Evaluating Record Linkage Software Using Representative Synthetic Datasets

Benmei Liu, Ph.D.
Mandi Yu, Ph.D.
Eric J. Feuer, Ph.D.

National Cancer Institute

Presented at the NAACCR Annual Conference
June 14th, 2018
Pittsburgh, PA
Outline

- Motivation
- Synthetic data generation
- Record linkage software and comparison
- Summary and discussion
• The NCI's Surveillance Epidemiology and End Results (SEER) program has been increasingly engaged in initiatives extensively utilizing record linkage techniques to capture additional medical information (e.g., treatment information, genetic tests, etc.) that cannot be obtained through traditional medical record abstraction.

• Variety of linkage software products exist and some are free accessible. In addition, to fulfill the linkage need by the Virtual Pooled Registry System, a new linkage software Match*Pro has been recently developed by the Information Management Services, Inc. for the NCI.

• Evaluations of the software using real data have restrictions due to unknown truth and limited data accessibility to the patient health identifiers beyond other restrictions related to the divisions and institutes requirements in sharing the dataset. Synthetic but representative datasets may facilitate the evaluation.
Study Goal

• To systematically test the usability of the new software Match*Pro 1.0 and compare it with CDC’s LinkPlus using representative synthetic datasets.
What is Synthetic data?

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

Characteristics of Generated Dataset:

– 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
For Record Linkage: Generate data based on specified distribution (uniform, Poisson, Zipf) while explicitly considering some appropriate assumptions such as attribute frequencies and possible errors occur in the real data.

- The data need to be generated are patient health identifiers such as SSN, first name, last name, gender, date of birth, mail address, phone number, etc.
GeCo (A Data Generator and Corruptor), in Python language:

A good choice for generating a synthetic data which is capable of receiving attribute frequency tables, dependent attributes, errors, and additional attributes by adding self-written codes, frequency tables, etc. to mimic a real cancer population.

- Reference:


Data generator steps:

- **Step 1**: Create user specified number of original records using look-up tables with real values and their frequencies and dependencies, or based on specific attribute generation rules.

- **Step 2**: Randomly select original records, modify with potential errors and store as duplicate records.

---

**Generate Synthetic Population Database Using GeCo**

**Attribute Generation Rules**

**Frequency Tables**

**Typographic Error Functions**

**Phonetic Error Rules**

**OCR Error Rules**

---

**Generate Original Records**

**Original Records**

**Generate Duplicate Records**

**Duplicate Records**

**Final Dataset**

---

- Generate Original Records
- Frequency Tables
- Typographic Error Functions
- Phonetic Error Rules
- OCR Error Rules
- Final Dataset

---

**Generate Synthetic Population Database Using GeCo**

**Attribute Generation Rules**

**Frequency Tables**

**Typographic Error Functions**

**Phonetic Error Rules**

**OCR Error Rules**

---

**Generate Original Records**

**Original Records**

**Generate Duplicate Records**

**Duplicate Records**

**Final Dataset**

---

- Generate Original Records
- Frequency Tables
- Typographic Error Functions
- Phonetic Error Rules
- OCR Error Rules
- Final Dataset
Errors incorporated in Data Generation

- Typographic errors at character level (insertion, deletion, swapping, substitution of a character, transposition of two adjacent characters)
- Optical Character Recognition (OCR) modifications are based on rules considering shape similarity among characters (e.g., ‘5’ and ‘s’, ‘6’ and ‘G’, ‘I’ and ‘1’)
- Phonetic errors include 7 components: Position, original pattern, substitute pattern, precondition, post condition, pattern existence condition, start existence condition
Two sets of data with different error rates (~1% and 15%):

- Each contains 400,000 total records with the following attributes:
  Unique record ID, first name, last name, date of birth, gender, race, SSN, phone number, city-state, and zipcode.

- Randomly split each dataset into two sub files with similar size.

- The Original record ID (call it TrueID) can be derived from the unique record ID and will be used to verify the true matches.

- After deduplication using TrueID:
  - Clean (1% error) File 1 contains 130,619 records;
  - Clean (1% error) File 2 contains 130,944 records.
  - True matched pairs are 61,563.

  - Dirty (15% error) File 1 contains 130,633 records;
  - Dirty (15% error) File 2 contains 130,628 records.
  - True matched pairs are 61,291.
### Display of the Generated Data in One File

<table>
<thead>
<tr>
<th>TrueID</th>
<th>Unique record ID</th>
<th>first name</th>
<th>last name</th>
<th>Race</th>
<th>city_state</th>
<th>Zipcode</th>
<th>date of birth</th>
<th>SSN</th>
<th>SSN_random</th>
<th>Phone number</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>000000</td>
<td>rec-000001-org</td>
<td>annabelle</td>
<td>lease</td>
<td>white</td>
<td>Newton city Massachusetts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>000001</td>
<td>rec-000001-dup-0</td>
<td>lucka</td>
<td>ang</td>
<td>white</td>
<td>Augusta-Richmond Conty consolidated go</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>000002</td>
<td>rec-000002-org</td>
<td>lauren</td>
<td>hodor</td>
<td>white</td>
<td>Visalia city California</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>000003</td>
<td>rec-000003-org</td>
<td>james</td>
<td>lubulva</td>
<td>white</td>
<td>Kent city Washington</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>000004</td>
<td>rec-000004-org</td>
<td>mazarieah</td>
<td>claire</td>
<td>white</td>
<td>Huntington Park city California</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>000005</td>
<td>rec-000005-org</td>
<td>harrison</td>
<td>deniss</td>
<td>white</td>
<td>Anaheim city California</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>000006</td>
<td>rec-000006-dup-1</td>
<td>apey</td>
<td>vep</td>
<td>white</td>
<td>Coral Gables city Florida</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>000007</td>
<td>rec-000007-dup-0</td>
<td>levy</td>
<td>wald</td>
<td>asian</td>
<td>Philadelphia city Pennsylvania</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>000011</td>
<td>rec-000011-org</td>
<td>jared</td>
<td>coulson</td>
<td>black</td>
<td>Los Angeles city California</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26033</td>
<td>02-05-1947</td>
<td>300-68-7518</td>
<td>600-48-1867</td>
<td>727-860-5406</td>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>missing</td>
<td>26-03-1983</td>
<td>073-68-2255</td>
<td>516-52-5992</td>
<td>725-165-8000</td>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17754</td>
<td>02-11-1979</td>
<td>250-67-5899</td>
<td>272-70-1206</td>
<td>814-827-9063</td>
<td>male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21035</td>
<td>30-10-2008</td>
<td>387-28-3190</td>
<td>803-76-9533</td>
<td>919-973-8767</td>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59349</td>
<td>26-02-1970</td>
<td>580-66-6304</td>
<td>290-41-4575</td>
<td>731-562-4752</td>
<td>male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9758</td>
<td>25-08-1967</td>
<td>540-67-2774</td>
<td>937-02-7850</td>
<td>610-773-8440</td>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60647</td>
<td>22-05-1931</td>
<td>318-67-9806</td>
<td>800-22-0980</td>
<td>863-921-4640</td>
<td>male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25665</td>
<td>27-08-1969</td>
<td>525-35-7137</td>
<td>293-72-9557</td>
<td>719-345-4939</td>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Attributes:**
- TrueID, Unique record ID, first name, last name, Race, city_state, Zipcode, date of birth, SSN, SSN_random, Phone number, gender.
Record Linkage Software Used in Cancer Registries

- **LinkPlus version 2.0**
  - Developed by CDC in support of CDC’s National Program of Cancer Registries (NPCR)
  - Free download from: https://www.cdc.gov/cancer/npcr/tools/registryplus/lp_tech_info.htm

- **MatchPro 1.0 (Just been released this month)**
  - Developed by IMS for NCI in support of the Virtual Pooled registry Cancer Linkage System
  - Free download from: https://surveillance.cancer.gov/matchpro/download
Features Comparison of Match*Pro and LinkPlus

- Both software use probabilistic linkage and require fixed-width or delimited format text files as input, and both use GUI. Match*Pro also accepts csv files as input.

- LinkPlus can only load up to 30,000 records each time for manual review. After manual review, the results can be exported to a csv file. The manual review feature may become nonresponding for large data files, though the matching results will automatically be saved into a text file.

- Match*Pro is designed to handle large datasets and has better features in manual review. It can export the manual review results into csv file along with a sas program to read the csv file into sas.

- Match*Pro has more flexibility in setting the blocking parameters and have more choices on matching methods. It supposes to run very fast for automated linkage process with huge data files.
Record Linkage Setup

- **Blocking parameters**: first name and last name soundex; zipcode (first 4 characters), date of birth (last 4 characters), phone number (first 3 characters)

- **Matching parameters**: first name, last name, gender, date of birth, zipcode, city_state, phone, SSN

- **Cut off value**: 1
LinkPlus Snapshot
Match*Pro Snapshot
Record Linkage Results

- Data with 1% error rates (each input file contains ~130k records, Match*Pro blocking sensitivity=2):

<table>
<thead>
<tr>
<th>Software</th>
<th>true matches cutoff</th>
<th>total matched pairs</th>
<th>True Positive</th>
<th>False Positive</th>
<th>False Negative</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>running time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MatchPro</td>
<td>61,56 3</td>
<td>70,57 7</td>
<td>60,238</td>
<td>10,33 9</td>
<td>1,325</td>
<td>0.854</td>
<td>0.978</td>
<td>0.912</td>
<td>~4 min</td>
</tr>
<tr>
<td></td>
<td>61,56 &gt;=14. 5</td>
<td>60,16 6</td>
<td>59,748</td>
<td>418 1,815</td>
<td>0.993</td>
<td>0.971</td>
<td>0.982</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LinkPlus</td>
<td>61,56 3</td>
<td>62,75 8</td>
<td>61,563</td>
<td>1,195 0</td>
<td>0.981</td>
<td>1.000</td>
<td>0.990</td>
<td></td>
<td>~12 mins</td>
</tr>
<tr>
<td></td>
<td>61,56 &gt;=10</td>
<td>61,51 2</td>
<td>61,502</td>
<td>10 61</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Record Linkage Results

- Data with 15% error rates (input file contains ~130k records, Match*Pro blocking sensitivity=2):

<table>
<thead>
<tr>
<th>Software</th>
<th>true matches</th>
<th>cutoff</th>
<th>total matched pairs</th>
<th>true positives</th>
<th>false positives</th>
<th>false negatives</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match*Pro</td>
<td>61,291</td>
<td>&gt;=14.5</td>
<td>54,837</td>
<td>54,289</td>
<td>548</td>
<td>7,002</td>
<td>0.990</td>
<td>0.886</td>
<td>0.935</td>
</tr>
<tr>
<td>LinkPlus</td>
<td>61,291</td>
<td>&gt;=10</td>
<td>60,317</td>
<td>60,307</td>
<td>10</td>
<td>984</td>
<td>1.000</td>
<td>0.984</td>
<td>0.992</td>
</tr>
</tbody>
</table>
Summary and Next Steps

- Synthetic data provides a useful data source for testing record linkage software.
- Both Match*Pro and LinkPlus produce very good linkage quality.
- The quality of the final record linkage results may depend on user’s pre-set up value of the cutoff point and user chosen blocking variables and matching methods.
- Match*Pro has the advantage of handling huge datasets.
- Will do further testing on larger data sets (>=1 million records in each input file, may take several hours to 10+ hours to run) and vary the set up.
- Will test the deduplication quality of both software.
Acknowledgement

- Sepideh Mosaferi, University of Maryland
- Will Howe, IMS
- Andy Lake, IMS
Thank you!

Contact info:
Benmei Liu
liub2@mail.nih.gov