavsky, A.M. (2005). Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association,* 100, 1123-1132.

Zieschang, K. (1990). Sample weighting methods and estimation totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association,* 85, 986-1001.

Appendix A

Detailed Review of Papers on Calibration and Composite Estimation

Appendix A
Detailed Review of Papers on Calibration and Composite Estimation

Merkouris (2004) discusses composite estimation with calibration under the framework that one has two independent surveys with independent control totals. In the simplest case, there are variables $\mathbf{X_1}$ from Survey 1 with control totals available, variables $\mathbf{X_2}$ from Survey 2 with control totals available, and a single variable *Z* which is present on both surveys for which the best estimator is desired. Merkouris presents a composite estimator of the single variable Z as follows (equation (8) in his paper):

$$\hat{Z}_s^{CR} = \varphi\hat{Z}_1 + (1-\varphi)\hat{Z}_2 + \widehat{\boldsymbol{\beta}}_{\mathbf{1}}\left[\varphi\left(\mathbf{t}_{\mathbf{x_1}} - \widehat{\mathbf{X}}_{\mathbf{1}}\right)\right] + \widehat{\boldsymbol{\beta}}_{\mathbf{2}}\left[(1-\varphi)\left(\mathbf{t}_{\mathbf{x_2}} - \widehat{\mathbf{X}}_{\mathbf{2}}\right)\right]$$

where $\hat{Z}_1$ and $\hat{Z}_2$ are the simple Horvitz-Thompson estimators of *Z* from Surveys 1 and 2 respectively, $\hat{\beta}_1$ and $\hat{\beta}_2$ are regressions of *Z* on $\mathbf{X_1}$ and *Z* on $\mathbf{X_2}$ within the two surveys respectively, $\mathbf{t}_{\mathbf{x_1}}$ and $\mathbf{t}_{\mathbf{x_2}}$ are control totals for $\mathbf{X_1}$ and $\mathbf{X_2}$ respectively, $\widehat{\mathbf{X}}_{\mathbf{1}}$ and $\widehat{\mathbf{X}}_{\mathbf{2}}$ are the Horvitz-Thompson estimators for $\mathbf{X_1}$ and $\mathbf{X_2}$ respectively, and φ is a compositing factor as follows:

$$\varphi = \frac{\tilde{n}_1}{\tilde{n}_1 + \tilde{n}_2}$$

$\tilde{n}_1$ and $\tilde{n}_2$ are effective sample sizes for *Z* from surveys 1 and 2 respectively. Essentially we are compositing two calibrated estimators of Z with calibration based on possibly differing sets of auxiliary vectors $\mathbf{X_1}$ and $\mathbf{X_2}$, which we assume are both unbiased (or their biases if any are corrected by the calibrations on $\mathbf{X_1}$ and $\mathbf{X_2}$), and the compositing factor is based on a proxy for relative variances.

Merkouris (2004) then generalizes this to a q-vector $\boldsymbol{Z}$ of common survey variables and generalizes φ to a q x q matrix Φ, with composite estimator then

$$\begin{aligned} \widehat{\boldsymbol{Z}}_s^{CR} &= \Phi\widehat{\boldsymbol{Z}}_1^R + \left(\mathrm{I}_q - \Phi\right)\widehat{\boldsymbol{Z}}_2^R \quad = \\ &= \widehat{\boldsymbol{Z}}_1^R + \left(\mathrm{I}_q - \Phi\right)\left(\widehat{\boldsymbol{Z}}_2^R - \widehat{\boldsymbol{Z}}_1^R\right) \\ &= \widehat{\boldsymbol{Z}}_2^R + \Phi\left(\widehat{\boldsymbol{Z}}_1^R - \widehat{\boldsymbol{Z}}_2^R\right) \end{aligned}$$

with I_q the q x q identity matrix, $\widehat{\boldsymbol{Z}}_1^R$ the calibrated estimator of $\boldsymbol{Z}$ from survey 1 (including the calibration to $\mathbf{t}_{\mathbf{x}_1}$ control totals) and $\widehat{\boldsymbol{Z}}_2^R$ the calibrated estimator of $\boldsymbol{Z}$ from Survey 2 (including the calibration to $\mathbf{t}_{\mathbf{x}_2}$ control totals). Note that the composite estimator can be rewritten in the form of $\widehat{\boldsymbol{Z}}_1^R$ with a calibration to $\widehat{\boldsymbol{Z}}_2^R$ (as if it is auxiliary information), or as $\widehat{\boldsymbol{Z}}_2^R$ to $\widehat{\boldsymbol{Z}}_1^R$ (as if it is auxiliary information). Another way of formulating this is that we have three control constraints on the final weights: the weights for Survey 1 have to guarantee that the final weighted estimator of $\mathbf{X_1}$ are exactly equal to the control totals $\mathbf{t}_{\mathbf{x}_1}$, the final weights for Survey 2 have to guarantee that the final weighted estimator of $\mathbf{X_2}$ is exactly equal to the control totals $\mathbf{t}_{\mathbf{x}_2}$, and the weights for the combined survey for the common variables $\boldsymbol{Z}$ have to force equality between the Survey 1 final estimator of $\boldsymbol{Z}$ and the Survey 2 final estimator of $\boldsymbol{Z}$. Merkouris (2004) points out as an example the work at BLS on the Consumer Expenditure Survey (CES) as given in Zieschang (1990), where the two surveys were the Diary and Interview components of the CES.

Merkouris (2010) then extends this work to domain estimation. Merkouris (2010) presents three different estimators for a common set of variables $\boldsymbol{Z}$ between two surveys with control totals $\mathbf{X_1}$ and $\mathbf{X_2}$ as in Merkouris (2004), with the estimation restricted to possibly small domain d. The ***first*** estimator calibrates to $\mathbf{t}_{\mathbf{x}_1}$ and $\mathbf{t}_{\mathbf{x}_2}$, and composites the $\boldsymbol{Z}$ domain estimates (effectively equating the Survey 1 and Survey 2 estimates within the domain d). The ***second*** estimator calibrates to $\mathbf{t}_{\mathbf{x_{1d}}}$ and $\mathbf{t}_{\mathbf{x_{2d}}}$: separate domain-$d$ control totals for $\mathbf{X_1}$ and $\mathbf{X_2}$, also compositing the Z domain estimates (again effectively equating the Survey 1 and Survey 2 estimates within the domain d). This second estimator requires the existence of domain-level auxiliary information $\mathbf{t}_{\mathbf{x_{1d}}}$ and $\mathbf{t}_{\mathbf{x_{2d}}}$: this may or may not be available. If it is available and is of sufficient quality, Merkouris (2010) showed that the second estimator will have higher precision than the first estimator (under certain conditions), as one might expect. The ***third*** estimator is similar to the first in that it calibrates to $\mathbf{t}_{\mathbf{x}_1}$ and $\mathbf{t}_{\mathbf{x}_2}$, at the full-population level only, and composites the $\boldsymbol{Z}$ domain estimates, but also adds as a control the domain population sizes. This third estimator should perform better than the first, but not as well as the second in general.

Appendix B

Detailed Review of Papers on Multiple Frame Methods

Appendix B
Detailed Review of Papers on Multiple Frame Methods

Lohr (2011) presents a synopsis of the literature on multiple frame methods going back to Hartley (1962). These methods allow for more than two frames, but the OSHA/SOII context is dual frame, so this literature review presentation will be restricted to that. The dual frame estimator of a population total Y using notation from Lohr is:

$$\hat{Y}(\theta) = \hat{Y}_a^A + \theta * \hat{Y}_{ab}^A + (1-\theta) * \hat{Y}_{ab}^B + \hat{Y}_b^B$$

The superscripts A and B refer to the two frames which together cover the full population. The subscripts a, b, and ab represent the parts of the population covered by Frame A only, Frame B only, and the part of the population covered by both frames, respectively. The parameter θ combines the estimators for population ab derived from Frame A and Frame B respectively.

Most of the literature on dual frame surveys developed from the original Hartley (1962) paper assumes that both population totals $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$ are unbiased, so that whatever the choice of θ, the final estimator $\hat{Y}(\theta)$ will be unbiased. θ can be selected in order to minimize the variance, or with other issues in mind. The OSHA/SOII application cannot make that assumption. Our interest is in an approach that assumes one or both of $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$ are biased.

Brick et al. (2011) discusses this in the context of combining landline and cellphone telephone samples. The source of potential bias is differential response rates. Lohr and Brick (2014) provide a theory of optimal sample design allocation. Lohr and Brick assign sample sizes to the two frames based on unit variances, unit costs (screener and extended interview), and response rate differentials. This is a useful framework for deciding on sample sizes.

B.1 Combining Cellphone and Landline Frames

The Brick et al. (2011) paper covers the important application of combining household-level samples from landline telephone frames and cellphone telephone frames. The landline-only and cellphone-only portions of this household universe are straightforwardly dealt with through standard methods, and the issue for research is the large subpopulation of households who have both landline and cellphone telephones.

Brick et al. (2011) provide a multiple-frames type estimator of the overlap population ab:

$$\hat{y}_{ps,ab} = \lambda \frac{N_{ab}}{\hat{N}_{ab}^{A}} \hat{y}_{ab}^{A} + (1-\lambda) \frac{N_{ab}}{\hat{N}_{ab}^{B}} \hat{y}_{ab}^{B}$$

where N_{ab} is the population count for the overlap landline-cellphone household population, $\hat{N}_{ab}^{A}$ is the estimator of this count from Frame A (landline), $\hat{N}_{ab}^{B}$ is the estimator of this count from Frame B (cellphone), $\hat{y}_{ab}^{A}$ is the estimator from Frame A, and $\hat{y}_{ab}^{B}$ is the estimator from Frame B, and λ is a parameter between 0 and 1 to be determined.

The primary issue here is that each of $\hat{y}_{ab}^{A}$ and $\hat{y}_{ab}^{B}$ suffer from nonresponse bias. Brick et al. (2011) focus on one important aspect of this. The overlap population can be divided into a 'landline-mainly' and a 'cellphone-mainly' group, based on the household's usage of their telephones. The landline-mainly group can be expected to have a low cellphone response rate, and the cellphone-mainly group can be expected to have a low landline response rate. $\hat{y}_{ab}^{A}$ then can be expected to have a high landline-mainly subgroup response rate and a low cellphone-mainly subgroup response rate underlying it, leading to a bias favoring the landline-mainly portion of this landline-frame sample. $\hat{y}_{ab}^{B}$ then can be expected to have a low landline-mainly subgroup response rate and a high cellphone-mainly subgroup response rate underlying it, leading to a bias favoring the cellphone-mainly portion of this landline-frame sample.

The bias in $\hat{y}_{ps,ab}$ can be written as a function of the difference in the land-mainly and cell-mainly population means (within the overlap population), and the response rates mentioned above. A judicious selection of λ as a function of these response rates can offset and eliminate this bias. Another more straightforward approach is to divide the overlap population further into landline-mainly and cellphone-mainly components. With appropriate control totals utilized from the National Health Interview Survey, an unbiased poststratified version can be computed:

$$\hat{y}_{sep,ab} = \lambda_1 \frac{N_{ml}}{\hat{N}_{ml}^{A}} \hat{y}_{ab}^{A}(ml) + (1-\lambda_1) \frac{N_{ml}}{\hat{N}_{ml}^{B}} \hat{y}_{ab}^{B}(ml) + \lambda_2 \frac{N_{mc}}{\hat{N}_{mc}^{A}} \hat{y}_{ab}^{A}(mc) + (1-\lambda_2) \frac{N_{mc}}{\hat{N}_{mc}^{B}} \hat{y}_{ab}^{B}(mc)$$

where N_{ml} (N_{mc}) is the population count for the mainly-landline (mainly-cellphone) household population, $\hat{N}_{ml}^{A}$ $(\hat{N}_{mc}^{A})$ are the estimators of these counts from Frame A (landline), $\hat{N}_{ml}^{B}$ $(\hat{N}_{mc}^{B})$ are the estimators of these counts from Frame B (cellphone), $\hat{y}_{ab}^{A}(ml)$ $\left(\hat{y}_{ab}^{A}(mc)\right)$ is the estimator for

the mainly-landline (mainly-cellphone) household population from Frame A, $\hat{y}_{ab}^{B}(ml)$ $\left(\hat{y}_{ab}^{B}(mc)\right)$ is the estimator for the mainly-landline (mainly-cellphone) household population from Frame B, and λ is a parameter between 0 and 1 to be determined.

The parameters λ_1 and λ_2 can again be chosen freely without causing bias, allowing for their selection by efficiency considerations. As the sample sizes underlying $\hat{y}_{ab}^{A}(ml)$ may be larger than those underlying $\hat{y}_{ab}^{B}(ml)$ (for the landline-mainly group, one would expect a lot more contacts from the landline sample), a λ_1 closer to 1 will likely be more efficient. Likewise, as the sample sizes underlying $\hat{y}_{ab}^{A}(mc)$ may be smaller than that underlying $\hat{y}_{ab}^{B}(ml)$ (for the cellphone-mainly group, one would expect a lot more contacts from the cellphone sample), a λ_2 closer to 0 will likely be more efficient.

B.2 National Crime Victimization Survey

A scenario quite close to that of the OSHA/SOII issue is the context of the Bureau of Justice Statistics' (BJS) National Crime Victimization Survey (NCVS). The NCVS is an expensive national longitudinal study that starts from a nationally representative sample of households. The sampled households are then tracked and re-interviewed twice a year for a total of seven interviews. The first interview is in-person: the follow-up interviews by telephone. BJS is studying the possible use of a Companion Survey (CS) to improve estimation from the NCVS in small domains. This CS will be an Address-based sampling (ABS) approach, with mail and telephone data collection aspects. There will be only one interview (no follow-up).

The differences in mode between the main NCVS and CS will make combining the two parts of the study very challenging (as with SOII and OSHA). Major mode differences include significant data collection differences, and the recall period for the primary questions about victimization. NCVS asks about six-month periods between longitudinal interviews, and the CS will ask about a twelve-month period preceding the single interview. The NCVS is viewed as unbiased in this context, with the CS as suffering biases from the mode differences with NCVS.

Lohr and Brick (2012) present a multiple frames type methodology for the NCVS/CS merging. Their model is a random effects type model, non-Bayesian. Suppose $\bar{y}_d$ is the mean for domain d from NCVS, and $\bar{x}_d$ is the mean for domain d from CS. Then we assume:

$$\begin{pmatrix}\bar{y}_d\\ \bar{x}_d\end{pmatrix} \sim N\left[\begin{pmatrix}\theta_d\\ \theta_d+\eta_d\end{pmatrix}, \sigma^2\begin{pmatrix}n_{yd}^{-1} & 0\\ 0 & n_{xd}^{-1}\end{pmatrix}\right]$$

The random effects parameter θ_d is assumed to be the actual population value, i.e., the NCVS estimator is assumed to be unbiased. The random effects parameter η_d then measures the bias in CS. n_{yd} and n_{xd} are effective sample sizes for domain d for NCVS and CS respectively. The distribution of random effects is

$$\begin{pmatrix}\theta_d\\ \eta_d\end{pmatrix} \sim N\left[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}a & b\\ b & c\end{pmatrix}^{-1}\right]$$

μ_1 is the overall population mean, and μ_2 the overall population bias (of the CS estimates). The parameter a measures the degree of variability across domains in the population mean, and the parameter c measures the degree of variability across domains in the bias levels (small values of the parameters here correspond to large variability: they are precision parameters). Lohr and Brick (2012) chose to estimate the parameters μ_1, μ_2, a, b, c through maximum likelihood. σ^2 is treated as a fixed value and derived from design-based type calculations similar to Fay and Herriot (1979). There is also a version of these calculations with b assumed to be 0.

The estimates of θ_d and η_d can be computed in an Empirical Bayes type calculation based on the maximum likelihood estimates of the parameters and expressions for the conditional distributions of θ_d and η_d given $\bar{y}_d$ and $\bar{x}_d$. The estimate for η_d (bias of CS for a particular domain) is a linear combination of the 'raw' estimator $\bar{x}_d - \bar{y}_d$, the partially shrunken estimator $\bar{x}_d - \hat{\mu}_1$ and the fully shrunken estimator $\hat{\mu}_2$.

Lohr and Brick (2012) also recommend more direct estimators based on linear combinations of $\bar{y}_d$, $\bar{x}_d$, linear combinations of $\bar{y}_d$ and $\bar{x}_d - \hat{\mu}_2$, and linear combinations of $\bar{y}_d$ and $\bar{x}_d - \hat{\eta}_d$, with linear weights suggested from the model described above. Also an estimator based on assuming a multiplicative bias was developed, as well as estimators based on calibrating the CS estimator to NCVS estimators. These estimators were studied for various population scenarios through a simulation study.

It should be noted that DoJ did not follow through on the Companion Survey, after pilot studies were done. There is no followup then to the Lohr and Brick (2012) work, and NCVS continues as the one and only DoJ survey for crime victimization (as of 2017).

Appendix C

Detailed Review of Papers on Small-Area Area-Level Estimation

Appendix C
Detailed Review of Papers on Small-Area Area-Level Estimation

C.1 NHIS and BRFSS: Reweighting Approach

For Elliot and Davis (2005) the estimates of interest are cancer risk factor prevalences (such as smoking and mammogram prevalences), and the small areas are counties. NHIS has excellent national estimates of these prevalences (with high relative response rates and full coverage), but is not designed to provide county-level estimates. Sample sizes are small and there are restrictions on geographic data disclosure at the county (or even state) level[2]. The Behavior Risk Factors Surveillance System (BRFSS) provides much larger sample sizes at the state and county levels, but suffers from known bias from a much lower response rate and coverage restricted to landline telephone households.

The solution of Elliot and Davis (2005) rests on the observation that the biases inherent in BRFSS mostly come about from a distortion in the represented population arising from the low response rates and the undercoverage of cellphone-only and non-telephone households. The BRFSS sample is the basis for the county-level prevalence estimates, but NHIS is utilized to adjust the BRFSS samples but calibrating the BRFSS weights to NHIS, but not in the usual way of raking to control totals.

The approach is to put the records from the two surveys together (without any matching which is not possible). Write $S_i = \{a, b\}$ as the indicator of whether the household is in NHIS (survey a) or BRFSS (survey b). Write y_i as the value of a prevalence indicator of household i, write $\mathbf{x_i}$ as a set of covariates shared by both NHIS and BRFSS related to y_i, write $G_i = g$ as an indicator for household i being in small-area g, $A_i = r$ an indicator for household i being in region r (with region r identified and disclosed for both surveys: Census region would be the right level for NHIS), and d_i a final weight from BRFSS.

The Elliot and Davis (2005) concept is designed to adjust d_i by the ratio

[2] The public-use NHIS data set only has geographic information at the Census Region level crossed with seven urban-rural categories.

$$\frac{f(Y_i, \mathbf{X_i} = \mathbf{x_i} | S_i = a, G_i = g)}{f(Y_i, \mathbf{X_i} = \mathbf{x_i} | S_i = b, G_i = g)}$$

with $f(\,)$ a PDF. This is designed to 'fix' the distorted probability distribution inherent in BRFSS and bring it into alignment with NHIS. Using Bayes rule this can be rewritten as:

$$\frac{f(Y_i, \mathbf{X_i} = \mathbf{x_i} | S_i = a, G_i = g)}{f(Y_i, \mathbf{X_i} = \mathbf{x_i} | S_i = b, G_i = g)} = \frac{f(S_i = a | Y_i, \mathbf{X_i} = \mathbf{x_i}, G_i = g) / f(S_i = a | G_i = g)}{f(S_i = b | Y_i, \mathbf{X_i} = \mathbf{x_i}, G_i = g) / f(S_i = b | G_i = g)}$$

But this ratio cannot be computed as we do not have $G_i = g$ information for NHIS, so Elliott and Davis (2005) assume that $G_i = g$ can be replaced by $A_i = r$ (assuming certain conditional probabilities are equal across small areas), resulting in their NHIS-adjusted BRFSS weight being as follows:

$$w_i = d_i \frac{f(S_i = a | Y_i, \mathbf{X_i} = \mathbf{x_i}, A_i = r) / f(S_i = a | A_i = r)}{f(S_i = b | Y_i, \mathbf{X_i} = \mathbf{x_i}, A_i = r) / f(S_i = b | A_i = r)}$$

The analysis of Elliott and Davis (2005) demonstrates to their satisfaction (using other sources of information such as the Current Population Survey Tobacco Use Supplement for smoking prevalence) that using the NHIS-adjusted BRFSS weights as compared to the original BRFSS weights adjust for the biases in the prevalence estimates from the original BRFSS. However, the weights tend to be sometimes highly variable, leading to higher variances. Elliott and Davis developed a hybrid estimator which estimates mean squared error and then finds the best linear combination of original-BRFSS and NHIS-adjusted BRFSS estimator to minimize overall mean-squared error.

C.2 NHIS and BRFSS: Small-Area Estimation Approach

Raghunathan et al. (2007) continue research on using NHIS to adjust BRFSS county-level estimates of cancer-risk prevalence. In their approach, they gain full access to county-level data from NHIS, allowing for considerable improvement over the information base available to Elliott and Davis (2005), who only had access to the public-use version of NHIS with its highly restricted geographic-level information. They use as a starting point the vector of direct estimates (using NHIS and BRFSS final weights) $(p_{xjt} \quad p_{yjt} \quad p_{zjt})'$ where p_{xjt} is the direct estimate of prevalence among NHIS telephone households in county j, year t, p_{yjt} is the direct estimate of prevalence among

NHIS non-telephone households in county j, year t, and p_{zjt} is the direct estimate of prevalence among BRFSS households in county j, year t. Note that all BRFSS households are telephone households. Raghunathan et al. found considerable differences between p_{xjt} and p_{yjt} in general for smoking prevalence and mammogram usage prevalence: non-telephone households appear to be considerably different from telephone households in the prevalence of these risk factors. On the other hand, there were not drastic differences between p_{xjt} and p_{zjt} (NHIS telephone households and BRFSS telephone households). For these prevalence items at least, BRFSS lack of coverage of non-telephone households is a much bigger generator of bias than BRFSS non-response (assuming NHIS is a fully unbiased benchmark).

Raghunathan et al. (2007) proceed to adjust BRFSS direct estimates p_{zjt} by utilizing the following Bayesian small-area estimation model:

$$\begin{pmatrix} x_{jt} \\ y_{jt} \\ z_{jt} \end{pmatrix} = \begin{pmatrix} \arcsin\sqrt{p_{xjt}} \\ \arcsin\sqrt{p_{yjt}} \\ \arcsin\sqrt{p_{zjt}} \end{pmatrix} \sim$$

$$\sim N_3\left[\begin{pmatrix} \theta_{jt} \\ \varphi_{jt} \\ (1+\delta_{jt})\theta_{jt} \end{pmatrix}, \frac{1}{4}\begin{bmatrix} \tilde{n}_{xjt}^{-1} & \rho_t(\tilde{n}_{xjt}\tilde{n}_{yjt})^{-\frac{1}{2}} & 0 \\ \rho_t(\tilde{n}_{xjt}\tilde{n}_{yjt})^{-\frac{1}{2}} & \tilde{n}_{yjt}^{-1} & 0 \\ 0 & 0 & \tilde{n}_{zjt}^{-1} \end{bmatrix}\right]$$

with $\tilde{n}_{xjt}$ $(\tilde{n}_{yjt},\ \tilde{n}_{zjt})$ being effective sample sizes (taking into account design effects from the sample design, weights, etc.) for the three estimates $(p_{xjt},\ \ p_{yjt},\ \ p_{zjt})$ respectively. The original percentages are transformed using the traditional arcsin-square root variance stabilizing transformation. Important parameters in the model are θ_{jt}—the true (transformed) telephone household prevalence in county j, time t; φ_{jt}—the true (transformed) non-telephone household prevalence in county j, time t; δ_{jt}—the bias term in the BRFSS telephone household prevalence; and, ρ_t—a correlation within NHIS between telephone and non-telephone household prevalences, assumed constant over counties.

There is a further hierarchical model for the three parameters $(\theta_{jt},\ \ \varphi_{jt},\ \ \delta_{jt})$ as follows:

$$\begin{pmatrix} \theta_{jt} \\ \varphi_{jt} \\ \delta_{jt} \end{pmatrix} \sim N_3(\boldsymbol{\beta}\mathbf{U}_{jt}, \boldsymbol{\Sigma})$$

$\mathbf{U}_{jt}$ is a vector of county-level covariates designed to explain the county-level variation in these parameters. These county level characteristics include county percentages by race, education, poverty, employment, social services, and other estimates of economic-related characteristics of the county. There are further noninformative priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

The Raghunathan et al. (2007) model is a true small-area estimation model in that there are design-based estimates of the variance built in to the model that are treated as fixed values. The inherent biases between NHIS and BRFSS at the county level as represented in the vector $(\theta_{jt}, \quad \varphi_{jt}, \quad \delta_{jt})$ are carefully disentangled using county-level sociodemographic and economic information.

C.3 Kim, Park, and Kim: Small-Area Estimation Combining Surveys with Fixed Bias

Kim et al. (2015) provide basic small-area estimation theory combining two surveys. This theory is primarily motivated by combining two sources for employment information in South Korea: The Korean Labor Force (KLF) survey and the Local Area Labor Force Survey (LALF). The KLF survey has about 7,000 households with no measurement error (the gold standard), and the LALF has about 200,000 households with measurement error from its rougher field collection. The KLF is a second-phase sample from the LALF, and they generate small-area estimates for 227 small areas (called 'Gu') within South Korea.

Their basic model can be summarized as follows:

$$\begin{pmatrix} \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \bar{X}_h \end{pmatrix} + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

with $\boldsymbol{\bar{X}_h}$ the target population mean in small-area $\boldsymbol{h}$, $\boldsymbol{\bar{x}_h}$ the estimate from the gold standard survey, $\boldsymbol{\bar{y}_{1h}}$ the estimate from the large survey with measurement error, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ coefficients relating $\boldsymbol{\bar{y}_{1h}}$ to $\boldsymbol{\bar{X}_h}$, and $\boldsymbol{a_h}$, $\boldsymbol{b_h}$, $\boldsymbol{\bar{e}_{1h}}$ variance components. Assuming parameters are known the final Empirical Bayes type shrinkage estimator is

$$\hat{\bar{X}}_h = \boldsymbol{\alpha_h \bar{x}_h} + (\mathbf{1} - \boldsymbol{\alpha_h})\boldsymbol{\beta_1^{-1}}(\boldsymbol{\bar{y}_{1h}} - \boldsymbol{\beta_0})$$

with $\boldsymbol{\alpha_h}$ generated based on the relative precision of the small-sample gold-standard estimator, and the large-sample estimator adjusted for bias. Parameters are estimated using maximum likelihood: the approach is non-Bayesian.

The estimation approach is well-developed. Kim et al. (2015) apply it to the Korean labor force surveys, but assume in that application that β_0 is equal to 0. The deficiency in the Kim et al. approach from the standpoint of the OSHA/SOII application is that the bias is assumed to be equal across all small areas. If this assumption is judged to be satisfactory, then the Kim et al. approach is definitely a possibility.

C.4 Small-Area Estimation Theory: General

Ybarra and Lohr (2008) present a theoretical development under the classical non-Bayesian small-area estimation framework. They work with the Fay-Herriot model

$$y_i = X_i^T \beta + v_i + e_i \quad \begin{bmatrix} v_i \\ e_i \end{bmatrix} \sim N_2 \begin{bmatrix} \sigma_v^2 & 0 \\ 0 & \psi_i \end{bmatrix}$$

ψ_i is a design-based variance. y_i is an unbiased estimator of $Y_i = X_i^T \beta + v_i$. If all the parameters are known, the traditional Fay-Herriot predictor for Y_i is

$$\hat{Y}_{iFH} = \hat{\gamma}_{iv} y_i + (1 - \hat{\gamma}_{iv}) X_i^T \hat{\beta}$$

with $\hat{\gamma}_{iv} = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \psi_i)$, and $\hat{\beta}, \hat{\sigma}_v^2$ are estimators of β, σ_v^2 respectively.

Ybarra and Lohr (2008) allow for X_i to be measured with error: potentially both bias and variance. Write MSE $(\hat{X}_i)$ as a matrix C_i. If $\beta^T C_i \beta > \sigma_v^2 + \psi_i$, then the Fay-Herriot predictor will have higher MSE than the simple direct estimator y_i. As X_i in many cases is coming itself from an estimate which may have the same level of error as y_i, this should be a concern.

Their alternative estimator under these conditions is as follows:

$$\hat{Y}_{iFH} = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i) \hat{X}_i^T \hat{\beta}_w$$

with $\hat{\gamma}_i = \frac{\hat{\sigma}_v^2(w)+\hat{\beta}_w^T C_i \hat{\beta}_w}{\hat{\sigma}_v^2(w)+\hat{\beta}_w^T C_i \hat{\beta}_w+\psi_i}$ and $\hat{\sigma}_v^2(w)$ and $\hat{\beta}_w$ are consistent estimators of σ_v^2 and β respectively allowing for the extra variance in $\hat{X}_i$.

Appendix D

Detailed Review of Papers on Hierarchical Models for Combining Surveys

Appendix D
Detailed Review of Papers on Hierarchical Models for Combining Surveys

D.1 Bayesian Models at the Estimate Level: No Covariates

Manzi et al. (2011) provides a Bayesian model for combining small area smoking prevalence estimates. The population estimate of interest is smoking prevalence, and the small areas are 48 Local Areas (LAs) in Eastern England. There are seven different surveys which are providing estimates for these 48 Local Areas. Three of the seven estimates are from Acxiom, which are providing estimates from a UK National Shoppers Survey. This survey has very low response rates and coverage issues: the methodology for carrying out the survey is actually unpublished. One of the seven surveys is a commercial community insights survey on household expenditure on tobacco (called 'CACI'). Three of the seven surveys are from the Health Survey for England (HSE), a major nationally representative survey of households (with two of the HSE surveys deriving the LA estimates from a small-area type model). The Acxiom surveys have large enough data sets from each LA, but an unknown degree of nonsampling error, and have prevalence estimates that are less than the others. CACI has prevalence estimates larger than the others. The HSE surveys have presumably much lower nonsampling error, but the sample sizes for the LAs are small and the estimates are based partially on small-area type models (and are also several years older).

The Manzi et al. (2011) model is as follows. Let y_{ij} be the smoking prevalence in LA i ($i = 1, \ldots, 48$) and from data source j ($j = 1, \ldots, 7$), in percentage terms (y_{ij} is 0 to 100). Let θ_i be the true smoking prevalence in LA i, with a uniform prior on the interval [0,100], but with the extra important proviso that the mean value of the θ_i is $\bar{\theta} = 23$[3]: an estimate for smoking prevalence for Eastern England from the UK General Household Survey (with assumed high accuracy and low bias). This General Household Survey estimate is being treated as the 'truth' and is assumed to be a fixed value (an overstatement as to its accuracy: it has at least sampling error associated with it).

Let δ_{ij} be the bias in the prevalence estimate for LA i, data source j. The conditional joint model for y_{ij} and δ_{ij} is

[3] This is implemented by assuming a uniform prior for each θ_i, i =1.„.47, and then assuming the last LA value θ_{48} is equal to 48*$\bar{\theta}$ – $\sum_{i=1}^{47} \theta_i$.

$$y_{ij} | \delta_{ij} \sim N(\theta_i + \delta_{ij}, \sigma_{ij}^2) \quad \delta_{ij} \sim N(\mu_j, \tau_j^2)$$

with σ_{ij}^2 the variance of the estimate ij (treated as a fixed value), μ_j the overall mean bias from source j, and τ_j^2 the variance of the LA biases from the mean bias for source j. Noninformative priors are posited for the μ_j and τ_j^2. Posterior distributions are generated for all of the relevant parameters using a Markov Chain Monte Carlo (MCMC).

Manzi et al. (2011) develop a second model which allows for correlations in the δ_{ij} between sources j_1 and j_2 (the covariance between δ_{ij_1} and δ_{ij_2} is posited to be $\rho\tau_{j_1}\tau_{j_2}$—there is one single universal correlation parameter ρ). This wrinkle doesn't change the final posterior means of the θ_i much, but increase the posterior variance.

Manzi et al. (2011) also develop a non-Bayesian version of this which omits the Bayesian priors and fits the model as a two-way ANOVA, doing effectively Empirical Bayes type estimates of the resultant θ_i. A second version of this fits the model as a mixed effects model:

$$y_{ij} = \theta_i + \mu_j + \delta_{ij}^* + \varepsilon_{ij}$$

with θ_i and μ_j as fixed effects, and $\delta_{ij}^* \sim N(0, \tau_j^2)$, $\varepsilon_{ij} \sim N(0, \sigma_{ij}^2)$, with τ_j^2 and σ_{ij}^2 also as fixed parameters, and imposing the constraint that $\bar{\theta}$ should be equal to 23 (the GHS estimate). The fixed parameters can be estimated using maximum likelihood for example, and Empirical Bayes methods utilized to provide BLU predictors of δ_{ij}^* for example. It should be noted that this non-Bayesian version of the model is very close to the model from Lohr and Brick (2012), with the only major difference that in Lohr and Brick one of the sources is assumed to be unbiased (i.e., one source is assumed to have a μ_j equal to 0).

Both the Manzi et al. (2011) model and the Lohr and Brick (2012) model essentially estimate overall bias simply by studying differences between the source means. If the overall sample sizes are large enough this may be reasonable. Of course comparing the source means only allows one to measure relative biases (differences in expectations). To get to absolute biases, one must either as in Lohr and Brick assume up front that one source is unbiased, or as in Manzi et al. one must have an independent, high-precision estimate of the overall mean (their $\bar{\theta}$ equal to 23).

Then, for domains, the idea is that the domain biases are all fairly close to the overall bias: whatever the source of bias is in a source, it does not differ systematically across domains. This crucial

assumption is what makes it possible to get a handle on the domain-level biases in the context of small sample sizes in the domains.

D.2 Bayesian Models at the Estimate Level: Corn Yields from the National Agricultural Statistical Services.

Wang et al. (2011) work with three separate estimates of corn yield from NASS as follows:

- OYS: Objective Yield Survey. This is based on a sample of acres with corn yield in June of the year, for five months August through December, in high corn-producing states;
- AYS: Annual Yield Survey. This is based on farm interviews conducted monthly for four months August through November in every state (from a list of producers screened in June as having planted corn);
- DAS: December Annual Survey. Interviews conducted of farmers in December at the end of the season across all corn farmers: large sample, and across all states.

The DAS is accepted by NASS as being relatively unbiased, and is assumed in models as being unbiased[4]. A fully Bayesian approach is used to bring these three estimators together, using the following model:

$$[\text{true yield}, \Theta_d, \Theta_p | \text{OYS}, \text{AYS}, \text{DAS}] \propto$$
$$\propto [\text{OYS}|\text{true yield}, \Theta_d] [\text{AYS}|\text{true yield}, \Theta_d] [\text{DYS}|\text{true yield}, \Theta_d] [\text{true yield}|\Theta_p] [\Theta_p] [\Theta_d]$$

with Θ_d, Θ_p parameters for the data model and the process model respectively.

The data models $[\text{OYS}|\text{true yield}, \Theta_d]$ and $[\text{AYS}|\text{true yield}, \Theta_d]$ assume arbitrary monthly biases for both surveys, but allow for an autocorrelation model across the months to pool the monthly estimates. The model $[\text{DYS}|\text{true yield}, \Theta_d]$ assumes DAS has as its expectation the true yield. These data models alone will not be enough to do anything more than make the final estimates rest primarily on DYS (with OYS and AYS simply estimating their respective biases), but there is also a substantive process model $[\text{true yield}|\Theta_p]$ which models the true yield as a linear function of

[4] Nandram et al. (2014) indicate that OYS may be biased due to technical reasons regarding the way that the corn crops are measured. AYS may be biased because the large producers who may have differing yield patterns are left off of the AYS frame to control burden (they participate in many NASS surveys).

exogenous variables such as July rain and temperatures, and corn planted by mid-May. Priors on the parameters are non-informative.

Nandram et al. (2014) provide further development of these models for this application for state-level estimates. In this paper, they explore the issue of constraining the state-level to overall estimates. One traditional approach to this is through calibration: calibrating the state-level estimates for overall national estimates directly. Nandram et al. do this benchmarking through a Bayesian approach instead. The assigned models for the state-level estimates have built into them the constraint that they add to national estimates. The national estimates themselves are brought in using a prior distribution that is concentrated on the realized value. This is somewhat ad hoc, but it works practically.

Appendix E

Detailed Review of Papers on Data Linkage

Appendix E
Detailed Review of Papers on Data Linkage

E.1 Data Linkage: Traditional Methods and Mixture Models

Christen (2012) provides a useful summary of data matching in a general context. His Sections 3 and 4 describe current methodologies for blocking and indexing: necessary procedures for dividing up data sets into digestible blocks (or reasonable sort orders) so that too many pairs do not need to be compared. Section 5 describes current methods for actually evaluating similarity of strings (names in particular). Section 6.3 presents probabilistic classification for deciding on whether or not two candidate records should be linked based on their similarity evaluation. A basic ratio of conditional probabilities as given from the original Fellegi and Sunter (1969) paper is as follows:

$$R = \frac{P(\gamma \in \Gamma \mid r \in M)}{P(\gamma \in \Gamma \mid r \in U)}$$

where γ is an agreement pattern based on the pooled comparison variables among the universe of all possible such patterns Γ, r is a candidate pair of records, M is the set of all record pairs corresponding to true matches (one single individual in the population), and U is the set of all record pairs not corresponding to true matches. The remainder of Section 6 explores alternatives to this basic Fellegi and Sunter approach. Section 7 provides an overview of methods for evaluating quality and complexity. Section 10 provides an inventory of currently available data matching systems.

Bohensky et al. (2010) study 33 studies which carry out data linkage with evaluation of the differences between linked and unlinked records (among a larger set of 612 studies who did data linkage), in the medical research literature. They found a large percentage of studies that found some imbalance in linkage rates by gender, age, race, socioeconomic status and health status.

Winglee et al. (2005) provide a case study of a traditional Fellegi-Sunter type approach from the Medical Expenditures Panel Study (MEPS). The linkage is between the annual MEPS medical event files from 1996, 1997, and 1998. Each annual set consists of a household file with self-reports of medical events from the household members, and a parallel file from the medical provider of these same events. Linking is done between these files using event dates and duration (hospital stay length), and medical condition and procedure codes. The MEPS application uses the standard Fellegi-Sunter approach to assign cases, assuming independence of the matching indicator vector. A

weight is generated which expresses the degree of matching of potential match pairs. The critical issue is to set cutoffs X for matching weights x_i deciding whether to designate a pair i as matched. Any selected cutpoint generates a certain percentage of false positives and false negatives, and deciding on this tradeoff is the subject of the Winglee et al. work. They apply a 'gold-standard' method of working with pairs assigned by experienced field-collection personnel. They also apply an approach from Belin and Rubin (1995), as well as a simulation approach.

Under the Belin and Rubin approach, the weights x_i that are assigned to pairs to represent the degree of matching are studied as a mixture distribution conditional on whether the pair represents a true match (z_i=1), or not (z_i=0). The conditional distributions for x_i given z_i may be quite different. With the two component distributions comprising the mixture distribution are fully parameterized and fit based on the gold-standard training data, this mixture model can then be inverted to provide conditional distributions for z_i given x_i, and provide cutoffs for x_i as to assigning pairs for new files (the mixture distribution allows a computation of expected false positive and false negative rates based on setting different potential cutoffs based on x_i (assign z_i=1 if $x_i > X$, assign z_i=0 if $x_i \leq X$).

Winglee et al. carry through an analysis along these lines then on the MEPS data, generating weights x_i based on a training sample (weights assigned by the linkage program on a set of true matches z_i=1 and true non-matches z_i=0 assigned through manual review by knowledgeable data managers).

Winglee et al. utilize a third approach for setting the cutoffs in a rational way. In this approach, Monte Carlo samples are generated from a theoretical distribution and sample distributions are then examined, measuring then the variability in the false positive and false negative rates resulting from differing cutoff values. This third approach is the least accurate, but it is the least expensive. Generating training data sets through manual checking by data managers, and/or using a theoretical approach such as the mixture models from Belin and Rubin, are more expensive and cannot be done annually.

Larsen and Rubin (2001) present a mixture model for determining whether of not a particular pair is in the M or the U sets. Their mixture likelihood is defined as follows:

$$p(\mathbf{y}|\pi,\theta) = \prod_{l=1}^{L}\left(\sum_{g=1}^{2}\pi_g\pi_{l|g}\right)^{n_l}$$

where $y = (y_1, \ldots, y_l, \ldots, y_L)$ are the possible patterns of agreement or disagreement (the same as y in Fellegi and Sunter), $g = 1,2$ are the two sets M and U from Fellegi and Sunter, π_g is the (overall) probability of being in set g, and $\pi_{l|g}$ is the probability of pattern l given the M or U sets. n_l is the total sample having pattern l. Each pattern l is based on binary agreement/nonagreement for K different matching items (e.g., first name, address, etc.), so that $L = 2^K$. Note that the sample units here are all possible pairs (not single record on a single files): only pairs can have patterns of agreement or disagreement. Larsen and Rubin carry through on this model using maximum-likelihood (the EM algorithm) rather than a fuller Bayesian approach. Observed resolved pairs (from clerical review) with known values of y and known assignment to M or U become the data which allows for estimation of the parameters. New pair assignments for unresolved cases can be determined from the model to 'predict' unresolved pairs as being in M or U. The model can actually be used to resolve cases into M, U and an intermediate undetermined class which can be then sent back for clerical review (and resolution of these can then be used to refit the model and re-estimate the parameters). The $\pi_{l|g}$ can be a product of K independent probabilities of matching (independence model), or it can be a more complicated interdependent linking of the K item matches (e.g., first name and last name matching are not independent). In their Census data file examples, Larsen and Rubin fit a variety of different models. In many of their cases, the $\pi_{l|g}$ models differ for $g = 1, g = 2$.

E.2 Data Linkage: Bayesian Methods

Goldstein et al. (2012) present a theory (with a simulation study) for linking two files when the data linkage is not definite. The two files are the 'File Of Interest' (FOI) and the 'Linked Data File' (LDF). The FOI is the primary file of which analysis is done, and the LDF provides auxiliary information for some further fields which are then appended to the FOI. When the link is clear, the LDF contributes fields to the linked FOI records directly. When the link is unclear, they outline what they call the 'traditional probabilistic record linkage' (with some references), in which a linkage is made using matching variables which are shared in common between the FOI and LDF. A probability p_{ij} is computed via a model of the relationship between the matching variables and the existence of a true match (or not) and a data analysis informed by the model, which generates the estimated probability that LDF record j is the match for FOI record i. There may be j=1,…,n_i candidate LDF records, and we have $\sum_{j=1}^{n_i} p_{ij} = 1$. We can write $\mathbf{p_i}$ as a vector of $\{p_{ij}\}$ representing our 'link probabilities' for linking FOI record i to the LDF. This basic structure they call the 'traditional probabilistic record linkage' approach.

Goldstein et al. converts $\mathbf{p_i}$ into a prior probability for a Bayesian-type multiple imputation approach. They append to $\mathbf{p_i}$ a second factor $f\left(y_{ij}^{A|B}\right)$ which is a more traditional multiple imputation maximum likelihood which treats the data being brought over from the LDF records as missing data (the 'A' data), and models them based on other fields within the FOI or even from the LDF (the 'B' data). The final set of posterior probabilities $\boldsymbol{\pi_i} = \{\pi_{ij}\} \propto f\left(y_{ij}^{A|B}\right)\mathbf{p_i}$. These posterior probabilities are then utilized in a way that is similar to traditional probabilistic record linkage. A threshold is taken for the pi and if any j-records exceed the threshold, then the j-record with the highest probability is chosen as the linked record to record i. If no records exceed the threshold, then no matches are taken and the FOI record has its data filled in by regular imputation.

Goldstein et al. carries out a simulation study that demonstrates the superiority of their mixed approach to the traditional probabilistic record linkage approach.

Steorts et al. (2016) moves beyond Goldstein et al. (2012) to a full-scale Bayesian approach. Under the Steorts et al. approach, we have k files, and assume for simplicity p categorical fields in common for linking, with M_ℓ levels for each field ℓ, ℓ =1,… p. Write $\mathbf{x_{ij}}$ as the vector of matching variables of length p for the jth record in file i. They posit a latent vector $\mathbf{y}_{j'}$ as a vector of true values for the j'th 'true' individual in the population, where j'=1,…,N, with N the total number of individuals corresponding to the records in the k files. They posit a random variable λ_{ij} (vector $\boldsymbol{\Lambda}$) which points to the true individual j'=1,…,N corresponding to record ij. An important indicator is $z_{ij\ell}$ which is 1 if there is an error in field $x_{ij\ell}$ (or 0 otherwise). $\boldsymbol{\theta_\ell}$ is a vector of multinomial probabilities. β_ℓ is the probability $z_{ij\ell}$=1. $\delta_{\boldsymbol{y}_{\lambda_{ij}\ell}}$ indicates a point mass at the value of $\boldsymbol{y}_{\lambda_{ij}\ell}$ (which is the value of the categorical vector for the true individual j' corresponding to record ij). The Steorts et al. model is as follows:

$$x_{ij\ell} \mid \lambda_{ij}, \boldsymbol{y}_{\lambda_{ij}\ell}, z_{ij\ell}, \boldsymbol{\theta_\ell} \sim \begin{cases} \delta_{\boldsymbol{y}_{\lambda_{ij}\ell}} & z_{ij\ell} = 1 \\ \mathrm{MN}(1, \boldsymbol{\theta_\ell}) & z_{ij\ell} = 0 \end{cases}$$

$$z_{ij\ell} \sim \mathrm{Bernoulli}(\beta_\ell)$$
$$\boldsymbol{y}_{j'\ell} | \boldsymbol{\theta_\ell} \sim \mathrm{MN}(\mathbf{1}, \boldsymbol{\theta_\ell})$$
$$\boldsymbol{\theta_\ell} \sim \mathrm{Dirichlet}(\mu_\ell)$$
$$\beta_\ell \sim \mathrm{Beta}(a_\ell, b_\ell)$$
$$\boldsymbol{\pi}(\boldsymbol{\Lambda}) \propto \mathbf{1}$$

MN represents the multinomial distribution. The parameters μ_ℓ, a_ℓ, b_ℓ are assumed known. The model is fit using the Gibbs sampler, and posterior probabilities of linkage can be then generated. It may be necessary to do blocking (dividing the data sets into pieces and allowing links only between designated pieces) to allow the Gibbs sampler to run in a reasonable time with reasonable memory. The posterior probabilities of linkage are used then to decide on best pairings (ij,$i'j'$). Their model is tested against a set of data sets generated from the National Long Term Care Survey, a longitudinal study of the health of elderly individuals, where the real link is known, and the model is found to perform well with a relatively limited set of error-prone linking variables (date of birth, sex, state of residence, and regional office).

Gutman et al. (2013) provide a further Bayesian model. Assuming for simplicity there are two files with equal numbers of records and every record on File A has a matching record on File B, a random variable C_j is defined which indicates the correct permutation of the second file records to match all of the first file records one to one. A conditional distribution of C_j given all of the field values ($\mathbf{Y_A}$ for fields only on File A, $\mathbf{Y_B}$ for fields only on File B, and $\mathbf{Z}$ being variables on both files)) and a set of model parameters can be defined. Likewise a likelihood function of for $\mathbf{Y_A}$, $\mathbf{Y_B}$, $\mathbf{Z}$ can be defined conditional on the particular C_j and the parameters. This conditional distribution and the likelihood, along with Bayesian priors, can be iterated to a solution. Once C_j is defined, the likelihood is easy to state. Once the $\mathbf{Y_A}$, $\mathbf{Y_B}$, and $\mathbf{Z}$ distributions are specified and the parameters defined, probabilities for C_j can be given. This approach has the virtue of being theoretically clean (though dealing with files of different sizes leads to complexities), but the sample universe for C_j is immense: all possible permutations of 1 through N, where N is the file size. This daunting sample universe can be reduced by blocking on some of the $\mathbf{Z}$ variables, reducing the set of potential pairs and permutations to a much more manageable number. Gutman et al. apply this approach to linking the Medicare file (from the Centers for Medicare and Medicaid Services) to the Vital Statistics Mortality (VSM) records compiled from death certificates nationwide from the National Center of Health Statistics. The goal was to link information about cause of death of decedents in the US (from the VSM), with information about Medicare expenditures from the Medicare file about these decedents. A gold standard link between these files (e.g., social security number) was not available to the researchers. Blocking was done on age, sex, race, month and day of week of death, and state and county of residence. This blocking generated small cells, so that there were not too many units. In some cases, exact matches were possible, but to represent the population correctly the data set was not restricted to exact matches only (which tended to be in small blocking cells). The final models are fit directly and analyze Cause of Death (multinomial logistic regression), Place of Death (multinomial logistic regression), Medicare Part A expenditures (linear regression), and Medicare Part A and B expenditures (linear regression).

Tancredi and Liseo (2015) present a further Bayesian model in the context of linear regression analysis. The approach of Steorts et al. (2016) and Gutman et al. (2013) is followed in that a matrix of all possible match pairs is specified, and the conditional distribution of the variables given this matrix is specified, and the conditional distribution of the matrix given these other variables is specified. The overall model is fit in an iterative process using the Gibbs sampler. As in Steorts et al. there is an allowance for noise: the two data files having slightly differing observations even though the match is a true ones. Conditional on a true match, there still can be differing values for particular linking variables. Tancredi and Liseo depart somewhat from the other papers in strongly including the analysis of interest directly into the parameterization. This means ultimately that matching the pairs will be partially driven by the analysis model itself: pairs are given a higher probability of being true in a particular iteration if they happen to match well the currently fitted regression model. This ties the pairing to a particular model for a particular set of variables, which may not be realistic for OSHA/SOII.

Gutman et al. (2015) provide a further application along the lines of Gutman et al. (2013) and Tancredi and Liseo (2015). In this application, records from the Rhode Island Department of Corrections are linked with records from the Miriam Hospital HIV care program to track the experience of recently released HIV+ prisoners in Rhode Island. The basic idea of having a latent variable for a true link between two records, and generating a conditional distribution of the variables of interest given one realization of 'true pairs' is implemented. Their model for deciding on true pairs is partially dependent on outcome variables only on one data set (rather than simply on the incidence of matching agreements between variables on both data sets). Multiple imputations are generated which represent various 'true pairs' from the posterior distribution for the potential pairings. It should be noted that this analysis is dependent on a 'gold standard analysis' being done on a small subset of the two files, so that there is firm data on true and false pairings that can go into the accurate development of the model.

E.3 Adjusting for Imperfect Data Linkage

There are a number of recent papers where there is a (non-Bayesian) analysis of the effects of imperfect data linkage on analysis, and on ways of adjusting for this effectively. This work generally focuses on regression analysis between a dependent variable y (from one file), and predictors X (from another file imperfectly linked record by record to the first file). This imperfect linking can lead to bias in the regression coefficient estimate. The approach of these papers focus on assigning a

probability of correct linking. For example, Lahiri and Larsen (2005) work with the following standard simple regression model:

$$y_i = \mathbf{x_i'}\boldsymbol{\beta} + \epsilon_i \quad i = 1, \ldots, n \quad E(\epsilon_i) = 0.$$

The imperfect linking is modeled by positing a z_i random variable as follows:

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \text{ for } j \neq i. \end{cases}$$

The regression is then done implicitly with z_i as the dependent variable rather than the unobserved y_i, and results in a biased estimator of $\boldsymbol{\beta}$. If the $\mathbf{q_i}$=(q_{i1},…, q_{in}) vector is known or can be estimated, then a weighted regression utilizing the $\mathbf{q_i}$ vector will result in a fully unbiased estimator of $\boldsymbol{\beta}$.

Kim and Chambers (2012) posit probabilities of true and false linkings $\mathbf{q_i}$ based on an exchangeable linking model (within particular blocks all of the q_{ii} are equal, and all of the $q_{ij}, j \neq i$, are equal). If this assumption is reasonable, then the $\mathbf{q_i}$ vector can be fully specified, and unbiased estimators of b are readily calculable.

Chipperfield et al. (2011) extend the basic approach as given in Lahiri and Larsen (2005) to categorical dependent variables. They study contingency tables and logistic regression. For example, they recommend adjusting for imperfect linkages using maximum likelihood under the standard logistic regression generalized regression logistic link iterative fitting, but proceed by replacing the dichotomous y dependent variable with a perturbed version of y:

$$\tilde{y}_i = \begin{Bmatrix} y_i^* \hat{p}_{xy^*} + (1 - \hat{p}_{xy^*})\tilde{v}_i & i \notin s_c \\ y_i^* & i \in s_c \text{ and } \delta_i = 1 \\ \tilde{v}_i & i \in s_c \text{ and } \delta_i = 0 \end{Bmatrix}$$

with s_c being the set of linked records, δ_i indicating a successful linking or not (based on clerical review), y_i^* is the value of y from the linked record, $\tilde{v}_i$ is the expected value of y from the most current iteration of the maximum likelihood, and $\hat{p}_{xy^*}$ is the probability of being correctly linked for the particular pair (x being the predictor variables, and y being the dependent variable value from the linked record). $\hat{p}_{xy^*}$ is computed from the clerical review records. As in the regular regression as

given in Lahiri and Larson (2005), the insertion of these predicted probabilities of accurate linkage into the maximum likelihood iterations, if done accurately, will eliminate the bias from invalid links.

Hof and Zwinderman (2015) provide a similar approach for carrying out maximum likelihood for a general model in the presence of imperfect linkages. Their model is different in the sense that they define a likelihood function for all possible pairs between the two matched files. One file (File A with n records) is providing the y_i values (the dependent variable), and the other file (File B with m records) is providing the $\mathbf{x_i}$ predictor variable values. They define a a d_{ij} random variable which is equal to 1 if record i in File A matches record j on File B, and is 0 otherwise. Their likelihood in the general case (before simplifying assumptions, and leaving out parameters) is:

$$\prod_{i=1}^{n}\prod_{j=1}^{m}\{\ L(y_j, \mathbf{x_i}, \mathbf{g_{ij}}|d_{ij}=1\ )Pr(d_{ij}=1)\ \ +\ \ L(y_j, \mathbf{x_i}, \mathbf{g_{ij}}|d_{ij}=0\ )Pr(d_{ij}=0)\ \}$$

$\mathbf{g_{ij}}$ is a vector of binary agreements and disagreements between record i in File A and record j on File B underlying the matching process. Later in the paper it is assumed that $\mathbf{g_{ij}}$'s distribution is independent of y_j and $\mathbf{x_i}$, which allows the probabilities for $\mathbf{g_{ij}}$ and d_{ij} to be done on their own based on standard methods for PRL. The likelihood portion for $d_{ij}=0$ (no match) is much simpler: y_j and $\mathbf{x_i}$ are assumed to be independent.

A likelihood function which is based on all possible pairs between two files can certainly be massive. Hof and Zwinderman (2015) work with a realistic real world example: matching first and second pregnancies from the Perinatal Registry Netherlands registry. There were 393,302 first deliveries and 312,871 second deliveries, so the total number of potential record pairs was 1.2×10^{11}. They reduced this considerably by taking only the roughly 100,000 pairs with a predicted probability of matching greater than 0.01. This type of fudge would be necessary in practice to make this procedure practicable, unless the data sets are small.

Appendix F

Detailed Review of Papers on Data Fusion

Appendix F
Detailed Review of Papers on Data Fusion

F.1 Data Fusion: Moriarity and Scheuren (2001)

Moriarity and Scheuren (2001) provide an overview of data fusion (statistical matching). Their starting point is a paper by Kadane from 1978 which is reprinted with their 2001 paper.

Kadane (2001) sets out a theoretical approach for the basic scenario of having one file with records with an **X** vector and a **Y** vector, and a second file with records with an **X** vector and a **Z** vector, with the final goal of doing analysis of the joint distribution of (**X**, **Y**, **Z**). One can match entirely on **X** (taking records with the same or very similar **X** vector values), and put together on a synthetic record the components $\mathbf{y_1}$ from the one file record and $\mathbf{z_1}$ from the other file record. Creating a synthetic file in this way will produce a file that has the right marginal distributions for **X**, **Y**, and **Z**, and the right correlations for **X** and **Y**, **X** and **Z**, but the file will effectively have **Y** and **Z** conditionally independent given **X** (an artifact of the file's construction).

Kadane (2001) admits there is no information available from the two files themselves about the right conditional distribution between **Y** and **Z**. Progress can only be made by making assumptions about the conditional correlation between **Y** and **Z**. Kadane puts forward a Bayesian approach for eliciting a prior for this correlation. The correlations between **X** and **Y** and **X** and **Z** force certain bounds on the possible correlations of **Y** and **Z** (for example, if the correlations between univariate X and Y and univariate X and Z are both very high, then it is not possible for the correlations between Y and Z to be close to 0), but these constraints may in many cases be quite loose. Once a correlation is designated (one draw from the prior distribution), one can proceed to construct an artificial data set. For the file with **Z** missing, one can produce the conditional expectation for **Z** given $\mathbf{x_i}$ and $\mathbf{y_i}$, using the assigned covariance matrix relating $\mathbf{x_i}$, $\mathbf{y_i}$, and $\mathbf{z_i}$, and for the file with **Y** missing, the conditional expectation for **Y** given $\mathbf{x_i}$ and $\mathbf{z_i}$, using the assigned covariance matrix relating $\mathbf{x_i}$, $\mathbf{y_i}$, and $\mathbf{z_i}$ (Kadane works within multivariate normality so that the full joint distribution is determined by conditional expectations and conditional variances and covariances). The resultant augmented Z-missing file has records $\mathbf{w_j} = (\mathbf{x_j}, \mathbf{y_j}, \hat{\mathbf{z}}_\mathbf{j})$, (with $\hat{\mathbf{z}}_\mathbf{j}$ the conditional expectation), and the augmented Y-missing file has records $\mathbf{v_i} = (\mathbf{x_i}, \hat{\mathbf{y}}_\mathbf{i}, \mathbf{z_i})$, (with $\hat{\mathbf{y}}_\mathbf{i}$ the conditional expectation). The matching of the two augmented files is done using a Mahalanobis distance $d_{ij} = (\mathbf{w_j} - \mathbf{v_i})'(\mathbf{S_1} + \mathbf{S_2})^{-1}(\mathbf{w_j} - \mathbf{v_i})$ where $\mathbf{S_1}$ and $\mathbf{S_2}$ are the variance matrices of $\mathbf{v_i}$ and $\mathbf{w_j}$ respectively. The matching process then produces the synthetic file with records ($\mathbf{x_i}$, $\mathbf{y_i}$, $\mathbf{z_i}$) created by pairing the closest $\mathbf{w_j}$'s and $\mathbf{v_i}$'s.

Moriarity and Scheuren (2001) criticize this approach as not really reproducing the prior correlation between **Y** and **Z**. In simulation studies they show that it produces a synthetic file with a different correlation between **Y** and **Z** than the posited correlation. They fix this inadequacy by proposing a methodology in which the missing **Y** values and the missing **Z** values are imputed first using conditional distributions based on the prior draw of the **Y**,**Z** correlation. Once the files are augmented in this way, the matching procedure will produce a synthetic file which retains the marginal distributions of **X**, **Y**, and **Z**, the joint distribution of (**X**,**Y**) from the Z-missing file, and the joint distribution of (**X**,**Z**) from the Y-missing file, as well as the posited (**Y**,**Z**) correlation.

F.2 Data Fusion: Rässler (2002)

Rässler (2002) presents an overview of statistical matching her monograph. In her Section 2 she presents an overview of the 'frequentist statistical matching' approach which matches data vectors (x_i, z_i) and (y_i, z_i) on matching vector z_i[5]. Rässler makes a distinction between the true joint pdf

$$f(x, y, z) = f_{Y,Z}(y, z) f_{X|Y,Z}(x|y, z)$$

and the induced joint pdf from the statistical matching operation:

$$\tilde{f}(x, y, z) = f_{Y,Z}(y, z) f_{X|Z}(x|z)$$

Note that the induced pdf has a conditional probability of x on z rather than a conditional probability of x on y, z as would be correct. Likewise the true and induced covariances of **X** and **Y** are, respectively,

$$Cov(X, Y) = E\{Cov(X, Y|Z)\} + Cov\{E(X|Z), E(Y|Z)\}$$
$$\widetilde{Cov}(X, Y) = Cov\{E(X|Z), E(Y|Z)\}$$

In Section 3, Rässler (2002) presents an overview of traditional approaches including data fusion in Europe and in the US and Canada. Included are 'unconstrained matching', 'constrained matching', 'categorically constrained matching', as well as clustering methods. Suppose for example we have one file ('recipients') with eight records and x and z fields, and a second file ('donors') with six

[5] Note here we match the notation in Rässler (2002) and contradict the notation in Section 3.1 and E.1 (to keep faith with the reference).

records and y and z fields. Under 'unconstrained matching' we create a file with eight records based on the recipients matched by zi vectors on the donors, with donors being taken with no regard as to how many times a donor is matched. This leads to a distortion of the marginal y distribution, which is perceived as problematic by some researchers.

Constrained matching controls the selection of donors in such a way as to preserve the y distribution in the linked file. This can be done by 'exploding the donor file': creating multiple donor records from the initial donor file, and then allocating these multiple donors carefully to retain the marginal distribution. Categorically constrained matching tightens the matching process by categorizing the fields and tightening further the matching process, possibly also including auxiliary information from external sources for calibration. Rässler presents references for each of these approaches.

In her Section 4.5 Rässler (2002) presents a solution to the statistical matching problem through Bayesian multiple imputation, which is called 'non-iterative Bayesian-based multivariate imputation'. The two files are combined, and the missing **X** data in the X-missing file and the missing **Y** data in the Y-missing file are assumed to be the blocks of data which are 'missing at random' conditional on **Z**.

A multivariate normal model is assumed with noninformative priors, except for the conditional correlation of **X** and **Y** given **Z**. This conditional correlation cannot be estimated in any way from the data, and has to be assigned in a fairly arbitrary way, using priors or relevant auxiliary information of some type. With this input, all other component random variables in the model have posterior distributions generated in the usual way for the multivariate normal model with noninformative priors: variances have as their posterior mean sample variances, regression parameters are drawn from standard posterior distributions, covariances are drawn after the critical **X**,**Y** conditional correlation is set (or drawn from a prior distribution), and then the missing **X** and **Y** blocks are drawn from the appropriate posterior distributions conditional on draws of parameters from their posterior distributions.

Appendix G

Detailed Review of Papers on Imputation

Appendix G
Detailed Review of Papers on Imputation

G.1 Imputation: Self-Reports and Clinical Data from NHIS and NHANES.

Raghunathan (2006) and Schenker et al. (2010) work with a problem similar to the OSHA/SOII application. The self-report items for both NHIS and NHANES provide self-reported hypertension, diabetes, and obesity for sampled persons (the incidence is based on how persons answer particular questions: for example for obesity the incidence is based on self-reported height and weight; for diabetes on direct questions about the disease). In all of these cases, the self-report incidence is lower than the clinical incidence from the NHANES clinical evaluations.

Schenker et al. (2010) create then something that fills in for a clinically based evaluation health status for the NHIS for these three health statuses, which will not suffer from the self-reporting bias. They do this through generating multiple imputations on the NHIS for each record using the self-report status and other covariates which are shared between NHIS and NHANES. These multiple imputations are based on fitting a model to the clinically based evaluations in NHANES, using as predictors the self-report statuses and other covariates. There is no direct linking between NHIS and NHANES records. The model though depends on NHANES having the same self-report questions for the NHANES records as NHIS that can be used to generate a completely relevant model. There is no overlap of records, but complete overlap of the questionnaire items allowing for the construction of an NHANES-based multiple imputation model that is completely relevant for the NHIS records.

One special aspect of this model is that they fit the model on subsets of the NHANES data set which are defined using a 'propensity for being an NHIS sample record'. Each NHANES observation is assigned a propensity for being in the NHIS sample based on a logistic regression model using covariates common to both surveys. These predicted propensities then are used to define the subsets. Separate clinical-evaluation propensity models are fit for each subset. The motivation for this is the fear that the model may not be specified perfectly, and differences in the distributions of the covariate space between NHANES and NHIS could result in the model fit from NHANES extrapolating poorly to portions of the NHIS covariate space not well-represented in NHANES. By doing local model fits on these subsets, the models that are fit to NHANES within the subset should be within the covariate bounds of the corresponding NHIS data subset, reducing

biases from extrapolation. This type of thing is done in pseudo-experimental studies matching pseudo-treated to pseudo-controls, where the propensity used is the propensity to be in the treatment group. Schenker et al. (2010) apply that theory in this application.

Raghunathan (2006) is an earlier version of this theory which is proposed to deal with the general issue of self-reported data, and applies it to the NHIS-NHANES case for hypertension status. The later Schenker et al. (2010) research develops the idea fully.

G.2 Imputation: Incomplete Data on Hospice-Use from the CanCORS.

Table G-1 from He et al. (2014) below summarizes their data set. Information on use of a hospice can come from Medicare claims of the patient if the patient is 65+ years old or from medical records (e.g., doctor, hospital). There is missing data in both sources, and the sources contradict each other.

Table G-1. Table 1 from He et al. (2014): 3,027 CanCORS lung and colorectal cancer patients who died within 15 months of diagnosis.

Whether Patient Utilized Hospice Services	Medicare claims Yes	Medicare claims No	Medicare claims missing
Medical records Yes	395	54	260
Medical records No	445	617	646
Medical records Missing	136	116	358

He et al. (2014) construct a latent variable model where Y_O is the true value (equal to 1 if the patient used hospice services, 0 if the patient did not), and Y_{R1} and Y_{R2} are the reported values from Medicare claims and medical records respectively. There is a set of covariates X which include the cancer type, cancer stage, sex, age, race, gender, education, and other health aspects. There is a random effect for site. The four aspects of their model were:

- A probit regression model for the true Y_O predicted by the covariates;
- A probit regression model for Y_{R1} given Y_O and the covariates;
- A probit regression model for Y_{R2} given Y_O and the covariates.
- If Y_O=0, then Y_{R1}=Y_{R2}=0.

Since Y_O is unobservable, a latent variable type Bayesian framework is necessary. The fourth assumption is a key assumption, and allows for the identifiability of the model according to He et al. (2014). This assumption assumes no 'overreporting'. If the patient has not utilized hospice, then neither Medicare nor the medical records will indicate hospice use. The direction of an error is assumed to be always in the direction of underreporting: true hospice use may not be indicated in either a Medicare claim or a medical record, for a variety of reasons. A missing value on the other hand can mean anything. He et al. (2014) in their discussion section point to other literature where this assumption is not made (in similar applications), and weak identifiability of the models results.

G.3 Imputation: Combining Cancer Registry Data with Followup Survey Data

The Yucel and Zaslavsky (2005) approach is as follows. Define $y_{(R)1}$, $y_{(R)2}$ as the cancer-registry reported chemotherapy incidence in 'S1' (the validation sample), and 'S2' (the remainder of the relevant years' cancer registry data not sampled in the validation sample), respectively, and define $y_{(O)1}$, $y_{(O)2}$ as the true chemotherapy incidence from the two data sources S1 and S2 reported by the presiding physician. Note that we assume $y_{(O)1}$ is fully observed for the respondents to the validation survey, and $y_{(O)2}$ is not observed.

Yucel and Zaslavsky (2005) proceed by effectively imputing $y_{(O)2}$ for all members of S2. Note that $y_{(O)1}$ must also be imputed for nonrespondents to the validation survey. With this mass imputation of chemotherapy onto the cancer registry, they proceed to fit a model of two-year survival on the registry with chemotherapy as one predictor (the other predictors are in better shape in terms of reliability on the cancer registry). Note that sometimes reported values in the registry $(y_{(R)1})$ are effectively overruled by the validation study $(y_{(O)1})$.

They make one key assumption very similar to that made by He et al. (2014) as described in Section 7.2: that a report on the cancer registry of a patient receiving chemotherapy ($y_{(R)1} = 1$ or $y_{(R)2} = 1$) must correspond to a true chemotherapy ($y_{(O)1} = 1$ or $y_{(O)2} = 1$). Errors in the cancer registry are only in the direction of failing to report chemotherapy that has occurred, not falsely reporting chemotherapy that has not occurred. Beyond this assumption, the model was a fairly standard Bayesian random effects probit regression.

He and Zaslavsky (2009) further extend the Yucel and Zaslavsky (2005) approach to a multivariate y-variable (a vector of L therapy outcomes, rather than a single outcome).

G.4 Mass Imputation: Combining Complex Surveys by Imputing to the Full Population

Dong et al. (2014b) propose combining surveys through recreating the frame by filling in by imputation all of the unsampled (and sampled but nonresponding) population members. They utilize a nonparametric Bayesian Bootstrap for this: drawing with replacement from the empirical sample distribution. Unequal weighting and unit nonresponse is dealt with by including final weights in this bootstrap. Clustering is dealt with by bootstrapping entire clusters. Once this process is complete, one has a full bootstrapped frame from the survey that can be combined with the bootstrapped frame from the other surveys.

Dong et al. (2014b) apply this theory to estimate health insurance prevalence: the percentage of persons who have no health insurance, those who are publicly insured (Medicare or Medicaid primarily), and those who are privately insured. The three sources are the National Health Insurance Survey (NHIS), the Medical Expenditure Panel Survey (MEPS), and the Behavior Risk Factors Surveillance System (BRFSS). These are all bootstrapped up to the population level and a combined frame created, and an estimator of health insurance prevalence is generated from the combined frame. Dong et al. (2014a) is a companion paper from the same authors that provides details on the Bayesian Bootstrap.