9

Chapter

Case Studies -- II

Chair: Ann Hardy, Centers for Disease Control

Clifford Adelman

Jimmy Hwang ◆ Bo Kolody ◆ William A. Vega

Susan S. Jack

9

Chapter

The Thin Yellow Line: Editing and Imputation in a World of Third-Party Artifacts

Clifford Adelman, U.S. Department of Education

Abstract

Each of the longitudinal studies of the National Center for Education Statistics includes a file of transcripts from colleges, commmunity colleges and trade schools attended by survey participants. The transcripts are gathered at about age 30, coded by a contractor, and delivered to NCES. Given the idiosyncratic record-keeping practices of 2,500 institutions (in the most recent collection) and inconsistencies in coding of graduate students who usually do not know what they are looking at, the delivered files are a tangle of contradictions.

The editorial process takes 12-15 months to complete, and is carried out with interagency support from the National Science Foundation.

This paper both reports and demonstrates what has been learned from the editing of two such samples, the development of decision rules, and the feedback of the decision rules into the initial coding process. More importantly for data quality and standards, the paper demonstrates where the line between editing and imputation lies in such an archive, and how the survey data guides the editor in making (rare) decisions to impute with respect to key variables.

The case in point is the most important variable for student records, namely, the credential/degree earned.

The Thin Yellow Line: Editing and Imputation in a World of Third-Party Artifacts

Clifford Adelman, U.S. Department of Education

Introduction

The tasks of editing and the occasions of imputation in data sets using third-party documents raise an ontological and epistemological issue in the same breath. When one imputes a phenomenon, one asserts its existence; and it is legitimate to ask for measures of confidence that the phenomenon, in fact, exists. The very process of imputation implies that the phenomenon did not emerge *ex nihilo*, rather is dependent or derived. The epistemological challenge lies in the identification of qualities of other, known phenomena on which the imputation depends or from which it is derived. The strength of those qualities and the logic of derivation determine how you know the phenomenon. The greater the strength of those qualities and the longer and more stable the histories of relationships among the variables described in the data set, the less likely you are guessing when information is missing. The less likely the guessing, the more the task is dominated by "editing" and the less by "imputation."

On the other hand, as the strength of these qualities diminishes to near zero, the greater the leaps of faith. At some point, the chasm between a phenomenon and its potential representation is so wide that only imagination can cross it. As much as we value imagination in the history of civilization, its place in data set construction is rather limited.

Editing is always involved in imputation, since the process of data editing identifies the missing. But it is difficult to describe the point on a continuum at which the balance of editing and imputation tips toward the latter. The task is somewhat akin to establishing a passing score on a "high stakes" test such as a licensure exam. A great deal of empirical evidence is assembled, and replication of results with different populations at different times is a necessary procedure. Both test publishers and test users wish to minimize the cases of false positives and false negatives. Minimize. A passing score is a guide, not an absolute. The passing score is like that thin yellow line down the middle of a country road: you want to keep the right traffic going in the right direction on either side.

In some data sets, e.g., those that produce the CPI, there are high stakes consequences of false imputations. Even though individuals are not being judged with the same consequences as a licensure examination, individuals are directly affected by the CPI. In other national data sets used in the course of policy-setting by states, institutions, and organizations, the stakes are not as high, and individuals are more indirectly affected. While these data sets are constructed and used to estimate aggregates, the direct and indirect effects on individuals argue that whenever we approach the mixture of editing and imputation, confidence levels are critical standards.

Third-Party Artifacts: the Historian's Stuff

The mass of data editing is conducted on the second-party level. That is, information is collected directly from subjects or their agents. The mechanism is a survey or an unobtrusive measure that is unmediated. I have a questionnaire, you are interviewed, and your responses are directly transcribed or transcribed by an interviewer who follows very standard and tight protocols. In the language of data editing, these protocols function at the capture stage. Or you engage in a discrete activity, such as filing for unemployment benefits, that leaves a direct, unmediated trace.

A third-party artifact involves a different order of evidence. Archaeologists, anthropologists and historians know it well. You engage in activities that are recorded for *sui generis* purposes. They are recorded in formats and symbolisms that can best be described, in Geertz's phrase, as "local knowledge." They are inscribed on documents we can call "artifacts." At some future time, these artifacts are discovered, collected from many sources, and re-recorded in a standardized format by a third-party. The original artifact is thus twice removed from the form in which it appears in a database. Depending on the bureaucratic protocols of the collection, the data can be edited in either a coding or post-coding phase.

Take, for example, the debarkation lists of boats arriving on the eastern seaboard in the National Period of our history. In the 1820s and 1830s, customs agents in Charleston, Baltimore, Philadelphia, New York, and Boston recorded information on the nature and destination of arriving immigrants. In no two ports was there a standardized form for doing so. Sometimes we got full names; sometimes not; sometimes gender, age, occupation, relationships and ethnicity; sometimes not -- or, in the case of the Irish, negative ethnic stereotypes and a sorting based on skin color. Data for key variables are always missing. There is no one port for which they are complete.

There is one exception to the missing variables: the name of the ship and its arrival date.

If we are building a modern database from these lists, we have a phenomenological choice: we can accept the classifications made by the customs agents as reflecting the views/perceptions of the customs agents -- in which case, we'd be writing a database that is more about the customs agents than the immigrants; or we can look for ways to fill in the information. We cannot imagine information of this type. And it is very difficult to impute.

But if we examine the port records on the other side of the Atlantic, we find that the embarkation lists are often more detailed than the debarkation lists. The level of detail was particularly high in ports such as Hamburg and Rostok. The Library of Congress possesses some of this material. For a complete examination, the investigator takes the name of the ship and checks the registries until the port of origin can be determined; goes to the port of origin, rummages around the archives, and finds the manifest. Sweat, toil, tedium. Nice travel, but hardly suited to an instant electronic environment, and not in the habits of data teams that have to release the monthly *Consumer Price Index* at 8:30 on a Thursday morning, hot-decking the price of laundry detergent in Seattle on the basis of analogous products.

It took Fernand Braudel and his associates 20 years to write *The History of the Mediterranean in the Age of Philip II*, meticulously hunting down meteorological data for the entire basin covering a period of a millenium, let alone records of caravans and harvests. We could save Braudel a lot of time today. The question is whether we want to and how. For our task, in many respects, would be just as tedious. Third-party documents are often like that.

The Case of College Records

We all generate unobtrusive records in our lives, records that become part of the grist for analyses of economic, social, and public health issues and trends. The case discussed here is that of college transcripts, a type of record now generated for over half the adult population between the ages of 19 and 35. One of our jobs in the research and statistics division of the U. S. Department of Education is figuring out what goes on in that vast and sometime amorphous enterprise called U. S. higher education, and whether, where, and how our individual and collective investments in higher education have convincingly measurable impacts on our worklives, citizenship, and adult development.

To determine what goes on, we could always perform a content analysis of college catalogues. Unfortunately, these documents tend to be higher education's contribution to American fiction and ought to be placed on appropriate shelves in the library. Or we could ask individuals, in the course Computer-Assisted Telephone Interviews (CATI) or paper/pencil interviews, what types of education they pursued after high school, in what kinds of institutions. That strategy, as we've discovered, leads to what can euphemistically be described as exaggeration--but, as we shall see, not always, at least in the matter of degrees earned.

In terms of what students study, we can examine enrollment surveys conducted by learned and professional societies. What we discover quickly, though, is that enrollments are not students. Rather, they represent the same students cycling themselves through many courses within the field(s) covered by the enrollment survey. No learned society will admit that fact because each is concerned with getting the maximum share of what we ex-deans call "enrollment mix." Having eliminated enrollment surveys, we can try course schedules. Like the catalogues, these at least show what was really offered. Unfortunately, the evidence of course schedules does not indicate whether enough students registered for a course to make it a "go." I made the George Washington University course schedule look fairly interesting a couple of semesters in a row with a course on quantitative historical methods that used those immigration lists, focused on women's roles, and wound up with a collective class project using the first *Women's Who's Who* (1916). Exactly one person signed up both times and the course was cancelled.

So we turn to transcripts. They don't lie, they don't exaggerate, they don't forget. But they are a mess. And they are even more a mess because, to arrive in a national database, they are re-coded in a standardized form by graduate students working for a contractor. Graduate students are supposed to be smart; but in the matter of the documents at issue, they are not fluent in the histories of the variables nor experienced in translating the oftimes idiosyncratic formats and signs used on those documents. Put more simply, they have little idea of what they are looking at. The editor's job is to spot and fix their errors.

For certain tasks, such as coding courses into over 1,000 course categories from an empirically-derived taxonomy, we gave the graduate student coders the assistance of "search strings." Given the existence of certain words in the title of the course, the search string presents the coder with a range of possibilities for coding. The coders choose. But the problem with search strings is that they do not provide decision rules for context. The coders either choose incorrectly or resort to residual categories for "unknowns" on 20 percent of the entries.

For example, if they are presented with a title such as "Composition and Conversation" in the junior year of a student at a selective college and code it as English Composition, I will wager they are wrong. In that situation, I look for context and derivation, and will immediately scan for the foreign language courses on the transcript that may set the context. If the title were merely "Composition," I would look for music or studio art as a guide to correct coding.

This example illustrates the editorial process, post-coding. If the context determines that "Composition and Conversation" most likely applies to a Russian language course and I recode it according, I am not changing the reality, rather making the mark of reality "fit" or "represent" the reality more accurately than the form in which the mark was delivered. But course titles such as "TEN BADTAB TEN" or GREEN BOX WORKSHOP or RAGS TO RICHES or THE GOOD LIFE or (yes) GOOD BOOKS, I would rather leave alone.

National Samples, Unique Institutions

We have taken two national samples of college transcripts in the course of longitudinal studies. I have edited both of them, and, in each case, the editing process took two years. The samples are very robust. The first (known as the NLS-72) involved 12,600 students, 19,500 transcripts, and 485,000 courses. The second ("High School and Beyond/Sophomore Cohort") was smaller, but has taken no less time: 8,400 students, 13,300 transcripts, 320,000 courses. For each course, there are 18 variables to which I must pay attention. For each transcript record, another 10. Into this mix, I can import other variables from CATI interviews, paper and pencil surveys, and high school records. The purpose of importing is to guide the decision-making process in determining the accuracy of data coding and entry.

While much common-sense guides decision rules, specialized knowledge is absolutely necessary. When the coders read, on an MIT transcript, "Math 1," and code it as a remedial course, you wonder how much common sense can be impaired. But when they read a sequence from an engineering student at, let us say, Wisconsin, who has Calc 1, 2, 3, 4 and they code all of those courses as Calculus, you can forgive their ignorance. There is a big difference between elementary functions and infinite series, and that difference is important for understanding the careers of engineering students. A lot of engineering deans and advisers want to know. The data editor cannot be a copy editor, rather someone who has to know a great deal about how specific colleges, community colleges, and trade schools work. This knowledge sets up a web of dependencies on which the editing decisions rest.

In Search of Accuracy: Consulting the Source

Unlike researching debarkation lists in the early 19th century, we have another choice with contemporary data bases: we can call the source. Given the uniqueness of institutions, accuracy in editing requires contacting their registrars to assist in interpretation. Third-party data from similar contemporary sources allow for such a procedure. Surveys of the current or recent recorded activities of individuals in health facilities using different record-keeping systems would be a good analogue.

Following the contractor's delivery of the tape for the High School & Beyond college transcript file, we made a list of schools where there appeared to be a great deal of inconsistency and contradiction in matters of credits, grades, dates, and course titles. There were 700 schools out of 2,500 where these problems were rampant. We telephoned them. It took three months to get the guidelines. Sometimes, the registrars didn't know the answers to our questions. After all, colleges are in a market, and try to grab a niche, finding any way they can to be unique. The evidence of such dubious niches include complex credit value systems or academic calendars of such a nature that students probably need pacemakers to tell them when to go to class.

Why is accuracy important and how much should one sacrifice accuracy for timeliness? We get bad legislation if we are not accurate. The Student Right-to-Know Act (1990) is a premier example. Under this act, colleges are required to report rates of graduation (or, in the case of community colleges, persistence). We understood the problem with this legislation: over half the undergraduates in this U.S. attend more than one institution, and (as it turns out), more than 20% change institutions across state lines (rendering it impossible for any state higher education authority to track them). The student may start in a college in South Carolina but graduate from a college in North Carolina. The first school is penalized by the propaganda of published graduation rates under Student Right-to-Know. And the cost to both institutions to produce information that few parents or students actually use exceeds the benefits.

But we were not there in time to testify on this legislation because the data sets were riddled with errors. Time to degree; average credits to degree. State legislators deserve--at the least--an accurate national tapestry that provides some norms. If they don't get it, or if the data are sloppy, somebody will suffer. Our accuracy is also critical to interpretation of labor market data. I have asked field interviewers from the Census Bureau and the Bureau of Labor Statistics how they know when the person who--given a reasonable demographic profile--says he/she is a doctor is, in fact, a physician? A data editing system within our framework would inquire whether, according to the evidence of transcripts, the person possessed the requisite educational credentials and history. If these credentials are missing, the person is probably either a physician's assistant or some other kind of "doctor," and their occupation code should be edited accordingly.

The Mediator and Code 590

Again, unlike the 19th century debarkation list case, we can also go back to the third-party when the same reality is represented in both literal (unmediated) and symbolic (mediated) form. The third-party is responsible for the mediated form that is delivered on tape as a database. When we don't understand the symbol, we can ask for the literal.

For example, in looking at the occupations of individuals in the NLS-72 database when these people were 32/33 years old (in 1986), I saw that a substantial number received the occupation code "590." The coding manual, a collection of symbols used in the mediating process, did not list "590." I telephoned the contractor and asked them what "590" signified. After a pause, the response came back: "Craftsmen in the Military." This was a strange response, particularly as 65% of the people in the "Code 590" bin held bachelor's degrees (the evidence came from the transcripts). I then asked the contractor to send "the literals," the direct transcriptions of what the respondent wrote (or said, if the data collection method was CATI) on the questionnaire. It turned out that one-third of the respondents assigned to Code 590 were active-duty military, one-third were civilian employees of the Department of Defence, and one-third belonged in occupation/industry categories that had nothing to do with the military.

The correct way to represent the occupation/industry of an accountant at Bolling Air Force Base is occupation=accountant and industry=U.S. military. Unfortunately, the coding scheme used by the mediator did not distinguish the military from the most aggregate notion of government employer. The results of the Code 590 inquiry thus included not only the corrections for hundreds of miscoded cases, but also a recasting of the "industry" variable to disaggregate the military from civilian government agencies.

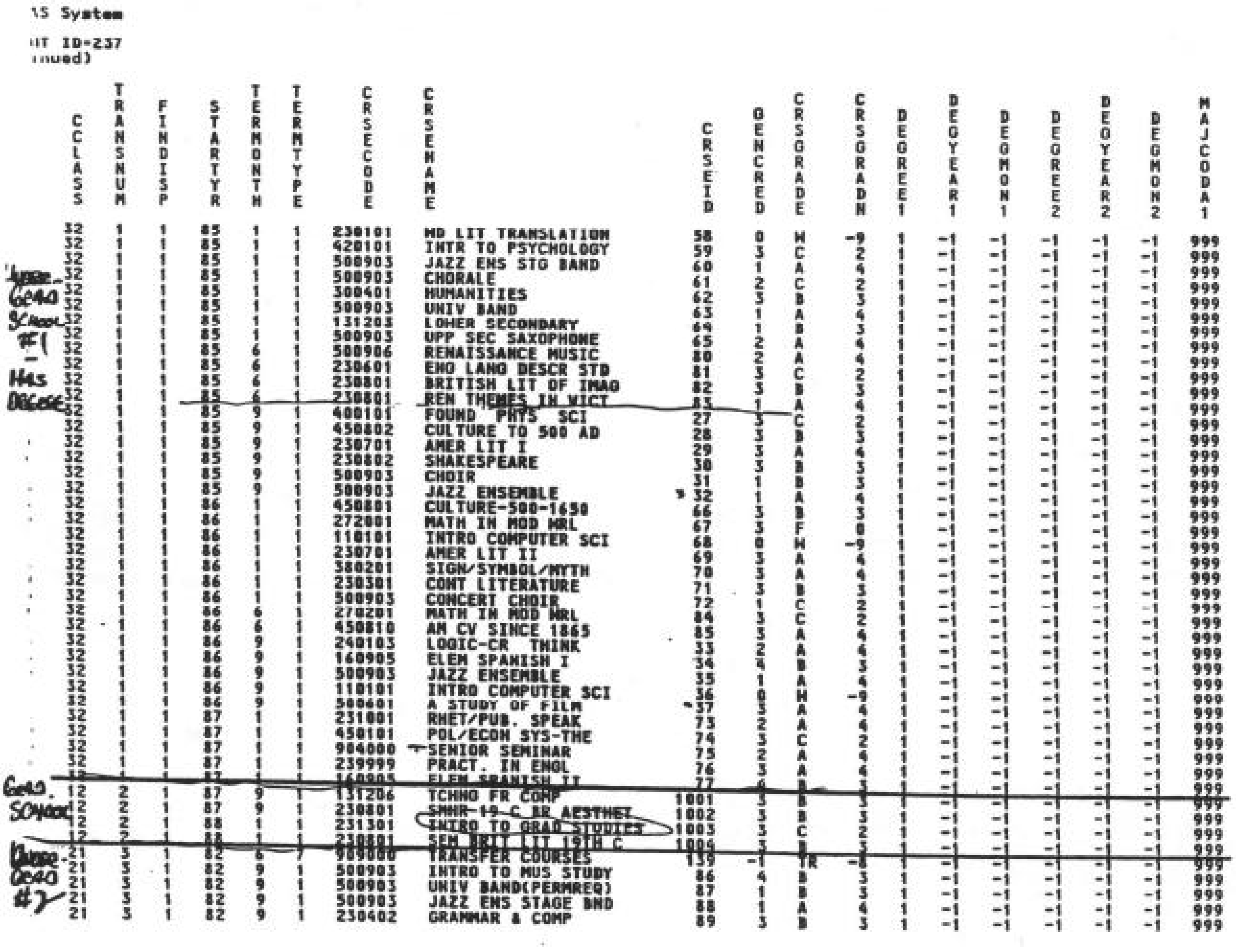
Errors and Imputation

In these brief accounts of contexts overriding search strings, consulting the sources, and sending the "literals" all I have described is editing. Imputation did not enter these transactions.

What is a true imputation in this business, and where does the limit lie? We are invited to impute, I contend, only in those cases that cannot be labelled "errors" or the result of errors, and that involve missing information.

Let us illustrate with variations on the most critical issue in the review of national college transcript samples: whether a student received a degree. Figure 1 is an actual page from the computerized records of a student as we created those records from the tape delivered by the mediator. The page does not represent the entire record for this particular student, but contains enough information to illustrate the case.

Figure 1.--Excerpt From Sample Student Transcript Records



```
\S System
IT ID=237
nued)

                 CCLASS TRANSNUM FINDISP STARTYR TERMONTH TERMTYPE CRSECODE CRSENAME                 CRSEID GENCRED CRSGRADE CRSGRADN DEGREE1 DEGYEAR1 DEGMON1 DEGREE2 DEGYEAR2 DEGMON2 MAJCODA1

                   32      1       1      85      1        1      230101  MD LIT TRANSLATION          58      0      M       -9        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      1        1      420101  INTR TO PSYCHOLOGY          59      3      C        2        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      1        1      500903  JAZZ ENS STG BAND           60      1      A        4        1      -1       -1      -1      -1       -1     999
Under-             32      1       1      85      1        1      500903  CHORALE                     61      2      C        2        1      -1       -1      -1      -1       -1     999
Grad               32      1       1      85      1        1      300401  HUMANITIES                  62      3      B        3        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      1        1      500903  UNIV BAND                   63      1      A        4        1      -1       -1      -1      -1       -1     999
School             32      1       1      85      1        1      131203  LOWER SECONDARY             64      1      B        3        1      -1       -1      -1      -1       -1     999
#1                 32      1       1      85      1        1      500903  UPP SEC SAXOPHONE           65      2      A        4        1      -1       -1      -1      -1       -1     999
-                  32      1       1      85      6        1      500906  RENAISSANCE MUSIC           80      2      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      6        1      230601  ENG LANG DESCR STD          81      3      C        2        1      -1       -1      -1      -1       -1     999
Has                32      1       1      85      6        1      230801  BRITISH LIT OF IMAG         82      3      B        3        1      -1       -1      -1      -1       -1     999
Degree             32      1       1      85      6        1      230801  REN THEMES IN VICT          83      1      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      9        1      400101  FOUND  PHYS  SCI            27      3      C        2        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      9        1      450802  CULTURE TO 500 AD           28      3      B        3        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      9        1      230701  AMER LIT I                  29      3      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      9        1      230802  SHAKESPEARE                 30      3      B        3        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      9        1      500903  CHOIR                       31      1      B        3        1      -1       -1      -1      -1       -1     999
                   32      1       1      85      9        1      500903  JAZZ ENSEMBLE               32      1      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      1        1      450801  CULTURE-500-1650            66      3      B        3        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      1        1      272001  MATH IN MOD WRL             67      3      F        0        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      1        1      110101  INTRO COMPUTER SCI          68      0      M       -9        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      1        1      230701  AMER LIT II                 69      3      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      1        1      380201  SIGN/SYMBOL/MYTH            70      3      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      1        1      230301  CONT LITERATURE             71      3      B        3        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      1        1      500903  CONCERT CHOIR               72      1      C        2        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      6        1      270201  MATH IN MOD WRL             84      3      C        2        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      6        1      450810  AM CV SINCE 1865            85      3      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      9        1      240103  LOGIC-CR  THINK             33      2      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      9        1      160905  ELEM SPANISH I              34      4      B        3        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      9        1      500903  JAZZ ENSEMBLE               35      1      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      9        1      110101  INTRO COMPUTER SCI          36      0      M       -9        1      -1       -1      -1      -1       -1     999
                   32      1       1      86      9        1      500401  A STUDY OF FILM             37      3      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      87      1        1      231001  RHET/PUB. SPEAK             73      2      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      87      1        1      450101  POL/ECON SYS-THE            74      3      C        2        1      -1       -1      -1      -1       -1     999
                   32      1       1      87      1        1      904000  SENIOR SEMINAR              75      2      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      87      1        1      239999  PRACT. IN ENGL              76      3      A        4        1      -1       -1      -1      -1       -1     999
                   32      1       1      87      1        1      160905  ELEM SPANISH II             77      3      B        3        1      -1       -1      -1      -1       -1     999
Grad.              12      2       1      87      9        1      131206  TCHNG FR COMP             1001      3      B        3        1      -1       -1      -1      -1       -1     999
School             12      2       1      87      9        1      230801  ~~SMNR-19-C BR AESTHET~~  1002      3      B        3        1      -1       -1      -1      -1       -1     999
                   12      2       1      88      1        1      231301  INTRO TO GRAD STUDIES     1003      3      C        2        1      -1       -1      -1      -1       -1     999
                   12      2       1      88      1        1      230801  SEM BRIT LIT 19TH C       1004      3      B        3        1      -1       -1      -1      -1       -1     999
Under-             21      3       1      82      6        7      909000  TRANSFER COURSES           139     -1      TR      -8        1      -1       -1      -1      -1       -1     999
Grad               21      3       1      82      9        1      500903  INTRO TO MUS STUDY          86      4      B        3        1      -1       -1      -1      -1       -1     999
#2                 21      3       1      82      9        1      500903  UNIV BAND(PERMREQ)          87      1      B        3        1      -1       -1      -1      -1       -1     999
                   21      3       1      82      9        1      500903  JAZZ ENS STAGE BND          88      1      A        4        1      -1       -1      -1      -1       -1     999
                   21      3       1      82      9        1      230402  GRAMMAR & COMP              89      3      B        3        1      -1       -1      -1      -1       -1     999
```

The contractor interviewed the student (the process is carried out prior to and separate from the gathering of the transcript). The student said she had a bachelor's degree. She listed the three schools she attended. We requested transcripts from the schools and received all three of them. The student has 135 credits on Transcript #1 from a liberal arts college (CCLASS=32). Neither a degree nor degree date are indicated, despite what appears to be a decent academic record and a course entitled, "SENIOR SEMINAR" in the year one would expect a 1982 high school graduate to be receiving a college degree. There is also a graduate school transcript (#2). The editorial process spots all these characteristics, and considers the student's claim to a degree against the missing information about the degree in the third-party presentation.

The evidence allows us to impute a bachelor's degree, a major in English, and a degree date of May, 1987 (her last term of undergraduate attendance was a semester that began in January of 1987, she entered graduate school in September of that year, and schools with semester systems hold commencements in May). There are virtually no degrees of freedom in this imputation. Our confidence level is very high.

On the continuum of balance between editing and imputation, there is more of the former than the latter in this case. Why? There are three historical relationships between the evidence and the receipt of a bachelor's degree that are strong enough to say that what appears to be "missing" is more the result of oversight (a form of "error"): entry to graduate school in an academic discipline; numbers of credits earned; and senior seminars in a pattern of courses with a dominant field (English literature) and an ostensible GPA comfortably at or above the norm.

How do I know that similar cases will produce similar results? The editor of a data set based on third-party artifacts has an ethical obligation to examine a sample of original artifacts in cases where such critical signs are missing. For the High School & Beyond college transcript sample, I looked at over 100 records where degrees were in question. In 70 percent of these cases, the degree was, in fact, indicated on a transcript, usually on the back side of a page the data entry person never turned over. That evidence was sufficient to justify similar imputations.

Consider a second variation in the case of our 1987 English Lit major: what if there was no graduate school transcript, and all we had were two fragmentary records with poor grades, remedial work and lots of withdrawals? No matter what degree the student claimed she had earned, there would be no clues, nothing from which the claim could be validated with confidence. The decision to leave the third-party presentation of the record alone is wholly an editorial decision: there are no errors.

And a third variation: what if the degree-bearing transcript was submitted by the college as a "blocked transcript"? A blocked transcript indicates only the degree, degree date, and major. No course work, credits, or grades are included. The "blocked transcript" is the student's choice, and we must respect that choice. At the same time, our data standards indicate that it is irresponsible to present degree data without course and credit data. So these data, which are missing but not as a result of error, are imputed.

In this imputation, we employ what lawyers call "custom and usage" guidelines, that is, special knowledge of organizational behavior and rules in colleges and universities. We know the student earned at least 120 semester credits (the accepted -- and empirical -- minimum for a bachelor's degree), a degree in English (with, by custom, at least 30 credits in the major), that the degree was awarded in May, 1987 and that the student graduated from high school in the spring of 1982. Furthermore, the student has told us, in the surveys, what she was doing (school, work, other) for every month since 1982. We can thus enter blocks of courses and credits, by term, e.g.,

IMPUTED UNDERGRAD COURSEWORK	1985 SPRING	15 CREDITS
IMPUTED MAJOR COURSEWORK	1985 FALL	15 CREDITS.

The former entry receives a special course code for imputed courses of unknown discipline. The latter entry receives the course code, within a field, that covers unknown, missing, and residual cases. As for grades, only "CR" is entered, along with a flag that tells the software to add the credits but not to include the course in the computation of GPA. These entries are put in a flat file which, when completed, is appended to the master data set.

In entries such as these, imputation outweighs editing, but must be limited to variables and values that can be asserted with confidence. How do we know that the balance has shifted? The phenomena did not exist previously in the universe of representations we call a database and could not be created algorithmically by reference to existing signs in that data base. At the same time, they are not imaginary phenomena: they are dependent and derived, and hence are clearly on the imputation side of that thin yellow line.

Invitation

Exploratory papers usually invite readers to further research and reflection. This occasion is no exception. To develop theory and guidelines for imputation in data sets built from third-party artifacts requires investigations in a variety of fields. Recent economic history, public health, and mass communications are all inviting areas. Of these, only public health has been compelling enough to produce systematic data collection of national scope. It is obviously the next terrain for charting the thin yellow line. ■

9

Chapter

Sampling Design and Estimation Properties of a Study of Perinatal Substance Exposure in California

Jimmy Hwang, University of California (San Diego),
Bo Kolody, San Diego State University, and
William A. Vega, University of California (Berkeley)

Abstract

During the period of 1992-93, a group of researchers from UC Berkeley, San Diego State University, the State Department of Alcohol and Drug Programs, and the University of California at San Diego conducted a comprehensive survey on substance abuse problems among pregnant women in California. A fully probabilistic stratified cluster sample was used to estimate the prevalence of perinatal drug exposure for the state of California. Included in the sampling plan were 29,494 pregnant women presenting for delivery in 202 hospitals, which were sampled from 602 hospitals throughout the state of California. Urine specimens were taken from women presenting for delivery and later linked by code number to demographic variables, tobacco use and prescribed drug data gathered at intake. Urine specimens were then shipped and tested at a NIDA certified lab. Based on the survey results, the study further projected that there were about 67,361 perinatal exposures to one or more non-prescribed drug, including alcohol, and 52,346 exposures to tobacco in California.

The findings of the study have many significant clinical and public health implications. The purpose of the presentation is to offer several practical experiences in data editing and exposition from the study. The discussion will be useful in the area of applied statistics and the implications of sampling survey. Since the study was the first of this kind in substance abuse programs and in sampling targeted subjects, the presentation will illustrate its innovative and unique sampling design of the study. The presentation will also present the problems and solutions, advantages and disadvantages of employing certain weighting procedures in adjusting the estimates. A detailed description is provided about the sampling process, its rationale, sampling factors, statistical estimation, and computational procedures.

Sampling Design and Estimation Properties of a Study of Perinatal Substance Exposure in California

Jimmy Hwang, University of California (San Diego),
Bo Kolody, San Diego State University, and
William A. Vega, University of California (Berkeley)

Introduction

A study was conducted according to a multistage probability sampling design to estimate the prevalence of perinatal substance exposure in California in 1992. The study used coded urine samples from 29,494 women presenting for delivery in 202 hospitals, screened for toxins; and later linked the results of toxicology by code number to the subjects' demographic variables and their reported use of tobacco and prescribed drugs. The study reported the survey results by age, marital status, county of residence, ethnicity and prenatal care history for state-wide and regional estimates. The findings have many significant clinical and public health implications (Vega et al., 1993a and 1993b). This paper presents a general discussion of the sampling process and statistical design of the study. The presentation is useful in the area of applied statistics and the implications of sampling survey.

Sampling Process and Its Rationale

The most important considerations governing the choice of the sampling design in the study involved several factors.

The sampling frame be representative of all births taking place in maternity hospitals in California. The practical constraint limited the study to pregnant women admitted to maternity hospitals at time of delivery. To ensure sampling efficiency and study feasibility, the study included only hospitals that had more than 10 births annually, and excluded federal hospitals, hospitals on military bases, hospitals that delivered babies on an emergency room only basis, and birthing centers. Births in these hospitals account for a proportion of about 2 percent of statewide births. Their exclusion would not bias the estimates.

The sampling procedure be fully probabilistic and thus yield population estimates with known sampling errors. Sampling of study subjects within a hospital was not based upon subject characteristics (e.g., race/ethnicity). The study defined all admissions within a specified time frame of March through October in 1992 as study subjects. This course of action necessitated a large sample size in order to have sufficient women from all ethnic groups enter the sample through a process of natural selection.

The study attempted to estimate regional differences in California. Since approximately 80% of births are in the ten counties with the highest number of births, a key objective of the study was to derive separate prevalence estimates for each of these ten counties as well as for the remaining forty-eight counties as aggregated into sampling strata. The strata design that was used conformed to geoadministrative county clusters previously established as Health Service Areas (HSAs). Although the possibility of using HSAs was considered, they were too numerous (N=26) for acceptable confidence intervals of subgroups. The county and county clusters (i.e., sampling strata), based on fourteen HSAs and the ten largest counties by birth population in California, constitute meaningful geographic divisions, for which separate estimates are both feasible and desirable. As a result, there were 11 counties (out of 58) that did not have a hospital in the study. Two of them had a hospital that did not want to participate. The others either did not have a hospital, or their one hospital was not randomly chosen. The stratum design allows for the absence of any one hospital from any one county in a stratum by allowing for grouping of results from hospitals in the other counties in the stratum. Based on the stratum design used in the study, the results obtained for any stratum can be generalized for any county in the stratum.

The sampling frame for the study was 583,487 births (98 percent of births statewide) in approximately 305 maternity hospitals in California. There were 21 sampling strata in the study, one for each HSA and large county and combining these 21 sub-samples into a single statewide sample. The sampling strata vary considerably in terms of their number of births.

For example, Los Angeles county recorded 206,457 births while Imperial county recorded 2,777 births. Given these wide disparities, the objective was to draw a sample sufficiently large for reasonably precise estimates in the small sampling strata while allowing for large n's in large sampling strata. A strictly proportionate to size allocation would render the larger strata too large or the smaller strata too small.

For Stratum #1.00, for example, the sampling fraction is n=1107 while in the largest sampling stratum (#11.00) n=4879. This strategy yielded an overall, statewide sample size n=29,200. Given the substantially larger sample size in Stratum #11.00 (Los Angeles county) the prevalence estimates were approximately twice as precise as in Stratum #1.00. This higher precision is desirable inasmuch as Los Angeles county accounts for about one third of statewide births. When combining stratum prevalence rates into a statewide rate, the higher precision for Los Angeles county, as well as for other large strata, minimized the impact of weighting, which was used to adjust for stratum size. Table 1 gives the actual numbers for each stratum.

Two-Stage Probability Sampling

Within each stratum hospitals form the clusters, the first stage sampling units. The method of systematic sampling was used to select hospitals within the stratum. A separate prevalence estimate could be made for each of the 21 strata. The size of each of these samples was proportionate to the number of births in the stratum.

The selection procedure began by listing hospitals ordered on ownership type. Ownership was used in order to assure that a representative proportion of women entering every type of hospitals was included in adequate numbers for the sampling design. Within type, hospitals were ordered on the annual number of births. A systematic sampling procedure was used to delete every third hospital from this ordered hospital list. Across the 21 strata, this procedure yielded a sample of 202 hospitals that were included in the study.

Table 1.--Target and Actual Samples in the Study

Stratum #	Number of Births	n=Target Samples to be Collected	Actual Number of Usable Samples	Counties
1.00	10,135	1,107	1,173	Northern CA
2.00	9,013	512	546	Golden Empire
2.34	20,969	1,358	1,429	Sacramento
3.00	12,710	1,253	1,337	North Bay
4.00	8,174	510	540	West Bay
4.38	14,589	1,198	1,198	San Francisco
5.00	11,401	1,333	1,267	Contra Costa
5.01	22,757	1,676	1,639	Alameda County
6.00	21,701	1,178	1,172	N.San Joaquin
7.00	31,380	1,575	1,657	Santa Clara
8.00	13,188	1,129	1,194	Mid Coast
9.00	20,036	1,154	1,158	Central CA
9.10	15,681	1,141	1,063	Fresno County
10.00	17,458	1,144	1,097	Ventura/S.B.
11.00	206,457	4,879	4,918	L.A. County
12.00	4,200	101	0	Inyo/Mono
12.33	22,813	1,241	1,278	Riverside County
12.36	28,434	1,547	1,530	San Bernardino Co.
13.00	53,678	2,682	2,714	Orange County
14.00	2,777	183	203	Imperial County
14.37	45,936	2,299	2,381	San Diego County
Totals	593,487	29,200	29,494	

The number of subjects sampled within selected hospitals was set to be directly proportionate to the number of births, specifically, the proportion of stratum births during the 1990-1991 fiscal year. To adjust for any disproportion due to slight over or under sampling by hospital size or ethnicity, weights were applied to conform the outcomes to the 1991-92 parameters on ethnic distributions in each hospital. To achieve the statewide estimates, appropriate weights were also applied to adjust for disproportion by stratum to conform the total sample to the statewide 1991-92 distributions on race/ethnicity by hospital and stratum.

Anonymity and Urine Testing

Subjects were selected in a manner designed to minimize selection bias and to ensure the anonymity of those from whom urine specimens were obtained. Starting on a given day, nurses were instructed to test all admitted patients until the sampling fraction for the hospital was met. Nurses collected urine specimens and basic descriptive information from each subject and recorded this information on a code sheet that contained no personally identifying information. The same code number was used on the code sheet and the urine-specimen label.

In accordance with national standards of nursing care, all patients in California hospitals are asked at the time of admission whether they currently smoke. There is no standard phrasing for this question. The answer does not indicate the frequency of extent of smoking. Information on patients' smoking was recorded on the code sheets; it was missing for 6.4 percent of the subjects in the sample. The direction of bias, if any, due to the missing data could not be judged.

Procedures and safeguards were established to ensure that no personally identifying information could be linked to the results of urine testing. The research protocol was reviewed and approved by the Committee for the Protection of Human Subjects, Health and Welfare Agency, State of California and by the Human Subjects Review Committee of the University of California, Berkeley, as well as by institutional review boards at the individual hosptials. The urine specimens were sent to a laboratory certified by the National Institute of Drug Abuse (PharmChem, Menlo Park, California). Table 2 lists the tests performed and the detection periods for each substance.

Each specimen was assayed by personnel who did not know its origin or the subject's demographic characteristics. They used enzyme-multiplied immunoassay techniques for all drugs and an enzymatic assay for alcohol (Test materials for these assays were manufactured by Syva, a subsidiary of Syntex, Palo Alto, Califorina.). If a screening test was **negative, i.e., failed to detect the presence of drugs or alcohol, the test results were reported to be negative and no further testing was conducted. If an enzyme-multiplied immunoassay was positive for any drug except a cannabinoid (marijuana), the result was confirmed by gas chromatography; this technique is commonly used by analytical toxicologists to confirm the presence of drugs or their metabolites in urine or other biologic fluids. If an assay was positive for a cannabinoid, the result was confirmed by high-performance thin-layer chromatography; the result of this technique correlates highly with that of gas chromatography, and the sensitivity of the test is comparable. When an immunoassay is combined with an appropriate confirmation assay that is chemically independent, the likelihood that a drug will be correctly identified is greater than 99 percent.

The pharmacological characteristics of alcohol differentiate it from other drugs. Its concentration in blood, breath, and urine can be estimated according to body weight if the amount ingested and the time elapsed are known; alcohol is rapidly absorbed into the bloodstream and distributed throughout body water. If an average-sized person in good health drinks 6 oz of beer (170 ml), 2 oz of wine (60 ml), or 0.5 oz of distilled spirits (15 ml), urine collected 1 to 1.5 hours later should contain at least 10 mg of alcohol per deciliter, a level used as the cutoff value for the study. A person who has one or two drinks at night and urinates after awakening in the morning would have a lower urinary alcohol level and not test positive. A person who consumes at least 1 oz of distilled spirits 2 to 2.5 hours before urine collection would test positive. Alcohol cannot be measured accurately in urine specimens containing glucose; to avoid confounding, the study considered such specimens to be negative for alcohol.

Statistical Estimation

A separate prevalence and errors estimate based on two-stage sampling design was made for each of the 21 strata. Calculation of standard errors took the sampling design into account by weighting values to conform the data to the distribution of births for the period 1991-1992 within hospitals, within regions, and statewide. The 95 percent confidence interval for each subgroup reflected the observed percentage of positive tests and the standard error of the estimate. To facilitate subgroup comparison, differences between proportions were tested, with the Student-Newman-Keuls ranges adjustment (a two-tailed test) for multiple comprisons, in analyses of prevalence according to racial or ethnic group and the duration of prenatal care. The comparison of the prevalence of tobacco use included the exact t-test value and probability for each substance.

Table 2.--Drug and Alcohol Testing Procedures

Drug	Screening Method	Cutoff	Method	Confirmation Cutoff	Detection Period
Alcohol	EA	10 mg/dl			
Ethanol			GC	10 mg/dl	Very short*
Glucose			GC	1 mg/dl	
Amphetamine	EMIT	1000ng/ml	GC	300 ng/ml	
Amphetamine			GC	300 ng/ml	12-72 hr
Methamphetamine			GC	300 ng/ml	12-72 hr
Barbiturates	EMIT	200ng/ml	GC	200 ng/ml	
Amobartital			GC	200 ng/ml	2-4 days
Butalbital			GC	200 ng/ml	2-4 days
Pentobarbital			GC	200 ng/ml	2-4 days
Phenobarbital			GC	500 ng/ml	up to 30 days
Secobarbital			GC	200 ng/ml	2-4 days
Benzodiazepines	EMIT	200ng/ml	GC	200 ng/ml	
ACB			GC	500 ng/ml	up to 30 days
MACB			GC	500 ng/ml	up to 30 days
Cannabinoid	EMIT	50ng/ml	HPTLC	50 ng/ml	
THC metabolite			HPTLC	50 ng/ml	up to 14 days**
Cocaine metabolite	EMIT	300ng/ml	GC	500 ng/ml	
Benzoylecgonine			GC	500 ng/ml	12-72 hr
Methadone	EMIT	300ng/ml	GC	300 ng/ml	
Methadone			GC	300 ng/ml	1-4 days
Opiates	EMIT	300ng/ml	GC	20 ng/ml	
Codeine#			GC	500 ng/ml	2-4 days
Hydromorphone			GC	1000 ng/ml	2-4 days
Morphine			GC	200 ng/ml	2-4 days
Phencyclidine	EMIT	25ng/ml	GC	200 ng/ml	
Phencyclidine			GC	200 ng/ml	up to 14 days**

Note: ACB denotes acetylbenzophenone, MACB Methylacetylbenzophenone, THC tetrahydrocanabinol; EA denotes Enzymatic assay, EMIT enzyme-multiplied immunoassay technique; GC denotes gas chromatography, and HPTLC high-performance thin-layer Chromatography.

The advantage of two-stage sampling is obvious in this study. We have the opportunity of obtaining some smaller number of observations that apear more efficient and hence produce more precise estimates. In our study, we first used systematic elimination to select n hospitals in each stratum. Then, for each selected unit, a random method was given for selecting the specified numbers of subjects from it. In finding the mean and variance of the prevalence estimate, averages were taken over all samples that were generated by the process. To calculate the average, we first averaged the estimate over all second-stage selections that were drawn from the n hospitals. Then we averaged over all possible selections of n hospitals by the study.

For an estimate, p, the method can be expressed as

$$E(p) = E_1[E_2(p)],$$

where E denotes expected or average value over all samples, E_2 denotes averaging over all possible second-stage selections from a fixed set of hospitals, and E_1 denotes averaging over all first-stage selections.

For the variance V(p), it is readily shown the following result (Cochran, 1977),

$$V(p) = V_1[E_2(p)] + E_1[V_2(p)] ,$$

where $V_2(p)$ is the variance over all possible subsample selections for a given set of units.

To illustrate the computational implications, for any subject in a hospital H, say y_{iH} , let y_{iH} be 1 if the subject possesses substance of our interest and 0 otherwise. Let n_H be the number of sampled subjects in the hospital H and N_H be the number of births in the hospital H. Then, the prevalence of a given substance for the hospital H, p_H, is defined as:

$$p_H = \frac{\sum_i y_{iH}}{n_H}$$

The prevalence of a given substance for the stratum A, p_A, is defined as:

$$p_A = \frac{\sum_H \sum_i W_A y_{iH}}{\sum_H n_H}$$

where W_A is the weighting value based on probability of sampling allocations of the hospitals for each stratum.

Alternatively, the prevalence for the stratum A may be calculated as:

$$p_A = \frac{\sum_H W_A n_H p_H}{\sum_H n_H}$$

This two-stage probability sampling involves two sources of the variance for the prevalence of the stratum: variance between hospitals and variance within hospitals. Although the weighting value is omitted from the following derivation of the variance, the weighted prevalence and weighted standard errors were used in the study.

For each stratum, the variance between hospitals may be defined as:

$$s_1{}^2 = \frac{\sum_H (p_H - P_A){}^\wedge 2}{(n_{sA} - 1)} * \frac{(1 - \frac{n_{sA}}{n_{tA}})}{n_{sA}} \quad ,$$

where n_{sA} is the number of sampled hospitals and n_{tA} is the number of the total hospitals within the stratum.

The variance within hospitals may be defined as:

$$s_2{}^2 = \frac{\sum_H n_H (p_H q_H)}{n_{sA}(n_H - 1)} \frac{(\frac{n_{sA}}{n_{tA}})(1 - \frac{n_H}{N_H})}{N_H n_{sA}} \quad .$$

Therefore, the variance of the prevalence is the summation of the variance between hospitals and the variance within hospitals, i.e., $s_1{}^2 + s_2{}^2$.

The standard error of the prevalence is bounded (at 95% level) by $1.96*(s_1{}^2+s_2{}^2)^{0.5}$.

A Computational Example

As a computational example, the Table 3 gives the prevalence and error estimate of using alcohol for African-American women at time of delivery in stratum #11.00 (Los Angeles county). To derive the error estimate, we need to have the following information for each selected hospital: prevalence (p_H), sampled subjects (n_H), and total births (N_H) (see column 2-4). We also need the following stratum information: the number of selected hospitals (n_{sA}), the number of total hospitals (n_{tA}), and stratum prevalence (p_A) (see column 5-7).

Computational Steps

- ❐ Calculate the squared differences between sample means and population mean (i.e., p_H and p_A). (Column 8)
- ❐ Divide Column 8 by (n_{sA}-1) to calculate the sampling variances. (Column 9)
- ❐ Adjust the sampling variances by probability factor, $(1-n_{sA}/n_{tA})/n_{sA}$, to yield the variances between hospitals. (Column 10)
- ❐ Calculate the product of $n_H(p_H)(1-p_H)$. (Column 11)
- ❐ Divide Column 11 by n_{sA}(nH-1) to yield the sample variances for each hospital. (Column 12)
- ❐ Adjust the sample variances by probability factor, i.e., $\frac{(\frac{n_{sA}}{n_{tA}})(1 - \frac{n_H}{N_H})}{N_H n_{sA}}$. (Column 13)

Table 3.--A Computational Example: Alcohol Use of African-American Women in Los Angeles County

HID (C1)	PS1 (C2)	NSAMP (C3)	BTOTAL (C4)	NSHOSP (C5)	NTHOSP (C6)	P01 (C7)	(C8)	(C9)	(C10)	(C11)	(C12)	(C13)
133	28.571	12	3330	50	77	11.689	2.85	0.06	0.00	244.90	0.45	0.00
134	0.000	3	1254	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
136	50.000	2	860	50	77	11.689	14.68	0.30	0.00	50.00	1.00	0.00
146	12.698	47	5330	50	77	11.689	0.01	0.00	0.00	521.04	0.23	0.00
151	8.947	63	2947	50	77	11.689	0.08	0.00	0.00	513.25	0.17	0.00
164	0.000	4	1048	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
168	0.000	1	1102	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
174	4.546	9	1300	50	77	11.689	0.51	0.01	0.00	39.05	0.10	0.00
177	0.000	1	4329	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
180	50.000	3	1939	50	77	11.689	14.68	0.30	0.00	75.00	0.75	0.00
182	0.000	1	1473	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
183	0.000	1	1763	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
187	0.000	1	1391	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
191	7.143	31	8309	50	77	11.689	0.21	0.00	0.00	205.61	0.14	0.00
195	66.667	5	1708	50	77	11.689	30.23	0.62	0.00	111.11	0.56	0.00
196	0.000	22	5534	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
204	6.849	27	2249	50	77	11.689	0.23	0.00	0.00	172.26	0.13	0.00
205	4.546	20	4229	50	77	11.689	0.51	0.01	0.00	86.78	0.09	0.00
207	0.000	1	973	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
213	50.000	3	2584	50	77	11.689	14.68	0.30	0.00	75.00	0.75	0.00
217	10.526	10	5475	50	77	11.689	0.01	0.00	0.00	94.18	0.21	0.00
219	19.231	12	14578	50	77	11.689	0.57	0.01	0.00	186.39	0.34	0.00
223	18.605	36	8225	50	77	11.689	0.48	0.01	0.00	545.16	0.31	0.00
227	13.636	24	5248	50	77	11.689	0.04	0.00	0.00	282.65	0.25	0.00
228	23.684	8	554	50	77	11.689	1.44	0.03	0.00	144.60	0.41	0.00
229	0.000	1	710	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
233	0.000	1	955	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
237	0.000	2	1500	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
240	0.000	1	2275	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
254	0.000	7	2269	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
261	0.000	1	1889	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
263	0.000	1	2457	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
265	0.000	5	2966	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
267	0.000	6	1322	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
275	0.000	1	662	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
277	0.000	1	2677	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
278	16.667	6	2430	50	77	11.689	0.25	0.01	0.00	83.33	0.33	0.00
279	0.000	1	589	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
280	0.000	2	1843	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
287	15.625	70	4922	50	77	11.689	0.15	0.00	0.00	922.85	0.27	0.00
289	0.000	1	1698	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
290	0.000	4	2609	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
297	0.000	15	4120	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
303	0.000	4	2266	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
305	0.000	1	1198	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
309	0.000	2	1600	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
310	0.000	1	1174	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
315	9.091	5	5412	50	77	11.689	0.07	0.00	0.00	41.32	0.21	0.00
702	0.000	1	952	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
729	0.000	5	2263	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00

Variance between hospitals = 1.7750 Variance between hospitals = 0.0667

Standard Error Estimate = 1.7750 + 0.0667 =1.8417

- ❐ The variance of the two-stage sampling is the summation of Column 10 and Column 13. The square root of the sum produces the standard error.
- ❐ The error estimate at the 95 percent confidence level is multiplying 1.96 by the standard error.

Results and Discussion

The total prevalence for any category of drug takes into account whether the women presenting for delivery had any drug administered prior to urine collection. The nurse filling out the data form was asked to check any category of drug that was administered. Those categories were Opiates, Benzodiazepines, Barbiturates, Other, and None. If a urine tested positive for any drug in any of these categories and if the nurse checked any of these categories as a previously administered drug, then the urine was not counted in the prevalence total for that category.

The Illicit Drug prevalence total was for any drug positive for the following substances: THC, Cocaine, Methamphetamine, Phencyclidine, and Heroin. The Non-illicit Drug prevalence total was for any drug positive for the following substances only if the nurse did not indicate that a drug in a drug category was administered prior to collection of the urine sample: Amphetamine, all Barbiturates, all Benzodiazepines, Methadone, Codeine, Hydromorphone, and non-illicit Morphine use. In other words, if the nurse indicated that a barbiturate was administrated prior to urine collection, then whatever urine positive PharmChem Laboratories detected (e.g., amobarbital) was neither counted in the Barbiturates total, nor in the Non-Illicit Drug Total. The prevalence total for Non-Illicit Drugs is the total of all non-illicit drugs not administered prior to urine collection.

Finally, totals for all drug categories as well as the total drug and alcohol prevalence will always be lower than the sum of the individual drugs for two reasons: first, multiple drug use, though not common, was still generally one-half of 1 percent; and second, PharmChem Laboratories reported all positives as they tested them. In reporting totals by drug category, urine samples from patients who had drugs administered prior to urine collection were not considered positive. PharmChem had no way of knowing which positive urine samples were due to prescribed drug use. This was determined only after matching the data forms with the urine samples.

Table 4 presents statewide prevalence rates overall and by race/ethnicity. From the table, the statewide prevalence rate of perinatal substance use among California women in 1992 was 11.35 percent. Illicit substance exposure was 3.49 percent, with subestimates of 1.11 percent for cocaine and 1.88 percent for marijuana. The sub-estimate for alcohol is 6.72 percent and the tobacco use 8.82 percent. Only about 0.48 percent had used more than one non-illicit or illicit drug. These results suggest that alcohol and tobacco use during pregnancy is quite common in California. About 1 in 20 pregnant women had used one or more non-illicit or illicit drug, not including alcohol and tobacco, 1 in 14 used alcohol, and 1 in 11 used tobacco in the hours or days before hospitalization for delivery.

From Table 4, the highest rate of alcohol, 11.5 percent, illicit drug use 11.9 percent, and tobacco use, 20.12 percent were found among African American women; contrasts between African Americans and other ethnic subgroups are statistically significant in every instance ($p < .05$). Cocaine prevalence was high at 7.79 percent, as was marijuana (THC metabolite) at 4.59 percent. One of 4 African American women tested positive for a licit or an illicit drug use at time of hospitalization for delivery.

Table 4.--Statewide Prevalence Rates Overall and by Race/Ethnicity

Sample Size n=	Overall	Asian	African	Hispanic	White	Other
	29494	1645	2280	13194	10615	1142
Substance	PP ET	PP ET	PP ET	PP ET	PP ET	PP ET
1. Alcohol	6.72(.30)	5.07(1.08)	11.58(1.34)	6.87(.44)	6.05(.23)	4.03(1.16)
2. Amphetamines	.66(.10)	.06(.12)	.19(.18)	.35(.10)	1.32(.11)	.24(.30)
3. Barbiturates	.26(.06)	.19(.22)	.23(.20)	.22(.08)	.32(.05)	.22(.28)
4. Benzodiazepines	.09(.04)	.00(.00)	.30(.22)	.03(.04)	.15(.04)	.01(.04)
5. THC Metabolite	1.88(.16)	.21(.22)	4.59(.88)	.61(.14)	3.25(.17)	1.21(.64)
6. Cocaine	1.11(.12)	.06(.12)	7.79(1.12)	.55(.12)	.60(.07)	.20(.26)
7. Methadone	.15(.04)	.00(.00)	.25(.22)	.16(.06)	.16(.04)	.06(.14)
8. Opiates	1.47(.14)	1.34(.56)	2.54(.66)	1.06(.18)	1.59(.12)	1.11(.62)
9. Phencyclidine	.04(.02)	.00(.00)	.16(.16)	.06(.04)	.01(.01)	.00(.00)
10. Total Drug Positives (Any illicit or non-illicit drug from 2-9)	5.16(.26)	1.82(.66)	14.22(1.46)	2.75(.28)	6.79(.24)	2.88(1.00)
11. Polydrug Use (Positive for more than one drug)	.48(.08)	.04(.10)	1.75(.54)	.27(.08)	.57(.07)	.17(.24)
12. Polydrug Use (Positive for both Alcohol and any drug)	.52(.08)	.05(.10)	1.77(.56)	.25(.08)	.56(.07)	.15(.24)
13. Total Any Positive (Alcohol or drugs)	11.35(.36)	6.84(1.24)	24.02(1.78)	9.37(.50)	12.28(.32)	6.76(1.48)
14. Illicit Drugs	3.49(.22)	.39(.30)	11.90(1.36)	1.51(.22)	4.92(.21)	1.57(.74)
15. Non-illicit Drugs	1.71(.16)	1.49(.60)	2.38(.64)	1.26(.20)	1.96(.13)	1.31(.68)
16. Tobacco	8.82(.34)	1.73(.66)	20.12(1.76)	3.29(.32)	14.82(.35)	4.81(1.30)

White non-Hispanic women had the second highest rates of licit and illicit drugs with the exception of alcohol. White non-Hispanic women had much higher rates of smoking, 14.92 percent, than all other ethnic sub-groups except for African Americans. One in eight White non-Hispanic women tested positive for a licit or an illicit drug.

Hispanic women had the second highest rates of alcohol use, 6.87 percent. However, they had lower prevalence rates for all other drugs. About one in ten Hispanic women tested positive for a licit or an illicit drug.

Asian and Pacific Islander women had negligible prevalence for all drugs except for alcohol, 5.07 percent. One in fourteen Asian and Pacific Islander women tested positive for a licit or an illicit drug.

Table 5 presents regional prevalence estimates of total drug and/or alcohol use in descending order. The table is designed to facilitate direct comparisons of regions statewide, and to illustrate how some regions have lower prevalence for some substances, and have the highest prevalence for other substances. No one or two regions consistently have the highest or the lowest prevalence for all the substances. In

Table 5.--Regional Prevalence Rates in Descending Order by Overall and Drug Use

	Overall		Alcohol		Illicit Drugs		Non-illicit		Tobacco
Sampling Region	%	Sampling Region	%	Sampling Region	%	Sampling Region	%	Sampling Region	%
Golden Empire (2.00)	17.54	Golden Empire (2.00)	10.67	North Bay (3.00)	8.03	Alameda Co. (5.01)	3.72	Northern CA (1.00)	21.04
Alameda Co. (5.01)	16.92	Alameda Co. (5.01)	10.05	Contra Costs Co. (5.00)	6.82	San Bern. Co. (12.36)	2.78	Sacramento Co.(2.34)	15.23
Contra Costa Co. (5.00)	16.44	Contra Costa Co. (5.00)	9.03	Sacramento Co.(2.34)	6.7	N. San Joaquin (6.00)	2.7	North Bay (3.00)	15.03
Sacramento Co.(2.34)	15.21	Sacramento Co.(2.34)	7.7	Alameda Co. (5.01)	6.13	Northern CA (1.00)	2.5	Contra Costa Co. (5.00)	14.82
San Bern. Co. (12.36)	14.62	Mid Coast (8.00)	7.7	Northern CA (1.00)	6.12	San Frans. Co. (4.38)	2.35	Golden Empire (2.00)	14.79
Northern CA (1.00)	14.04	Imperial Co. (14.00)	7.42	Golden Empire (2.00)	5.88	Sacramento Co.(2.34)	2.1	San Bern. Co. (12.36)	12.81
North Bay (3.00)	14.04	Fresno Co. (9.10)	7.29	San Bern. Co. (12.36)	5.36	Fresno Co. (9.10)	1.89	Central CA (9.00)	12.18
N. San Joaquin (6.00)	12.65	LA Co. (11.00)	6.94	N. San Joaquin (6.00)	4.06	Golden Empire (2.00)	1.77	N. San Joaquin (6.00)	11.97
Fresno Co. (9.10)	12.61	Marin/San Mat (4.00)	6.9	Central CA (9.00)	3.96	LA Co. (11.00)	1.75	Alameda Co. (5.01)	11.42
Mid Coast (8.00)	11.86	San Bern. Co. (12.36)	6.85	Fresno Co. (9.10)	3.72	Santa Clara Co. (7.00)	1.73	Riverside Co. (12.33)	10.63
Riverside Co. (12.33)	11.36	San Diego Co. (14.37)	6.68	Riverside Co. (12.33)	3.53	Contra Costa Co. (5.00)	1.67	Mid Coast (8.00)	9.67
San Frans. Co. (4.38)	11.12	Riverside Co. (12.33)	6.58	Sn Frans. Co. (4.38)	3.47	Riverside Co. (12.33)	1.46	San Diego Co. (14.37)	9.6
LA Co. (11.00)	10.77	Ven/S.B. (10.00)	6.49	Mid Coast (8.00)	3.37	Ven/S.B. (10.00)	1.44	Fresno Co. (9.10)	9.52
Ven/S.B. (10.00)	10.04	North Bay (3.00)	6.12	Santa Clara Co. (7.00)	3.21	Orange Co. (13.00)	1.36	Ven/S.B. (10.00)	7.81
Santa Clara Co. (7.00)	9.76	N. San Joaquin (6.00)	6.1	LA Co. (11.00)	2.6	Mid Coast (8.00)	1.01	Santa Clara Co. (7.00)	7.3
Central CA (9.00)	9.51	San Frans. Co. (4.38)	5.99	Marin/San Mat (4.00)	2.42	Imperial Co. (14.00)	0.98	San Frans. Co. (4.38)	6.6
Imperial Co. (14.00)	9.44	Northern CA (1.00)	5.96	San Diego Co. (14.37)	2.35	Central CA (9.00)	0.95	Orange Co. (13.00)	5.88
San Diego Co. (14.37)	9.44	Santa Clara Co. (7.00)	5.5	Ven/S.B. (10.00)	2.21	North Bay (3.00)	0.79	LA Co. (11.00)	5.84
Marin/San Mat (4.00)	9.36	Central CA (9.00)	4.78	Orange Co. (13.00)	2.06	San Diego Co. (14.37)	0.61	Marin/San Mat (4.00)	5.73
Orange Co. (13.00)	7.49	Orange Co. (13.00)	4.49	Imperial Co. (14.00)	1.04	Marin/San Mat (4.00)	0.53	Imperial Co. (14.00)	4.7
Inyo/Mono (12.00)		Did not participate in the study							

	Marijuana		Amphetamines		Opiates		Cocaine
Sampling Region	%	Sampling Region	%	Sampling Region	%	Sampling Region	%
North Bay (3.00)	6.36	San Bern. Co. (12.36)	2.8	Alameda Co. (5.01)	3.51	Alameda Co. (5.01)	3.21
Golden Empire (2.00)	4.48	Golden Empire (2.00)	1.73	San Bern. Co. (12.36)	2.6	San Frans. Co. (4.38)	2.15
Contra Costa Co. (5.00)	4.41	Riverside Co. (12.33)	1.45	Northern CA (1.00)	2.59	Contra Costa Co. (5.00)	2.04
Northern CA (1.00)	4.4	Sacramento Co.(2.34)	1.45	N. San Joaquin (6.00)	2.45	Fresno Co. (9.10)	1.96
Sacramento Co.(2.34)	3.89	Northern CA (1.00)	1.44	San Frans. Co. (4.38)	2.41	Sacramento Co.(2.34)	1.67
N. San Joaquin (6.00)	2.58	N. San Joaquin (6.00)	1.42	Sacramento Co.(2.34)	2.03	LA Co. (11.00)	1.4
Central CA (9.00)	2.55	Contra Costa Co. (5.00)	1.4	Golden Empire (2.00)	1.77	San Bern. Co. (12.36)	0.95
San Bern. Co. (12.36)	2.55	Central CA (9.00)	1.09	Contra Costa Co. (5.00)	1.64	Santa Clara Co. (7.00)	0.9
Mid Coast (8.00)	2.43	North Bay (3.00)	0.85	Fresno Co. (9.10)	1.55	Mid Coast (8.00)	0.87
Alameda Co. (5.01)	2.35	Imperial Co. (14.00)	0.52	Santa Clara Co. (7.00)	1.5	North Bay (3.00)	0.78
Santa Clara Co. (7.00)	2	Alameda Co. (5.01)	0.47	LA Co. (11.00)	1.3	Central CA (9.00)	0.76
Riverside Co. (12.33)	1.9	San Diego Co. (14.37)	0.45	Riverside Co. (12.33)	1.29	Orange Co. (13.00)	0.7
Ven/S.B. (10.00)	1.48	Ven/S.B. (10.00)	0.45	Orange Co. (13.00)	1.11	Marin/San Mat (4.00)	0.59
Fresno Co. (9.10)	1.46	Fresno Co. (9.10)	0.43	Mid Coast (8.00)	1.08	Riverside Co. (12.33)	0.56
San Diego Co. (14.37)	1.42	Marin/San Mat (4.00)	0.34	North Bay (3.00)	1.04	San Diego Co. (14.37)	0.53
Marin/San Mat (4.00)	1.22	Orange Co. (13.00)	0.32	Ven/S.B. (10.00)	0.99	Northern CA (1.00)	0.41
Orange Co. (13.00)	1.15	Santa Clara Co. (7.00)	0.31	Central CA (9.00)	0.72	Ven/S.B. (10.00)	0.38
San Frans. Co. (4.38)	1.15	Mid Coast (8.00)	0.25	Marin/San Mat (4.00)	0.7	N. San Joaquin (6.00)	0.33
LA Co. (11.00)	1.03	LA Co. (11.00)	0.22	San Diego Co. (14.37)	0.52	Golden Empire (2.00)	0.1
Imperial Co. (14.00)	0.52	San Frans. Co. (4.38)	0.18	Imperial Co. (14.00)	0.49	Imperial Co. (14.00)	0
Inyo/Mono (12.00)		Did not participate in the study					

sum, there is one outstanding feature of the comparisons: prevalence estimates are not related to rural or urban strata per se, but both urban and rural regions in the northern part of California are more likely to have higher total prevalence rates than those in the southern part of California.

For example, the four regions with the highest total prevalence rates (for either drugs and/or alcohol) are in the northern part of California, and nine of the first ten are in the central and northern part of California. The only exception is San Bernardino county. While some predominantly rural regions have high total drug and/or alcohol use, other have relatively low use, such as Imperical county. These results reveal one other interesting finding. There seem to be large variations in prevalence rates among counties within the same geographic area. Such is the case in the region surrounding San Francisco Bay. The relatively low prevalence levels of Marin-San Mateo (Stratum #4.00), 9.36 percent, contrast with those of Contra Costa (Stratum #5.00) and Alameda (Stratum #5.01) counties, 16.44 percent and 16.92 percent, respectively. The rates for the North Bay (Stratum #3.00), Santa Clara (Stratum #7.00) and San Francisco (Stratum #4.38) counties range at all points in between. In the southern part of California, similar results are found when comparing San Bernardino (Stratum #12.26) and Orange (Stratum #13.00) counties. These variations underscore the complexity of interpreting these findings, and the importance of conducting a more detailed analysis to understand the reasons for this distribution.

While it is evident that regions in the northern part of California generally have higher prevalence rates than regions in the southern part of California, this does not seem attributable to income differences among counties/regions. However, there probably is a direct link to sociodemographic composition of the regions. Regions in the southern part of California with proportionally large Hispanic populations had lower prevalence on illicit drugs and tobacco use because Hispanic women had a low statewide prevalence for these substances generally. However, it remains enigmatic why White non-Hispanic maternity patients or Asian and Pacific Islander maternity patients, for example, in the northern part of California seemed so much more likely to test positive for various substances during the final term of their pregnancy than women in the same race/ethnic group in the southern part of California.

It appears that there is a strong cultural basis for perinatal substance use in California that operates to minimize use, or influence the decision to use particular types of substances. There also appeared to be "hot spots" of concentrated substance use, such as cocaine in Alameda county, amphetamines in San Bernardino county, and marijuana in the North Bay HSA (Napa, Solano, and Sonoma counties). These important differences in regional characteristics underlying the estimates suggest that they have critical public health implications for targeting services.

References

Vega, William A.; Kolody, Bohdan; and Hwang, Jimmy (1993a). Prevalence and Magnitude of Perinatal Substance Exposures in California, *The New England Journal of Medicine*, 329:850-854.

Vega, William A.; Kolody, Bohdan; and Hwang, Jimmy (1993b). *Profile of Alcohol and Drug Use During Pregnancy in California, 1992,* California Health and Welfare Agency.

Cochran, William G. (1977). *Sampling Techniques*, John Wiley & Sons, 276. ■

9

Chapter

The Processing and Editing System of the National Health Interview Survey: The Old and New

Susan S. Jack, National Center for Health Statistics

Abstract

The National Health Interview Survey (NHIS) has been fielded continuously since 1958. The procedures for processing, editing, producing, and documenting clean data with precise documentation have evolved over time. The data "products" have also changed somewhat over time, consistent with evolving technology.

The forthcoming redesigned NHIS, which has been converted to Computer-assisted Personal Interviewing format using the CASES authoring system, is a natural outcome of this evolving technology. The concurrent challenge is to redesign a processing and editing system that retains the "spirit" and the positive aspects of the old system while benefiting from technologic advances, minimizing the amount of labor intensive editing, and producing new forms of documentation appropriate to the NHIS's widely varying audience.

This presentation will give an overview of the current data processing/editing procedures, which reflect a paper-and-pencil NHIS, potential human error in responding, recording, and keying of data using a mainframe system, according to detailed specifications prepared by subject matter and editing experts. It will also describe the process NHIS is using to design a new multi-dimensional system from raw data to clean, documented data release. Using personnel with

Abstract (Cont'd)

a wide variety of experience and expertise, interrelated work groups will set policies and priorities which should ultimately

- ❐ reduce the amount of time involved in edit specification and programming,
- ❐ allow for minor modifications as well as major changes involving cyclical modules,
- ❐ automate and simplify the documentation process, and
- ❐ decrease the turn-around time for clean data release.

The goal is to synthesize the best of the old NHIS policies with the experience of others undergoing the CAPIization process to produce an essentially "generic" editing/ processing system which can be used by other large complex surveys, particularly those within NCHS. ■