The Rakai Health Sciences Program (RHSP): Developing *a Data Warehouse*

Rakai Health Sciences Program
Improved Health Through Research

By: ANTHONY NDYANABO

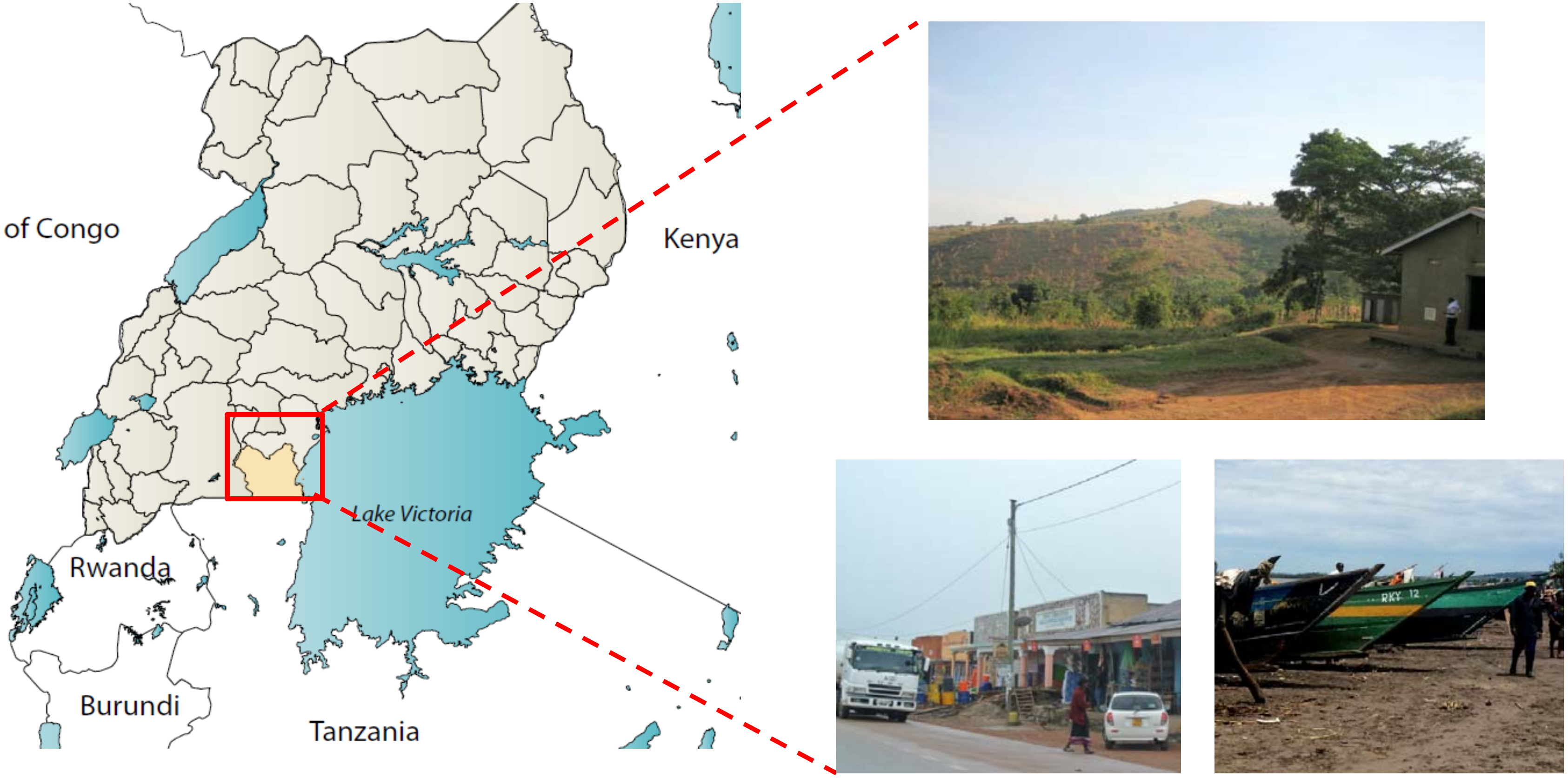
The Rakai Health Sciences Program (RHSP)

- The RHSP is currently one of the largest and oldest community-based non-profit research endeavors on HIV/AIDS, associated infections and reproductive health, in Africa.

- The program is currently conducting HIV prevention studies, basic and clinical laboratory research, and operations research/evaluation to develop improved strategies of service delivery in southern Uganda.

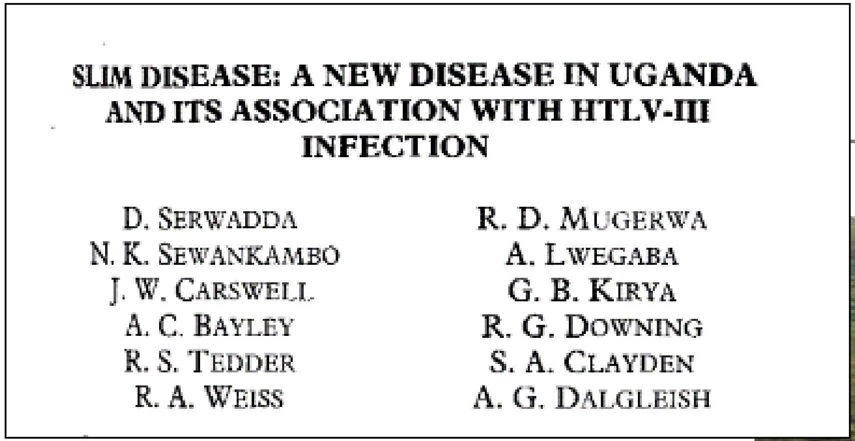
Rakai Region, Uganda

- Agrarian, trading, and fishing communities.



Why Rakai?

- D Serwadda, N Sewankambo, *et al* identified “slim disease” in Rakai in 1982.



SLIM DISEASE: A NEW DISEASE IN UGANDA AND ITS ASSOCIATION WITH HTLV-III INFECTION

D. Serwadda	R. D. Mugerwa
N. K. Sewankambo	A. Lwegaba
J. W. Carswell	G. B. Kirya
A. C. Bayley	R. G. Downing
R. S. Tedder	S. A. Clayden
R. A. Weiss	A. G. Dalgleish

“A new disease has recently been recognized in rural Uganda.”

Serwadda D, Sewankambo N, et al. Lancet 1985.

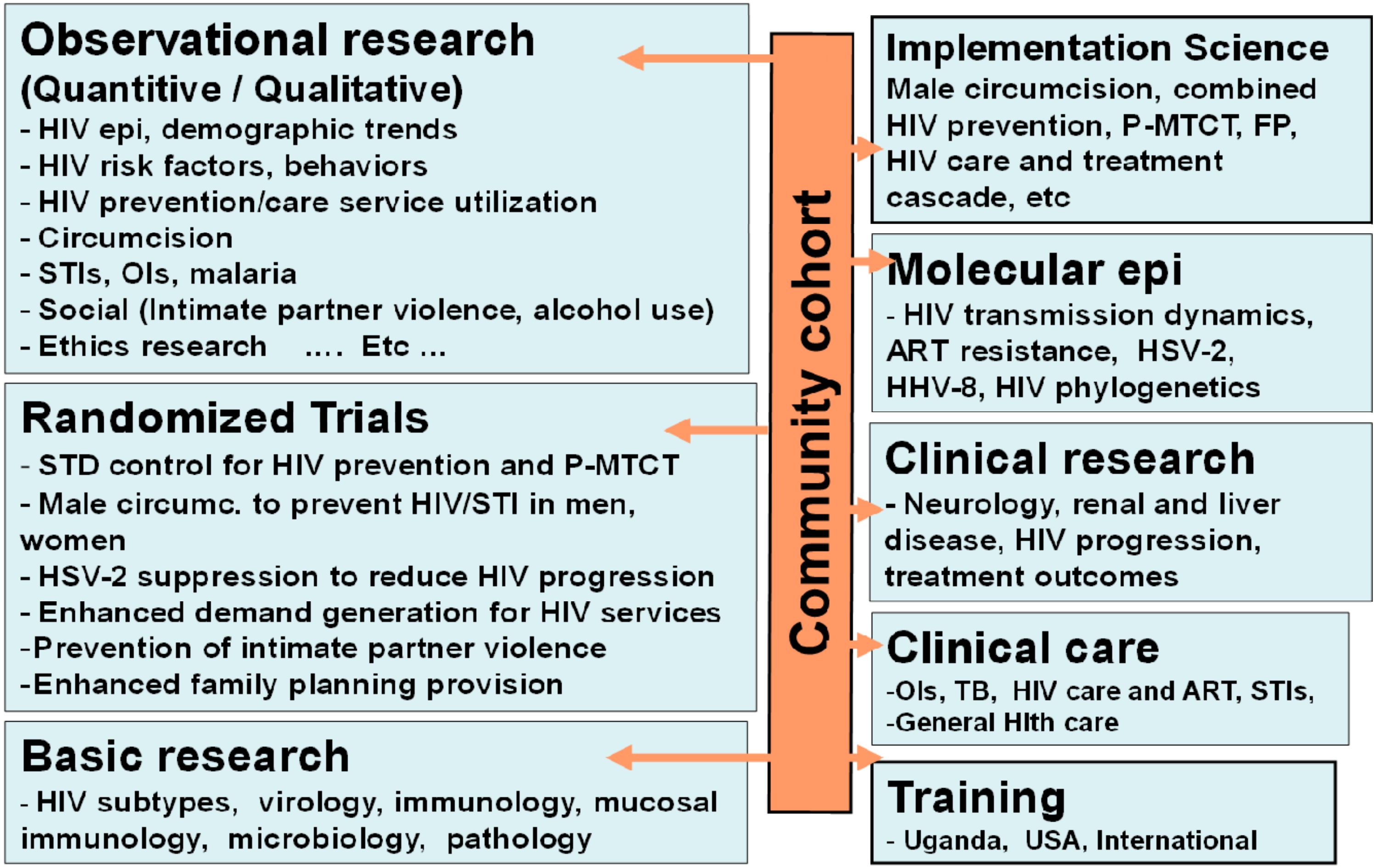
The Rakai Community Cohort Study (RCCS)

- Open population-based **household census and cohort survey** of 40 communities ongoing since 1994 conducted by the RHSP

- **Household Census**
  - All household residents included irrespective of age
  - Data obtained: residence status, births and deaths, family relationships, household assets/wealth, GPS coordinates for households and local built environment (e.g. schools, bars, hotels).

- **RCCS Cohort Survey**
  - Currently restricted to ages 15-49 years
  - Questionnaire: Detailed demographic, sexual network and behaviors, health status, and health care utilization questions
  - Specimen collection: Serum/plasma for HIV, STI, and other tests; buffy coat for human genetic testing

Rakai Community Cohort Study (RCCS)



Observational research (Quantitive / Qualitative)
- HIV epi, demographic trends
- HIV risk factors, behaviors
- HIV prevention/care service utilization
- Circumcision
- STIs, OIs, malaria
- Social (Intimate partner violence, alcohol use)
- Ethics research …. Etc …

Randomized Trials
- STD control for HIV prevention and P-MTCT
- Male circumc. to prevent HIV/STI in men, women
- HSV-2 suppression to reduce HIV progression
- Enhanced demand generation for HIV services
-Prevention of intimate partner violence
-Enhanced family planning provision

Basic research
- HIV subtypes, virology, immunology, mucosal immunology, microbiology, pathology

Community cohort

Implementation Science
Male circumcision, combined HIV prevention, P-MTCT, FP, HIV care and treatment cascade, etc

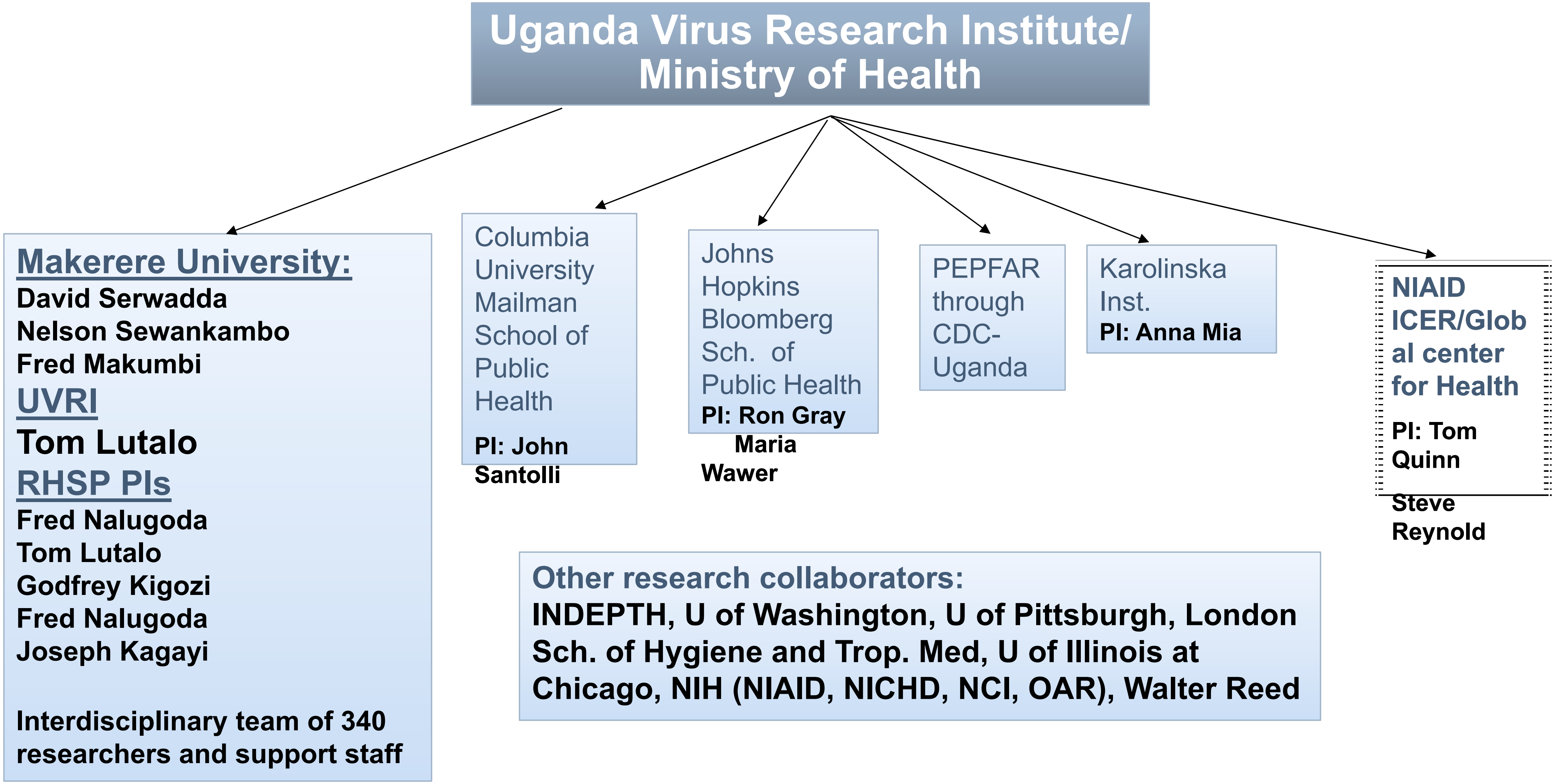
Molecular epi
- HIV transmission dynamics, ART resistance, HSV-2, HHV-8, HIV phylogenetics

Clinical research
- Neurology, renal and liver disease, HIV progression, treatment outcomes

Clinical care
-OIs, TB, HIV care and ART, STIs,
-General Hlth care

Training
- Uganda, USA, International

RHSP: A global collaboration



Uganda Virus Research Institute/ Ministry of Health

Makerere University:
David Serwadda
Nelson Sewankambo
Fred Makumbi
UVRI
Tom Lutalo
RHSP PIs
Fred Nalugoda
Tom Lutalo
Godfrey Kigozi
Fred Nalugoda
Joseph Kagayi

Interdisciplinary team of 340 researchers and support staff

Columbia University Mailman School of Public Health
PI: John Santolli

Johns Hopkins Bloomberg Sch. of Public Health
PI: Ron Gray
Maria Wawer

PEPFAR through CDC-Uganda

Karolinska Inst.
PI: Anna Mia

NIAID ICER/Global center for Health

PI: Tom Quinn
Steve Reynold

Other research collaborators:
INDEPTH, U of Washington, U of Pittsburgh, London Sch. of Hygiene and Trop. Med, U of Illinois at Chicago, NIH (NIAID, NICHD, NCI, OAR), Walter Reed

RCCS Data

- Data collected thus far:
  - 18 completed survey/census rounds (19th ongoing)
  - ~22,000 study participants currently
  - ~80,000 total participants
  - >500K archived laboratory samples
- Historically, data for each survey round was stored in own separate FoxPro Database (hundreds of tables!)
- Data requests can be overwhelming
  - Assembling longitudinal data sets challenging!

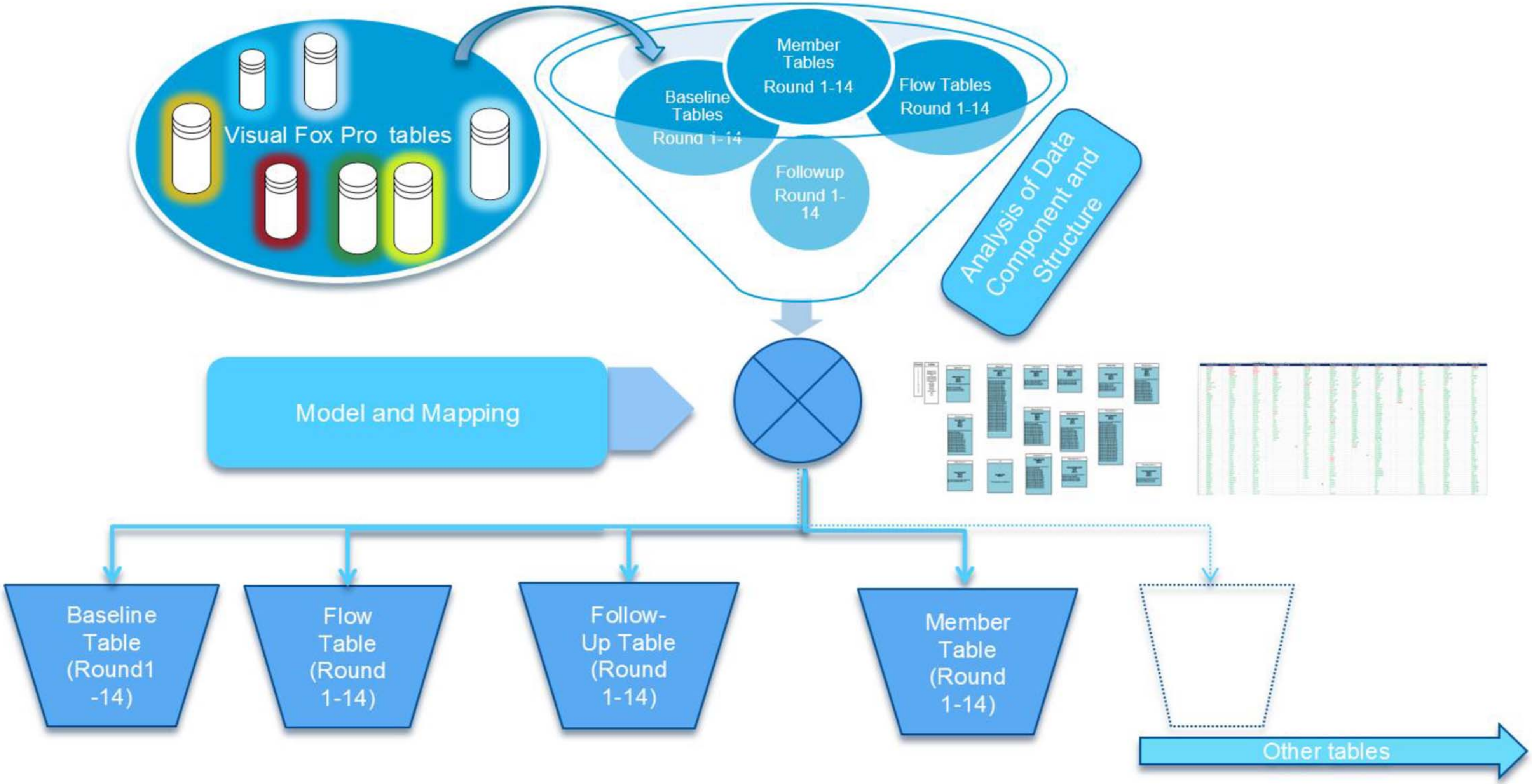
Why the data warehouse?

- To consolidate the available cohort study data rounds into a single data repository.

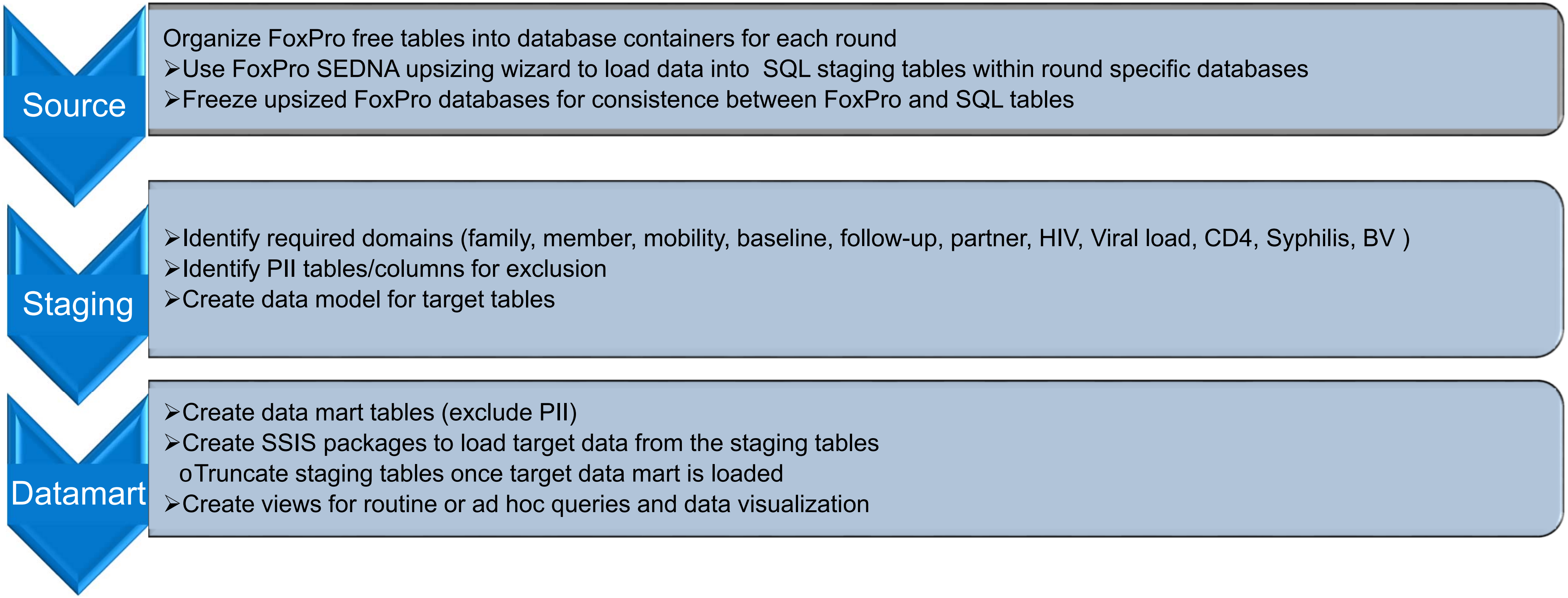
Specific objectives:

- Enable RHSP researchers and statisticians correlate and analyze all study round data from one source for timely fulfillment of data requests.
- Develop a framework to perform this task repeatedly as additional survey, laboratory and clinical data is collected.
- Standardize data retrieval and ensure reproducible reports.
- Enforce data access controls to ensure data fidelity overtime.
- Proper documentation for easy variable definition and data auditing.

Conceptual Framework


Visual Fox Pro tables
Member Tables Round 1-14
Baseline Tables Round 1-14
Flow Tables Round 1-14
Followup Round 1-14
Analysis of Data Component and Structure
Model and Mapping
Baseline Table (Round1 -14)
Flow Table (Round 1-14)
Follow-Up Table (Round 1-14)
Member Table (Round 1-14)
Other tables

Implementation(ETL)



Source

Organize FoxPro free tables into database containers for each round

- Use FoxPro SEDNA upsizing wizard to load data into SQL staging tables within round specific databases
- Freeze upsized FoxPro databases for consistence between FoxPro and SQL tables

Staging

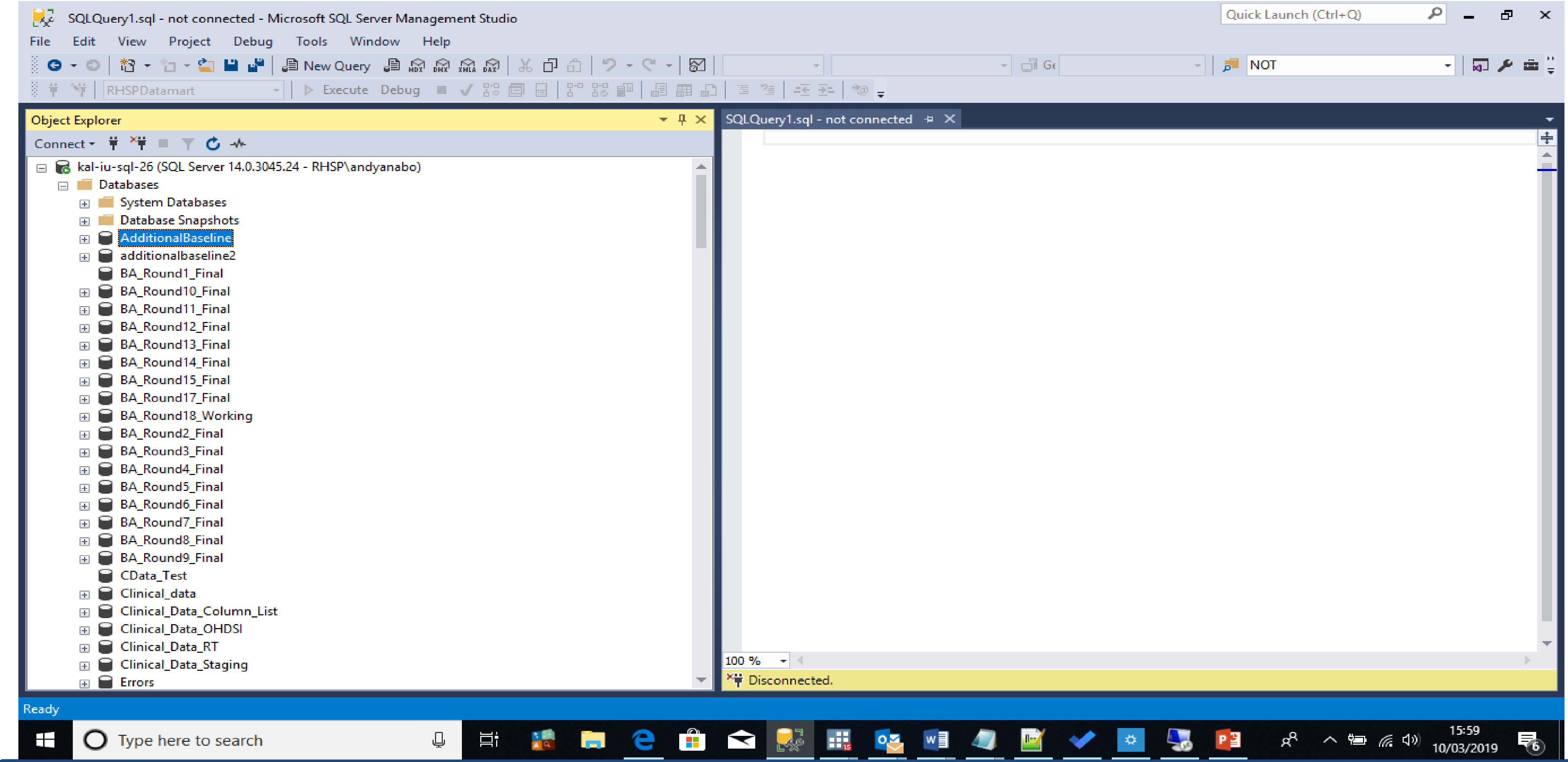
- Identify required domains (family, member, mobility, baseline, follow-up, partner, HIV, Viral load, CD4, Syphilis, BV)
- Identify PII tables/columns for exclusion
- Create data model for target tables

Datamart

- Create data mart tables (exclude PII)
- Create SSIS packages to load target data from the staging tables
  - Truncate staging tables once target data mart is loaded
- Create views for routine or ad hoc queries and data visualization

Staging tables

Rakai Health Sciences Program
Improved Health Through Research



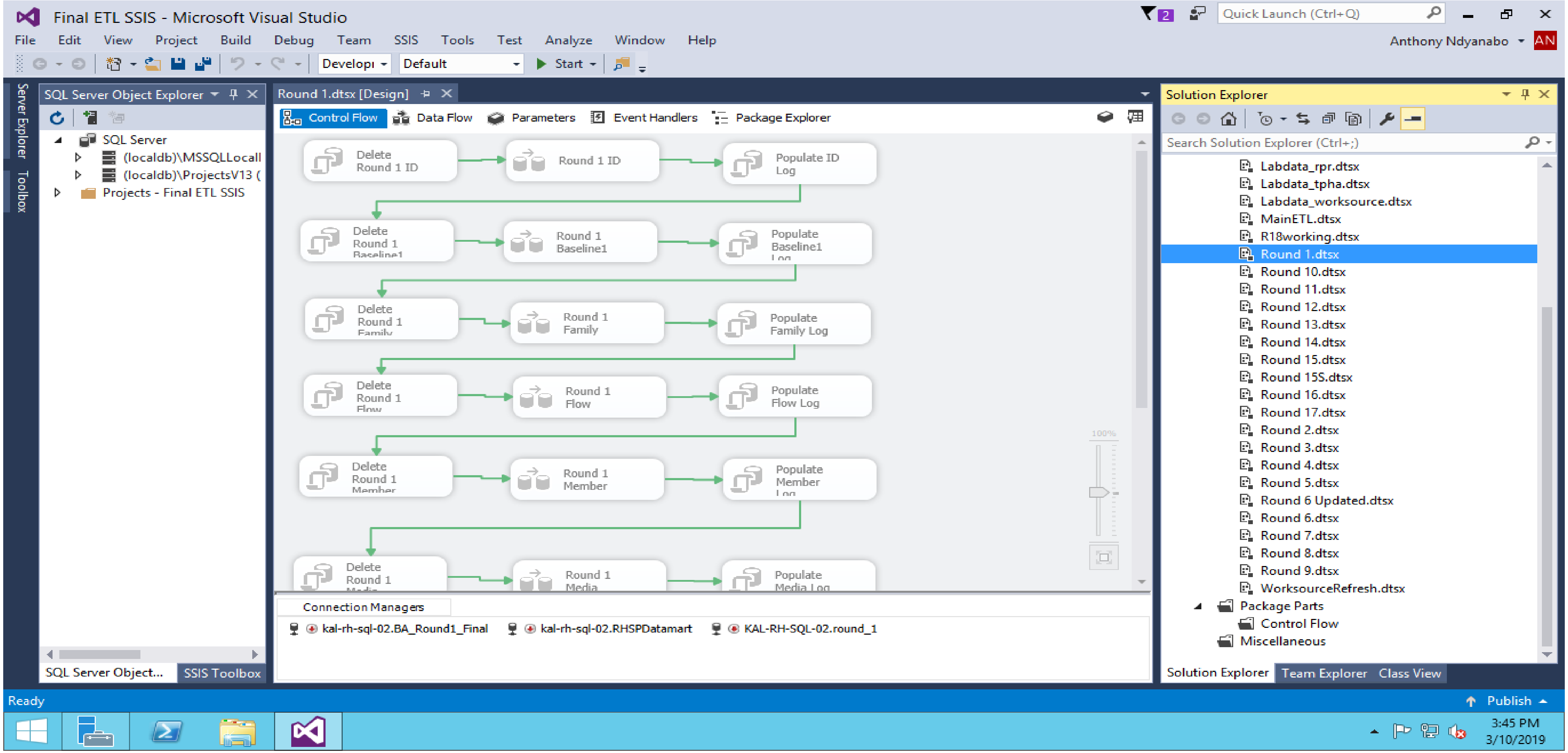
SQLQuery1.sql - not connected - Microsoft SQL Server Management Studio

```
⊟ kal-iu-sql-26 (SQL Server 14.0.3045.24 - RHSP\andyanabo)
   ⊟ Databases
      ⊞ System Databases
      ⊞ Database Snapshots
      ⊞ AdditionalBaseline
      ⊞ additionalbaseline2
        BA_Round1_Final
      ⊞ BA_Round10_Final
      ⊞ BA_Round11_Final
      ⊞ BA_Round12_Final
      ⊞ BA_Round13_Final
      ⊞ BA_Round14_Final
      ⊞ BA_Round15_Final
      ⊞ BA_Round17_Final
      ⊞ BA_Round18_Working
      ⊞ BA_Round2_Final
      ⊞ BA_Round3_Final
      ⊞ BA_Round4_Final
      ⊞ BA_Round5_Final
      ⊞ BA_Round6_Final
      ⊞ BA_Round7_Final
      ⊞ BA_Round8_Final
      ⊞ BA_Round9_Final
        CData_Test
      ⊞ Clinical_data
      ⊞ Clinical_Data_Column_List
      ⊞ Clinical_Data_OHDSI
      ⊞ Clinical_Data_RT
      ⊞ Clinical_Data_Staging
      ⊞ Errors
```

Disconnected.

SSIS packages

Rakai Health Sciences Program
Improved Health Through Research

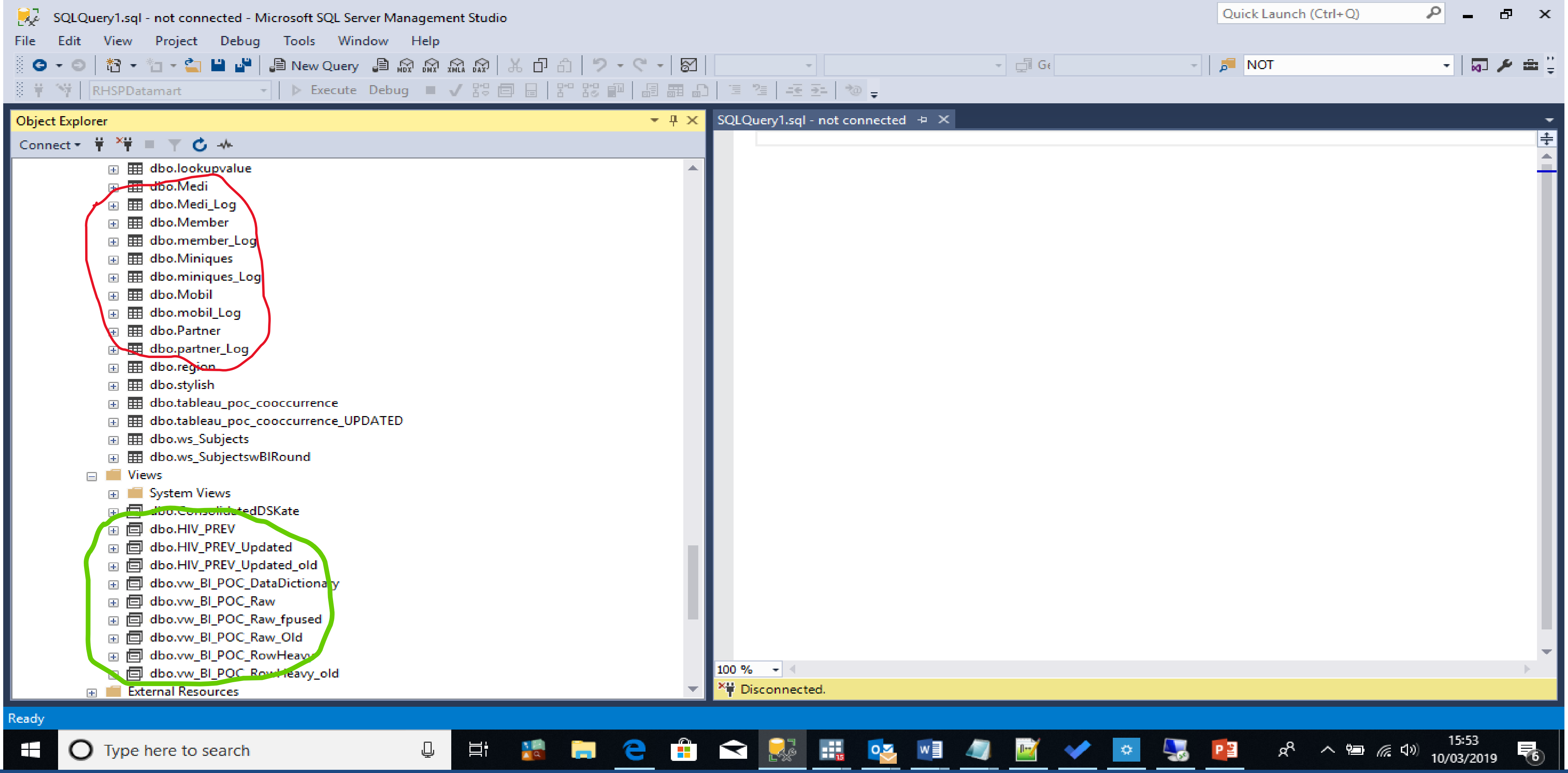


Final ETL SSIS - Microsoft Visual Studio

Labdata_rpr.dtsx
Labdata_tpha.dtsx
Labdata_worksource.dtsx
MainETL.dtsx
R18working.dtsx
Round 1.dtsx
Round 10.dtsx
Round 11.dtsx
Round 12.dtsx
Round 13.dtsx
Round 14.dtsx
Round 15.dtsx
Round 15S.dtsx
Round 16.dtsx
Round 17.dtsx
Round 2.dtsx
Round 3.dtsx
Round 4.dtsx
Round 5.dtsx
Round 6 Updated.dtsx
Round 6.dtsx
Round 7.dtsx
Round 8.dtsx
Round 9.dtsx
WorksourceRefresh.dtsx
Package Parts
Control Flow
Miscellaneous

Data mart tables and Views

Rakai Health Sciences Program
Improved Health Through Research



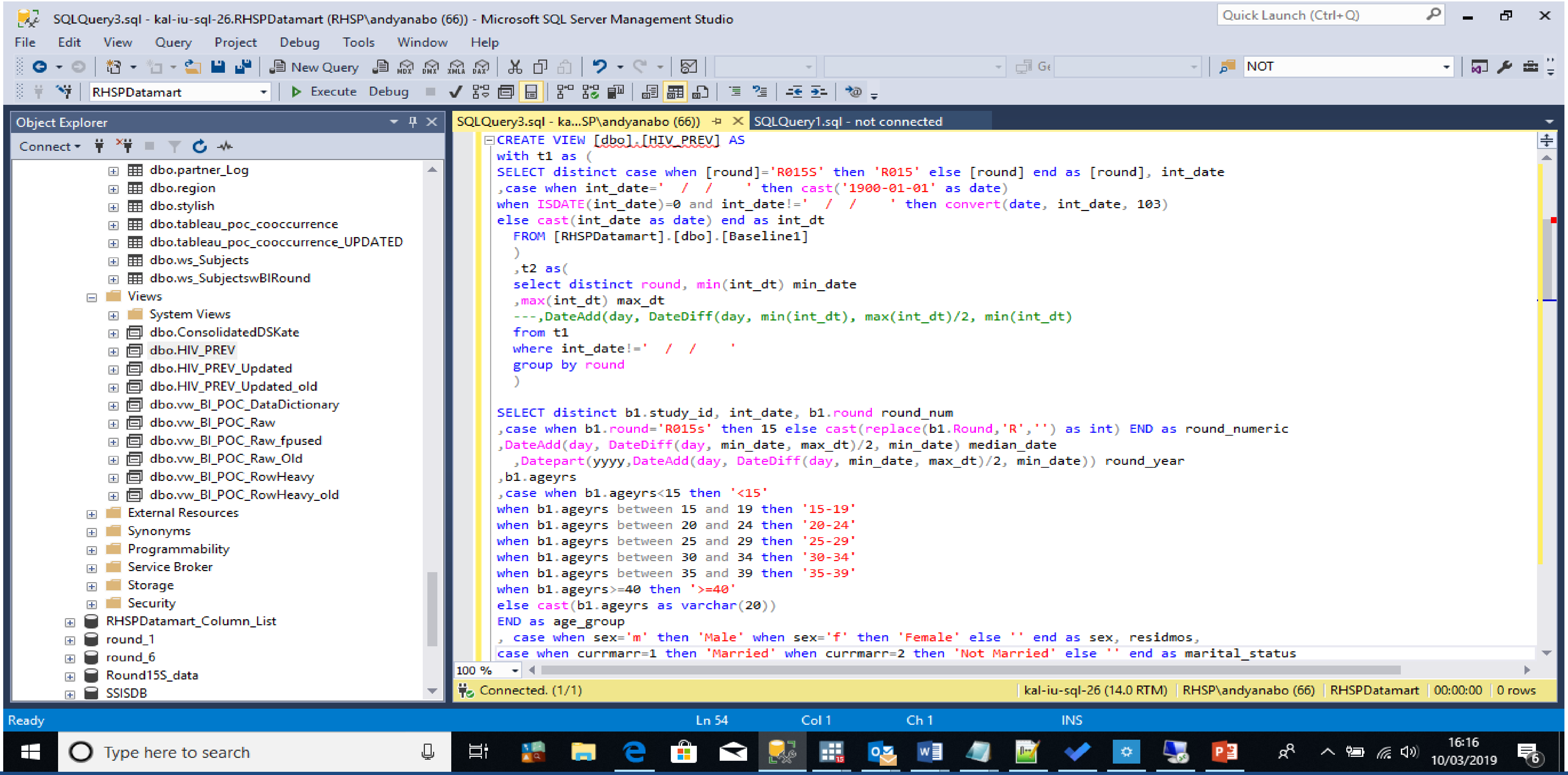
SQLQuery1.sql - not connected - Microsoft SQL Server Management Studio

- dbo.lookupvalue
- dbo.Medi
- dbo.Medi_Log
- dbo.Member
- dbo.member_Log
- dbo.Miniques
- dbo.miniques_Log
- dbo.Mobil
- dbo.mobil_Log
- dbo.Partner
- dbo.partner_Log
- dbo.region
- dbo.stylish
- dbo.tableau_poc_cooccurrence
- dbo.tableau_poc_cooccurrence_UPDATED
- dbo.ws_Subjects
- dbo.ws_SubjectswBIRound
- Views
  - System Views
  - dbo.ConsolidatedDSKate
  - dbo.HIV_PREV
  - dbo.HIV_PREV_Updated
  - dbo.HIV_PREV_Updated_old
  - dbo.vw_BI_POC_DataDictionary
  - dbo.vw_BI_POC_Raw
  - dbo.vw_BI_POC_Raw_fpused
  - dbo.vw_BI_POC_Raw_Old
  - dbo.vw_BI_POC_RowHeavy
  - dbo.vw_BI_POC_RowHeavy_old
- External Resources

Disconnected.

Ready

Business intelligence

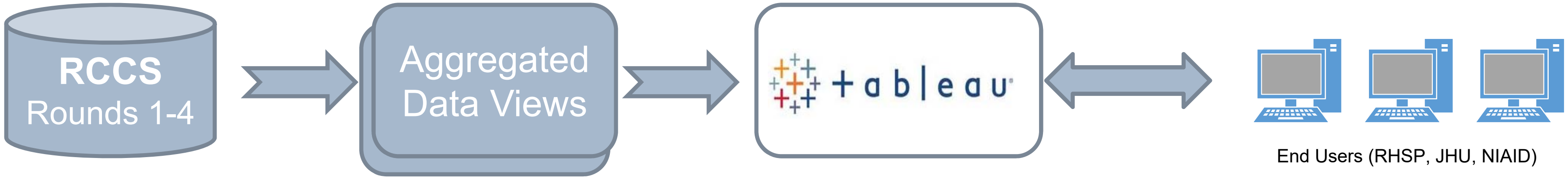


```sql
CREATE VIEW [dbo].[HIV_PREV] AS
with t1 as (
SELECT distinct case when [round]='R015S' then 'R015' else [round] end as [round], int_date
,case when int_date='  /  /    ' then cast('1900-01-01' as date)
when ISDATE(int_date)=0 and int_date!='  /  /    ' then convert(date, int_date, 103)
else cast(int_date as date) end as int_dt
  FROM [RHSPDatamart].[dbo].[Baseline1]
  )
  ,t2 as(
  select distinct round, min(int_dt) min_date
  ,max(int_dt) max_dt
  ---,DateAdd(day, DateDiff(day, min(int_dt), max(int_dt)/2, min(int_dt)
  from t1
  where int_date!='  /  /    '
  group by round
  )

SELECT distinct b1.study_id, int_date, b1.round round_num
,case when b1.round='R015s' then 15 else cast(replace(b1.Round,'R','') as int) END as round_numeric
,DateAdd(day, DateDiff(day, min_date, max_dt)/2, min_date) median_date
  ,Datepart(yyyy,DateAdd(day, DateDiff(day, min_date, max_dt)/2, min_date)) round_year
,b1.ageyrs
,case when b1.ageyrs<15 then '<15'
when b1.ageyrs between 15 and 19 then '15-19'
when b1.ageyrs between 20 and 24 then '20-24'
when b1.ageyrs between 25 and 29 then '25-29'
when b1.ageyrs between 30 and 34 then '30-34'
when b1.ageyrs between 35 and 39 then '35-39'
when b1.ageyrs>=40 then '>=40'
else cast(b1.ageyrs as varchar(20))
END as age_group
, case when sex='m' then 'Male' when sex='f' then 'Female' else '' end as sex, residmos,
case when currmarr=1 then 'Married' when currmarr=2 then 'Not Married' else '' end as marital_status
```

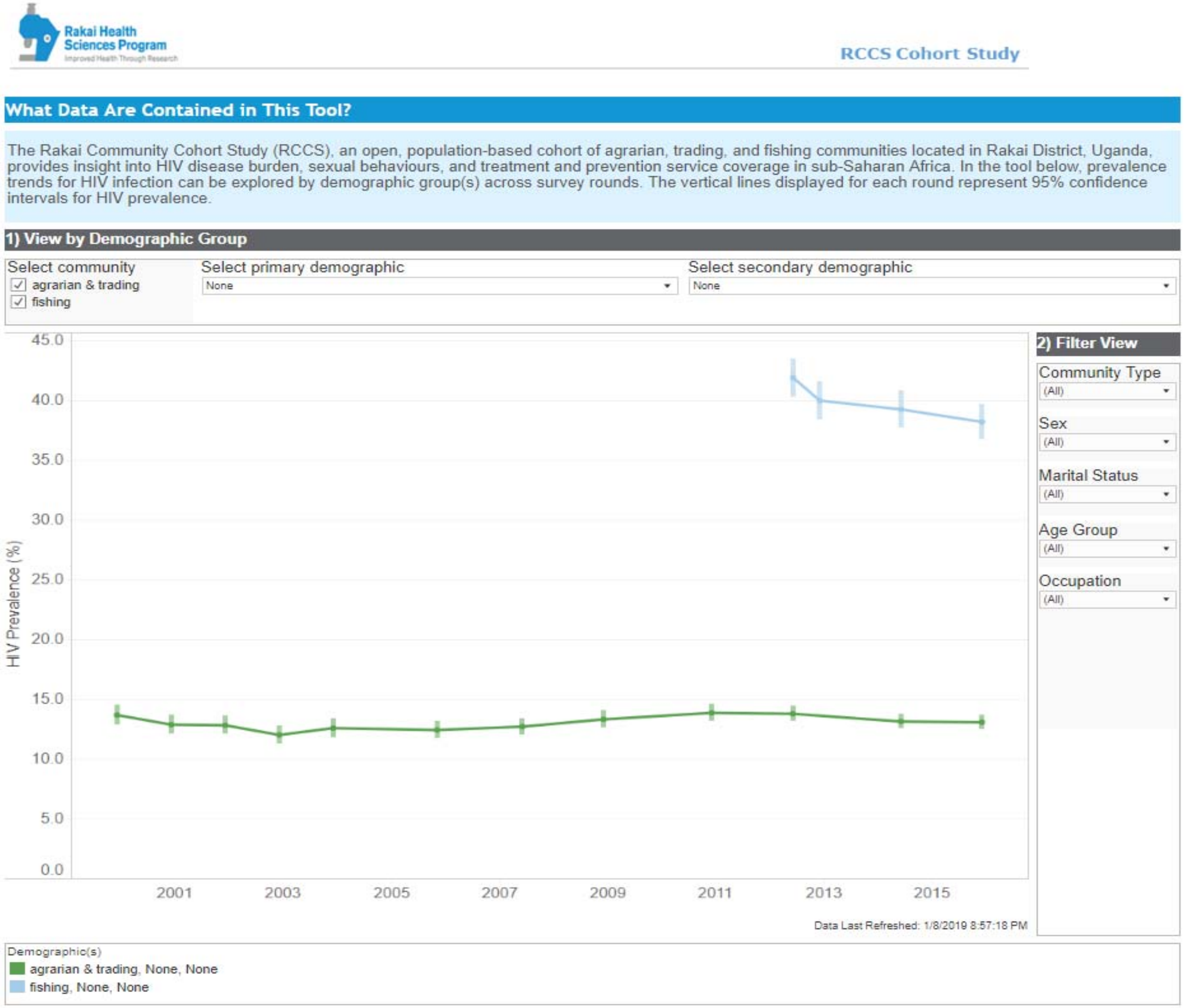
Data Presentation

- **Summarize** and visualize large amounts of historical data
- Only reveal **aggregate distributions** to mitigate risk of misuse
- Allow users to filter and represent data in **dynamic/interactive** ways
- Generate **new insights** that can lead to research hypotheses
- Leverage tableau public platform to enable access to broader research community



Current Dashboard

- Displays Prevalence trends from Rounds 6-17
- Separates Agrarian & Trading communities from Fishing communities
- Mirrors eligibility criteria from Nov 2017 NEJM article (https://www.nejm.org/doi/full/10.1056/NEJMoa1702150)
- Allows stratification by Sex, Age Group, Marital Status, and Occupation
- Currently adding Incidence, ART Coverage, and Male Circumcision Coverage
- Accessible for RHSP team via this link



RCCS Cohort Study

What Data Are Contained in This Tool?

The Rakai Community Cohort Study (RCCS), an open, population-based cohort of agrarian, trading, and fishing communities located in Rakai District, Uganda, provides insight into HIV disease burden, sexual behaviours, and treatment and prevention service coverage in sub-Saharan Africa. In the tool below, prevalence trends for HIV infection can be explored by demographic group(s) across survey rounds. The vertical lines displayed for each round represent 95% confidence intervals for HIV prevalence.

1) View by Demographic Group

Select community
☑ agrarian & trading
☑ fishing

Select primary demographic: None

Select secondary demographic: None

2) Filter View

Community Type: (All)

Sex: (All)

Marital Status: (All)

Age Group: (All)

Occupation: (All)

Data Last Refreshed: 1/8/2019 8:57:18 PM

Demographic(s)
🟩 agrarian & trading, None, None
🟦 fishing, None, None

Results!

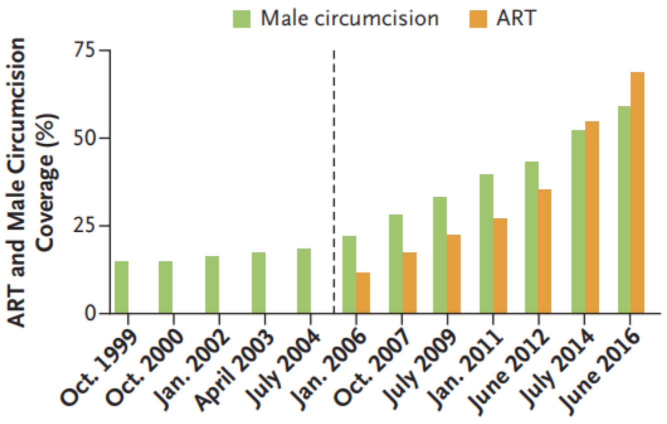
Rakai Health Sciences Program
Improved Health Through Research

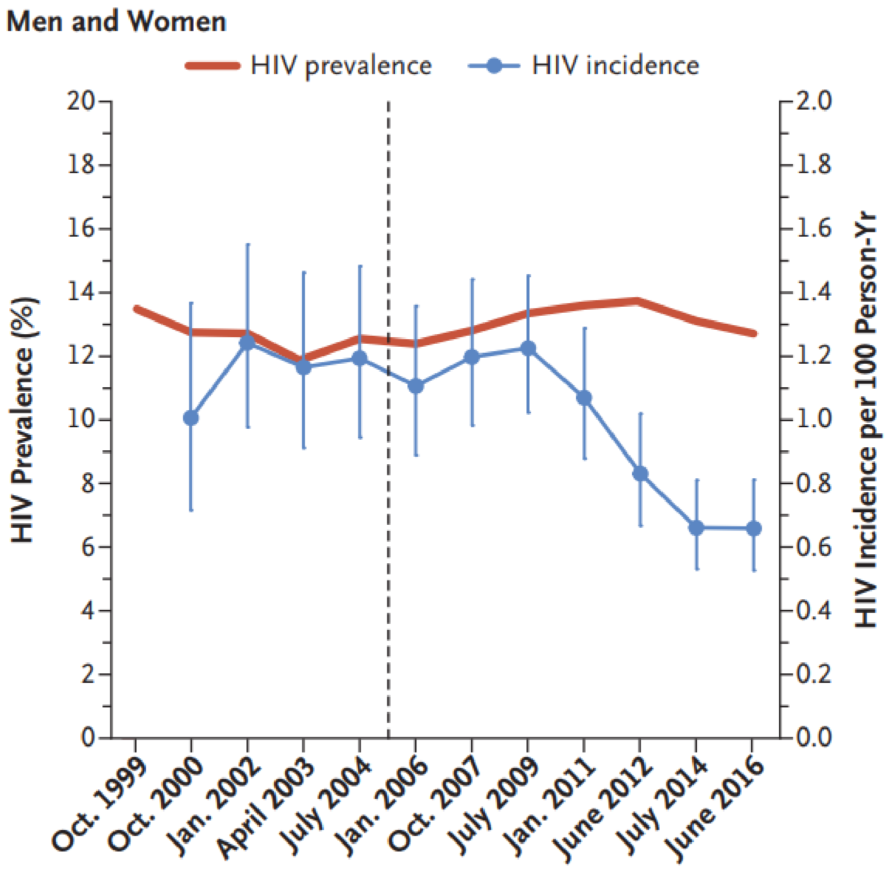
The NEW ENGLAND JOURNAL *of* MEDICINE

ORIGINAL ARTICLE

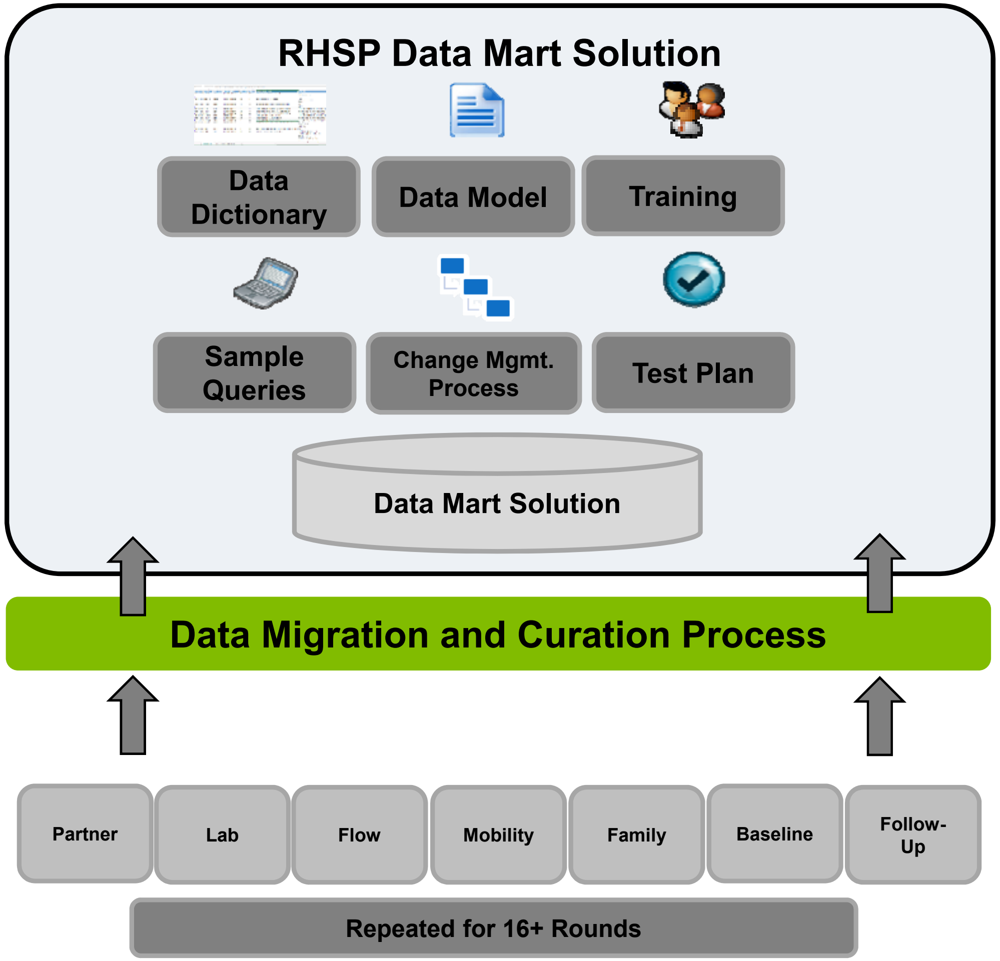
HIV Prevention Efforts and Incidence of HIV in Uganda

M.K. Grabowski, D.M. Serwadda, R.H. Gray, G. Nakigozi, G. Kigozi, J. Kagaayi, R. Ssekubugu, F. Nalugoda, J. Lessler, T. Lutalo, R.M. Galiwango, F. Makumbi, X. Kong, D. Kabatesi, S.T. Alamo, S. Wiersma, N.K. Sewankambo, A.A.R. Tobian, O. Laeyendecker, T.C. Quinn, S.J. Reynolds, M.J. Wawer, and L.W. Chang, for the Rakai Health Sciences Program*





RHSP Data Mart Solution



Allows researchers / sponsors to have easy access to accurate data

- Consolidated Data Mart reduces time and overhead required to fulfill data requests

Researchers and statisticians can correlate and analyze all study results from a single source

- Data Mart controls increase research validity

Develops a repeatable process framework to efficiently and consistently manage future data

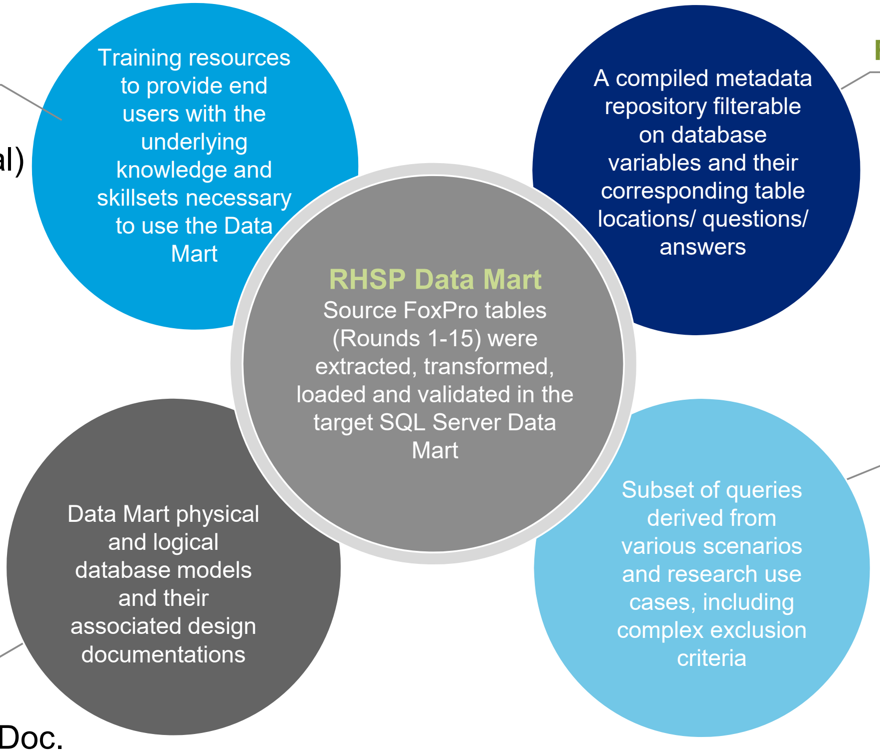
- Data Mart provides for standardized integration of new data

RHSP Data Mart Solution

Initial Rollout: Feb 2015

Training Materials

- **10** Video Tutorials (~150 minutes total)
- **2** User Manuals
- **2** User Test Plans
- **1** Training Plan



Training resources to provide end users with the underlying knowledge and skillsets necessary to use the Data Mart

A compiled metadata repository filterable on database variables and their corresponding table locations/ questions/ answers

RHSP Data Mart
Source FoxPro tables (Rounds 1-15) were extracted, transformed, loaded and validated in the target SQL Server Data Mart

Data Mart physical and logical database models and their associated design documentations

Subset of queries derived from various scenarios and research use cases, including complex exclusion criteria

RHSP Data Dictionary

- **750** Pages
- **9** Columns of Information

Library of Queries

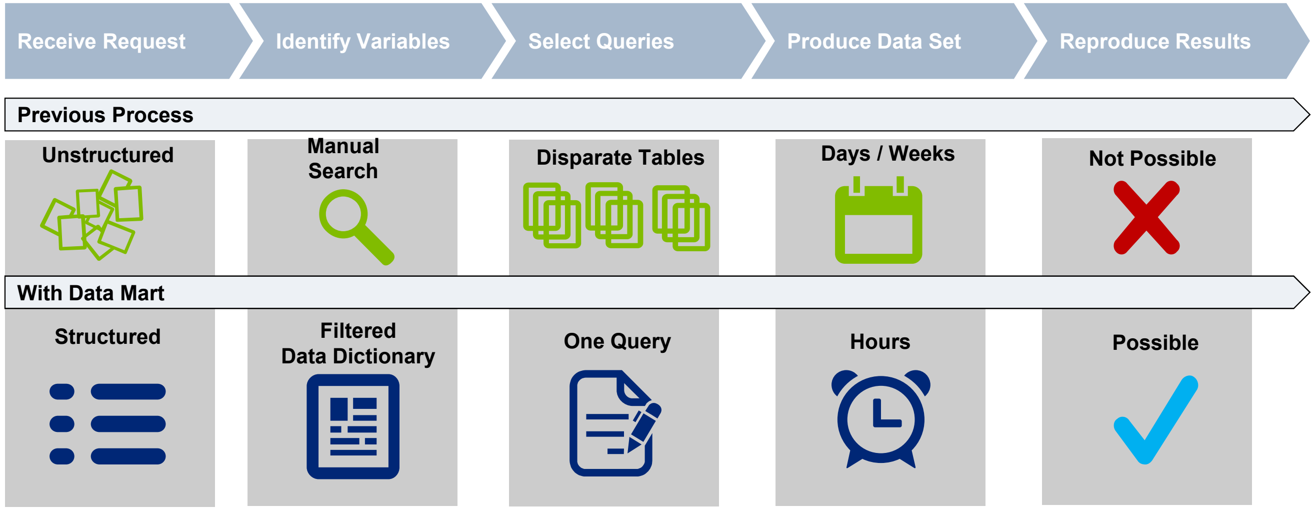
- **34** Total sample Queries
- **91** ETL Scripts
- **5** Sample Use Cases

Data Model/ Design

- Data Mart Model
- Data Mart Design Doc.
- Physical Database Model
- Data Mart Table Links

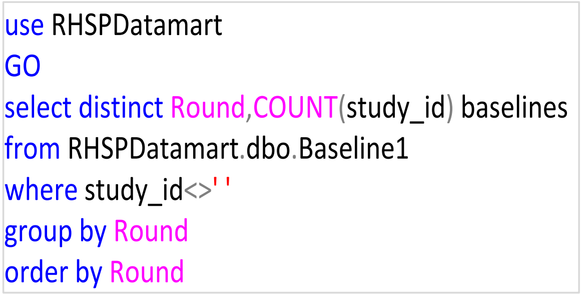
- **20** Tables
- **65,762** Subject Records
- **15+** Rounds of Research Data
- Track Changes Logs

RHSP Data Mart Solution



	Receive Request	Identify Variables	Select Queries	Produce Data Set	Reproduce Results
Previous Process	Unstructured	Manual Search	Disparate Tables	Days / Weeks	Not Possible
With Data Mart	Structured	Filtered Data Dictionary	One Query	Hours	Possible

Sample Query:



```
use RHSPDatamart
GO
select distinct Round,COUNT(study_id) baselines
from RHSPDatamart.dbo.Baseline1
where study_id<>''
group by Round
order by Round
```

Conclusion

- Moving towards more open information sharing with the wider community-allowing RHSP to be discovered through visual analytics and hopefully drive new collaborations and grants.
- Ready to harness new technologies to uncover meaningful insights in the data using predictive models, scenario analysis, and advanced analytics for timely and effective decision making.
- Legacy systems might have become outdated but setting up the infrastructure for modern systems requires high capital expenditure.

Acknowledgments

FOR GOD AND MY COUNTRY
THE REPUBLIC OF UGANDA
Ministry of Health

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NATIONAL INSTITUTES OF HEALTH

BILL & MELINDA GATES foundation

WE BUILD FOR THE FUTURE

THE WORLD BANK

INDEPTH · INDEPTH · INDEPTH

KAROLINSKA INSTITUTET ANNO 1810
Karolinska Institutet