

Technical Issues in Creating a Unique Client Identifier Part 2

DIG Technical Assistance on
Implementing Unique Client Identifiers
NASMHPD Research Institute, Inc.

Notes

- This presentation is simultaneously too detailed for a brief overview and not detailed enough for practical implementation guidance.
- Some slides will be only briefly discussed during the presentation (but available for later reference).

Notes

- For implementation purposes, greater detail on many of the concepts can be found in the draft document available for review during this meeting.

DRAFT

USING STATE ADMINISTRATIVE RECORDS TO MANAGE SUBSTANCE ABUSE TREATMENT SYSTEM PERFORMANCE

FIELD REVIEW COPY

Not for distribution, copy, or dissemination without permission from the authors.



U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
Center for Substance Abuse Treatment
www.samhsa.gov

Purposes of Unique Client IDs - UCIDs

- Unique Client Counts.
- Also called Client De-Duplication.
- The ability to determine the total number of unique individuals receiving services (across all providers in a State) is a key component of NOMs and the annual MH and SAPT federal block grant applications.
- A well-designed client identification protocol (in association with the deterministic and probabilistic deduplication routines) can help the State obtain such unduplicated client counts.

Purposes of Unique Client IDs

- Database Linking.
- The ability to collect and utilize data elements (DOB, name components, SSN components, etc.) is central to the ability of a State MH or AOD treatment agency to link to external databases (arrests, emergency room, social services, etc.) for purposes of outcome evaluation and shared-client analyses.
- Such linking analyses often involve the generation (directly or indirectly during the linking process) of synthetic client IDs generated from the client identifier data elements common to the datasets being linked.

Purposes of Unique Client IDs

- While unique client counts is generally considered the more immediate reason for the development and use of unique client IDs, States should develop unique IDs (or should at least collect a core set of client identifiers) that could be used at a later date for purposes of database linking within the State.

Conceptual Criteria for Unique Identifiers

- The American Society for Testing and Materials (ASTM), a standards development organization accredited by the American National Standards Institute, published the Standard Guide for Properties of a Universal Healthcare Identifier (UHID) [1998].
- An abbreviated list of desired features of a Unique Identifier include:

Conceptual Criteria for Unique Identifiers

- Accessible - Available when required
- Assignable - Assign when needed by trusted authority after properly authenticated request
- Atomic - No sub-elements having embedded meaning
- Concise - As short as possible
- Content-Free - No dependence on possibly changing or unknown information
- Controllable - Only trusted authorities have access to linkages between encrypted and non-encrypted identifiers
- Discriminative – Capable of discriminating among unique individuals
- Identifiable - Possible to identify the person with such properties as name, birth date, gender, by associating these with the identifier

Conceptual Criteria for Unique Identifiers

- Linkable – Capable of linking client records across databases
- Longevity - Designed to function for foreseeable future with no known limitations
- Permanent - Never to be reassigned, even after a holder's death
- Public - Meant to be an open data item--person can reveal it
- Secure - Can encrypt and decrypt securely
- Unambiguous - Minimizes risk of misinterpretation such as confusing number zero with letter O
- Unique - Identifies one and only one individual
- Universal - Able to support every living person for the foreseeable future

Common Approaches to Client Identification

- Social Security Number – SSN
- Centrally-Assigned Client Identifiers
 - Master Patient Index - MPI
 - Master Client Index - MCI
- Constructed Client Identifiers
 - Based on Fixed Client Characteristics
- Assigned Identifiers
 - Provider-specific (non-centrally-assigned) client identifiers

Social Security Number - SSN

- Some States may use Social Security Number (SSN) [or encrypted SSN] to uniquely identify clients.
- Many other States collect SSN, but do not use SSN as the primary client identifier.
- The most discriminating client identifier would be a “universal” ID, such as a verified full nine-character social security number.
- Readily available
- But privacy concerns
- Not all persons have an SSN (undocumented)
- SSN easy to fabricate
- Often not verified
- Frequent data entry errors

Social Security Number - SSN

- Alternatives:
- Last 4 digits of SSN [SSN4]
- Encrypted SSN
- Addition of a check digit to help detect errors in data entry or in transmission

Social Security Number - SSN

- In situations where SSN is not available for use as an identifier (due to privacy concerns, State policy, client preference not to disclose, MH or AOD treatment agency collects such but partner data linking agency does not, etc.), States and providers often choose to generate:
 - 1) A Master Patient Index or
 - 2) A Constructed Identifier based on concatenated client data elements (such as date of birth, gender code, perhaps the last four characters of the SSN [SSN4], and certain characters from the client's first and last names).

Centrally-Assigned Client Identifiers

Master Patient Index - MPI

Master Client Index - MCI

- Other States may use centrally-assigned Statewide-unique identifiers such as a Master Client Index or Master Patient Index (MCI, MPI).

Centrally-Assigned Client Identifiers – MPI, MCI

- In a typical MPI protocol, when a client presents for services during any given service episode, the provider enters client identifying information (SSN, Name, DOB, Gender, etc) into a central State database server.
- Based on the identifying information, the server scans the records of all existing clients Statewide, to determine if the client has been served previously by any provider in the State (including the current provider).
- If so, the central server returns to the provider the existing “Master Client ID” number.
- If not, the central server assigns a new Master Client ID to the prospective consumer (often a sequential autonumber “next value - one up” protocol).

Centrally-Assigned Client Identifiers – MPI, MCI

- This Master Client ID is intended to follow the specified individual across all providers over time.
- Advantage – No embedded information

Centrally-Assigned Client Identifiers – MPI, MCI

- Other Approaches to MPI - MCI
- Some States generate MPI strings based on encrypted (“hashed”) identifiers such as SSNs.
- For example, popular shareware encryption programs (such as MD5) can hash a 9-character SSN such as:
- 999754321 into a 32-character encrypted string such as:
- 2de1227bdd070b7c34b7a7067c14a707
- Occasionally, a State may use the encrypted SSN as its MPI, but more often the SSN is encrypted for transmission and the encrypted SSN serves as the input to the Statewide unique autonumber MPI protocol.
- For consumers that do not have, or refuse to provide a SSN, such protocols often create a synthetic pseudo SSN, composed of components such as client’s DOB, gender code, etc

Centrally-Assigned Client Identifiers – MPI, MCI

- In some States, the MPI is unique only to the MH entity, or the AOD entity, or both.
- In other States, the MPI is used across many health, social, and human service agencies.
- MPIs are useful for generating unique client counts within the entity or entities that use a particular MPI protocol.
- In some States, an MPI is assigned to only a subset of all clients receiving services (e.g., only for those State-funded clients whose services were 100% paid or coordinated by a particular program or service agency or Managed Care entity, etc).
- MPIs are not useful for database linking to other datasets that do not use the same MPI protocol, unless both the “source” database and the “target” database also collect other client identifier data elements (SSN4, DOB, Name elements, etc) on which to link.

Constructed Client Identifiers

- Constructed Client IDs are created using relatively fixed or immutable personal characteristics.
- Also called Synthetic Client IDs.
- Many States and providers generate (or at least have the potential to generate) primary or secondary client identifiers created from concatenated component client data elements representing “fixed” client characteristics (or relatively fixed characteristics) that are unlikely to change over time.

Common Data Elements Used to Generate Constructed Client IDs

- Social Security Number (SSN), or at least the last four characters of the SSN (SSN4)
- Date of Birth (DOB)
- First Name (or at least some characters or phonetic encoding of First Name) [Example: NYSIIS phonetic code]
- Last Name (or at least some characters or phonetic encoding of Last Name)
- Gender
- Middle Name or Middle Initial
- Race-Ethnicity

Common Data Elements Used to Generate Constructed Client IDs

- Other identifiers to the extent that they exist and are common between datasets to be linked:
- Other fixed client characteristics
- Medicaid Number
- Criminal Justice ID
- Drivers License Number

Constructed Client IDs

- One common “constructed” identifier based on “fixed” client characteristics might be composed of the following data elements:
 - First and third characters of client’s first name [F1F3]
 - Middle initial [MI]
 - First and third characters of client’s last name [L1L3]
 - Full eight character date of birth [DOB – MMDDYYYY]
 - Gender [M,F 1,2] and
 - Last four characters of client’s SSN [SSN4]

Constructed Client IDs

- Thus:
- TONY DORSEY HUTCHISON
- DOB=December 3, 1971 [12031971]
- Gender=Male
- SSN=869-93-2927
- Would have a constructed client ID of:
- TNDHT12031971M2927.
- Such constructed client IDs can be stored as a designated database field (primary client ID) or can be generated as needed for client deduplication or linking projects (secondary client ID)

Other Data Elements Less Frequently Used to Generate Constructed Client IDs

- FIXED DATA ELEMENTS
 - Client's county of birth (FIPS code etc) [Federal Information Processing Standards]
 - Client's city of birth (FIPS code or alpha components, such as the first 3 chars)
 - Client's birth last name (or components) [useful for name change situations] *
 - Client's birth first name (or components) *
 - Mother's birth (pre-marital) last name (or components)
 - Mother's first name (or components)
 - Father's first name (or components)
 - Other phonetic coding of names (such as Soundex)
- * Name changes (due to marriage, divorce, adoption, personal choice, witness protection)

Other Data Elements Less Frequently Used to Generate Constructed Client IDs

- SUBJECT TO CHANGE OVER TIME
- [Not Recommended]
- Client's county of residence
- Client's city of residence
- Client's zip code of residence
- Client's phone number

Other Data Elements Less Frequently Used to Generate Constructed Client IDs

- Some States do collect the “less frequent” data elements listed above throughout all of the State’s health and social services agencies.
- In these States, such data elements may be useful as client ID components and linkable variables (depending upon the accuracy and completeness of such data elements).
- However, most States do not appear to collect these “less frequent” data elements routinely in their State datasets. As such, these data elements are not emphasized in this review.

Constructed Client IDs

- Easy for client to remember and clinician to generate
- But privacy concerns
- Embedded information
- Susceptible to fraud – third party knowledge of client could generate the client's ID and use for fraudulent purposes (more often an issue with financial benefit programs)

Constructed Client IDs

- Many constructed identifier protocols are possible.
- Each data element used in the creation of a client ID (or used in the creation of a temporary client ID during deduplication and linking analyses) has strengths and weaknesses in terms of:
 - the availability of each data element in the dataset(s),
 - confidentiality concerns,
 - the amount of missing data associated with each data element,
 - the quality of the collected data, the reliability and stability of each data element (e.g., people change their last name or use nicknames, people change their race-ethnicity self identify, people change their addresses).

Provider-Entity-Specific Client IDs

- Some States allow each provider to develop and submit provider-specific (non-centrally-assigned) client identifiers that contain little or no fixed client characteristic data elements.
- Also called Assigned Identifiers
- Such provider-specific client IDs are (at best) unique within the provider entity only, not State-wide unique

Provider-Entity-Specific Client IDs

- By chance, a particular client ID string or protocol may be used by two or more provider agencies within a State.
- Thus, some States use provider-specific client identifiers and concatenate the provider/entity code to address this issue.
- However, this approach still will not yield accurate State-wide unique client counts, since a particular individual can receive services from more than one provider (under two different provider-client ID strings) within any given period of time (fiscal year etc).

Provider-Entity-Specific Client IDs

- Even within a provider entity, a given client may have more than one provider-specific client ID (often due to lack of previous client look-up capabilities or change in provider database software over time which may have resulted in historical client IDs not being re-synchronized to the new client ID protocol)
- Thus, Statewide counts based on Provider ID – Client ID strings typically overestimate the number of unique individual receiving services within a State in a given period of time.

Constructed Client IDs

- The majority of the following discussion will focus on Client IDs generated from fixed (or relatively fixed) client characteristics.

Business Rules, Analysis Rules, and Data Cleansing Considerations for Fixed Client Characteristics IDs

- Name Fields. In health and social services databases, the same individual may have multiple service records with discrepant first name, last name, and middle name fields.
- Many name discrepancies are due to name changes and can be addressed by alias fields.
- State datasets often include multiple alias name fields for first names, nicknames, last names, pre-marital names, married names, criminal alias names, etc.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Alias fields are valuable for record matching. For example, females' last names often change with marriage or divorce. Names can also change through adoption or personal preference.
- Alias fields (in association with deterministic and probabilistic matching routines) also provide an alternative to re-synchronization of fixed characteristics client IDs whenever a name change occurs.
- Many State datasets include up to three to six alias fields for first name and for last name, all of which are considered by the matching protocol either during the data preparation stage or during the linking algorithm or both.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Other name discrepancies include:
- the unavoidable character-by-character typographical errors
- character transpositions
- dropped characters
- added characters
- nicknames (Bob vs. Robert)
- transpositions of first and middle names
- transpositions of middle and last names
- homonym names (e.g., Gene and Jean)
- embedded names (Jo Anne vs. Joanne)
- hyphenated names (Zeta vs. Zeta-Jones vs. Jones)
- names composed of two or more words (De La Rosa and Running Deer).

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Data deduplication and data linking software typically will assess whether name discrepancies across database records might be due to reasons such as the above and, if so, will consider such records as representing the same unique individual.
- Best practices: Capitalize all character fields, remove all special characters (apostrophe, dash, spaces). Can be done at input stage, but should be done at analysis stage.
- Store appellations (Jr., III, etc) in separate fields, but use in analysis stage.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Social Security Number. In health and social services databases, the same individual may have multiple service records with discrepant social security numbers (SSNs).
- Such discrepancies can occur due to:
 - simple character-by-character typographical errors,
 - the clinician heard the SSN wrong or wrote it wrong,
 - the client may be confused and not remember his or her exact SSN,
 - transcription errors (1s look like 7s, 3s look like 8s),
 - transpositions (92 vs. 29), and
 - provision of deliberately bogus SSNs by some clients (especially criminal justice clients).

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Many service agencies do not require that the client present a verifiable Social Security Number
- e.g., SSN card or payroll stub with SSN.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Block bogus SSNs at input stage – 123-45-6789, 111-11-1111, 999-99-9999
- Generate synthetic pseudo SSNs for clients who do not provide (or refuse to provide) SSNs composed of components such as client's DOB, gender code, etc – 072319631
- Add a SSN status field to describe if the SSN was verified, if the SSN is synthetic, or client self-report.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- For all these reasons, the same individual may have more than one putative SSN within and across databases and across providers, over time.
- During the data preparation stage, most data linking programs will attempt to identify all the potential discrepant SSNs that may in fact belong to the same individual.
- Data deduplication and data linking programs cannot determine which of the multiple SSNs for a given individual is the “correct” SSN, but can identify clusters of SSNs that may represent the same person and use such information to effect a record match that may otherwise be missed.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Date of Birth Fields. The date of birth (DOB) for a given individual can be discrepant within and across databases over time.
- In addition to the usual typographical and transcription errors
- The MM and DD fields are frequently transposed (civilian dates vs. military dates)
- Clients may shave a year off their age (add a year to DOB)
- Clients may deliberately provide bogus DOBs
- Older clients may mis-report DOB due to memory problems
- Algorithms may be constructed to allow for such variations in the order and accuracy of month, day and year fields in providing weights for field matches on DOB.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Best Practices:
- Store date components [Month of Birth, MOB] and [Day of Month of Birth, DMB]
- With leading zeros (01 .. 09 .. 12) format
- (e.g., January is coded as “01”, not “1”)
- Such that future combinations of MM DD
- Are not ambiguous
- Does “111” represent “01-11” or “11-01”
- Block (or question) at input stage common bogus DOBS: 01-01-01 etc

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Note the necessity of using the full four-digit year in date elements.
- Many States currently assign unique client numbers to infants (born 2000 and later) e.g., children of clients in women's residential care or children receiving therapeutic child care while the parent is in intensive outpatient group.
- Also note that within a few years, States will be providing services to adolescents who were born in year 2000 and later.
- Thus, State will need the full four-character year of birth format to clearly distinguish clients born in the 1900s from clients born in the 2000s.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Gender. Gender is usually reliably coded, except for occasional typographical errors (e.g., “1” versus “2” code data entry error).
- In some situations, a clinician or data entry staffer may incorrectly presume a client’s gender, especially for persons with gender-neutral or gender-ambiguous first names (Shannon, Chris, Lynn, Pat, Sandy, Casey, Frankie, Bobbie, Billie, Jessie, etc).

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Race-Ethnicity Fields. For many reasons, a given person's coded race-ethnicity can be discrepant within and across databases:
 - A person may change his or her self-identified racial or ethnic group over time
 - Persons with mixed racial-ethnic heritage may select different racial-ethnicity labels over time
 - Clinicians may code (and miscode) their impressions of a client's race-ethnicity without asking the client

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Different databases may use different race-ethnicity codes that must be crosswalked prior to linking
- The same dataset may change the available race-ethnicity response codes over time
- Clients and staff often confuse race and ethnicity (resulting in inconsistent coding of persons as White vs. Hispanic, Asian vs. White, etc.)
- A moderate amount of inconsistent coding is to be expected on race-ethnicity.
- For this reason, race-ethnicity is not a particularly good linking variable.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Missing Values: A good practice for handling missing values in an identifier data element is to replace missing values (such as missing Middle Initials) with a non-alpha placeholder character (such as dash, “-” or an asterisk, “*”).
- Using a non-character placeholder value for missing data for middle initial (and for all other data elements) allows the creation of multi-element data strings, containing both known values and missing data placeholder characters as necessary.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- It is best to define the full synthetic ID field (and all component date elements) as a "character" or "text" or "alphanumeric" field (not as a numeric field) so that:
- the database users can input leading zeros, as well as
- character strings (M F codes, initials of first and last name, etc.).

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Consider adding a tie-breaker character (or field) to differentiate persons sharing a constructed ID, such as:
- Fraternal twins sharing same initials of FN and LN, same DOB, same gender.
- Other occasional shared constructed IDs between unrelated individuals.

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Maintenance of Constructed Client IDs:
- Client name has changed from one service episode to another
- Date of Birth was incorrectly coded in a previous episode
- Gender code or other ID components were incorrectly coded in previous episodes

Business Rules, Analysis Rules, and Data Cleansing Considerations

- Alternatives:
- Update constructed ID with new information (after verification), then re-synchronize all previous electronic client records to new ID (automated routines). [Note: paper files likely would contain both the old and new constructed client IDs].
- Maintain both old constructed ID and updated constructed ID in separate fields or look-up tables.
- Maintain historical constructed ID tables (with dates of each update) for each client over time.
- Continue using original incorrect constructed ID, but use alias name fields (in association with deterministic and probabilistic matching protocols) to cluster unique clients appropriately under a new temporary unique client ID.

Evaluation of Various Constructed Client ID Protocols

- This section discusses the discriminating power and effectiveness of various unique client identification protocols that a State may employ.
- The discussion that follows provides some observations and suggestions based on analyses of various MH-AOD treatment client datasets as well as synthetic test databases and offers some metrics that would allow a State to assess the discriminating power of its own potential client identification protocols.

Evaluation of Various Constructed Client ID Protocols

- While the observations that follow might be representative of a “typical” State MH-AOD treatment client database,
- the particular unique ID strategy that would work best for any given State will depend upon:
- the availability, completeness, and accuracy of the component data elements in each State's MH-AOD treatment client database,
- plus the availability, completeness, and accuracy of the various data elements in other target databases with which the State MH-AOD treatment agency would like to link.

Evaluation of Various Constructed Client ID Protocols

- In addition, the distribution of particular data elements and the total number of unique values for each data element in each State's database can affect the decision on the most efficient client ID for a State.

Discriminating Power

- Developing a test database. The general approach to assessing the discriminating power of a particular unique client identification scheme is to generate a dataset of uniquely-identified clients with each unique client's record populated with his or her identifiers, as available – SSN or partial SSN, names (or components of names), DOB, gender, etc.
- Configure the dataset to contain one record per uniquely identified individual.
- One approach to generating the unique client dataset is to employ the deterministic and probabilistic deduplication protocols.

Discriminating Power

- For example, a database of 344,730 client episode records (across all providers in the State over an eight year period) might be deduplicated by the deterministic and probabilistic protocols to yield 197,587 uniquely identified individuals.
- Select one record (for example, select the record representing the most recent service episode) for each of these 197,587 individuals.

Relative Discriminating Power

- Using this unique individual dataset, one can calculate a measure of the power of each client data element (DOB, gender, part of SSN, part of name, etc.) alone, or in combination, to discriminate among unique clients.
- Relative Discriminating Power values can range from 0% (worst) to 100% (best).
- For example, the full nine-character SSN should have a Relative Discriminating Power [RDP] value of close to 100%.
- The relative discriminating power of the last four characters of the SSN (SSN4) might be approximately 75%.

Relative Discriminating Power

- The relative discriminating power of DOB might be around 79%
- The relative discriminating power of a string composed of the first and third characters of the first name and the first and third characters of the last name (F1F3L1L3) might be approximately 74%.
- The relative discriminating power of “weak” client identifiers such as middle initial (MI), race-ethnicity (RCE), and gender (GEN) might be approximately 18%, 6%, and 4%, respectively.
- But in combination, a synthetic client identifier composed of elements such as F1F3L1L3DOBSSN4GEN might have a relative discriminating power approaching 99%.

Discriminating Power

- Note that the observations in this section are fairly universal and can be assumed to fairly reliably model the results that any given State might achieve, even without conducting a full evaluation.
- Note, however, that:
 - the completeness and quality of a given State's data,
 - the distribution of the values for each of the client data elements (for example, a client database where gender is distributed 50% male and 50% female will yield different results than a State client database that is 70% male and 30% female), and
 - the total number of records upon which the analysis is based
- can affect the discriminating power values for any given constructed client ID under consideration by a State.

Discriminating Power

- Thus, where possible, States may wish to consider conducting discriminating power analysis using the State's own client datasets.

Calculating Discriminating Power

- Discriminating Power is calculated as:
- $\ln\left(\frac{1}{\sum(p(i)^2)}\right)$, where $p(i)$ =proportion of total clients in each value of the data element or data element string under consideration and \sum = summation of $p(i)^2$ across all values of the data element or data element string.
- Below is an example of the calculation of Discriminating Power for the race-ethnicity data element from a particular MH-AOD treatment client database (approximately 60% white, 30% black, 7% Hispanic, and 3% other race-ethnicities).
- Note that race-ethnicity is not a particularly useful data element by itself for unique identification and linking purposes, but it has a small enough set of possible values (four race-ethnicity codes) that allow for easy illustration.

Calculating Discriminating Power

Discriminating Power Calculation for Race-Ethnicity - RCE			
Category	Count	p(i)	p(i)^2
White	118,809	0.6013	0.3616
Black	60,047	0.3039	0.0924
Hispanic	13,337	0.0675	0.0046
Other Race-Ethnicities	5,394	0.0273	0.0007
Total	197,587	1.0000	0.4592
Calculation	1/sum(p(i)^2)		2.178
Discriminating Power	ln(1/sum(p(i)^2))		0.778
Maximum Possible DP this Database	If Perfect Discrimination		12.194
Relative Discrim Power (RDP) for RCE	DP / Max Poss DB		6.4%

Calculating Discriminating Power

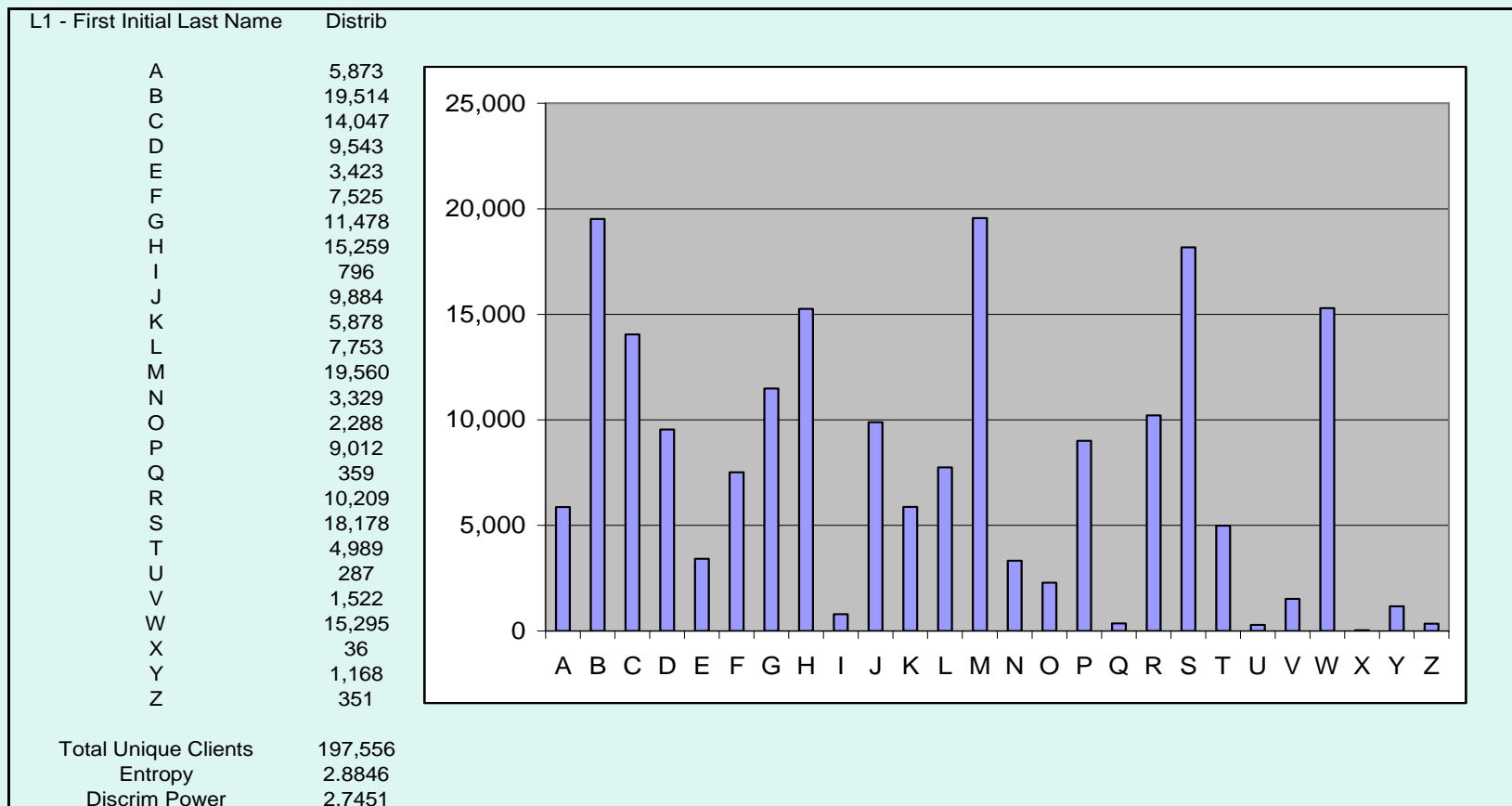
- The calculated discriminating power for the race-ethnicity data elements in this particular dataset is 0.778.
- Under theoretical conditions, a client identifier that perfectly discriminated all 197,587 individuals in this particular dataset would have a discriminating power value of 12.194 (the maximum possible discriminating power value can be calculated using the new unique identifier which is specific to each of the 197,587 individuals in this dataset).
- Thus, race-ethnicity in this particular database has a relative discriminating power of only 6.4% (which is to be expected, since race-ethnicity alone is insufficient to distinguish individuals)

Calculating Discriminating Power

- Race-ethnicity was used in this example for illustrative convenience, not as a recommended data element for unique client identifiers.
- Note that the particular calculated values for every data element (or combination of data elements) will vary depending on the number of unique individuals in the particular dataset and will vary depending on the distribution of the number of individuals in each category (e.g., the racial distribution of clients, in this example).

Another Example of Discriminating Power Calc

- Discriminating Power = $\ln\left(\frac{1}{\sum(p(i)^2)}\right)$, where $p(i)$ =proportion of total clients in each value of the data element or data element string under consideration and sum = summation of $p(i)^2$ across all values of the data element or data element string.
- Below is an example of the calculation of Discriminating Power using the L1 data element. [L1 is not a particularly useful data element by itself for linking and clustering purposes, but it has a small enough set of possible values (26, A-Z) that allow for easy illustration]. Note the variable distribution of the first initial of the last name.



Calculating Discriminating Power

- Discriminating Power is an overall measure of the power of a generated client ID (or data components thereof) to uniquely identify clients and discriminate among unique clients.
- Larger values of the Discriminating Power calculation indicate greater information provided by the data elements of the ID string for discriminating among unique individuals.
- Data elements with a large number of potential values, such as date of birth, will have greater discriminating power than data elements with few potential values (such as gender or race).

Calculating Discriminating Power

- However, the distribution of the data element values also affects the discriminating power.
- If a data element has significant missing data (e.g., SSN is missing for 70% of clients) or
- If the values of a data element are heavily skewed (such as a database where the gender variable is 80% male),
- Then such data elements will not be particularly useful for uniquely identifying or linking clients in this particular dataset.

Discriminating Power of Common Client Identifiers

- This section describes typical data elements used as identifiers in various MH-AOD treatment client databases and provides some general description of the discriminating power of each data element (or combinations of elements).

Social Security Number - SSN

- In a database of (for example) 197,587 unique individuals and 197,587 unique SSNs, a relative discriminating power of 100% would be expected.
- In practice, the value of the relative discriminating power calculation for SSN is usually some value less than 100%, since some SSNs in any database are bogus or typos and deterministic–probabilistic routines can (and do) occasionally incorrectly cluster clients together or fail to cluster persons together under a new unique ID that should have been clustered.
- In general, however, the Relative Discriminating Power (RDP) value for SSN in a database with minimal missing values for SSN will be 99% or better.

SSN4 – Last Four Characters of SSN

- For situations in which it is not possible to collect or link to the full nine-character SSN, use of the last four character of the SSN (SSN4) is an alternative data element with relatively high discriminating power, while preserving some level of confidentiality.
- In a large dataset, would expect to find an approximate even distribution of clients across all 9,999 possible SSN4 values (0-9 values for each character = $10 \times 10 \times 10 \times 10$ - minus one - no "0000" allowed), "0001" through "9999").
- The RDP for SSN4 in a database with minimal missing values for SSN4 is often around 75%.

Client Current Last Name - LN

- The Relative Discriminating Power of Last Name (LN) is often in the 50%-55% range.
- Note that there is not always a direct association between the number of values that a data element can take on and the relative discriminating power of that particular data element.
- In a large dataset, there may be approximately 24,000 last names.
- So why does Last Name (with 24,110 unique values in this dataset) have less relative discriminating power (53%)
- than the last four characters of the SSN (only 9,999 unique values, but a relative discriminating power of 75%)?

Client Current Last Name - LN

- The reason that the total number of possible values of a data element (last name, in this example) is potentially misleading as an indicator of discriminating power is because the simple count of possible last name values does not take into account the distribution of last names.
- The 1,000 most common last names in the US (e.g., Smith, Johnson, Williams, Jones) are shared by 43% of the US population and the top 30 last names in the US account for 11% of all last names used in the US.

Client Current Last Name - LN

- Thus, even though there are far more last names than there are possible combinations of the last four characters of SSN, many of these last names have relatively little discriminating power since so many people share a small set of a few last names.
- The 9,999 possible combos of the last four characters of the SSN, however are randomly assigned to persons.
- Thus, the distribution of the last four characters of SSN is relatively uniform with no predominance toward any one four-character string (e.g., no inordinately large number of persons with, for example, a "3667" SSN4 character string).
- Thus, the SSN4 string has more discriminating power than does last name.

Client Current Last Name - LN

- Last name (or portions of last name) as a component of a unique ID is subject to corruption and missed links if a person changes his or her last name over time.
- Example, a woman may receive MH-AOD treatment services under the last name of Smith at one provider, then years later, change her name due to marriage or divorce and receive services at another provider under the last name of Williams.
- The two providers using the same ID generation protocol that includes characters of the last name would thus inadvertently create two separate "unique" IDs for this one individual, which would need to be reconciled through deterministic and probabilistic matching later at the State agency level.

Client Current Last Name - LN

- Note: If using names (especially last names) or portions of names as part of client ID,
- need to consider that even if the State MH-AOD treatment agency is conscientious about synchronizing and maintaining the same last name components in an ID for a client who has changed names over time,
- the State MH-AOD treatment agency cannot be confident that other external client databases have the ability to synchronize name components over time for such clients
- (or if the external agency does have the ability, that such agency has synchronized to the same last name component as did the State MH-AOD treatment agency).

Client Current Last Name - LN

- Example, a specific female client may have had last names of Smith and Jones (through marriage, divorce, personal choice, adoption, etc.) over time.
- The State AOD treatment agency may have synchronized its last name component of the client ID to the Smith last name and maintained the Smith component in the person's ID over time.
- However the Mental Health agency at which she was also a client in the past may have the person listed as (or synchronized to) the Jones last name.

Client Current Last Name - LN

- Note: All client ID strings using full first name (FN), full last name (LN), or NYSIIS FN, or NYSIIS LN will be variable length IDs, since name lengths can vary.
- States using such name components as part of a synthetic ID may wish to establish a maximum estimated fixed length field for such IDs and right-fill shorter names with spaces, etc.
- Typical field length for LN = 20 characters

Client Current First Name - FN

- The relative discriminating power of first name (often in the range 44%-50%) is generally less than the RDP for last name.
- A large dataset might have 11,000-16,000 first names (compared to 24,000 or so last names).
- There are fewer first names in use in this country than last names.
- As with last names, the discriminating power of first name is less than the RDP of SSN4 (even though there are more distinct values of first name than SSN4) due to the unequal distributions of first names (and middle names) in the US (concentrations among James, Robert, John, Michael, Mary, Pat, Linda, Jennifer, etc.).

Client Current First Name - FN

- Also over the last several decades, there has been a compression of first names selected for babies in the US, as increasing numbers of parents opt for a relatively small set of trendy names (e.g., Jacob, Joshua, Emily, Madison) each year.
- Thus, identifiers that utilize first names or components of first names are likely to continue to decrease in discriminatory power unless this trend is reversed.
- Also note that in general, there are fewer unique first names for males in the US compared to the larger variety of first names for females.
- Thus, in datasets in which males predominate, the discriminating power of first name is reduced further.
- Typical field length for FN = 15 characters

Phonetic Encoding of Names

- Phonetic encoding protocols can be used to assess whether any two names under comparison “sound alike” in spite of different spellings or misspelling.
- The most commonly used phonetic protocols used are:
- NYSIIS (New York State Identification and Intelligence System) and
- Soundex

Phonetic Encoding of Names

- For example, compare the first names Katrina and Catreena.
- Are the names similar enough that they might represent the same individual?
- Under the NYSIIS phonetic coding scheme:
- KATRINA is “phonetically” coded as CATRAN and
- CATREENA is coded as CATRAN as well, suggesting a potential match.

Phonetic Encoding of Names

- Soundex encoded names consist of a letter and three numbers.
- Under Soundex,
- D'ANGELO is coded as D524 and
- DEANGELIS is coded as D524 as well,
- suggesting a potential match.

Phonetic Encoding of Names

- Note: NYSIIS and Soundex have multiple versions that can result in slightly different coding.
- NYSIIS is a more sophisticated phonetic protocol than Soundex.
- Other software packages use other phonetic encoding algorithms such as Metaphone and Double Metaphone.

NYSIIS Phonetic Transformation of Last Name (LNN)

- Use of phonetic transformations:
- Help reduce the impact of missed record linkages due to misspellings, and alternative spellings and
- Help provide a measure of confidentiality.
- Could also use Soundex, double metaphone, and other phonetic algorithms.

NYSIIS Phonetic Transformation of Last Name (LNN)

- Use of any phonetic transformation as part of an assigned ID will require that the assigning parties (clinicians, etc.) have access to an electronic version of the phonetic algorithm and that all assigning parties are using the same version of NYSIIS, etc.
- In most database linking projects, it will be necessary to apply the phonetic transformation algorithm to the names in target linkable datasets.

NYSIIS Phonetic Transformation of Last Name (LNN)

- The RDP for LNN will be less than that for the full last name,
- But still often in the 48%-50% range.

NYSIIS Phonetic Transformation of First Name (FNN)

- The issues addressed above for the last name (LNN) apply to the first name (FNN) as well.
- RDPs for FNN are often in the 38%-42% range.

First 4 Characters of Last Name (LN4)

- Note that if the intent of using partial names is to preserve some confidentiality, four-character names will not do that.
- With four character names (even with three-character names), it is often possible to guess client names and identities fairly easily.
- If a name is less than five characters, then the maximum available characters in the name will have been used in the LN4 data element.

First 4 Characters of Last Name (LN4)

- The first four characters of last name has less discriminating power than last four characters of SSN since last four characters of SSN are random and not constrained,
- whereas adjacent characters of a name:
- are constrained by commonality
- (e.g., SMIT representing Smiths)
- and constrained by phoneme rules of the language
- (e.g., HLGE is an improbable first four characters of a last name).

First 4 Characters of Last Name (LN4)

- The RDPs for LN4 (51%-52%)
- are often marginally higher than the RDPs for the phonetic transformation of last name (LNN)
- but at a cost in confidentiality.

First 4 Characters of First Name (FN4)

- FN4 provides very little confidentiality
- since many first names are four characters or shorter (e.g. John, Ann) and
- a full first name of any length is often easily guessed on the basis of the first four characters.
- RDPs for FN4 (41%-47%) are often marginally higher than the RDPs for the phonetic transformation of first name (FNN)
- but at a cost in confidentiality.

First 3 Characters of Last Name (LN3)

- The issues addressed above apply to this component as well.
- RDPs often range from 46%-47%.

First 3 Characters of First Name (FN3)

- The issues addressed above apply to this component as well.
- RDPs often range from 39%-42%.

First 3 Characters of NYSIIS Transformed LN (LNN3)

- RDPs often range from 37-38%.

First 3 Characters of NYSIIS Transformed FN (FNN3)

- RDPs often range from 32%-34%.

First 2 Characters of Last Name (LN2)

- The maximum theoretical number of possible values for LN2 is 677 ($26*26 + 1$ known missing value placeholder string),
- but not all adjacent combinations are possible
- e.g., “Qu” is possible as the first two characters of a last name, but “Qx” is not likely.
- One seldom finds more than 300-339 possible LN2 combinations in a typical US name database
- Note: apostrophes, hyphens and other name punctuation (such as O’ in O’Neal) should be removed in the data preparation stage prior to analysis.
- RDPs for LN2 are often 35%-36%.

First 2 Characters of First Name (FN2)

- The maximum theoretical number of possible values for FN2 is 677 ($26*26 + 1$ known missing value placeholder string), but not all adjacent combinations are possible
- e.g., “Qu” is possible as the first two characters of a first name, but “Qx” is not likely.
- One seldom finds more than 290-360 possible FN2 combinations in a typical US name database.
- Note: apostrophes, hyphens and other name punctuation (such as L' in L'Shaun) should be removed in the data preparation stage prior to analysis).
- Also some FNs are only one character, such as F Scott Fitzgerald.
- RDPs for FN2 are often 32%-33%.

First and Third Characters of Last Name (L1L3)

- RPCs for L1L3 are often 40%-41%.
- Note that the L1L3 approach yields more unique values and higher discriminatory power than the LN2 element, even though both elements are two characters in length.
- The L1L3 advantage is due to fact that two adjacent characters are more restrained in possible combinations,
 - e.g., an adjacent “WL” is seldom seen in a name,
 - but a “W_L” combination (as in Williams, Wales, etc.) is quite possible.
- If a client's last name is only one or two characters (e.g., “Yi”) then L1L3 will be only one alpha character long (“Y”) [or” Y*” with place-holder].

First and Third Characters of Last Name (L1L3)

- A State considering using any two-character name segment as a data element, may wish to consider L1L3 and F1F3 rather than LN2 or FN2,
- since the 1_3 approach yields more discriminating power and
- since the 1_3 approach yields some extra confidentiality.
- A three-character name element (e.g., LN3, FN3) offers even more discriminating power than L1L3, F1F3, but less confidentiality.
- Components can be scrambled for greater confidentiality (e.g., use L3L1 - third character of last name, followed by first character of last name).

First and Third Characters of First Name (F1F3)

- Note: If client's first name is only one or two characters (e.g., "Al") then F1F3 is only one alpha character long ("A") [or A*].
- RPCs for F1F3 are often 35%-38%.
- Again, components can be scrambled for greater confidentiality
- e.g., F3F1 - third character of first name, followed by first character of first name, etc.

First Character of Last Name (LN1)

- Number of expected unique values of LN1 = 27 (26 alpha characters plus a missing value indicator).
- RDPs for LN1 are generally 22%-23%.

First Character of First Name (FN1)

- Number of expected unique values of FN1 = 27 (26 alpha characters plus a missing value indicator).
- RDPs for FN1 are generally 22%-23%.

Client Current Middle Initial (MI)

- Number of expected unique values of MI = 27 (26 alpha characters plus a missing value indicator).
- Middle initial is often missing in up to 20%-30% of client records, since MI is often an optional field and some people do not have a middle name.
- In a database with, for example, 25% missing data on MI, 25% of the unique clients will, in effect, “share” the “missing” value indicator for MI.
- Such relatively large amounts of missing data will lower the relative discriminating power for the particular data element accordingly.

Client Current Middle Initial (MI)

- RDPs for MI in a dataset with minimal missing data (e.g., 1% missing) on MI may be around 23%.
- RDPs for MI in a dataset with approximately 27% missing values on MI may be around 18%.

Client Current Middle Initial (MI)

- Thus, MI is not particularly recommended as a data element since some people have no middle name and since middle initial can change in situations where a woman's birth middle name is supplanted by use of her maiden last name as her new middle name after marriage (and the possible reverse after divorce, etc.).
- But, if Middle Initial or Middle Name is collected, such information can be used to help resolve marital name change situations

Date of Birth (DOB)

- In large datasets (e.g., approximately 200,000 unique individuals or more)
- with relatively large age ranges (e.g., 60-plus year span of potential ages among the clients in any given year),
- DOB is approximately as discriminating as SSN4 (and usually slightly better than SSN4).
- DOB was coded in MMDDYYYY format in these examples.

Date of Birth (DOB)

- DOB in such databases often has greater discriminating power than last name even if there are more distinct values of last name than DOB.
- DOB will have greater discriminating power than last name due to its more uniform distribution of values
- Dates of birth are relatively evenly distributed
- Whereas last names have a skewed distribution, with a large number of persons sharing a relatively small number of last names.
- RDPs for DOB in databases of this type are often in the 78%-79% range.

Date of Birth (DOB)

- DOB may be less useful as a potential identifier element if the data set under consideration has a relatively compressed age range.
- For example, a database of middle school students (grades 6-8) would have only approximately 1,095 randomly distributed potential DOBs across an approximate three year range (365 days * 3 years).
- In such a database, DOB is likely to have less discriminating power than SSN4 (potentially 9,999 randomly distributed SSN4 values, depending on the number of unique records in the dataset).

Year of Birth (YOB)

- RDPs for YOB will depend upon the number of distinct years of birth present in the dataset(s) under review.
- In the case of a longitudinal dataset covering, for example, 8-10 years of client data, one may find 95 or more distinct years of birth in the dataset.
- Such longitudinal datasets, with client age distributions appropriate to a typical MH-AOD treatment clientele, may have RDPs for YOB around 30%-31%.

Month of Birth (MOB)

- RPCs for MOB are typically around 20%,
- reflecting the relatively even distributions of clients across the 12 potential months of birth.

Date of Month of Birth (DMB)

- RDPs for DMB are typically around 28%,
- reflecting the relatively even distributions of clients across the 31 potential days of the month of birth.

Current Zip 5 of Residence (ZIP)

- Most medium and large size States will have 600-700 or more five-character zip codes. RDPs for zip can range up to 41%-42%.
- However, use of ZIP as a component in a unique identifier is not recommended, since people change addresses frequently.

Current County of Residence (COR)

- Most medium and large size States will have 50-100 or more counties.
- RDPs for COR can range up to 20%-27%.
- However, use of COR as a component in a unique identifier is not recommended, since people change addresses frequently (and since some smaller States – and DC – have as few as 1-3 “counties”).
- Also note that COR would not be very discriminating in a geographically-limited client dataset.

Race-Ethnicity (RCE)

- A combined race and ethnicity code is not recommended
- since people may change their self-identified race-ethnicity over time and
- since target linkable datasets may not code race-ethnicity the same or as extensively or
- may not code ethnicity at all, or
- may combine Hispanic ethnicity as an “other” race, etc.
- Therefore race-ethnicity by itself is a weak client identifier.
- RPCs for race-ethnicity are often in the 5%-7% range.

Race-Ethnicity (RCE)

- Race-ethnicity becomes less useful as an identifier in databases that are skewed toward one or two race-ethnicity groups.
- In many client datasets, 70% or more of the client population may be coded under one race-ethnicity category
- and frequently 90% or more of a State's client population is coded under two race-ethnicity categories, e.g., 60% white, 30% black.

Gender (GEN)

- As with many of the above data elements, by itself, gender is a weak client identifier.
- RDPs for gender are typically in the 4% range.
- Gender does become useful as a client identifier data element when used in combination with other more discriminating data elements (such as SSN4, DOB, etc.).

Gender (GEN)

- However, gender becomes less useful as an identifier as a database population becomes increasingly skewed to one gender.
- For example, clients in juvenile detention facilities are often 85% male (or higher).
- Thus, a gender data element would not be particularly useful as part of a client ID in this database, nor would gender be a particularly efficient variable for linking this database to another.

Gender (GEN)

- Such a skewed distribution variable can however be useful in the exceptional case.
- For example, if you linked this juvenile incarceration database to a MH-AOD treatment database and found a possible record match in which both datasets indicated that the client was female, you would have a greater degree of “value-specific” confidence in that particular potential match than if any two records showed a link in which both records matched on male.

Multi-Element Client Identifier Strings

- The most effective approach to generating a synthetic potentially-unique client identifier is to combine multiple client data elements into one client identifier as a string of characters,
 - either directly as part of a formal client identifier or
 - indirectly as part of a temporary client identifier used during client deduplication and client data linking analyses.

Multi-Element Client Identifier Strings

- Obviously, the full nine-character SSN, used by itself or in combination with full first name, full last name, and full DOB, will provide the greatest discriminating power of all possible client IDs.
- In situations where full SSN, full names, and possibly full DOB are not available, various synthetic client IDs can be created using data elements such as SSN4, components of names, DOB or components of DOB, gender, etc.
- While several of the client data elements are weak discriminators individually, when used in particular combinations, a reasonably discriminative client identifier can be developed.

Multi-Element Client Identifier Strings

- The various client data elements described in this section can be combined in hundreds of potential configurations
- Such as L1L3 + F1F3 + DOB + SSN4 + GEN
- to generate a constructed client ID.
- Over 150 such constructed ID combinations were reviewed for this analysis.

Multi-Element Client Identifier Strings

- In an attempt to restrict the analyses to just those client data combinations with the potential to be reasonably discriminating, a series of four screening thresholds were used, the chief criteria of which were that:
 - the client ID string must yield a relative discriminating power of 97.5% or higher and
 - that incomplete or missing data on any of the component data elements in the data element string should not exceed 25%.

Multi-Element Client Identifier Strings

- Using these thresholds, the most discriminating client identifiers are summarized below,
- classified by whether the identifier required use of full SSN, full names, and other key data element that may or may not be present in any given dataset(s).

Best solutions with full SSN

- With any number of additional elements and any length
- **SSN + FN + LN + DOB** [Variable Length, Unwieldy]
- With smallest number of additional elements without significant loss in discriminating power
- **SSN + FN + LN**
- With shortest length without significant loss in discriminating power
- **SSN + FN1 + LN1 + DOB**

Best solutions with partial SSN

- With any number of additional elements and any length
- **FN + LN + DOB + SSN4**
- With least number of additional elements without significant loss in discriminating power
- **DOB + SSN4**
- With shortest length without significant loss in discriminating power
- **F1F3 + L1L3 + SSN4**

Best solutions without SSN and without partial SSN

- With any number of additional elements and any length
- **FN2 + LN3 + DOB + GEN**
- With least number of additional elements without significant loss in discriminating power
- **LN + DOB**
- With shortest length without significant loss in discriminating power
- **F1F3 + L1L3 + DOB**

Best solutions that do not use any SSN component or full first name or full last name (can include partial name or NYSIIS name or initials or component characters of a name)

- With any number of additional elements and any length
- **FN2 + LN3 + DOB + GEN**
- With fewest number of additional elements without significant loss in discriminating power
- **FN3 + LN3 + DOB**
- With shortest length without significant loss in discriminating power
- **F1F3 + L1L3 + DOB**

Best solution that does not use full FN or full LN
but does allow up to 3 characters each of FN
and LN and does not use SSN or partial SSN

- F1F3 + LN3 + DOB + GEN

Best solution that does not use full FN or full LN but does allow up to 3 characters each of FN and LN and does use partial SSN

- FN3 + LN3 + DOB + SSN4

Best solution that
does not use full SSN or partial SSN,
does not use full FN (though partial FN is OK),
does not use full LN (though partial LN is OK),
and does not use full DOB (though up to two
components of DOB is OK)

- **FN3 + LN3 + MOB + YOB + GEN**

Best solution that
does not use any SSN component and
does not use any name component
(no full names, no partial names, no initial, no NYSIIS)
with any number of additional elements and any length.

- No solution with sufficient discriminating power meeting all suggested effectiveness thresholds is available

Best solution that
does not use any SSN component, and
does not use any name component (no full names,
no partial names, no initial, no NYSIIS, etc), and
does not use any DOB component
with any number of additional elements, any length

- No solution with sufficient discriminating power meeting all suggested effectiveness thresholds is available

Minimum Elements Necessary for Effectiveness

- Note: All client identifiers that met the effectiveness screening thresholds had at least one or more of the following three components:
 - SSN (or component),
 - DOB (or component),
 - Last Name (or component).
- It was not possible in this analysis to construct a client identifier using any typically available elements other than SSN, DOB and LN (or components of any of these three elements) that would yield a sufficiently discriminating client ID.

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- The best specific client ID approach for any given State will depend on the particular component data elements collected by the State MH-AOD treatment agency and by its potential database linkage partners.
- However, a few general recommendations are possible.

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- In general, if a State has:
- reasonably clean data elements (correct and consistent spelling, minimal typos, minimal invalid codes, etc.) and
- complete data elements (minimal number of records with missing values),
- then almost any synthetic client ID components selected from data elements such as LN and FN (and variants), DOB, gender, SSN, and partial SSN, etc., will work almost equally well.
- Example: **F1F3 + L1L3 + SSN4 + DOB + GEN**

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- Thus, a State can adopt a synthetic client ID strategy based on the State agency's own internal client tracking needs and based on the availability of potential linking data elements in the external data sets to which the State MH-AOD treatment agency would like to link.

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- Beyond a certain level of complexity, all multi-element data strings perform similarly.
- After a point, massively complex multi-element client IDs do not appear to add much additional discriminating power.

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- States can select an ID strategy depending upon the availability, reliability, accuracy, and completeness of the component data elements in the State's own MH-AOD treatment client database and in the databases to which the State MH-AOD treatment agency would like to link.

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- A State may wish to select a relatively discriminating unique ID to meet its own internal client identification needs for service delivery purposes –
- Perhaps SSN,
- Perhaps a centrally assigned and administered master client index,
- Perhaps a synthetic ID based on relatively fixed client characteristics (e.g., FN2 LN3 DOB GEN).

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- In addition, for maximum flexibility to link to external datasets, a State may wish to consider collecting all of the following:
 - SSN (or partial SSN)
 - First name
 - Last name
 - Middle name
 - DOB
 - Gender
 - Race-ethnicity (in as many categories as possible).

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- States also should develop the ability to generate NYSIIS and Soundex phonetic translations of names
- Programming code for NYSIIS and Soundex and other phonetic transformations is readily available on the internet.

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- If possible and appropriate, States should collect other client identifiers that may be useful for matching against other data sets (e.g., identifiers such as Medicaid number, corrections number, driver's license number, etc.).
- From these data elements, a State MH-AOD treatment agency should be able to uniquely identify 99% or more of its clients and should be able to successfully link to almost any external client database (hospital discharge, arrests, social services, etc.).

General Observations Regarding the Discriminating Power of Unique Client Identifiers

- The success of any client ID strategy also is contingent upon the
 - reliability,
 - accuracy, and
 - completeness
- of the data elements used in the client ID
- Both on the part of the MH-AOD treatment agency and on the part of potential linking client database partners.

Missing Data

- The datasets used in this analysis have relatively little missing data.
- Missing data often can be minimized through the use of data input software that:
 - does not allow “mandatory” fields to be skipped, and
 - which prevents the entry (in real time) of most invalid data
 - (e.g., gender value of “4”) and
 - which blocks or questions most obviously bogus data
 - (e.g., DOB = 01/01/01 or SSN = 123-45-6789).

Missing Data

- As a result, the discriminatory power estimates for the data elements in the datasets used in this analysis are close to the “best case” obtainable for databases of this type.
- If a State dataset has significant amounts of invalid or bogus data values, or if a dataset has a lot of missing values for various data elements, the corresponding discriminatory power estimates will be lower.

Effectiveness of Various Unique Client Identifier Protocols and Data Linking Protocols

- All constricted client identifier protocols are subject to False Positives and False Negatives
- Example: Assume that a State MH-AOD treatment agency is using the following client data elements to deduplicate its own database (to obtain unique client counts) or is using the following data elements to link to, for example, an arrest dataset:
 - **F1F3L1L3 + MI + DOB + GEN + SSN4**

False Links

- False Links. The above identifiers would mistakenly link the following two records and would incorrectly determine that the following two records represent the same individual.
- MILLER DEONTA SHEPARD 12201965 M 780-70-7768
MLSED12201965M7768
- MILFORD DALTON SKETOE 12201965 M 896-32-7768
MLSED12201965M7768
- A deterministic-probabilistic deduplication-linking routine operating with the full identifiers would easily identify these two records as representing two different individuals.

False Links

- False links can never be completely eliminated but can be reduced to an “acceptable” level through careful selection of client identifier strings and linking protocols appropriate to the particular task at-hand.
- False positive rates under deterministic-probabilistic routines with moderately clean data (or better) are often in the 2%-3% range.

Missed Links

- Missed Links. The above client identifier protocol would fail to determine that the following two records represent the same individual.
- DOUGLAS OLIN FOWLER 01171963 M 780-81-6552
DUFWO011719636552
- OLIN DOUGLAS FOWLER 01171963 M 780-81-6552
OIFWD011719636552
- A deterministic-probabilistic deduplication-linking routine operating with the full identifiers would easily identify these two records as representing the same individual.

Missed Links

- A client identification protocol that generates a large number of missed link records would, in turn:
 - 1) generate an inflated estimate of the number of “unique clients” in the database (since the same individual , as in the example above, would be inappropriately coded under more than one client ID string), and
 - 2) would under-represent each identified client’s full service history (since portions of each client’s history would be inappropriately coded under separate unlinked client IDs, as above), and
 - 3) would reduce the chances for successful, complete links to external datasets (such as arrest databases, etc).

Missed Links

- Missed links can never be completely eliminated but can be reduced to an “acceptable” level through careful selection of client identifier strings and linking protocols appropriate to the particular task at hand.
- False negative rates even under deterministic-probabilistic routines with moderately clean data (or better) will often range from 8%-15%.

Effectiveness of Various Unique Client Identifier Protocols and Data Linking Protocols

- Different client identification protocols will yield varying rates of
- false links (false positive rates) and
- missed links (false negative rates).

Effectiveness of Various Unique Client Identifier Protocols and Data Linking Protocols

- A State can estimate the degree of false positives and false negatives various ways:
- 1) Conduct paper chart audits on a sufficiently large, representative sample of client files to estimate the false positive and false negative rates associated with its client ID protocol.

Effectiveness of Various Unique Client Identifier Protocols and Data Linking Protocols

- 2) Use probabilistic-deterministic client matching software to estimate the “true state” (best estimates of which clusters of client records represent the same individual) and calculate false positives and false negatives against this presumed “true state”

Effectiveness of Various Unique Client Identifier Protocols and Data Linking Protocols

- 3) Develop a synthetic test database containing a sufficiently large number of fully identified records to model its own real client database (if it were to contain full identifiers) and then assess the false positive and false negative rates for various constructed identifiers using the test dataset.

Interactions Between False Positive Rates and False Negative Rates.

- A State can make its constructed client identifier more conservative or more liberal by:
- Adding or subtracting data elements to the constructed ID
- Using more discriminating or less discriminating data elements (F1F3 instead of FN2 etc)
- Altering any probabilistic matching routines to accept fewer or more discrepancies in Name spellings, DOB discrepancies, SSN4 discrepancies etc to be considered as “matches”

Interactions Between False Positive Rates and False Negative Rates.

- Assume that a State's current client identification protocol has a
- False Negative Rate = 11.0% and a
- False Positive Rate = 3.7%.
- Assume that the State feels that the 11% false negative rate is unacceptable.
- This 11% false negative rate means 11% of the client records that should be clustered together are not (i.e., are missing 11% of the potential client record associations within the primary client database) and that 11% of potential links to an external dataset (such as arrests) would be missed.

Interactions Between False Positive Rates and False Negative Rates.

- The State tweaks its constructed client ID to be more “liberal” (using fewer or less discriminating data elements) to increase the number of links and reduce the number of missed links.
- The new liberal constructed client ID now has
- False Negative Rate = 5.0%
- False Positive Rate = 11.5%
- The False Negative Rate has gone down but the False Positive Rate has gone up.

Interactions Between False Positive Rates and False Negative Rates.

- In many situations, however, you will want to minimize the False Positive Rate
- For purposes of outcome studies, false positives are generally considered the greater problem
- For example, it is worse to incorrectly link one person's MH-AOD treatment service record to a different person's arrest record [a false link]
- than it would be to simply miss a potential match between this person's MH-AOD treatment service record and this person's arrest record that you failed to identify and link to [a missed link].

Interactions Between False Positive Rates and False Negative Rates.

- So now, the State tweaks its constructed client ID to be more “conservative” (using more and/or more discriminating data elements) to reduce the number of false links (minimize false positive rate).
- The new conservative constructed client ID now has
- False Negative Rate = 29.8%
- False Positive Rate = 0.2%
- The False Positive Rate has gone down (as desired) but the False Negative Rate has gone up.

Interactions Between False Positive Rates and False Negative Rates.

- Thus, while constructed client identification and linking protocols can be manipulated to reduce the false positive rate, the false positive rate can not be reduced without increasing the false negative rate (and the converse).
- Selecting a constructed client ID and linking protocol always involves a trade-off between false positives and false negatives.

Probabilistic and Deterministic Unique Client Identification and Client Data Linking Software Resources

- IDB algorithms developed by MEDSTAT for the Substance Abuse and Mental Health Services Administration's (SAMHSA) Integrated Database Project. Dan Whalen et al.
- <http://www.csat.samhsa.gov/idbse/lnkrptch1.asp>

Probabilistic and Deterministic Unique Client Identification and Client Data Linking Software Resources

- Link King, a public domain linkage program developed by Washington State's Division of Alcohol and Substance Abuse. Kevin Campbell et al.

<http://the-link-king.com/>

- Portions of the Link King linkage protocol were adapted from the IDB algorithms.

Probabilistic and Deterministic Unique Client Identification and Client Data Linking Software Resources

- Many additional commercial and public domain data linkage software products are available.
- See listing of products and projects employing data linkage software in the Administrative Data Guide draft document.
- Some products are complete packages, other products may focus on related sub-tasks, such as mailing list database deduplication or data cleansing.

Contact Information

- Dennis Nalty
- American Institutes for Research
- 101 Conner Drive
- Suite 301
- Chapel Hill NC 27514
- 919 918 2307
- dnalty@air.org