A SAS Program to Identify Duplicates in Clinical Data
A SAS Program to Identify Duplicates in Clinical Data

Xiaoli Lu
VA Cooperative Studies Program, Perry Point, MD 21902

ABSTRACT
Duplicates in a clinical trial or survey database could jeopardize the data quality and integrity and induce biased analysis results. These complications often happen in clinical trials, meta-analyses, and registry and observational studies. Common practice to identify possible duplicates involves sensitive personal information, such as name, social security number (SSN), date of birth, address, telephone number, etc.; however, access to this sensitive information is limited, and sometimes it is restricted. As a measure of data quality control, a SAS program was developed to identify duplicated individuals using non-sensitive information, such as age, gender, race, medical history, vital signs, and laboratory measurements. A probabilistic approach was used by calculating weights for data elements used to identify duplicates based on two probabilities, i.e. probability of agreement for an element among matched pairs and probability of agreement purely by chance among non-matched pairs. For elements with categorical values, agreement was defined as matching pairs sharing the same value, and for elements with interval values, agreement was defined as matching values within one percent of measurement precision range. Probabilities used to compute matching element weights were estimated using an expectation-maximization (EM) algorithm.

BACKGROUND

Duplicates
* Individual subject be recruited in one study more than one time (different sites ) or multiple similar studies
* Jeopardize data quality and integrity and induce biased analysis results, either inflating or masking treatment signals

Sensitive Information
Traditional methods to identify duplicates focus on the personal information: SSN, Name and initials, Birth date, Birth place, Address, Zip code, Telephone number etc. Access to these sensitive information is restricted.

Non-Sensitive Information
As a measure of data quality control, we propose a method to identify duplicated individuals using non-sensitive information, such as Age, Gender, Medical history, Vital signs, and Laboratory measurements etc.

Approaches
* Statistical technique to identify records that belong to the same individual without personal identifier
* Select common features from clinical trial datasets as matching variables
* Probabilistic (Matching variables high chance match)

Why Probabilistic Approach
* Two members that refer to the different individual can generate identical records
* Two members that refer to the same individual can generate different records

METHODS

Theoretical Model
Use probabilities of agreement and disagreement between matching variables (Fellegi and Sunter method)
Let \( \alpha \) and \( \beta \) be the corresponding records for a and b. Assume \( k \) features available to be used for the identification if \( a = A, b = B \) and \( \alpha(a) = (a_1, a_2, \ldots, a_k) \) and \( \beta(b) = (b_1, b_2, \ldots, b_k) \), the information refers to the same variables, then a comparison vector \( \gamma \) would be \( \gamma = (\gamma_{a(a),b(b)}, \gamma_{a(a),b(b)}, \ldots, \gamma_{a(a),b(b)}) \). Assume the comparison vector \( \gamma(a,b) \) to be a random variable, the conditional probabilities producing \( \gamma \) given \( (a, b) \in U \) are:

\[
\begin{align*}
P_{\gamma(a,b)} &= P(\gamma(a,b) \mid (a,b) \in M) = \sum P(\gamma(a,b), \gamma(b,a)) \cdot P(\gamma(a,b), \gamma(b,a)) \\
\text{Weights for each matching variable can be computed based on the two conditional probabilities:}
\end{align*}
\]

When a variable matches: \( W_i(y_i) = \log P_{\gamma_{a(a),b(b)}} - \log P_{\gamma_{a(a),b(b)}} \)
When a variable does not match \( W_i(y_i) = \log(1-P_{\gamma_{a(a),b(b)}}) - \log(1-P_{\gamma_{a(a),b(b)}}) \)
Calculate the matching score \( W(y) = W_1 + W_2 + \ldots + W_i \)

Step 1: Data Preparation
- Select matching variables
  - Categorical variables (Gender, Race, etc)
  - Continuous variables (Age, Height, DBP, etc)
- Combine matching datasets
- Block dataset for record pairing
  - Error minimum variables to reduce comparing pairs
  - Gender, Marital status, Race etc.
- Compute Z statistics for continuous variables
  - \( Z = \frac{\text{Pair Difference} - 0}{\text{Standard Deviation}} \)
  - \( Z < 0.1 \rightarrow \text{Match} \)
  - \( Z \geq 0.1 \rightarrow \text{Unmatch} \)
- Generate matching pair dataset
  - Match = 1; Unmatch = 0;

Step 2: Constructing vector \( Y, G \)
- \( Y = (Y_1, Y_2, \ldots, Y_n) \)
  - Indicator vector of 0.1 for \( r \)th pair and \( v \)th variable
- \( G = (G_{a(a),b(b)}) \)
- If \( n_{a(a),b(b)} = 1 \rightarrow \text{Same individual} \)
- If \( n_{a(a),b(b)} = 1 \rightarrow \text{different individual} \)
- Bayes' rule for conditional probabilities

\[
L_x = \frac{\prod_{i=1}^{k} P(y_i \mid a(a), b(b))}{\prod_{i=1}^{k} P(y_i \mid \alpha, \beta)} = a \cdot \frac{L_0}{L_1}
\]

Step 3: EM Algorithm to Validate the Probabilities
- E Step: to obtain the estimates of \( \hat{\alpha} \) and \( \hat{\beta} \)
- M Step: to update the conditional probabilities and \( \gamma \)

Step 4: Generate Matching Scores
- Compute weights
  - \( W_{\alpha(a),b(b)} = \log P_{\gamma_{a(a),b(b)}} - \log P_{\gamma_{a(a),b(b)}} \)
  - \( W_{\alpha(a),b(b)} = \log(1-P_{\gamma_{a(a),b(b)}}) - \log(1-P_{\gamma_{a(a),b(b)}}) \)
- Compute scores
  - \( y_{\alpha(a),b(b)} = \sum W_{\alpha(a),b(b)} \)
- Generate a list based on the scores

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.
A SAS Program to Identify Duplicates in Clinical Data

Xiaoli Lu
VA Cooperative Studies Program, Perry Point, MD 21902

METHODS CONTINUED

Test Data Sets
- Hypertension Survey Data
- Matching variables:
  - Gender
  - Race
  - Marital Status
  - Family History
  - Year of Diagnosis

- Hypertension Clinical Trial Data
- Matching variables:
  - Gender
  - Race
  - Marital Status
  - Family History
  - Year of Diagnosis
  - Years in School

Duplicate Record Detection: A Survey
Nora Race

Matching variables:
- Age
- Marital Status

No exact matching found

Height
Ahmed K.

Probabilistic record linkage is a valid and transparent tool to
Hypertension Type
- DBP

Deterministic results

Programming outline
* SAS program developed to implement the method
* PROC SQL to generate matching pairs
* PROC IML to compute match/mismatch probabilities

Deterministic results

Survey data
No exact matching found

RESULTS CONTINUED

Probabilistic results

Score list for the survey data

<table>
<thead>
<tr>
<th>Obs</th>
<th>No1 record</th>
<th>No2 record</th>
<th>age_m</th>
<th>height_m</th>
<th>hype_m</th>
<th>dbp_m</th>
<th>ecgv_m</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64404285</td>
<td>290</td>
<td>68903217</td>
<td>451</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4.89095</td>
</tr>
<tr>
<td>2</td>
<td>65701779</td>
<td>335</td>
<td>64403967</td>
<td>335</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4.89095</td>
</tr>
<tr>
<td>3</td>
<td>64403967</td>
<td>335</td>
<td>65701779</td>
<td>335</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4.89095</td>
</tr>
<tr>
<td>4</td>
<td>64403967</td>
<td>335</td>
<td>65701779</td>
<td>335</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4.89095</td>
</tr>
<tr>
<td>5</td>
<td>64403967</td>
<td>335</td>
<td>65701779</td>
<td>335</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4.89095</td>
</tr>
<tr>
<td>6</td>
<td>64403967</td>
<td>335</td>
<td>65701779</td>
<td>335</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4.89095</td>
</tr>
<tr>
<td>7</td>
<td>64403967</td>
<td>335</td>
<td>65701779</td>
<td>335</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4.89095</td>
</tr>
</tbody>
</table>

CONCLUSIONS

- A method developed to identify duplicates using non-sensitive information
- Couple of duplicates found in the survey data, not in the clinical trial data
- Probabilistic approach is more sensitive than the deterministic approach

REFERENCES

- Nora Meraya, Johannes B. Reitsma, Anita C.J. Ravellia, Gouke J. Bonsel. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number *Journal of Clinical Epidemiology* 60 (2007) 883-891
- John S. Lawson Record Linkage Techniques for Improving Online Genealogical Research using Census Index Records *ASA Section on Survey Research Methods* 3297-3303

* Disclaimer: The contents do not represent the views of the U.S. Department of Veterans Affairs
SAS® GLOBAL FORUM 2017
April 2 – 5 | Orlando, FL