Epi 239 – Real World Data

Elvin Geng

UCSF

2019

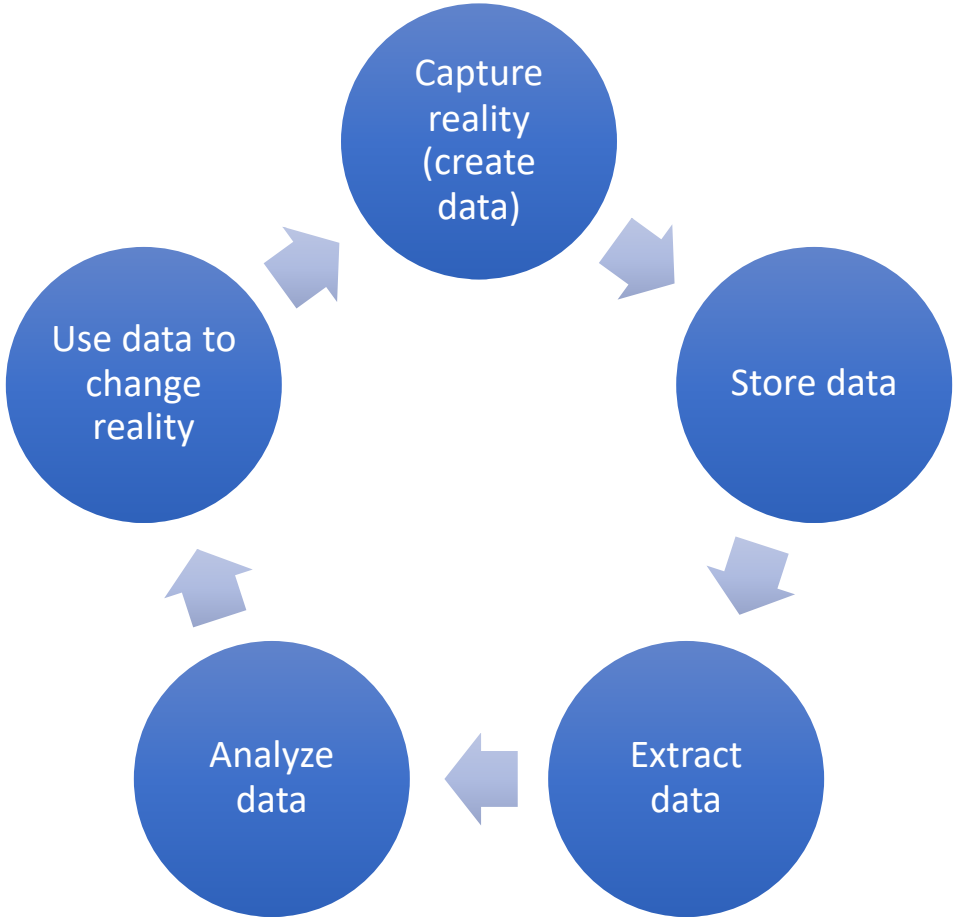
Outline

- Implementation science depends on real world data
- Administrative data is one important source of real life data
- Understanding how to use this data is an practical skill in implementation science
- Discuss “real world” data lifecycle
  - Capturing reality (creating data)
  - Storing data
  - Extracting data
  - Analyzing data from non-research sources
- Challenges and opportunities
- Use cohorts and or studies that I have worked in as case studies

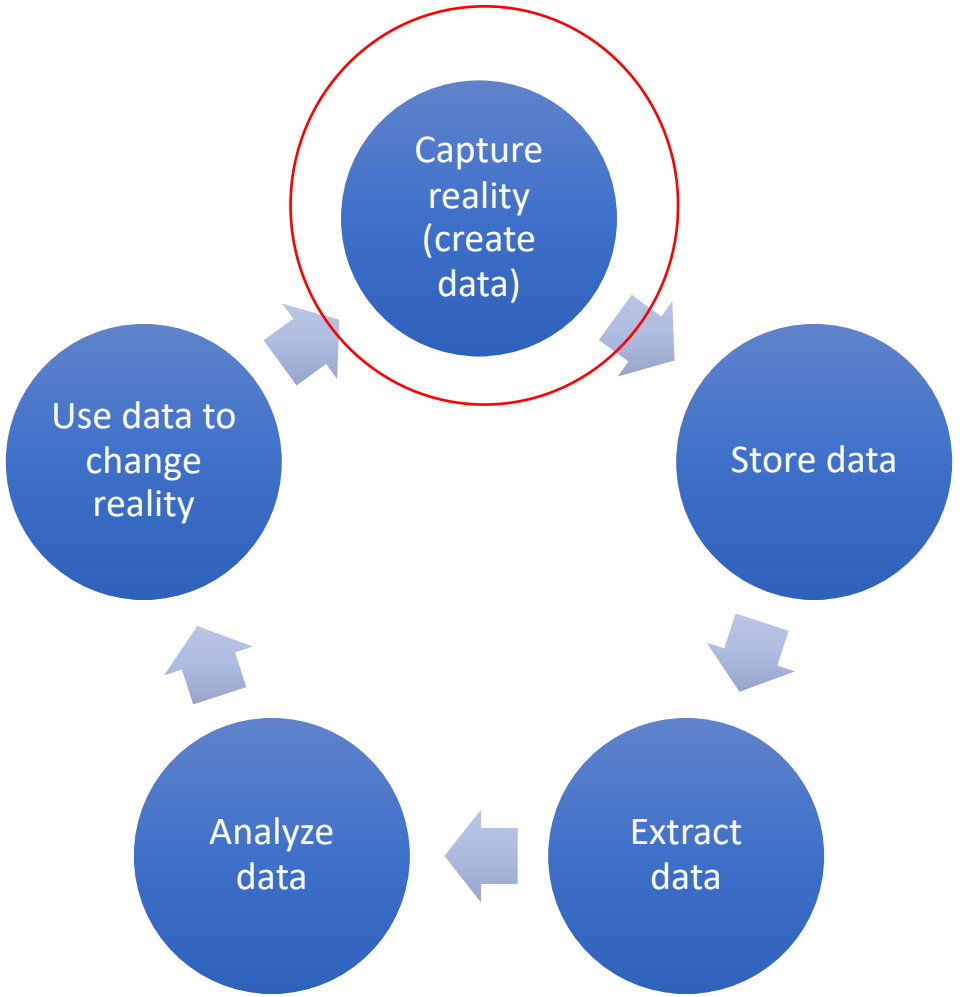
Phronesis

Practical lessons about using real world secondary data

Data lifecycle



Data lifecycle – data creation



Widespread capture of reality with data

- Data used to be the domain of researchers
- Today data capture is ubiquitous
  - Clinical electronic medical records
  - Administrative databases
  - Uber / facebook
  - The purpose is not expressly for research
  - *How much of your life is "data" (a lot)*

Opportunities and challenges in use of health data from care settings

- Advantageous
  - Timely
  - Cheap (low marginal cost)
    - Drug trial – 10000$ a patient
  - “Real world”
  - Large
  - Diverse
- Disadvantages
  - Does not map perfectly to concepts in research questions
  - Big data = small biases effect inferences

Choice A

- Decide what you want
- Go get the ingredients
- But what if Whole Foods is closed?

Choice B

What do you want?

…..I don’t know, what you have?

What is your research question?

- Traditional way to get a research question: first decide what you want then go get ingredients / data

- Alternative way to get a research question: first look to see what ingredients / data you have then decide what you want to eat

Many existing databases for analysis public available

- https://www.gapminder.org/data/
- https://www.healthdata.gov/
- https://data.gov.uk/

HIGHLIGHTS

Safety at Sea – U.S. Coast Guard Marine Casualty and Pollution Data for Researchers

The U.S. Coast Guard (USCG) is responsible for investigating reportable marine casualties, accidents, and serious marine incidents. The relevant mission statement and specific regulations can be found on the USCG Investigations Division homepage. After an incident has been reported, it is entered into a national database of all marine casualty and pollution incidents. These important historical records can then be accessed by researchers interested in understanding maritime safety, accident prevention, or trends in certain types of maritime incidents through time. Other agencies interested in maritime transportation performance measures rely on the USCG data to examine incident trends on U.S. waterways. Files for the Marine Casualty and Pollution Data for Researchers datasets can be downloaded directly from the U.S. Coast Guard Homeport website by following the drop-down menu options on the homepage **Missions: Investigations**: Marine Casualty Pollution Investigations page at https://homeport.uscg.mil/missions/investigations/marine-casualty-pollution-investigations.

As described by the USCG, "the Marine Casualty and Pollution Data files provide details about marine casualty and pollution incidents investigated by Coast Guard Offices throughout the United States. The database can be used to analyze marine accidents and pollution incidents by a variety of factors including vessel or facility type, injuries, fatalities, pollutant details, location, and date. The data collection period began in 1982 for marine casualties and 1973 for polluting incidents, and is ongoing. Documentation includes entity and attribute descriptions along with suggested solutions to general marine pollution, vessel casualty, and personnel injury and death questions."

Visitors to the USCG Homeport data download site should note that there are three files available fore download, but it is the second file on the list (named MISLE_DATA.zip) that contains all available marine casualty data from January 2002 – July 2015. The files extracted from MISLE_DATA.zip can be opened with most standard spreadsheet editing software programs.

Source: USCG Marine Casualty and Pollution Data, https://homeport.uscg.mil/missions/investigations/marine-casualty-pollution-investigations

ARTICLES | 19 MAY 2009

The President's Emergency Plan for AIDS Relief in Africa: An Evaluation of Outcomes FREE

Eran Bendavid, MD; Jayanta Bhattacharya, MD, PhD

Article, Author, and Disclosure Information

PDF | CITATIONS | PERMISSIONS

Published: *Ann Intern Med.* 2009;150(10):688-695.
DOI: 10.7326/0003-4819-150-10-200905190-00117

Abstract

Background: Since 2003, the President's Emergency Plan for AIDS Relief (PEPFAR) has been the most ambitious initiative to address the global HIV epidemic. However, the effect of PEPFAR on HIV-related outcomes is unknown.

Objective: To assess the effect of PEPFAR on HIV-related deaths, the number of

SEE ALSO

Evaluating the President's Emergency Plan for AIDS Relief: Time to Scale It Up

Evaluating Outcomes of the President's Emergency Plan for AIDS

designated all other sub-Saharan African countries with a generalized HIV epidemic as the control group.

Data Sources

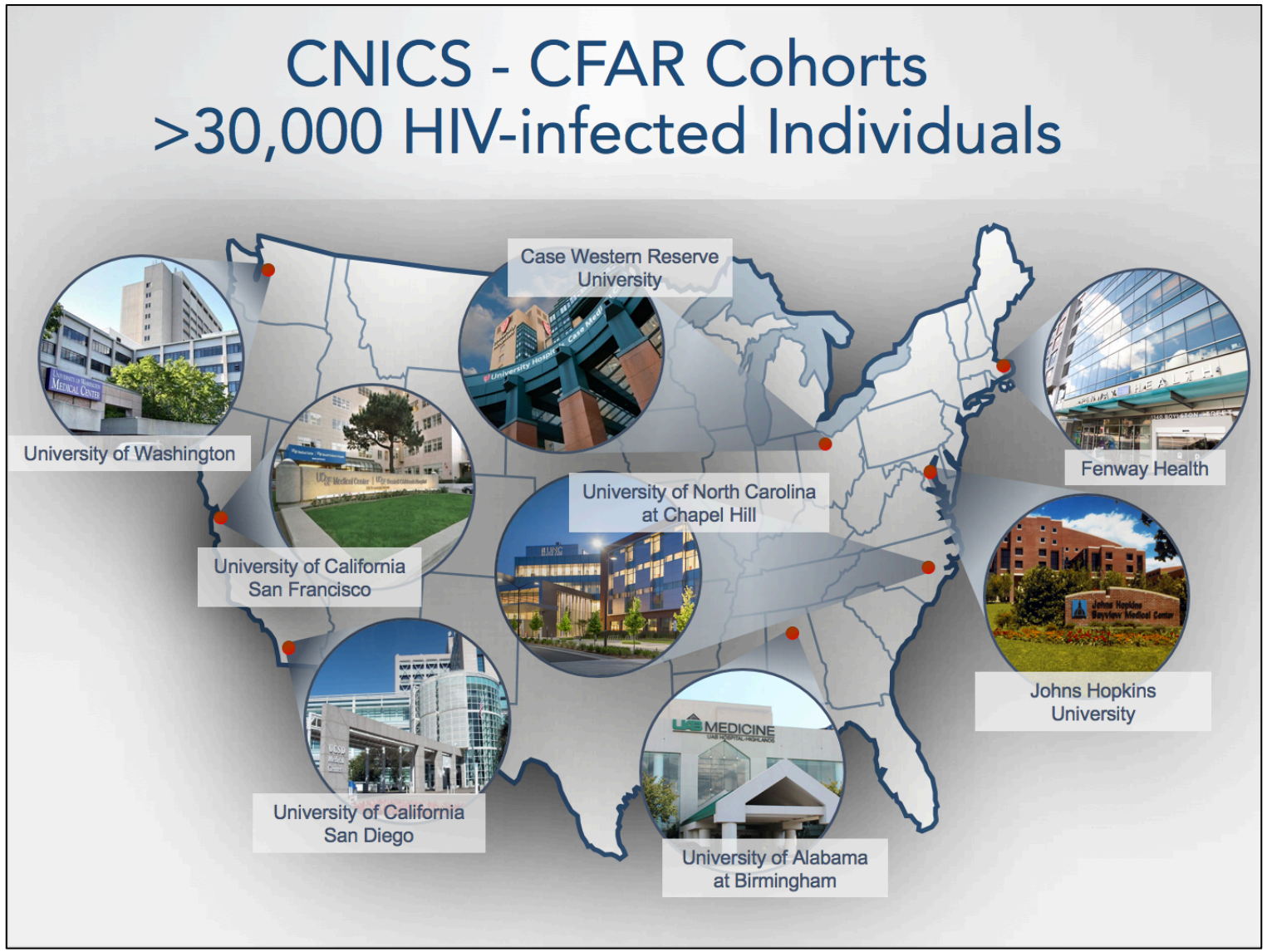
The joint United Nations Programme on HIV/AIDS (UNAIDS) has monitored the Declaration of Commitment on HIV/AIDS by working with individual countries to measure the global epidemiology of HIV and AIDS (16). We used country and year epidemiologic data obtained through UNAIDS from 1997 to 2007 as the outcome variables for this study. The UNAIDS determines the prevalence trends by using sentinel and population-based surveys. The prevalence estimates, vital registry data, and model-based calculations are then used to determine mortality and the number of people living with HIV. All the estimates are given uncertainty bounds that depend on the quality of the primary data and the strength of the assumptions used in the estimation process (17).

We obtained data on Global Fund allocation of resources, population structure, governance indicators, life expectancy, and per capita gross domestic product from publicly available databases (10, 18–20).

Study Periods and Outcomes

We chose 2 study periods for this analysis: an early period, from 1997 to 2002, before PEPFAR began, and a late period, from 2004 to 2007, during PEPFAR's

CNICS - CFAR Cohorts
>30,000 HIV-infected Individuals



Clinic-based database

- Network of integrated “clinical systems”
  - “unique” electronic records of patient care experience
- Anyone can request data from the consortia

Article types
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Clear all

Show additional filters

Search results

Items: 1 to 20 of 273

« First ‹ Prev Page 1 of 14 Next › Last »

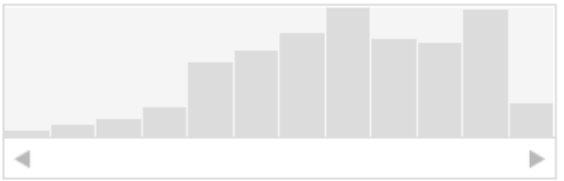
1. Association of Biomarker Clusters With Cardiac Phenotypes and Mortality in Patients With HIV Infection.
   Scherzer R, Shah SJ, Secemsky E, Butler J, Grunfeld C, Shlipak MG, Hsue PY.
   Circ Heart Fail. 2018 Apr;11(4):e004312. doi: 10.1161/CIRCHEARTFAILURE.117.004312.
   PMID: 29615435
   Similar articles

2. Recent Abacavir Use Increases Risk of Type 1 and Type 2 Myocardial Infarctions Among Adults With HIV
   Elion RA, Althoff KN, Zhang J, Moore RD, Gange SJ, Kitahata MM, Crane HM, Drozd DR, Stein JH, Klein MB, Eron JJ, Silverberg MJ, Mathews WC, Justice AC, Sterling TR, Rabkin CS, Mayor AM, Klein DB, Horberg MA, Bosch RJ, Eyawo O, Palella FJ Jr; North American AIDS Cohort Collaboration on Research and Design of IeDEA.
   J Acquir Immune Defic Syndr. 2018 May 1;78(1):62-72. doi: 10.1097/QAI.0000000000001642.
   PMID: 29419568
   Similar articles

3. Measurement of depression treatment among patients receiving HIV primary care: Whither the truth?
   DiPrete BL, Pence BW, Grelotti DJ, Gaynes BN.
   J Affect Disord. 2018 Apr 1;230:50-55. doi: 10.1016/j.jad.2017.12.068. Epub 2018 Jan 3.
   PMID: 29407538

Sort by:
Best match | Most recent

Results by year



Download CSV

Find related data
Database: Select
Find items

Search details
R24 AI067039[Grant Number]

To use it properly, you got to know what's in it

- Questions to data or…
- Don't make a BLT if you have no bacon

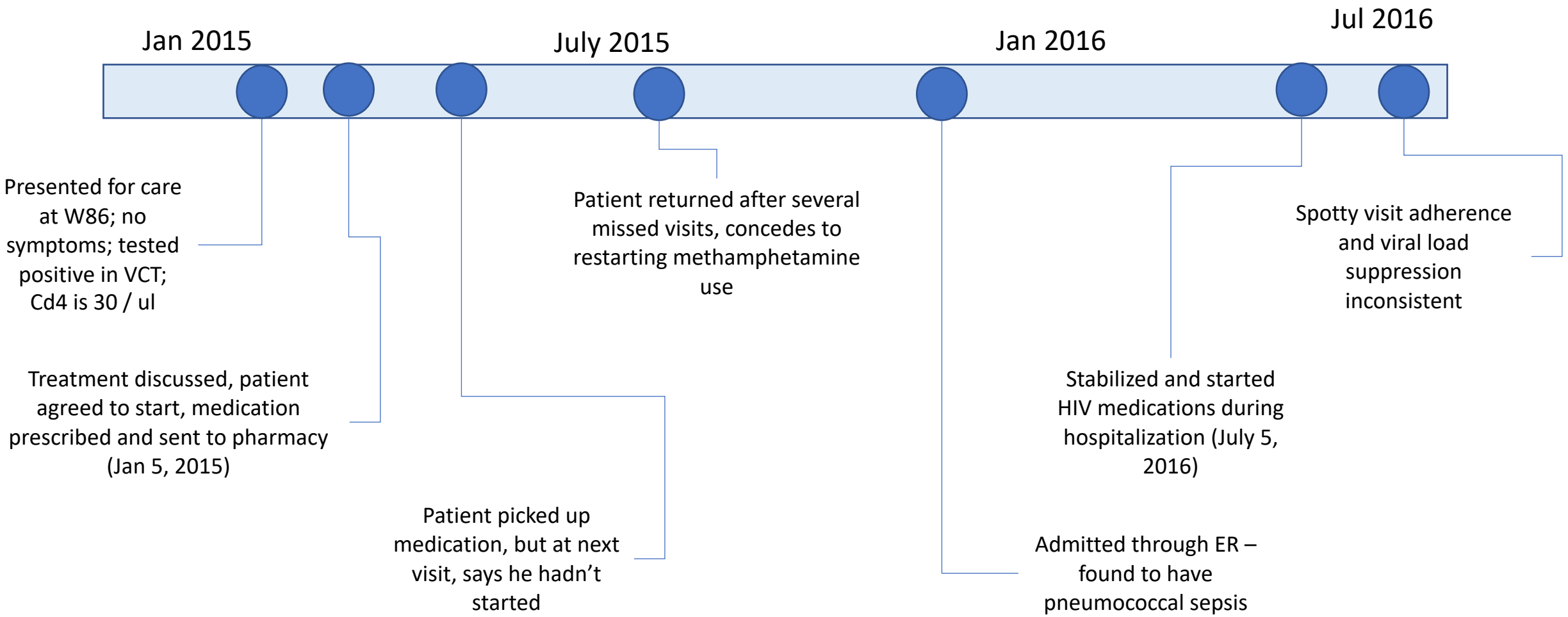
Good or bad question?

- Do the integrase inhibitors lead to more rapid viral suppression? What are the predictors of more rapid suppression in real-world populations?

TabMedication

id	date	Medication active	Medication discontinuation
1	4354		
1	4376	EFV	
1	4395	EFV	
1	4420	EFV	
1	4510	EFV	
1	4554	EFV	
1	4576	EFV	
1	4695	EFV	
1	4720	EFV	
1	4810	EFV	
1	4850	EFV	

When is the start date?



When does the patient suppress?

- When does the patient get measured?

“Adjudication…”

- Medication start date is actually unknown in the database
- Large scale effort to “correct” the start date
  - What if they never come back?
  - What if the provider doesn’t know or doesn’t document that delayed start?

Don’t promise a BLT if you don’t have bacon

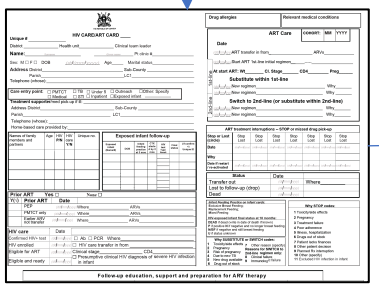
- The effect of drug x on rapidity of suppression a bad question (a biological question)
- Asking a question that can be answered with your data even if it is not exactly what you originally had in mind is better than asking a question that can’t be answered
- Iterative
- But… the questions you can answer are more interesting (in my opinion)
- In reality many times people go ahead and make the BLT and then are surprised when guests complain when it is served without bacon

The omnivore’s dilemma……

where does your food come from and do you know what is in it?


LOGIN
USERNAME
PASSWORD
LOG IN
IeDEA
Who We Are
News
Calendar
Regions
Working Groups
International Epidemiology Databases to Evaluate AIDS
IeDEA
Each region collaborates with clinical sites to identify and define key variables, harmonize and effectively analyze the data to generate large datasets. These data can be merged across regions to address research questions related to the impact of the global ART rollout on HIV-related clinical outcomes.
IEDEA
GLOBAL COHORTS COMPOSED OF
>1,700,000
PERSONS LIVING WITH HIV/AIDS
Who We Are
The IeDEA network is an international research consortium established in 2005 by the National
Public News
Prevent HIV, test and treat all - WHO
Dec 4 2016 - 7:06am
Publications
HIV and Aging: Demographic Change in the Asia-...
4/15/17
Upcoming Events
Jun 06
Data Harmonisation WG call
Jun
Executive Committee

Clinical care



Data collected on a clinical encounter form

Data entry

Data pooling at national and subnational scale (possible derivation)

International pooling

Unique # ☐☐☐☐☐☐☐ **HIV CARE/ART CARD** Name ____________________

Date Check if scheduled. Write in alternate pick-up if ill	Follow-up date	Duration in months since first starting ART/ since starting current regimen	Wt Ht f child (0-5yrs) record +/- oedema	If Pregnant EDD?PMTCT? Write gestation in weeks and ANC # FP/no FP If FP write method(s) If child record MUAC Write age in months if ≤5 yrs	TB status If TB Rx, start /stop date (mm/yyyy) and District TB reg #	Potential SIDE EFFECTS	New OI, Other PROBLEMS If child, include nutritional problems	Function Work/Playing Amb Bed	WHO clinical stage	CPT Adhere Dose/# of days Prescribed	INH # pills dispensed	Other meds dispensed (including nutritional supplements)	ARV drugs (incl. prophylaxis) Adhere/ Why	Regimen/ Dose/ # of days dispensed	Investigations CD4 If < 5, record CD4% +/- severe	Hgb, RPR, CXR, TB sputums, Infant Ab/PCR, other	Refer or consult or link/provide (including nutritional support and infantfeeding) If hospitalized, # of days	Name of attending clinician
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt Ht Oedema		TB Status mm/yy Reg No.								ADH Why	REGIMEN DOSE No. of Days				
☐			Wt		TB Status									REGIMEN				

What does this mean?

- Information collected for clinical care – recording CD4 count of that row isn’t bad
- But not a good way to capture the CD4 for research purposes

Immunodeficiency at the Start of ART: A Global View

Klea Panayidou[1], Ole Kirk[2], IeDEA collaboration[1], and COHERE collaboration[2,3]

[1]Institute of Social & Preventive Medicine (ISPM), University of Bern, Switzerland;
[2]Centre for Health and Infectious Disease Research, Department of Infectious Diseases and Reumatology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark;
[3]Institut de Santé Publique, d'Epidémiologie et de Développement – INSERM U987, Université de Bordeaux, Bordeaux Cedex, France.

Matthias Egger
ISPM, University of Bern
Phone: +41 31 631 35 11
egger@ispm.unibe.ch
www.ispm.ch
www.iedea-sa.org

Abstract Number: 1094

Objectives

- To describe the CD4 cell count at the start of ART in adults in low-, middle-, and high-income countries.
- To examine the impact of changes in World Health Organization (WHO) guidelines on when to start ART.

Methods

- Countries were classified into low-, middle-, and high-income countries according to World Bank definitions.
- We included adults (age ≥16 years) with known gender and ART start date from a clinic or program participating in one of the following collaborative studies:
  - The International epidemiologic Databases to Evaluate AIDS (IeDEA), cohorts from: North America, the Caribbean, Central and South America, Asia-Pacific, Central, East, Southern and West Africa.
  - The Collaboration of Observational HIV Epidemiological Research in Europe (COHERE), cohorts: AHIVCOS, AMACS, ANRS CO2 SEROCO, ANRS CO4 FHDH, ANRS CO6 PRIMO, ANRS CO8 APROCO-COPILOTE, ATHENA, St. Pierre, Brussels Cohort, CASCADE, ClinSurv, Cologne/Bonn, CoRIS, Danish HIV Cohort Study, EuroSida, ICC, ICONA, Italian Master, Modena, PISCIS, San Raffaele, SHCS, UK CHIC.

Results

- **746,646 patients** from **41 countries** were included in the analysis (Figure 1, Table 1): 47,155 from North America (2 countries), 15,738 from the Caribbean, Central and South America (7), 2,028 from Asia Pacific (1), 539,064 from sub-Saharan Africa (19) and 142,661 from Europe (12).
- The median number of patients per country was 6,562 (IQR 2,556 - 18,116), median age was 36.2 years (IQR 30.3 - 43.5) and the percentage of females was 53%.
- CD4 counts were missing (and imputed) in 28% of adults in low-income, 21% in lower middle-income, 18% in upper middle-income and 13% in high-income countries.
- Trends in median CD4 counts at start of ART from 2002, when ART was scaled up globally, were similar in low-income and upper middle-income countries (Table 1).

	Low-income		Lower middle-income		Upper middle-income		High-income	
No. of patients	194,005		232,158		128,639		191,844	
No. of countries	14		6		6		15	
Median year of starting ART	2009		2008		2009		2005	
Median age	36		35.3		34.7		38.9	
% Women	64%		62%		64%		24%	
Median CD4 count (cells/μL)	Male	Female	Male	Female	Male	Female	Male	Female
2002	63	78	88	112	104	124	201	209
2004	82	102	101	128	114	136	199	206
2006	105	130	117	148	124	148	220	224
2008	129	161	135	171	139	167	262	258

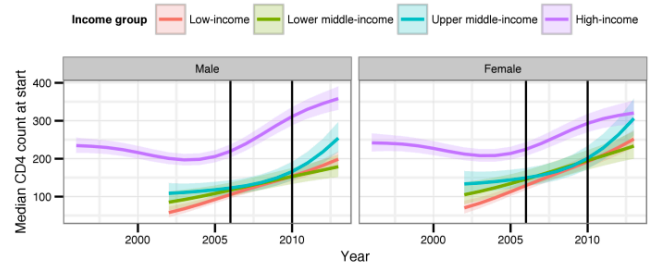


Figure 2. Modeled median CD4 by gender and country income group. Vertical lines indicate the changes in WHO guidelines on when to start ART.

Period	Low-income	Lower middle-income	Upper middle-income	High-income
2002 - 2005	14.5 (5.5 - 22.1)	4.7 (-13.7 - 18.9)	5.3 (-6.2 - 15.5)	2.9 (-5.2 - 10.6)
2006 - 2009	10.7 (5.9 - 15.7)	11.2 (3.6 - 18.7)	6.7 (-1.7 - 15.4)	18.7 (12.0 - 25.4)
2010 - 2014	16.7 (9.3 - 25.0)	13.6 (1.7 - 27.6)	26.3 (13.0 - 41.9)	21.3 (10.9 - 32.8)

Table 2. Increase of median CD4 count (95% CI) at start of ART per year for income group and periods corresponding to WHO guidelines.

Conclusion

- Median CD4 count at start of ART increased in most countries, but remained below 350 cells/μL in all low- and middle-income countries in 2013. Substantial effort and resources continue to be needed to achieve earlier implementation of ART globally.

Garbage in, Garbage out

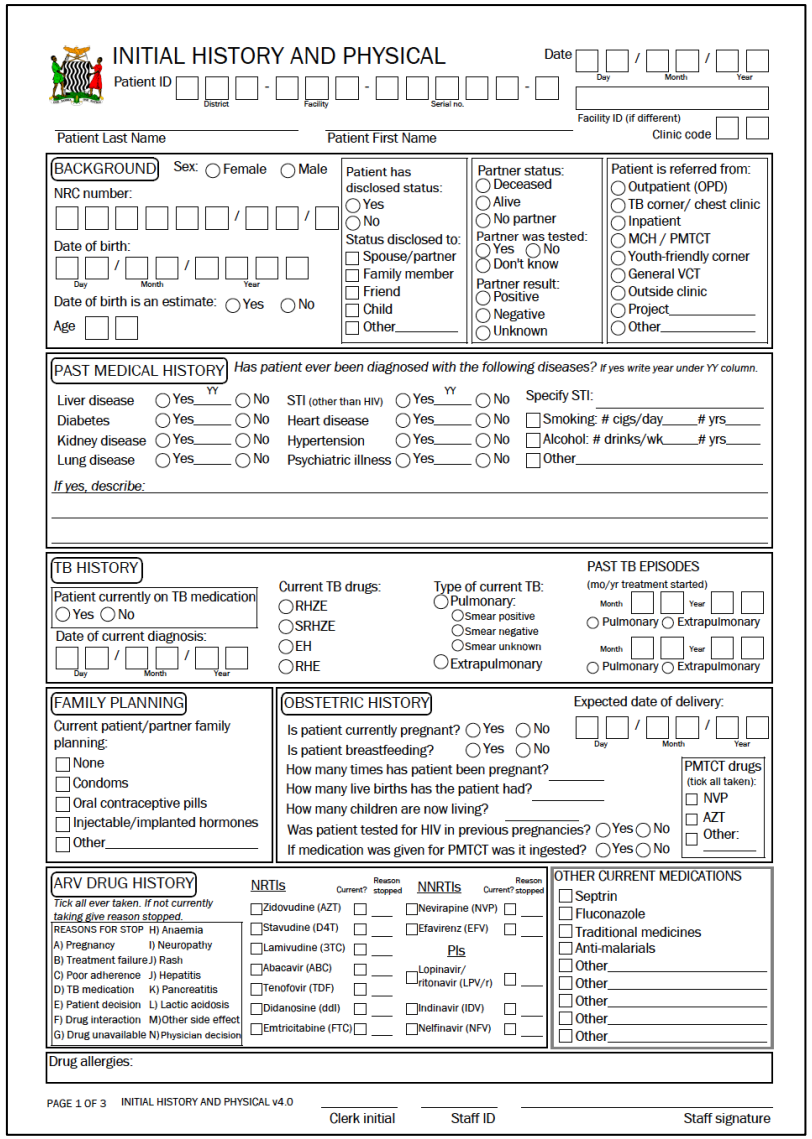
(common insult for clinic-based data, data blaming)

If you don’t want garbage, you got to know what you are looking for and what you are not (recognize garbage)

(Rebuttal, investigator blaming)

Step 1 to getting usable data:

The eye of the beholder (health care wokers)



INITIAL HISTORY AND PHYSICAL

Patient ID ☐☐☐ - ☐☐☐ - ☐☐☐☐☐ - ☐ (District - Facility - Serial no.)

Date ☐☐ / ☐☐ / ☐☐ (Day / Month / Year)

Facility ID (if different) ☐☐☐☐☐

Patient Last Name _____ Patient First Name _____ Clinic code ☐☐

BACKGROUND

Sex: ◯ Female ◯ Male

NRC number: ☐☐☐☐☐☐ / ☐☐ / ☐

Date of birth: ☐☐ / ☐☐ / ☐☐☐☐ (Day / Month / Year)

Date of birth is an estimate: ◯ Yes ◯ No

Age ☐☐

Patient has disclosed status: ◯ Yes ◯ No

Status disclosed to: ☐ Spouse/partner ☐ Family member ☐ Friend ☐ Child ☐ Other_____

Partner status: ◯ Deceased ◯ Alive ◯ No partner

Partner was tested: ◯ Yes ◯ No ◯ Don't know

Partner result: ◯ Positive ◯ Negative ◯ Unknown

Patient is referred from: ◯ Outpatient (OPD) ◯ TB corner/ chest clinic ◯ Inpatient ◯ MCH / PMTCT ◯ Youth-friendly corner ◯ General VCT ◯ Outside clinic ◯ Project_____ ◯ Other_____

PAST MEDICAL HISTORY *Has patient ever been diagnosed with the following diseases? If yes write year under YY column.*

	YY		YY	
Liver disease	◯ Yes_____	◯ No	STI (other than HIV) ◯ Yes_____	◯ No
Diabetes	◯ Yes_____	◯ No	Heart disease ◯ Yes_____	◯ No
Kidney disease	◯ Yes_____	◯ No	Hypertension ◯ Yes_____	◯ No
Lung disease	◯ Yes_____	◯ No	Psychiatric illness ◯ Yes_____	◯ No

Specify STI: _____

☐ Smoking: # cigs/day_____# yrs_____

☐ Alcohol: # drinks/wk_____# yrs_____

☐ Other_____

If yes, describe:

TB HISTORY

Patient currently on TB medication ◯ Yes ◯ No

Date of current diagnosis: ☐☐ / ☐☐ / ☐☐☐☐ (Day / Month / Year)

Current TB drugs: ◯ RHZE ◯ SRHZE ◯ EH ◯ RHE

Type of current TB: ◯ Pulmonary: ◯ Smear positive ◯ Smear negative ◯ Smear unknown ◯ Extrapulmonary

PAST TB EPISODES (mo/yr treatment started)

Month ☐☐ Year ☐☐ ◯ Pulmonary ◯ Extrapulmonary

Month ☐☐ Year ☐☐ ◯ Pulmonary ◯ Extrapulmonary

FAMILY PLANNING

Current patient/partner family planning:

☐ None ☐ Condoms ☐ Oral contraceptive pills ☐ Injectable/implanted hormones ☐ Other_____

OBSTETRIC HISTORY

Expected date of delivery: ☐☐ / ☐☐ / ☐☐☐☐ (Day / Month / Year)

Is patient currently pregnant? ◯ Yes ◯ No

Is patient breastfeeding? ◯ Yes ◯ No

How many times has patient been pregnant? _____

How many live births has the patient had? _____

How many children are now living? _____

Was patient tested for HIV in previous pregnancies? ◯ Yes ◯ No

If medication was given for PMTCT was it ingested? ◯ Yes ◯ No

PMTCT drugs (tick all taken): ☐ NVP ☐ AZT ☐ Other: _____

ARV DRUG HISTORY

Tick all ever taken. If not currently taking give reason stopped.

REASONS FOR STOP: A) Pregnancy; B) Treatment failure; C) Poor adherence; D) TB medication; E) Patient decision; F) Drug interaction; G) Drug unavailable; H) Anaemia; I) Neuropathy; J) Rash; J) Hepatitis; K) Pancreatitis; L) Lactic acidosis; M)Other side effect; N) Physician decision

NRTIs	Current?	Reason stopped
☐ Zidovudine (AZT)	☐	___
☐ Stavudine (D4T)	☐	___
☐ Lamivudine (3TC)	☐	___
☐ Abacavir (ABC)	☐	___
☐ Tenofovir (TDF)	☐	___
☐ Didanosine (ddI)	☐	___
☐ Emtricitabine (FTC)	☐	___

NNRTIs	Current?	Reason stopped
☐ Nevirapine (NVP)	☐	___
☐ Efavirenz (EFV)	☐	___
PIs		
☐ Lopinavir/ritonavir (LPV/r)	☐	___
☐ Indinavir (IDV)	☐	___
☐ Nelfinavir (NFV)	☐	___

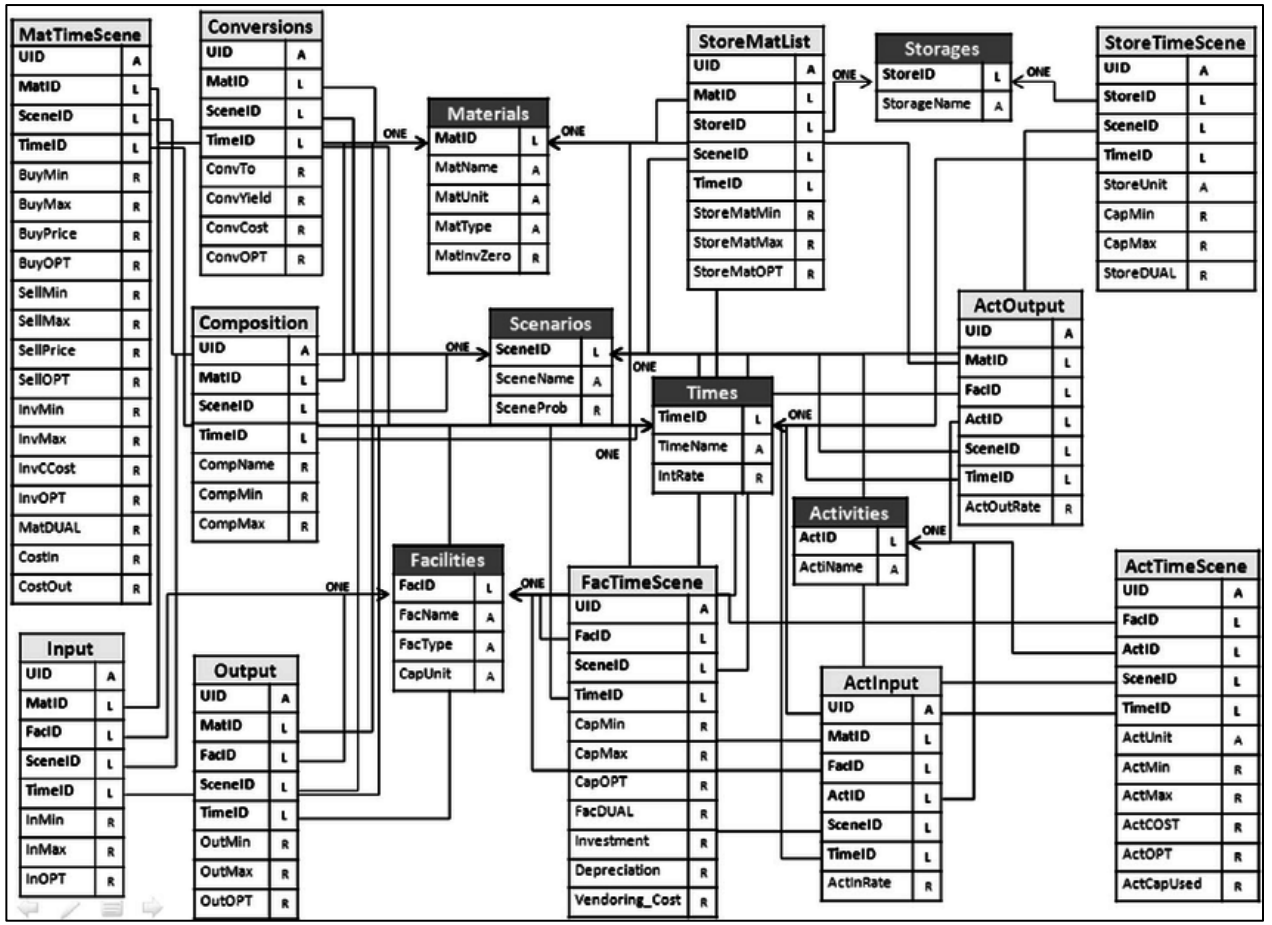
OTHER CURRENT MEDICATIONS

☐ Septrin ☐ Fluconazole ☐ Traditional medicines ☐ Anti-malarials ☐ Other_____ ☐ Other_____ ☐ Other_____ ☐ Other_____ ☐ Other_____

Drug allergies: _____

Clerk initial _____ Staff ID _____ Staff signature _____

The eye of the beholder (data manager)



Communications: Clinical Investigators and Data Managers

Clinical investigators are from Venus and Data managers are from Mars

- Investigator
  - You understand what you want to do
  - You understand what the measurements mean
  - You must define the ingredients you want with precision
  - You need to spell out the ingredients
- Data manager
  - Must extract data from a database
  - But does not know exactly what you want
  - Does not understand what a “viral load” is
  - The concept is named 14334 in the database
  - The data dictionary says

  “Plasma HIV RNA (Roche TM V12) = 14334”

Don’t assume…

- The data manager knows where to find “the data” (that you are looking for)
- The data manager know whether what they’ve given you matches what you wanted
- *The data manager can read your mind*

Need a “contract”

Step 1 for Investigator

- Generally you have access to the front end
- If you can, get a hold of a database schema or the tables
- Surprisingly hard to get

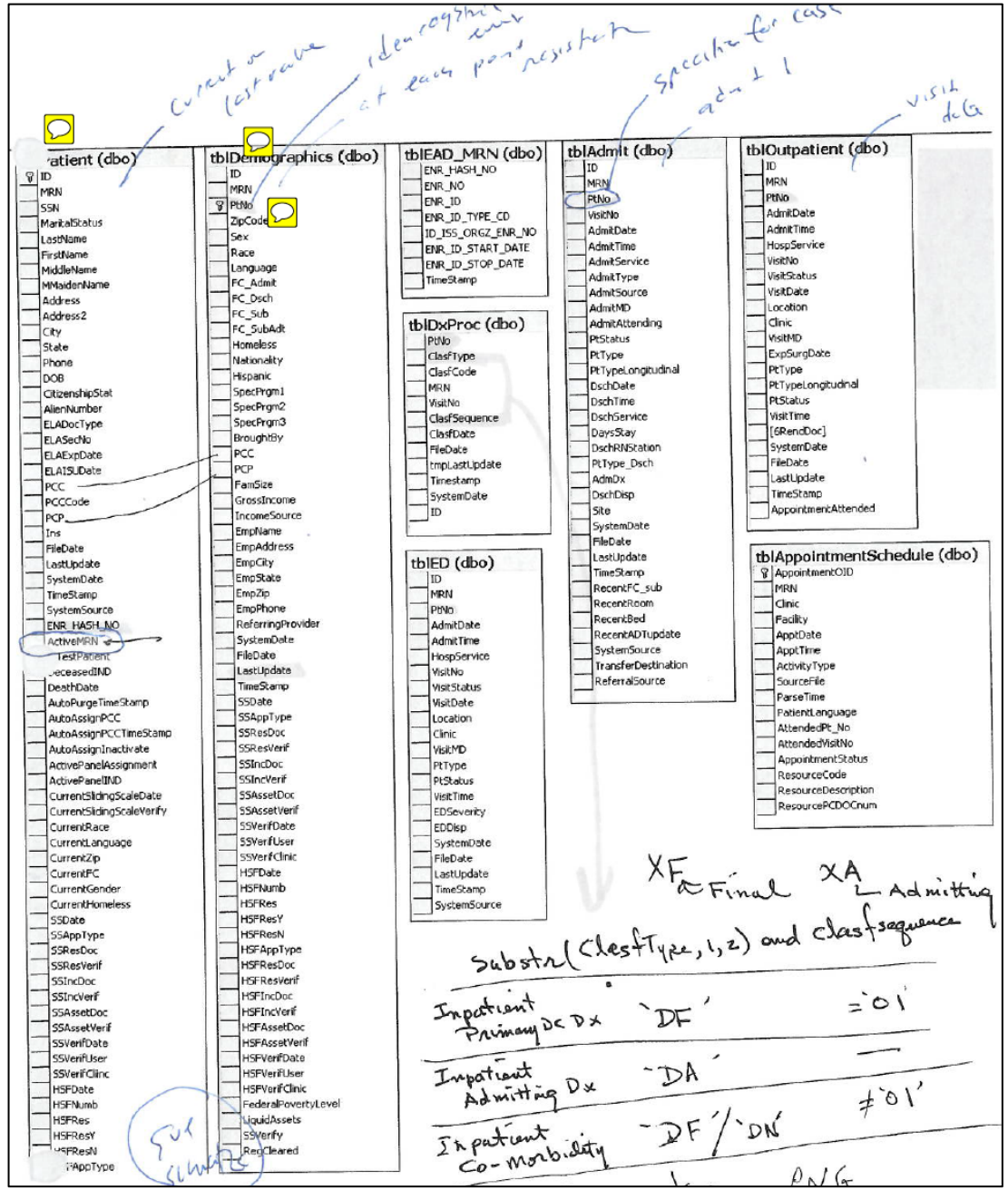
Step 2 – Decide how you will define your cohort of interest

- Do you obtain the variables to do so or give instructions for cohort definition?
- CNICS cohort inclusion criteria: two primary care visits within a year after January 1, 2001
- But ”primary care” is not a variable in the database
- Logic of the visit type classification
  - Resource code – one of any number of providers
  - Location – Ward 86
  - Date of visit
- Prefer to derive yourself

Defining the cohort

- Give me all patients who had a primary care visit after January 1, 2000
- (unfulfillable request)

- Give me all visits after January 1, 2001, and the accompanying resource code
- (maybe fulfill-able request)


tblDemographics (dbo)
ID
MRN
PtNo
ZipCode
Sex
Race
Language
FC_Admit
PC_Dsch
FC_Sub
FC_SubAdt
Homeless
Nationality
Hispanic
SpecPrgm1
SpecPrgm2
SpecPrgm3
BroughtBy
PCC
PCP
FamSize
GrossIncome
IncomeSource
EmpName
EmpAddress
EmpCity
EmpState
EmpZip
EmpPhone
ReferringProvider
SystemDate
FileDate
LastUpdate
TimeStamp
SSDate
SSAppType
SSResDoc
SSResVerif
SSIncDoc
SSIncVerif
SSAssetDoc
SSAssetVerif
SSVerifDate
SSVerifUser
SSVerifClinic
HSFDate
HSFNumb
HSFRes
HSFResY
HSFResN
HSFAppType
HSFResDoc
HSFResVerif
HSFIncDoc
HSFIncVerif
HSFAssetDoc
HSFAssetVerif
HSFVerifDate
HSFVerifUser
HSFVerifClinic
FederalPovertyLevel
LiquidAssets
SSVerify
RecCleared
tblEAD_MRN (dbo)
ENR_HASH_NO
ENR_NO
ENR_ID
ENR_ID_TYPE_CD
ID_ISS_ORGZ_ENR_NO
ENR_ID_START_DATE
ENR_ID_STOP_DATE
TimeStamp
tblDxProc (dbo)
PtNo
ClasfType
ClasfCode
MRN
VisitNo
ClasfSequence
ClasfDate
FileDate
tmpLastUpdate
Timestamp
SystemDate
ID
tblED (dbo)
ID
MRN
PtNo
AdmitDate
AdmitTime
HospService
VisitNo
VisitStatus
VisitDate
Location
Clinic
VisitMD
PtType
PtStatus
VisitTime
EDSeverity
EDDisp
SystemDate
FileDate
LastUpdate
TimeStamp
SystemSource
tblAdmit (dbo)
ID
MRN
PtNo
VisitNo
AdmitDate
AdmitTime
AdmitService
AdmitType
AdmitSource
AdmitMD
AdmitAttending
PtStatus
PtType
PtTypeLongitudinal
DschDate
DschTime
DschService
DaysStay
DschRNStation
PtType_Dsch
AdmDx
DschDisp
Site
SystemDate
FileDate
LastUpdate
TimeStamp
RecentFC_sub
RecentRoom
RecentBed
RecentADTupdate
SystemSource
TransferDestination
ReferralSource
tblOutpatient (dbo)
ID
MRN
PtNo
AdmitDate
AdmitTime
HospService
VisitNo
VisitStatus
VisitDate
Location
Clinic
VisitMD
ExpSurgDate
PtType
PtTypeLongitudinal
PtStatus
VisitTime
[6RencDoc]
SystemDate
FileDate
LastUpdate
TimeStamp
AppointmentAttended
tblAppointmentSchedule (dbo)
AppointmentOID
MRN
Clinic
Facility
ApptDate
ApptTime
ActivityType
SourceFile
ParseTime
PatientLanguage
AttendedPt_No
AttendedVisitNo
AppointmentStatus
ResourceCode
ResourceDescription
ResourcePCDOCnum
PCC
PCPCode
PCP
Ins
FileDate
LastUpdate
SystemDate
TimeStamp
SystemSource
ENR_HASH_NO
ActiveMRN
TestPatient
DeathDate
AutoPurgeTimeStamp
AutoAssignPCC
AutoAssignPCCTimeStamp
AutoAssignInactivate
ActivePanelAssignment
ActivePanelIND
CurrentSlidingScaleDate
CurrentSlidingScaleVerify
CurrentRace
CurrentLanguage
CurrentZip
CurrentPC
CurrentGender
CurrentHomeless
SSDate
SSAppType
SSResDoc
SSResVerif
SSIncDoc
SSIncVerif
SSAssetDoc
SSAssetVerif
SSVerifDate
SSVerifUser
SSVerifClinic
HSFDate
HSFNumb
HSFRes
HSFResY
HSFResN
Substr(ClasfType, 1, 2) and clasfsequence
Inpatient Primary Dc Dx `DF` =`01`
Inpatient Admitting Dx `DA` —
Inpatient Co-morbidity `DF`/`DN` ≠`01`

Codify decisions and formally communicate with data manager

- Specify name (might be different than the “native” name of the variable
- Specify format (date, numeric, text, coded with response categories)
- Avoid highly derived variables – strive to get more granular (can’t go other way)
- Ask for ”time invariant variables” in one table (e.g., sex, dob, etc.)
- As for all ”repeated measures” in separate “long tables” (e.g., appointments, visits, CD4 values, hemoglobin levels, etc.
- Make sure to obtain variables from individuals even if those precede the cohort entry date

STARTs Analysis Datasets Codebook

Data Files

File Name	Description
File 1	Wide, patient level characteristics
File 2	Wide, clinic level characteristics
File 3	Long, patient level characteristics
File 4	Long, clinic level characteristics
File 5	Wide, patient level VL info

File 1

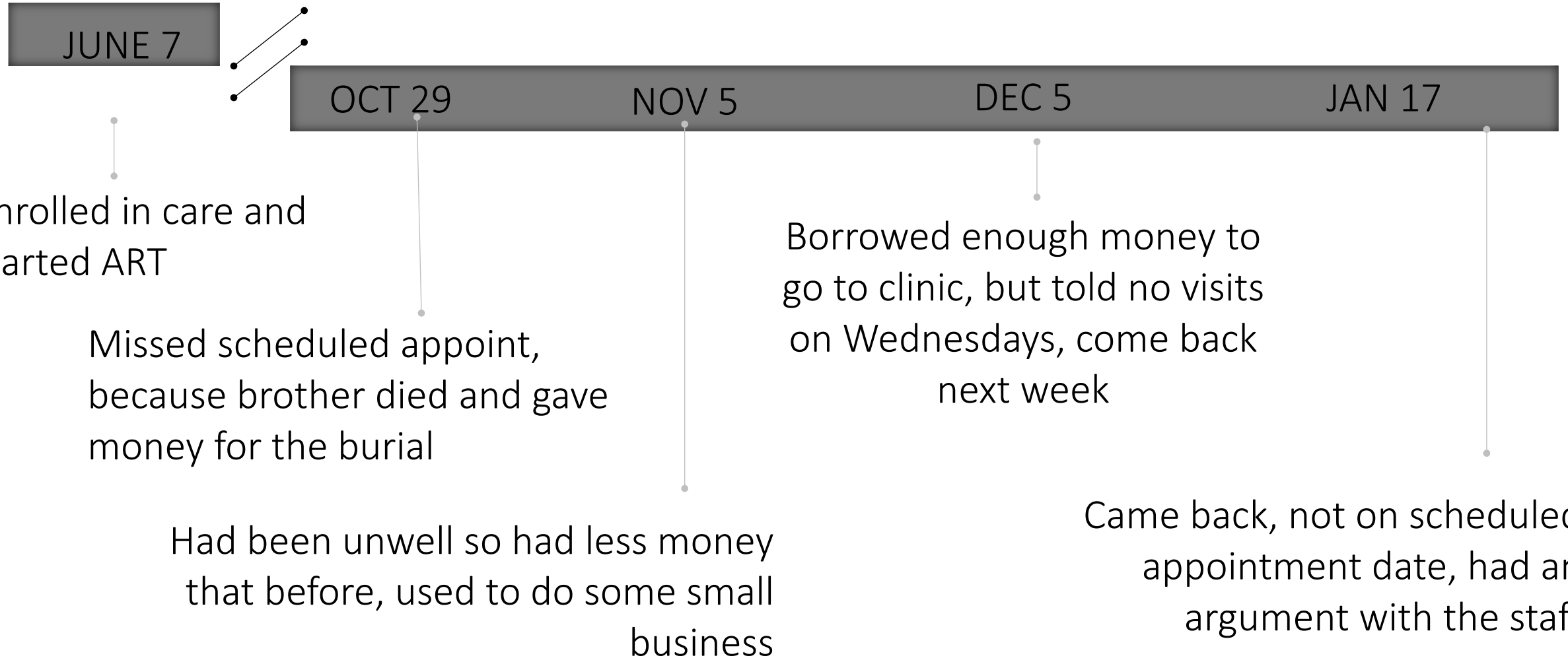
Variable Name	Format	Description	Notes
id	Text	Patient id	ptclinicno
sex	Numeric with codes: 1 = male 2 = female	Patient gender	
dob	Date	Date of birth	
age	Numeric	Patient age at date of eligibility for ART	used dob and feligd to calculate. In case where dob was missing and age was not: calculate DOB using (enrollment date – age)
clinic	Text	Name of clinic patient received care in at study commencement	
carentpt	Numeric with codes: 1=PMTCT 2=Medical 3=TB 4=STI 5=Under5 6=Inpatient 7=Outreach	Care entry point	

Other tips…

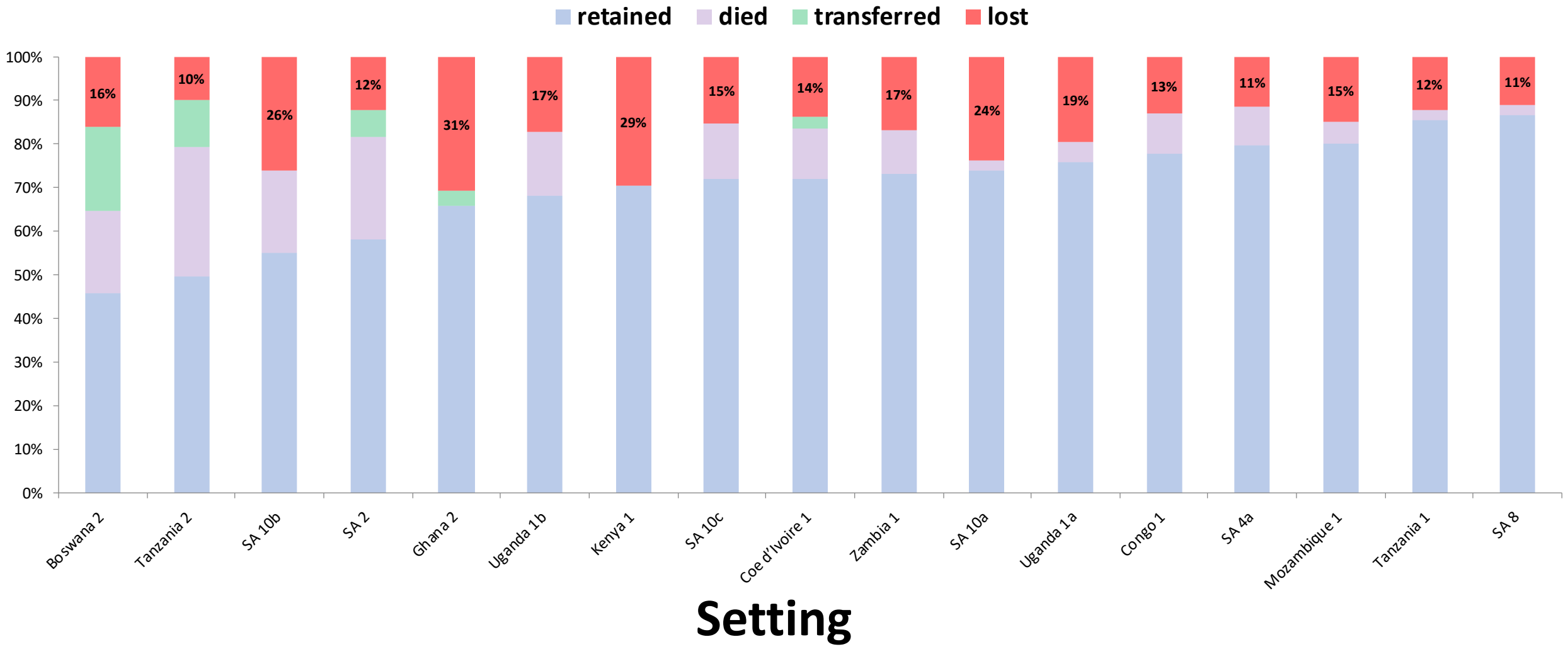
- Meet data manager face-to-face
- Codebook specification is iterative
- Data extraction is iterative

Once you know your ingredients you know what’s missing

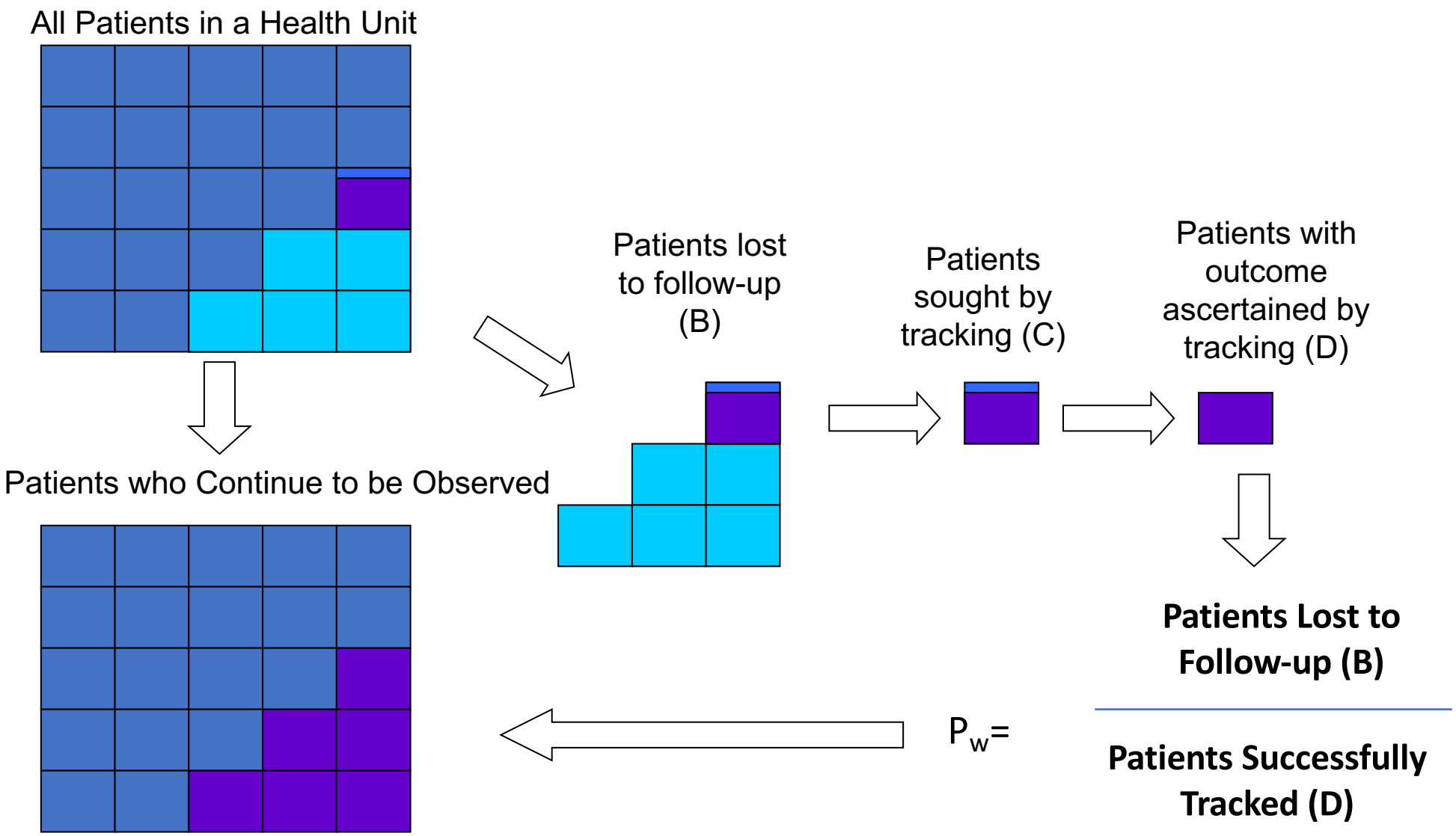
In-depth Interviews of Patient-Reported Reasons for Disengagement



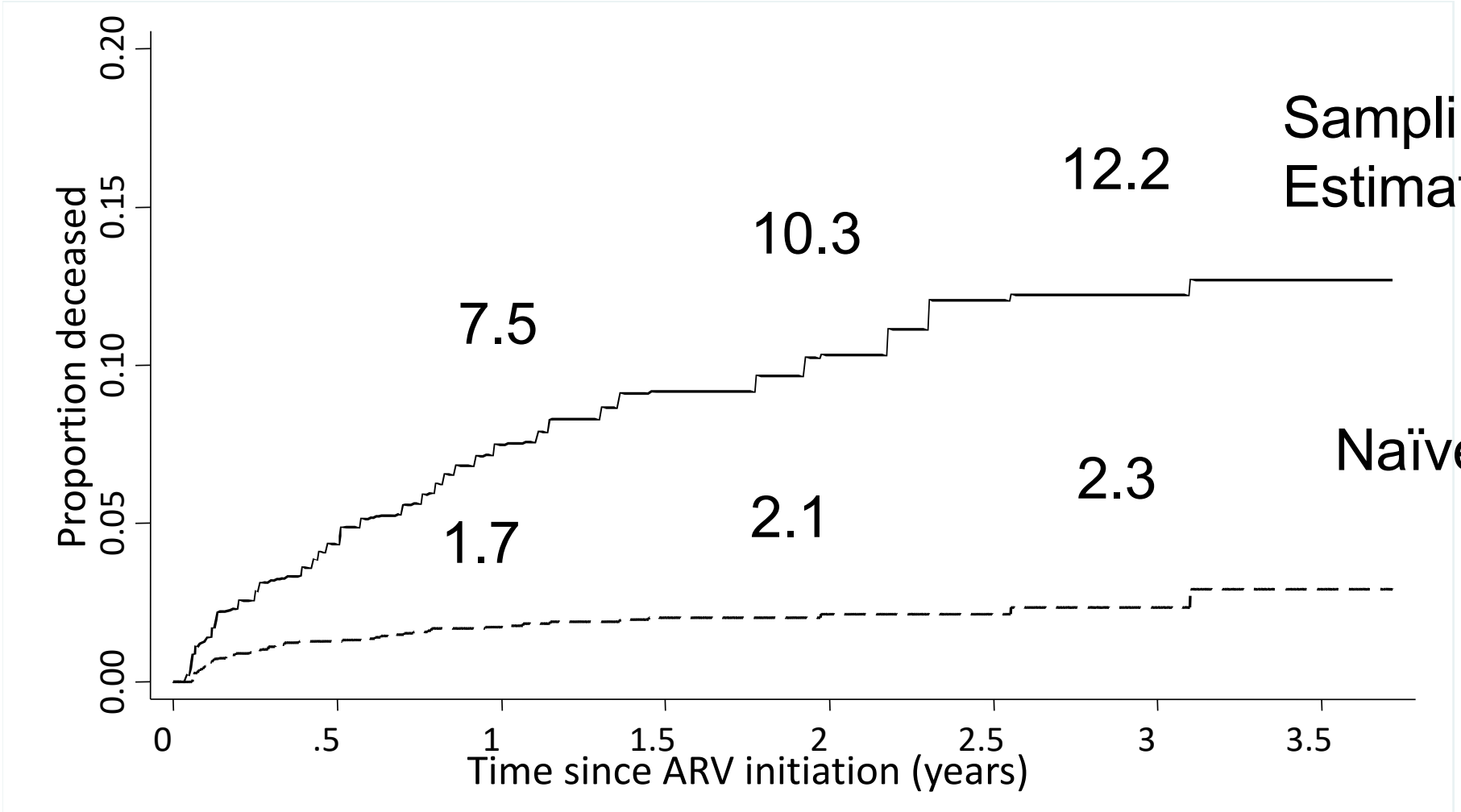
Understanding Retention in the Cascade of HIV Care



A Sampling-based Solution?

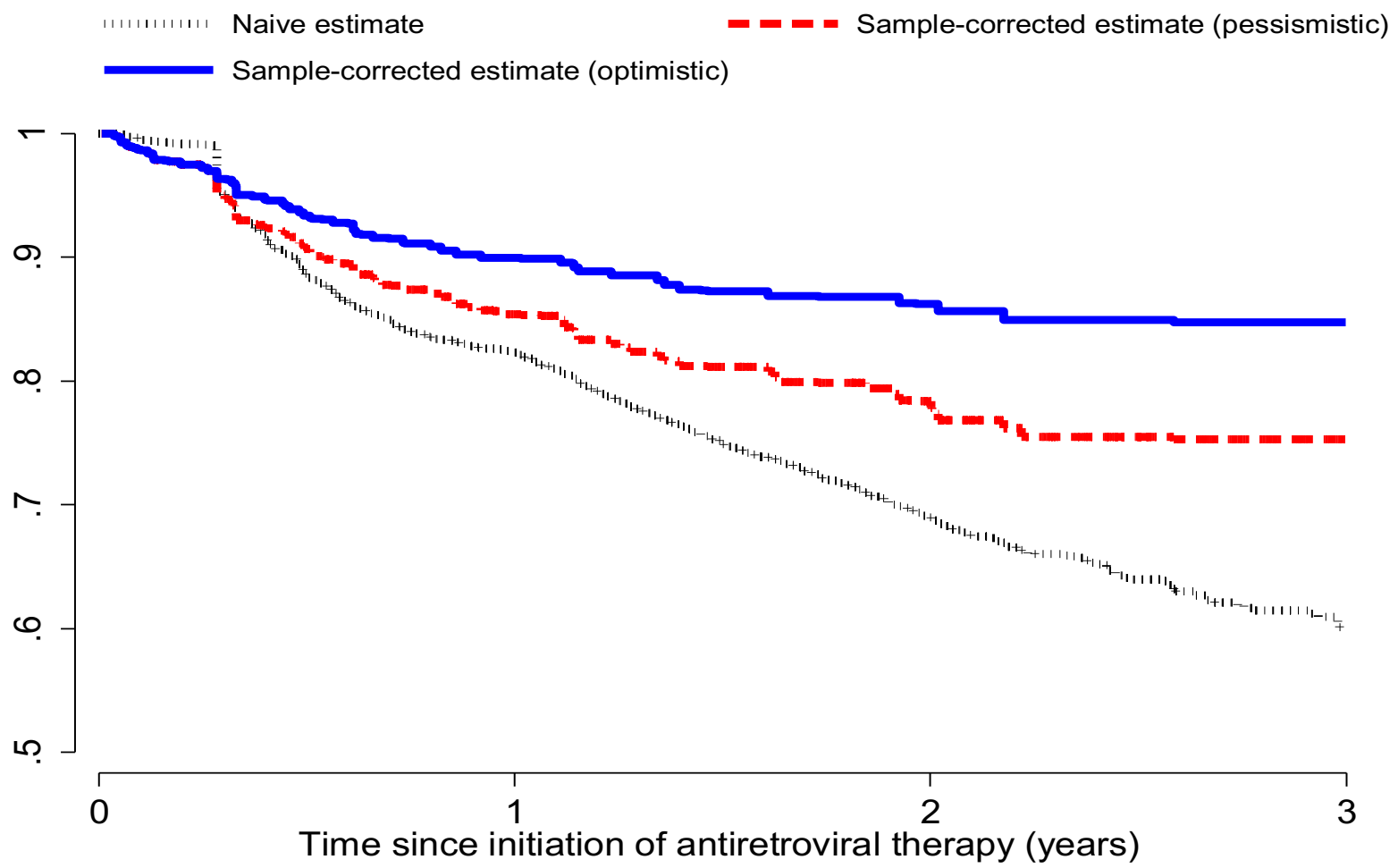


Sampling-based Estimates of Mortality in Mbarara, Uganda (N=3,628)



Geng et al., JAMA 2008; 300(5):506-507

Sampling-based Estimates of Retention in Care (N=3,628)



86.2% (95%CI: 82.9–89.5)

78.9% (95%CI: 75.2–82.6)

68.9% (95%CI: 67.1–70.8)

Geng et al., PLoS One 2011 6(7): e217972011

Sampling-weighted Factors Associated with Mortality (N=3,628)

	Un-weighted (outcomes as observed)			Sample-weighted (resampled outcomes)		
	HR	**95% CI**	**P-value**	**HR**	**95% CI**	**P-value**
Age (per 10 years)	**1.18**	**0.87-1.59**	**0.29**	**1.37**	**1.04-1.81**	**0.03**
Male sex	**1.86**	**1.06-3.26**	**0.03**	**1.02**	**0.57-1.83**	**0.96**
Weight < 40 kg	**1.64**	**0.67-4.03**	**0.28**	**3.04**	**1.58-5.85**	**<0.01**
Pre-ART CD4 count						
<= 50 cells/mm^3	**ref**			**ref**		
51-100 cells/mm^3	**0.56**	**0.25-1.24**	**0.15**	**0.58**	**0.24-1.36**	**0.21**
101-200 cells/mm^3	**0.24**	**0.08-0.68**	**<0.01**	**0.21**	**0.08-0.56**	**<0.01**
> 200 cells/mm^3	**0.15**	**0.03-0.67**	**<0.01**	**0.2**	**0.06-0.73**	**0.02**
Distance from home to clinic (per 10 km)	**0.93**	**0.85-1.01**	**0.1**	**1.01**	**0.94-1.09**	**0.77**
WHO clinical stage 4	**1.25**	**0.47-3.28**	**0.65**	**0.82**	**0.43-1.55**	**0.54**
Year of ART initiation						
2004	**ref**			**ref**		
2005	**0.69**	**0.37-1.29**	**0.25**	**1.91**	**1.07-3.41**	**0.03**
2006	**0.50**	**0.22-1.10**	**0.09**	**0.88**	**0.37-2.08**	**0.78**
2007	**0.27**	**0.06-1.22**	**0.09**	**0.15**	**0.03-0.71**	**0.02**

Geng et al., TMIH (15) s1:63–69 2010

Cooking with what's in the kitchen is hard…

- You have to be a good chef to make something good with what's in the kitchen.
  - Creative
  - Appropriate (good judgement)
  - Iron Chef

Real world data for implementation science

- Measurement issues in real world data are complicated
- One kind of phronesis of implementation science
- No clear single source of definitive information
- Experiential (kind of like being a resident)
- Extremely valuable..