

# An Overview of Patient Matching

---

## Shaun Grannis, MD MS

Medical Informatics Research Scientist,  
Regenstrief Institute

Assistant Professor of Family Medicine,  
Indiana University School of Medicine

U.S. Population Health Technical Work  
Group Co-Chair,  
Health Information Technology  
Standards Panel



# Challenges

---

- Recording Errors
  - Phonetic
  - Typographical
- Identifiers change
  - Last Name
  - Address
  - Phone
- Sharing Identifiers (SSN, etc.)



# Barriers to Accurate Patient Matching

---

- Recording Errors
  - Phonetic ("Shaun", "Sean", "Shawn")
  - Typographical  
(S**m**ith → S**n**ith, "0**7**" → "0**1**")
- Changing Identifiers
  - Last Name (Marriage)
  - Geographic location (Home address, etc)
- Sharing Identifiers (SSN, etc.)
- Identifiers Limited or Unavailable



# Ideal Identifier Characteristics

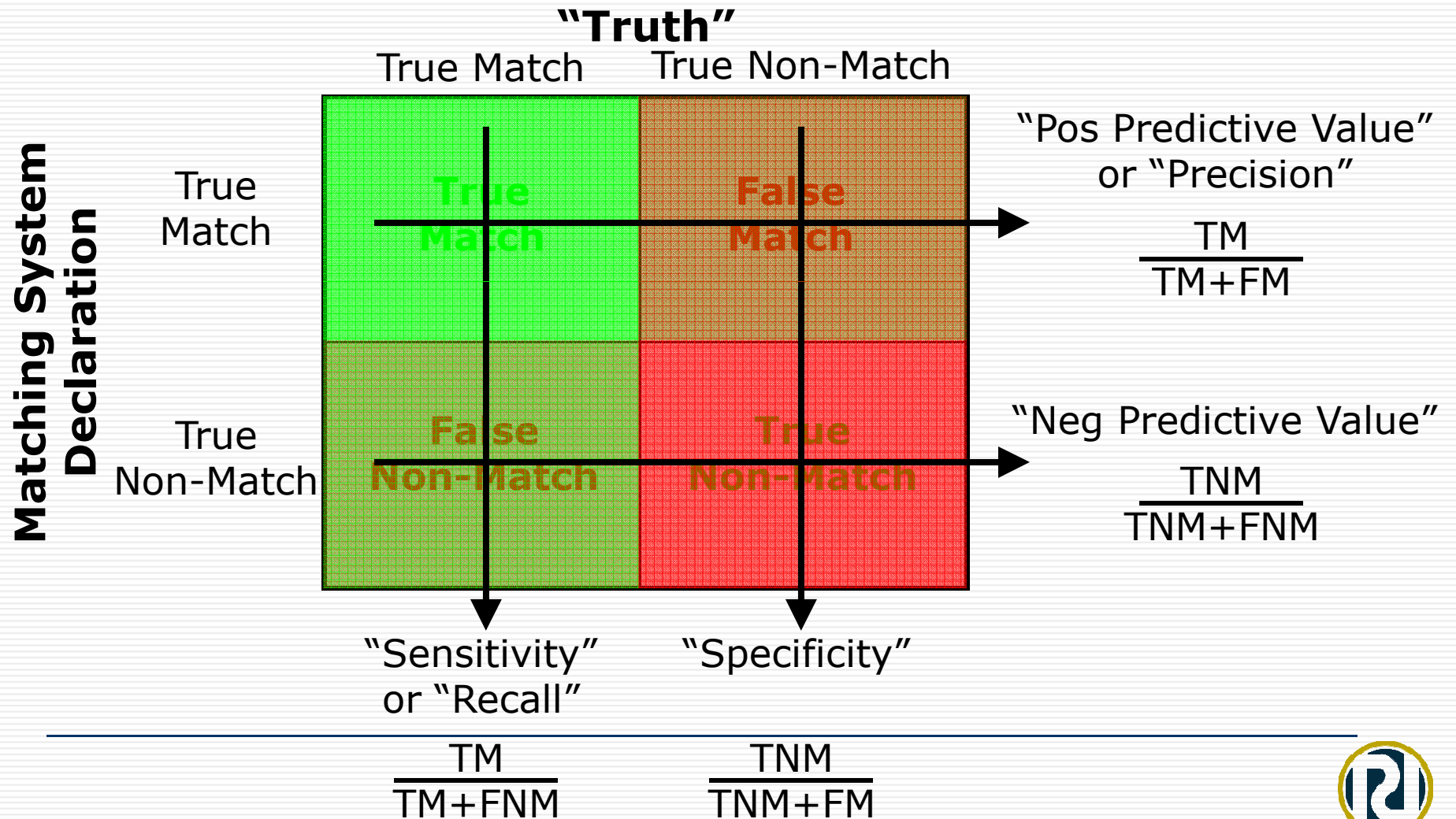
---

- **Unique**  
(eg, fingerprint, Iris, DNA, National ID)
- **Ubiquitous**  
(eg, Name, DOB, Sex, Eye Color)
- **Unchanging**  
(eg, DOB, Sex, Given Name, DNA)
- **Uncomplicated**  
(eg, Name, DOB, Sex)
- **Uncontroversial**  
(eg, avoid sensitive data)
- **Easily and Inexpensively Accessible**

**No identifier  
meets all of these  
characteristics**



# Patient Matching Terminology



# Patient Matching Terminology

---

- **Potential Pairs/Potential Links**  
Record-pairs that have not been declared a match or non-match
- **Blocking/Grouping**  
Method to limit search space for potential links, usually by forcing exact match with one or more fields. (Analogous to sorting socks by color before pairing)
- **Field Agreement Weight/Score**  
Value assigned when two fields are declared to agree
- **Field Disagreement Weight/Score**  
Value assigned when two fields are declared to disagree
- **Record Pair Score/Composite Score/Global Score**  
Value derived from individual field contributions (typically the product or sum of field weights)
- **Score Threshold**  
record pair score above which a match is declared and/or below which a non-match is declared



# Potential Solutions

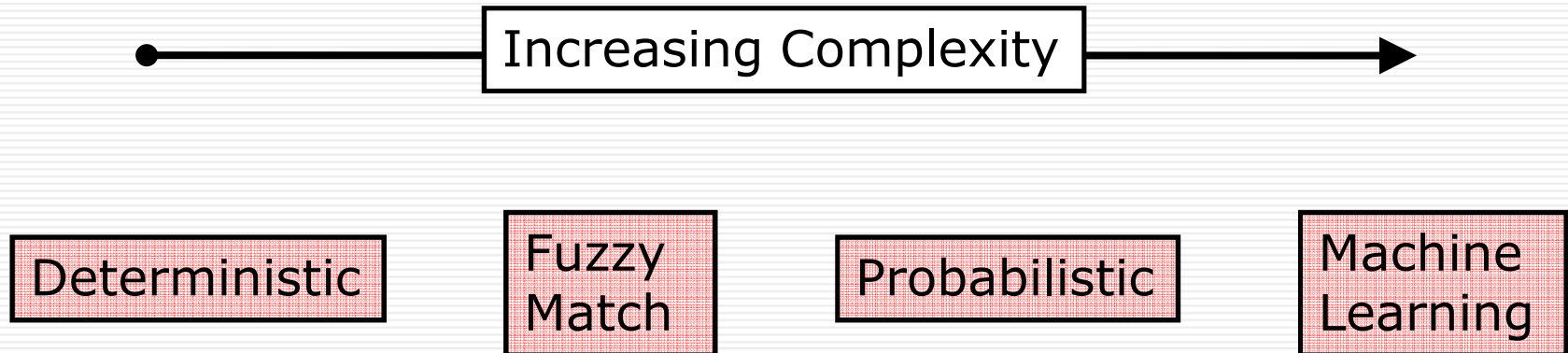
---

- National Patient Identifier
  - Recording errors
  - Sharing ID's
  - Lost ID's
  - Controversial (in some regions)
- Biometrics
  - Require proprietary hardware for all data generators
  - How secure?
  - Privacy concerns



# Patient Matching Methodologies

---



# Deterministic

---

- 'Rules-based' or 'Heuristic'
- Accuracy is highly dependent on presence of discriminating identifiers (national or local ID, etc)
- Rule-based, eg declare a match if exact match on:
  - National ID + DOB
  - Full Name + Address
  - etc.



# Fuzzy Match

---

- Non-exact agreement, allows for errors:
  - “If last name agrees on first 6 characters then declare agreement”
  - “If birth date is within 1 month, then declare agreement”
- To loosen agreement, string comparators or phonetic transformation functions may be used:
  - Soundex - Phonetic
  - NYSIIS - Phonetic
  - Levenshtein Edit Distance - Comparator
  - Jaro-Winkler Comparator - Comparator
  - Longest Common Sub-sequence - Comparator



# Probabilistic/Machine Learning

---

- Implements a statistical model for matching
- A common model is Felligi-Sunter maximum likelihood model
- Establish parameters for model using machine learning algorithms (EM) or bootstrap review
- Maximum Entropy Model also used



# Patient Matching Methodologies

---

## Deterministic/Heuristic

- Rapid Implementation
- Simple calculations
- Relies on accurate and consistent data
- May not generalize well to other data sets

## Probabilistic

- Complex implementation
- Computationally intensive
- More forgiving of data errors
- Algorithms adapt to data being linked



# Probabilistic (F-S) Example

---

- Among the 10 true-links, the last names agreed in 9/10 pairs (e.g. one of the last names was misspelled)
- This represents a 90% AGREEMENT RATE for last name among TRUE LINKS.
- Similarly, among the 90 non-links, last names agreed (by random chance) in 2/90 pairs
- This represents a 2% AGREEMENT RATE for last name among NON-LINKS.



# Probabilistic (F-S) Example

---

$\frac{90\%}{2\%} = 45$  ■ Records that agree on last name are 45 times more likely to be a true-link than a non-link

- Weights for each field are combined to form a composite record pair score.
- Field disagreement contributes a negative weight, and reduces the overall record pair score.

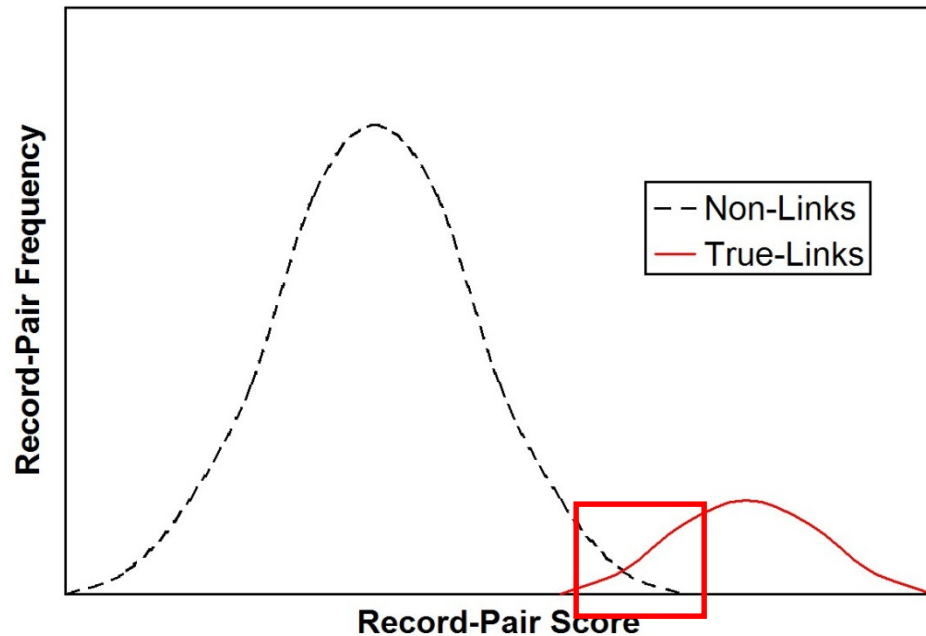


# Probabilistic (F-S) Example

- 1 Generate Record-Pairs: Each record pair is assigned a score.  
2 A histogram of scores may look like:

File 1	File 2
Record A	Record X
Record B	Record Y
Record C	Record Z

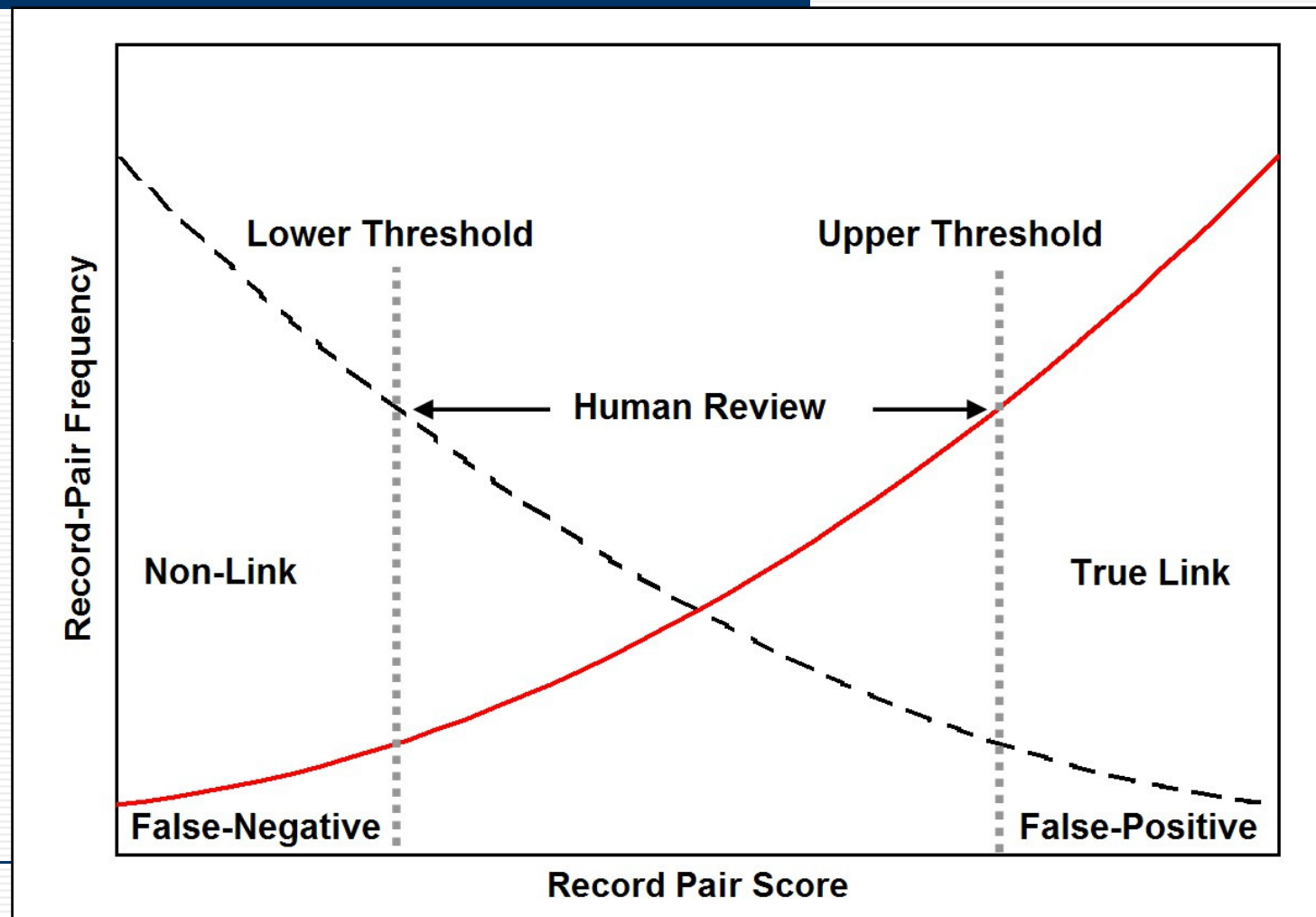
Potential Record Pairs



Which are true links?



# Probabilistic Linkage Overview: Human Review Thresholds



# Global Patient Registry Matching Algorithm Features

---

- One record per assigned patient number per institution
- Create logical links between each of these records
- Match using social security number, patient name, birth date, gender
- Use string comparator algorithm and phonetic transformations for near name matches
- Implements the concept of a “patient group”



# Global Patient Registry

---

Assigning

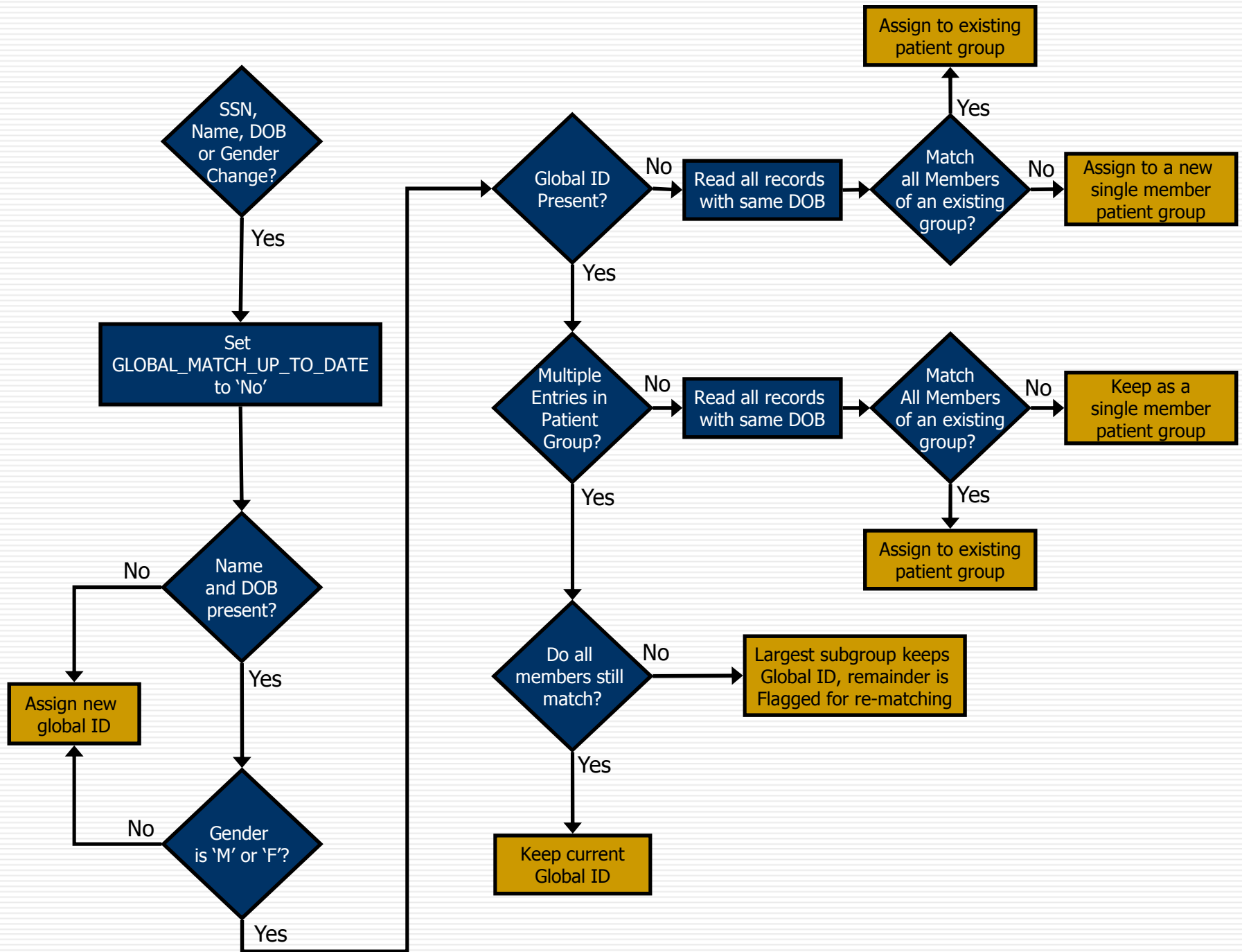
Authority   Global #   Local Pat #   Patient Name   Birth date   Sex

Hospital A	99-1	231456	Sinkwell, Ralph J	12-2-59	M
Hospital B	123-0	A47239	Sinkwell, RJ	2-12-59	M
Hospital A	99-1	1032115	Sinkweil, Ralph	12-2-59	
Hospital C	101-0	A3276	Fredrick, Alice	4-14-78	F
Hospital A	101-0	2314590	Fredrick, Alyce	4-14-78	F



Global ID is for internal indexing only  
– not publicly exposed







Back Search Favorites

Address <http://kite.wishard.edu:7106/REGEN/0/37F2/load/top.subdoc> Go Links

Google Search Web PageRank 313

Select a patient Browse Patient Record Other **Select a patient»Select Patient**

Hide Menu Select another patient Show patient data Praxis Logout Help

**Select Patient**

### Patient Lookup By Name

Selection(s):

	Patient_Name	Hospital#	Birth_Date	SocSec#	S Phone	Moms_Nam
1)			29-Aug-86	On file		
2)			27-Apr-80	On file		
3)			14-Aug-71	Not on file		
4)			20-Jul-82	Not on file		
5)			11-May-79	On file		
6)			03-Sep-75	On file		
7)			03-Sep-75	On file		
8)			24-Aug-98	Not on file		
9)			01-Jul-77	On file		
10)			19-Aug-65	Not on file		

# Results Retrieval: Patient Selection Dialog

Back | Search | Favorites | Media | 2002 blocked | AutoFill | Option

Address: http://kite.wishard.edu:7134/REGEN/0/4A59/load/top.subdoc

Age: 18 years [WISHARD]

Select a patient | Browse Patient Record | Other | Select a patient»Select Patient

Hide Menu | Cancel selection | Accept choices | Praxis | Logout | Help

Select Patient

Choose electronic patient records to display for this patient

Selection(s): **ALL**

Institution	Patient ID	Name	Birth	SSN
<input type="checkbox"/> All				
<input type="checkbox"/> 1) CLARIAN	6596000700	SHARON, JENNIFER	29-Aug-86	No

Results Retrieval:  
Global Patient Group Display

Choose electronic patient records to display for this patient

Selection(s): **ALL**

Institution	Patient ID	Name	Birth	SSN
<input type="checkbox"/> All				
<input type="checkbox"/> 1) CLARIAN	6596000700	SHARON, JENNIFER	29-Aug-86	No
<input type="checkbox"/> 2) CLARIAN	7147000000	WELLS, JENNIFER	29-Aug-86	No
<input type="checkbox"/> 3) COMMUNITY	0033000000	WELLS, JENNIFER	29-Aug-86	Yes
<input type="checkbox"/> 4) MARION_COUNTY	0090000000	WELLS, JENNIFER	29-Aug-86	No
<input type="checkbox"/> 5) WISHARD	1040000000	WELLS, JENNIFER	29-Aug-86	Yes

# Bibliography - Theory

---

- Fellegi IP, Sunter SB. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.
- Dunn HL. (1946) Record Linkage. *Am J Public Health*. 36, 1412-1416.
- Newcombe HB. (1988) *Handbook of Record Linkage, Methods for Health and Statistical Studies, Administration, and Business*. Oxford University Press.
- Newcomb HB, Kennedy JM, Axford SJ, James AP. (1959) Automatic Linkage of Vital Records. *Science*, 130, 954-959.
- Gill, L., *Methods for Automatic Record Matching and Linking and their use in National Statistics*. Her Majesty's Stationary Office, Norwich, 2001.
- Porter E, Winkler W. Approximate String Comparison and its Effect on an Advanced Record Linkage System. *Record Linkage Techniques--1997: Proceedings of an International Workshop and Exposition*. National Academy Press, Washington DC 1999.
- Public Health Informatics Institute. *The unique records portfolio*. Decatur, GA: Public Health Informatics Institute, 2006.



# Bibliography:

## Applications and Research (1)

---

- Christen P. Febrl: A freely available record linkage system with a graphical user interface. Submitted to the Australasian Workshop on Health Data and Knowledge Management (HDKM), Wollongong, January 2008.
- Potosky A, Riley G, Lubitz J, et al. Potential for Cancer Related Health Services Research Using a Linked Medicare-Tumor Registry Database. *Medical Care* 1993;31(8):732-748.
- Whalen D, Pepitone A, Graver L, Busch JD. Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. SAMHSA Publication No. SMA-01-3500. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, July 2000.
- Liu S, Wen SW. Development of Record Linkage of Hospital Discharge Data for the Study of Neonatal Readmission. *Chronic Diseases in Canada* 1999; 20(2):77-81.
- Pates R, Scully W, et al. Adding Value to Clinical Data by Linkage to a Public Death Registry. *MedInfo* 2001;10(Pt 2):1384-8



# Bibliography:

## Applications and Research (2)

---

- Lynch BT, Arends WL. Selection of a surname coding procedure for the SRS record linkage system. Washington, DC: US Department of Agriculture, Sample Survey Research Branch, Research Division, 1977.
- Newman T, Brown A. Use of Commercial Record Linkage Software and Vital Statistics to Identify Patient Deaths. J Am Med Inform Assoc. 1997 May-June; 4 (3): 233-237.
- Schadow G, McDonald CJ Maintaining Patient Privacy in a Large Scale Multi-Institutional Clinical Case Research Network. AMIA Proceedings (2002 Submission).
- Public Health Informatics Institute. (2006). The Unique Records Portfolio. Decatur, GA: Public Health Informatics Institute
- Sideli R, Friedman C. Validating Patient Names in an Integrated Clinical Information System. Symposium on Computer Applications in Medical Care, Washington, DC. November 1991:588-592.



# Bibliography:

## Applications and Research (3)

---

- Miller PL, Frawley SJ, Sayward FG. IMM/Scrub: a domain-specific tool for the deduplication of vaccination history records in childhood immunization registries. *Computers and Biomedical Research* 2000;33:126–143.
- Salkowitz SM, Clyde S. De-duplication technology and practices for integrated child-health information systems. Decatur, GA: All Kids Count, Public Health Informatics Institute, 2003.
- Van Den Brandt PA, Schouten LJ, Goldbohm RA, Dorant E, Hunan PMH. Development of a record linkage protocol for use in the Dutch Cancer Registry for epidemiological research. *Int J Epidemiol* 1990; 19:553-8.
- Grannis SJ, Overhage JM, McDonald CJ. Analysis of Identifier Performance Using a Deterministic Linkage Algorithm. *Proc AMIA Symp* 2002:305-9.
- Grannis SJ, Overhage JM, McDonald CJ. Analysis of a Probabilistic Record Linkage Technique without Human Review. In: *Proceedings of American Medical Informatics Association Fall Symposium*; 2003; Washington, D.C.; 2003.
- Integrating the Health Care Enterprise. (2006) Patient Identifier Cross-Reference (PIX) and Patient Demographic Query (PDQ) HL7 v3 Transaction Updates. Available at: [http://www.ihe.net/Technical\\_Framework/upload/IHE\\_ITI\\_TF\\_Suppl\\_PIXPDQ\\_HL7v3\\_PC\\_2006\\_08\\_15.pdf](http://www.ihe.net/Technical_Framework/upload/IHE_ITI_TF_Suppl_PIXPDQ_HL7v3_PC_2006_08_15.pdf)

