Multiple imputation for national public-use datasets and its possible application for gestational age in United States Natality files

Jennifer D. Parker and Nathaniel Schenker

National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

Correspondence:
Jennifer D. Parker, Office of Analysis and Epidemiology, Room 6107, National Center for Health Statistics, Hyattsville, MD 20782, USA.
E-mail: jdparker@cdc.gov

Conflicts of interest:
the authors have declared no conflicts of interest.

Summary

Parker JD, Schenker N. Multiple imputation for national public-use datasets and its possible application for gestational age in United States Natality files. *Paediatric and Perinatal Epidemiology* 2007; **21**(Suppl. 2): 97–105.

Multiple imputation (MI) is a technique that can be used for handling missing data in a public-use dataset. With MI, two or more completed versions of the dataset are created, containing possibly different but reasonable replacements for the missing data. Users analyse the completed datasets separately with standard techniques and then combine the results using simple formulae in a way that allows the extra uncertainty due to missing data to be assessed. An advantage of this approach is that the resulting public-use data can be analysed by a variety of users for a variety of purposes, without each user needing to devise a method to deal with the missing data. A recent example for a large public-use dataset is the MI of the family income and personal earnings variables in the National Health Interview Survey. We propose an approach to utilise MI to handle the problems of missing gestational ages and implausible birthweight–gestational age combinations in national vital statistics datasets. This paper describes MI and gives examples of MI for public-use datasets, summarises methods that have been used for identifying implausible gestational age values on birth records, and combines these ideas by setting forth scenarios for identifying and then imputing missing and implausible gestational age values multiple times. Because missing and implausible gestational age values are not missing completely at random, using multiple imputations and, thus, incorporating both the existing relationships among the variables and the uncertainty added from the imputation, may lead to more valid inferences in some analytical studies than simply excluding birth records with inadequate data.

Keywords: *missing data*, *gestation*, *multiple imputation*, *birth records*.

Introduction

Accurate information on the length of pregnancy in vital records is necessary for determining preterm delivery rates, creating fetal growth curves, and other programmatic and research purposes. Missing data and inaccuracies have affected the utility of gestational age information on the US Natality public-use datasets produced by the National Center for Health Statistics (NCHS) for some time.[1,2] In the 2003 public-use file, for example, approximately 1% of birth records have no gestational age stated.[2] In addition, analysts using the NCHS public-use files often identify a small percentage of birth records with gestational age–birthweight combinations that are deemed inconsistent;[3–6] the number flagged as inconsistent in an analysis depends on the method of determination.

As infants whose birth records have implausible data appear more likely to be *high risk* than infants with complete and clinically plausible data[7,8] – i.e. they are not a simple random sample of the whole cohort – simply deleting the implausible records may lead to selection bias.[9–11] A study of exclusion methods used to address implausible gestational age, for example, showed that although some associations were

unaffected by record exclusion, gestation-specific relative risks of infant mortality for high-risk vs. low-risk mothers varied considerably by method of exclusion.[8] Furthermore, inasmuch as risk factor information is available on birth records, retention of cases with missing and implausible gestational age data could be potentially useful for our understanding of high-risk births.

Our paper focuses on the potential use of multiple imputation (MI) for the problem of missing and inaccurate gestational age data in the US Natality public-use datasets. With MI, we propose to view the dual issues of missing and inaccurate gestational age data within the framework of a 'missing data' problem. We do not suggest any modifications to current data editing or imputation procedures for the public-use datasets; rather, we propose an approach for handling the missing and suspect gestational age values that remain in the dataset after the current in-house editing and imputation have been accomplished. The results of the MI would reside in auxiliary public-use datasets, separate from the original files.

This paper is organised as follows. First, we introduce MI and provide examples where the method has previously been applied to public-use datasets. Next, we turn to implausible and missing gestational age records, briefly describing some approaches for the identification of possibly inaccurate records. Finally, we combine these ideas and offer a proposed outline for the creation of public-use analytical datasets with MIs of gestational age.

Overview of multiple imputation

Imputation refers to assigning a value for each missing observation, allowing the retention of all records in analyses. This approach to missing data has additional advantages. First, using the same approach to handling missing data on public-use datasets provides consistency across different scientific questions and objectives, as well as across varying degrees of statistical expertise and computing power. As a result, inferences can be assessed in the context of different questions and analytical approaches without the confusion that arises from the different ways of handling missing data. Another advantage of imputation for public-use datasets is the potential for using information in the imputation procedure that may not be available on the public-use files, such as geographical detail. These reasons, among others, led Rubin to conclude that 'modelling the missing data must be, in general, the data constructor's responsibility'.[12] In the case of public-use US Natality files, the data constructor would be the NCHS.

Many different approaches have been used to impute missing data. For example, a rather naïve approach replaces each missing value on a variable with the variable's overall mean. A more complicated approach uses multiple regression equations estimated from the set of observations with complete cases to obtain predicted values for the missing observations. Hot-deck imputation replaces each missing value with that from a similar record (known as a donor case) which has a valid reported value for the element of interest. Currently, the NCHS imputes a small number of gestational age values for the US Natality public-use datasets using a hot-deck method when the month and year of the last menstrual period (LMP) are present but the day is missing.[1,13]

Little gave a detailed discussion of issues in creating imputations for large datasets.[14] Two major considerations are that all observed values should be taken into account to the extent possible, and that random draws of the missing values from an appropriate distribution should be used. If observed variables are not taken into account, biases can occur to the extent that the missing data depend on the variables. Replacing missing values by point estimates (e.g. means or regression predictions) rather than random draws can distort estimates of quantities that are not linear in the data, such as variances and correlations.

A primary disadvantage of *single* imputation, that is, imputing only one value for each missing observation, is that typically the imputed values are treated as if they were true values in analyses, so that point estimates, their estimated variances and subsequent inferences do not adequately reflect the added uncertainty due to the assignment of a plausible, yet not actual, value for each missing datum. This is true even if random draws are used and all observed values are taken into account. MI is the assignment of two or more values for each missing datum.[11,12,15,16] The additional imputations, generally drawn at random from an appropriate distribution, enable the incorporation of the uncertainty of the imputation procedure into the analysis. Often five imputations are sufficient for adequate estimation.[12]

Creation of multiple, say 5, imputations for the missing values results in 5 completed datasets. (The values that are not missing stay constant across the 5

datasets.) An analyst of the multiply imputed data carries out the estimation procedure that would have been used for complete data 5 times, once with each completed dataset. The result is 5 point estimates and their 5 estimated variances. The final MI point estimate is just the average of the 5 point estimates, and the final MI variance estimate is a simple combination of the average of the 5 variance estimates and the variance of the 5 point estimates. (The inclusion of a component for the variance of the point estimates across imputations expresses the uncertainty due to missing data.) Although an MI analysis can be carried out by entering the results of the 5 analyses into a spreadsheet and then using spreadsheet functions to implement the simple rules for combining the 5 sets of results (or, alternatively, by using a macro within the software package that would be used for complete data), the procedure can be cumbersome to apply. Fortunately, several statistical software packages, including Stata,[17] SUDAAN[18] and SAS,[19] have capabilities to analyse multiply imputed data more automatically.

In MI, each set of imputations for the missing values in a dataset is ideally drawn from an approximate predictive distribution conditional on the observed values. To implement this process, distributional parameters are estimated from a statistical model developed using observed data. The parameter estimates, in turn, can be used to predict plausible replacement values for the missing data. The model used for imputing the missing data can be quite complicated, but the essential idea for drawing a single set of imputations can be illustrated with a simple example.

Consider a dataset of n cases and two variables, gestational age Y and birthweight X, for which m cases are missing gestational age data and $n–m$ have complete data. If we fit a simple linear regression model, $Y = \beta_0 + \beta_1 X + \varepsilon$, where ε has a normal distribution with mean 0 and variance σ^2, to predict gestational age as a function of birthweight using the $n–m$ cases with complete data, we can create a set of imputations for the m cases with missing gestational ages in two steps. In the first step, values β_0^*, β_1^* and σ^{*2} are drawn randomly from the joint posterior distribution of the regression parameters. (This can be approximated by using a scaled inverse chi-square distribution for drawing σ^{*2} and a bivariate normal distribution for drawing β_0^* and β_1^* given σ^{*2}.[20])

In the second step, for each of the m cases with missing gestational age, say case i, we impute the missing value of Y using $Y_i^* = \beta_0^* + \beta_1^* X_i + \varepsilon_i^*$, where X_i is the value of X for case i, and ε_i^* is drawn from a normal distribution with mean 0 and variance σ^{*2}. The first step reflects the uncertainty due to the fact that the imputation model was fitted to just a sample of data, and the second step reflects the uncertainty given the fitted model. To create multiple, say 5, imputations for the missing values, the two-step procedure is repeated independently 5 times.

Examples of multiple imputation for public-use datasets

MI has been used to handle missing data in the context of public-use files in a number of applications.[16] In addition to the need to retain the general-use characteristic of national public-use datasets, these datasets tend to be large and often have complex survey designs, which complicate the creation of multiply imputed files. In a recent multi-year project at the NCHS, MIs were created for missing family income and personal earnings data in the National Health Interview Survey (NHIS). The NHIS is an annual household interview survey designed to monitor the health of the US population through the collection and analysis of data on a broad range of health topics, tabulated by a number of demographic factors. Information from the NHIS, for instance, aids our understanding of relationships between income and health, including the health and health care of low-income populations. However, in recent years, detailed family income values have been missing for nearly one-third of survey respondents.[21,22]

The creation of the imputed income files was complicated by several factors, including information having been collected at both the person and family level, structural dependencies among the variables in the survey (e.g. unemployed individuals have no earnings), and the necessity of placing reasonable bounds on the imputed values. Also, the predictor variables used to create the imputations were of many types (categorical, continuous, ordinal) and had small percentages of missing values which needed to be imputed as well. An iterative procedure based on publicly available software (IVEware, available at http://www.isr.umich.edu/src/smp/ive) was used to create the MIs. Although imputations of several variables in the NHIS were created as part of the procedure, and in-house geographical variables were used in the models, only family income (and its ratio to the Federal poverty threshold), employment status and

personal earnings variables were released for public use. Since the release of the imputed income files on the Internet, several hundred downloads for data years 2002 and 2003 have been tallied (personal communication, Diane Makuc, NCHS).

An earlier application to a national health survey was the MI of several variables in the Third National Health and Nutrition Examination Survey (NHANES III).[23] The NHANES III MI research project began in 1992 when the NCHS assembled a team of expert statisticians to evaluate different options for handling missing data in the NHANES datasets. Over the next several years, various imputations and simulations of the NHANES III data files were evaluated.[24] In 2001, a public-use version of the file was released for research purposes, with many key variables multiply imputed.[23,24] As with other in-house imputation efforts, intended for public-use files, the NHANES III imputations were developed using detailed demographic and geographical information unavailable to the general public.

Additional MI applications for US public-use data include handling missing data in the Survey of Consumer Finances (SCF),[25,26] calibrating industry and occupation codes across censuses,[27] and handling missing blood alcohol test results in the Fatality Analysis Reporting System.[28] Public-use files from the SCF, a survey conducted every 3 years by the U.S. Federal Reserve Board, have been released with MIs for missing data since the 1989 survey.[26]

Identification of implausible values for gestational age

Like income in the NHIS, gestational age is an important variable for assessing health disparities. Furthermore, its imputation for US Natality data is complicated by several intrinsic factors, including large and small differences in data availability among States in this State-based data system. In the 2003 public-use dataset, for example, data were reported by some States using the revised U.S. Standard Certificate of Live Birth, while other States reported data using the older certificate.[2] As in other MI projects, relevant variables from the Natality dataset are in different formats and are subject to their own missing values. However, the biggest complicating factor in the imputation of gestational age is the uncertainty about which records need to be imputed due to implausible values.

As mentioned earlier and described more fully elsewhere, the birthweight and gestational age relationship identifies implausible couplings of the two variables that suggest reporting error. For early gestational ages, birthweight appears to follow a bimodal distribution, with the primary distribution centred on a clinically reasonable birthweight and a secondary distribution, to the right of the primary, with a mean birthweight more consistent with higher gestational ages. Less commonly, at higher gestational ages, the mean birthweight is lower than would be expected assuming monotonically increasing fetal growth, suggesting probable misclassification of some term births to post-term gestational ages. Although measurement and reporting errors in birthweight, rather than gestational age, are possible for an individual birth record, this scenario is deemed generally less likely than errors in recorded gestational age, as birthweight is measured and recorded at birth whereas LMP, the most commonly used measure of gestational age, is subject to recall bias. The clinical estimate of gestation, used as the recorded gestational age in approximately 5% of records on the public-use Natality dataset, is subject to other problems, including being inconsistently obtained and reported.[29]

A variety of solutions have been proposed to handle suspected inaccuracies in gestational age reporting for birth records with unusually high or low birthweights. Some solutions have been specifically applied towards a particular study question, such as the creation of fetal growth curves[3–5,30–33] or the understanding of preterm delivery rates between groups[6,34] or over time.[34,35] One general approach is to define criteria for plausibility, excluding or reassigning records that fall outside reasonable bounds.[3,4,7] For example, Zhang and Bowes used the assumption that birthweight follows a normal distribution at each gestational age and deleted observations, or, when available, substituted clinical estimates of gestation when birthweights deviated from a normal probability plot.[4] Studies that have compared these truncation approaches have concluded that higher-risk births are more likely to be flagged for deletion or substitution than lower-risk births, potentially affecting analyses of temporal trends and other comparative studies of pregnancy outcomes.[7,8]

Other work to identify inaccuracies has taken a mixture-model approach[36–38] to describe the observed primary and secondary modes of birthweight, using the results to calculate adjusted statistics, delete

records, or reassign gestational ages. Mixture models are fitted when the data are thought to come from two or more subpopulations. In addition to separate parameters for each subpopulation, mixture models include 'mixing proportions' to estimate the relative contribution of each subpopulation to the overall population. In mixture models for birthweight described below, one subpopulation, referred to as the primary component, comprises the infants with accurately recorded gestational ages. Additional subpopulations, the secondary components, are composed of infants with inaccurately recorded gestational ages. Specifications for the secondary components have included: limiting births to gestational ages of ± 4 weeks relative to those recorded,[35] limiting births to a single alternative centred at term or 40 weeks' gestation,[39] or assuming births come from some unspecified gestational age.[33]

In the context of public-use data, differences among various users in approaches and models for handling implausible gestational ages could very well lead to various studies with valid inferences; however, comparisons among the studies might be hindered by the lack of comparable adjustments for implausible gestational ages. Thus, MI by the data producer could have the advantage of helping to promote comparability.

Multiple imputation of gestational ages

We propose to use MI to handle both implausible and missing gestational ages in the US Natality public-use files, which is probably more complicated than handling only missing gestational ages or identifying implausible records with a truncation approach. This is in the spirit of recent work by Ghosh-Dastidar and Schafer,[40] termed 'multiple edit/multiple imputation', in which MI was used to adjust for both measurement error and missing data, and of earlier work by Little and Smith,[41] who illustrated editing and imputation of outlying data values with the Annual Survey of Manufacturers.

An application of MI of gestational ages in a specific study was performed by Hediger and colleagues, who investigated children's growth using the NHANES III (unimputed files) linked to birth certificate data.[42] Their objective was to understand associations between size at birth and growth markers later in childhood. Some of the linked birth records, needed to calculate relative size at birth, had incomplete or improbable gestational age information, and it was thought that limiting the study to complete cases could bias inferences. To address this issue prior to their analysis, Hediger and colleagues applied the Zhang and Bowes criteria[4] for determining improbable birthweight–gestational age combinations and then performed MI for the missing and flagged records using a standard regression model. Their comparison of the MI and complete-case results gave them confidence to present MI results in an article.

For the Natality public-use data, which is intended for users with diverse objectives, we will probably extend the approach of Hediger and colleagues in two ways. First, a more detailed model for predicting implausible and missing gestational ages will probably be used. As discussed by Rubin, it is desirable in principle for an imputation model for missing data in a public-use dataset to include as predictors all variables that will ultimately be used by analysts of the data (although this ideal is seldom achieved completely).[12] Second, we will probably take steps to reflect uncertainty in deciding which gestational ages are implausible.

A complete application of MI for the gestational age problem would allow the assessment of uncertainty due to: (a) not knowing precisely which reported gestational ages were correct, (b) not knowing the true values for the incorrectly reported gestational ages, and (c) not knowing the values for the unreported gestational ages. To account for source (a), for each reported gestational age, an indicator variable Z of correctness, with $Z = 1$ if the gestational age is correct and $Z = 0$ if it is not correct, would be imputed. Then, to account for source (b), for each reported gestational age with imputed value $Z = 0$, an alternative gestational age would be imputed. (If $Z = 1$ were imputed, the reported gestational age would not be replaced.) Finally, to account for source (c), for each unreported gestational age, a value would be imputed. Creation of multiple, say 5, imputations would involve repeating the imputation sequence independently 5 times. Note that in addition to the imputed values of gestational age being allowed to vary across the MIs, the imputed correctness statuses (Z) for each case with a reported gestational age would be allowed to vary as well. Thus, for one imputation, a case may have $Z = 1$ and thus retain its reported gestational age, whereas for another imputation, the case may have $Z = 0$ and thus have an imputed gestational age. It is through this variability in Z that uncertainty about the correctness of each case would be reflected in the MIs.

The creation of imputations would involve models for predicting whether a reported gestational age is correct, for predicting the value of an incorrectly reported gestational age, and for predicting the value of an unreported gestational age. As described earlier, the model for correctness of reported gestational ages could be a mixture model for the primary and secondary distributions of birthweight given gestational age, which could be used to estimate the probability of each case's gestational age being correct. Variations on the mixture model could include the addition of covariates, the use of specific and non-specific gestational age alternatives for the secondary distribution, and the use of alternative forms of the model. Alternatively, other methods for estimating the probability that a particular birthweight–gestational age combination is plausible could be considered.

Regression-type models could be used both for predicting the value of an incorrectly reported gestational age and for predicting the value of an unreported gestational age. Some variables available on the US Natality dataset likely to be relevant for these models include: infant birthweight, the clinical estimate of gestational age, prenatal care indices, marital status, maternal race, birth order, maternal education and maternal age. Moreover, as mentioned above, imputations have been found to be more useful if they are created using as many of the variables that could be used by future analysts as possible. As studies of neighbourhood and contextual effects on birth outcomes are increasing, for example, including county-level covariates, such as median income, could improve the imputations. Variations on the regression models could include the addition, deletion or transformation of covariates, as well as the use of alternative forms of the model and variance structure.

To summarise, an outline of a potential algorithm for imputing for implausible and missing gestational ages is as follows:

1 Fit a mixture model using the birth records with reported, even though possibly inaccurate, gestational ages.
2 Use the fitted mixture model from step 1 to impute the value of an indicator variable Z of correctness for each reported gestational age, where $Z = 1$ if a gestational age is correct and $Z = 0$ if it is not correct. The probability of $Z = 1$ would be calculated from the relative likelihoods of a particular record being in the primary distribution and considered correct, or being in the secondary distribution and considered incorrect. (Note that imputing a set of indicators for all of the birth records with reported gestational ages would involve the two-step procedure for drawing from a predictive distribution that was outlined in the *Overview of Multiple Imputation* section.)
3 Fit a prediction model for gestational age using the birth records from step 2 with imputed values $Z = 1$.
4 Use the fitted prediction model from step 3 to impute values of gestational age for the birth records with reported gestational ages but imputed values $Z = 0$, as well as for the birth records with unreported gestational ages. (Once again, the two-step procedure for drawing from a predictive distribution would be used.)

The algorithm just outlined is based on the use of a mixture model to help identify implausible gestational ages in step 2, but not to assign alternative values for cases with imputed values $Z = 0$ in step 4 (the alternative values are imputed using a separate prediction model). Thus, the way a mixture model is used in the preceding algorithm is in the spirit of the work by Platt *et al*.[39] and Tentoni *et al*.[33] If instead, in the spirit of Oja *et al.* [35] we wanted to use a mixture model directly to assign alternative values for cases with imputed values $Z = 0$, then steps 2–4 could be modified as follows:

2 (a) Use the fitted mixture model from step 1 to impute the value of an indicator variable Z of correctness for each reported gestational age, where $Z = 1$ if a gestational age is correct and $Z = 0$ if it is not correct. (b) For each birth record with $Z = 0$, use the mixture model to impute an alternative value of gestational age.
3 Fit the prediction model for gestational age using the birth records from step 2, with the reported gestational ages for those records with imputed values $Z = 1$ and the imputed gestational ages for those records with imputed values $Z = 0$.
4 Use the fitted prediction model from step 3 to impute values of gestational age for the birth records with missing gestational ages.

(As mentioned earlier, creating the imputations in steps 2 and 4 would involve the two-step process outlined in the *Overview of Multiple Imputation* section.)

Discussion

The public-use product would be a small set of completed datasets (say 5, one for each of the MIs), each containing three variables: (1) a record identifier; (2) a categorical indicator of whether or not the gestational

age had been imputed and, if it had been imputed, whether the original gestational age had been missing or implausible; and (3) the corresponding imputed or original gestational age value. Each of the completed datasets could be linked to the original public-use file to obtain covariates, as well as original gestational age values for those that were imputed. As mentioned previously, data users would replicate their analysis for each completed dataset and combine the results using established methods, either with their own calculations or with available software.[17–19] Although resources for MI would be targeted towards current datasets, once the methodology and programs were developed, the creation of multiply imputed datasets for earlier years for the analysis of trends would be straightforward.

To evaluate the MI approach, comparative analyses for a small number of pregnancy outcomes and demographic subgroups would be needed,[7,8] especially for the pregnancy outcomes and subgroups most carefully observed for clinical and public policy. Finding the right imputation models would be the biggest challenge of the project. If the models were very wrong, analyses using the multiply imputed files could be subject to *information bias,* that is, the use of gestational age values that do not, in fact, correspond to the distribution of true gestational ages.[9,10] However, the 'model' that specifies that the missing and implausible data should be deleted prior to analysis can lead to *selection bias*.[9,10] Inasmuch as the prediction models incorporate information about the underlying associations, using the imputations would seem to be less biased, on average, than the unimputed data, although the actual effects of MI on sources of bias would depend on the particular study. Furthermore, the assumptions underlying the imputation method need only affect the final inferences through their effect on the imputed values, not for the whole dataset. A full MI of gestational age on US Natality datasets would be an iterative process, repeating the imputation and evaluation steps. The least complex formulations of the models would be considered and understood first. More complicated formulations could follow. The process would enable us to assess the sensitivity of the models and the results under different scenarios and assumptions as the analysis progresses.

An important use of gestational age data is to assess preterm delivery rates.[5,34] As noted by others, preterm delivery rates based on ultrasound data are higher than those based on LMP; that is, more term infants are identified as preterm than preterm infants identified as term when early ultrasound information is compared with LMP.[43–47] The reasons for discrepancies between LMP and ultrasound estimates are not fully known, but could be due, in part, to associations between growth and gestational age[47] or bleeding early in pregnancy.[48] However, like the data-based editing methods described previously which delete or replace suspect gestational ages,[3,4,6,33,35,36] MI would be likely to lower the overall preterm delivery rate by replacing implausibly low gestational ages with higher gestational age values. MI could, nevertheless, provide less biased comparisons of preterm delivery over time as well as across demographic and geographical groups. Many factors associated with preterm delivery are also associated with implausible and missing gestational age data.

Gestational age information is also used to create fetal growth curves (birthweight-for-gestational-age references). In turn, upper and lower percentiles of growth are estimated from the curves to identify growth restricted and macrosomic infants. Consequently, care is needed to ensure that the imputations do not smooth away the actual clinical variability and unnecessarily limit the range of birthweights at each gestational age. Single imputations, based on assigning the most likely value of gestational age – be it 40 weeks, ± weeks, or another value – are at greater risk of oversmoothing the distributions than MI, which allows for more variability, albeit not necessarily at the extreme ends of the distribution.

A related concern with imputation (either single or multiple) is the potentially large number of births that would be reassigned into a more likely gestational age despite having an observed gestational age that is clinically possible; this reassignment would be more common for infants with birthweights near the tails of the observed distribution and would affect the upper and lower percentiles. To address this issue, Platt and colleagues considered only 40 weeks as the secondary distribution; in their framework, births with suspiciously high birthweights for a reported gestational age would have had to be more consistent with the birthweight distribution at 40 weeks to be flagged as potentially in error.[39] With MI, we assume that the multiple applications of the imputation algorithm will lead to a set of imputed and/or observed gestational age values for each birth record that reflect both the recorded birthweight and gestational age on the birth certificate as well as the likelihood of the combination being clinically plausible.

The proposed methods are targeted towards identifying and imputing gestational age values that are considered too low relative to the corresponding birthweight – that is, actual near-term or term births with reported preterm gestational ages. Platt *et al*.[39] also use their approach to identify possible term infants with reported post-term gestational ages. Extending this methodology to better separate the actual preterm births from the small-for-gestational-age births among those with reported term gestational age is more challenging given the preponderance of term births and their variation in birthweight, but it could be explored.

Public-use datasets have many uses. Obtaining valid and informative inferences is an important use, although maintaining the integrity of the individual birth records could be considered another. MI, by design, will not replace the missing or implausible gestational age with the most likely value for an individual birth record; rather, it will provide a set of likely alternative values that can be used to improve inferences for the overall birth cohort as well as for subgroups of interest. Counts of births for small geographical areas or for small demographic subgroups may not be as robust as counts for larger areas or groups. However, the original gestational age value would be included in the public-use database along with the multiple imputations, allowing users to make their own decisions. Furthermore, MI, in contrast to single imputation or simple reassignment with a clinical estimate, may have a perceptional advantage; assigning two or more values to a birth record lessens the impression that the gestational age was actually 'changed'.

While the ultimate utility of MI for a public-use US Natality dataset will depend on a variety of factors, it is hoped that the process can provide some insights into the birthweight–gestational age relationship, as well as a better understanding of the mechanisms underlying the reporting of gestational age.

References

1 National Center for Health Statistics. Technical Appendix. *Vital statistics of the United States, 2003, Vol. 1, Natality*. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Available at: http://www.cdc.gov/nchs/data/TechApp03_1-09.pdf

2 National Center for Health Statistics.*Documentation of the Detailed Natality Public Use File for 2003.* Available at: http://www.cdc.gov/nchs/births.htm

3 Alexander GR, Himes JH, Kaufman RB, Mor J, Kogan M. A United States national reference for fetal growth. *Obstetrics and Gynecology* 1996; **87**:163–168.

4 Zhang J, Bowes WA Jr. Birth-weight-for-gestational-age patterns by race, sex, and parity in the United States population. *Obstetrics and Gynecology* 1995; **86**:200–208.

5 Overpeck MD, Hediger ML, Zhang J, Trumble AC, Klebanoff MA. Birth weight for gestational age of Mexican American infants born in the United States. *Obstetrics and Gynecology* 1999; **93**:943–947.

6 Kiely JL. What is the population-based risk of preterm birth among twins and other multiples? *Clinical Obstetrics and Gynecology* 1998; **41**:3–11.

7 Joseph KS, Kramer MS, Allen AC, Mery LS, Platt RW, Wen SW. Implausible birth weight for gestational age. *American Journal of Epidemiology* 2001; **153**:110–113.

8 Parker JD, Schoendorf KC. Implications of cleaning gestational age data. *Paediatric and Perinatal Epidemiology* 2002; **16**:181–187.

9 Delgado-Rodriguez M, Llorca J. Bias. *Journal of Epidemiology and Community Health* 2004; **58**:635–641.

10 Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research*. Belmont, CA: Lifetime Learning Publications, 1982.

11 Little RJA, Rubin DB. *Statistical Analysis with Missing Data*, 2nd edn. New York: John Wiley & Sons, 2002.

12 Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**:473–489.

13 Taffel S, Johnson D, Heuse R. A method of imputing length of gestation on birth certificates. *Vital and Health Statistics. Series 2* 1982; **93**:1–11.

14 Little RJA. Missing data adjustments in large surveys. *Journal of Business and Economic Statistics* 1988; **6**:287–301.

15 Rubin DB. *Multiple Imputation of Nonresponse in Surveys*. New York: John Wiley & Sons, 1987.

16 Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine* 1991; **10**:585–598.

17 StataCorp. *Stata Statistical Software: Release 9.0*. College Station, TX: Stata Corporation, 2005.

18 Shah BV, Barnwell BG, Gieler GS. *SUDAAN User's Manual: Release 7.5*. Research Triangle Park, NC: Research Triangle Institute, 1997.

19 SAS Institute Inc. *SAS/STAT 9.1 User's Guide*. Cary, NC: SAS Institute Inc., 2004.

20 Box GEP, Tiao GC. *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley, 1973.

21 Schenker N, Raghunathan TE, Chiu PL, Makuc DM, Zhang G, Cohen AJ. Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association* 2006; **101**:924–933.

22 Schenker N, Raghunathan TE, Chiu PL, Makuc DM, Zhang G, Cohen AJ.*Multiple Imputation of Family Income and Personal Earnings in the National Health Interview Survey: Methods and Examples*. Available at: http://www.cdc.gov/nchs/data/nhis/tecdoc1.pdf

23 National Center for Health Statistics.*NHANES III Multiply Imputed Data Set User's Guide.* July 2001. Available at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHANES/NHANESIII/7A/doc/nh3mi.pdf

24 Schafer JL, Ezzati-Rice TM, Johnson W, Khare M, Little RJA, Rubin DB. The NHANES III Multiple Imputation Project. In: *Proceedings of the Survey Research Methods Section of the American Statistical Association* 1996; pp. 28–37.
25 Kennickell AB. Imputation of the 1989 Survey of Consumer Finances: stochastic relaxation and multiple imputation. In: *Proceedings of the Survey Research Methods Section of the American Statistical Association* 1991; pp. 1–10.
26 Board of Governors of the Federal Reserve System. *Codebook for 2004 Survey of Consumer Finances*. Available at: http://www.federalreserve.gov/PUBS/oss/oss2/2004/codebk2004.txt
27 Clogg C, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple imputation of industry and occupation codes in census public use samples using Bayesian logistic regression. *Journal of the American Statistical Association* 1991; **86**:68–78.
28 Subramanian R.*Transitioning to Multiple Imputation – A New Method to Estimate Missing Blood Alcohol Concentration (BAC) Values in FARS*. Technical report DOT HS 809 403, January, 2002. Available at: http://www.madd.org/docs/nhtsa_impute_method.pdf
29 Alexander GR, Tompkins ME, Petersen DJ, Hulsey TC, Mor J. Discordance between LMP-based and clinically estimated gestational age: implications for research, programs, and policy. *Public Health Reports* 1995; **110**:395–402.
30 Gruenwald P. Growth of the human fetus. Normal growth and its variation. *American Journal of Obstetrics and Gynecology* 1966; **94**:1112–1119.
31 Williams RL, Creasy RK, Cunningham GC, Hawes WE, Norris FD, Tashiro M. Fetal growth and perinatal viability in California. *Obstetrics and Gynecology* 1982; **59**: 624–632.
32 Kramer MS, Platt RW, Wen SW, Joseph KS, Allen A, Abrahamowicz M, *et al.* A new and improved population-based Canadian reference for birthweight for gestational age. *Pediatrics* 2001; **108**:E35.
33 Tentoni S, Astolfi P, De Pasquale A, Zonta LA. Birthweight by gestational age in preterm babies according to a Gaussian mixture model. *BJOG* 2004; **111**:31–37.
34 Vahratian A, Buekens P, Bennett TA, Meyer RE, Kogan MD, Yu SM. Preterm delivery rates in North Carolina: are they really declining among non-Hispanic African Americans? *American Journal of Epidemiology* 2004; **159**:59–63.
35 Oja H, Koiranen M, Rantakallio P. Fitting mixture models to birth weight data: a case study. *Biometrics* 1991; **47**:883–897.
36 Everitt BS. An introduction to finite mixture distributions. *Statistical Methods in Medical Research* 1996; **5**:107–127.
37 McLachlan G, Peel D. *Finite Mixture Models*. New York: John Wiley & Sons, 2000.
38 McLachlan GJ. Mixture modeling for cluster analysis. *Statistical Methods in Medical Research* 2004; **13**:347–361.
39 Platt RW, Abrahamowicz M, Kramer MS, Joseph KS, Mery L, Blondel B, *et al.* Detecting and eliminating erroneous gestational ages: a normal mixture model. *Statistics in Medicine* 2001; **20**:3491–3503.
40 Ghosh-Dastidar B, Schafer JL. Multiple edit/multiple imputation for multivariate continuous data. *Journal of the American Statistical Association* 2003; **98**:807–817.
41 Little RJA, Smith PJ. Editing and imputation for quantitative survey data. *Journal of the American Statistical Association* 1987; **82**:58–68.
42 Hediger ML, Overpeck MD, McGlynn A, Kuczmarski RJ, Maurer KR, Davis WW. Growth and fatness at three to six years of age of children born small- or large-for-gestational age. *Pediatrics* 1999; **104**:E33.
43 Kramer MS, McLean FH, Boyd ME, Usher RH. The validity of gestational age estimation by menstrual dating in term, preterm, and postterm gestation. *JAMA* 1988; **26**:3306–3309.
44 Yang H, Kramer MS, Platt RW, Blondel B, Bréart G, Morin I, *et al.* How does early ultrasound scan estimation of gestational age lead to higher rates of preterm birth? *American Journal of Obstetrics and Gynecology* 2002; **186**:433–437.
45 Henriksen TB, Wilcox AJ, Hedegaard M, Secher NJ. Bias in studies of preterm and postterm delivery due to ultrasound assessment of gestational age. *Epidemiology* 1995; **6**:533–537.
46 Savitz DA, Terry JW Jr, Dole N, Thorp JM Jr, Siega-Riz AM, Herring AH. Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination. *American Journal of Obstetrics and Gynecology* 2002; **187**:1660–1666.
47 Hediger ML, Scholl TO, Schall JI, Miller LW, Fischer RL. Fetal growth and the etiology of preterm delivery. *Obstetrics and Gynecology* 1995; **85**:175–182.
48 Gjessing HK, Skjaerven R, Wilcox A. Errors in gestational age: evidence of bleeding early in pregnancy. *American Journal of Public Health* 1999; **89**:213–218.

