Chapter 10

Record Linkage Methods Applied to Health Related Administrative Data Sets Containing Racial and Ethnic Descriptors

Selma C. Kunitz, Clara Lee, and Rene C. Kozloff, Kunitz and Associates, Inc.
Harvey Schwartz, Agency for Health Care Policy and Research

Abstract

In response to the lack of easily retrievable clinical data to address health services and medical effectiveness questions, especially as they relate to racial/ethnic minorities, the Center for Information Technology (CIT), Agency for Health Care Policy and Research (AHCPR) recently sponsored a project on record linkage methodology applied to automated medical administrative datasets containing racial and ethnic identifiers (Contract 282-94-2005). The primary objectives of the project were to:

> *link patient-level related datasets that contain racial and ethnics descriptors;*
> *and assess the value of the linked data to address medical effectiveness research questions that focus on the quality, effectiveness, and outcomes from care for minority populations.*

KAI, AHCPR's contractor, received approval from the State of New York's Department of Health to utilize the Statewide Planning and Research Cooperative System (SPARCS) files Discharge Data Abstract (DDA) and Uniform Billing files (UBF), which contain all acute hospital discharge and claims data, the SPARCS Ambulatory Surgical files, and the Cardiac Surgery Reporting System (CSRS) files, a research dataset. KAI received files for the 1991, 1992, and 1993 time periods. The files were successfully linked by patient, "visits" across the time periods. While the linked data appear to be of high quality, the process of obtaining and linking the data is lengthy. Additionally, these administrative health care data sets contain millions of records that document all hospital stays and thus, identifying appropriate subpopulations for a particular research question is a time and resource-consuming effort.

While the administrative health care datasets may be useful in answering questions about charges, length of stay, and other health service issues, their current utility may be less useful in answering clinical questions for minority populations. These datasets can be used to explore potential associations among diagnoses, treatment, and outcome variables. However, understanding the mediating factors and the decision-making variables that result in patient care may not be possible. For example, the results of diagnostic tests such as angiograms are not generally recorded in these datasets, thus limiting the ability to carefully subgroup patients by disease severity. With consideration for the potential utility of these datasets, however, there are several recommendations that emanate from the study.

This talk will briefly describe the research questions posed, linkage process, findings, and recommendations for additional action and policy considerations.

Introduction

The ability to link automated health data records is of critical importance in our rapidly changing health care system. In a managed care and cost containment environment, researchers require reliable and valid data collected over time and across providers that describe patient characteristics and the location, process, cost, quality, and outcome of care to analyze which procedures are effective and produce satisfactory patient outcomes. Approaches and methods to linking records across time and providers are needed to provide information to policy makers, health plans, practitioners, consumers, and patients to make decisions about accessing, using, and paying for care, as well as the effectiveness of that care.

Background

In response to the lack of easily retrievable clinical data to answer medical effectiveness questions, especially as they relate to racial/ethnic minorities, the Agency for Health Care Policy and Research (AHCPR) sponsored a project on "Record Linkage Methodology Applied to Linking Automated Data Bases Containing Racial and Ethnic Identifiers to Medical Administrative Data Bases" under AHCPR contract number 282-94-2005 (Kunitz and Associates, Inc., 1996). This linkage demonstration project contributed to AHCPR's research goals by reviewing and adding to record linkage methodology; illustrating the value of this methodology; assessing the need for further development; and providing guiding principles to developers. The primary objectives of this record linkage methodology project were to: link two patient- level related data sets that contain racial and ethnic descriptors; and assess the value of the linked data to address medical effectiveness research questions that focus on the quality, effectiveness, and outcomes from care for minority populations.

Data Sets

AHCPR's contractor, KAI, a health research firm, identified data sets to use for assessing the value of linking administrative health related data bases to support medical effectiveness research in minority populations. KAI received approval from New York State's Department of Health (NYSDOH) to utilize the Statewide Planning and Research Cooperative System (SPARCS) files Discharge Data Abstract (DDA) and Uniform Billing files (UBF), the SPARCS Ambulatory Surgery files, and the Cardiac Surgery Reporting System (CSRS) files. KAI received files for 1991, 1992, and 1993. The selected systems and data files are briefly described as follows:

- **SPARCS (State Wide Planning and Research Cooperative System)** is a system maintained by the NYSDOH. The Discharge Data Abstract files (DDA) contain all acute hospital discharge data and the Uniform Billing Files (UBF) contain all acute hospital billing records. Data about surgeries performed at hospital-based ambulatory care centers and certified diagnostic and treatment free-standing centers are maintained in the Ambulatory Surgery files. The data are used for planning and research. Three of the files extracted from SPARCS for this project were the DDA, UBF, and the Ambulatory Surgery file.

  NYSDOH staff combined the acute hospital DDA and UBF data files by individual hospital stay for this project. Thus, we received both matched and unmatched records from the DDA and UBF for 1991 - 1993. Because the files were selected based on DDA variables, unmatched records are those that are in the DDA file but do not have a corresponding match in the UBF. A completeness level of 95% is typically achieved in SPARCS files, a figure that is supported by our research, as seen in Figure 1. Yet, those records which are unmatched may reflect not only missing UBF records, but also incorrect information which may have hindered the original matching process performed by the

NYSDOH.

Figure 1. -- Matching Rates Between DDA and UBF Records

Year	Total	Matched	Unmatched	
			N	%
1991	626,222	594,302	31,290	5%
1992	699,246	663,323	35,923	5%
1993	714,583	677,778	36,805	5%

KAI obtained 31,290 unmatched records out of a total of 626,222 for 1991, 35,923 unmatched out of a total of 699,246 for 1992, and 36,805 unmatched records out of 714,583 for 1993. The selection process did not enable KAI to receive data which was in the UBF file but absent in DDA. In addition, we received the Ambulatory Surgery files for these three years.

- **Cardiac Surgery Reporting System (CSRS)** is a voluntary reporting system of all in-hospital cardiac surgeries. It contains risk factors, clinical descriptors and procedure data and is used as a research data set. We received these files for 1991-1993.

Figure 2 summarizes the size of the original data files. The DDA/UBF files contain between 2.5 and 3 million records each year. The ambulatory surgery files were not segregated by year and contain slightly more than two million records. The CSRS data files are also summarized by year and contain a considerably smaller number of records because of the more narrow focus of the records on cardiac surgery.

Figure 2. -- Summary of Sizes of Complete Data Files

Data Set	Year(s)	Number of Records
SPARCS	1991	1,687,521
	1992	1,677,948
	1993	1,660,109
Ambulatory Surgery	1991-1993	2,121,542
Cardiac Surgery Reporting System (CSRS)	1991	19,783
	1992	21,592
	1993	22,491

Research Question

One of the primary goals of this project was to determine whether a medical effectiveness research

question could be successfully addressed by the linked data. The selected *research question* for this project relates risk factors, treatment and outcome of cardiovascular disease to minority status:

> *Are the racial/ethnic differences in mortality and morbidity from coronary heart disease related to racial/ethnic differences in treatment?*

The working *hypothesis* stated that minorities are less likely to receive surgical treatment for coronary artery disease and, therefore as a group, experience higher incidence of cardiovascular morbidity and mortality than the majority U.S. population. The cohort was to be extracted from the linked SPARCS and CSRS data sets. The linked data sets were to contain records for 3 years, 1991-1993. Males and females aged 45-75 who were assigned a diagnosis of ischemic heart disease (ICD-9 codes 410 - 414) were to be included.

Confidentiality Approach

One of the primary issues in acquiring the New York State files was data confidentiality. Technically, the problems of confidentiality of data are often addressed by suppressing, encrypting or compressing information. In these data sets primary identifiers such as name, address, and telephone number were removed or suppressed from the files and secondary identifiers such as Medical Record Number (MRN), Admission Number, and Physician License Numbers (PLNs) were encrypted consistently across files and years to aid the matching process. Typically, confidentiality restrictions hinder the matching of large data sets. Identifiers such as name, address, and medical record number are important in order to be confident that the correct linkages are being made. If only demographic data and broad geographic identifiers are available such as gender, race, age and zip code, then a large group of people may have the same characteristics with the result that their records inaccurately matched.

Cardiac Subset -- Identification and Issues

The original research plan specified the use of ICD-9 codes 410 - 414 to address the research question. The low yield on initial matches, however, indicated that we needed to expand these codes to obtain a more complete record match between the SPARCS and CSRS files. Therefore, for the linkage process, the codes were expanded to include: 390.xx - 459.xx - disease of the circulatory system; 212.7x - benign neoplasm of the heart; 745.xx - bulbus cordis anomalies and other cardiac anomalies; 861.0x - injury to the heart without open wound to thorax; 861.1x - injury to the heart with open wound to thorax; 901.xx - injury to thoracic aorta; and 996.0x - mechanical complication of cardiac device. Figure 3 summarizes the number of potential patients on the DDA/UBF files using the ischemic heart disease codes (ICD-9 410 - 414) and an expanded set of codes.

Figure 3. -- Universe of Patient Records In DDA/UBF

DDA/UBF File Year	Initial Universe of ICD-9 Codes - 410-414	Expanded Universe of ICD-9 Codes - 390-459
1991	170,779	626,222
1992	189,198	699,246
1993	190,497	714,583

Record Linkage

The linkage software used for this project was MatchWare Technology Incorporated's (MTI) *Automatch*, developed by MTI's founder, Matthew Jaro (Jaro, 1997). MTI was KAI's subcontractor and its linkage experts collaborated with KAI's clinical researchers in conducting this project.

Several steps were involved in the data preparation process prior to performing the record matching or linking process. Fields that are common to the files had to be identified and recoded, where necessary, for potential use in the linkage process. Common person and event fields included for all three data sets were MRN, sex, date of birth, patient county, hospital identification number, diagnosis, procedure code and date, and Physician License Number (PLN). Fields common to two of the three files included age, patient zip code and state, admit date, discharge date, and payor.

As an example of recoding needs, race codes on the CSRS files were converted to correspond to SPARCS codes as shown in Figure 4.

Figure 4. -- Race Code Conversions

Description	SPARCS Race	CSRS Race
Asian or Pacific Islander	1	8
Black	2	2
Hispanic	3	8
Native American	4	8
Other	5	8
White	6	1

Linkage Objective

The linkage objective was to build a longitudinal, comprehensive patient history that captured clinical encounters over time and across care settings. Thus, records for the same patient were linked in two ways: matches were performed within each of the three data sets; and matches were performed between the DDA/UBF files and CSRS and between the DDA/UBF and Ambulatory Surgery files.

Steps in Record Linkage

Steps in the linkage process included identifying duplicate records; running preliminary matches as an iterative process to determine which fields yielded the most appropriate matches; identifying appropriate cut-off weights; and running the final linkage.

Duplicate records were identified on each of the files with no file having duplicates that exceeded 1% of the records. Automatch's method for determining most effective variables and probability weights to match across files were evaluated in preliminary iterative match runs. The process was iterative and consisted of selecting key variables for each match strategy, producing preliminary matched pairs, examining matched pairs with marginal match weights, and revising the parameters to better discriminate between apparent true and false matches. For the final matches specific probabilities of agreement were determined based on the preliminary matches. The match cutoff weight was chosen so that the estimated absolute odds of a true match for record pairs with that match weight were 95:5; i.e., a confidence level of .95 of a true match.

Linkage Data Quality Analysis

The linkage results were reviewed for data reliability and validity. First, the same variables on linked and unlinked records were compared to assess internal consistency and reliability. Agreement was 99% or greater for all variables except for date of principal procedure (67%) and admission number (83%); MRN, zip code, county, and other procedure each exhibited an agreement rate of 93%. Principal procedure as well as other procedure differences may reflect differences in reimbursement categories that were changed on the UBF for payment advantages. Admission number and MRN are scrambled by computer and any clerical error such as a transposition of numbers in the original MRN yields an inconsistent scrambled MRN. Likewise, transposition of numbers in zip code and county can yield mismatches.

The DDA and UBF responses for linked and unlinked records were then compared for the same patient. The responses are fairly consistent across DDA and UBF subfiles and between linked and unlinked records with slight differences in reimburser and diagnoses, which could be a function of the research question reflected in the linked files.

The DDA variables were selected for matching and were compared for linked and unlinked patient records, because of their tendency to be more reliable in the clinical area. In the linked records, patients are older (age $\geq$ 65 - 71% versus 59% for unlinked records), most likely reflecting the research question which focuses on cardiac diagnoses. Racial characteristics are similar as are ethnicity and gender.

Linked and unlinked records for Ambulatory Surgery patients were also compared. Analysis showed a greater percentage of the linked records to have a higher proportion of angina as the primary diagnosis while in the unlinked files there was a higher proportion of arterial disease, perhaps reflecting procedures performed in ambulatory surgery, i.e., angiograms. There were more Medicare reimbursers in the linked records which is consistent with differences in age groups. Other fields show no differences. Linked records compared with unlinked records for the CSRS patients showed a greater proportion of persons over 65, most likely reflecting the diagnostic groups of research interest. There were no gender, race, or ethnicity differences in the linked and unlinked records, reflecting similar patient populations.

The general consistency between the DDA and UBF subgroups and the consistency between linked and unlinked records within each of the data sets demonstrate the reliability of the matching and indicates that the linked records generally reflect the file population.

Racial Subsets

Responses across racial subgroups for DDA variables were reviewed. As expected, more Blacks, Asians and other minorities are treated in the New York City area (over 70%) than other parts of the state. Payment also differs, with a higher proportion of Whites on Medicare (69% versus 46% for Blacks and Others and 40% for Asian Americans. A higher proportion of Blacks and other minorities have Medicaid as the primary reimburser (Blacks -- 26%, Whites -- 5%, Asian Americans -- 25%, Other -- 27%). Blacks have a higher proportion of diabetes (4% versus 1% for Whites, 2% for Asian Americans and Other) and hypertension diagnosis (5% versus 1% for Whites, and 2% for Asian Americans and Other), and a slightly lower proportion of myocardial infarctions (Whites -- 11%, Blacks -- 7%, Asian Americans -- 10%, Other -- 11%) as principal diagnosis. Responses for other variables for linked and unlinked records by racial and ethnic categories are consistent, indicating that the linked file is a representative subset of the larger file.

Research Subsets

The research subsets, defined as the original diagnoses categories, 410.xx - 414.xx, were examined next. Comparing the DDA and UBF records on the SPARCS data set for linked and unlinked records indicates that age is higher on the linked records (age $\geq$ 65 = 75%) than on the unlinked records ($\geq$ 65 = 59%), reflecting the cardiac procedure research question. Also reflecting the research question is the larger number

of patients on Medicare in the linked data set (74% versus 59% in the unlinked data set). Comparisons between linked and unlinked records in the ambulatory surgery research files indicates no significant differences between the two subsets.

A review of responses for racial and ethnic subgroups for the linked and unlinked subsets in the DDA research file indicates that in both Whites are significantly older (78% Whites in the linked subset and 66% Whites in the unlinked subset are 65 or older). In the other racial categories, however there is a larger proportion under 65 (Blacks -- 42%; Asian Americans -- 33%; and Other -- 56%) in both linked and unlinked subgroups. The age differences between White and minority racial subgroups are also reflected in the proportion of patients on Medicare. There do not appear to be other major differences between White and minority subgroups. These trends are also reflected in the differences between Hispanic and non-Hispanic subgroups.

Linked Data Sets and the Research Question

Preparing the data to answer the research question was a complex process despite the fact that record linkage had taken place. The primary reason for the complexity of the process is that the research question focuses on outcome while the linkage focused on diagnosis. The linkage focus on diagnosis appears logical because it is how patients are generally categorized for health services and clinical research. However, medical effectiveness questions often focus on outcomes and thus, within diagnoses, outcome is an important patient characteristic. The research question, while resulting in a complex subject identification procedure, was typical of many medical effectiveness questions. The amount of time, then, needed for progressing from a linked data set to analyses for outcomes research, is several months and should be built into the research planning process.

Data and Linkage Issues

Several issues related to health care data sets and application of linkage methodology were identified:

- **Purpose. --** The purpose of the primary data collection endeavor impacts on the quality of specific variables and on their utility for linkage and their relevance for addressing a medical effectiveness question. For example, primary diagnosis frequently differed between the DDA and UBF subfiles. The diagnoses in the DDA is driven by clinical practice while in the UBF it is driven by reimbursement. Variables such as age and date of birth, gender, county of residence, hospital identification number, MRN, admission date, and procedure date may not be consistent across billing and discharge administrative records as well as the research records for several reasons: accuracy is not important for billing, discharge, and some research; an individual's high anxiety state; and family members reporting information under stress. Further, discharge abstracts generally reflect clinical diagnoses more accurately, while billing data typically reflect charge justification.

- **Encryption. --** Encrypting the Medical Record number (MRN), admission numbers, and physician license numbers degrades the efficiency of the matching software. The matching software used in this study can take into account slight differences among identifiers such as transposition of characters and adjust the match for them. However, since the encryption process scrambles identifiers or assigns a sequential number to records, the software is not dealing with actual numeric identifiers, which may have typographical errors. Thus this feature of the software is not useful for electronically encrypted or created numbers. The degradation was demonstrated in the first matching pass between the ambulatory surgery file and the DDA/UBF file. The MRN in the DDA/UBF file is defined as ten characters and was encrypted as such. The MRN in the Ambulatory Surgery file is defined as seventeen characters in which the first ten characters actually contain the MRN and the last seven characters are spaces. When the initial match between the DDA/UBF file and the Ambulatory

Surgery file took place there were no matches. The resolution involved the recreation of the Ambulatory Surgery File using only the first ten characters of the MRN in the encryption process. If however, the MRNs had been provided without being encrypted, the software could have adjusted for the spaces at the end of the original MRN in the Ambulatory Surgery File.

- **Race and Ethnicity Codes. --** The race and ethnicity codes are not always accurate as demonstrated by all observations for a particular New York State hospital which contained a race code of 5 and ethnicity code of 2 for all patient records. Additionally, the state SPARCS programmer indicated that there were software problems for RACE and ETHNICITY for certain hospitals that affected accuracy.

- **Dependent Relationships Among Variables. --** Certain pairs of patient and provider variables are strongly dependent on each other. For example, MRN is frequently hospital-specific and physicians are generally associated with only a few hospitals, thus PLN (Physician License Number) and Hospital Identification Number are also strongly dependent as shown statistically by *chi square* and *uncertainty coefficient* tests. The *Automatch* software requires that only one member of each dependent pair is used as a match variable because of relative odds of a true match calculation. For example, if both date of birth and age were used in a matching process, the calculated match weight would overstate the relative odds of a true match by exactly the contribution of the second occurrence. While date of birth and age represent the same concept, hospital and physicians may be logically independent entities although statistically associated. The nature of association in health related records should be considered in the matching process and perhaps, a different statistical approach used for these data.

- **Matching Variables. --** A related issue is determining what variables provide the greatest yield during the blocking and matching procedures. Linking is generally dependent upon person identifiers such as name and address, and date of birth, as well as on procedure and diagnosis codes from health related records. Since name and address were omitted from the files used to preserve personal privacy, other variables assumed greater importance. The clinical research staff, experienced with clinical data, recommended the use of age and date of birth, gender, county of residence, hospital identification number, MRN, admission date, and procedure date. The researchers pointed out that procedure and diagnoses codes can vary between administrative and clinical data sets because of reimbursement interests and are more likely to be accurate in clinical files. Identification of the variables most appropriate for linking health related files is still an open research issue.

- **Type and Number of Variables Utilized for Linking. --** Personal identifiers such as name and address are frequently used in census and vital statistics linkage efforts. Since these variables are not present on the health files, other variables that appear in several files and have a high probability of accuracy must be identified. Some examples are hospital identification number, admission date, and zip code. Additionally, linkage software experts often argue for numerous variables upon which to link. We found that the health- related data sets were more frequently linked with fewer discrepancies in the matching records when fewer variables are used. Thus the percentage of "true" matches was higher with fewer variables or, conversely, the number of false positives was lower. However, the total number of matched records was fewer.

- **Experience from Other Applications. --** Experience and assumptions gathered from other applications of linkage methodology such as census data cannot necessarily be applied to health- related data. Thus, for health- related data, multidisciplinary teams of linkage software programmers and health researchers need to develop appropriate linkage algorithms and to identify variables pertinent for linking these files.

Findings

Despite time delays and other issues, the files were successfully linked and the data were used to address the above hypothesis that pertains to care among minority populations. General findings are as follows:

- **Data quality** in the administrative and research files generally appears high and the data are potentially useful for health services research.

- **Both the linkage process and the analytic phase** for large data sets are lengthy and resource consuming. The practicality of linking large health- related data sets needs to be balanced against the number of years the data will be useful. If data can be used to support research for three to five years, then the linkage overhead expense may be justifiable. Costs of linking large data sets, then need to be balanced against the potential benefits.

- **Linking is only the first step** when the data are to be used to address research questions. The linkage process identifies a set of unique indexes for each of the patient records in each of the linked files. Depending upon the focus of the research question, it is necessary to carefully review the data files and the index files, which consumes both time and computer processing. Since the data files for large data sets must reside on mainframe computers, it also is a costly process.

  In this project, in which those subjects with the same diagnoses who received cardiac surgery are compared to those who did not, patients with relevant diagnoses had to be identified to form a subgroup from the SPARCS DDA/UBF files. The subgroup had to then be identified on the index files, determined whether linked or not linked to the CSRS file, and then found on the CSRS files. These steps precede any analytic procedures and represent the complexity of data management procedures that are associated with the analysis of the linked files.

- **Utility of administrative data sets** in answering medical effectiveness questions is variable. Clearly, identifying diagnoses, treatment, and outcome at a general level is possible and meaningful. The data set can be used to explore potential associations among diagnoses, treatment, and outcome variables. However, understanding the mediating factors and decision making variables that result in a patient proceeding to surgery or not may not be possible. For example, the results of an angiogram for a patient with ischemic heart disease are not recorded in SPARCS DDA/UBF or in CSRS. Thus, understanding why some patients who have angiograms proceed to surgery and others do not

is not possible.

Recommendations

This project yielded the following recommendations:

- **Utilize Linking Techniques for Projects With a Three- to Five-Year Life. --** Because of the time, labor, and financial costs of linking large data sets, it would appear practical to utilize linking techniques for data that can be analyzed over a period of three to five years.

- **Continue Methods Research. --** Issues in data dependence and optimal variables for use in linking health related data sets should be addressed in additional research projects.

- **Multidisciplinary Teams. --** The need for utilizing multidisciplinary teams composed of health researchers, programmers, and linkage experts was demonstrated in the linkage process.

- **Linking Prior to Research Use. --** Future efforts may enlarge record linkage before data are released from the agency that holds authority for the data to avoid degradation of data from scrambling or encryption. Linking prior to release across agencies raises issues of data sharing, protection of privacy, and other operational issues that must be addressed.

- **Recognize Time Needed for Research. --** Research efforts using linked data sets must allocate sufficient time and manpower resources to identify and extract the suitable subpopulation for a specific research question.

Selected References

Kunitz and Associates, Inc. (1996). Record Linkage Methodology Applied to Linking Automated Data Bases Containing Racial and Ethnic Identifiers to Medical Administrative Data Bases. Unpublished Final Report.

Jaro, Matt (1997). MatchWare Product Overview, *Record Linkage Techniques – 1997*, Washington, DC: National Academy Press.

Schwartz, H.; Kunitz, S.; Jaro, M.; Therlault, G.; and Kozloff, R. (1996). Studying Treatment Variation among Minority Populations via Linked Administrative and Clinical Data Sets, *Proceedings of the Section on Social Statistics, American Statistical Association.*