Statistical Modeling for Errors in Record Linkage Applied to SEER Cancer Registry Data

Michael D. Larsen (GWU)
Collaborators:
Will Howe, Nicola Schussler (IMS),
Benmei Liu, Valentina Petkov, Mandi Yu (NCI)

SSC Methodology Symposium 2016,
Wednesday, March 23 at 10:30am-12:00pm
Session 6A

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; several registry areas.
- 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.
- 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 (also MEMORY)
  - 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)

- State
- First Name (Soundex)
- Last Name (Soundex)
- SSN
- Date of Birth

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
- Middle Name
- Last Name
- SSN
- Date of Birth

Methods: additional requirements for linkage to be accepted

In-house development based on SAS (by IMS)

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 = intermediate criteria

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

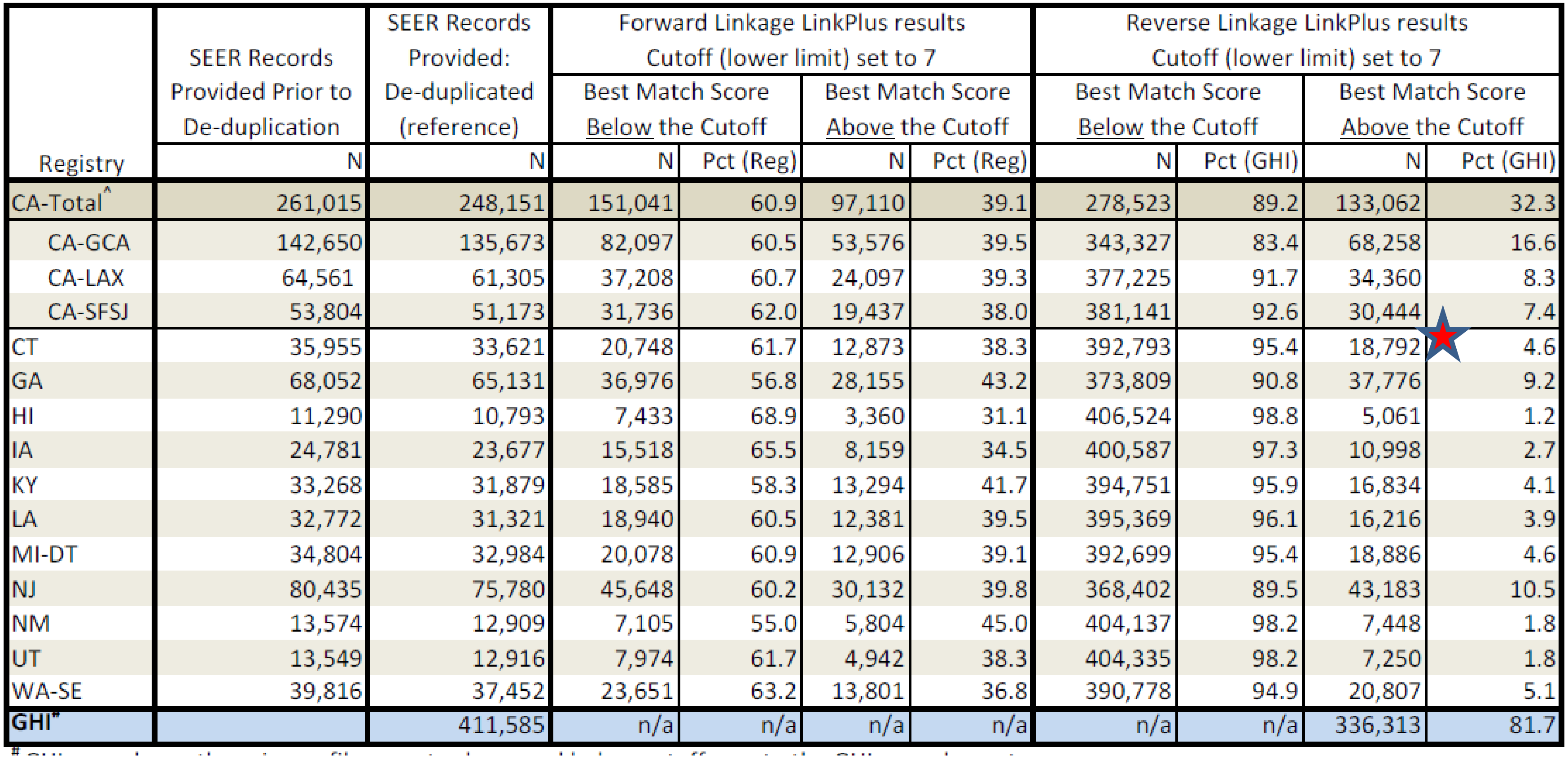
1. How accurate is the linkage?
2. What affects the quality of the linkage?

Evaluation Study

Manual review design: basic review

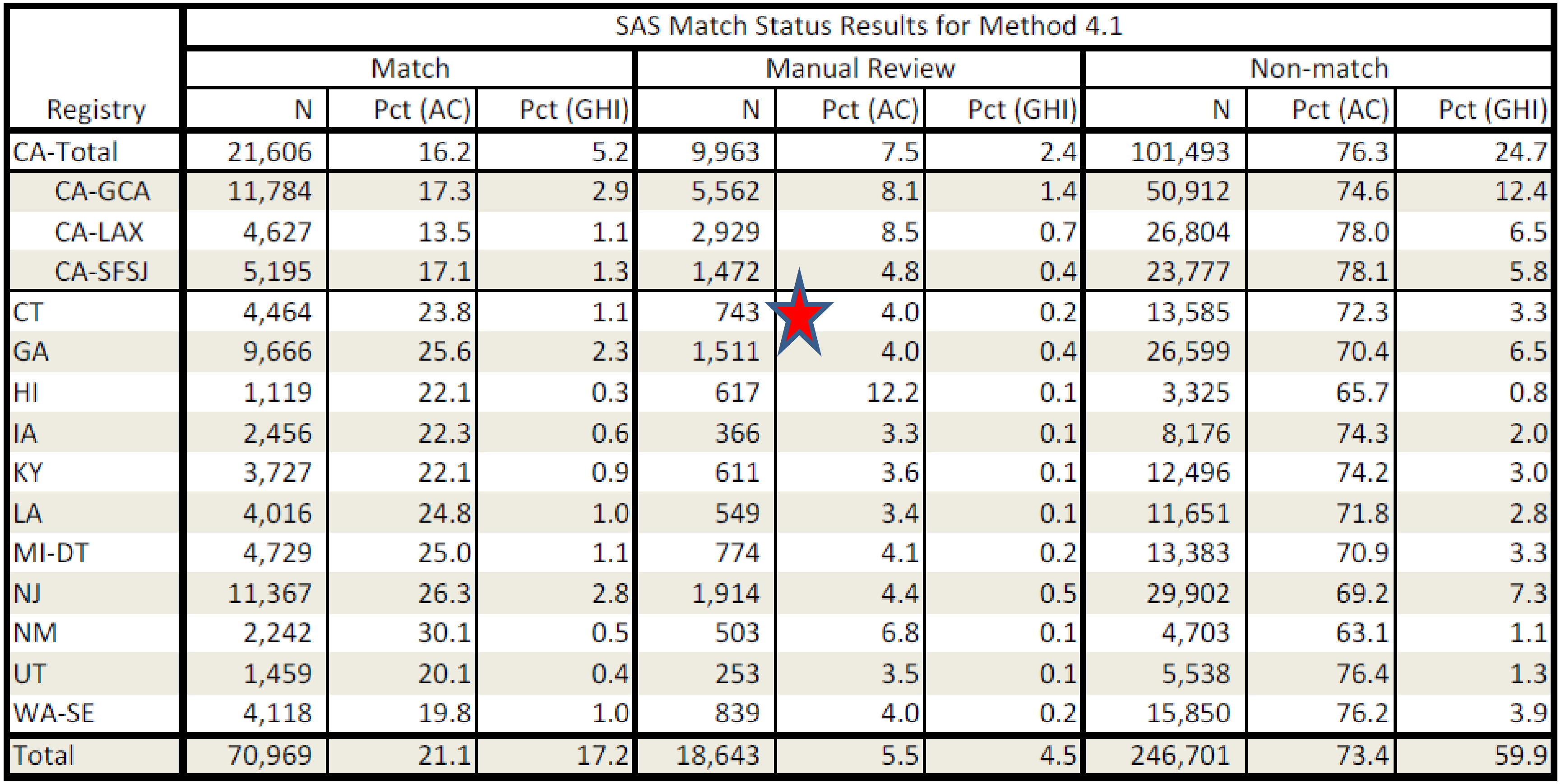
- Review all 18,643 potential matches that score above 7 and are classified as “manual view”

All records
CT: n=18,792 pairs above 7 cutoff



			Forward Linkage LinkPlus results Cutoff (lower limit) set to 7				Reverse Linkage LinkPlus results Cutoff (lower limit) set to 7			
	SEER Records Provided Prior to De-duplication	SEER Records Provided: De-duplicated (reference)	Best Match Score <u>Below</u> the Cutoff		Best Match Score <u>Above</u> the Cutoff		Best Match Score <u>Below</u> the Cutoff		Best Match Score <u>Above</u> the Cutoff	
Registry	N	N	N	Pct (Reg)	N	Pct (Reg)	N	Pct (GHI)	N	Pct (GHI)
CA-Total^	261,015	248,151	151,041	60.9	97,110	39.1	278,523	89.2	133,062	32.3
CA-GCA	142,650	135,673	82,097	60.5	53,576	39.5	343,327	83.4	68,258	16.6
CA-LAX	64,561	61,305	37,208	60.7	24,097	39.3	377,225	91.7	34,360	8.3
CA-SFSJ	53,804	51,173	31,736	62.0	19,437	38.0	381,141	92.6	30,444	7.4
CT	35,955	33,621	20,748	61.7	12,873	38.3	392,793	95.4	18,792 ★	4.6
GA	68,052	65,131	36,976	56.8	28,155	43.2	373,809	90.8	37,776	9.2
HI	11,290	10,793	7,433	68.9	3,360	31.1	406,524	98.8	5,061	1.2
IA	24,781	23,677	15,518	65.5	8,159	34.5	400,587	97.3	10,998	2.7
KY	33,268	31,879	18,585	58.3	13,294	41.7	394,751	95.9	16,834	4.1
LA	32,772	31,321	18,940	60.5	12,381	39.5	395,369	96.1	16,216	3.9
MI-DT	34,804	32,984	20,078	60.9	12,906	39.1	392,699	95.4	18,886	4.6
NJ	80,435	75,780	45,648	60.2	30,132	39.8	368,402	89.5	43,183	10.5
NM	13,574	12,909	7,105	55.0	5,804	45.0	404,137	98.2	7,448	1.8
UT	13,549	12,916	7,974	61.7	4,942	38.3	404,335	98.2	7,250	1.8
WA-SE	39,816	37,452	23,651	63.2	13,801	36.8	390,778	94.9	20,807	5.1
GHI#		411,585	n/a	n/a	n/a	n/a	n/a	n/a	336,313	81.7

Best matches: Processing by Method 4.1
CT: n=743 manual review



Registry	SAS Match Status Results for Method 4.1								
	Match			Manual Review			Non-match		
	N	Pct (AC)	Pct (GHI)	N	Pct (AC)	Pct (GHI)	N	Pct (AC)	Pct (GHI)
CA-Total	21,606	16.2	5.2	9,963	7.5	2.4	101,493	76.3	24.7
CA-GCA	11,784	17.3	2.9	5,562	8.1	1.4	50,912	74.6	12.4
CA-LAX	4,627	13.5	1.1	2,929	8.5	0.7	26,804	78.0	6.5
CA-SFSJ	5,195	17.1	1.3	1,472	4.8	0.4	23,777	78.1	5.8
CT	4,464	23.8	1.1	743	4.0	0.2	13,585	72.3	3.3
GA	9,666	25.6	2.3	1,511	4.0	0.4	26,599	70.4	6.5
HI	1,119	22.1	0.3	617	12.2	0.1	3,325	65.7	0.8
IA	2,456	22.3	0.6	366	3.3	0.1	8,176	74.3	2.0
KY	3,727	22.1	0.9	611	3.6	0.1	12,496	74.2	3.0
LA	4,016	24.8	1.0	549	3.4	0.1	11,651	71.8	2.8
MI-DT	4,729	25.0	1.1	774	4.1	0.2	13,383	70.9	3.3
NJ	11,367	26.3	2.8	1,914	4.4	0.5	29,902	69.2	7.3
NM	2,242	30.1	0.5	503	6.8	0.1	4,703	63.1	1.1
UT	1,459	20.1	0.4	253	3.5	0.1	5,538	76.4	1.3
WA-SE	4,118	19.8	1.0	839	4.0	0.2	15,850	76.2	3.9
Total	70,969	21.1	17.2	18,643	5.5	4.5	246,701	73.4	59.9

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)

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

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 was set to give only one best match
  - 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

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
- Modeling efforts – what is impact of record linkage on subsequent analyses

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