Published in final edited form as:
Demography. 2015 February ; 52(1): 329–354. doi:10.1007/s13524-014-0358-x.

Can We Spin Straw Into Gold? An Evaluation of Immigrant Legal Status Imputation Approaches

Jennifer Van Hook,
The Population Research Institute, The Pennsylvania State University, 601B Oswald Tower, University Park, PA 16802, USA

James D. Bachmeier,
Department of Sociology, Temple University

Donna Coffman, and
The Methodology Center, The Pennsylvania State University

Ofer Harel
Department of Statistics, University of Connecticut

Jennifer Van Hook: jxv21@psu.edu

Abstract

Researchers have developed logical, demographic, and statistical strategies for imputing immigrants' legal status, but these methods have never been empirically assessed. We used Monte Carlo simulations to test whether, and under what conditions, legal status imputation approaches yield unbiased estimates of the association of unauthorized status with health insurance coverage. We tested five methods under a range of missing data scenarios. Logical and demographic imputation methods yielded biased estimates across all missing data scenarios. Statistical imputation approaches yielded unbiased estimates only when unauthorized status was jointly observed with insurance coverage; when this condition was not met, these methods overestimated insurance coverage for unauthorized relative to legal immigrants. We next showed how bias can be reduced by incorporating prior information about unauthorized immigrants. Finally, we demonstrated the utility of the best-performing statistical method for increasing power. We used it to produce state/regional estimates of insurance coverage among unauthorized immigrants in the Current Population Survey, a data source that contains no direct measures of immigrants' legal status. We conclude that commonly employed legal status imputation approaches are likely to produce biased estimates, but data and statistical methods exist that could substantially reduce these biases.

Keywords

Legal status; Unauthorized; Immigration; Imputation; Simulation

"Interest in immigrants' socioeconomic characteristics from scientists, policy-makers, and the public has run ahead of the availability of the data to address these

> interests. The most serious omission from data sets is information on legal status…."
>
> —Clark and King (2008:295)

The capacity for immigration scholars to produce research results of major social and policy significance remains hampered by a lack of population and survey data allowing the identification of immigrants' legal status. Large-scale, nationally representative surveys that are most commonly used to study the foreign-born population—such as the American Community Survey (ACS)—distinguish between naturalized citizens and noncitizens, but they do not inquire about the legal status of the latter. Surveys that have included such measures are limited by the fact they are typically relatively small, regionally targeted, and/or focused on a particular subpopulation of immigrants (Bachmeier et al. 2014). As a result, important questions—such as the extent to which unauthorized status threatens the well-being of immigrant families, the role undocumented immigrants play in the labor market, and their economic and fiscal impacts—remain largely unaddressed (Clark and King 2008; Clark et al. 2009; Massey and Bartley 2005).

Faced with these data limitations, researchers have developed logical, demographic, and statistically based strategies for imputing the legal status of immigrants in the aforementioned nationally representative surveys (e.g., Batalova et al. 2014; Heer and Passel 1987; Marcelli 2004; Marcelli and Heer 1997, 1998; Passel and Cohn 2009; State Health Access Data Assistance Center 2013). At present, such methods are the only means through which much-needed avenues of research on legal status can be opened. However, the conditions under which these methods yield unbiased estimates of the characteristics of the unauthorized foreign-born population have never been tested.

We address this question using Monte Carlo simulations based on the Survey of Income and Program Participation (SIPP), an underutilized source of information on the foreign-born population and the only nationally representative survey with questions about immigrants' legal status. We provide empirical tests of whether, and under what conditions, it is possible to impute legal status on the basis of commonly available socioeconomic and demographic survey items: to spin straw into gold, so to speak. We first review existing legal status imputation methods, including a recently developed method that employs multiple imputation using pooled survey samples. We subsequently present simulation results that compare the various approaches with respect to the degree that they yield unbiased estimates of the association of unauthorized status with insurance coverage, an important predictor of access to health care, and thus a potential source of cumulative disadvantage (e.g., Bustamante et al. 2012; Javier et al. 2010; Ku 2009; Sommers 2013; Stevens et al. 2010). Although insurance coverage serves primarily as an example used to test the various imputation methods, our analyses do provide new estimates of the level and geographic distribution of coverage among unauthorized immigrants.

The results show that it is not possible to spin straw into gold. All the approaches that we tested produced biased estimates. Some methods failed in all circumstances, and others failed only when the "joint observation" condition was not met, meaning that the imputation method was not informed by the association of unauthorized status with the dependent

variable. Nevertheless, we also show that these methods could be improved if external ("prior") information about legal status were available. Additionally, in an example using the Current Population Survey (CPS), we demonstrate the utility of the best-performing method for increasing statistical power when the joint observation condition is met.

Background: Methods for Measuring Immigrants' Legal Status

Several panels of the SIPP (1996, 2001, 2004, and 2008) asked foreign-born respondents their immigration status when they entered the United States and whether they had since adjusted their status (U.S. Census Bureau 2013). Assessments of the quality of the legal status data from the SIPP further reveal that they are likely to produce an accurate portrayal of the unauthorized population. Despite moderately high levels of missing data, the demographic characteristics of the unauthorized immigrants in the SIPP closely match residual estimates produced by the Department of Homeland Security (DHS) and the Pew Hispanic Center (Pew) (Bachmeier et al. 2014).

However, the SIPP is not appropriate for all research questions involving unauthorized migration. Its sample is too small for some types of analyses (e.g., state-level). Additionally, although the SIPP includes detailed information about income, poverty, and public assistance receipt, it provides much less information about health, education, and fertility, all of which are highly significant topics. Other surveys collect data on legal status, but none are nationally representative. They are limited to a specific metropolitan area (e.g., Los Angeles Family and Neighborhood Study, Los Angeles County Mexican Immigrant Residency Status Survey, Los Angeles County Household Survey), state (e.g., California Health Interview Survey), occupational group (e.g., National Agricultural Workers Survey), immigrants who legalized (e.g., Legalized Population Survey, New Immigrant Survey), or immigrants who returned to Mexico (e.g., Mexican Migration Project).

To compensate for the scarcity of survey data on legal status, several researchers have developed creative ways to impute legal status in surveys lacking such measures. In the most basic terms, imputation methods involve assigning legal status to immigrants in a survey sample lacking measures of legal status on the basis of information provided by outside knowledge about the characteristics of legal immigrants or independent data sources, such as a survey with direct measures of legal status. Most of these methods do not account for uncertainty but rather treat imputed values as if they were true.

The most widely publicized imputation-based results are those produced by Passel and published by the Pew Hispanic Center. Originally developed by Passel and Clark and published by the Urban Institute (1998), this method imputes legal status for foreign-born respondents of the CPS in order to produce detailed descriptive profiles of the unauthorized population, such as poverty rates, unemployment rates, educational attainment, and occupational composition (e.g., Passel 2006; Passel and Cohn 2009).

The method on which Pew Hispanic Center estimates are based combines a variety of techniques—logical imputation, statistical imputation, and weighting adjustments—to assign legal status. Because it attempts to match legal status assignments with external information about immigration policy and residual (demographic) estimates of the unauthorized

population, we classify it as a "demographic accounting method." This method first identifies those who are very likely to be legally resident on the basis of indicators of legality, such as U.S. citizens, veterans, and those in occupations that make it nearly impossible for them to be unauthorized.[1] The remaining noncitizens are the pool of potentially unauthorized immigrants: a group that contains a mixture of legal noncitizens and unauthorized immigrants. To further distinguish the unauthorized from legal noncitizens, the method assigns legal status to the remaining noncitizens based on an estimated probability of being unauthorized, which is calculated from the occupational distribution by age, sex, and state of residence of unauthorized immigrants in the Legalized Population Survey (LPS). The LPS is a 1989 survey of those who applied for legalization under the main provisions of the 1986 Immigration Reform and Control Act (IRCA). After additional data editing to ensure that the status assignments correspond with U.S. immigration law for families, the sampling weights of respondents are adjusted to match control totals, derived from residual estimates (e.g., Passel 2006) of the number of unauthorized immigrants by state and national origin.

Demographic accounting estimates, particularly those produced by Pew/Passel, are based on the meticulous application of demographic methods and have come to be trusted and widely cited outside of academia. However, the method has never been evaluated. The specific details of the Pew/Passel method are not publicly available, thus making it difficult for other researchers to replicate the method. Beyond the lack of transparency, its reliance on LPS data raises concern largely because the LPS was collected more than two decades ago, and it represents only the unauthorized who applied for legalization under the nonagricultural worker provisions of IRCA. The LPS thus excludes those who qualified for the other major legalization program under IRCA (the Special Agricultural Workers program), and it overrepresents Mexicans and those who arrived in the United States before 1982, and who were then concentrated in the American Southwest to a far greater degree than is true today (Durand et al. 2005).

Others have used survey-based statistical imputation to assign unauthorized status (Caponi and Plesca 2014; Capps et al. 2013; Heer and Passel 1987; Marcelli and Heer 1997, 1998). Statistical imputations use the associations between a set of predictors and unauthorized status from a survey that includes questions about immigrants' legal status (the donor sample) to assign legal status to foreign-born respondents in surveys lacking such measures (the target sample). Statistical imputations have employed both single and multiple imputation approaches as the basis for prediction. Heer and Passel (1987) were among the first to use a single-imputation approach. In subsequent developments, Marcelli and Heer (1997) estimated a logistic regression model predicting unauthorized status as a function of duration of U.S. residence, educational attainment, age, and sex in the 1994 Los Angeles County Household Survey (the donor sample). They then used this model to estimate the predicted probability of being unauthorized among immigrants in the Los Angeles County

[1]Indicators of legality include U.S. citizenship, migration from countries and periods that correspond with known patterns of refugee flows, being newly arrived with characteristics that would qualify for certain visa categories, working in occupations or industries that require legal status; receipt of public assistance or social services, and having moved to the United States before 1982 (thus qualifying for IRCA legalization).

1990 Census (the target sample). Finally, they used the predicted probabilities in subsequent analyses that estimated the relationship between unauthorized status and labor force and welfare outcomes (Marcelli and Heer 1997, 1998).

In contrast to the single-imputation method, researchers have begun to employ cross-survey multiple imputation to impute variables that are completely missing in one data set but observed in another (Rässler 2004; Rendall et al. 2013; Resche-Rigon et al. 2013; Schenker et al. 2010). Multiple-imputation approaches are preferred because they account for the uncertainty in imputed data (Little and Rubin 2002). This approach has recently been used to impute legal status. To produce a profile of health insurance coverage and other social and economic characteristics among the unauthorized, Capps et al. (2013) pooled the SIPP (the donor sample) with the ACS (the target sample), and multiply imputed unauthorized status for all foreign-born observations in the ACS. A Minnesota policy analysis group (State Health Access Data Assistance Center 2013) used a similar approach in examining immigrants' access to health insurance coverage.

As long as a suitable donor sample is available (such as the SIPP), statistical imputation methods can be readily replicated, unlike the detailed algorithms involved in logical and demographic accounting methods. However, the bias and precision of the resulting estimates remain unclear. Importantly, Rendall and his colleagues (2013) argued that the success of the cross-survey multiple-imputation method depends on two conditions. First, the target and donor samples must be drawn from the same universe. In other words, the population-level associations producing the donor sample should be statistically identical in the target sample. Second, to avoid identification problems, every pair of variables must be jointly observed in one data set or the other to enable the estimation of the covariance for all pairs. Taking an example developed by Rässler (2004), say we have two data sets. Variables X and Z are observed in the first, Z and Y are observed in the second, and X and Y are never jointly observed. If the correlation between Z and X is .9, and that between Z and Y is .8, then the estimated correlation between X and Y is mathematically bounded between 0.4585 and 0.9815—a wide range. Without additional information, no value in this range is a better estimate than another. Additionally, Rodgers (1984) showed that only very high correlations approaching 1.0 will narrow the range considerably.

When applied to legal status imputations, the joint observation requirement effectively limits the analytic variables to those that are jointly observed with legal status. For example, if an analyst were interested in estimating the effects of unauthorized status on insurance coverage, and if legal status were completely missing in the target sample, then insurance coverage *must be* observed in both the donor and target samples. If insurance coverage were completely missing in the donor data, then legal status and insurance coverage would never be jointly observed. If both the same universe and joint observation conditions must always be met, this would cast doubt on methods that violate them, including most imputation methods employed in past research.

Here, we evaluate the prevailing approaches to imputing legal status. We do not attempt to replicate and evaluate specific imputation methods, such as the precise methodology from which Pew Hispanic Center estimates are derived, largely because such methods change

over time as researchers refine their methodologies and data inputs—and, as noted, they can be difficult to replicate. Rather, we evaluate and compare five general approaches (explained in the Imputation Methods section). We tested multiple variations of each of these methods in preliminary analyses, but due to space constraints, we present the results for only the best-performing variants.

We conducted Monte Carlo simulations that evaluate whether, and under what conditions, estimates of the association between imputed unauthorized status and insurance coverage are unbiased. By varying the imputation method, the simulations identify the optimal method. We alter the missing data patterns in the simulation data to assess the performance of the methods when the joint observation condition is not met. We further assessed how much the methods would improve if prior information about immigrants' legal status were available beyond that already included in most demographic surveys, whether through administrative record linkages, new survey questions, or information from an auxiliary survey.

Throughout, we assessed the robustness of the results across different dependent variables by varying, in simulated data, the magnitude of the association between unauthorized status and health insurance coverage. Imputation methods may perform well when the association between unauthorized status and the dependent variable is consistent with socioeconomic and demographic characteristics (e.g., the unauthorized have lower levels of insurance coverage than legal immigrants, which is consistent with their lower levels of education and income). However, imputation methods may be less able to detect "surprises," such as when unauthorized immigrants exhibit unique or exceptional outcomes.

Methodology

Data

We used the SIPP as a basis for generating data and establishing true population values for the simulations. The SIPP is a longitudinal survey of the U.S. noninstitutionalized population conducted by the U.S. Census Bureau (2013). Every few years, the SIPP draws a new panel of households (i.e., 1996, 2001, 2004, and 2008). All individuals in these households are then followed up every four months for three to four years. Panel respondents in each wave are asked a set of core questions primarily about labor force activity, income, and program participation. In addition, respondents are administered wave-specific topical modules. In all panels from 1996–2008, including the 2004 panel on which we rely for our simulations, the second wave of data collection includes a series of questions about migration, which includes questions about country of birth, year of arrival, citizenship, and visa status. Although SIPP is longitudinal, each wave can be weighted with cross-sectional weights to represent the current U.S. population. The final SIPP sample from which our simulated data were generated was restricted to 8,898 foreign-born respondents age 16 and older who were interviewed in Wave 2 of the 2004 panel. Children as well as persons born abroad to U.S. citizen parents were excluded.

Measures

The weighted means for all analytic variables from the SIPP are shown in Table 1 for the total foreign-born sample in the SIPP as well as separately by three legal status designations —"probably legal," "ambiguous," and "unauthorized"—the definitions for which are provided in the Imputation Methods section. Further description of the two simulated insurance coverage variables—insurance 2 and 3—is also provided later herein.

Unauthorized Legal Status—The key independent variable used in our analyses is a dichotomous indicator of unauthorized legal status (= 1 if unauthorized). The SIPP asked questions about immigrants' legal status at the second interview. Foreign-born respondents were asked whether they were citizens, and all noncitizens were asked about their status upon arrival. Immigrants could select one of six arrival statuses: three types of legal permanent resident (LPR) status; and three non–LPR statuses, including refugee/asylee status, legal nonimmigrants (e.g., student or tourist visas), and "other." Finally, noncitizen, non-LPR arrivals were asked whether they have adjusted to LPR status since first immigrating. Following others (Greenman and Hall 2013; Hall et al. 2010), we infer that the group of persons arriving with "other" status and who have not adjusted to LPR status overwhelmingly consists of unauthorized immigrants. To address data handling challenges presented both by relatively high rates of missing data on immigration related items in the SIPP and by the suppression by the Census Bureau of detailed visa status categories in the arrival status item in the public-use data, we have employed similar methods reported in previous research (Bachmeier et al. 2014; Greenman and Hall 2013; Hall et al. 2010).

Insurance Coverage—The dependent variable, "insurance 1," is a dichotomous indicator of insurance coverage (employer, other private, Medicaid, and other public = 1).

Controls—All models of insurance coverage included the following controls: income-to-poverty ratio (logged), educational attainment (years), Mexican place of birth, years of U.S. residence, age, sex, number of functional limitations, and self-rated health (fair/poor vs. better health).

Simulations

We conducted Monte Carlo simulations to evaluate the bias of the estimate of the association of imputed unauthorized status with the outcome (health insurance coverage) under different scenarios. For the simulation exercises, we assumed that the true association (i.e., the expected population value) of unauthorized status with insurance coverage is the association observed in the SIPP. Bias is the difference between the true and estimated association. The validity of the SIPP-based measure of legal status (or how close the SIPP-based measure comes to reality) is also an important question, but we set it aside here because it is not central to our question concerning bias and because we have already addressed it in another article (Bachmeier et al. 2014).

Each simulation involved three steps. First, we drew 10,000 cases with replacement from a self-weighted version of the 2004 SIPP (i.e., expanded in proportion to sampling weights). Second, we randomly divided the sample in half, with the first half representing the donor

sample, and the second half representing the target sample. This randomization ensured that the donor and target samples are drawn from the same universe. We coded unauthorized status to missing in the target sample, and for some simulations, we coded the dependent variable and/or a key independent variable (the income-to-poverty ratio) to missing in the donor sample. Both samples included a common set of control variables. The data structure of the donor and target samples is illustrated in Table 2. Third, we imputed unauthorized status in the target sample using one of five imputation methods, and then estimated a multivariate model predicting insurance coverage as a function of imputed unauthorized status and controls on cases in the target sample.

For each simulation, we repeated all three steps 500 times to estimate the coefficient and standard error for imputed unauthorized status (i.e., the average of the coefficient and standard error across the 500 replications). We estimated bias as the average difference between the estimated coefficient and the expected population value; we estimated relative bias as bias divided by the expected value. The level of acceptable bias varies by application. As a rule of thumb, we favor methods that produce estimates with lower bias and relative bias, and we describe estimates as "unbiased" if they fall within 10 % of the expected value (i.e., relative bias falls in the range of −.10 to .10).

Imputation Methods—As summarized in Table 3, we varied the imputation methodology to evaluate the comparative performance of a set of approaches under a range of conditions.

The logical-imputation method is similar to the first part of the Pew/Passel method. We tested it separately from the demographic accounting method because it is sometimes used to proxy legal status in policy analyses (e.g., Bohn et al. 2014; Bozick and Miller 2014; Flores 2010; Kaushal 2006; Potochnick 2014). It codes as legal those in the target data who have characteristics that make it very unlikely they are unauthorized (i.e., those who are "probably legal"), and all others as unauthorized. In our simulations, the "probably legal" included U.S. citizens and others with indicators of legality, such as employment by the U.S. government, a history of military service, or receipt of Social Security income; we used similar indicators of legality as the Passel/Pew method (see footnote 1 for a complete list). These indicators are measured in most major demographic and health surveys (e.g., ACS, CPS, NHIS). As already shown in Table 1, roughly one-half (51 %) of all foreign-born in the SIPP can be logically imputed as legal by these criteria. Among the remaining unclassified foreign-born respondents, one-half are actually unauthorized (24 % of all foreign-born), and the rest are neither unauthorized nor have characteristics that signify legality and therefore have "ambiguous" status.

The demographic accounting method has similarities to the full Pew/Passel method in that it combines elements of the logical- and single-imputation methods, and its estimates are forced to match target values of the percentage of unauthorized immigrants. Those classified as "probably legal" in the target sample were coded as legal (51 %). Among the remaining foreign-born, the single-imputation method was employed to assign unauthorized status. To do this, we estimated a logistic regression model predicting unauthorized status as a function of several regressors[2] in the donor data source. Importantly, the dependent variable (insurance coverage) was included in the prediction model in simulations in which the

dependent variable was jointly observed with unauthorized status. The estimated coefficients were then applied to immigrants in the target data to derive for each person a predicted probability of being unauthorized. Each individual's predicted probability was compared with a random draw from a uniform distribution: if the predicted probability was greater than the random draw, the individual was assigned unauthorized status. This assignment process continued until a targeted percentage (24 %, or about one-half of the non–probably legal, as observed in the SIPP) was coded as unauthorized.[3]

The single-imputation method is similar to the approach taken by Heer and Marcelli (Marcelli 2004; Marcelli and Heer 1997, 1998), and more recently by Caponi and Plesca (2014). A logistic regression model predicting unauthorized status was estimated on the donor data, using the same predictors as for the demographic accounting method. In the target data, foreign-born persons were coded as unauthorized if a random draw from a uniform distribution was less then their predicted probability of being unauthorized, and all others were coded as legal. Unlike the demographic accounting method, no attempt was made to first code people as "probably legal," and the percentage assigned as unauthorized was derived from the percentage unauthorized in the donor data, not from a predetermined target.

The cross-survey multiple-imputation (CSMI) method is similar to the approach taken by Capps et al. (2013). It pools the donor and target samples and treats the absence of an unauthorized status indicator in the latter as a missing data problem to be addressed by multiple-imputation techniques. Specifically, missing values in the target data were imputed using multiple chained equations (StataCorp 2013). Following common practice (Rubin 1987), 10 data sets were created, and results were summarized using the *mi* routines available in Stata version 12 or higher. We used the same predictors as in the single-imputation method to inform the imputations.

Finally, the logical cross-survey multiple-imputation (logical-CSMI) method combines elements of three methods. We tested this approach because we wondered whether the multiple-imputation method could be improved if it were informed by outside (i.e., logical) information about legal status. As with the CSMI method, the donor and target samples were pooled, and missing data were imputed with multiple chained equation techniques. In addition to the predictors used by the other methods, the imputation model was informed by the respondents' classification as "probably legal" and predicted probability of being unauthorized (based on a single prediction model). Specifically, the predicted probability was coded to 0 for those who are "probably legal" and was then included as a predictor in

[2]Regressors include insurance coverage, all the controls described earlier, and several additional variables: marital status, spouse's citizenship, occupational status, English proficiency, parental status, household size, homeownership, employment status, occupation, state of residence, and selected squared and interaction terms.

[3]We tested variations where we altered which half of the non–probably legal were coded as unauthorized: those most likely to be unauthorized, the most disadvantaged, and a random half. None performed better than the demographic accounting method described here.

the imputation model.[4] Like the single-imputation and CSMI methods, the percentage assigned as unauthorized was derived from the donor data, not a predetermined target.

Joint Observation of the Dependent and Independent Variables With Unauthorized Status

We assessed whether bias depends on whether (1) the dependent variable, (2) an important independent variable (income-to-poverty ratio), and (3) both the dependent and independent variables are jointly observed with unauthorized status. In simulations in which these variables are treated as jointly observed, they are observed along with unauthorized status in the donor data set, as shown in Table 2. In simulations in which they are never jointly observed, we recoded them to missing in the donor sample prior to carrying out the imputation.

Strength of Association Between Unauthorized Status and Insurance Coverage

Robust imputation methods should produce unbiased estimates of the association of unauthorized status regardless of how surprising the result is. To assess the robustness of the methods, we created two variants of insurance coverage, one with weak ("insurance 2") and another with strong ("insurance 3") associations with unauthorized status; we used these along with the original measure of insurance coverage as dependent variables in our models. We created the variants by recoding insurance coverage among those with ambiguous legal status (i.e., legal noncitizens in the donor sample that do not have characteristics of the "probably legal" population). As observed in the SIPP, insurance coverage is lowest for unauthorized immigrants, higher among those with "ambiguous" status, and highest among those who are "probably legal." To weaken the association, we randomly reduced coverage among "ambiguous" immigrants, thus reducing the insurance gap between the unauthorized and all legal immigrants (see Table 1). To strengthen the association, we randomly increased coverage among "ambiguous" immigrants, thus increasing the gap.

Results

As noted earlier, the associations observed in the SIPP represent the expected population values. The first row of Table 4 reports ordinary least squares (OLS) coefficients from the model estimated on the SIPP data. Because insurance coverage is treated continuously, the interpretation of coefficients is that of a linear probability model. In the case of "insurance 1" (observed), the coefficient for unauthorized status is −0.152, indicating that insurance coverage is about 15 percentage points lower among the unauthorized than the authorized after accounting for the control variables. As designed, unauthorized status is more weakly associated with "insurance 2" ($\beta = -0.058$) and more strongly associated with "insurance 3" ($\beta = -0.236$).

The simulations presented in Table 4 assess the imputation methods under optimal circumstances: that is, when all variables are jointly observed with unauthorized status. For

[4]We also tried (1) coding unauthorized status to 0 for those who are "probably legal" before multiply imputing, (2) including "probably legal" and the predicted probability separately in the imputation model, and (3) coding the "probably legal" as legal and those with very high probabilities of being unauthorized (>.8) as unauthorized prior to multiple imputation. None of these variations outperformed the logical-CSMI method.

this scenario, both the logical and demographic accounting methods produce biased estimates for two of the three dependent variables. The logical-imputation method completely fails to pick up variations in the association between unauthorized status and the dependent variables, estimating the strongest association where the expected association is weakest ("insurance 2"; relative bias = 3.259), and the weakest association where the expected association is strongest ("insurance 3"; relative bias = −0.743). The demographic accounting method does not perform much better. For example, the bias in the model predicting "insurance 2" is −0.10 (relative bias = 1.757).

In contrast, the single-imputation, CSMI, and logical-CSMI methods yield virtually unbiased estimates for all three dependent variables, with bias never exceeding ±5.5 % of the expected value. The two methods using multiple imputation (CSMI and logical-CSMI) yield larger standard errors than the single-imputation method because single imputation treats imputed values as true and therefore underestimates the true variance (Little and Rubin 2002).

We next evaluate the effects of violating the joint observation assumption. As shown in Table 5, the logical and demographic accounting methods are less sensitive to the missing data pattern than the other methods because they do not rely heavily (and the logical imputation does not at all rely) on the missing variables to impute legal status. Nevertheless, both methods produce coefficients with large biases across all missing data scenarios for at least two of the three dependent variables.

In contrast, the single-imputation, CSMI, and logical-CSMI methods are sensitive to the missing data pattern. For these methods, the estimates are virtually unbiased when the dependent variable is jointly observed with unauthorized status (either "DV and IV jointly observed" or "DV jointly observed, IV never jointly observed"). Even when a key independent variable (income-to-poverty ratio) is missing from the donor data, bias remains somewhat low and less than ±20 % of the expected value. Supplementary analyses suggest that this holds only when the imputation model is "inclusive" (i.e., containing "everything but the kitchen sink"). When the imputation model is parsimonious and includes only the variables used in the model of insurance coverage, the estimates are more biased (results available upon request), which is consistent with Collins et al. (2001). Finally, when the dependent variable is not jointly observed with unauthorized status in the donor data (either "DV never jointly observed, IV jointly observed" or "DV and IV never jointly observed"), bias is much greater than for the other missing data scenarios—in one case, exceeding 80 % of the expected value. For "insurance 1" and "insurance 3" (for which the expected coefficient is large and negative), the coefficients overestimate health coverage for unauthorized immigrants relative to legal immigrants.

Using Prior Information to Improve Imputations

The results thus far show that when the joint observation requirement is met, the statistical imputation approaches, particularly those employing multiple imputation, yield unbiased estimates, at least in the particular scenarios we tested. However, when the joint observation requirement is not met, none of the methods produce unbiased estimates across all three outcomes, including the logical-CSMI method and all its variants (see footnote 4). This led

us to explore whether the incorporation of prior information into the imputation methods would improve the estimates even when the joint observation condition is not met.

We first considered the methods that rely on logical imputation: the logical, demographic accounting, and logical-CSMI methods. Our data permit us to logically impute about one-half (51 %) of the foreign-born as "probably legal," but we wondered how much bias would be reduced if a higher percentage were logically imputed (i.e., if additional information enabled us to identify more legal immigrants). Logically imputing a higher percentage would be possible because additional indicators of legality from administrative record linkages are available in restricted data sets (e.g., possession of a valid Social Security number), and new survey questions could be added to surveys. To explore this question, we reran the simulations while randomly increasing the percentage of those with ambiguous status that are classified as "probably legal" to 50 %, 75 %, and 90 % (meaning that the non–"probably legal" group was composed of 66 %, 80 %, and 90 % unauthorized, respectively). We confined our tests to the most problematic scenario in which the dependent variable is never jointly observed with unauthorized status. As shown in Table 6, bias decreases substantially across all methods and dependent variables as the percentage logically imputed increases. Nevertheless, even when as many as 90 % of those with ambiguous status are logically imputed as "probably legal," relative bias remains moderately high for "insurance 2," reaching 46.2 % of the expected value in the case of the demographic accounting imputation method, although the absolute bias is low (−0.028).

This approach offers a potential solution to analysts who are unable to meet the joint observation requirement. However, most demographic and health data do not typically include enough indicators of legality to classify such a high proportion as "probably legal." Additionally, federal statistical agencies appear to have become more, not less, restrictive in their willingness to release sensitive data and administrative record linkages, especially concerning immigrants and their statuses. We therefore explored yet another option available to researchers when the joint observation condition is not initially met. Following Rässler (2004), the CSMI method is likely to yield less-biased estimates if an auxiliary data set with jointly observed measures of unauthorized status and the dependent variable were pooled with the donor and target data sets. Typically, an auxiliary data set is one that contains the necessary variables but may not be drawn from the same universe as the target sample. To illustrate, one might pool the SIPP with a national health survey to estimate the association of legal status with chronic health conditions. Because legal status is measured only in the SIPP and chronic health conditions are measured only in the health survey, the two are never jointly observed. However, one could add an auxiliary data set that includes both unauthorized status and chronic health conditions, such as the California Health Interview Survey (2013), to satisfy the joint observation requirement.

To evaluate this approach, we pooled target and donor data (wherein unauthorized status and the dependent variable are never jointly observed) with a third equal-sized data set containing both variables. In practice, it would be difficult to locate an auxiliary data set that is drawn from the same universe as the other two data sets, so we assessed how the results differ when all three data sets are drawn from the same universe (the SIPP) versus when the auxiliary data set is drawn from a different universe (specifically, from Californians in the

SIPP). The results, shown in Table 7, indicate that bias is reduced to nearly 0 when all three data sets are drawn from the same universe. Bias is somewhat higher (but still quite low) when the same-universe assumption is violated. It is likely that the level of bias depends on a variety of factors, such as the number of variables common to the three data sets, the relative sample sizes of three data sets, and how different their universes are. Given space constraints, identifying and evaluating the impact of these factors extends beyond the scope of this article.

Using the Cross-Survey Multiple Imputation Method to Increase Statistical Power In this last section, we demonstrate how the CSMI method could be used with currently available public-use data to increase statistical power when the joint observation requirement is met. We selected CSMI over other methods because it yielded less-biased estimates than the logical and demographic accounting methods and is easier to implement than the equally well-performing logical-CSMI method. Although CSMI generally should not be used to examine outcomes that are unobserved in the donor data (i.e., without prior information), it can be used to increase sample size and power, which is extremely valuable for producing estimates by detailed characteristics, such as state of residence, country of birth, or year of entry cohort.

To demonstrate, we applied the CSMI method to actual data, using the 2004 SIPP data as the donor sample and the 2004 March CPS as the target sample. The CPS is conducted throughout the year by the U.S. Census Bureau on approximately 60,000 civilian U.S. housing units; thus, the sample is very large and has the capacity to produce estimates for many states, something that is not feasible with the smaller SIPP sample. The CPS lacks a measure of legal status, but includes many of the same variables as the SIPP (including all the predictors and the dependent variable in our example), which we coded identically as the SIPP variables. Thus, the joint observation condition can be met. Additionally, the CPS relies on nearly the same sampling frame as the SIPP. As with the analytical SIPP sample, our final CPS sample was restricted to foreign-born adults age 16 and older, excluding those born abroad of U.S.-born parents (N = 21,214). A comparative profile of the SIPP and CPS foreign-born samples on standard socioeconomic and demographic variables provides support for the same universe assumption. On nearly every dimension of comparison, the samples are virtually identically distributed (available upon request).

To implement the multiple-imputation method, we pooled the SIPP and CPS, multiply imputed unauthorized status for cases in the CPS on the basis of a large set of predictors (including insurance coverage), and estimated the percentage of unauthorized with insurance coverage by state/region in the CPS data. To demonstrate the sensitivity of the estimates to the specification of the imputation model, we first excluded the dependent variable—insurance coverage—from the predictors in the imputation model (i.e., treating it as if it were never jointly observed). In a second analysis, we included it. Finally, in the third, we estimated the imputation model separately by state/region (i.e., essentially a fully interactive model), thus allowing for interactions between state/region and all other predictors.

Results are reported in Table 8. In the first set of columns, means and standard errors are reported for states and regions in the SIPP in which there are at least 250 observations. For

example, as observed in the SIPP, approximately 42 % of unauthorized immigrants in California had health insurance coverage in 2004. The second column reports the corresponding percentage estimated in the CPS when unauthorized status has been imputed without health insurance coverage being included in the model. Despite greater precision signified by lower standard errors, these estimates are significantly different from the SIPP-based estimates in approximately one-half (8 of 14) of the state/regions and tend to overestimate coverage among unauthorized immigrants. For example, the estimate for California derived from the CPS when insurance was not included in the imputation model is fully 10 percentage points higher than that estimated by the SIPP.

Health insurance coverage is allowed to be jointly observed in the final two columns of Table 8. Estimates in the third set of columns include state/region as a variable in the imputation model, and estimates in the fourth set of columns are derived from a model estimated separately within the 14 states/regions. In these two columns, estimates of insurance coverage among the unauthorized are much more in line with the SIPP-based estimates compared with the scenario in which coverage is never jointly observed with unauthorized status. Just 1 of the 14 states/regions has estimates that are significantly different from the SIPP-based estimates both when state/region is included as a variable and when separate models are estimated within states/regions; and in both instances, the magnitude of the difference is smaller compared with the scenario in which coverage is never jointly observed.

Conclusions

Research on immigration, immigrant incorporation, and immigration policy has been stymied by inadequate data on legal status. To compensate for the lack of individual-level indicators of legal status, researchers have tried to impute legal status in order to examine the socioeconomic and demographic characteristics of unauthorized immigrants. Although imputation-based results appear to be widely accepted (especially in policy settings), the degree to which such methods produce biased estimates have not been empirically tested. In this article, we used Monte Carlo simulations to evaluate a variety of imputation approaches under a range of conditions.

Our simulations revealed significant limitations in all the methods we tested. The logical and demographic accounting approaches produced biased estimates even under the most favorable scenarios regarding missing data. They were especially poor at detecting surprising results, such as unusually low or high insurance coverage among unauthorized relative to legal immigrants. The statistical imputation methods (single-imputation, CSMI, and logical-CSMI) produced unbiased estimates when unauthorized status was jointly observed with the dependent variable. However, when this condition was not met, these methods tended to overestimate coverage among the unauthorized.

These biases arose when the imputation method failed to incorporate information about the association of unauthorized status with the dependent variable. To understand how problematic this is, consider the surprising but well-established finding that Hispanics have lower mortality than non-Hispanic whites despite their lower socioeconomic status (SES)

(Markides and Eschbach 2005). If we knew only that low SES is associated with higher mortality and that Hispanics have lower SES, we would erroneously conclude that Hispanics have higher mortality. Direct observation of the association between Hispanic origin and mortality (i.e., joint observation) is necessary to detect the truth. The critical nature of the joint observation requirement means that legal status imputation is not appropriate for analyses of outcomes that are not measured alongside legal status in the donor data set.

However, our simulations also showed that external data about legal status (i.e., prior information) could be used to reduce bias, even when joint observation does not occur. As the percentage of those classified as "probably legal" increased, bias in the estimates produced by the logical-imputation, demographic accounting, and logical-CSMI imputation methods decreased substantially. Similarly, CSMI estimates were improved by incorporating information about the association between unauthorized status and the dependent variable from an auxiliary data set, even when it was drawn from a different universe (California) than the other two data sets.

What do the results imply about prior research? Unfortunately, most legal status imputation methods violate the joint observation assumption and are therefore likely to produce estimates that are biased to an unknown degree. To our knowledge, no existing applications of the logical and single-imputation methods use information about the dependent variable to inform the imputation. Although the demographic accounting approach used by Pew/Passel is harder to evaluate, it also appears to violate the joint observation assumption because it imputes legal status on the basis of only a few predictors in the LPS survey (occupational distribution by age, sex, and state of residence), yet examines imputed legal status across a wide variety of outcomes.

Beyond this, prior approaches have clearly violated the same-universe requirement. For example, Marcelli and Heer (1997, 1998) used a local Los Angeles–based sample to impute unauthorized status for respondents in a sample that is representative of a larger geographic area; and Pew/Passel used an older sample (the LPS) to impute unauthorized status decades later. Similarly, Caponi and Plesca (2014) used data from the New Immigrant Survey (NIS) (Jasso et al. 2000), a nationally representative sample of legal permanent residents (LPRs) admitted to the United States in the early 2000s, to impute the legal status of all immigrants in the ACS. Although our tests involving the usage of an auxiliary data set suggest that the same-universe condition is less critical than the joint observation condition, we caution that more research is necessary to assess the conditions under which the same-universe assumption can be relaxed.

What do the results mean for future research on legal status? All the imputation methods we tested were limited in one way or another, suggesting that rather than continuing to try to spin gold from straw, the research community should increase efforts to improve data on immigrants. The most inexpensive and timely way to accomplish this would be to permit administrative record linkages to be used to logically impute legal status. Such data linkages already exist, but as we note earlier, U.S. federal statistical agencies have been reluctant to permit researchers to use these linkages to proxy legal status. A more expensive, but perhaps more ethically acceptable, route would be to add questions about legal status in surveys.

Recent evaluations of the quality of information gathered from survey questions on legal status are promising, and suggest that the addition of such items to questionnaires are unlikely to compromise survey response rates (Bachmeier et al. 2014).

Absent better data, the CSMI method has promise, as long as the joint observation condition is met. As demonstrated in our CPS example, this method has the potential to increase statistical power, thus enabling analyses for detailed subgroups and geographies. Although the joint observation condition appears to constrain the applicability of the CSMI method, the SIPP nevertheless includes a very rich set of outcomes in many topical modules. To the extent that the SIPP can be pooled with larger data sets with common variables, there are numerous opportunities to expand research on the unauthorized population. Of course, the SIPP is not perfect. It appears to undercount unauthorized immigrants, and handling its data on legal status can be challenging. When using the SIPP in research on legal status, we recommend following the methods outlined in Bachmeier et al. (2014). Additionally, when implementing the CSMI method, it is important that donor and target data sample from the same universe, although as noted earlier, our tests suggest that this restriction could be less important than the joint observation requirement. We refer readers to Rendall et al. (2013) for guidance on how to test for violations of this assumption. Finally, it is important that analysts employ best practices for multiple imputation, such as specifying the correct functional form, including all analytically important variables, and appropriately accounting for clustered observations in the imputation model (Allison 2002; Rubin 1987).

Research on immigration to the United States remains plagued by the lack of data available to researchers, precisely at a time when public policy discussions are most in need of input by social scientists. Even more problematic is the fact that the limited existing knowledge that we have about the characteristics of the unauthorized population has been derived from imputation approaches that the simulation exercises reported here have shown to yield biased estimates. Nevertheless, the simulation exercises have also demonstrated that social scientists have at their disposal reliable data and statistical methods for imputing legal status in large-scale surveys lacking such measures, insofar as important conditions are met. Although these conditions might appear to limit the utility of the CSMI method, it can nevertheless be employed to substantially expand the body of literature on the incorporation of the unauthorized population beyond the limited number of studies that currently comprise it.

Acknowledgments

This research was support by grants from the National Institutes of Health (RC2 HD064497, P01 HD062498, and 2R24HD041025). We thank Michelle Frisco, Molly Martin, Nancy Landale, Claire Altman, Susana Sanchez, and the anonymous reviewers for helpful comments.

References

Allison, PD. Missing data. Thousand Oaks, CA: Sage; 2002.

Bachmeier JD, Van Hook J, Bean FD. Can we measure immigrants' legal status? Lesson from two U.S. surveys. International Migration Review. 2014; 48:538–566. [PubMed: 25525285]

Batalova, J.; Hooker, S.; Capps, R. DACA at the two-year mark: A national and state profile of youth eligible and applying for deferred action. Washington, DC: Migration Policy Institute; 2014.

Bohn S, Lofstrom M, Raphael S. Did the 2007 Legal Arizona Workers Act reduce the state's unauthorized immigrant population? Review of Economics and Statistics. 2014; 96:258–269.

Bozick R, Miller T. In-state college tuition policies for undocumented immigrants: Implications for high school enrollment among non-citizen Mexican youth. Population Research and Policy Review. 2014; 33:13–30.

Bustamante AV, Fang H, Garza J, Carter-Pokras O, Wallace SP, Rizzo JA, Ortega AN. Variations in healthcare access and utilization among Mexican immigrants: The role of documentation status. Journal of Immigrant and Minority Health. 2012; 14:146–155. [PubMed: 20972853]

California Health Interview Survey. CHIS 2013–2014 sample design. Los Angeles, CA: UCLA Center for Health Policy Research; 2013.

Caponi V, Plesca M. Empirical characteristics of legal and illegal immigrants in the USA. Journal of Population Economics. 2014; 27:923–960.

Capps, R.; Bachmeier, JD.; Fix, M.; Van Hook, J. A demographic, socioeconomic, and health coverage profile of unauthorized immigrants in the United States. Washington, DC: Migration Policy Institute; 2013.

Clark RL, Glick JE, Bures RM. Immigrant families over the life course: Research directions and needs. Journal of Family Issues. 2009; 30:852–872.

Clark RL, King RB. Social and economic aspects of immigration. Annals of the New York Academy of Sciences. 2008; 1136:289–297. [PubMed: 18579888]

Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods. 2001; 6:330–351. [PubMed: 11778676]

Durand, J.; Massey, DS.; Capoferro, C. The new geography of Mexican immigration. In: Zuniga, V.; Hernandez-Leon, R., editors. New destinations: Mexican immigration in the United States. New York, NY: Russell Sage Foundation; 2005. p. 1-22.

Flores SM. State dream acts: The effect of in-state resident tuition policies and undocumented Latino students. Review of Higher Education. 2010; 33:239–283.

Greenman E, Hall M. Legal status and educational transitions for Mexican and Central American immigrant youth. Social Forces. 2013; 91:1475–1498.

Hall M, Greenman E, Farkas G. Legal status and wage disparities for Mexican immigrants. Social Forces. 2010; 89:491–513.

Heer DM, Passel JS. Comparison of two methods for estimating the number of undocumented Mexican adults in Los Angeles County. International Migration Review. 1987; 21:1446–1473. [PubMed: 12280920]

Jasso G, Massey DS, Rosenzweig MR, Smith JP. The New Immigrant Survey Pilot (NIS-P): Overview and new findings about U.S. legal immigrants at admission. Demography. 2000; 37:127–138. [PubMed: 10748994]

Javier JR, Huffman LC, Mendoza FS, Wise PH. Children with special health care needs: How immigrant status is related to health care access, health care utilization, and health status. Maternal and Child Health Journal. 2010; 14:567–579. [PubMed: 19554437]

Kaushal N. Amnesty programs and the labor market outcomes of undocumented workers. Journal of Human Resources. 2006; 41:631–647.

Ku L. Health insurance coverage and medical expenditures of immigrants and native-born citizens in the United States. American Journal of Public Health. 2009; 99:1322–1328. [PubMed: 19443829]

Little, RJA.; Rubin, DB. Statistical analyses with missing data. 2. New York, NY: John Wiley and Sons; 2002.

Marcelli EA. Unauthorized Mexican immigration, day labour and other lower-wage informal employment in California. Regional Studies. 2004; 38:1–13.

Marcelli EA, Heer D. Unauthorized Mexican workers in the 1990 Los Angeles County labour force. International Migration. 1997; 35:59–83. [PubMed: 12292469]

Marcelli EA, Heer D. Unauthorized Mexican immigration and welfare: A comparative statistical analysis. Sociological Perspectives. 1998; 41:279–302. [PubMed: 12294132]

Markides KS, Eschbach K. Aging, migration, and mortality: Current status of research on the Hispanic paradox. Journals of Gerontology: Series B. 2005; 60(Special Issue 2):S68–S75.

Massey DS, Bartley K. The changing legal status distribution of immigrants: A caution. International Migration Review. 2005; 39:469–482.

Passel, JS. The size and characteristics of the unauthorized migrant population in the US. Washington, DC: Pew Hispanic Center; 2006.

Passel, JS.; Clark, RL. Immigrants in New York: Their legal status, incomes, and taxes. Washington, DC: Urban Institute; 1998.

Passel, JS.; Cohn, DV. A portrait of unauthorized immigrants in the United States. Washington, DC: Pew Hispanic Center; 2009.

Potochnick S. How states can reduce the dropout rate for undocumented youth: The effects of in-state resident tuition policies. Social Science Research. 2014; 45:18–32. [PubMed: 24576624]

Rässler S. Data fusion: Identification problems, validity, and multiple imputation. Austrian Journal of Statistics. 2004; 33(1–2):153–171.

Rendall MS, Ghosh-Dastidar B, Weden MM, Baker EH, Nazarov Z. Multiple imputation for combined-survey estimation with incomplete regressors in one but not both surveys. Sociological Methods & Research. 2013; 42:483–530.

Resche-Rigon M, White IR, Bartlett JW, Peters SAE, Thompson SG. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. Statistics in Medicine. 2013; 32:4890–4905. [PubMed: 23857554]

Rodgers WL. An evaluation of statistical matching. Journal of Business and Economic Statistics. 1984; 2:91–102.

Rubin, DB. Multiple imputation for nonresponse in surveys. New York, NY: John Wiley and Sons; 1987.

Schenker N, Raghunathan TE, Bondarenko I. Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. Statistics in Medicine. 2010; 29:533–545. [PubMed: 20029804]

Sommers BD. Stuck between health and immigration reform—Care for undocumented immigrants. New England Journal of Medicine. 2013; 369:593–597. [PubMed: 23883331]

StataCorp. Stata statistical software: Release 13. College Station, TX: StataCorp LP; 2013.

State Health Access Data Assistance Center. State estimates of the low-income uninsured not eligible for the ACA Medicaid expansion. Minneapolis, MN: University of Minnesota; 2013. (Issue Brief No. 35)

Stevens GD, West-Wright CN, Tsai KY. Health insurance and access to care for families with young children in California, 2001–2005: Differences by immigration status. Journal of Immigrant and Minority Health. 2010; 12:273–281. [PubMed: 18780183]

U.S. Census Bureau. Survey of Income and Program Participation. Washington, DC: U.S. Census Bureau; 2013.

Table 1

Sample descriptive statistics (weighted means) for all analytic variables, 2004 SIPP

Variables	All Foreign-born	Probably Legal	Ambiguous Status	Unauthorized
Legal Status				
Probably legal	0.511	1.000	0.000	0.000
Ambiguous status	0.248	0.000	1.000	0.000
Unauthorized	0.241	0.000	0.000	1.000
Dependent Variables				
Insurance 1: Observed	0.684	0.848	0.586	0.435
Insurance 2: Ambiguous have low coverage	0.644	0.848	0.425	0.435
Insurance 3: Ambiguous have high coverage	0.725	0.848	0.751	0.435
Controls				
Income-to-poverty ratio (logged)	0.627	0.845	0.568	0.225
Years of education	11.745	12.573	11.333	10.414
Mexican-born	0.346	0.214	0.448	0.519
Years of U.S. residence	17.273	23.764	12.551	8.364
Age	42.959	48.845	37.863	35.718
Sex (male = 1)	0.498	0.465	0.519	0.547
Number of functional limitations	0.571	0.839	0.279	0.303
Self-rated health (fair/poor = 1)	0.122	0.167	0.078	0.072
N	8,898	4,633	2,201	2,064

Notes: Data and sample are from the 2004 Survey of Income and Program Participation, Topical Module 2. The sample is restricted to foreign-born persons age 16 and older.

Table 2

Data structure used for simulations, by missing data pattern

Missing Data Pattern (joint observation with unauthorized status)	Unauthorized Status	DV (insurance coverage)	IV (income- to-poverty ratio)	Controls			
				X_1	X_2	…	X_7
DV and IV Jointly Observed							
Donor ($N \approx 5{,}000$)	X	X	X	X	X	…	X
Target ($N \approx 5{,}000$)	.	X	X	X	X	…	X
DV Jointly Observed; IV Never Jointly Observed							
Donor ($N \approx 5{,}000$)	X	X	.	X	X	…	X
Target ($N \approx 5{,}000$)	.	X	X	X	X	…	X
DV Never Jointly Observed; IV Jointly Observed							
Donor ($N \approx 5{,}000$)	X	.	X	X	X	…	X
Target ($N \approx 5{,}000$)	.	X	X	X	X	…	X
DV and IV Never Jointly Observed							
Donor ($N \approx 5{,}000$)	X	.	.	X	X	…	X
Target ($N \approx 5{,}000$)	.	X	X	X	X	…	X

Notes: X = completely observed; . = completely missing; DV = dependent variable; IV = independent variable.

Table 3

Description of imputation methods

Method	Description
1 Logical	In target data, those appearing to be “probably legal” are coded as legal and the remaining as unauthorized.
2 Demographic Accounting	In target data, those appearing to be “probably legal” are coded as legal. Among remaining, single-imputation method used to assign targeted percentage as unauthorized. The rest are coded as legal.
3 Single Imputation	In target data, unauthorized = 1 if random draw from uniform distribution < predicted probability of being unauthorized; unauthorized = 0 otherwise; predicted probability is generated from a prediction equation estimated on the donor sample.
4 Cross-Survey Multiple Imputation (CSMI)	Pool donor and target data. Impute legal status using multiple imputation (multivariate imputation by chained equations).
5 Logical Cross-Survey Multiple Imputation (Logical-CSMI)	Pool donor and target data. Predicted probability of being unauthorized is coded as 0 for those who are “probably legal,” and this variable is included in the imputation model (multivariate imputation by chained equations).

Table 4

Simulated coefficients of unauthorized status in models predicting insurance coverage by imputation method (N = 10,000; 500 repetitions; DV and IV jointly observed)

	Insurance 1 (observed)				Insurance 2 (weak association)				Insurance 3 (strong association)			
Imputation Method	**β**	**SE**	**Bias**	**Relative Bias**	**β**	**SE**	**Bias**	**Relative Bias**	**β**	**SE**	**Bias**	**Relative Bias**
Expected Value	−0.151				−0.057				−0.237			
1 Logical	−0.164	0.014	−0.013	0.086	−0.245	0.014	−0.187	3.259	−0.061	0.013	0.176	−0.743
2 Demographic Accounting	−0.225	0.015	−0.073	0.485	−0.158	0.016	−0.101	1.757	−0.256	0.015	−0.019	0.079
3 Single Imputation	−0.153	0.015	−0.002	0.014	−0.060	0.015	−0.003	0.045	−0.238	0.014	−0.001	0.005
4 CSMI	−0.149	0.027	0.002	−0.016	−0.057	0.027	0.001	−0.015	−0.233	0.026	0.004	−0.016
5 Logical-CSMI	−0.150	0.025	0.002	−0.012	−0.061	0.025	−0.003	0.055	−0.231	0.024	0.006	−0.025

Notes: Coefficients are estimated on the target sample only ($N \approx 5{,}000$). Bias = estimate – expected value; Relative Bias = bias/expected value; CSMI = cross-survey multiple imputation.

Table 5

Simulated coefficients of unauthorized status in models predicting insurance coverage by imputation method and missing data pattern (N = 10,000; 500 repetitions)

Method	Missing Data Pattern (joint observation with unauthorized status)	Insurance 1 (observed)				Insurance 2 (weak association)				Insurance 3 (strong association)			
		β	SE	Bias	Relative Bias	β	SE	Bias	Relative Bias	β	SE	Bias	Relative Bias
Expected Value		−0.151				−0.057				−0.237			
1) Logical	DV and IV jointly observed	−0.164	0.014	−0.013	0.086	−0.245	0.014	−0.187	3.259	−0.061	0.013	0.176	−0.743
	DV jointly observed, IV never jointly observed	−0.164	0.014	−0.013	0.086	−0.245	0.014	−0.187	3.259	−0.061	0.013	0.176	−0.743
	DV never jointly observed, IV jointly observed	−0.164	0.014	−0.013	0.086	−0.245	0.014	−0.187	3.259	−0.061	0.013	0.176	−0.743
	DV and IV never jointly observed	−0.164	0.014	−0.013	0.086	−0.245	0.014	−0.187	3.259	−0.061	0.013	0.176	−0.743
2) Demographic Accounting	DV and IV jointly observed	−0.225	0.015	−0.073	0.485	−0.158	0.016	−0.101	1.757	−0.256	0.015	−0.019	0.079
	DV jointly observed, IV never jointly observed	−0.231	0.015	−0.080	0.526	−0.170	0.016	−0.113	1.963	−0.257	0.015	−0.020	0.085
	DV never jointly observed, IV jointly observed	−0.176	0.015	−0.024	0.161	−0.184	0.016	−0.127	2.205	−0.145	0.015	0.092	−0.389
	DV and IV never jointly observed	−0.175	0.015	−0.023	0.155	−0.180	0.016	−0.123	2.143	−0.141	0.015	0.096	−0.405
3) Single Imputation	DV and IV jointly observed	−0.153	0.015	−0.002	0.014	−0.060	0.015	−0.003	0.045	−0.238	0.014	−0.001	0.005
	DV jointly observed, IV never jointly observed	−0.157	0.015	−0.006	0.039	−0.068	0.015	−0.010	0.179	−0.240	0.014	−0.003	0.011
	DV never jointly observed, IV jointly observed	−0.097	0.015	0.054	−0.357	−0.082	0.015	−0.025	0.435	−0.098	0.014	0.139	−0.588
	DV and IV never jointly observed	−0.096	0.015	0.056	−0.368	−0.081	0.015	−0.023	0.407	−0.093	0.014	0.144	−0.606
4) CSMI	DV and IV jointly observed	−0.149	0.027	0.002	−0.016	−0.057	0.027	0.001	−0.015	−0.233	0.026	0.004	−0.016
	DV jointly observed, IV never jointly observed	−0.151	0.042	0.000	0.000	−0.049	0.043	0.008	−0.142	−0.242	0.039	−0.005	0.021
	DV never jointly observed, IV jointly observed	−0.125	0.102	0.026	−0.172	−0.105	0.099	−0.047	0.827	−0.125	0.102	0.112	−0.473
	DV and IV never jointly observed	−0.107	0.132	0.044	−0.294	−0.082	0.127	−0.024	0.426	−0.106	0.135	0.131	−0.553
5) Logical-CSMI	DV and IV jointly observed	−0.150	0.025	0.002	−0.012	−0.061	0.025	−0.003	0.055	−0.231	0.024	0.006	−0.025
	DV jointly observed, IV never jointly observed	−0.153	0.035	−0.001	0.007	−0.052	0.038	0.006	−0.097	−0.240	0.031	−0.003	0.013

Method	Missing Data Pattern (joint observation with unauthorized status)	Insurance 1 (observed)				Insurance 2 (weak association)				Insurance 3 (strong association)			
		β	SE	Bias	Relative Bias	β	SE	Bias	Relative Bias	β	SE	Bias	Relative Bias
	DV never jointly observed, IV jointly observed	−0.073	0.119	0.078	−0.515	−0.066	0.116	−0.008	0.146	−0.055	0.113	0.182	−0.767
	DV and IV never jointly observed	−0.040	0.133	0.111	−0.735	−0.023	0.129	0.034	−0.591	−0.034	0.127	0.203	−0.856

Notes: Coefficients are estimated on the target sample only ($N \approx 5{,}000$). Bias = estimate – expected value; Relative Bias = bias/expected value; CSMI = cross-survey multiple imputation.

Table 6

Simulated coefficients of unauthorized status in models predicting insurance coverage, by imputation method and percentage classified as "probably legal," (N = 10,000; 500 repetitions; DV never jointly observed, IV jointly observed)

Percentage of Those With Ambiguous Status Classified as "Probably Legal"	Insurance 1 (observed)				Insurance 2 (weak association)				Insurance 3 (strong association)			
	β	SE	Bias	Relative Bias	β	SE	Bias	Relative Bias	β	SE	Bias	Relative Bias
Expected Value	−0.151				−0.057				−0.237			
Logical Imputation												
0 %	−0.164	0.014	−0.013	0.086	−0.245	0.014	−0.187	3.259	−0.061	0.013	0.176	−0.743
50 %	−0.143	0.014	0.009	−0.056	−0.140	0.014	−0.083	1.439	−0.128	0.013	0.109	−0.459
75 %	−0.142	0.014	0.009	−0.059	−0.100	0.014	−0.042	0.739	−0.173	0.013	0.065	−0.272
90 %	−0.147	0.014	0.005	−0.030	−0.075	0.015	−0.017	0.304	−0.208	0.014	0.029	−0.122
Demographic Accounting												
0 %	−0.176	0.015	−0.024	0.161	−0.184	0.016	−0.127	2.205	−0.145	0.015	0.092	−0.389
50 %	−0.168	0.015	−0.016	0.107	−0.138	0.015	−0.081	1.410	−0.178	0.015	0.059	−0.249
75 %	−0.164	0.015	−0.012	0.080	−0.109	0.016	−0.051	0.893	−0.204	0.015	0.033	−0.139
90 %	−0.162	0.015	−0.011	0.069	−0.085	0.016	−0.028	0.480	−0.227	0.015	0.010	−0.043
Logical-CSMI												
0 %	−0.073	0.119	0.078	−0.515	−0.066	0.116	−0.008	0.146	−0.055	0.113	0.182	−0.767
50 %	−0.116	0.094	0.036	−0.237	−0.125	0.093	−0.067	1.170	−0.094	0.096	0.143	−0.603
75 %	−0.140	0.059	0.011	−0.072	−0.110	0.059	−0.053	0.919	−0.172	0.063	0.065	−0.275
90 %	−0.149	0.030	0.002	−0.015	−0.075	0.031	−0.018	0.307	−0.223	0.029	0.014	−0.061

Notes: Coefficients are estimated on the target sample only ($N \approx 5{,}000$). Bias = estimate – expected value; Relative Bias = bias/expected value; CSMI = cross-survey multiple imputation.

Table 7

Simulated coefficients of unauthorized status in models predicting insurance coverage estimated by CSMI method with and without an auxiliary data set (N = 15,000; 500 repetitions; DV never jointly observed, IV jointly observed)

	Insurance 1 (observed)				Insurance 2 (weak association)				Insurance 3 (strong association)			
Method	**β**	**SE**	**Bias**	**Relative Bias**	**β**	**SE**	**Bias**	**Relative Bias**	**β**	**SE**	**Bias**	**Relative Bias**
Expected Value	−0.151				−0.057				−0.237			
Cross-Survey Multiple Imputation (CSMI)												
Without auxiliary data set	−0.125	0.102	0.026	−0.172	−0.105	0.099	−0.047	0.827	−0.125	0.102	0.112	−0.473
With auxiliary data set												
Auxiliary data set has same universe as target and donor data (United States)	−0.153	0.027	−0.001	0.008	−0.060	0.027	−0.003	0.051	−0.236	0.026	0.001	−0.004
Auxiliary data set has different universe (California)	−0.171	0.027	−0.019	0.128	−0.048	0.027	0.009	−0.161	−0.237	0.027	0.000	0.001

Notes: Coefficients are estimated on the target sample only ($N \approx 5{,}000$). Bias = estimate − expected value; Relative Bias = bias/expected value; CSMI = cross-survey multiple imputation.

Table 8

Proportion of unauthorized immigrants with insurance coverage estimated by CSMI method, by state/region and imputation model specification

			Pooled 2004 SIPP and CPS					
			Insurance Never Jointly Observed		Insurance Jointly Observed			
	2004 SIPP		State/Region in Imputation Model		State/Region in Imputation Model		Imputed Separately by State/Region	
State/Region	Mean	SE	Mean	SE	Mean	SE	Mean	SE
California	0.416	0.023	0.524	0.010***	0.442	0.009	0.434	0.011
Florida	0.462	0.038	0.555	0.010*	0.449	0.008	0.468	0.011
Illinois	0.560	0.068	0.553	0.015	0.502	0.019	0.568	0.023
New Jersey	0.427	0.050	0.567	0.015**	0.463	0.017	0.458	0.020
New York	0.513	0.041	0.590	0.010	0.503	0.009	0.478	0.020
Texas	0.357	0.035	0.360	0.015	0.324	0.010	0.338	0.010
Other Pacific/Southwest (AZ, HI, NV, NM)	0.334	0.037	0.546	0.015***	0.407	0.016	0.412	0.016*
Northwest (CO, ID, MT, OR, UT, WA, WY)	0.456	0.039	0.473	0.012	0.400	0.009	0.420	0.029
Midwest (IN, IA, KS, MN, NE, ND, OK, SD)	0.533	0.045	0.564	0.008	0.543	0.014	0.550	0.020
Great Lakes Region (MI, OH, WI)	0.692	0.062	0.553	0.018*	0.556	0.029*	0.589	0.022
Northeast (CT, ME, MA, NH, PA, RI, VT)	0.518	0.043	0.671	0.008***	0.540	0.011	0.497	0.021
Mid-Atlantic (DE, DC, MD, VA)	0.583	0.054	0.556	0.017	0.476	0.024	0.521	0.030
Southeast (GA, NC, SC, WV)	0.259	0.035	0.406	0.014***	0.324	0.012	0.312	0.013
Deep South (AL, AK, KT, LA, MS, MO, TN)	0.453	0.062	0.627	0.016**	0.492	0.023	0.552	0.035

Notes: The sample is pooled. N = 30,112; SIPP N = 8,898; CPS N = 21,214; SIPP includes at least 250 foreign-born per state/region; CSMI = cross-survey multiple imputation. Significance tests are based on the SE of the difference between the estimate from the 2004 SIPP and the pooled SIPP-CPS:

$$SE_{diff} = \sqrt{\left(SE_{SIPP^2} + SE_{pooled^2}\right)}.$$

* $p < .05$;

** $p < .01$;

*** $p < .001$