fhop

FAMILY HEALTH OUTCOMES PROJECT

UNIQUE IDENTIFIERS, DISCUSSION, RECOMMENDATIONS AND TESTING

INTRODUCTION

In 1991, a legislative appointed committee in California regarding Health Care for Women, Children, and Adolescents (AB99) recommended that the state adopt a set of data elements which would constitute a unique patient identifier for use by health and welfare programs. Unique identifiers allow information about individuals to be linked or shared across data bases for the purposes of data analysis, program planning, development and evaluation, client tracking and client eligibility and enrollment across programs. In response to this mandate, The Family Health Outcomes Project (FHOP) began investigating approaches to collecting data on unduplicated clients and longitudinal tracking in order to make a recommendation to the California DHHS/MCH branch for establishing a unique personal identifier and a common/minimum data set. This unique identifier will be used to establish an integrated information system to support policy analysis and program decisions within the Department, as well as be useful to providers, local and community agencies, and benefit consumers of health care services.

FHOP has developed criteria for selecting a unique identifier through a process that included a literature review, convening a group of experts to participate in the Unique ID Subcommittee, a survey of county MCAH directors and state program directors on uses of client tracking systems and preferences on approaches, a confidentiality and ethical literature review and participation on various state committees also studying unique ID such as the school linked data project (CIDC) and the California Health Information for Policy Project (CHIPP).

BACKGROUND

The purpose of a unique identifier is to allow information about individuals to be linked or shared across data bases for the purposes of:

- Surveillance and client tracking
- Identification of duplicate records
- Needs assessment, program planning, development and evaluation Client eligibility determination & enrollment across programs
- Monitoring health/status & outcomes

A Unique ID can be an assigned number such as the Social Security Number or a constructed number from a set of defined data elements such name, birthdate, etc.

Critical to the use of a Unique ID is maintaining and respecting a client's right to confidentiality and privacy. Unless informed consent has been obtained, no information should be shared in a form which identifies or links it to an individual.

FHOP developed the set of evaluation criteria for unique ID options in a consistent and logical manner. Key selection criteria include:

- **Universality**

A unique identifier should be available to anyone, *portable* (can be carried at all times), and *endure over time.* It must be relatively easy produce; for example, a plastic Id card is portable and relatively easy to reproduce - if it is not lost or stolen.

Demographic identifiers (e.g., name, birthdate mother's name) are also easy to produce **and are commonly used by clients.**

- **Invasiveness**

Unique identifiers must be sensitive to a client's privacy and not perceived as being intrusive when assigned or solicited.

- **Flexibility**

Software and hardware requirements should satisfy many computing environments (mainframe, local area network, PC), programming complexity for implementation and maintenance should be minimal and connecting capability (to large data sets or across multiple data sets) should be relatively easy.

- **Discriminative Capability**

A unique identifier should be *unique, accurate* and capable of *linking* client information across multiple data collection systems.

- **Confidentiality**

A clients right to privacy must not be threatened. A unique ID should *minimize the risk of breaching the confidentiality* of the client by agencies not given explicit permission to gain access to the record by the client.

- **Financial Feasibility**

A unique identifier should be economical to implement and maintain

The unique Id research identified five *major* unique ID models. The five models were then evaluated using the above criteria by a statewide advisory group convened by FHOP. In addition, this information was shared with the California Inter-agency Data Collaboration Project, a group of key decision makers from education, social service, mental health and data system agencies.

A description of the five models and a summary of how each option met the selection criteria follows:

1. Client/Master Index

The client index, developed by the California Data Systems Branch, is a computer generated number assigned to an individual as they enroll in publicly funded health programs in California.

The index is encoded and stored on a plastic Id card magnetic strip issued to program recipients and is, of course subject to being lost, forgotten, stolen sold, or counterfeited! Provisions for these eventualities (at least in California) are uncertain and duplicate, inaccurate records will be inevitable.

Built in a mainframe environment, the index will be expensive to maintain, change, and local access and cost are a concern.

Confidentiality may be a problem if the index is linked to other large databases on the State Mainframe.

2. Social Security Number

Social Security Number is assigned by the Newborn infants are now electronically issued birth) in almost all states. The SSN has been identifier for private industry and many state and

Social Security Administration. a SSN at birth (enumeration at used for many decades as an local programs.

Non- Taxpayers and undocumented newcomer populations have no incentive to obtain a SSN.

SSN performs poorly in terms of accuracy. There has been a relatively high (1 in 26 inaccuracy rate reported in the literature and as high as 7% in an internal study done at the San Francisco Health Department) of the same SSN numbers being assigned to different individuals and of individuals obtaining multiple Social Security Numbers. Numbers are easily made up by those evading tax payments and those who have lost their number.

The SSN is now widely used by financial and government institutions who are not bound by the same confidentiality and privacy standards as health agencies.

3. Biometric approaches (finger prints, retinal scans, etc.)

Biometric approaches to unique identification have been used by law enforcement agencies, licensing agencies and some local registries for many years. Biologic identifiers are virtually universal (e.g., almost everyone has a fingerprint) tend not to change over time and, are unparalleled in being able to uniquely identify an individual! The use of biometric identification in the forensic sciences has proven their reliability.

The cost of developing (or purchasing) and maintaining the software to connect optical scanners to existing databases requires a significant initial investment. Such an investment would only be justified when detection of fraud and duplication would result in significant cost savings

Obviously the largest concern with biometric identifiers is the perception of these procedures as being offensive and the association of the process with law enforcement, Immigration and IRS agencies. In fact, since law enforcement agencies are exempt from many of the confidentiality laws, the risk of breach of confidentiality/privacy rights is great.

4. Common Patient Identifier

The common select stable place of birth, patient identifier (CPI) is an alpha-numeric variable derived from demographic data elements such as: birth name, mother's name, birth date, gender.

Specified demographic variables do not require plastic cards, or memorization and are unlikely to change with time.

CPIs can function in any computing environment and have good discriminative capability.

5. Virtual Identifier

The virtual identifier or "black box" is a computerized probabilistic matching program that utilizes the same select demographic data elements as the CPI. A *temporary* unique identifier is constructed only for the time it takes for data linkages to occur. After the task is accomplished, the identifier is destroyed.

The virtual ID meets the selection criteria in the same way as the CPI except that since it is transient, the identifier as a data element does not exist after the

identifying or matching routine is finished and therefore this option provides the highest degree of security.

Each model was scored from 1 to 3 on each criteria. Table 1 presents the scores. (Highest possible score is 36 and some criteria have more than one component.)

TABLE 1 Five Unique ID Models Criteria Scores

IDENTIFIER	UNIVERSALITY (Portability, Durability	FLEXIBILITY	DISCRIMINATIVE CAPABILITY	CONFIDENTIALITY	FINANCIAL FEASIBILITY	NON INVASIVENESS	TOTALS
CLIENT INDEX	UNIVERSALITY PORTABILITY 2 DURABILITY 1 **SUBTOTAL 4**	LANGUAGE 1 SYSTEMS 1 **SUBTOTAL 2**	DISCRIMINATIVE 3 LINKAGE 2 **SUBTOTAL 5**	UNAUTH. ACCESS: WITHIN AGENCY 3 EXTERN. AGENCY 1 **SUBTOTAL 6**	HARDWARE MAINT. AND SUPPORT 1 SOFTWARE MAINT. AND SUPPORT 1 **SUBTOTAL 2**	PERCEPT. OF NON INVASIVENESS 3 **SUBTOTAL 3**	**22**
BIOMETRIC SCANNING	UNIVERSALITY PORTABILITY 3 DURABILITY 3 **SUBTOTAL 9**	LANGUAGE 3 SYSTEMS 3 **SUBTOTAL 6**	DISCRIMINATIVE LINKAGE 3 **SUBTOTAL 6**	UNAUTH. ACCESS: WITHIN AGENCY 3 EXTERN. AGENCY **SUBTOTAL 4**	HARDWARE MAINT. AND SUPPORT 1 SOFTWARE MAINT. AND SUPPORT 2 **SUBTOTAL 3**	PERCEPT. OF NON INVASIVENESS 1 **SUBTOTAL 1**	**29**
SOCIAL SECURITY NUMBER (SSN)	UNIVERSALITY PORTABILITY 2 DURABILITY 2 **SUBTOTAL 6**	LANGUAGE 3 SYSTEMS 3 **SUBTOTAL 6**	DISCRIMINATIVE LINKAGE 2 **SUBTOTAL 4**	UNAUTH. ACCESS: WITHIN AGENCY 2 EXTERN. AGENCY **SUBTOTAL 3**	HARDWARE MAINT. AND SUPPORT 3 SOFTWARE MAINT. AND SUPPORT 3 **SUBTOTAL**	PERCEPT. OF NON INVASIVENESS 2 **SUBTOTAL 2**	**27**
COMMON PATIENT IDENTIFIER (CPI)	UNIVERSALITY 3 PORTABILITY 3 DURABILITY 3 **SUBTOTAL 9**	LANGUAGE 3 SYSTEMS 3 **SUBTOTAL 6**	DISCRIMINATIVE 3 LINKAGE 3 **SUBTOTAL 6**	UNAUTH. ACCESS: WITHIN AGENCY 1 EXTERN. AGENCY 2 **SUBTOTAL 3**	HARDWARE MAINT. AND SUPPORT 3 SOFTWARE MAINT. AND SUPPORT 2 **SUBTOTAL 5**	PERCEPT. OF NON INVASIVENESS 3 **SUBTOTAL 3**	**32**
VIRTUAL IDENTIFIER	UNIVERSALITY 3 PORTABILITY 3 DURABILITY 3 **SUBTOTAL 9**	LANGUAGE 3 SYSTEMS 3 **SUBTOTAL 6**	DISCRIMINATIVE 3 LINKAGE 3 **SUBTOTAL 6**	UNAUTH. ACCESS: WITHIN AGENCY 3 EXTERN. AGENCY 3 **SUBTOTAL 6**	HARDWARE MAINT. AND SUPPORT 3 SOFTWARE MAINT. AND SUPPORT 2 **SUBTOTAL 5**	PERCEPT. OF NON INVASIVENESS 3 **SUBTOTAL 3**	**35**

Consensus was reached to recommend the use of either the Common Patient Identifier (CPI) or the Virtual Identifier. The CPI is useful at the local level for clinical case management and longitudinal tracking.

The Virtual Identifier is preferable for use at the State level where linkage of records for surveillance purposes does not require specific client identification.

Recommendations

In order to actualize this approach, FHOP recommended that a set of core data elements with standard definitions and an additional set of "confirmatory elements" be collected on every client by any California agency delivering services.

The advantage of this approach is that it provides the potential to construct a common patient identifier, a virtual identifier or the ability to simply use the data elements for matching and linkage. The criteria of universality, discriminative capability, confidentiality, and flexibility are best satisfied by this recommendation.

The following are the recommended minimum core identifying data elements selected after a review of literature on data linkage (See attached data definitions):

- **BIRTH NAME OF CLIENT**
- **DATE OF BIRTH** (month, day, year)
- **GENDER**
- **MOTHER'S FIRST NAME**
- **PLACE OF BIRTH (County if in California, State if out of state, Country if out of USA)**

The following are the verifying or confirmatory data elements that should also be collected for situations where unduplicated records are not successfully identified with the core elements or where a database does not contain one or more of the core data elements (See attached data definitions):

- **CLIENT'S COUNTY OF RESIDENCE**
- **CLIENT ALIASES (SUCH AS NICKNAME)**
- **FATHER'S NAME**
- **MOTHER'S MAIDEN NAME**
- **OTHER CLIENT NUMBERS***
- **SOCIAL SECURITY NUMBER* ZIP OF CLIENT'S RESIDENCE**

*Use of the Social Security Number and other client numbers allows continuity with existing systems and supports the California client index system.

TESTING

FHOP has been testing the core identifying elements as to:

- How *uniquely* they identify a person or have *discriminative validity*
- How well they discriminate one person from another *(reduce duplication)*
- How successfully they link other databases (operational validity)
- How easily they an be implemented in a clinical setting

The testing of the Core Identifying Elements occurred in two phases. Phase One evaluated discriminative validity. Phase Two evaluated the core elements' success as linkers. Phase Three will include testing in a clinical environment when the Common Application Transaction System is piloted in California.

PHASE ONE --- tested discriminative validity and the ability to reduce duplication

Data Set: 1992 California Automated Birth Certificate
# of Cases: 602, 269 records

PROCEDURE

An object-oriented approach using the core data elements along with a matching algorithm was employed. Using this approach, the first step was to determine the relative weights of each variable (data element). This was done through a technique called blocking (well described in the data linkage literature) where each variable is tested individually to determine its specificity or weight in limiting the size of the database and increasing accuracy.

An algorithm was then built using the core elements and their weights to determine a probabilistic match. Once the variables and sequence were selected, a string variable could be constructed. This alpha-numeric string can act as a virtual unique ID or can be recoded as a number to form a Common Patient Identifier. See Table 2.

This object-oriented approach adds each data element according to a specific algorithm or weighting, and thereby progressively reduces the number of duplicate records.

Table 2: PRELIMINARY RESULTS of the PHASE ONE TEST

KEY	Total Records	Number of Unique Records	p(Unduplicate)**
All	602269	600463	0.99843
Full Name (First, Last)	602269	574492	0.95388
Partial full name (FN_3/LN_3)*	602269	562850	0.93455
Last Name	602269	448504	0.74469
First Name	602269	330935	0.54948
Mother's First Name	602269	288757	0.47945
DOB	602269	365	0.00061
Place of Birth	602269	56	0.00001
Gender	602269	2	0.000004
Matching String: (Gender/DOB/POB/FN3/LN_3)*	602269	601323	0.99805

Data Source: California Automated Birth Certificate Database (1992 Data)
Sample Size: 602269

*** Definitions:** FN _ 3 = Partial 1st Name (3 Characters), LN_3 = Partial Last Name (3 Characters),MN_3 = Partial Mother's First Name (3 Characters)
**p(Unduplicate) = Unduplicated Count/ #Total (Probability of Unique Key)

The additive weights or the relative value of each of the core elements in establishing a unique identity was also determined do that if an element was missing from a data set, a substitution or combination of elements could be used.

The blocking order used was to first block by year of birth, in this case 1992, followed by gender, place, place of birth, and month of birth, etc.

Results

- The combination of all core identifying elements had a 99.843% successful unduplication rate.

  The weighted order of the core identifying elements was established.

- There is not a significant difference between using a full name or partial name to determine unduplicated counts.

  A composite string variable also had a 99.805% successful unduplication rate.

PHASE TWO

Tested the core elements in a second large data set:

1992 Newborn Screening Records
600.071 records

Evaluated the linking ability of the core elements

RESULTS

- Using the same algorithm on the Newborn Screening Database, an unduplicated rate of 99.8 % resulted.
- However, the records 46.88% of the newborn screening records were missing first names, or had dummy first names such as baby girl or newborn boy.
- The first attempt at linking the two data sets where there was a high rate of missing first names yielded only a 95% match.
- If we replaced the first name data element with the first 3 digits of the *Place of Birth ZIP code* (increasing the specificity of place of birth) a valid substitute was established to link the two databases and a 98.2% match rate resulted.

DISCUSSION

What the results illustrate is that a random use of core identifying elements as identifiers will not yield unduplicated counts. By using new technologies as this object-oriented approach, we can use the same familiar demographic elements and get an almost 100% match.

NEXT STEPS

As a result of this work, a number of California programs have already adopted the core elements: WIC, Adolescent Family Life Program and the Common Registration and Eligibility System.

The California Center for Health Statistics and the Primary Care and Family Health Division has recommended to the California Department of Health Services Executive Committee that this approach be adopted throughout the Health Department as well as the entire Health and Welfare Agency.

PHASE THREE TESTING

Phase three testing will field test the core identifying elements to determine their practical validity by assessing the difficulty in recording and obtaining the elements from clients in clinical settings. In order to increase the level of confidentiality of the identifying elements at the State level, the algorithms used in this study could be used to test creating a virtual identifier with an object-oriented approach as discussed above.

Standard Definitions of Data Elements

The following data definitions are recommended by the California Interagency Collaboration Project.

The approach to coding client names is to use a single convention with indicate the type of name**;** for example, birth name, alias, aka, legal.

The coding definition of mother's name includes first, middle and last because it is assumed that agencies would need the entire name. However, only the first name will be needed for this purpose.

Note: Place of birth has three definitions depending on the place of birth of the client.

The approach to coding client Identification numbers Is similar to client name with one coding definition and the use of qualifiers; for example, social security number, client ID number, etc.

The following recommended data definitions are presented In order of their listing in the text.