

PERFORMANCE EVALUATION OF THE STRUCTURED
POPULATION EVENT HISTORY REGISTER (SPEHR)
RELATIONAL SCHEMA FOR MANAGING LONGITUDINAL
HEALTH AND POPULATION DATA

Africa Centre for Health and Population Studies: June 30, 2009

Benjamin D. Clark
Samuel J. Clark

London School of Hygiene and Tropical Medicine
University of Washington, University of Colorado at Boulder, University of the Witwatersrand

We appreciate the generous support provided by:

- Agincourt Health and Demographic Surveillance System Site, South Africa
- London School of Hygiene and Tropical Medicine
- University of Washington
- United States National Institutes of Health grants:
 - 1 R03 AG028640 1
 - 1 K01 HD057246-01
 - 1 R01 HD054511-01 A1

Specific Questions

- Will SPEHR accommodate the full range of longitudinal data in a real DSS of non-trivial size and duration?
- Are data equivalent in the non-SPEHR and SPEHR schemas?
- Is it possible to conduct the same analysis on the two schemas and produce identical results?
- Is performance of the SPEHR-based database acceptable

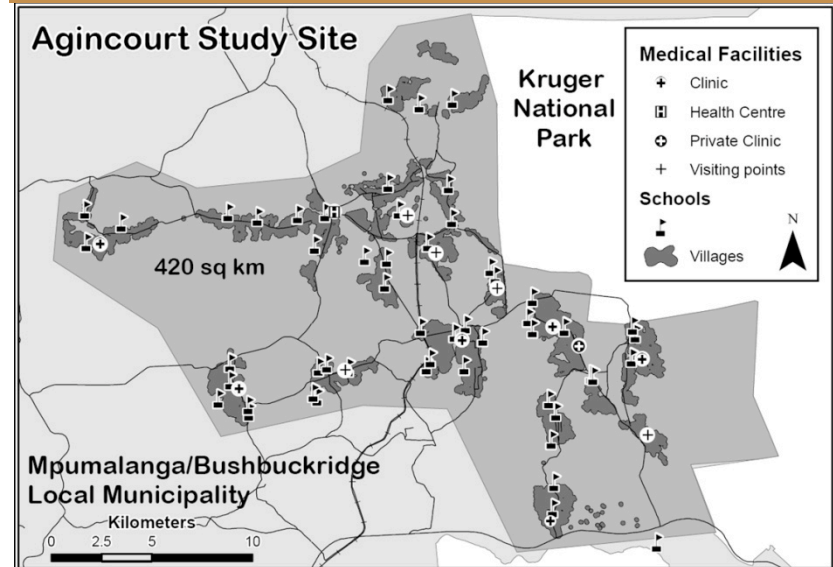
Outline

- Introduction to Agincourt HDSS
- Introduction to the Agincourt data model (RDM) and the SPEHR data model
- Sample database (RDM) construction used for this project
- Schema conversion process
- Sample data extraction
- Retrieval performance

Agincourt HDSS

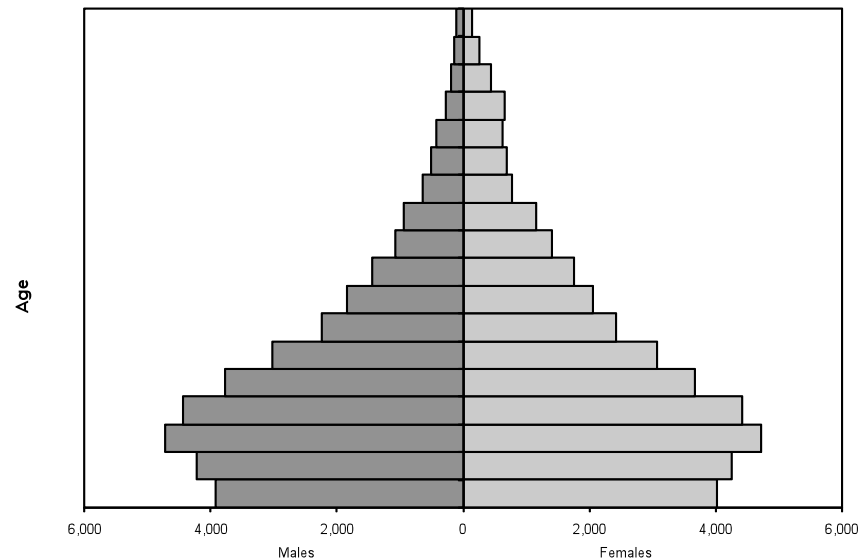
- Demographic surveillance site in northeast South Africa
- Data collecting through annual census rounds starting in 1992
- ~70,000 people under observation
- Data describes vital events; births deaths, migrations, and more
- Additional population-wide data collected at individual and population levels
- Additional nested projects collect data using DSS population as a sampling frame for targeted studies.
- For more info visit: <http://www.agincourt.co.za/DataSection/index.htm>

Agincourt Study Site (South Africa)



Agincourt population (2004)

Agincourt Population Pyramid for 2004



Basic Structure of Reference Data Model (RDM) Schema

- Traditional relation structure of DSS data: Reference Data Model (RDM)
 - Episode (interval)-based with geographic residency and social group membership tables
 - One table per event; birth, death, migrations (in/out)
 - Individuals table
 - Dwelling locations and village tables for geography
 - Individual and household status observations for additional data

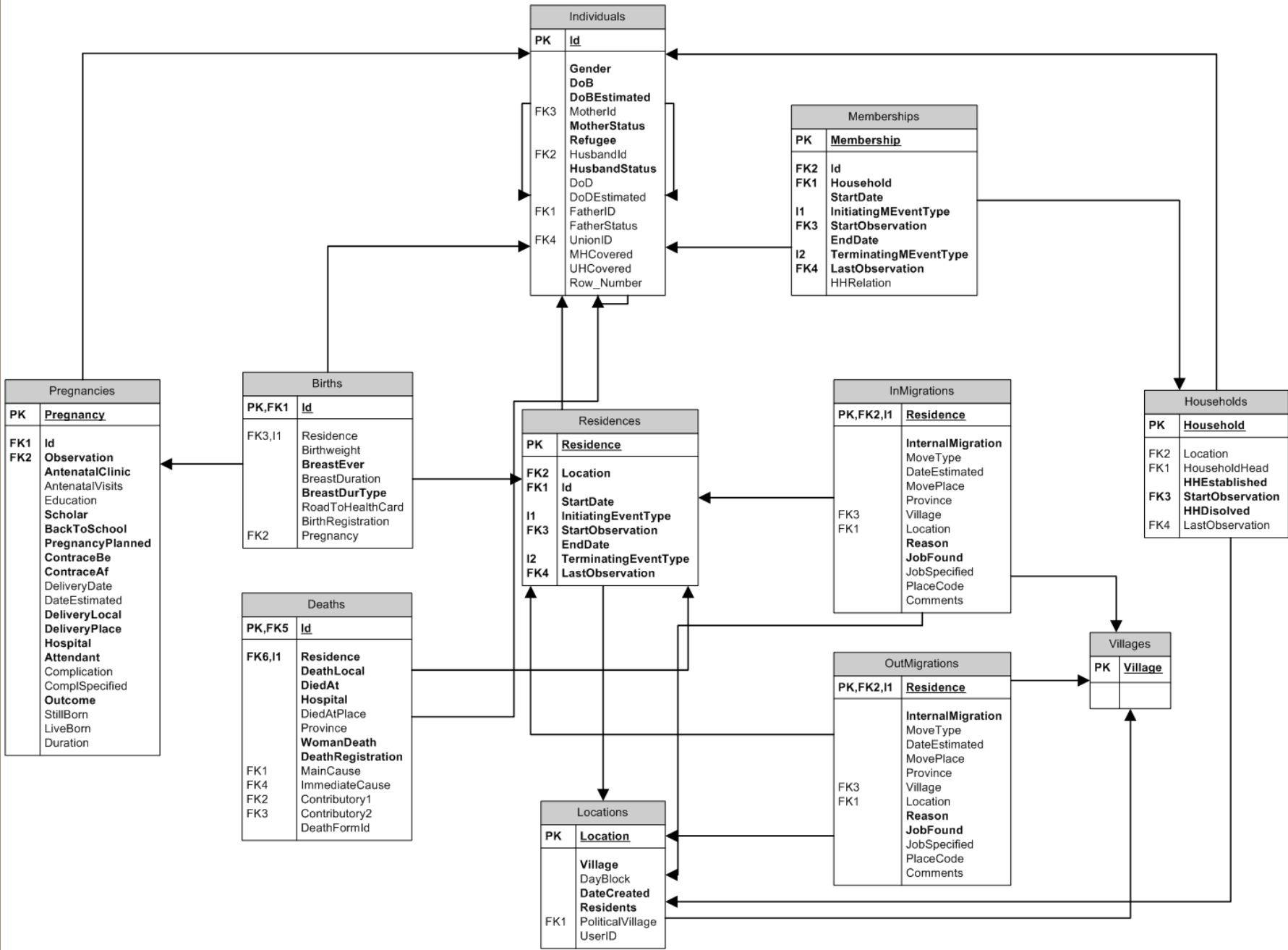
Advantages of RDM

- Easy to conceptualize
- Reasonably fast data retrieval

Disadvantages of RDM

- Dates de-normalized, i.e. duplicated everywhere
 - Date of birth in 4 different tables
 - Date of death in up to 4 different tables
 - Migration dates in up to 3 different tables
- Difficult to extend the model to collect new event types or exposures
- Difficult to reconcile duplicated entities
- Requires change to data structure for each additional set of data
- Difficult to temporally constrain
- Not self-documenting
- Difficult to meta-program against (multiple paths between tables)

BASIC RDM DATA MODEL (SIMPLIFIED)



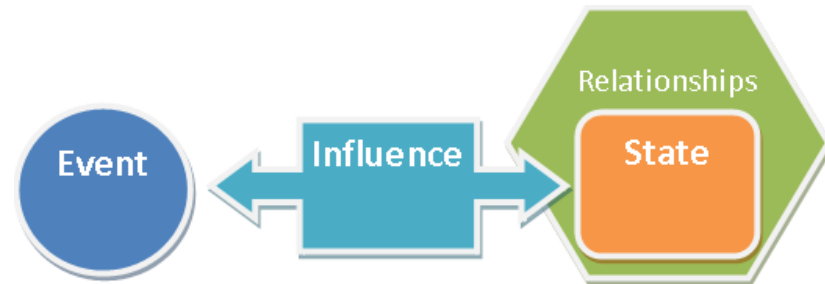
Basic Structure of Structured Population Event History Register (SPEHR) Schema

States: Temporally persisting objects, e.g. people, places, social groups

Events: Time points that mark change, e.g. births deaths, status observations

Influences: specify what influence an event has on an instance of a state. Events can influence multiple states in different ways but can have only one influence on any one instance of a state; i.e. mother gives birth, baby is born, father gains a child, social group gains a member

Relationships: non temporal collections of states with specific roles in the defined collections

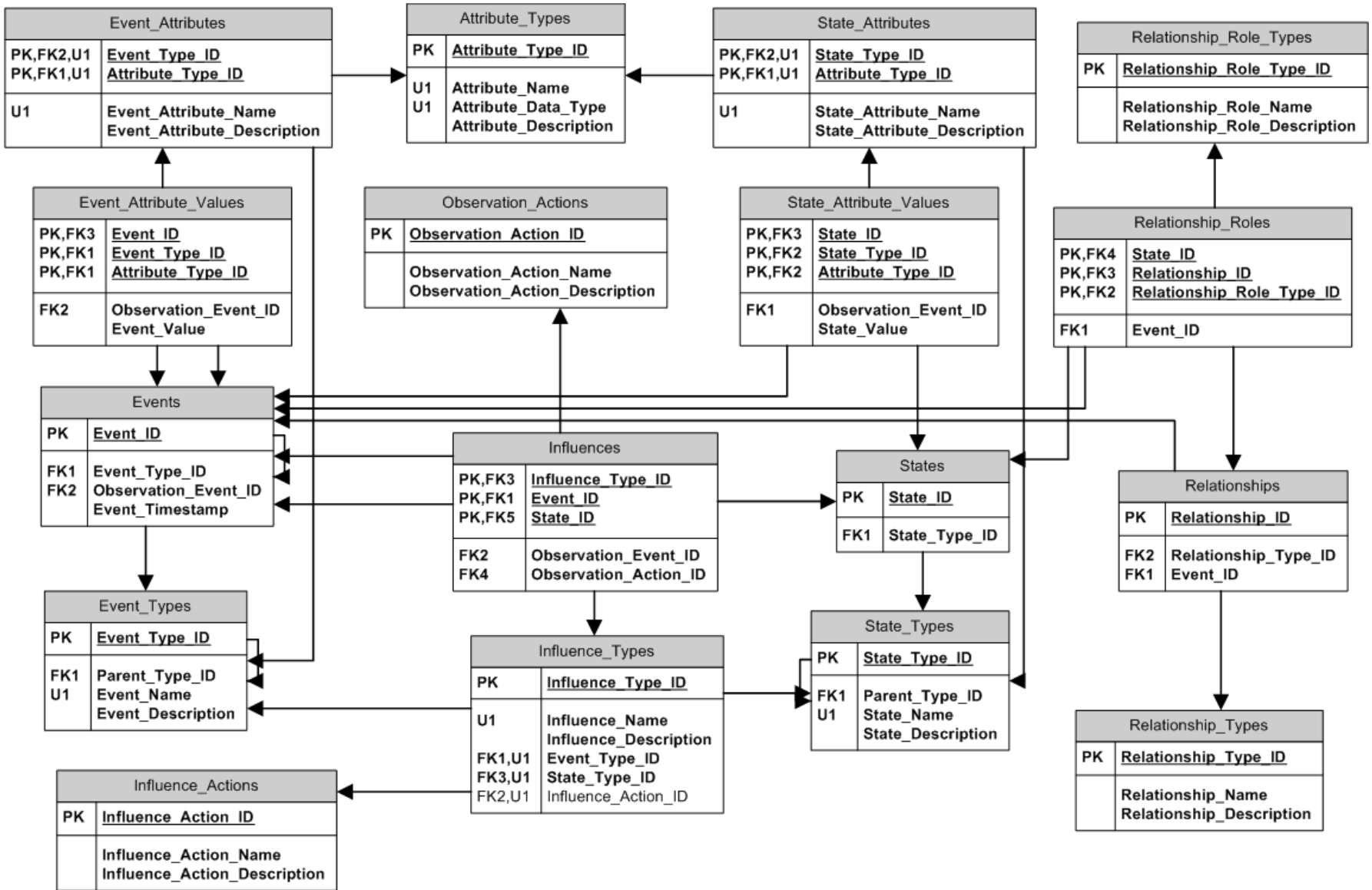


Advantages of SPEHR

- Temporally normalized with no event date duplication
- Add dates in a single table allowing for metadata-driven temporal constraints
- Static structure while allowing for dynamic data, easy to extend with new data collections without changing the data model (describe and store) → schema invariance
- Self documenting
- Easy to program against in a metadata-driven style

Disadvantages of SPEHR

- Slow for data extractions
- Conceptually more difficult, i.e. very abstract
- Storage size larger (could be optimized)



Reason for using a sample database

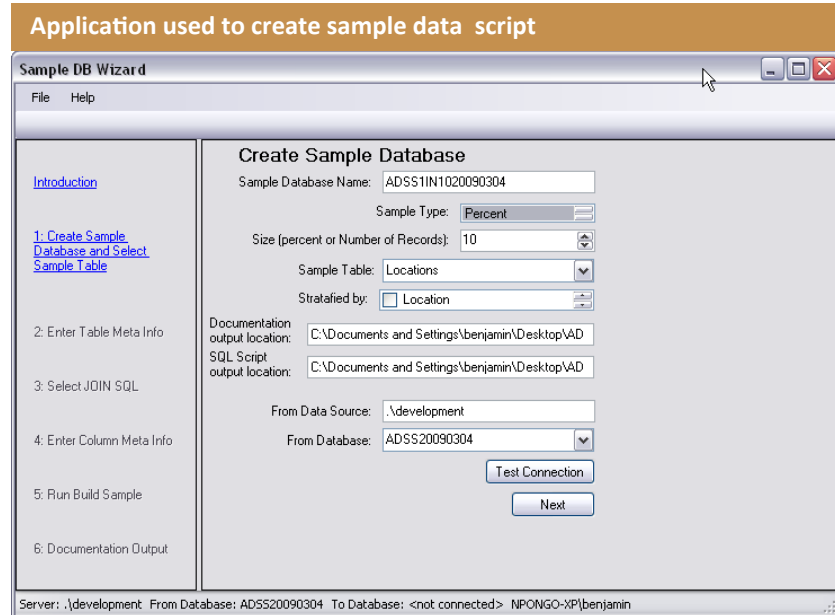
- Smaller and more manageable for development purpose. Faster build times when debugging conversion process
- Anonymous so can be shared as output of this project

Objective of the sample database

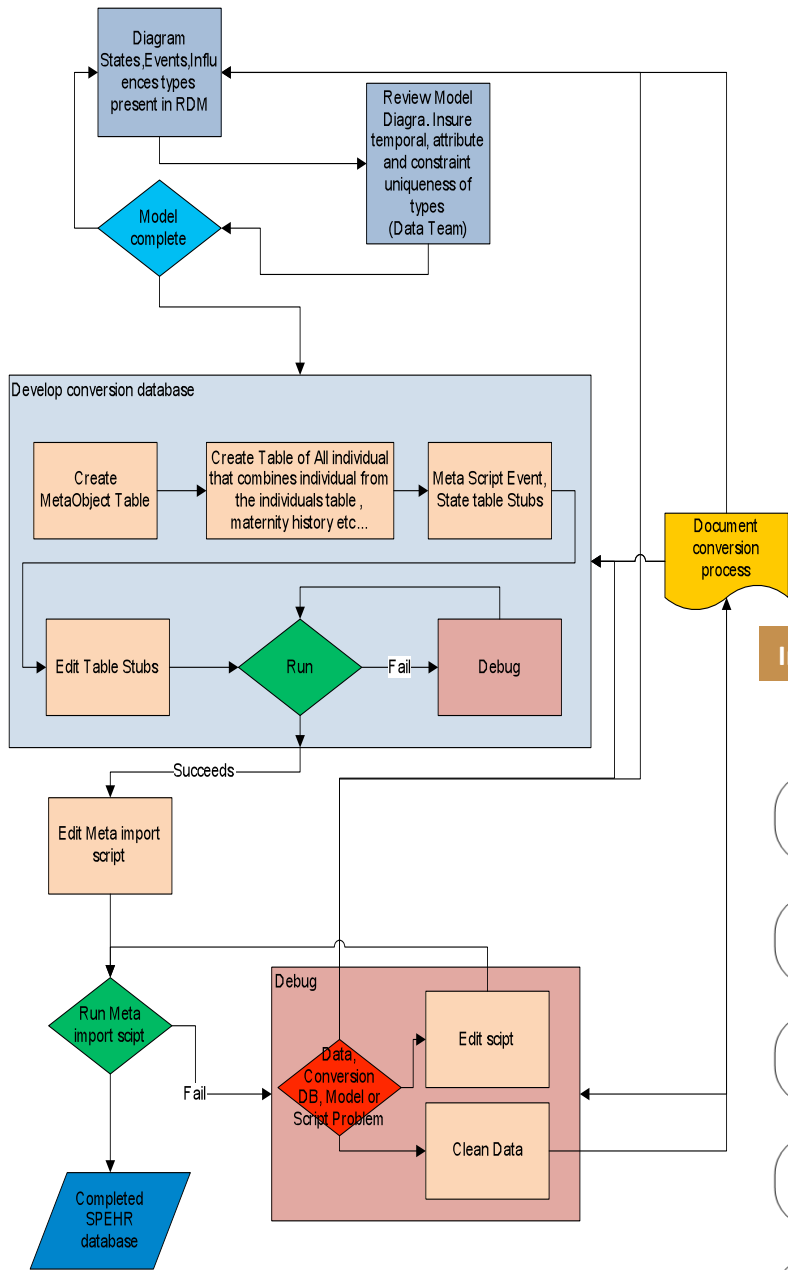
- Smaller and more manageable for development purposes. Faster build times when debugging a conversion process
- Anonymous so can be shared as output of this project

Process

- Sample **locations** table
 - Represents the unit of observation during data collection
 - Stratified by village
- Decide tables to include in the sample database, either as sampled or in entirety, e.g. code tables
- Decide columns of each table to include, excluding identifiers and sensitive variables. Anonymize other columns using various methods, including all table primary keys so that the sample database cannot be joined back onto the original.
- Select appropriate JOIN to join on base sample table (location)

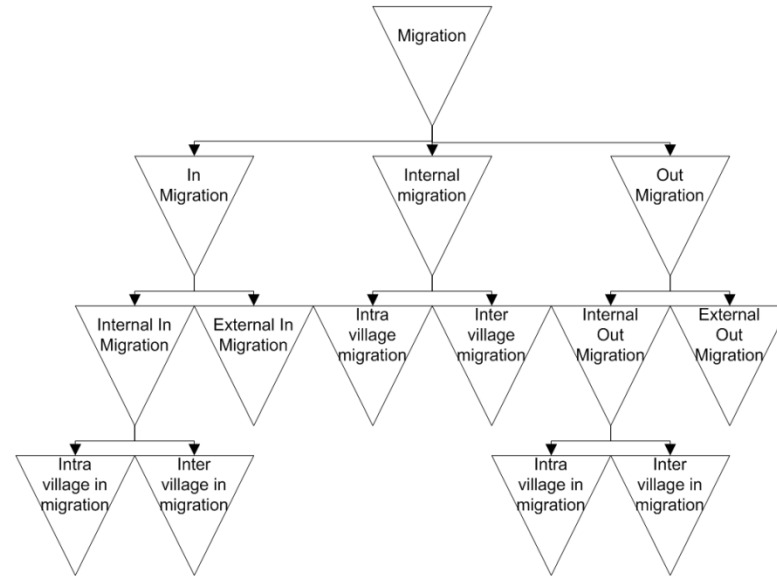


Conversion Process model



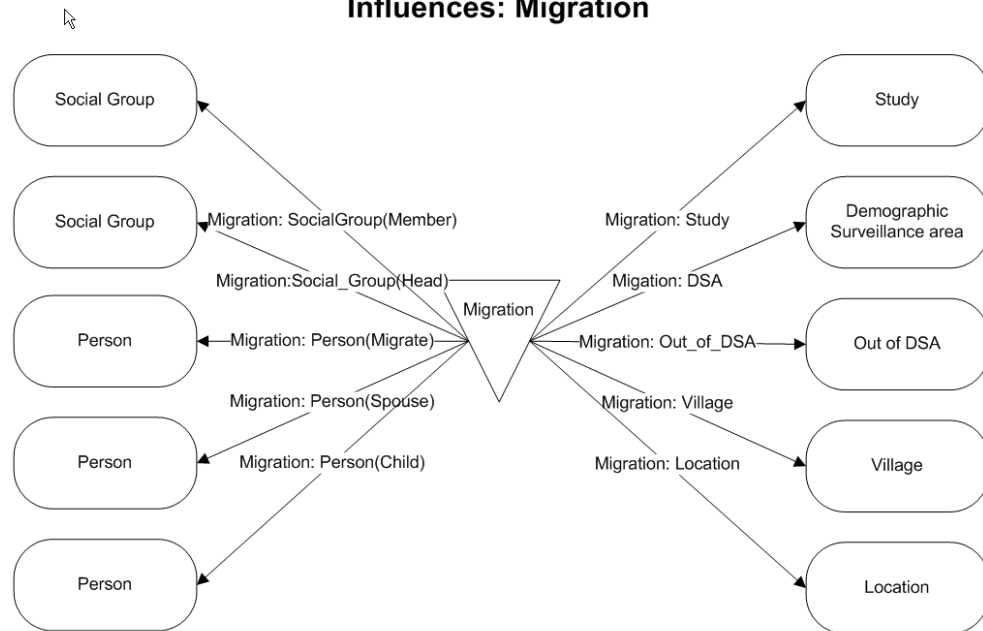
Event Hierarchy Model

Event: Migrations



Influence

Influences: Migration



Methods and Results

Methods

Produce the INDEPTH minimum data set of person years by age and sex by year for each year of the study and event counts. The dataset allows for the most basic demographic rates and statistics to be calculated.

Real difference in scripts used to extract data sets was the need to create the residence episode table in SPEHR, which already exists in the RDM. After that the scripts were basically identical for person year calculations.

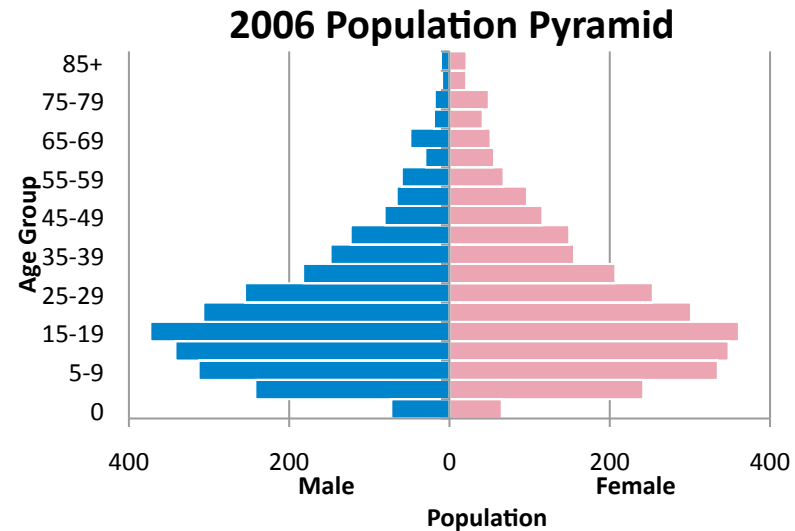
Event counts were simpler in SPEHR.

This process exposed dirty data in the RDM**Results:**

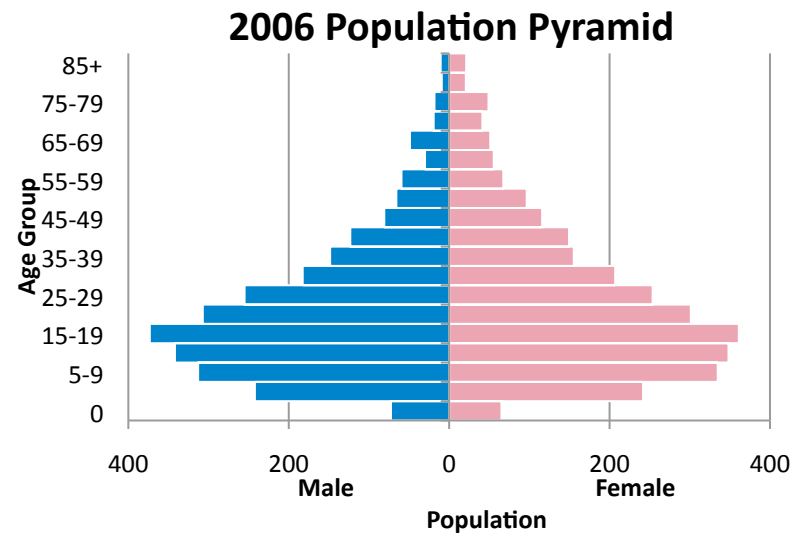
Identical dataset where event counts were concerned. Slight differences in person years for a few records in the 4th decimal place.

SPEHR took about twice as long to calculate, 45 sec (RDM) to 1.5 min (SPEHR)

RDM 2006 Population Pyramid



SPEHR 2006 Population Pyramid



Performance Test results

Methods

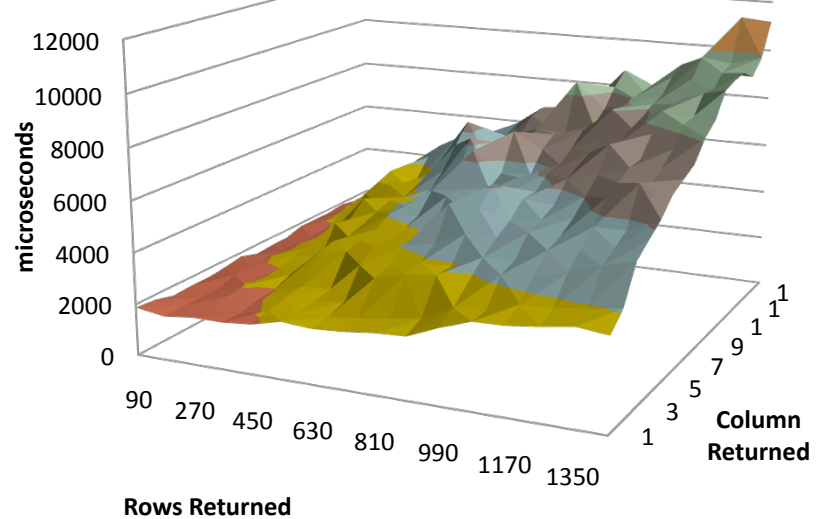
Averaged 100 retrievals of 1 to 16 columns of data by 16 different levels of rows returned ranging from 90 to 1430

Results

- SPEHR much slower but feasible
- SPEHR scales better (implications for large repositories used to pool multisite data)
- In the real world, SPEHR seems to take about twice as long as the RDM

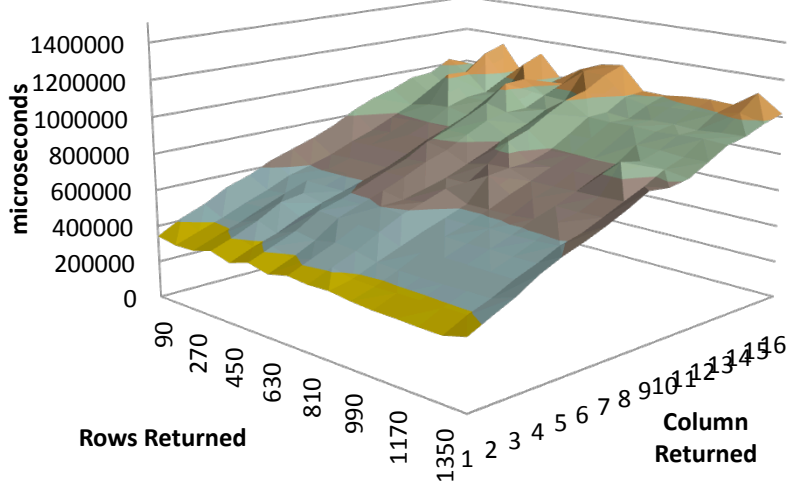
RDM Data Retrieval Performance

Average of 100 runs after caching



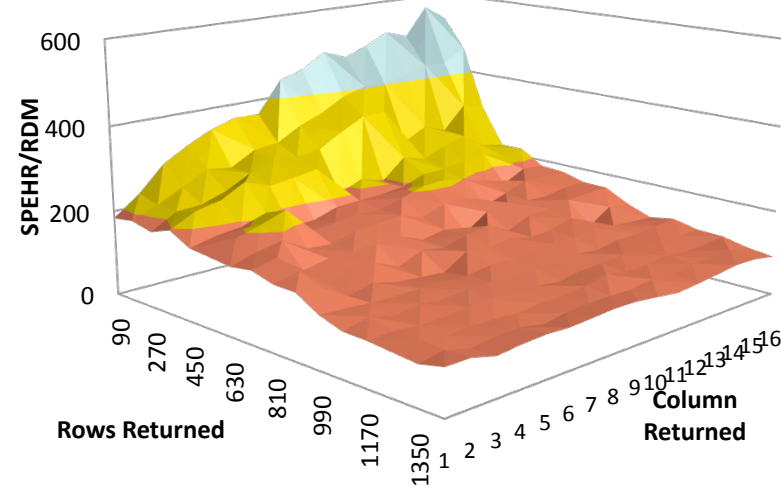
SPEHR Data Retrieval Performance

Average of 100 runs after caching



Comparison of RDM to SPEHR

Ratio of SPEHR To RDM



Answers to Specific Questions

- Will SPEHR accommodate the full range of longitudinal data in a real DSS of non-trivial size and duration ?

YES

- Are data equivalent in the non-SPEHR and SPEHR schemas ?

YES

- Is it possible to conduct the same analysis on the two schemas and produce identical results ?

YES

- Is performance of the SPEHR-based database acceptable ?

YES, but it is slow for small databases