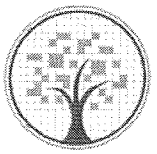
MAINTAIN AND EXPAND
THE HEALTHCARE COST AND UTILIZATION PROJECT (HCUP)
Contract No. 290-00-0004



H·CUP

HEALTHCARE COST AND UTILIZATION PROJECT

FINAL REPORT FOR LINKAGE VARIABLE PILOT STUDY
DELIVERABLE #336.1

November 18, 2005

Submitted to:
Jenny Schnaier, Project Officer
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

Submitted by:
Medstat
5425 Hollister Ave.
Suite 140
Santa Barbara, CA 93111

TABLE OF CONTENTS

EXECUTIVE SUMMARY I

INTRODUCTION 1

Purpose 1

Overview of potential methods 2

CANDIDATE STUDY DESIGNS 2

Option 1: Conduct probabilistic linking on SID/SASD/SEDD data for several states. 3

Option 2: Develop model using external (non-HCUP) data. 5

Option 3: Develop model using existing SID/SASD/SEDD data for one or more states. 7

Option 4: Use data from a state with gold standard. 10

PROPOSED STUDY DESIGN 12

Summary of alternative designs 12

Proposed design candidate 13

DISCUSSION 14

Implications for data analyses 14

Implications for data acquisition 14

Implications for Partner workgroups 14

REFERENCE 15

INDEX OF TABLES

Table 1. Comparison of two hypothetical records 3

Table 2. Possible outcomes for record pairs compared to a gold standard 7

Table 3. Expected results for potential linking variables 8

Table 4. Preliminary results for sample inpatient data – Linkage accuracy for FEMALE+DOB+ZIP composite identifier 9

Table 5. Preliminary results for sample inpatient data – Comparison of linkage accuracy for potential identifiers 9

Table 6. Potential Linkage Elements Collected by Selected HCUP Partner Organizations 10

Table 7. Estimated staffing levels for option 3 13

EXECUTIVE SUMMARY

The ability to link observations associated with the same individual is essential to accurate analysis of health care data over time and across settings. This document outlines four options for examining the creation and analysis of alternate methods for linking observations in HCUP databases. These approaches include

- Probabilistic linking
- Development of a model based on external data
- Analysis of existing HCUP data, and
- Development of routines for use with HCUP source data.

Preliminary results are presented for option 1 (probabilistic linking) and option 3 (analysis of existing HCUP data), with the most encouraging results obtained for option 3.

The proposed study would treat the HCUP person identifier (PNUM) as a "gold standard" and use a variety of demographic and other person-level data elements to "predict" whether two record pairs represent the same person (i.e., predict whether two records share the same PNUM). Although individual data elements are not expected to be accurate, combinations of these variables have the potential to accurately re-identify individuals for subsequent analysis. Estimated resource requirements are included for each option, with more detailed estimates included for the proposed study.

INTRODUCTION

The purpose of this document is to describe and propose a pilot study that will test the efficacy of various data elements for identifying patients across databases and over time. These data elements could be used alone or in combination to examine health service issues such as readmission patterns, costs and quality of care across settings, and classification of individuals that are members of populations of interest. Specifically, this document addresses an assortment of methods for linking person-level records in order to ascertain which records represent the same individual.

Our primary goal for the pilot study is to develop an understanding of the methods and issues associated with conducting an extensive analysis of record linkage methods across states, databases, and years. In the process of designing and conducting the pilot study, we expect to shed light on which methods are likely to yield useful research-related results and to provide preliminary results for certain types of analyses.

Purpose

There are several reasons for undertaking this study; chief among them is to explore the feasibility of using a combination of demographic elements as a surrogate for the HCUP person identifier, especially for states that do not provide Person Number (PNUM). For the 2001 data year, PNUM is available for fewer than half the states (15 out of 32 SID states). Availability of a person-identifier is useful for health service research, and the absence of such an identifier may prohibit or limit the types of analyses that can be conducted. Although other potential identifiers (e.g., date of birth, gender) are available, any given demographic data element is typically not sufficient to identify an individual. Presumably, it would be possible to combine two or more demographic data elements to create a composite identifier. For example, DOB, ZIP Code, and gender (FEMALE) might be used to create a composite identifier. While this approach is trivial from a data processing standpoint, it is complicated by data submission patterns across states, databases, and over time. Even though most facilities collect similar demographic data (i.e., elements found on a UB-92), there are marked differences between states in terms of what information is provided and the consistency of that information. These patterns are outlined below and listed in Appendix A.

Data Availability and Consistency

HCUP data are affected by data availability issues in two significant ways: (a) missing values for certain observations, and (b) data elements that are included for some states/databases/data years but are not included for other states/databases/data years.

For example, it is not uncommon for the RACE data element to be missing or invalid for 20 percent (or more) of the observations in a given HCUP intramural database (Appendix A). While the presence of race/ethnicity information may assist in matching records, the absence of such information should not negatively impact the accuracy of record linkage. Other data elements appear to be missing for entire population segments. For instance, Florida collects (but does not supply) dates of birth (DOB) for patients who are more than 11 years of age. As a result, DOB appears as missing for approximately 90 percent of Florida observations.

Further complicating factors include the variability in composition of state databases from year to year, and states that submit more than one data type (e.g., inpatient and ambulatory surgery) may have different data standards across databases. Potentially useful data elements for creating patient level matches such as RACE and medical record number (MRN) are not consistently supplied by state data sources across databases and over time. Appendix A summarizes the data availability and completeness for all states and databases for the 2001 data year.

As a result of this source-specific variability, it may be necessary to develop linking methods for different combinations of states, years, and even databases. A composite identifier that relies on DOB would obviously not work when applied to Florida data. Likewise, it is possible to construct an "identifier" that is so generic (e.g., relies solely on FEMALE and ZIP) as to be useless in a research context.

The noted differences between HCUP data sources suggest that there may be several different approaches to evaluating linkage methods. Given our understanding of record linkage methods and our familiarity with HCUP data, we have developed four different designs that could be used to examine methods for linking observations with and across databases.

Overview of potential methods

Previous deliverables and memos have described general classes of linking techniques, and detailed descriptions of specific methods can be found in HCUP Deliverable #215, "Linkage Variable Memo – Final Report." Brief descriptions of these methods are excerpted below.

Match merging

This method represents the most basic form of linking, with several variations possible depending on the research design. Perhaps the most intuitive and common implementation consists of joining two data sets and matching only those records where there is agreement on a designated data element. In the context of working with HCUP data, a simple analysis would involve finding records that match on either a single data element (e.g., PNUM) or a set of data elements (e.g., FEMALE and ZIP and DOB). There are several practical considerations that limit the accuracy of this technique. First, match merging relies on the presence and accuracy of multiple data elements. An error or transposition in any one could result in an error, either by linking records that should not be linking or by missing links.

Deterministic linking

In match merging, records are compared using several data elements, and equal weight is assigned to each data element in the comparison process. In practice, an exact match on DOB is far less likely than an exact match on gender. As such, it is reasonable to assign greater importance to links that are unlikely to occur as a result of chance, and less importance to links that could occur by chance. This is typically done by constructing a scoring algorithm and assigning somewhat arbitrary weights to specific linkage combinations.

Probabilistic linking

Probabilistic linking uses population-based frequencies to determine how likely any given match might be. These probabilities can be calculated via a number of different methods but typically result in weights for a match on a given variable and also weights for a non-match on that variable. Although most implementations use population-based characteristics to estimate the various parameters, probabilities can be calculated a priori based on the characteristics of a known population.

CANDIDATE STUDY DESIGNS

In this section we outline the four candidate study designs and provide a description of each.

Option 1: Probabilistic linking. In this design, commercial record linkage software would be used to conduct probabilistic linking of records in one or more state databases. Probabilistic linking is widely used for comparing lists of persons from the same population when a unique identifier is not available.

Option 2: Develop model using external data. A second approach would entail the use of a non-HCUP dataset to develop a model (or set of models) for linking records given the presence of different linking variables. This model could then be applied to HCUP data, and the results would be compared to those obtained from the external data set.

Option 3: Develop model using existing SID/SASD/SEDD data for one or more states. This option would utilize existing HCUP data and would develop an exhaustive set of analyses using all possible linking variables, alone and in combination. In order to evaluate the accuracy of these analyses, PNUM would be used as the "gold standard."

Option 4: Use data from a state with gold standard. The final candidate design involves using source data from an HCUP Partner. Since some HCUP Partners collect information that is not included in the HCUP databases, access to this information could help refine linkage methods and/or provide a reliable gold standard.

Each design entails some use of HCUP data. Given results of PNUM analyses (HCUP Deliverable #333), it would also be useful to present results separately for newborns, maternity discharges, and "all other" populations.

Option 1: Conduct probabilistic linking on SID/SASD/SEDD data for several states.

One of the most obvious choices for this pilot study is to employ probabilistic record linkage techniques. Probabilistic linkage techniques are widely used in domains from health care to mass marketing. Briefly stated, these methods allow for two or more records to be linked, even if the there are only partial matches among the information in the records. Most probabilistic routines have sophisticated methods for dealing with partial matches on "unreliable" fields such as name and address. For example, relatively simple linkage methods (e.g., match merging) would not link the two records below because they do not match exactly on either name or address.

Table 1. Comparison of two hypothetical records

	List A	List B
Name	Jon Busch	John Bush
Address	120 Record Linkage Street	120 Recrd Link St.
ZIP	31428	31428

However, probabilistic routines (and sophisticated parsing routines associated with probabilistic linking) would likely take into account the subtle differences in spelling for name, and the apparent abbreviation of street name. Whereas match merging techniques ask "are these records the identical?" probabilistic routines ask "how likely is it these records are the same?" While records in the example above are not letter-for-letter exact matches, it would seem likely that these represent the same person.

Design

This design option entails using probabilistic linking methods to (a) search for observations associated with the same individual, and (b) link records from one data source to another (e.g., SID and SASD data for the same state and data year). A logical first step would be to create a state-specific master patient list based on the results of probabilistic linking. This listing could be accomplished with commercial off-the-shelf (COTS) record linkage software already available to Medstat/AHRQ staff. This software (LinkageWiz) allows for records to be linked using a number of potential linking variables, including user-defined values. The software also provides default weights for each data element and allows for user defined weights and thresholds.

The patient index would be constructed for a given state by using the software to scan for apparent "duplicates" based on available linkage variables. Observations that appear to represent the same person would be flagged and an arbitrary person ID would be assigned to any records that are linked. Depending on the number of links, some or all of the linkages could be reviewed for consistency and face validity. Since there is no gold standard, accuracy would be evaluated by raters.

Cost

The chief costs associated with this option are labor for conducting the probabilistic linking and the review phase. Software costs have been paid from existing project funds. Depending on the state, conducting the linking effort would be minimal, requiring between one and four person days, and review could take up to an additional four days.

Advantages

Although the underlying methods are complex, an advantage of this approach is the simplicity. As noted previously, the software included default weights and thresholds for most linking variables (e.g., ZIP, DOB). User-specified fields could be created. These fields would allow for linking using information in the race/ethnicity fields, and custom weights can be assigned to those fields. This approach allows for an almost unlimited combination of linkage variables and weights.

Disadvantages

The chief disadvantage is the inability to easily evaluate the accuracy of the linking routines. In the absence of a gold standard, the only ways to establish the validity of the process are to evaluate the results via clerical review or some type of post hoc analytic comparisons. Both are relatively time-consuming and based on analysis of incomplete data. While probabilistic population estimation (Banks and Pandiani, 2001) can be used to assess the population overlap between two data sets, this is an aggregate-level measure and does not directly assess the accuracy of any given linkage method.

A related concern is that it can be difficult to ascertain the optimal set of weights and thresholds for a given data set. Although the software provides default starting values, there is no straightforward mechanism for computing custom weights.

As noted under advantages, a disadvantage of this approach is that it allow for an almost unlimited combination of linkage variables and weights. That is, the linkage phase can quickly become a lengthy series of "tinkering" to determine the best weights. Although this process may be possible for relatively small datasets, it can translate to impossibly long cycles for large databases.

Preliminary Results

In order to evaluate the feasibility of using probabilistic linking for intramural HCUP databases, we used LinkageWiz on a sample dataset (2001 Nebraska SID). While the initial setup was relatively easy and the linkage was performed rather quickly (approximately 2-20 minutes, depending on settings), the results were heavily dependent on weights and thresholds. The initial runs found zero "duplicates" (the software's term for a match) because the default threshold was set too high for a certain data element. After some experimentation, we were able to modify the linking criteria and matched anywhere between 13 percent and 72 percent of the observations in the database. In a very few cases, the software identified records that did not match on medical record number (MRN) but matched on other key data elements. For those linkage results that appear most valid, inspection of detail indicates that the majority of the records are being linked based on identical values for FEMALE, ZIP, and DOB. In effect, the

software was conducting a match-merge on these three variables and was not exploiting the full capabilities of the software.

Option 2: Develop model using external (non-HCUP) data.

Another option is to develop a model using non-HCUP data in which data elements are used alone or in combination to predict matches. In order to proceed with this option, an external data set with fields similar to those in the HCUP databases would need to be identified. While there are a variety of vendors that license claims level databases, Medstat is most familiar with its own commercially available product, MarketScan®. Therefore, the balance of this option outlines a study design based on the MarketScan® Commercial Claims and Encounters (CC&E) Database. Where appropriate, considerations about non-Medstat products are included.

The MarketScan® CC&E Database, the largest of the MarketScan® Research Databases, reflect the combined healthcare service use of individuals covered by Medstat clients nationwide. To help these clients manage their healthcare, Medstat builds databases comprising the healthcare experience of the clients' covered populations, as well as information about the populations themselves and the providers that serve them. We apply sophisticated methods for standardizing, integrating, and enhancing the data— both when the individual client databases are constructed, and again when these databases are combined to form the MarketScan® Research Databases.

The CC&E Database contains the healthcare experience of several million individuals annually (more than 5 million in 2002). Healthcare services for these are provided under a variety of fee-for-service (FFS), fully- and partially-capitated health plans, including exclusive provider organizations, preferred provider organizations, point of service plans, indemnity plans, and health maintenance organizations.

The data that make up the CC&E and Medicare Databases are stored in several tables: the inpatient admissions table, the inpatient services table, the outpatient services table, the population table, the outpatient pharmaceutical claims table, the REDBOOK supplement table, the enrollment summary table, and the enrollment detail table. Additional details for the tables most suitable for use in the model development are listed below.

- **The Inpatient Admissions Table:** The Inpatient Admissions Table contains records that summarize information about a hospital admission. Medstat constructs this table after identifying all of the service records associated with an admission (e.g., the hospital claims, physician claims, surgeon claims, and claims from independent labs). Similar information (such as payments for professional services) is then summed across the claims. The admission record includes the principal procedure and diagnosis, Major Diagnostic Category (MDC), and Diagnosis Related Group (DRG). It also includes all diagnoses and procedures (as many as 14 each) found on the service records.
- **The Inpatient Services Table:** The Inpatient Services Table contains the individual claims that are summed to create the inpatient admission records. An admission identifier on both the Inpatient Admissions and the Inpatient Services Tables identifies the claims that make up each admission record.
- **The Enrollment Summary Table:** This table provides, for a subset of the population, a single record for each period of continuous enrollment for each enrollee, showing enrollment start and end dates and most prevalent demographic (e.g., age group, gender, region, etc.) and plan information.

Design

Using MarketScan® CC&E data elements, various linkage algorithms can be constructed, which a different number of data elements in each algorithm. The goal will be to develop an algorithm using one year of data and then 'test' each algorithm on another year of data to determine the number of matches. The algorithm that performs the best will then be applied to the HCUP data. Results from the HCUP data can then be compared those obtained in the non-HCUP data.

Cost and level of effort

To proceed with this option, Medstat would need to access the external data set. The costs associated with the licensing of commercial claims data sets are significant. To develop a sufficiently reliable model, several years of data should be examined. However, the price of commercial data sets are customarily based on a single year of data, with some cost savings for multi-year 'purchases'. In addition to the direct costs associated with the purchase of external data sets, considerable effort would be required to become familiar with the data. One might assume that there would be a learning curve associated with HCUP programmers becoming familiar with the nuances of an external data set. An alternative would be to employ non-HCUP programmers (e.g., programmers familiar with Medstat's MarketScan® data).

Advantages

Many external data sets include a rich collection of data elements. In addition to information about inpatient claims, the MarketScan® CC&E Database contains outpatient and pharmaceutical claims. An algorithm developed using the MarketScan® inpatient claims, could also be tested on outpatient claims. If the model was able to predict "true" matches in outpatient claims, the algorithm could also be applied against the HCUP SASD. The MarketScan® Databases also include a Medicare product and a Medicaid product. Additional validity testing could be carried out to determine whether the model holds for these 'special populations'.

Many external data sets include an existing gold standard. The MarketScan® CC&E Database, for example, has undergone rigorous quality control processes. Demographic data are verified against independent employer data to ensure that employee claims are correctly identified.

Disadvantages

The majority of claim-based datasets are developed using data for specific populations and usually pertain to certain insured groups. For example, there are external data sets for employed populations, Medicaid only populations and Medicare populations. However, unlike HCUP data, most external data sets do not contain information for various payers. Consequently, we may be unable to generalize a model created using a single external data set to the HCUP data. Medstat licenses commercial, Medicaid, and Medicare databases, however, these separate databases would require significant effort to create a single standardized data set.

As previously stated, access to external data may be time-consuming and/or expensive. The licensing fees associated with a single year of a data or single year of commercial data set (commercial product) ranges between $75,000 and $100,000. Licensing fees for multiple years of data would be similarly expensive. There would also be contracting costs associated with accessing external data. Other data sources are likely to have similar costs.

The availability of data elements to 'end-users' of external data sets may not be sufficient for the development of a high quality model. Most data vendors have agreements that required the re-coding of certain person-level data elements. For example, in the Medstat MarketScan® CC&E data, only the first three digits of the ZIP Code are available to end-users. Race/ethnicity is another variable that may not be consistently available across data sets.

Option 3: Develop model using existing SID/SASD/SEDD data for one or more states.

Under this scenario, we would employ a variation of match-merging/deterministic techniques and analyze existing intramural HCUP databases for state(s) with relatively valid PNUMs. In these analyses PNUM would serve as a gold standard, and we would develop a set of models that use person-level data elements to construct surrogate identifiers. The accuracy of the surrogate identifiers would be compared to the linkage accuracy obtained by matching records based on PNUM.

Design

A necessary first step in this design is to identify at least one state with a "good" PNUM and sufficient availability of demographic and other person-level data. While this may be difficult to quantify, results included in previous deliverables (HCUP Deliverable #333) should aid in this task. Possible data sources include SID/SASD/SEDD data for HCUP states, as well as the planned analytic dataset to be constructed for the study of ED visits, revisits, and inpatient referrals.

Once we have identified a data source, using PNUM as a gold standard, we could identify true matches. Recognizing that PNUM is not 100 percent accurate, it does represent the single most accurate identifier available. We would then conduct a series of analyses using demographic and other potentially identifying elements. The most basic design would entail joining each observation in a data set to every other observation in that data set, and comparing the record pairs as to whether they (a) match on PNUM, and (b) whether they match on the demographic variables. Thus, for each comparison we would know if they match on PNUM and if they match on each demographic data element. This results in four possible outcomes for a given pair of records:

- **The pair matches on PNUM and matches on the demographic data element**. These are true positives.
- **The pair matches on PNUM but does not match on the demographic data element**. These are false negatives.
- **The pair does not match on PNUM but does match on the demographic data element**. These are false positives.
- **The pair does not match on PNUM and does not match on the demographic data element**. These are true negatives.

Table 2. Possible outcomes for record pairs compared to a gold standard

		Gold Standard		Total
		Link	**Non-link**	
Linkage Decision	**Link**	True Positive	False Positive	$N_{1.}$
	Non-link	False Negative	True Negatives	$N_{2.}$
	Total	$N_{.1}$	$N_{.2}$	$N_{..}$

In addition to the rates noted in the table above, we can calculate the following additional measures:

- **Match sensitivity** = True positives / $N_{.1}$
- **Match specificity** = True negatives/ $N_{.2}$
- **Positive predictive value (PPV)** = True positives/ $N_{1.}$

- **Negative predictive value (NPV)** = True negatives/ N_2..

This process would be repeated independently for each data element and for each combination of data elements. Each data element could be evaluated in terms of their individual accuracy and accuracy when combined with other data elements.

Table 3. Expected results for potential linking variables

	Measure					
Identifier	**False positive**	**False negative**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
FEMALE+DOB+ZIP	LOW	LOW	MID	MID	MID	MID
DOB+ZIP	LOW	LOW	MID	MID	MID	MID
FEMALE+DOB	MID	MID	MID	MID	MID	MID
FEMALE+ZIP	MID	MID	MID	MID	LOW	MID
DOB	HIGH	HIGH	LOW	LOW	LOW	MID
ZIP	HIGH	HIGH	LOW	LOW	LOW	MID
FEMALE	HIGH	HIGH	LOW	LOW	LOW	MID

Cost and resources

Depending on the nature and number of data sources selected, this study design would be relatively inexpensive. There would be no data acquisition costs and project staff are familiar with the data. Assuming no more than three states are selected and the states have a moderate number of observations, programming time should require no more than 10 days of a mid-level programmer. It is worth noting that the resources required for these analyses are extremely dependent on the number of observations in the dataset. Since each observation in the dataset is compared every other observation, a relatively small data set such as the Nebraska SID can result in several billion comparisons. Although not unfeasible given current capabilities, it might be qualitatively different to process a database with four million records, since that would translate into approximately eight trillion comparisons.

Advantages

An apparent advantage of this approach is the convenience associated with access to data with well known and documented limitations. It would require relatively little effort to start processing data, and analyses could be completed relatively quickly. Furthermore, however limited they might be, analyses should be relatively straightforward and easy to interpret.

This approach would also provide researchers with a tool for estimating the accuracy of relatively simple linkage methods they might use in the course of basic and applied research. For example, a researcher might be interested in linking records using FEMALE, ZIP, and DOB. Results from this analysis could be used to estimate the proportion of accurate links.

Disadvantages

There are at least two obvious disadvantages of this approach. The first is the lack of a true gold standard that would allow for measurement of the relative accuracy of the linkage variables. As noted in relation to option 1, probabilistic population estimates (PPE) can be used to gauge the overall reasonableness of the results but not the accuracy of specific linkages. While the

analyses might be methodologically thorough, the results are likely to be somewhat inconclusive. While this drawback can be mitigated by the careful selection of data sources, it is an intrinsic feature of the approach.

A second disadvantage of this approach is that it is limited to using existing data, and there will almost certainly be instances where this approach cannot be usefully applied. For example, constructing a composite identifier that depends on DOB will be of little use in working with Florida data, since DOB is missing for a significant proportion of the observations in Florida HCUP databases.

Preliminary Results

Although preliminary, we were able to conduct analyses on a subset of inpatient data. Results are presented below based on 1,000 observations from the 2001 Nebraska SID.

Table 4. Preliminary results for sample inpatient data – Linkage accuracy for FEMALE+DOB+ZIP composite identifier

	Gold Standard (PNUM)		
Linkage Decision	**Match**	**Non-Match**	**Total**
Match	213	3	216
Non-Match	15	499,269	499,284
Total	228	499,272	499,500

Table 5. Preliminary results for sample inpatient data – Comparison of linkage accuracy for potential identifiers

	Measure					
Identifier	**False positive (percent)**	**False negative (percent)**	**Sensitivity (percent)**	**Specificity (percent)**	**PPV (percent)**	**NPV (percent)**
FEMALE+DOB+ZIP	1.39	0.00	93.42	100	98.61	100
DOB+ZIP	2.29	0.00	93.42	100	97.71	100
FEMALE+DOB	7.69	0.00	100	100	92.31	100
FEMALE+ZIP	99.57	0.00	93.42	90.21	0.43	100
DOB	13.96	0.00	100	99.99	86.04	100
ZIP	99.76	0.00	93.42	82.18	0.24	100
FEMALE	99.92	0.00	100	45.98	0.08	100

As expected, each demographic data element alone is insufficient for use as an identifier, with the possible exception of DOB, which results in a false match approximately 14 percent of the time. However, various combinations of these data elements can yield relatively reliable pseudo-identifiers. Most notably, the combination of FEMALE+DOB+ZIP yields a false positive rate of 1.39 percent on the sample data. Put another way, this composite identifier would appear to be about 98 percent as accurate as PNUM. Although far from conclusive, these results hold some promise regarding the ability to create composite identifiers from demographic variables. While encouraging, we would expect the accuracy to diminish as the population size increases.

Option 4: Use data from a state with gold standard.

A noted drawback of Option 3 is the reliance on existing HCUP data sources and their associated limitations. Although convenient, this does not take full advantage of the person- level information collected by most HCUP Partners. This relatively ambitious option would use source data from an HCUP Partner with a true "gold standard" identifier to (a) determine which combination of data elements could produce the best identifier, and (b) which method(s) produce the best identifier. Although not included in the pilot, follow-on work could examine methods for using these linkage variables to create a state-specific patient index. Such an index could utilize an arbitrary person identifier, and should be allowable under HIPAA's research provisions.

Design

The final study design candidate would involve working with existing source data from one or more HCUP Partners. Although selection criteria vary, the most important criterion is that the HCUP Partner has a "gold standard" person-identifier. The analysis effort would provide states with software similar to that developed in Option 3, whereby different linkage variables would be constructed from source data (i.e., data collected by HCUP Partners but not necessarily submitted for HCUP processing) and evaluated in terms of reliability and precision. Rather than evaluate matches on demographic data elements against matches on PNUM, results of various analyses would be compared to the gold standard.

Several states collect (but do not submit) potential identifiers including:

- Direct Person Identifier (MRN or equivalent)
- Last Name
- First Name
- Middle Initial
- Patient Sex
- DOB
- Patient Marital Status
- Patient Address
- City
- State
- ZIP Code

These data elements are collected on the standard UB-92 and HCFA 1500 and are collected by most states with a data collection mandate. The table below summarizes some of the identifiers collected for selected states, as well as the reasons this information is not released.

Table 6. Potential Linkage Elements Collected by Selected HCUP Partner Organizations

State	SSN	Other Person ID	Name	Street Address	Comments	Reasons for Non-Release
California	Collected	None	None	None	OSHPD releases Record Linkage Number, which is an encrypted SSN	OSHPD may not release data in a form that will identify an individual (Civil Code Section 1798.24-1798.24b)
Nebraska	Collected	unknown	First name,	Patient City	Prior to April 2004, Nebraska released	Person identifiable data is no longer

State	SSN	Other Person ID	Name	Street Address	Comments	Reasons for Non-Release
			Last name		an "encrypted common patient ID" that was created using a 256 byte one-way hashing of the patient's first name, patient's last name, date of birth, and social security number	released by the Nebraska Hospital Association pursuant to their interpretation of the HIPAA privacy rule.
New York	Collected	Medical record number Policy number	First name, Last name	Patient Residence Address: - Address line 1 - Address line 2 - City - State - ZIP Code	SPARCS creates a Unique Personal Identifier and supplies this element to HCUP. This composite field is composed of portions of the patient's last name, first name, and social security number. This field, in conjunction with the Patient's Sex and the Patient's Birth Date is designed to provide probabilistic matching criteria for individual patient records for longitudinal analysis without compromising the confidentiality of the record.	SPARCS does not release data elements that pertain to a particular individual's facility stay, which, if disclosed, would constitute an unwarranted invasion of personal privacy, as stipulated in Department regulations, Title 10, NYCRR 400.18.
Utah	Collected	Medical record number	*Collected as available:* First name Last name Middle initial	Address is collected, but only ZIP Code is required	Utah Department of Health releases encrypted SSN and MRN to HCUP. There is no documentation of the Utah RODS dataset.	Undetermined. Non-release is presumably related to policies established by the Health Data Committee (HDC)

Provided a state collects and retains this information, it may be the case that they cannot or will not submit this level of detail, especially in a post-HIPAA environment. Our preferred approach would be to process data at Medstat or AHRQ, since this allows for maximum flexibility in processing and analysis. However, it is feasible that Medstat could provide software and/or technical assistance to a state with an appropriate data set. The state could then conduct the analyses and release summary-level information to Medstat and AHRQ.

At a minimum, analyses would include those described as part of option 3, whereby various data elements are used alone and in combination to "predict" whether observations will match the gold standard. Depending on the nature of the available data and resources, this could also be extended to include more sophisticated techniques such as probabilistic linking.

Cost

This is the most expensive option, in terms of both costs and resources. This proposal would entail gaining access to the most sensitive person-level data collected. This is also the most analytically-intense activity, and as such would require extensive development and data processing resources.

Advantages

There are several advantages to recommend this approach. Notably, this is the only option that would allow for precise and exhaustive analyses of person-level HCUP data relative to a gold standard identifier.

Such an effort would also foster future collaboration with state(s) willing to participate, and could lead to the development of a linkage variable/identifier that meets the need of researchers while protecting the privacy of individuals.

Disadvantages

The most prominent disadvantage of this approach is that it requires access to the most sensitive person-level data held by HCUP Partners. Obtaining access to this data would almost certainly require additional data use agreements, and would likely involve an innovative data processing/analysis arrangement, whereby analyses would be conducted by either Medstat or state staff using state resources.

PROPOSED STUDY DESIGN

Summary of alternative designs

Although each of the candidate design outlined above has some merit, we believe options 3 and 4 are the most promising, with a preference toward option 3. Option 1 would provide little advantage over more straightforward linking techniques. Option 2 would entail the relatively resource-intensive acquisition of third-party data, resulting in little overlap with HCUP data.

By contrast, option 3 offers a comparatively expedient method for evaluating a range of easily created identifiers. Option 4 extends this analysis to the state-level and allows for more extensive analyses. Although more ambitious, this may also be impractical given current privacy and confidentiality practices.

Preliminary analyses were conducted for options 1 and 3. With respect to option 1, these analyses suggest that while probabilistic linking may be useful with certain data sets, many of the benefits of this technique are lost on existing HCUP databases. Probabilistic methods appear to be best suited for analyses involving multiple possible linking variables, especially those where "approximate" matches (e.g., John vs. Jon) are likely to suggest links that otherwise might not be made.

By contrast, results of the preliminary analyses for options 3 are promising and merit further investigation. Although limited in scope and sample size, these preliminary findings appear to work well on sample data, and may scale well enough to be used in larger populations.

Proposed design candidate

Our recommended design is Option 3, whereby potential identifiers in existing HCUP databases would be analyzed relative to a pseudo-standard (PNUM). Although less ambitious than Option 4, this is more likely to produce useful results in the near-term.

Expected cost, level of effort, and timeline

Task 1 – Selection of states and preparation of state data. This task involves identifying states with PNUM and performing preliminary data screening for each state. At a minimum, we would create analytic data sets that do not include duplicate observations. We would also delete "suspicious" PNUMs (e.g., PNUMs that appear more than 20 times per year), since these might lead to spurious results. We estimate that this task would be completed within three weeks of the project start date, and would require approximately 20 labor days.

Task 2 – Design analyses. The basic approach to data analysis is outlined in the description of Option 3 (above). It may be possible to extend this design to include data elements such as MRN and RACE, despite missing values for some observations. It may also be possible to include variations of existing data elements such as DOB +/- 1 day and 3-digit ZIP Code. We estimate that this task would be completed within four weeks of the project start date, and would require approximately 20 labor days.

Task 3 – Draft report outline. For this task we would prepare an outline for the report that describes the purpose, method, and expected results of the pilot study. We estimate that this task would be completed within six weeks of the project start date, and would require approximately 6 labor days.

Task 4 – Data processing and analysis. This task is central to the overall activity, and as such represents the bulk of the labor and elapsed time. For this task we would analyze data from all states/databases that include PNUM for the 2001 data year. Summary results would be generated for each state/database and included as an attachment to the report. We estimate that this task would be completed within 20 weeks of the project start date, and would require approximately 50 labor days.

Task 5 – Draft report. This activity would result in a draft report which describes the purpose, methods, results, and implications of the linkage task. A project deliverable would be submitted to AHRQ with 22 weeks of the project start date. We estimate that this task would require approximately 25 labor days.

Task 6 – Final report. The task involves revising and resubmitting the pilot study report. A project deliverable would be submitted to AHRQ within 30 weeks of the project start date. We estimate this would require approximately 10 labor days.

Table 7. Estimated staffing levels for option 3

	Task 1 – Selection of states	Task 2 – Design analyses	Task 3 – Draft report outline	Task 4 – Data processing and analysis	Task 5 – Draft report	Task 6 – Final report
Programmer/Analyst	1 day per state/database	-	-	2 days per state/database	2	1
Senior Researcher	1	2	5	0.5 days per state/database	15	5
Statistician	-	3		2	2	1

Research Assistant	2		1	5	5	3

Proposed methods

Since our proposed design option relies on the availability of PNUM, our data sources for these analyses would consist of all states for which PNUM is available. Although now somewhat dated, our preference would be to use 2001 since the potential linkage issues are well-documented for that data year (see Appendix A). Alternatively, we would update the information in Appendix A to reflect 2001 and 2002 data, and analyze data for the 2002 data year.

Proposed analyses of source data

Analyses would be similar to those performed in the description of option 3, where each potential identifier is used to "predict" PNUM – both independently and in combination with other potential identifiers. The specific data elements will vary from state to state, but will typically include FEMALE, ZIP, and DOB. Where possible, we would extend this to include (a) variations of these data elements, and (b) provisions for analysis of data elements that may not always be available.

Results would be presented in a format similar to that shown in Tables 4 and 5, and would allow for a comparison of different "models" for each state/database. The principal dependent measures are the false positive rate and the false negative rate. In the context of the proposed study, desirable properties of any identifier are low values for both the false positive rate and false negative rate.

DISCUSSION

The proposed pilot study would allow for a comprehensive analysis of potential linking variables across a range of HCUP states and databases. While it may or may not yield an easily computed and reliable identifier, at the very least it can provide an estimate as to the accuracy of various linkage techniques.

Implications for data analyses

Should the proposed study yield a pseudo-identifier that is approximately as accurate as PNUM, this identifier could allow for certain analyses (e.g., readmission studies) to be performed for states that do not submit a PNUM data element.

Implications for data acquisition

One of the on-going goals of data acquisition is to encourage HCUP Partners to supply a HIPAA-compliant person identifier. At present, not all states provide this type of data element. Although the proposed study holds some promise for analysis of existing data, it does little to promote development of unique but HIPAA-compliant person identifiers. Accordingly, the proposed methodology neither adds nor subtracts to the level of effort required to obtain data elements from HCUP Partners.

Implications for Partner workgroups

While other design options would allow for the construction of unique person identifiers, this approach would also require additional efforts to obtain direct patient identifiers not currently supplied to HCUP. Alternatively, a Partner workgroup could be convened to study the best mechanism(s) for creating a direct patient identifier.

REFERENCE

Banks, S.M. and Pandiani, J.A. (2001) Probabilistic population estimation of the size and overlap of data sets based on date of birth. *Statistics in Medicine*; 20:1421–1430.