Linking Maternal and Child Health Data to Create a Comprehensive Longitudinal Dataset: The Florida Experience

Hamisu Salihu, Ph.D., Principal Investigator
University of South Florida College of Public Health
R01 HS19997-01
9/30/2010 – 9/29/2013 (3 years)

April 6, 2011

Specific Aims

1. To create an **expanded clinically enhanced maternal-infant dataset** for the State of Florida by augmenting the current statewide hospitalization data files through linkages to other data sources.
2. To **validate the created dataset** in Specific Aim 1 through a rigorous process that will establish confidence in the use of the dataset.
3. To demonstrate the utility of the newly created, enriched dataset in conducting comparative effectiveness analysis using early term elective delivery as a case study.

Data Sources

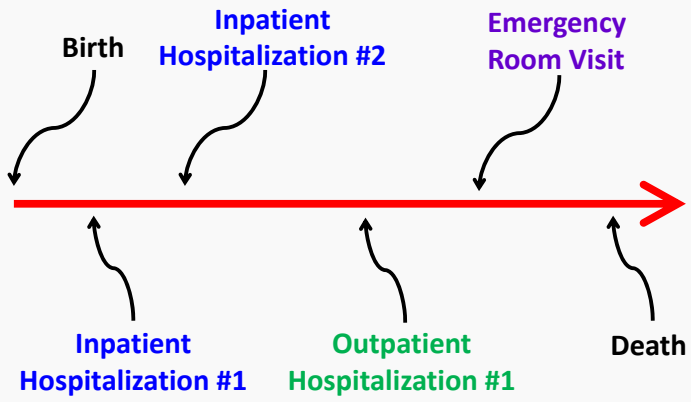
FDOH *(Office of Vital Statistics)* Data

Live Birth Data	Mortality (Infant and Mom) Data	Fetal Death Data
1998-2009	1998-2009	1998-2009

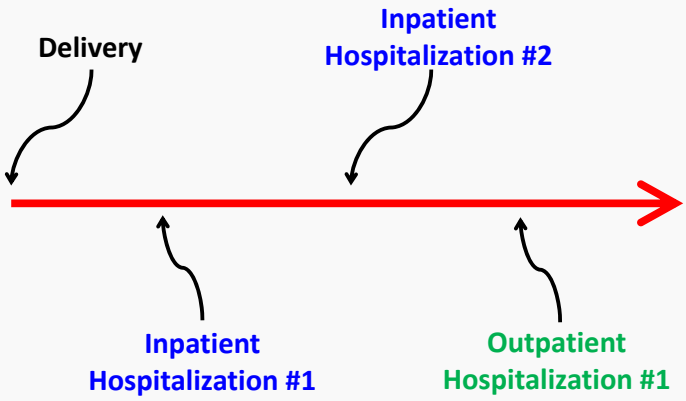
Agency for Health Care Administration Data

Inpatient Hospitalization Data	Outpatient/ Ambulatory Hospitalization Data	Emergency Department Data	Hospital Financial Data
1998-2009	1998-2009	***2005***-2009	1998-2009

"Follow" Infants Over Time Through Linkage



"Follow" Moms Over Time Through Linkage



Special Challenges to Our Data Linkage

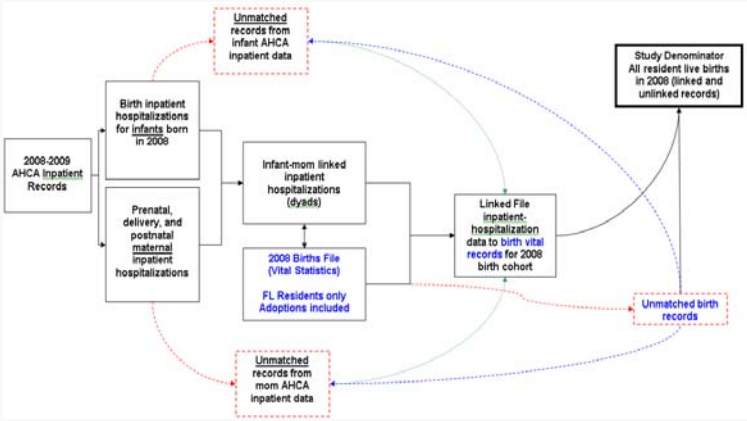
- **Birth vital records** contain a significant amount of identifying information
- **Hospital records** (inpatient, ambulatory, ED) for the infant contain limited identifying information
  - ✓ No infant SSN, name, address
  - ✓ Primary identifier is mother's SSN (INFANTLINK), but it is missing >10% and disproportionately among certain subgroups
  - ✓ Previous investigation reveal that maternal SSN has a typo or transposition in over 1,000 instances (**ASSUME** identifiers have errors)
  - ✓ Missing mother's date of birth, a key linking and/or confirmatory variable

Our Approach to Linking AHCA to VS

- **Stage I**
  - ✓ Within the inpatient hospital discharge data, we first attempt to link infants to their mothers (so called **dyad** links) with the primary goal of obtaining maternal DOB, an important linking variable (**FIND** other identifying information)
- **Stage II**
  - ✓ Link these dyad pairs to birth vital records, now incorporating infant's and mom's DOB, mom's SSN, and facility of birth as the primary linking variables
- **Stage III**
  - ✓ Attempt to link infant and mom hospitalizations that did not link to a maternal record from Stage #1 directly to the birth record

Example of Overarching Linkage Approach

Live birth records to inpatient birth hospitalizations in 2008

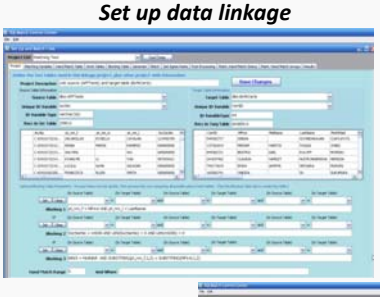


Software To Facilitate Data Linkage

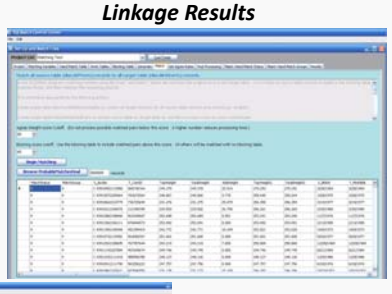
- LinkSolv
- AutoMatch
- LinkageWiz
- FRIL
- LinkPlus
- Link King
- SQL Match
- FEBRL
- SQL Server (SSIS)
- **SAS**, SPSS, Stata, S-Plus, R
- …many more!

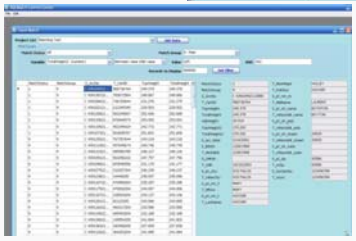
SQL Match

Set up data linkage



Linkage Results

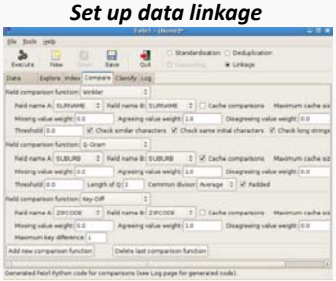




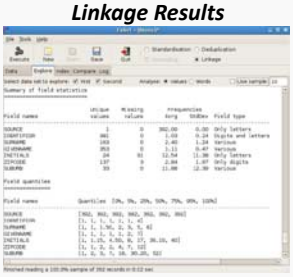
Manual Review

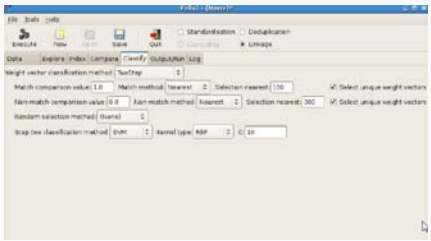
Freely Extensible Biomedical Record Linkage

Set up data linkage

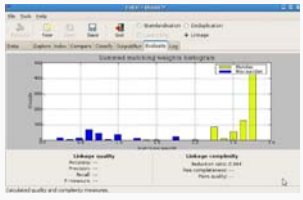


Linkage Results



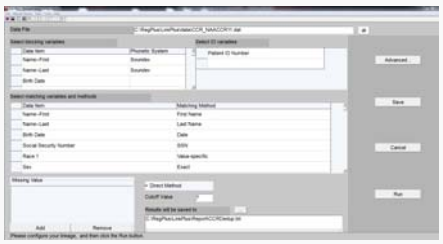


Manual Review and Summary

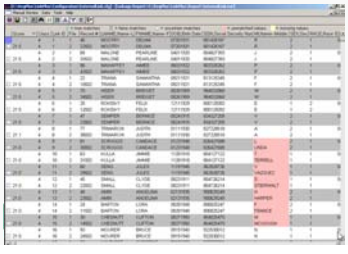


Link Plus

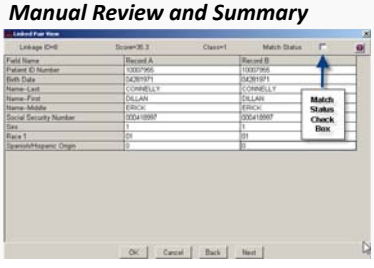
Set up data linkage



Linkage Results



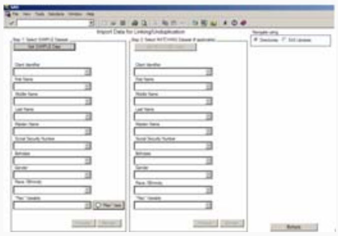
Manual Review and Summary

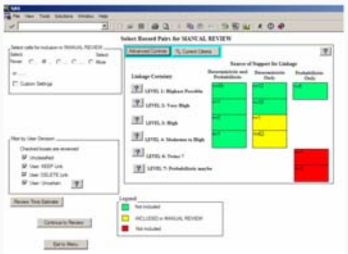


Link King

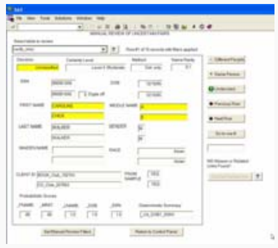
Set up data linkage

Select Variables



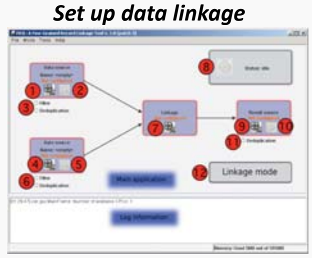


Manual Review

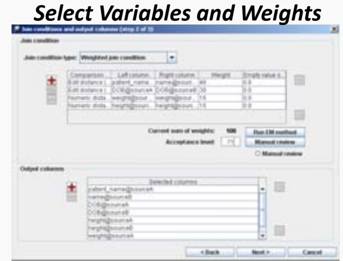


Fine-Grained Record Linkage (FRIL)

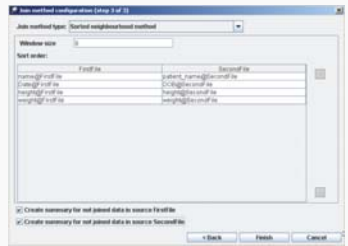
Set up data linkage



Select Variables and Weights

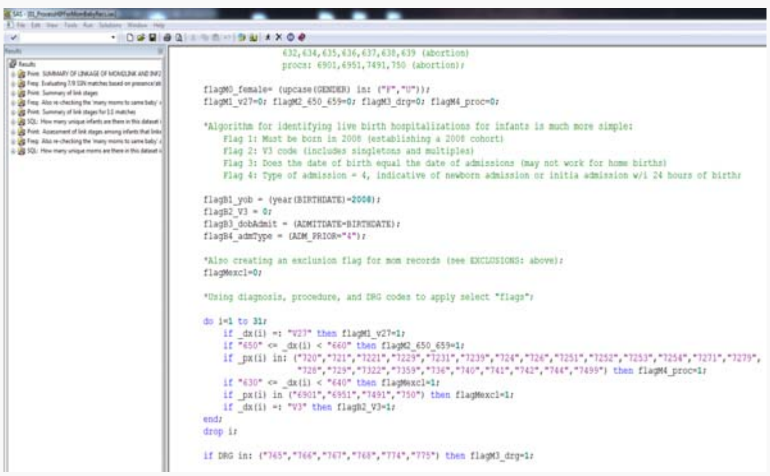


Join Method (Blocking, SNM)



Manual Review

But our choice…SAS



```
                632,634,635,636,637,638,639 (abortion)
                procs: 6901,6951,7491,750 (abortion);

flagM0_female= (upcase(GENDER) in: ("F","U"));
flagM1_v27=0; flagM2_650_659=0; flagM3_drg=0; flagM4_proc=0;

*Algorithm for identifying live birth hospitalizations for infants is much more simple:
    Flag 1: Must be born in 2008 (establishing a 2008 cohort)
    Flag 2: V3 code (includes singletons and multiples)
    Flag 3: Does the date of birth equal the date of admissions (may not work for home births)
    Flag 4: Type of admission = 4, indicative of newborn admission or initia admission w/i 24 hours of birth;

flagB1_yob = (year(BIRTHDATE)=2008);
flagB2_V3 = 0;
flagB3_dobAdmit = (ADMITDATE=BIRTHDATE);
flagB4_admType = (ADM_PRIOR="4");

*Also creating an exclusion flag for mom records (see EXCLUSIONS: above);
flagMexcl=0;

*Using diagnosis, procedure, and DRG codes to apply select "flags";

do i=1 to 31;
    if _dx(i) =: "V27" then flagM1_v27=1;
    if "650" <= _dx(i) < "660" then flagM2_650_659=1;
    if _px(i) in: ("720","721","7221","7229","7231","7239","724","726","7251","7252","7253","7254","7271","7279",
                "728","729","7322","7359","736","740","741","742","744","7499") then flagM4_proc=1;
    if "630" <= _dx(i) < "640" then flagMexcl=1;
    if _px(i) in ("6901","6951","7491","750") then flagMexcl=1;
    if _dx(i) =: "V3" then flagB2_V3=1;
end;
drop i;

if DRG in: ("765","766","767","768","774","775") then flagM3_drg=1;
```

USF HEALTH

Linking Mechanics

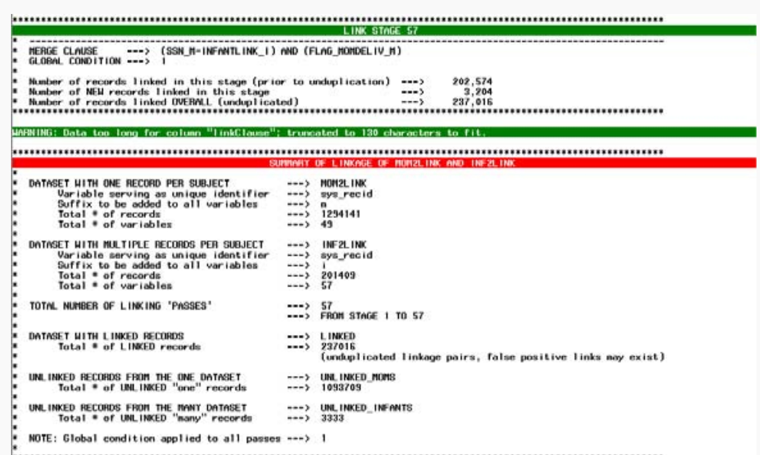
- Developed a SAS macro
- **Hierarchical**, **stepwise** series of linking stages, using various combinations of variables, proceeding from highest to lowest confidence
  - ✓ Exact and partial matching, **linking *with* replacement**
  - ✓ Primarily **deterministic**, includes **probabilistic** elements
  - ✓ **CREATES** potential matches
- Coding algorithm to calculate a "linking confidence" score to **GRADE** matches
  - ✓ Also incorporate a "delivery confidence" score
- Records above a certain score are **SELECTED** as links, borderline scores require **manual validation**
  - ✓ May find false + we need to **CORRECT**
  - ✓ We try to minimize manual review

USF HEALTH

Linking Mechanics

- We do not use **blocking**
  - ✓ Too concerned about flawed data
  - ✓ Linking approximately 230,000 birth hospitalization records to approximately 1.4 million "women" records using the merging macro takes approximately 1 hour
    - – Will sacrifice extra time for greater sensitivity
- SAS
  - ✓ Not as automated or "point-and-click" as other software
  - ✓ Extremely **customizable** through coding
  - ✓ Easy to incorporate a large number of variables (Link King)
  - ✓ Easy to allow **"crossover" links**
    - – Mom's SSN in AHCA links to father's SSN in vital stats
  - ✓ Can process extremely large datasets quickly given powerful computers

USF HEALTH



```
****************************************************************************************
                                   LINK STAGE 57
*  -------------------------------------------------------------------------------------
*  MERGE CLAUSE     ---> (SSN_M=INFANTLINK_I) AND (FLAG_MOMDELIV_M)
*  GLOBAL CONDITION ---> 1
*
*  Number of records linked in this stage (prior to unduplication) --->    202,574
*  Number of NEW records linked in this stage                      --->      3,204
*  Number of records linked OVERALL (unduplicated)                 --->    237,016
****************************************************************************************

WARNING: Data too long for column "linkClause"; truncated to 130 characters to fit.

****************************************************************************************
                      SUMMARY OF LINKAGE OF MOM2LINK AND INF2LINK
*
*  DATASET WITH ONE RECORD PER SUBJECT          ---> MOM2LINK
*       Variable serving as unique identifier   ---> sys_recid
*       Suffix to be added to all variables     ---> m
*       Total # of records                      ---> 1294141
*       Total # of variables                    ---> 49
*
*  DATASET WITH MULTIPLE RECORDS PER SUBJECT    ---> INF2LINK
*       Variable serving as unique identifier   ---> sys_recid
*       Suffix to be added to all variables     ---> i
*       Total # of records                      ---> 201409
*       Total # of variables                    ---> 57
*
*  TOTAL NUMBER OF LINKING 'PASSES'             ---> 57
*                                               ---> FROM STAGE 1 TO 57
*
*  DATASET WITH LINKED RECORDS                  ---> LINKED
*       Total # of LINKED records               ---> 237016
*                                                    (unduplicated linkage pairs, false positive links may exist)
*
*  UNLINKED RECORDS FROM THE ONE DATASET        ---> UNLINKED_MOMS
*       Total # of UNLINKED "one" records       ---> 1093709
*
*  UNLINKED RECORDS FROM THE MANY DATASET       ---> UNLINKED_INFANTS
*       Total # of UNLINKED "many" records      ---> 3333
*
*  NOTE: Global condition applied to all passes ---> 1
*
****************************************************************************************
```

USF HEALTH

Additional Challenges

- **Disentangling multiples** (twins, triplets, etc)
  - ✓ No infant SSN, no names in hospital data
  - ✓ Multiples will share all mom characteristics
  - ✓ Ordering of variables in AHCA does not match birth order
  - ✓ Can use **sex** to differentiate between opposite-sex dizygotic twins
  - ✓ Can use diagnosis codes that reflect 500 gram **birth weight categories** to disentangle same sex multiples that may differ in birth weight
  - ✓ For multiples that have the same sex, similar birth weights, it may be impossible to determine, given the available data, which hospital record goes with which birth record
    - Investigating other options
      - Random assignment
      - Allocation to "family" as unit

Next Steps

- Finalize enhancements to linkage of birth record to birth hospitalizations
- Link in **post-birth hospitalizations**, ambulatory records, and ED data
  - ✓ Challenging for those with missing/incorrect maternal SSN
- Develop an identifier crosswalk to link in **cost-to-charge ratios (CCRs)** from CMS to convert hospital charges to costs