Evaluation of Record Linkage for SEER breast cancer registries to Oncotype DX tests

Michael D. Larsen, Will Howe, Nicola Schussler, Benmei Liu, Valentina Petkov, Mandi Yu

FCSM Research Conference2015,
Wednesday, December 3 at 10:30am-12:15pm
Room 146A

Outline

1. SEER breast cancer cases and GHI Oncotype DX test
   - GHI=Genomic Health Incorporated
2. Record linkage
3. Manual review design
4. Results
5. Conclusions

Disclaimer: The opinions presented in this talk are those of the author and not necessarily any other person or organization.

SEER breast cancer registries

- Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI)
- Population-based cancer registries. 30% of the US
- Patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status
- Goals and details online at **seer.cancer.gov**.

Genomic Health OncoType DX test

- **Oncotype DX, was developed by Genomic Health, Inc. (GHI) in 2004**.
- Indicated in early stage breast cancer (hormone receptor positive, negative lymph nodes) to **stratify the risk** of distant recurrence and to help predict the benefit of chemotherapy added to hormonal therapy.
- In 2008 the assay was validated for node positive HR+ breast cancer and for DCIS in 2011.
- **Oncotype DX is the most frequently used Multigene Signature Assay in breast cancer in the US** and has been recommended by oncology practice guidelines for early stage diseases (negative lymph nodes, hormonal receptor positive, HER2 negative and tumor size > 0.5cm) since 2008.
- SEER started to collect Oncoype DX and other multigene since 2010.
- **Given the prognostic and predictive significance of Oncotype DX it is important for SEER registries to continue collecting these variables**.
- Quality and completeness of the data can be greatly improved if information is obtained directly from the labs performing molecular/genomic testing.
- **GHI is the only lab in US that carries out Oncotype DX. This fact makes it an ideal target to test linkages of laboratory results to SEER data**.

Record linkage of SEER and GHI files within registry areas

- Identify pairs of records that pertain to the same person. Combine information from two sources for true links.
- Turn comparisons on variables into a ***score*** for similarity
- High scores = likely match; Low score= likely nonmatch.
- Errors are made because of errors in data, missing values, and non-uniqueness
- Middle ground: Clerical review is possible

LinkPlus 3.0 Beta Software

- Probabilistic record linkage program developed at CDC's Division of Cancer Prevention and Control in support of CDC's National Program of Cancer Registries (NPCR). Free online
- Based on Fellegi and Sunter (1969 JASA)

Overall Linkage Procedure

- Two-step match
  - First-step: LinkPlus to obtain the scores
  - Second-step: in-house developed SAS program to further refine the matches
  - We experimented with a few LinkPlus cutoffs to balance the sensitivity of throwing away true matches or the amount of clerical review efforts
  - De-duplicate SEER to patient-level; match those to GHI cases; once pairs of records are determined to be the same person; associate the records in two datasets

LinkPlus settings: Blocking variables

Blocking Variables: If the records match exactly on ANY of these fields, the match will be assigned a score (fairly broad)

- Address State (Phonetic method: none)
- First Name (Soundex)
- Last Name (Soundex)
- SSN (Phonetic method: None)
- Date of Birth (Phonetic method: None)

LinkPlus settings: Matching variables

Matching Variables: Used for score calculations. Exact matches get a higher score than partial matches. The exact scoring algorithms are in the LinkPlus black box. For each record in the primary file, only the match with the best score is kept:

- First Name (match method: first name)
- Middle Name (middle name)
- Last Name (last name)
- SSN (ssn)
- Date of Birth (date)

Methods: additional requirements for linkage to be accepted

In-house development based on SAS (by IMS)

Method 4.1 (various criteria were initially investigated).

Of those pairs that score above 7:

Match = exact match on first and last name and at least 2 of the following: date of birth, SSN, (phone number or street address)

Manual Review = exact or partial matches on 3 of the following: first name, last name, DOB, SSN, phone, address* (city & state)

- or exact match on SSN and partial match on 1 of the following: first name, last name, DOB, phone
- or exact match on phone number and partial match on 1 of the following: first name, last name, DOB, SSN

Non-match = failed to match exactly/partially on 3 of the following: first name, last name, DOB, SSN, phone, address* (city & state)

* Address is not checked for partial matches

Research questions

1. How accurate is the linkage?
2. What affects the quality of the linkage?

Evaluation Study

Manual review design: basic review

- Review all 18,643 potential matches that score above 7 and are classified as “manual view”

All records

- For example, in Connecticut: n=18,792 pairs above 7 cutoff

Best matches: Processing by Method 4.1

- For example, in Connecticut:
- n=743 manual review

Manual review design: additional pairs

1. Sample some records that score 6-7
2. Sample some records that score above 7 and are "match" by additional criteria
3. Sample some records that score above 7 and are "non match" by additional criteria
4. Also OncoType=YES in SEER but not matched (*n*=103)

Additional effort (1-3) spread across participating registries.

Results: Number of matched pairs

		n	Link	Nonlink
1	Score 5-6	1,999	0	1,999 (100%)
2	Score >7, Designated match	1,998	1,998 (100%)	0
3	Score >7, Designated nonmatch	1,998	0	1,998 (100%)
4	Score >7, Manual review group	18,644	12,783 (70%)	5,661 (30%)
	Total	24,742	14,781 (60%)	9,858 (40%)

Groups 1, 2, and 3 are proportional samples by registry

Group 4 is N=population size of all record pairs

Conclusion: score of 7 is a good cut point

Match and nonmatch additional criteria are accurate

Manual review is pretty important

Match rate varies by registry

- There was variability
- Was it due to differences by region or difference by procedure for declaring matches?
- This will be further investigated

Validation Result for SEER says OncotypeDX=Yes in 4 registries

- 103 BC cases with OncotypeDX=Yes did not have a match – lack of matching variables
- 680 BC cases with OncotypeDX=Yes and possible match were rejected based on clerical review – again lack of matching variables

In total, 2,112 BC cases with OncotypeDX=Yes were not matched to GHI tests: 8.3% of all OncotypeDX tests (also varied by registry)

Study of variables used in linkage

Several variables were created using in-house SAS for the LinkPlus pairs

- City, State, Street: nonmatch, match, missing [3]
- DD, MM, YYYY: 3 versions + minor + transpose [5]
- SSN, Phone: 5 versions + JW [6]
- Last: 6 versions + contains [7]
- DOB: 6 versions + MD_swap [*not used here*]
- Middle: 7 versions + 2 comparisons to last [9]
- First: 9 versions + 2 comparisons to middle [11]
  - *Jaro-Winkler distance not used here*

Predicting Score

- R-squared for predicting score using main effects of 10 variables is 73%
- All variables have 2 or more statistically significant levels for predicting score
- Impact on score if a pair is nonmatching on …

State	**-0.49**		**Middle**	**-0.14**
SSN	**-0.20**		**Phone**	**-0.12**
Last	**-0.20**		**First**	**-0.06**
Year	**-0.19**		**Street**	**-0.05**
Day	**-0.17**		**Month**	**-0.04**

Predicting Match via Logistic Regr.

- Accuracy for predicting match (using estimated probability above 0.6) is 92%
- All variables have 2 or more statistically significant levels for predicting match
- Impact of nonmatch on linear scale …

SSN	**-5.86**		**Street**	**-2.63**
Year	**-4.34**		**Month**	**-2.61**
Last	**-3.55**		**State**	**-1.94**
Day	**-3.50**		**First**	**-1.76**
Phone	**-2.74**		**Middle**	**-1.00**

Limitations

- LinkPlus gives only one best match and a score
  - A second or third record might be a near match and help one decide whether to accept the best
- You must do your own comparison of fields separately to incorporate that information
- Review of records was not blinded – reviewers knew which batch records were in and linkage score – difficult to avoid this

Three issues for further study

- **Dates**: Date of test should be relatively soon after Date of Diagnosis, but sometimes it is delayed (e.g., payment). Challenging to use.
- **Multiple primary tumors** could create duplicate people in SEER (and possibly GHI?), but linkage by tumor should be possible.
- **Movers**: always a concern for address and phone

Summary

- Record linkage effectively identified most of the pairs between SEER breast cancer cases and GHI's Oncotype DX database.
- LinkPlus has some limitations as has been noted.
  - Limited to 10 matching variables
  - Memory limitation
- Variability by SEER registry will be studied
- Quality of variables and how they are pre-processed is considered key factor in success of record linkage
- Some interesting results on predicting score and match, but more to do.

Future

- Ongoing work to establish performance and reporting standards for NCI record linkage projects
- Comparing other record linkage software and methods of handling inexact agreement on fields of information

Thanks!

- Thanks to organizers and FCSM and the chair and discussant of this session
- Thanks to my coauthors and collaborators (NCI, IMS)
- Funding under contract to NCI
- ***Thanks to all who did manual review in the several SEER registry offices!***

mlarsen@bsc.gwu.edu