SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

A SAS Program to Identify Duplicates in Clinical Data

USERS PROGRAM

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
* These complications often happen in clinical trials, meta-analyses, and registry and observational studies

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
  - Deterministic (All Matching variables exact match)
  - 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 α (a) and β(b) be the corresponding records for a and b. Assume κ features available to be used for the identification If a$\in$A, b$\in$B and α(a) = (α1, α2,…,$\alpha\kappa$) and β(b) = (β1, β2,…,$\beta\kappa$), the information refers to the same variables, then a comparison vector γ would be $\gamma = \gamma[\alpha(a),\beta(b)]=\{\gamma^1[\alpha(a),\beta(b)],\ldots, \gamma^\kappa[\alpha(a),\beta(b)]\} =(\gamma_1,\ldots,\gamma_\kappa)$ Assume the comparison vector $\gamma[\alpha(a),\beta(b)]$ be a random variable, the conditional probabilities producing γ given (a, b) $\in$ M and (a, b) $\in$ U are:

$$P_{M(\gamma)} = P\{\gamma[\alpha(a),\beta(b)]\,|\,(a,b)\in M\} = \sum_{(a,b)\in M} P\{\gamma[\alpha(a),\beta(b)]\}\bullet P[(a,b)\,|\,M\} \qquad P_{U(\gamma)} = P\{\gamma[\alpha(a),\beta(b)]\,|\,(a,b)\in U\} = \sum_{(a,b)\in U} P\{\gamma[\alpha(a),\beta(b)]\}\bullet P[(a,b)\,|\,U\}$$

Weights for each matching variable can be computed based on the two conditional probabilities:

When a variable matches: $W_k(\gamma_k) = \log P_{M(\gamma_k)} - \log P_{U(\gamma_k)}$ When a variable does not match $W_k(\gamma_k) = \log(1-P_{M(\gamma_k)}) - \log(1-P_{U(\gamma_k)})$

Calculate the matching score $W(\gamma) = W_1 + W_2 + \ldots + W_k$

Procedures

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 = (Pair Difference – 0) / Standard Deviation
  - Z < 0.1 → Match
  - Z ≥ 0.1 → Unmatch
- Generate matching pair dataset
  - Match = 1; Unmatch = 0;

Step 3: EM Algorithm to Validate the Probabilities

- E Step: to obtain the estimates of $\hat{g}_{rm}$ and $\hat{g}_{ru}$
- M Step: to update the conditional probabilities and $\hat{\pi}$

$$\hat{p}_{m(vt)} = \frac{\sum_{r=1}^{R}[\hat{g}_{rm}, y_{rvt}]}{\sum_{r=1}^{R}\hat{g}_{rm}} \qquad \hat{p}_{u(vt)} = \frac{\sum_{r=1}^{R}[\hat{g}_{ru}, y_{rvt}]}{\sum_{r=1}^{R}\hat{g}_{ru}} \qquad \hat{\pi} = \frac{\sum_{r=1}^{R}\hat{g}_{rm}}{R}$$

Step 2: Construct comparing vector { Y_r, G }

- $Y_r = (y_{r1}, y_{r2}, \ldots, y_{rk})$
  Indicator vector of 0,1 for r^{th} pair and v^{th} variable
- $G = (g_{rm}, g_{ru})$
  If $g_{rm} = 1 \rightarrow$ Same individual
  If $g_{ru} = 1 \rightarrow$ different individual
- Bayes' rule for conditional probabilities

$$\hat{g}_{rm} = \frac{\hat{\pi}\prod_{j=1}^{v}\prod_{i=1}^{c_v} p_{m_{ij}}^{y_{rvj}}}{\hat{\pi}\prod_{j=1}^{v}\prod_{i=1}^{c_v} p_{m_{ij}}^{y_{rvj}} + (1-\hat{\pi})\prod_{j=1}^{v}\prod_{i=1}^{c_v} p_{u_{ij}}^{y_{rvj}}} \qquad \hat{g}_{ru} = \frac{(1-\hat{\pi})\prod_{j=1}^{v}\prod_{i=1}^{c_v} p_{u_{ij}}^{y_{rvj}}}{\hat{\pi}\prod_{j=1}^{v}\prod_{i=1}^{c_v} p_{m_{ij}}^{y_{rvj}} + (1-\hat{\pi})\prod_{j=1}^{v}\prod_{i=1}^{c_v} p_{u_{ij}}^{y_{rvj}}}$$

Step 4: Generate Matching Scores

- Compute weights

$$W_{(v)Match} = \log(p_{m(vi)}) - \log(p_{u(vi)})$$

$$W_{(v)Differ} = \log(1-p_{m(vi)}) - \log(1-p_{u(vi)})$$

- Compute scores

$$S_{(r)} = \sum_{v=1}^{k} W_{(v)}$$

- Generate a list based on the scores

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
    - Age
    - Medication
    - Height
    - DBP
    - Years in School
- ***Hypertension Clinical Trial Data***
  - Matching variables:
    - Gender
    - Race
    - Hypertension Type
    - Height
    - Age
    - DBP

Programming outline

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

RESULTS

Deterministic results

Clinical trial data

Obs	No1	record	No2	record2	age_m	height_m	hype_m	dbp_m	ecgv_m
1	64404285	290	68903217	451	1	1	1	1	1

Survey data

No exact matching found

RESULTS CONTINUED

Probabilistic results

Score list for the survey data

Obs	No1	record	No2	record2	on_meds_m	height_m	yr_sch_m	dbp_m	mh_10_m	score
1	5841605	84	5841606	85	1	1	1	1	3	4.11795
2	5840711	175	5841028	176	1	1	1	1	3	4.11795
3	5841652	314	5841711	317	1	1	1	1	3	4.11795
4	5841652	314	5841710	316	1	1	1	1	3	4.11795
5	5841652	314	5841704	315	1	1	1	1	3	4.11795
6	5841704	315	5841711	317	1	1	1	1	3	4.11795
7	5841704	315	5841710	316	1	1	1	1	3	4.11795
...										
157	5830345	4924	5860667	4925	1	0	1	3	3	1.72507
158	5860896	1715	5860953	1716	0	0	3	1	3	1.26986
159	6570376	1034	6571118	1036	1	0	0	1	1	0.67963
160	5861157	3356	6570538	3358	1	0	0	1	1	0.67963
161	5840059	216	5840060	217	3	1	3	3	3	0.14488
162	5830730	4374	5841152	4375	3	1	3	3	3	0.14488

Score list for the clinical trial data

Obs	No 1	record	No2	record2	age_m	height_m	hype_m	dbp_m	ecgv_m	score
1	51204844	46	61803889	154	0	1	1	1	1	4.89095
2	58604548	83	65700478	356	0	1	1	1	1	4.89095
3	61801995	135	65703710	391	0	1	1	1	1	4.89095
4	61803625	152	65701810	371	0	1	1	1	1	4.89095
5	63001922	183	65703567	390	0	1	1	1	1	4.89095
6	64400395	235	64405756	305	0	1	1	1	1	4.89095
7	64400949	241	64409658	342	0	1	1	1	1	4.89095
8	64402091	261	64403967	285	0	1	1	1	1	4.89095
9	64404285	290	68903217	451	1	1	1	1	1	4.89095
10	64409288	335	65701779	369	0	1	1	1	1	4.89095
...										
86	64401617	253	68905654	475	0	0	0	1	1	4.77301
87	64402112	262	65700082	348	0	0	0	1	1	4.77301
88	64402342	269	68901871	430	0	0	0	1	1	4.77301
89	65700004	344	68905209	470	0	0	0	1	1	4.77301
90	65701616	368	65702695	384	0	0	0	1	1	4.77301

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

- Fellegi, I. P., and A. B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association* (1969)64, 1183-1210.
- Nora Meraya, Johannes B. Reitsmab, Anita C.J. Ravellia, Gouke J. Bonselc 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
- L. J. Cook, L. M. Olson, J. M. Dean Probabilistic Record Linkage: Relationships between File Sizes, Identifiers, and Match Weights *Methods of Information in Medicine* (2001); 40: 196–203
- Ahmed K. Elmagarmid Duplicate Record Detection: A Survey *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* (2007) 19:1 : 1-16
- 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