Generating Synthetic Population Data for Testing Record Linkage Procedures

Sepideh Mosaferi

National Cancer Institute (NCI) Presentation
University of Maryland College Park

August 25, 2016

OUTLINE

- Why generate synthetic cancer population data?
- How to generate synthetic cancer population data?
- Evaluating Febrl using generated synthetic cancer population data... This will be shown by running the software.
- Summary and Discussion

Background and Motivation

- Linking the SEER cancer registry data with different administrative or commercial data sources through probabilistic software becomes an interested area within the SRP at NCI
- Variety of free accessible linkage software such as LinkPlus (developed by CDC), BigMatch (developed by Census Bureau) and Febrl (developed by the Austria group) exist
- Challenges arise in choosing a suitable and reliable linkage software and evaluating the linkage quality
- Evaluations using real data have restrictions due to unknown truth, limited data accessibility to the patient health identifiers, and other restrictions related to the divisions and institutes requirements in sharing the data set
- An alternative choice is generating synthetic (artificial) but representative data sets based on information from a real data

Background and Motivation (Cont'd)

Using Synthetic data enables researchers to:

- Fully and consistently compare the performance of several record linkage software or other procedures because the truth is known
- Bring the possibility of different research when the unit level information is not available, but the population characteristics are accessible

What Is Synthetic Data Set?

Any production of data applicable to a given situation which is not obtained by a direct measurement.

Characteristics of Generated Data Set:

- depending upon the purpose of researcher
- exhibiting similar statistical characteristics to the real data
- preserving the frequency distributions of attributes, the occurrences and frequencies of typographical and other errors and variations
- preserving the dependencies or relationship among elements of attributes in the real data

Approaches to Generate Synthetic Data Set for Different Purposes

- For protecting confidentiality: Partially or fully replacing sensitive original values with synthetic ones by perturbing the real data through an explicit statistical model and preserving the original statistical inferences. For example, generate synthetic census tracks for SEER to preserve confidentiality.
- For record linkage: Generate data based on a specified distribution (Uniform, Poisson, Zipf) while explicitly considering some appropriate assumptions such as attribute frequencies and possible errors occur in the real data.
    - □ Attributions need to be generated are mainly patient health identifications such as SSN, first name, last name, date of birth, mail address, phone number, etc.

Data Generator Software for the Record Linkage Purpose

- DBGEN (A Database Generator): in C language: This software is reasonably good for the purpose of generating US mailing addresses; it is not appropriate for generating the cancer population.
  http://www.cs.utexas.edu/users/ml/riddle/data.html
- GeCo (A Data Generator and Corruptor): in Python language: A good choice for generating a synthetic data which is capable of receiving frequency tables, dependent attributes, errors, and cancer attributes by adding self-written codes, frequency tables, etc. to mimic a real cancer population.
  http://dmm.anu.edu.au/geco/

Generated Cancer Population Data Set by (GeCo)

For the NCI synthetic data

N=20,000: number of records

C=10,000: number of original records and duplicates

Duplicates are randomly distributed among original individuals. This does not mean each individual should be necessarily to be repeated one time; some individuals might not be duplicated and some others might just be duplicated one or two times depends upon the maximum value for the number of duplication, which is 3 in our case.

The maximum modifications per attribute is 1, and the number of modifications per record is 5 based on our definition.

Generated Attributes

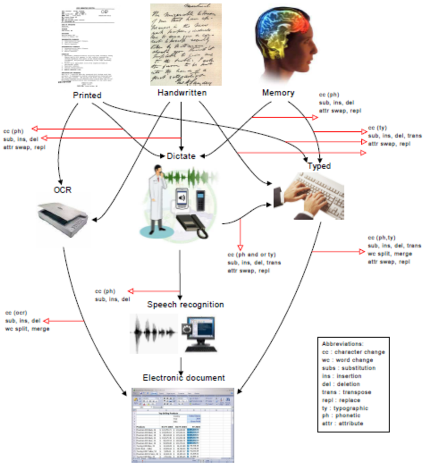
- First name, Last name, Race, City state, Zip code (based on frequency tables obtained from public website)
- Date of birth, SSN, Phone number (generated using computer code assuming certain distribution)
- Four cancer depend attributes (based on frequency tables from public cancer website):
    - □ Cancer site cross by Gender; Cancer site cross by State; Cancer site cross by Ethnicity; Cancer site cross by Age-group

Incorporated Errors in the Generated Data set (GeCo)

Usual errors in administrative records are considered:

- ocr-variation error: optical character recognition error; e.g.: l (the letter) instead of 1 (the number)
- qwerty-keyboard error: errors during typing by PC
- phonetic-variations error: speech sound disorders over the phone
- given-name misspell and surname misspell
- missing values
- swapping: two fields contents replaced
- insertion, deletion, substitution, and transposition

Different Sources of Errors

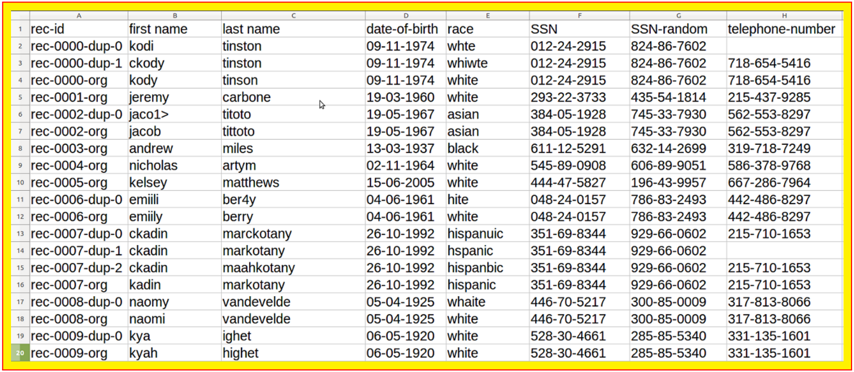


Abbreviations:
cc : character change
wc : word change
subs : substitution
ins : insertion
del : deletion
trans : transpose
repl : replace
ty : typographic
ph : phonetic
attr : attribute

SIX Different Scenarios Considered as

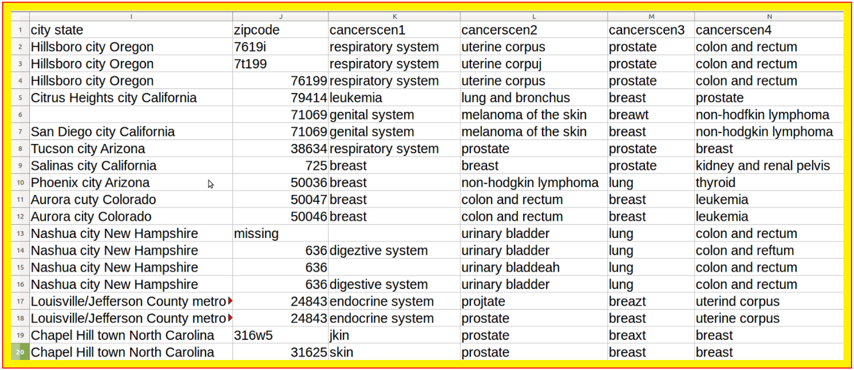
1. Uniform distribution for generating duplication and assigning more errors to the attributes (dirty situation I).
2. Uniform distribution for generating duplication and assigning less errors to the attributes (clean situation I).
3. Poisson distribution for generating duplication and assigning more errors to the attributes (dirty situation II).
4. Poisson distribution for generating duplication and assigning less errors to the attributes (clean situation II).
5. Zipf distribution for generating duplication and assigning more errors to the attributes (dirty situation III).
6. Zipf distribution for generating duplication and assigning less errors to the attributes (clean situation III).

Example of the Generated Data



	A	B	C	D	E	F	G	H
1	rec-id	first name	last name	date-of-birth	race	SSN	SSN-random	telephone-number
2	rec-0000-dup-0	kodi	tinston	09-11-1974	whte	012-24-2915	824-86-7602	
3	rec-0000-dup-1	ckody	tinston	09-11-1974	whiwte	012-24-2915	824-86-7602	718-654-5416
4	rec-0000-org	kody	tinson	09-11-1974	white	012-24-2915	824-86-7602	718-654-5416
5	rec-0001-org	jeremy	carbone	19-03-1960	white	293-22-3733	435-54-1814	215-437-9285
6	rec-0002-dup-0	jaco1>	titoto	19-05-1967	asian	384-05-1928	745-33-7930	562-553-8297
7	rec-0002-org	jacob	tittoto	19-05-1967	asian	384-05-1928	745-33-7930	562-553-8297
8	rec-0003-org	andrew	miles	13-03-1937	black	611-12-5291	632-14-2699	319-718-7249
9	rec-0004-org	nicholas	artym	02-11-1964	white	545-89-0908	606-89-9051	586-378-9768
10	rec-0005-org	kelsey	matthews	15-06-2005	white	444-47-5827	196-43-9957	667-286-7964
11	rec-0006-dup-0	emiili	ber4y	04-06-1961	hite	048-24-0157	786-83-2493	442-486-8297
12	rec-0006-org	emiily	berry	04-06-1961	white	048-24-0157	786-83-2493	442-486-8297
13	rec-0007-dup-0	ckadin	marckotany	26-10-1992	hispanuic	351-69-8344	929-66-0602	215-710-1653
14	rec-0007-dup-1	ckadin	markotany	26-10-1992	hspanic	351-69-8344	929-66-0602	
15	rec-0007-dup-2	ckadin	maahkotany	26-10-1992	hispanbic	351-69-8344	929-66-0602	215-710-1653
16	rec-0007-org	kadin	markotany	26-10-1992	hispanic	351-69-8344	929-66-0602	215-710-1653
17	rec-0008-dup-0	naomy	vandevelde	05-04-1925	whaite	446-70-5217	300-85-0009	317-813-8066
18	rec-0008-org	naomi	vandevelde	05-04-1925	white	446-70-5217	300-85-0009	317-813-8066
19	rec-0009-dup-0	kya	ighet	06-05-1920	white	528-30-4661	285-85-5340	331-135-1601
20	rec-0009-org	kyah	highet	06-05-1920	white	528-30-4661	285-85-5340	331-135-1601

Example of the Generated Data



	I	J	K	L	M	N
1	city state	zipcode	cancerscen1	cancerscen2	cancerscen3	cancerscen4
2	Hillsboro city Oregon	7619i	respiratory system	uterine corpus	prostate	colon and rectum
3	Hillsboro city Oregon	7t199	respiratory system	uterine corpuj	prostate	colon and rectum
4	Hillsboro city Oregon	76199	respiratory system	uterine corpus	prostate	colon and rectum
5	Citrus Heights city California	79414	leukemia	lung and bronchus	breast	prostate
6		71069	genital system	melanoma of the skin	breawt	non-hodfkin lymphoma
7	San Diego city California	71069	genital system	melanoma of the skin	breast	non-hodgkin lymphoma
8	Tucson city Arizona	38634	respiratory system	prostate	prostate	breast
9	Salinas city California	725	breast	breast	prostate	kidney and renal pelvis
10	Phoenix city Arizona	50036	breast	non-hodgkin lymphoma	lung	thyroid
11	Aurora cuty Colorado	50047	breast	colon and rectum	breast	leukemia
12	Aurora city Colorado	50046	breast	colon and rectum	breast	leukemia
13	Nashua city New Hampshire	missing		urinary bladder	lung	colon and rectum
14	Nashua city New Hampshire	636	digeztive system	urinary bladder	lung	colon and reftum
15	Nashua city New Hampshire	636		urinary bladdeah	lung	colon and rectum
16	Nashua city New Hampshire	636	digestive system	urinary bladder	lung	colon and rectum
17	Louisville/Jefferson County metro	24843	endocrine system	projtate	breazt	uterind corpus
18	Louisville/Jefferson County metro	24843	endocrine system	prostate	breast	uterine corpus
19	Chapel Hill town North Carolina	316w5	jkin	prostate	breaxt	breast
20	Chapel Hill town North Carolina	31625	skin	prostate	breast	breast

Evaluating Febrl Using Generated Synthetic Cancer Population Data

□ Febrl (Freely Extensible Biomedical Record Linkage) is a software for data cleaning, deduplication and record linkage in Python code.

□ Since we have created true identifier in the synthesized data set, we are able to check the quality of linkage via *OptimalThreshold* classifier which requires the true match status of all compared record pairs to be known (i.e. it is a supervised classifier). We evaluate both deduplication part and record linkage part of the software.

□ We chose Febrl for this study within the limited time because Febrl has not been explored by the NCI researchers before and it was an often cited software in the record linkage literature.

Entity Resolution via Febrl

Step One: Blocking

We have adopted two standard sequential blocking strategies as follows:

Last name (Double-Metaphone)	**Block 1**
Zip code (without putting any constraints) and First Name (Soundex with three parameters)	**Block 2**

Records are not compared if they disagreed on the first set of blocking item and also disagreed on one or two of the second set of blocking items.

Entity Resolution via Febrl (Cont'd)

Step Two: Field Comparison

We have considered 8 attributes:

- First name and Last name: Winkler method
- Race: Q-gram with parameter length of Q 2 and common divisor of average
- Zip code: Key-diff with maximum difference of 1 between the digits of two fields for pair record
- Date of birth: Date method which compares a pair of dates and we considered the maximum difference of 5 days before and after of two dates for the error tolerance.

Entity Resolution via Febrl (Cont'd)

Step Two: Field Comparison

The remained attributes:

- City-state: Q-gram with parameter length of Q 2 and common divisor of average.
- Telephone: Key-diff with maximum difference of 2 digits.
- Cancer attribute: Winkler

Entity Resolution via Febrl (Cont'd)

Step Three: Classification

For weight vector (or comparison weight) classification of each pair, we have considered the Fellegi-Sunter algorithm which usually has been used by practitioners and in software.

For employing this algorithm, we need to define an upper and lower threshold.

We assigned two values 3 and 5 for the lower and upper thresholds based on the trial-error method.

Criteria for Evaluating Linkage Software

linkage quality: Accuracy, Precision, Recall, and F-measure.

linkage complexity: Reduction Ratio, Pairs Completeness, and Pairs Quality.

Accuracy =(TP+TN)/(TP + FP + TN + FN)
Precision=(TP)/(TP+FP)
Recall= (TP)/(TP+FN)
F-measure= 2 × (precision × recall)/(precision + recall)
Reduction Ratio=1- $\{N_b/|A| \times |B|\}$
Pairs Completeness=N_m/M
Pairs Quality=TP/(number of given weight vectors)

Summary and Discussion

- We have explored different tools to generate synthetic cancer population for record Linkage purpose;
- The GeCo software with modified codes is readily available to adopt for generating synthetic cancer population based on different true information;
- The generated synthetic data can be used to fully evaluate different record linkage software;
- Our evaluation indicates that the linkage quality of Febrl may not be good when we have large data set with complicated fields which exist in reality;
- It could be beneficial to compare the quality of Febrl with other linkage software such BigMatch and LinkPlus.
- For evaluating different linkage software, one needs to consider the same method including blocking, classifying, etc; across the software.

Thank You!

Questions: smosafer@umd.edu