Unduplication and Data Cleaning for Immunization Registry Records

Project Overview

11/30/99

Perry L. Miller, MD, PhD (PI)

Center for Medical Informatics
Yale University School of Medicine

(CDC grant U1W/CCU114707)

Overview: Three Main Topics

1. **Exploring the use of AutoMatch to perform deduplication using demographic information.**

2. **Building IMM/Scrub, a pilot tool for deduplication of vaccination history records within a single patient record.**

3. **Exploring how an immunization forecasting program (IMM/Serve) can accommodate incomplete vaccination histories (containing missing doses).**

Topic 3:

Exploring the use of AutoMatch to perform deduplication using demographic information (in progress)

AutoMatch

- **a commercial probabilistic program designed for database record deduplication**
- **focuses on using demographic data**

Three Immunization Registries

- **Oregon ALERT**

  - 600,000+ patient records (ALERT as a whole)
  - ~9,000 patient records (Clatsop county, pilot focus)

- **Multnoma County (contains Portland, OR)**

  - ~134,000 patient records

- **Philadelphia, PA**

  - ~186,000 patient records

Oregon ALERT (1968 data):
Missing Demographic Data Elements
(431,024 records)

Field	% Missing	
Last Name	0	
First Name	0	
Date of Birth	0.1	
Sex	0	
Middle Name	37	
Name Suffix	99.9	
State of Birth	89	
Last Name at Birth	90	
Social Security Number	96	
Medicaid Number	95	
Mother's Last Name	100	
Mother's First Name	100	
Mother's Middle Name	100	
Mother's Name Suffix	100	
Mother's Maiden Name	72	
Mother's HBsAG Status	0	(99.9% "unknown”)
Race	0	(52% "unknown”)
Ethnicity	0	(94% "unknown”)
Language Written	0	(86% "unknown”)
Language Spoken	0	(86% "unknown”)

AutoMatch:
Probabilistic Record Deduplication

- **An AutoMatch "run" consists of several "passes"**

- **Each pass requires specifying:**

  - a blocking variable
  - one or more variables to be used in matching
  - a matching threshold

- **Each pass identifies a NEW set of matches (possible duplicate records), not including matches identified by previous passes**

The AutoMatch Blocking Variable

- **Used to divide the matching problem into multiple subproblems**

- **With no blocking:**

  430,000 x 430,000 = 184,900,000,000 matches performed

- **Blocking using DOB:**

  - **assuming 1,000 DOB values:**

    1,000 x (430x430) = 184,900,000 matches performed

  - **assuming 4,300 DOB values:**

    4,300 x (100x100) = 43,000,000 matches performed

An Example AutoMatch Run

Pass 1:	**- block: DOB**	**85 matches**
	- match: Last name	**(83 dups, 2 twins)**
	First name	
	Middle name	
	Maiden name	
Pass 2:	**- block: DOB**	**12,240 new matches**
	- match: Last name	**(first 100 all dups)**
	First name	
	Middle name	
Pass 3:	**- block: DOB**	**4,858 new matches**
	- match: Last name	
	First name	
Pass 4	**- block: DOB**	**25,103 new matches**
	- match: Last name	

[4 hours on a 200 Mhz Pentium Pro PC, 128 MB]

Ongoing Research

- **Explore Automatch strategies for the test immunization registries.**

- **Determine sensitivity and specificity of the different data elements, alone and in combination.**

- **Explore how one might analyze two vaccination histories to assist in demographic deduplication, based on the relationship between the two histories, e.g.:**

  - identical,
  - subset,
  - complementary,
  - conflicting,
  - number of common vaccinations, etc.