Comparison of Record Linkage Software for De-duplicating Patient Identities in California’s Prescription Drug Monitoring Program

Susan Stewart
Division of Biostatistics
Department of Public Health Sciences
November 2019
Objectives

1. Understand the importance of accurate record linkage in a prescription drug monitoring program.
2. Become familiar with methods to evaluate the accuracy of record linkage software.
3. Know which patient metrics are most affected by the use of specific record linkage software.
Background

• Poisoning: leading cause of injury death in US:
• Drugs: cause of most poisoning deaths
  – Both pharmaceutical and illicit
• Drug-poisoning death rates more than tripled from 1999-2016

NCHS Fact Sheet, October 2018

1Significant increasing trend from 1999 to 2016 with different rates of change over time, p < 0.001.

22016 rate for male was significantly higher than for female, p < 0.001.


1Significant increasing trend from 1999 to 2016 with different rates of change over time, p < 0.05.
2Significant increasing trend from 1999 to 2006, then decreasing trend from 2006 to 2016, p < 0.05.
Prescription Drug Monitoring Program (PDMP)

- Statewide registry of dispensed prescriptions
  - Includes controlled substances
  - Implemented in 49 states
  - Can be checked by prescribers and pharmacists

- California’s PDMP
  - Started in 1939
  - Current version: Controlled Substance Utilization and Review System (CURES)
Significance

• PDMP data can be used to prevent overdose deaths
  – By identifying potentially risky prescribing and dispensing patterns and outlier patient behavior
  – By monitoring potentially risky population trends

• Therefore, accurate linkage of PDMP records is essential
- Patient entity resolution is performed in CURES to provide the following features:
  - Patient safety alerts to prescribers (new alerts produced daily)
  - De-identified data for researchers

- CURES receives approximately 155K new prescription records daily.

- With this new data, the analytics engine must reconcile patient, prescriber, and dispenser entities across the 1TB database every night.
- Once the data is de-duplicated nightly, the analytics engine identifies the resolved persons’ current prescriptions based on date filled and number of days supply.

- The resolved persons’ current prescription medicinal therapy levels are calculated and compared against pre-established thresholds.

- Therapy levels exceeding those thresholds trigger Patient Safety Alerts to current prescribers.
- The de-duplicated data also contributes to the quarterly and annual systematic production of 58 California county and one statewide de-identified data sets for use by public health officers and researchers.

- This data enables counties to
  - calculate current rates of prescriptions,
  - examine variations within the state, and
  - track the impact of safe prescribing initiatives.
- CURES is a “home grown” PDMP system. This means that the CA PDMP has full access and visibility to how the CURES system operates and functions. After employing a custom-built entity resolution methodology, the CA PDMP wanted to have its de-duplication approach evaluated.

- One of the purposes of the evaluation is to help inform the CA PDMP on areas for strength and weakness. The CA PDMP plans to pursue implementing improvements in this challenging area.
CURES Record Linkage Evaluation Project

• Collaborators
  – California DOJ: Mike Small, Tina Farales
  – UC Davis: Garen Wintemute, Stephen Henry
  – California Dept. of Public Health: Steve Wirtz

• Funding
  – Bureau of Justice Assistance: 2015-PM-BX-K001
  – CDC: U17CE002747
Goal

• Compare record linkage programs with respect to
  – Accuracy in de-duplicating a subset of patient identities
  – Identification of excessive opioid use and outlier behavior

• Challenges
  – No unique patient identifier
  – Variation in identity fields for an individual
  – Hundreds of millions of records
### Example

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Sex</th>
<th>DOB</th>
<th>Address</th>
<th>Zip Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stephen</td>
<td>Henry</td>
<td>Male</td>
<td>05/11/77</td>
<td>2450 48th Street</td>
<td>95817</td>
</tr>
<tr>
<td>Steven</td>
<td>Henry</td>
<td>Male</td>
<td>05/11/77</td>
<td>2450 48th St.</td>
<td>95817</td>
</tr>
<tr>
<td>Henry</td>
<td>Stevens</td>
<td>Male</td>
<td>11/05/77</td>
<td>2450 48th St., Apt. 2</td>
<td>95817</td>
</tr>
<tr>
<td>Steve</td>
<td>Henry</td>
<td>Male</td>
<td>05/11/87</td>
<td>2405 48th Street</td>
<td>95807</td>
</tr>
</tbody>
</table>

Are these the same person?
Methods
Compare Record Linkage Programs

- CURES 2.0 custom-built program
  - SAS application
- The Link King: [http://www.the-link-king.com/index.html](http://www.the-link-king.com/index.html)
  - SAS application
  - Microsoft Windows stand-alone application
  - Microsoft Access application
Approach

• Start with exact matching of prescription record identifiers
  – Decreases size to ~60 million records

• Link within smaller geographic areas
  – Test dataset: patient identities for prescriptions filled in 2013 in 2 zip3s
    • 1 in Northern California, 1 in Southern California
    • ~500,000 records
Entity resolution

1) Compare pairs of records to determine whether they match
2) Assign a score to indicate match quality
3) Determine which records correspond to the same entity based on match results
Fields Available to Match

- First name
- Last name
- Date of birth
- Gender
- Address
  - Street address
  - City
  - Zip code (5 digits)
Manual Review

• Matches identified by one or more of the programs at any level of certainty were included in the full dataset of paired records
• Paired records were stratified by level of certainty
  – From high to low confidence in a match
• 5 reviewers inspected a stratified random sample of 720 paired records
  – Blinded to software certainty ratings
  – “Truth” determined by majority opinion
Statistical Analysis

• Assessed accuracy of software using stratified sample weighted to full set of paired records
  – **Sensitivity**: proportion of true matches identified by the program (aka recall)
  – **Positive predictive value**: proportion of identified matches that are true matches (aka precision)
• Determined the optimal cut-point distinguishing between matches and non-matches for each program
• Assessed relative importance of specific identity fields in distinguishing matches from non-matches by each program
• Computed PDMP patient alerts and CDC metrics for the patient entities identified by each program
Results

- Total of 365,503 record pairs identified as possible matches by at least one program from a sample of 557,861 identity records
  - Total pairs = \( \binom{557,861}{2} \) = 155.6 billion

<table>
<thead>
<tr>
<th>Software</th>
<th>Possible Matched Pairs (initially identified)</th>
<th>Patient Entities (using optimal cut-point)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custom-built</td>
<td>97,695</td>
<td>482,786</td>
</tr>
<tr>
<td>The Link King</td>
<td>122,884</td>
<td>467,454</td>
</tr>
<tr>
<td>Link Plus</td>
<td>363,590</td>
<td>452,116</td>
</tr>
<tr>
<td>LinkSolv</td>
<td>130,017</td>
<td>460,594</td>
</tr>
</tbody>
</table>
Agreement between Record Linkage Software and Manual Review

<table>
<thead>
<tr>
<th>Software</th>
<th>PPV (%)</th>
<th>95% CI</th>
<th>Sensitivity (%)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custom-built</td>
<td>Est.</td>
<td>94.9%</td>
<td>94.1-95.7%</td>
<td>73.0%</td>
</tr>
<tr>
<td>The Link King</td>
<td>97.9%</td>
<td>96.7-99.2%</td>
<td>94.8%</td>
<td>93.8-95.8%</td>
</tr>
<tr>
<td>Link Plus</td>
<td>93.5%</td>
<td>92.3-94.7%</td>
<td>83.6%</td>
<td>81.5-85.8%</td>
</tr>
<tr>
<td>LinkSolv</td>
<td>93.1%</td>
<td>91.7-94.5%</td>
<td>95.3%</td>
<td>94.8-95.8%</td>
</tr>
</tbody>
</table>

Note: CI=confidence interval; PPV=positive predictive value
Match by manual review: at least 3 of 5 reviewers rated pair as probably or definitely the same person
Importance of Date of Birth

Percent of Paired Identities with the Same DOB by Match Status

- Manual Review
- Custom-built
- The Link King
- Link Plus
- LinkSolv

- Match
- Non-Match
Importance of Last Name

Percent of Paired Identities with the Same Last Name by Match Status

- Manual Review
- Custom-built
- The Link King
- Link Plus
- LinkSolv

- Match
- Non-Match
Importance of Zip Code

Percent of Paired Identities with the Same Zip Code by Match Status

- Manual Review
- Custom-built
- The Link King
- Link Plus
- LinkSolv

Match | Non-Match
# Number of Patient Alerts

<table>
<thead>
<tr>
<th>PDMP Alert Scenario</th>
<th>Software</th>
<th>Patient Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Currently prescribed &gt;90 MMEs/day</td>
<td>Custom-built</td>
<td>3426</td>
</tr>
<tr>
<td></td>
<td>The Link King</td>
<td>3434</td>
</tr>
<tr>
<td></td>
<td>Link Plus</td>
<td>3444</td>
</tr>
<tr>
<td></td>
<td>LinkSolv</td>
<td>3435</td>
</tr>
<tr>
<td>Obtained prescriptions from ≥6 prescribers or ≥6 pharmacies in last 6 months</td>
<td>Custom-built</td>
<td>1993</td>
</tr>
<tr>
<td></td>
<td>The Link King</td>
<td>2211</td>
</tr>
<tr>
<td></td>
<td>Link Plus</td>
<td>2524</td>
</tr>
<tr>
<td></td>
<td>LinkSolv</td>
<td>2329</td>
</tr>
<tr>
<td>Currently prescribed opioids &gt;90 consecutive days</td>
<td>Custom-built</td>
<td>3039</td>
</tr>
<tr>
<td></td>
<td>The Link King</td>
<td>3138</td>
</tr>
<tr>
<td></td>
<td>Link Plus</td>
<td>3097</td>
</tr>
<tr>
<td></td>
<td>LinkSolv</td>
<td>3140</td>
</tr>
<tr>
<td>Currently prescribed both benzodiazepines and opioids</td>
<td>Custom-built</td>
<td>2923</td>
</tr>
<tr>
<td></td>
<td>The Link King</td>
<td>2955</td>
</tr>
<tr>
<td></td>
<td>Link Plus</td>
<td>2989</td>
</tr>
<tr>
<td></td>
<td>LinkSolv</td>
<td>2976</td>
</tr>
</tbody>
</table>
## CDC Metrics

<table>
<thead>
<tr>
<th>CDC Metric</th>
<th>Software</th>
<th>Value per Quarter or 6-Month Period</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Period 1</td>
<td>%diff.</td>
<td>Period 2</td>
</tr>
<tr>
<td>Average dose of &gt; 90 MMEs in quarter*</td>
<td>Custom-built</td>
<td>8.89</td>
<td>0</td>
<td>8.33</td>
</tr>
<tr>
<td></td>
<td>The Link King</td>
<td>8.76</td>
<td>-1.5</td>
<td>8.22</td>
</tr>
<tr>
<td></td>
<td>Link Plus</td>
<td>8.91</td>
<td>0.2</td>
<td>8.33</td>
</tr>
<tr>
<td></td>
<td>LinkSolv</td>
<td>8.78</td>
<td>-1.2</td>
<td>8.25</td>
</tr>
<tr>
<td>Obtained prescriptions from ≥5 prescribers and ≥5 pharmacies in 6 months†</td>
<td>Custom-built</td>
<td>18.15</td>
<td>0</td>
<td>13.68</td>
</tr>
<tr>
<td></td>
<td>The Link King</td>
<td>20.44</td>
<td>12.6</td>
<td>16.74</td>
</tr>
<tr>
<td></td>
<td>Link Plus</td>
<td>25.16</td>
<td>38.6</td>
<td>20.34</td>
</tr>
<tr>
<td></td>
<td>LinkSolv</td>
<td>22.39</td>
<td>23.4</td>
<td>18.25</td>
</tr>
<tr>
<td>Overlap of opioid prescriptions in quarter‡</td>
<td>Custom-built</td>
<td>16.70</td>
<td>0</td>
<td>17.53</td>
</tr>
<tr>
<td></td>
<td>The Link King</td>
<td>17.14</td>
<td>2.6</td>
<td>18.04</td>
</tr>
<tr>
<td></td>
<td>Link Plus</td>
<td>17.55</td>
<td>5.1</td>
<td>18.45</td>
</tr>
<tr>
<td></td>
<td>LinkSolv</td>
<td>17.30</td>
<td>3.6</td>
<td>18.20</td>
</tr>
<tr>
<td>Overlap of benzodiazepine and opioid prescriptions in quarter‡</td>
<td>Custom-built</td>
<td>9.72</td>
<td>0</td>
<td>9.96</td>
</tr>
<tr>
<td></td>
<td>The Link King</td>
<td>9.89</td>
<td>1.7</td>
<td>10.15</td>
</tr>
<tr>
<td></td>
<td>Link Plus</td>
<td>10.12</td>
<td>4.1</td>
<td>10.38</td>
</tr>
<tr>
<td></td>
<td>LinkSolv</td>
<td>9.97</td>
<td>2.6</td>
<td>10.24</td>
</tr>
</tbody>
</table>

*% of patients  
†per 100,000 population  
‡% of patient prescription days
Discussion

• All 4 record linkage programs were reasonably accurate in identifying matches and non-matches
  – Most accurate: the Link King and LinkSolv
  – Least accurate: custom-built program
Importance of Matching Fields

- Date of birth was very important to human reviewers, but less so to the custom-built program and Link Plus.
- Agreement in last name was more important to the custom-built program than to human reviewers and the other 3 programs.
  - Double last names and switched first & last names were less likely to be included in matches by the custom-built program.
- Agreement in zip code was more important to the custom-built program and Link Plus than to the others.
Patient Alerts and Metrics

• Effects of using specific software were greatest on the identification of outlier patients who obtained prescriptions from a large number of prescribers and/or pharmacies
  – Prescriptions from multiple prescribers and/or pharmacies are likely to result in multiple identity records, which must be linked
Limitations

• Small scope of evaluation
  – Half a million records from geographically separated areas
  – Used default settings where available

• Changes to current linkage methods in production would require further testing for feasibility and accuracy
Conclusions

• Certain publicly and commercially available record linkage programs linked identity records more accurately than a custom-built application
  – It is not necessary to build a record linkage system from the ground up
  – It is necessary to conduct a test of any proposed software with manual review of matches to ascertain their accuracy
Thank you!

• For further information, please contact
  Susan Stewart: slstewart@ucdavis.edu