Strengthening the Role of Cancer Registries in Cancer Control through Privacy Preserving Record Linkage (P3RL)

Kerri M. Clough-Gorr, DSc, MPH[1,2];Kurt Schmidlin, DMD, MPH[2];
Adrian Spoerri, PhD, MPH[2]

[1] National Institute for Cancer Epidemiology and Registration (NICER), Zurich, Switzerland

[2] Institute of Social and Preventive Medicine (ISPM), University of Bern, Switzerland

OVERVIEW

OBJECTIVE

To demonstrate the utility of Privacy Preserving Probabilistic Record Linkage (P3RL) in cancer registration applications.

INTRODUCTION

Prevention, early detection, and effective treatment of cancer will continue to be critical activities worldwide with cancer registration and cancer epidemiology research occupying central roles. The ability to link cancer-related data sources using key discriminating information such as patient name ethically without breach of confidentiality – P3RL – has the potential to transform the role of cancer registration in cancer control by making available cancer-related data anonymously (privacy of cancer patients completely safeguarded) and thus heretofore not previously accessible.

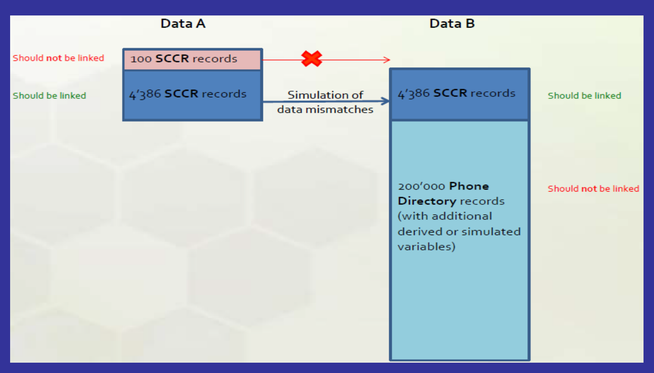
METHODS

DATA SOURCES

Data for this exercise came from the Swiss Childhood Cancer Registry (SCCR) and the Swiss phone directory (SPD).

DATA PROCEDURES

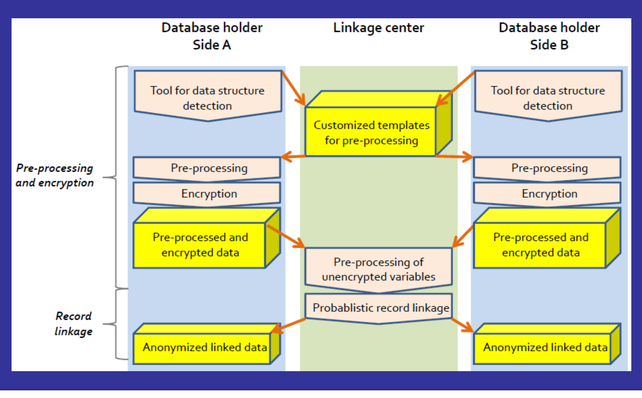
Two datasets were created: (A) 4,486 SCCR records, and (B) 200,000 SPD plus 4,386 SCCR records with simulated errors.



Sixteen variables were used for linkage.

Variable	DATA A (SCCR)		DATA B (Phone directory)	
Id	√ existing		≈ simulated	simulated
Name 1, name 2	√ existing	processed from name, maiden name	√ existing	processed from name, maiden name
First name 1, first name 2	√ existing	processed from firstnames	√ existing	processed from firstnames
Civil status	√ existing		≈ simulated	simulated based on data A distribution
Nationality	√ existing		» allocated	allocated based on surname, distribution data A and census 2000
Nationality, binary	► derived	derived from nationality	► derived	derived from nationality
Date of birth	√ existing		≈ simulated	simulated based on data A distribution
Gender	√ existing		» allocated	allocated based on first names, distribution equivalent to data A
Tumor category	√ existing	processed from 2 tumor variables	≈ simulated	simulated based on data A
Zip code	√ existing		√ existing	
Community code	√ existing		► derived	derived from zip code
Agglo-district, Grossregion, language region	► derived	derived from community code	► derived	derived from community code
Abroad (residence)	√ existing		► derived	derived
Diagnosis date	√ existing		- N/A	
Age at 31.10.2001	► derived	derived from date of birth	► derived	derived from date of birth
Assumed reference date	► derived	derived from diagnosis date and modification date	≈ simulated	fixed: 31.10.2001

Data were linked in 2 steps: (1) pre-processing and encryption with Bloom filters; then (2) probabilistic linkage.



RESULTS

LINKAGE

Four linkages were completed. Three linkage variations were compared with P3RL results.

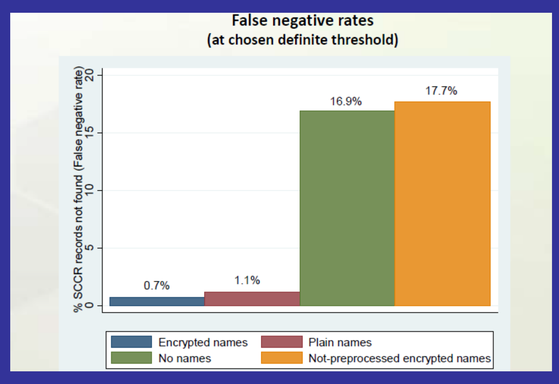
Linkage 1: P3RL

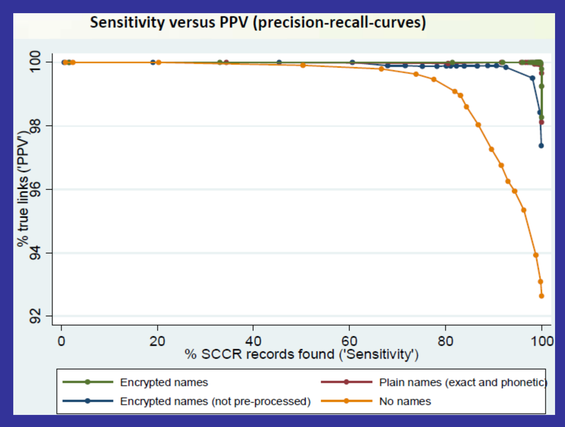
Linkage 2: with unencrypted names

Linkage 3: without names

Linkage 4: without pre-processing

- P3RL is very similar to linkage using unencrypted names.
- P3RL is much better than linkage without names.
- P3RL with pre-processing is much better than linkage without.
- P3RL has very low proportion of false positive and false negative pairs.
- Using P3RL SCCR records found and true links near 100% (i.e. sensitivity and positive predictive value [PPV] near 1).
- Using P3RL the few SCCR records that could not be linked were either records with ≥1 simulated errors (often "severe" error like gender or date of birth) or variables had many missing values.





CONCLUSION

These results indicate that P3RL is a valid and useable method of linking cancer-related data; nearly as good as linkage using unencrypted names. P3RL may be useful resolving ethical tensions regarding use of cancer registration data. Thus potentially further integrating cancer registration in cancer epidemiology research and cancer control.

SELECTED REFERENCES

1. Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc.* 1969;64:1183-1210.
2. Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. *Stat Med.* May 30 2002;21(10):1485-1496.
3. Herzog TN, Scheuren FJ, Winkler WE. *Data quality and record linkage techniques*. New York, NY: Springer; 2007.
4. DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. *J Biomed Inform.* 2010;43(1):24-30.
5. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. *BMC Med Inform Decis Mak.* 2009;9:41.

IMPACT

In practice P3RL is important because access to privacy protected data (such as patient names for linking records) is normally restricted to sites with data ownership (i.e. not allowed to share patient names between data sources). Ethically patients' identities and personal data must be strictly protected yet information regarding their cancer experience is vital to society for developing and implementing adequate cancer prevention and control programs as well as for apposite health service planning.

Please contact kerri.clough-gorr@nicer.org for more information on this or related projects.