

Matching Patient Data

2nd Nationwide Health Information Network Forum:
Health Information Network Security and Services

October 16-17, 2006

Panel Discussion

Don Grodecki, President, Browsersoft Inc, Chief Architect 

Connecting for Health Consortium

Dave “Casey” Webster, Chief Architect

IBM Consortium



Data Needs for Matching Patients without Unique Identifiers

- **Information Exchange between disparate healthcare systems depend on ability to match patient identities without benefit of common identifiers**

- **The Problem:**
 - Given sets of identifying information (matching variables) ...
 - e.g. Name, Date of Birth, Address, SSN (perhaps), ...
 - For a set, or multiple sets, of patient records ...
 - Determine which of the records are for the same patient

Theory of Probabilistic Matching

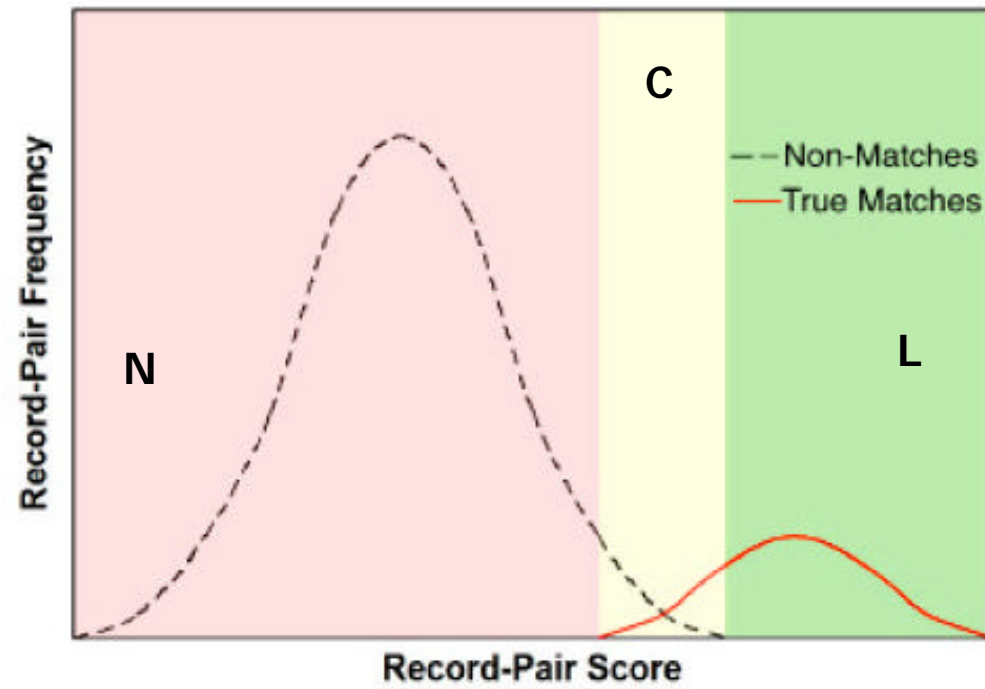
- **A Matching Rule divides the set of all possible record pairs into three sets:**
 - L: (Matched, or Linked)
 - N: (Not matched)
 - C: (not determined, needs Clerical review)
- **The Sensitivity (m) of a rule is the probability that the rule declares a match when there really is a match**
 - $1-m$ is the probability of a false negative
- **The Specificity ($1-u$) of a rule is the probability that the rule predicts a non-match when there really is a non-match**
 - u is the probability of a false positive
- **Obviously we want both Sensitivity and Specificity to be high**
- **As an example, using the stated value of “Gender” to decide a match has high Sensitivity (0.99..) but low Specificity (0.5)**

Matching Patient Data



Theory of Probabilistic Matching

- Standard practice is to build a rule using the weighted sum of the values of comparators that each evaluate the match of a single matching variable and assign a value between "0" and "1" to the match.
- The L, C, and N sets are determined from cutoff values applied to the combined score.



Theory of Probabilistic Matching

- $u \sim u_1 \times u_2 \times \dots \times u_k \times \dots \times u_K$
 - so, we minimize false positives by comparing a sufficient number of independent variables with high specificity
- Fellegi and Sunter (1969) proved that the “optimal” weight for the comparator for independent variable “k” is:
 - $\log_2(m_k)/\log_2(u_k)$
- “Optimal” in the sense that the L and N sets are maximally “distinct”

Process

- **Data Cleaning**

- The possibility of matching is greatly enhanced by pre-processing the variables using specific algorithms for each variable
 - Remove most punctuation in names
 - Removing bad values: e.g. "9999...", "0000..."

- **Standardization**

- Upper/lower case
- Mapping nicknames to standard names
- USPS address processing services

- **Pre-processing**

- Computing phonetically encoded values, e.g. Soundex

- **Blocking**

- Optimizing database queries by a-priori requiring some exact matches

- **Post-Processing**

- Using nearness operators on a set of candidate matches

- **Clerical Intervention**

- Manual processing, a-priori and/or on-the-fly

Patient Matching Errors and their Impacts

- **Bad match ("false positive")**
 - Violation of privacy of wrongly matched individual
 - Data returned could impact diagnosis and/or treatment
 - Clinician and patient faith in the system adversely impacted

- **Missed Match ("false negative")**
 - Missing data could be important to diagnosis or treatment (recurring symptoms, allergies, repeated tests)
 - Clinicians won't trust a system they perceive as delivering partial information

Patient Matching Challenges that Affect Accuracy

- **No universal patient identifier**
 - Nor would one work, reference Great Britain

- **Demographics change**
 - Americans age 18-65 average 1 move every 5-6 years
 - Every year 35% of Americans age 20-30 move
 - Telephone numbers change frequently/Multiple numbers common
 - Name changes due to marriage, divorce, other

- **Cultural Impact**
 - Soundex, Metaphone based on names of European descent
 - Cultural diversity impacts "near-match" algorithms
 - Longest Common Substring, Levenshtein Edit Distance do poorly on names like "Lee", "Li", "Leigh"

Patient Matching Challenges that Affect Accuracy

- **Quality of data**

- Name suffixes (Jr, Sr, III, etc) are often omitted
- Compound (hyphenated) last names increasingly common
- Missing middle name does not imply lack of a middle name
- Names often have multiple spellings or variants
 - Smith/Smythe, Mac/Mc, Dave/David

- **Special Cases**

- Single names ("Cher", "Bono")
- George Foreman

Matching Patient Data



Architectural Approaches

<p>Centralized Matching</p> <ul style="list-style-type: none">+/- All demographics available, but perhaps not populated+ Matches will be consistent across the entire NHIN- Requires centralized database, privacy- Performance may be an issue	<p>Local (Community) Matching</p> <ul style="list-style-type: none">+ Community has personal knowledge of patients, which can aid in matching- Available demographics limited to what each community "knows"- Match success depends on community- Need to link individuals across communities
<p>Homogenous Matching (Single System)</p> <ul style="list-style-type: none">+ Algorithms and data are consistent+ Same input always results in same result- Simplifies administration and validation	<p>Eclectic Matching (Multiple Systems)</p> <ul style="list-style-type: none">+ Leverages existing matching systems that already work and may have large clerical investment+ Lowers barrier to entry for some org's- Tuning and administration require coordinated effort- Same input may result in different results due to differences in underlying matching algorithms
<p>Deterministic Matching</p> <ul style="list-style-type: none">+ Much smaller chance of false matches- More potential matches will be missed- Without a UPI or similar identifier, requires manual or external confirmation	<p>Probabilistic (Stochastic) Matching</p> <ul style="list-style-type: none">+/- Tradeoff between potentially missing a match vs returning a mismatch

Matching Patient Data



Architectural Approaches

<p>Persistent Matching</p> <ul style="list-style-type: none">+ Once a match is made, it is permanent+ Potential for a-priori clerical review- Requires all systems involved be able to (logically or physically) persist the match- Potential for disagreement on clerical matches	<p>Transient Matching</p> <ul style="list-style-type: none">- Matches occur "on-the-fly" and could result in different matches over time+ Easier to integrate existing systems
<p>All-or-Nothing</p> <ul style="list-style-type: none">+/- Returns either a match or nothing+ Simplifies use of the results+ Better privacy of patient list- Higher rate of false negatives	<p>List of Candidates</p> <ul style="list-style-type: none">+ Returns potential matches with a match probability and allows end user to choose+ Fewer false negatives- Potentially more false positives- Potentially exposes another patient's data

Matching Patient Data



Questions for Discussion

- Is there an allowable threshold of “false positives”?
- What is the minimum acceptable threshold for “false negatives”?
 - How does the age of the data affect this threshold?
- Is further matching necessary to tie providers to patients?