IHME

Institute for Health Metrics and Evaluation

The Mathematics of Verbal Autopsy

how to track the leading causes of death globally

Abraham D Flaxman

July 11, 2016

UNIVERSITY of WASHINGTON

THE LANCET

Subscribe | Register | Login

Search for [ ] in [All Fields] GO Advanced Search

Home | Journals | Specialties | Clinical | Global Health | Multimedia | Conferences | Information for | Healthcare Jobs

Global Burden of Disease Study 2010

Published Dec 13, 2012

Executive summary

The Global Burden of Disease Study 2010 (GBD 2010) is the largest ever systematic effort to describe the global distribution and causes of a wide array of major diseases, injuries, and health risk factors. The results [illegible] infectious diseases, maternal and child illness, and malnutrition [illegible] fewer deaths are premature illnesses than they did twenty years ago. As a result, [illegible] are dying every year, [illegible] more young and middle [illegible] adults are living and suffering from disease and injury, as [illegible] diseases such as cancer and heart disease, [illegible] the dominant causes of death and disability worldwide. Since 1970, men and women worldwide gained slightly more than ten years of life expectancy overall, but they spend more years living with injury and illness.

GBD 2010 consists of seven Articles, each containing a wealth of data on different aspects of the study (including data for different countries and world regions, men and women, and different age groups), while accompanying Comments include reactions to the study's publication from WHO Director-General Margaret Chan and World Bank President Jim Yong Kim. The study is described by *Lancet* Editor-in-Chief Dr Richard Horton as "a critical contribution to our understanding of present and future health priorities for countries and the global community."

Audio

MP3 Audio (1): 1 2 3

▶ 00:00/00:00

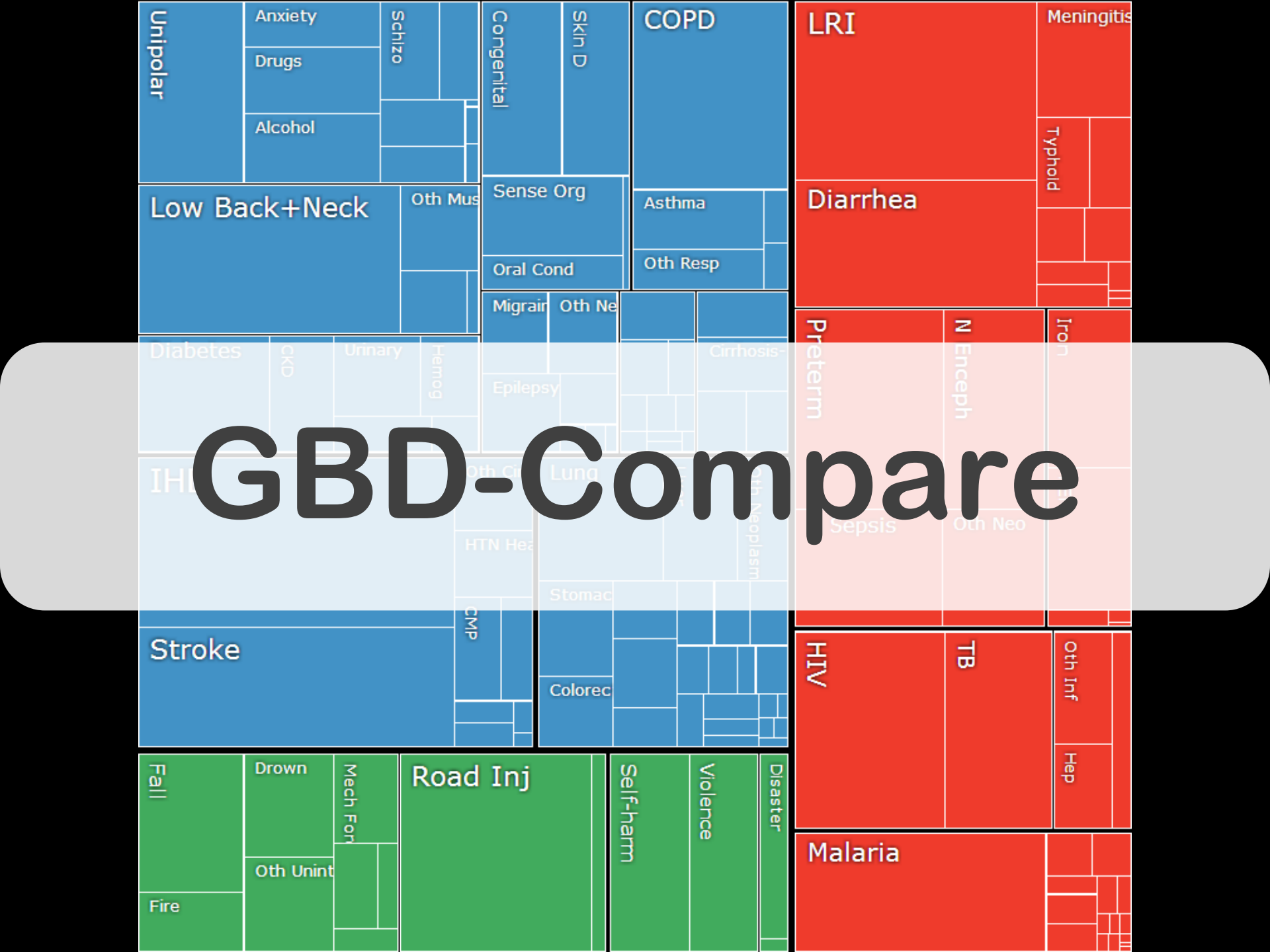
Global Burden of Diseases
Richard Horton with a background and overview of GBD 2010.
Download this audio (8.10Mb)

Comments

Interactive graphs and figures interpret the GBD 2010 data

Click on the image below to view the interactive graphs and

DALYs = YLL + YLD


Unipolar
Anxiety
Drugs
Alcohol
Schizo
Congenital
Skin D
COPD
Asthma
Oth Resp
Sense Org
Oral Cond
Low Back+Neck
Oth Mus
Migrain
Oth Ne
Diabetes
CKD
Urinary
Hemog
Epilepsy
Cirrhosis-
IHD
Oth Ci
Lung
HTN Hea
Stomac
CMP
Stroke
Colorec
LRI
Meningitis
Typhoid
Diarrhea
Preterm
N Enceph
Iron
Sepsis
Oth Neo
HIV
TB
Oth Inf
Hep
Malaria
Fall
Fire
Drown
Oth Unint
Mech For
Road Inj
Self-harm
Violence
Disaster
GBD-Compare

Outline of minitutorial

Overview of Disease Burden Measurement

Deep Dive into Cause-of-Death Estimation

The Challenges to Come

Figure 7. GBD 2010 Data and Model Flow Chart

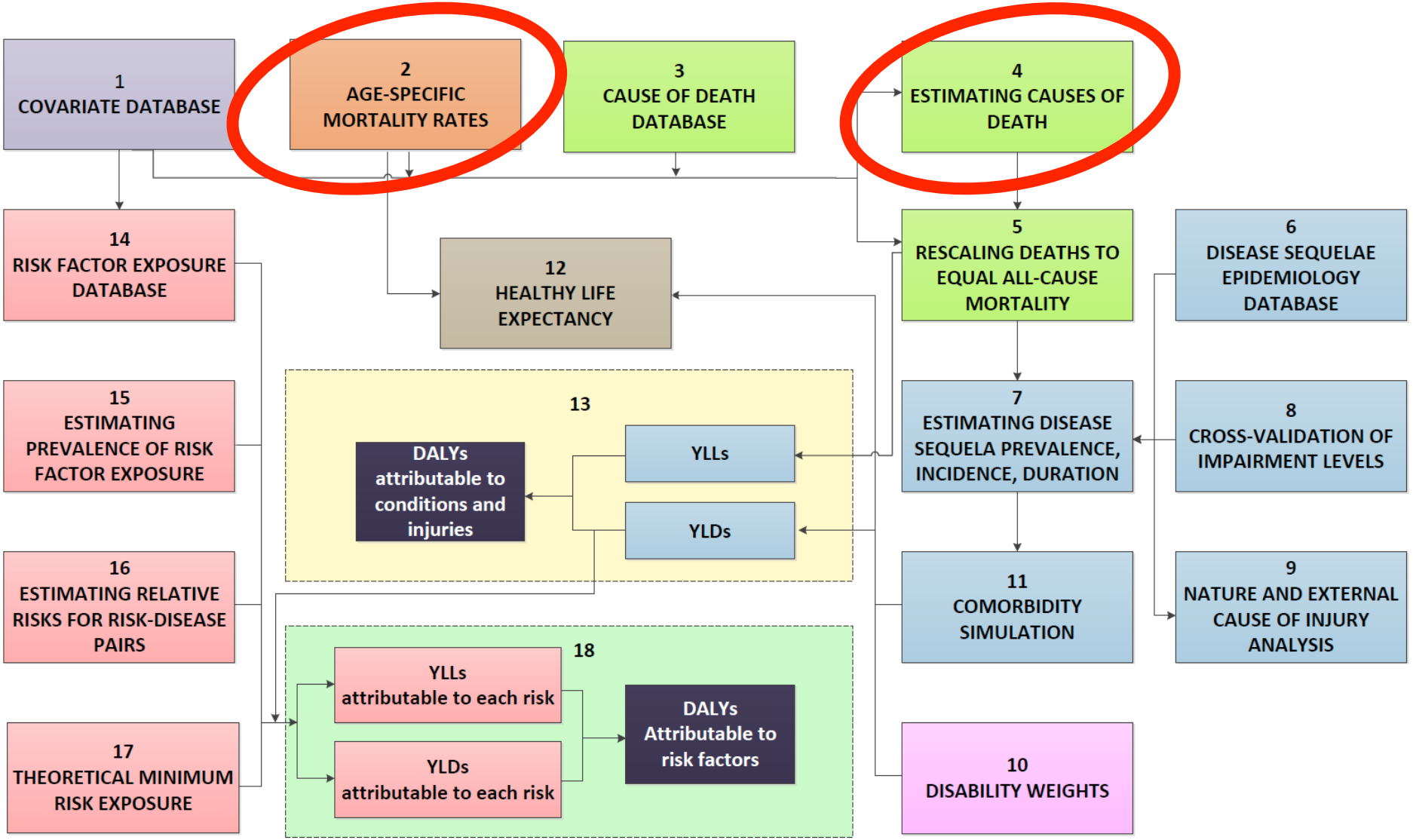
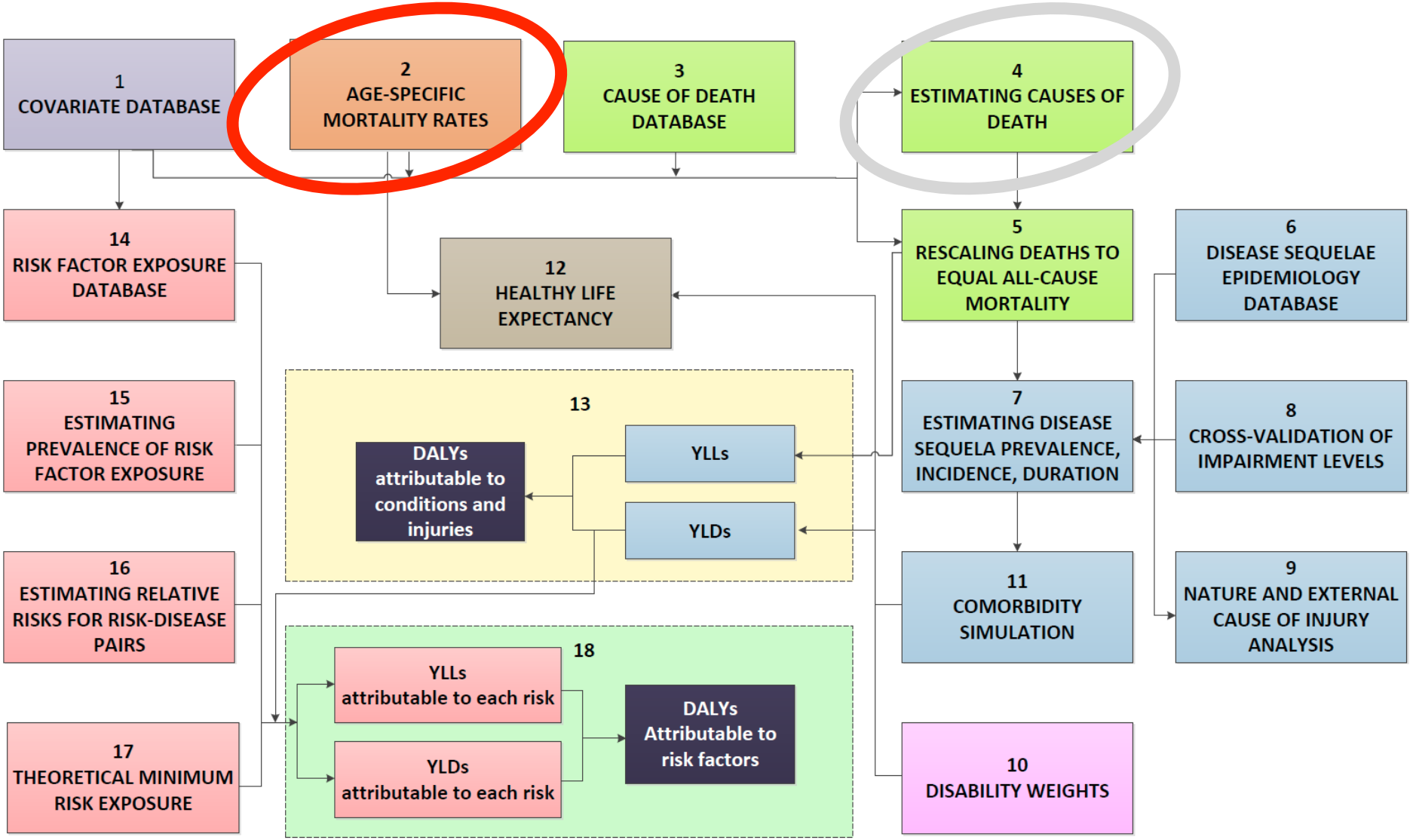
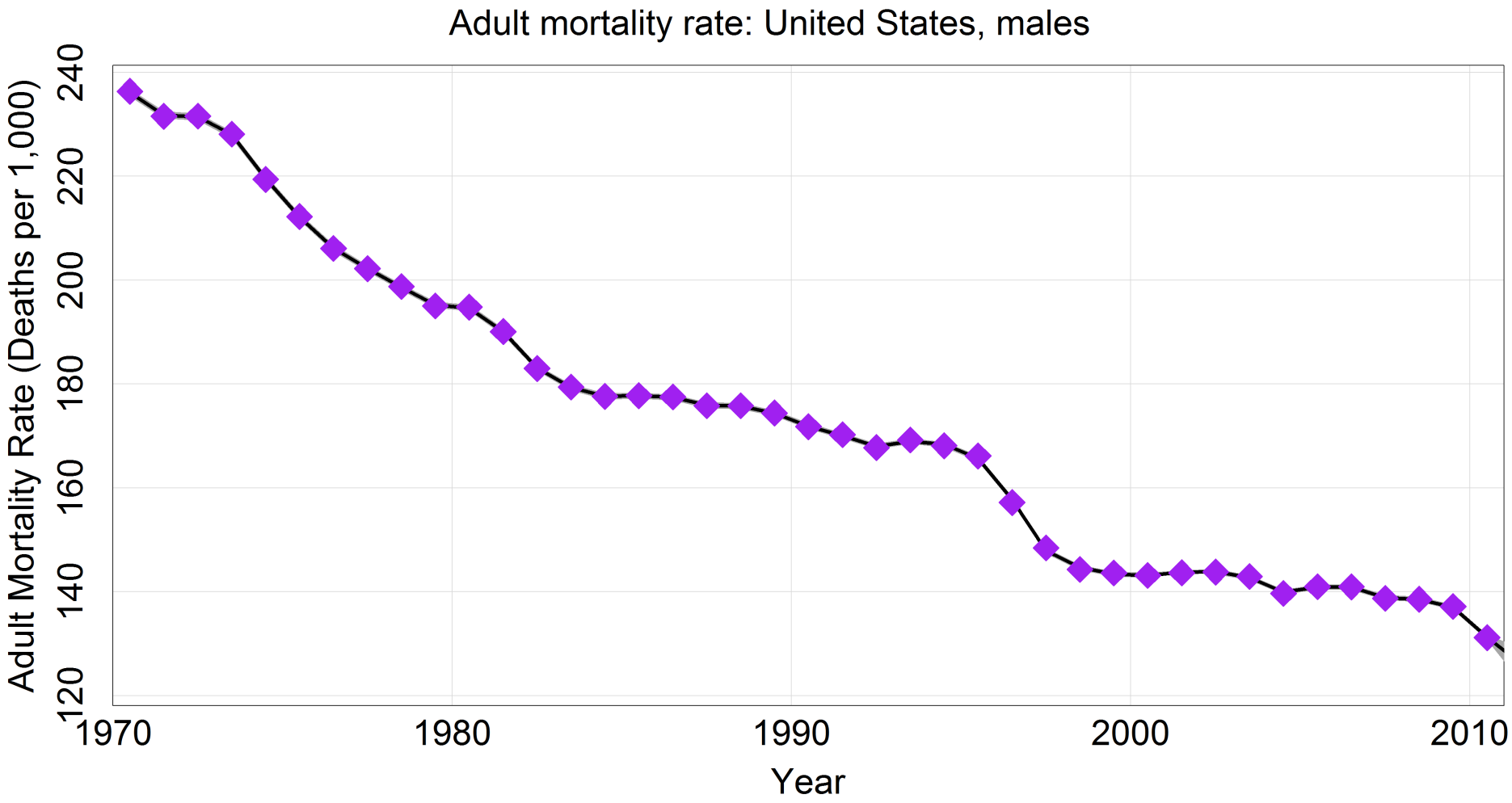
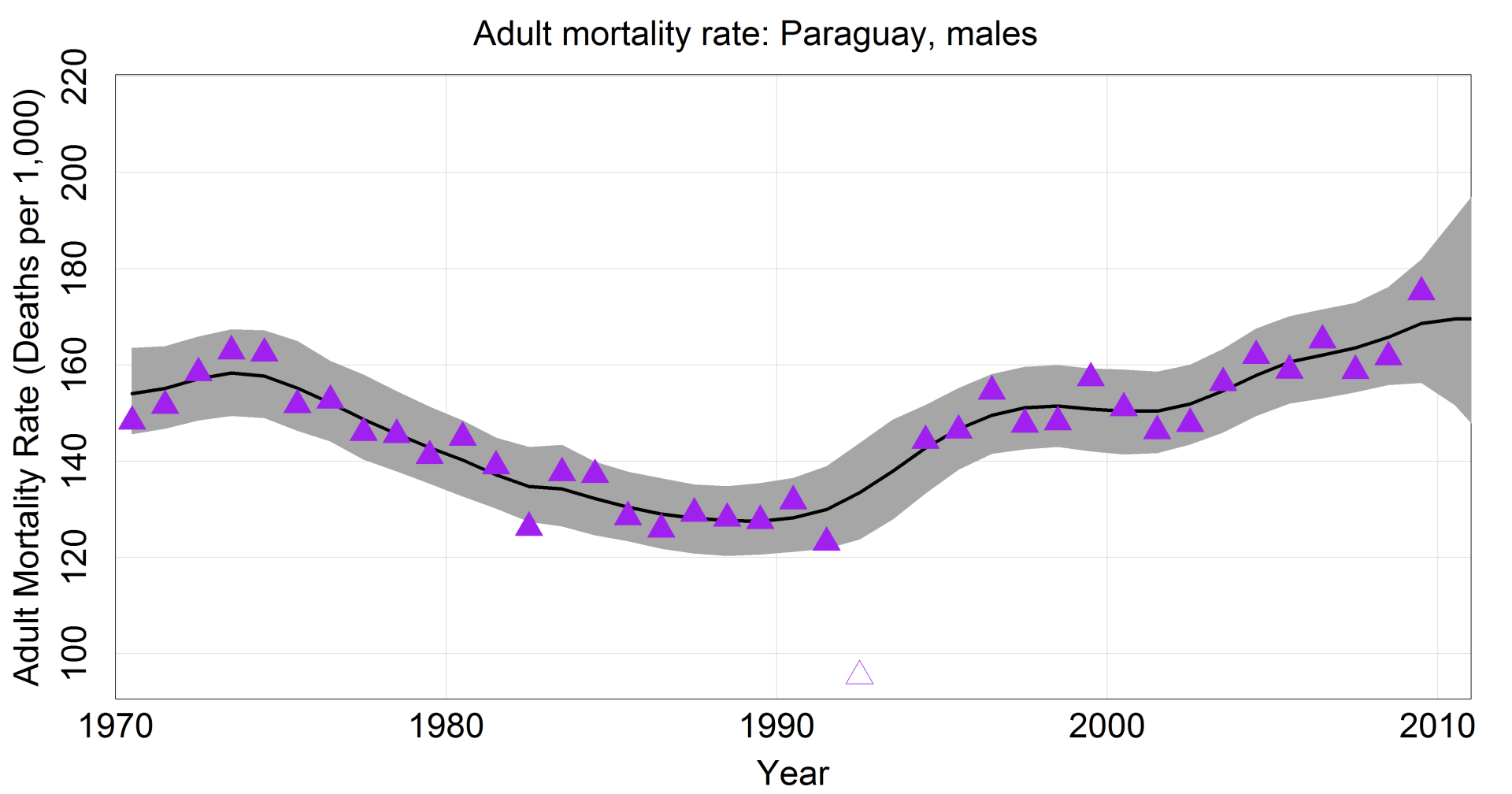


Figure 7. GBD 2010 Data and Model Flow Chart



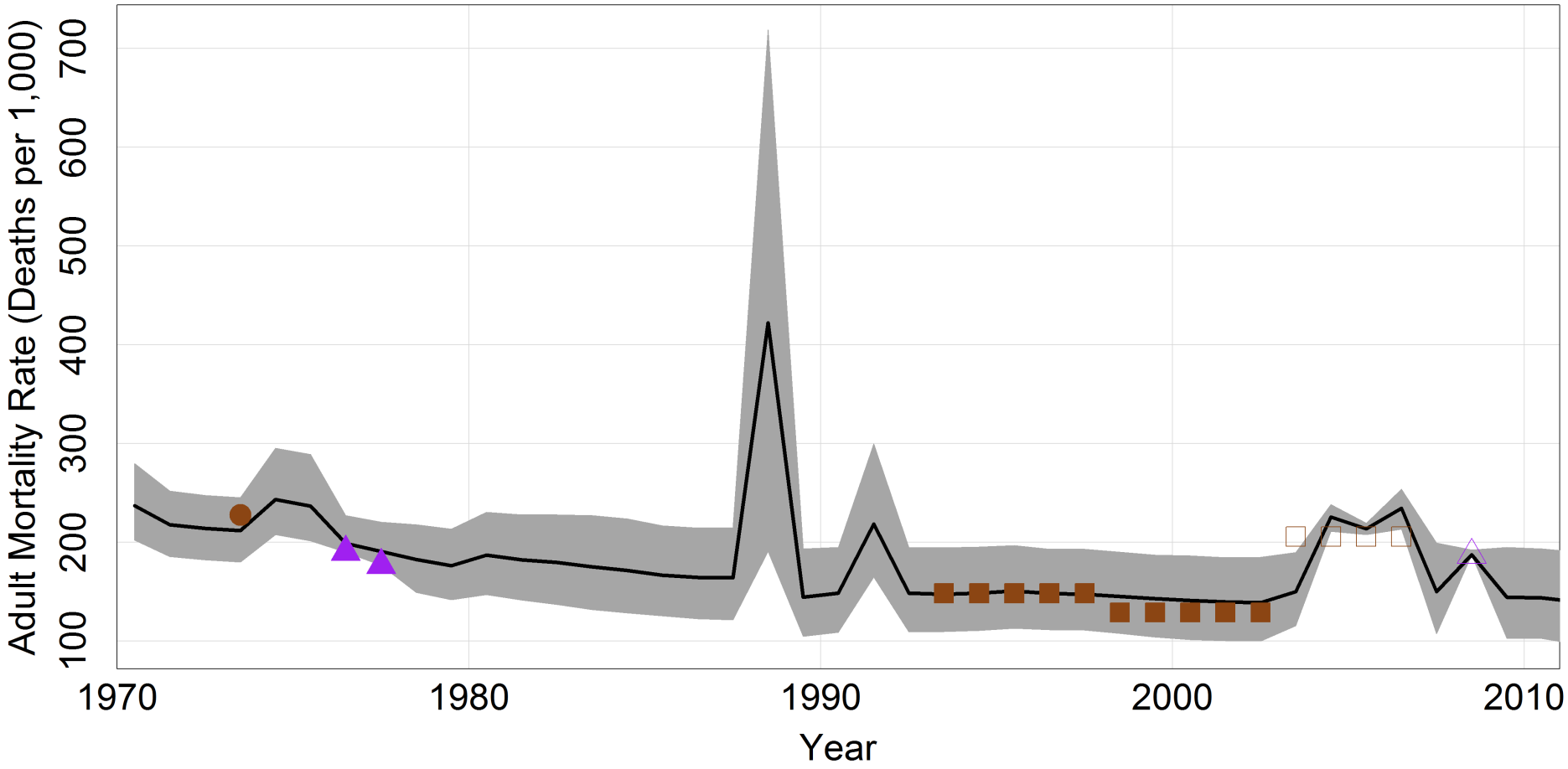


*Hollow points indicate data excluded from the analysis



*Hollow points indicate data excluded from the analysis

Adult mortality rate: Iraq, males

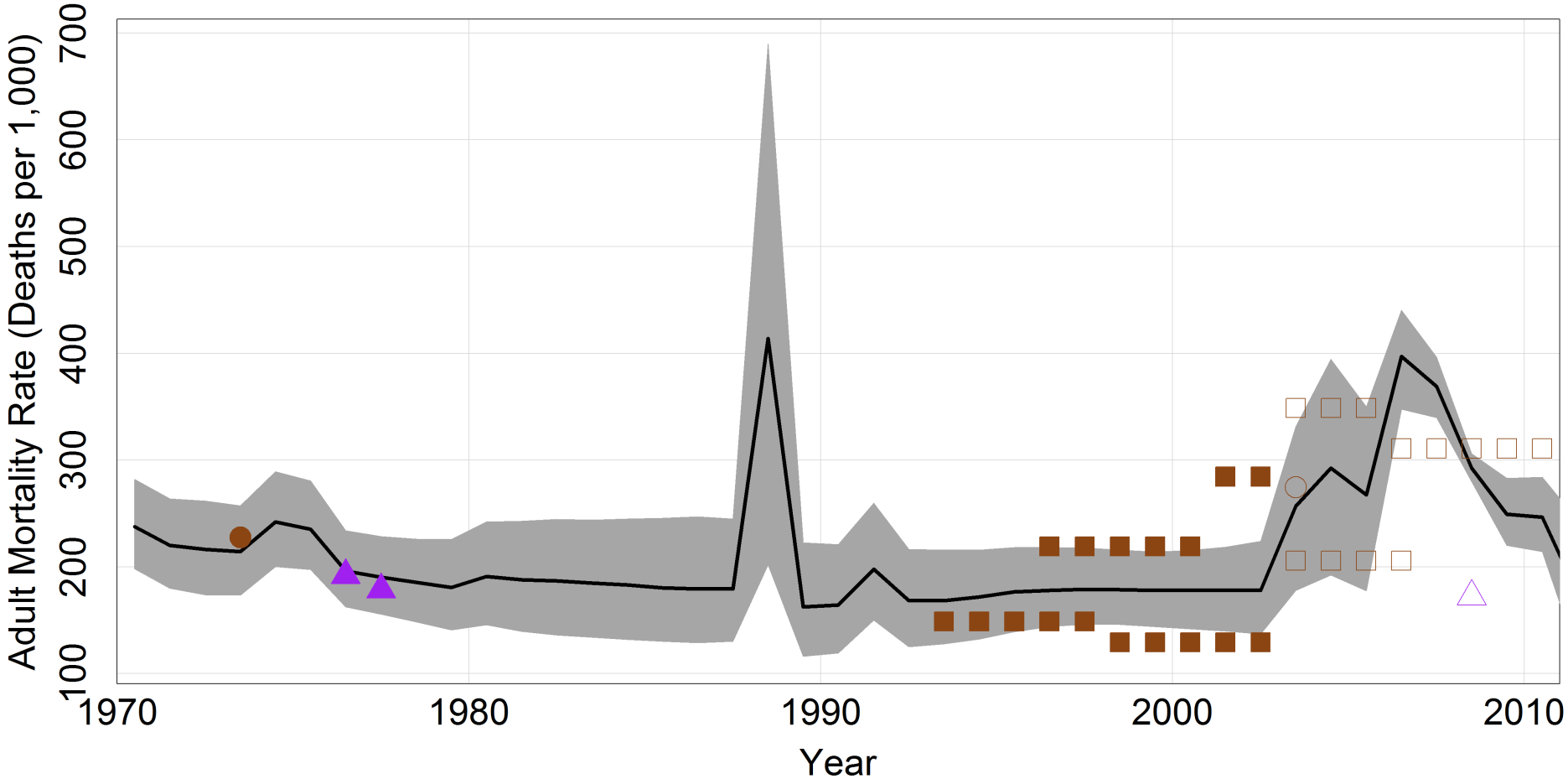


Data Source and Type:

- Gaussian Process Regression with Uncertainty
- Vital Registration - DDM Adjusted
- Iraq Family Health Survey - Sibling History
- Demographic Sample Survey - Unadjusted

*Hollow points indicate data excluded from the analysis

Adult mortality rate: Iraq, males

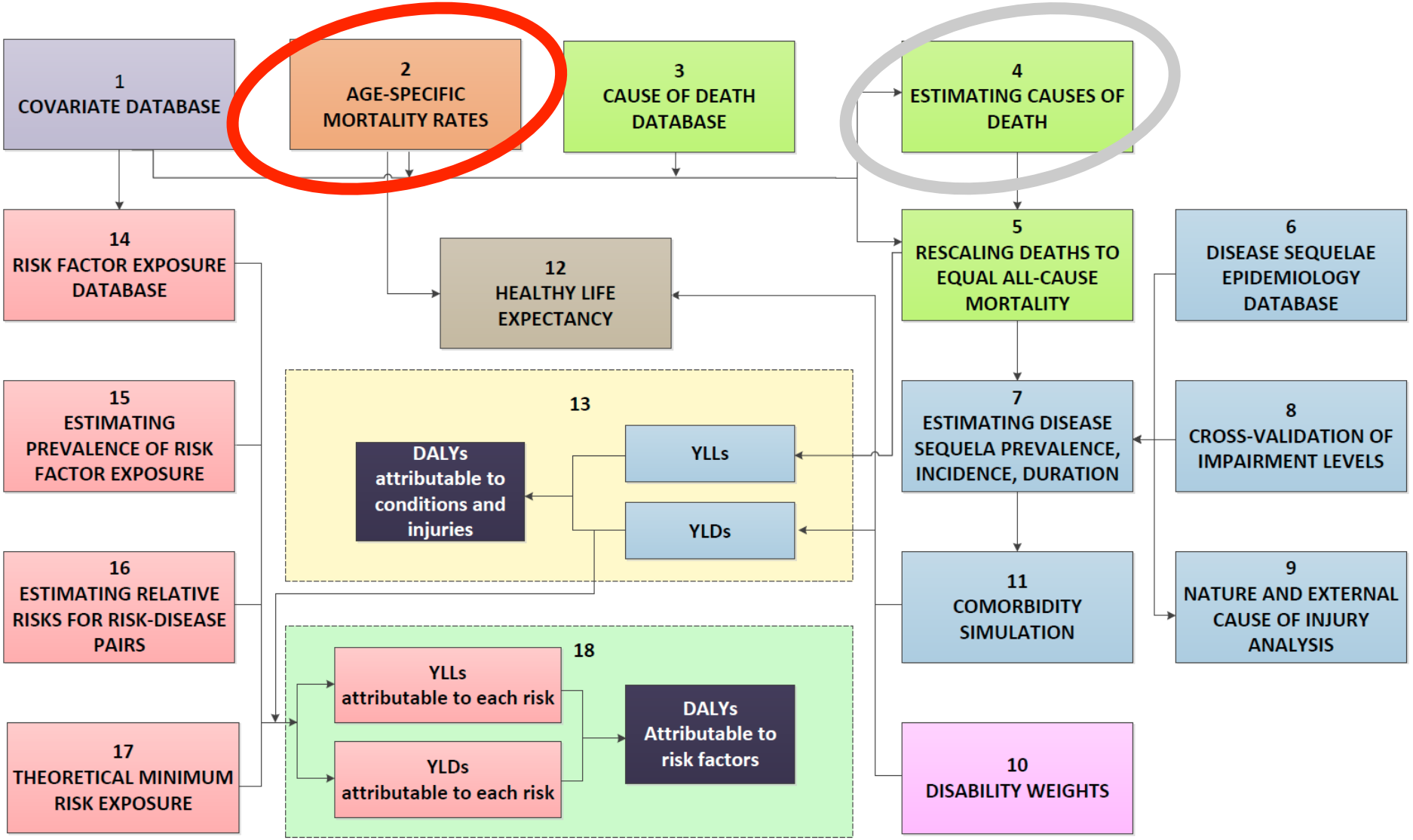


Data Source and Type:

- Gaussian Process Regression with Uncertainty
- Vital Registration - DDM Adjusted
- Iraq Family Health Survey - Sibling History
- Demographic Sample Survey - Unadjusted

*Hollow points indicate data excluded from the analysis

Figure 7. GBD 2010 Data and Model Flow Chart



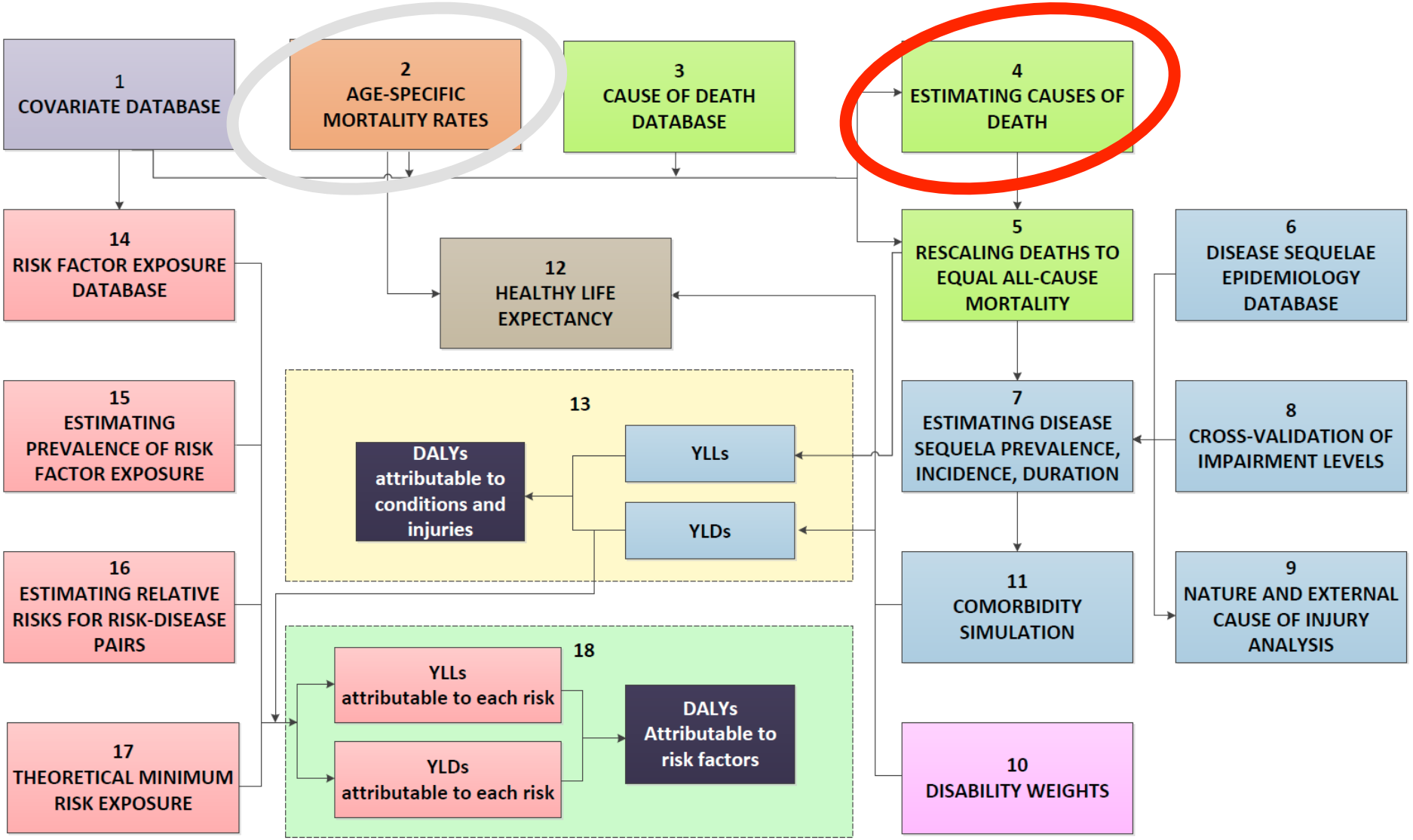
PBF: Application Process

To apply:

You can apply to the Post-Bachelor Fellowship using our online application.

Deadline: January 10, 2016

Figure 7. GBD 2010 Data and Model Flow Chart



INTERNATIONAL FORM OF MEDICAL CERTIFICATE OF CAUSE OF DEATH

Cause of death		Approximate interval between onset and death
I Disease or condition directly leading to death*	(a) Pulmonary embolism	hours
	due to (or as a consequence of)	
Antecedent causes Morbid conditions, if any, giving rise to the above cause, stating the underlying condition last	(b) Pathological fracture of Femur	2 days
	due to (or as a consequence of)	
	(c) Secondary malignant neoplasm (femur)	2 months
	due to (or as a consequence of)	
	(d) Malignant neoplasm of breast (nipple)	1 year
II Other significant conditions contributing to the death, but not related to the disease or condition causing it	Essential hypertension	5 years
	Obesity	10 years

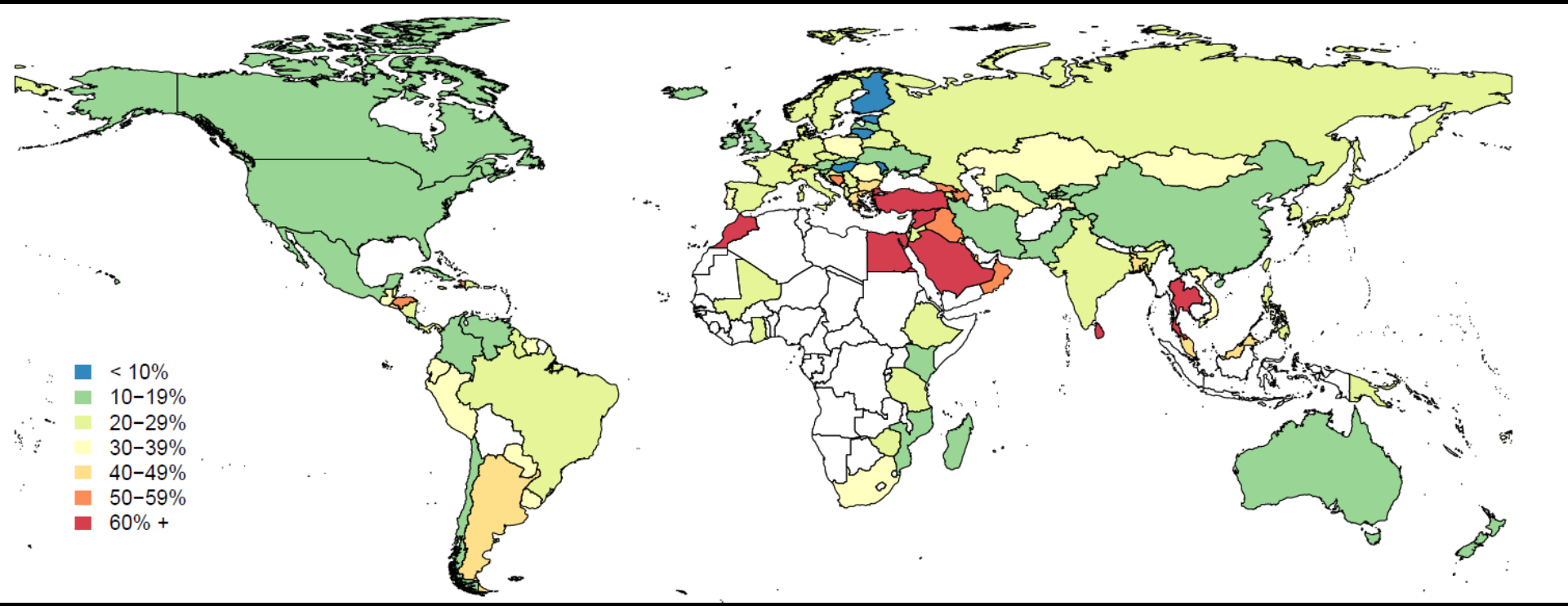
**This does not mean the mode of dying, e.g. heart failure, respiratory failure. It means the disease, injury, or complication that caused death.*

INTERNATIONAL FORM OF MEDICAL CERTIFICATE OF CAUSE OF DEATH

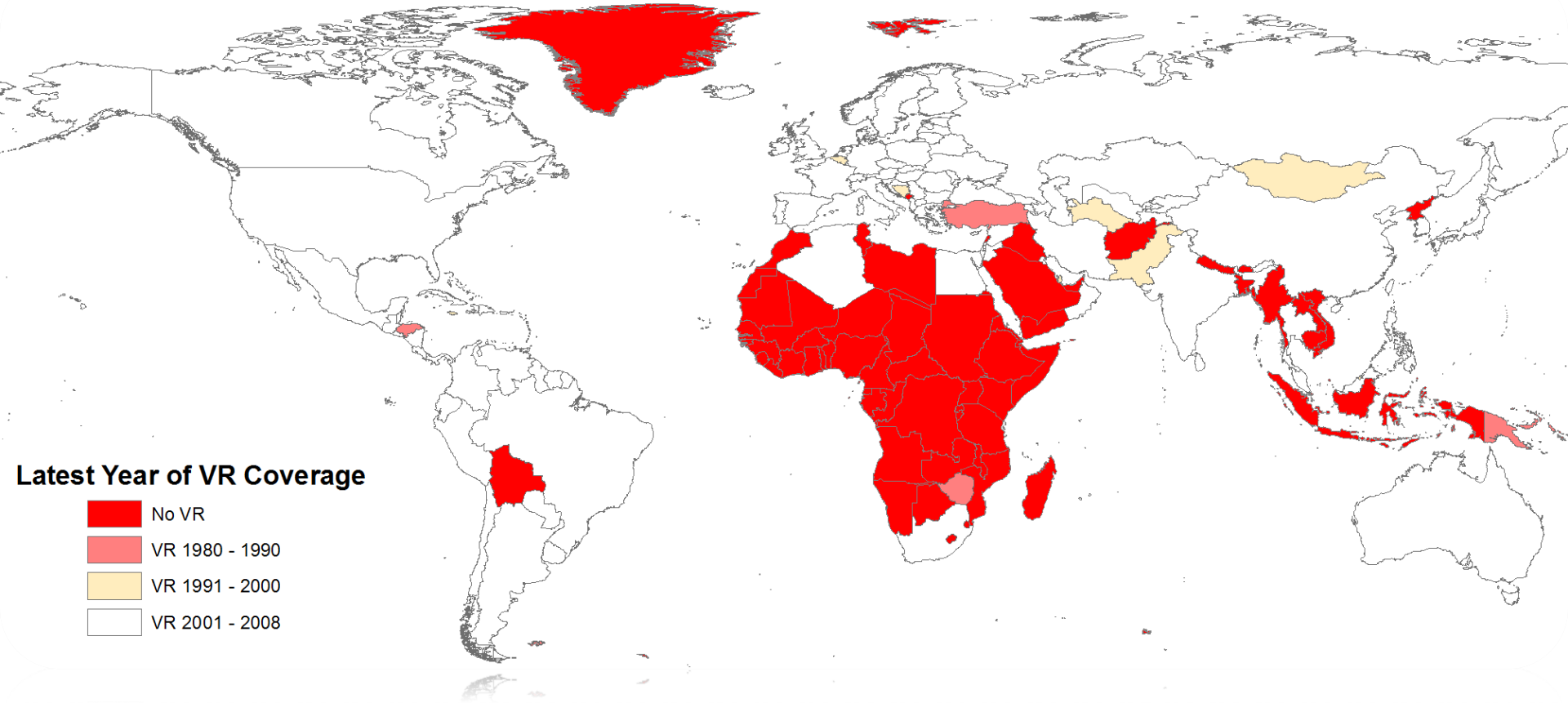
Cause of death		Approximate interval between onset and death
I Disease or condition directly leading to death*	(a) Acute myocardial infarction	1 hour
	due to (or as a consequence of)	
Antecedent causes Morbid conditions, if any, giving rise to the above cause, stating the underlying condition last	(b) Essential Hypertension	1 year
	due to (or as a consequence of)	
	(c) Diabetes mellitus 2	14 years
	due to (or as a consequence of)	
	(d) Hypothyroidism	24 hours
II Other significant conditions contributing to the death, but not related to the disease or condition causing it		

**This does not mean the mode of dying, e.g. heart failure, respiratory failure. It means the disease, injury, or complication that caused death.*

Percent of “Garbage” Death Certificates



Death Registration Coverage



How would you fill in these blanks?

- I'm about to tell you how I do, but my solution is not as good as I would like. So I'll give you a minute to think about how you would do it.

- For real, go ahead.

- Then we will "pair and share". This whole bit of *active learning* will take about 7 minutes.

- If you have something good, tweet it to me: @healthyalgo #SIAMAN16 (or email abie@uw.edu)

Verbal Autopsy

SECTION 5
MODULE 1. GENERAL ILLNESS LEADING TO DEATH
SPECIFIC QUESTIONS TO ELICIT SYMPTOMS AND SIGNS OF THE LAST ILLNESS

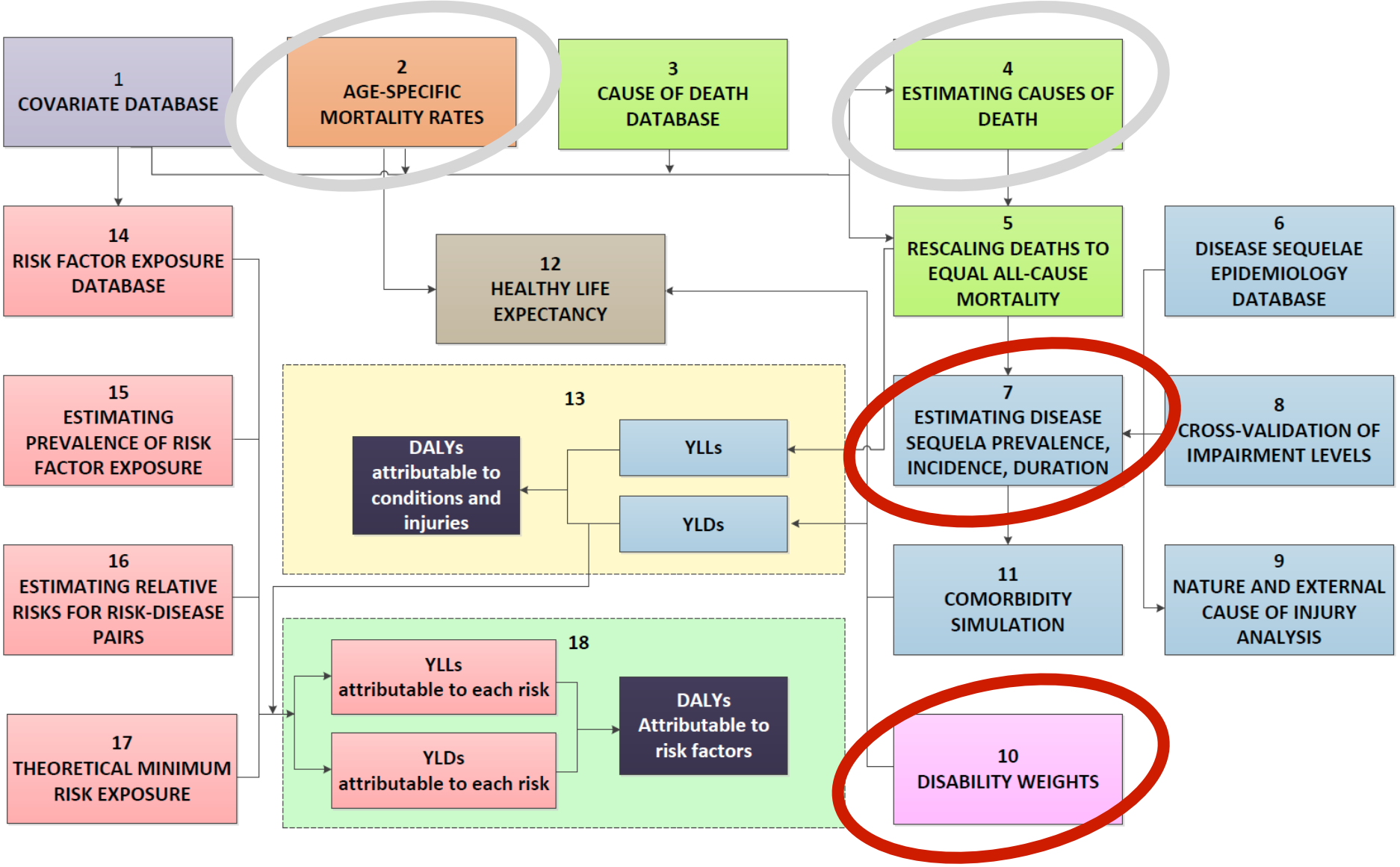
NO.	QUESTIONS AND FILTERS	CODING CATEGORIES	SKIP
501	Did _____(NAME) have fever during her last illness?	YES.....1	
		NO 2	502
		DON'T KNOW.....8	502
501A	How many days/months before her death did the fever start and end?	START |____|____| |____|____|	
		MILD 2	
		DON'T KNOW/UNSURE.....8	
501C	Was the fever continuous or on and off?	CONTINUOUS.....1	
		AFTER EVERY 1 - 2 DAYS.....2	
		AT NIGHT ONLY 3	
		OTHER.....7	

SECTION 4. DESCRIPTIVE REPORT OF ILLNESS AND EVENTS THAT LED TO THE DEATH

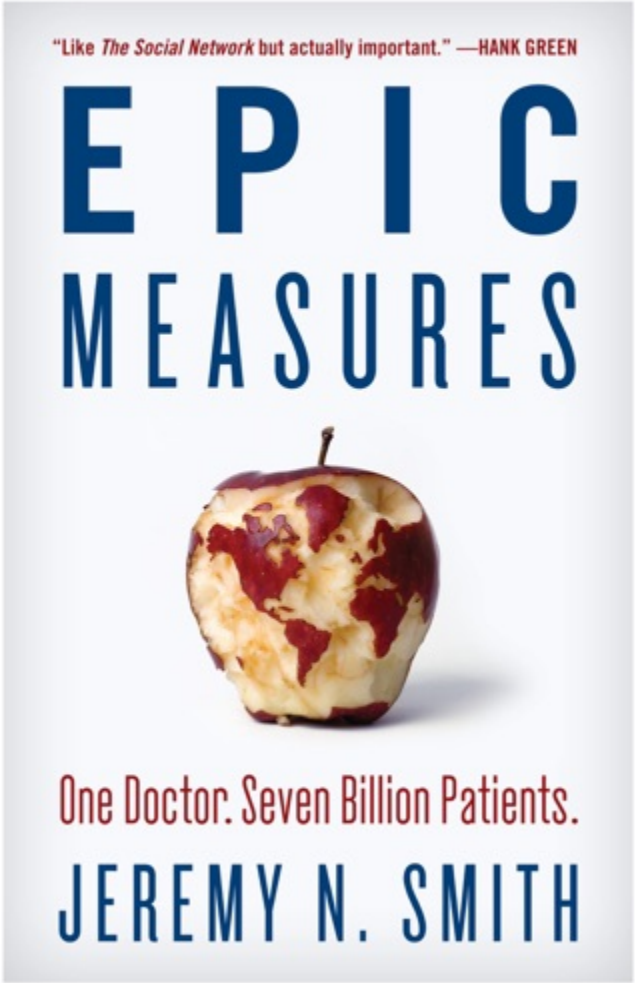
401. Explain to the respondent that we would like to hear the details about everything that happened during the last illness before _____ death starting from the beginning of the ilness and also about what happened during the final hours of the woman's death.

Verbatim:

Figure 7. GBD 2010 Data and Model Flow Chart



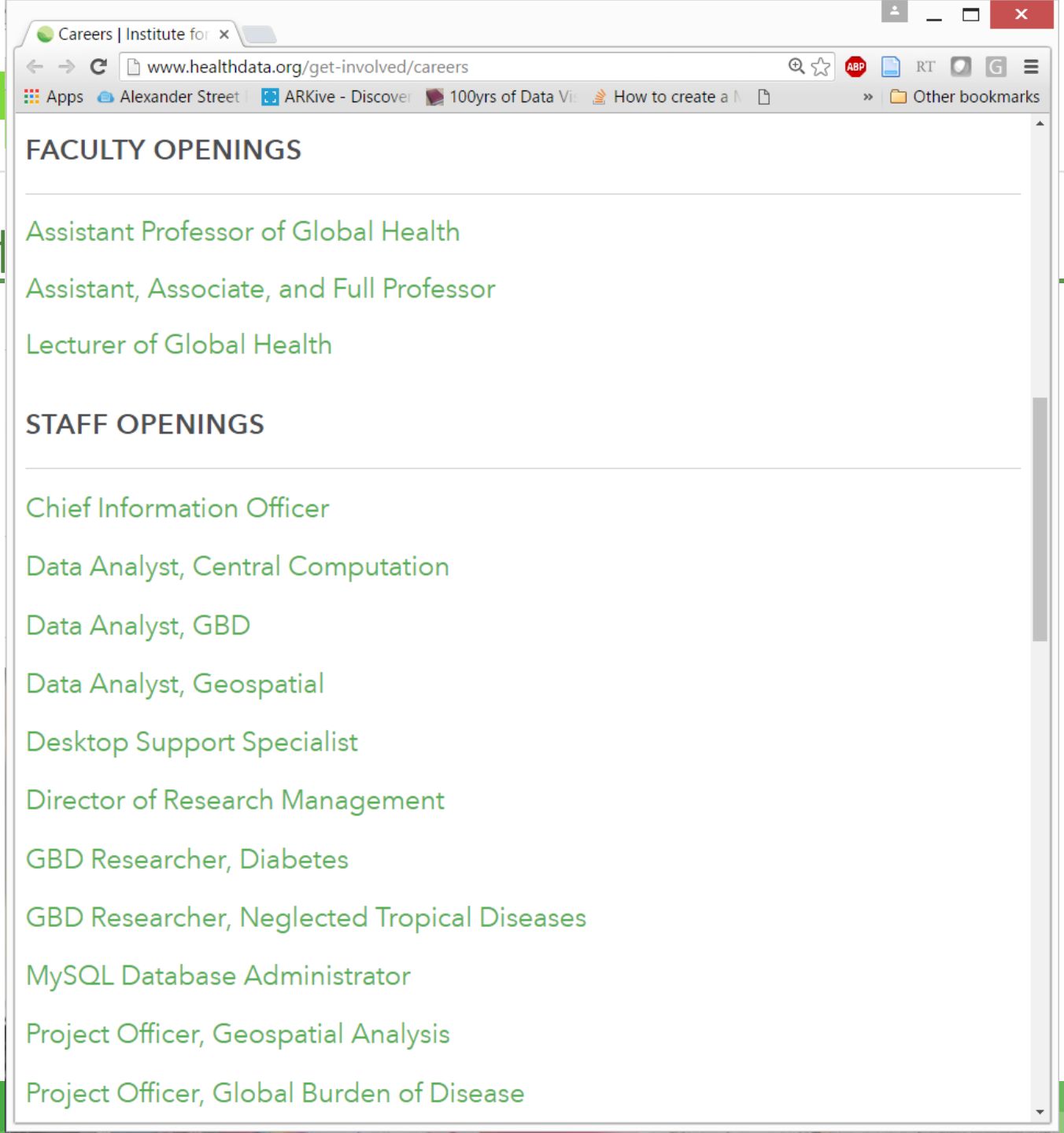
Book length version of this half of the mini-tutorial:



"Jeremy Smith's engaging story of a man obsessed with the numbers, and the mortal dramas they tell, reads like a novel and is better than any textbook or survey of this planet's health." ---*Paul Farmer*

IHMI

htt d/



FACULTY OPENINGS

Assistant Professor of Global Health

Assistant, Associate, and Full Professor

Lecturer of Global Health

STAFF OPENINGS

Chief Information Officer

Data Analyst, Central Computation

Data Analyst, GBD

Data Analyst, Geospatial

Desktop Support Specialist

Director of Research Management

GBD Researcher, Diabetes

GBD Researcher, Neglected Tropical Diseases

MySQL Database Administrator

Project Officer, Geospatial Analysis

Project Officer, Global Burden of Disease

Outline of minitutorial

Overview of Disease Burden Measurement

Deep Dive into Cause-of-Death Estimation

The Challenges to Come

Outline for this half of minitutorial

Verbal autopsy

Machine learning for predicting cause of death

Tools of the trade: python, sklearn, github, binder, software carpentry, …

Example VA response (this data is real)

Deceased was 53 Year Old Male, with:

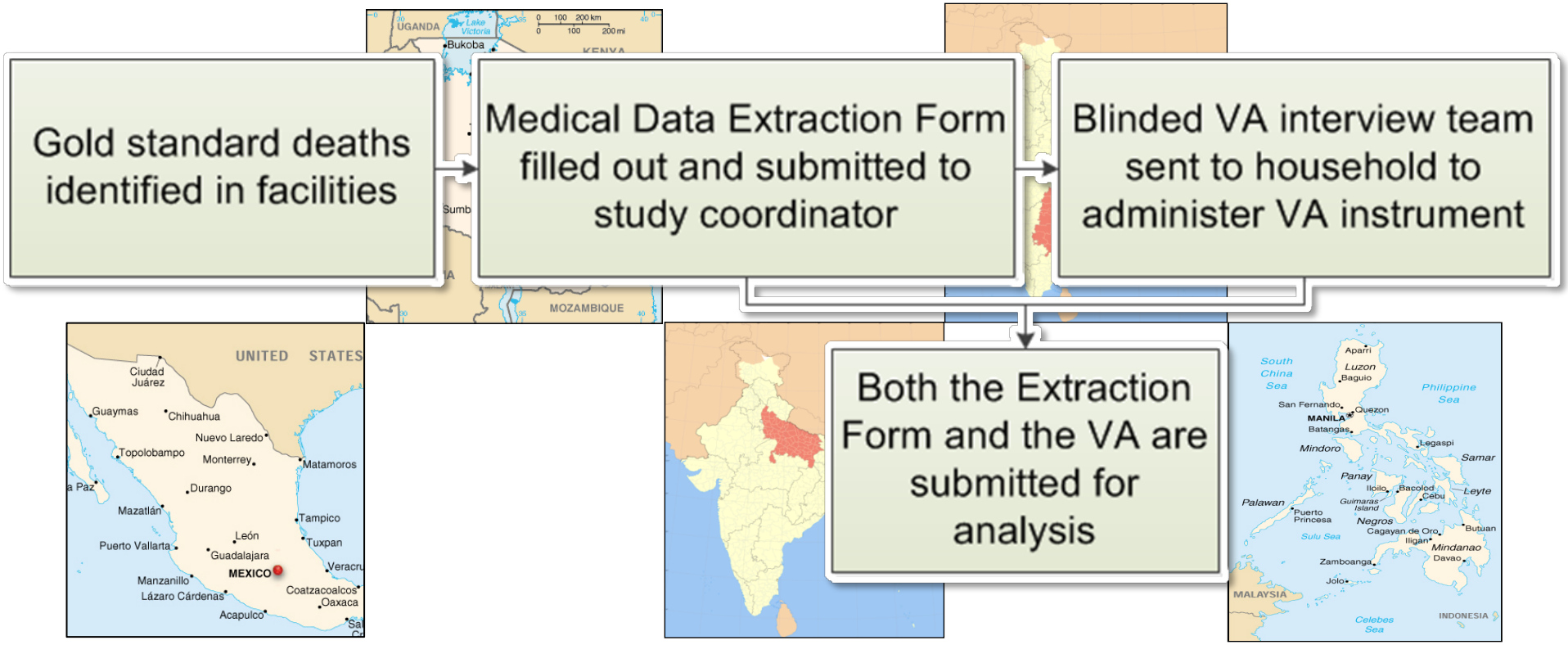
- Asthma
- Heart Disease
- Hypertension
- Ankle Swelling
- Puffiness of the Face, All Over His Body

Underlying Cause: COPD

- Used Tobacco
- Drank Low Amount of Alcohol
- Free Text: Asthma, Breath, Heart, Lung, Swell, Water

PHMRC VA Validation Dataset (GC-13)

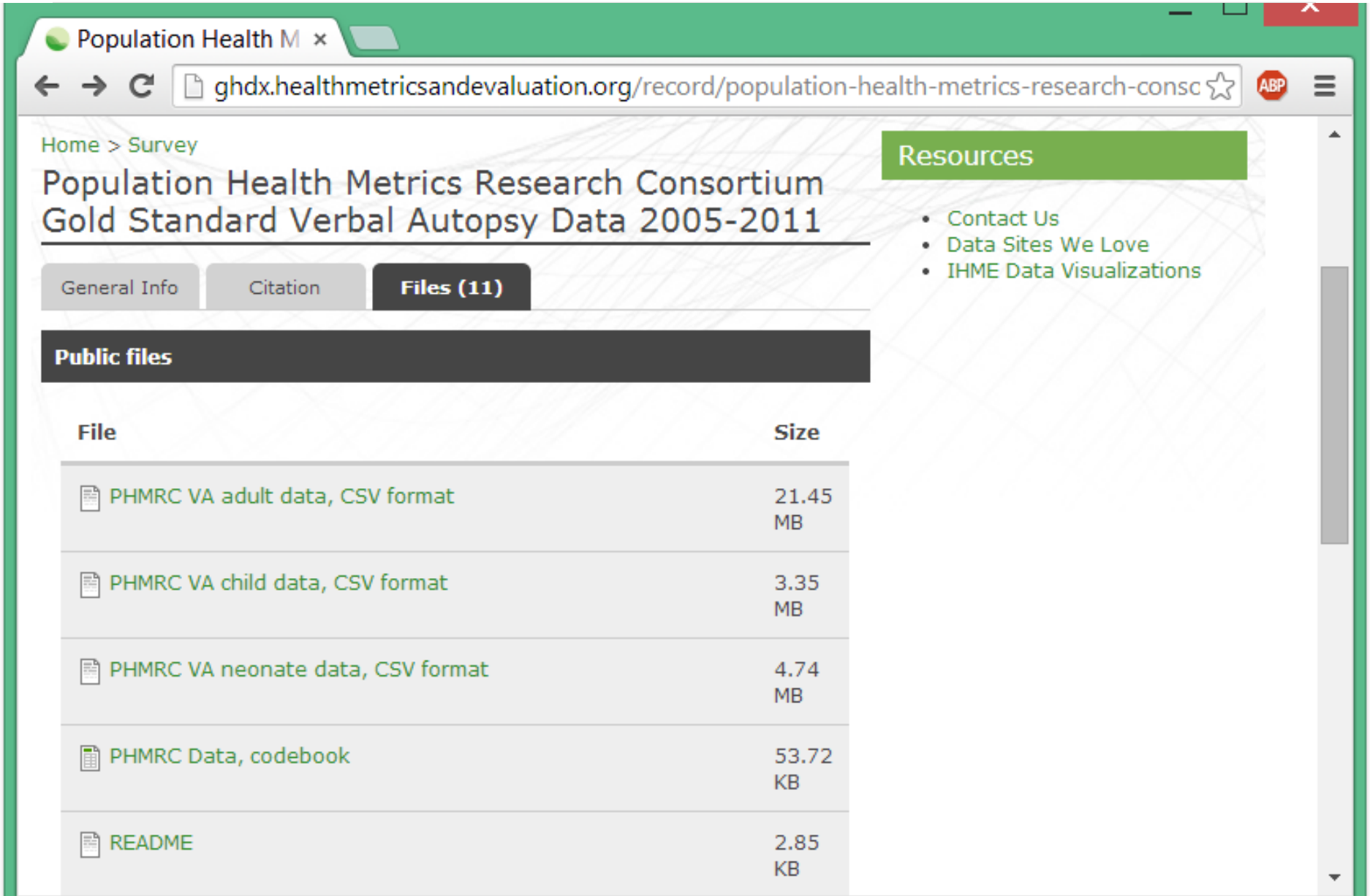
Population Health Metrics Research Consortium (PHMRC) study was part of the Bill & Melinda Gates Foundation Grand Challenges in Global Health (GC-13, to be specific).



Deaths with CoD known and VA collected

Site	Adult		Child		Neonate		Total
	Level 1	Level 2	Level 1	Level 2	Level 1	Level 2	
AP	1,285	269	385	66	376	1	**2,382**
Bohol	998	262	234	30	374	0	**1,898**
Dar	1,556	162	366	106	1,047	2	**3,239**
Mexico	1,373	215	124	4	313	2	**2,031**
Pemba	266	31	156	105	261	3	**822**
UP	1,277	142	412	87	251	1	**2,170**
Total	**6,755**	**1,081**	**1,677**	**398**	**2,622**	**9**	**12,542**

Labeled data from GC-13



Outline for this half of minitutorial

Verbal autopsy

Machine learning for predicting cause of death

Tools of the trade: python, sklearn, github, binder, software carpentry, …

Live Coding (with a net)

Find it on GitHub – t.co/9ARJpW2XVC

Use it on Binder – mybinder.org
(cf. cloud.sagemath.com)

Want to learn more/skill-up students?

HOME ABOUT WORKSHOPS LESSONS GET INVOLVED BLOG CONTACT SEARCH

software carpentry

Teaching basic lab skills for research computing

Our Workshops ›

Find or host a workshop.

Our Lessons ›

Have a look at what we teach.

Get Involved ›

Help us help researchers.

“hello, world” of Scientific Python

Jupyter (IPython) Notebook

Matplotlib

Numpy

Pandas

Scikit-Learn

An aside on REPRODUCIBLE RESEARCH

For more, attend this invited talk:

Victoria Stodden

Implementing reproducibility in computational science

Thursday, 2 PM

Outline for this half of minitutorial

Verbal autopsy

Machine learning for predicting cause of death

Tools of the trade: python, sklearn, github, binder, software carpentry, …

What is “machine learning”?

For my purposes, ML means something very specific:

Although the **framework** for mapping from VA interviews to cause-of-death **is fixed**, the **details** are **learned from data**.

Machine learning methods

RESEARCH Open Access

Performance of the Tariff Method: validation of a

RESEARCH Open Access

Simplified Symptom Pattern Method for verbal

RESEARCH Open Access

Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards

Abraham D Flaxman[1*], Alireza Vahdatpour[1], Sean Green[2], Spencer L James[1] and Christopher JL Murray[1] for the Population Health Metrics Research Consortium (PHMRC)

“Horserace” paper

Murray et al. *BMC Medicine* 2014, **12**:5
http://www.biomedcentral.com/1741-7015/12/5
doi:10.1186/1741-7015-12-5

Applied Machine Learning for SIAM

Let us now do something hands-on again, **make predictions** of cause-of-death from VA interviews, **using sklearn** for machine learning methods

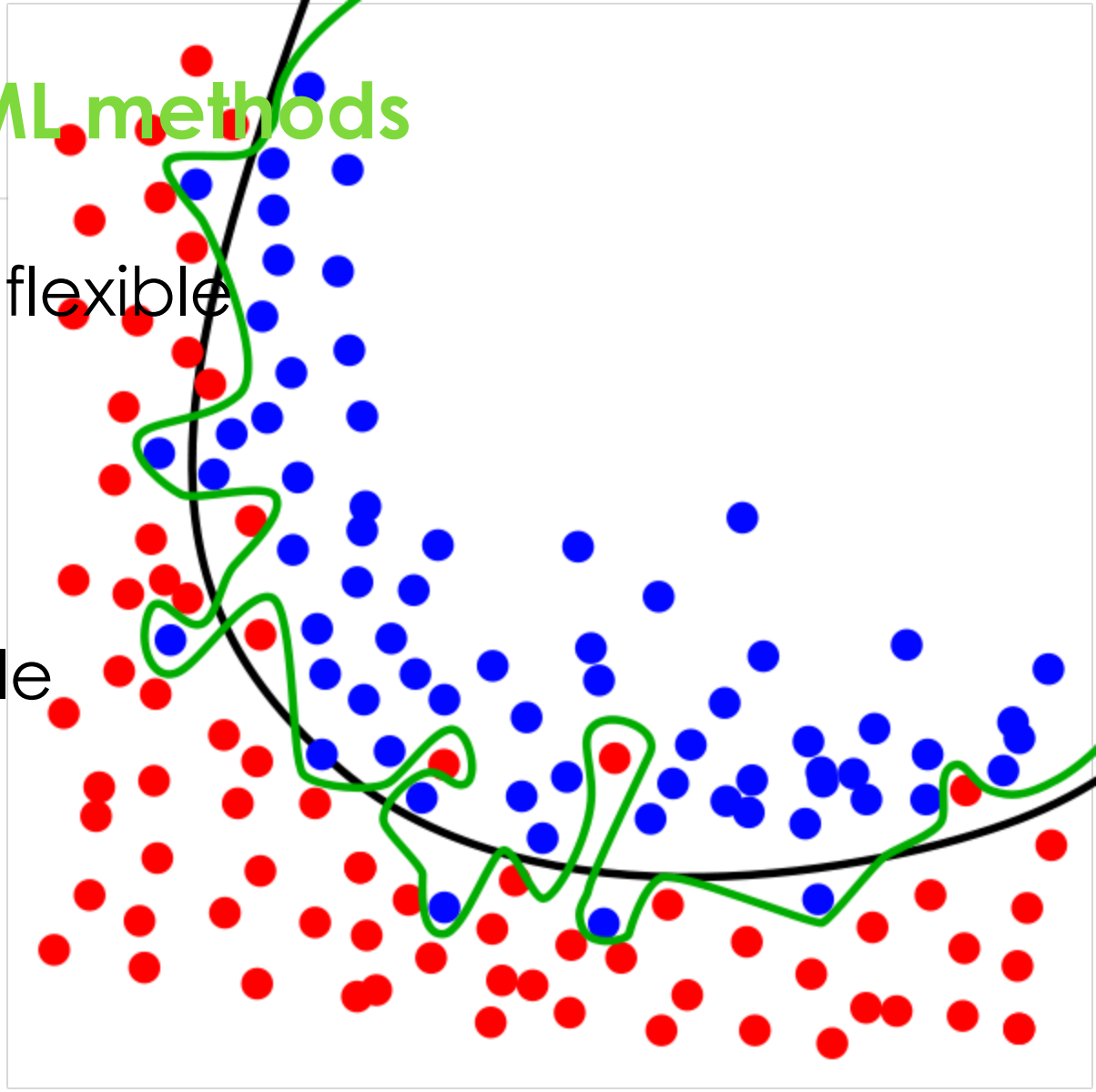
Out-of-Sample Predictive Validity

Evaluation of ML methods

ML methods are flexible

Very flexible

Almost too flexible



Standard Approach 10x 10-fold C-V

Example in Binder

Very cool recent alternative: **reusable hold-out** approach developed by differential privacy crowd, perhaps mentioned in Cythia Dwork's invited talk this morning.

Question remains: what to optimize?

Quantity we have been calculating is "accuracy", which is not really what I am interested in.

Back to active learning

On what metric of prediction quality *should* I be focused?

Write, pair, share (time permitting).

If you have something good, tweet it to me: @healthyalgo #SIAMAN16 (or email abie@uw.edu)

Two metrics for prediction quality

Individual-level performance:

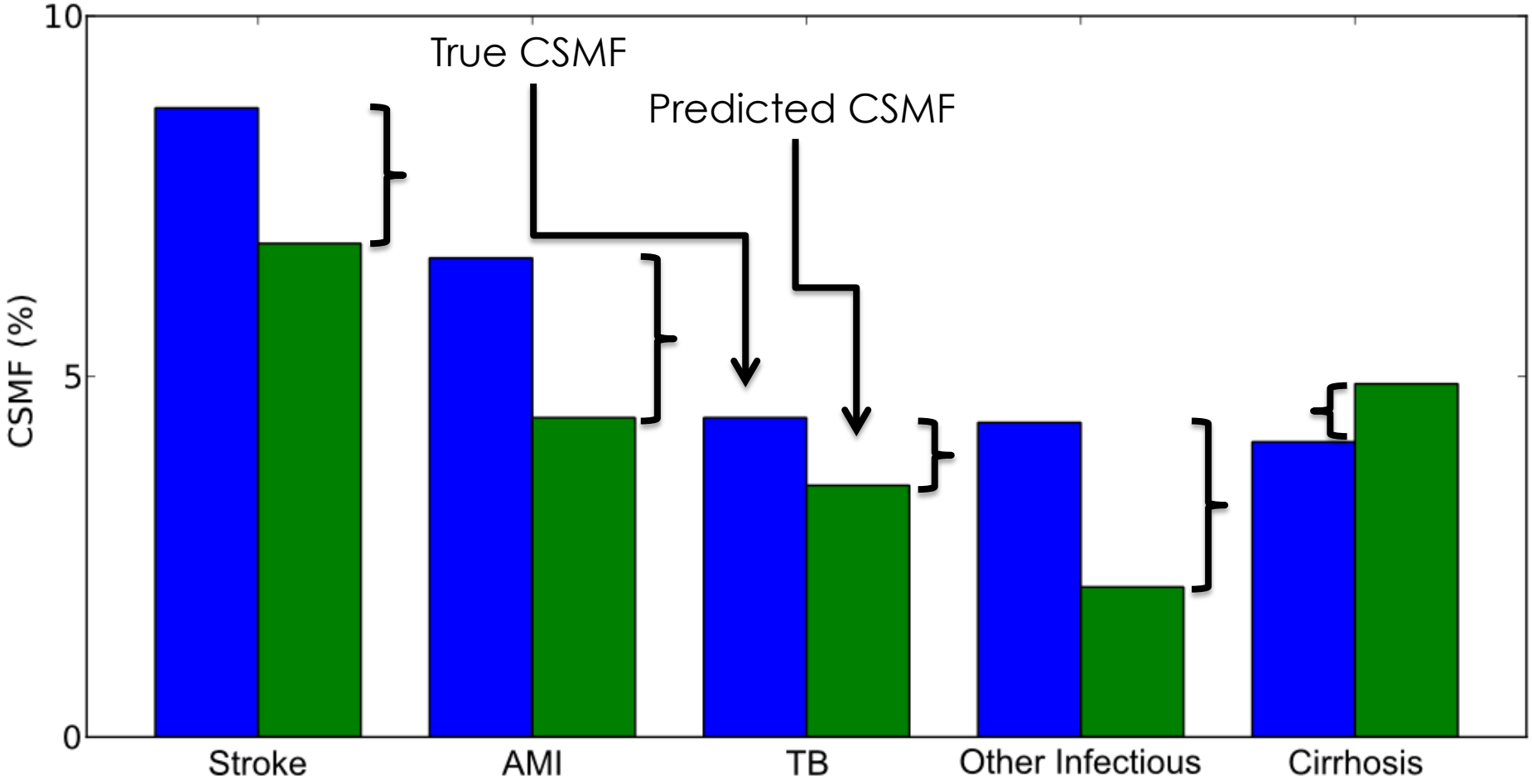
- Chance-corrected concordance (CCC)

Population-level performance:

- Cause-specific mortality fraction (CSMF) accuracy

Population-level quality

CSMF Accuracy, $1 - \frac{\sum_j \left| \text{CSMF}_j^{\text{true}} - \text{CSMF}_j^{\text{pred}} \right|}{\text{CSMF Max Error}}$



More live coding

It's too much to really implement it.

Let's just spend a minute to try to make it work, with a pretty secure scaffold.

[If time permits, followed by a word on **Test-driven Development** (TDD)]

Out-of-sample validation

Really being out-of-sample is tricky for CSMF Accuracy

Unusual part here

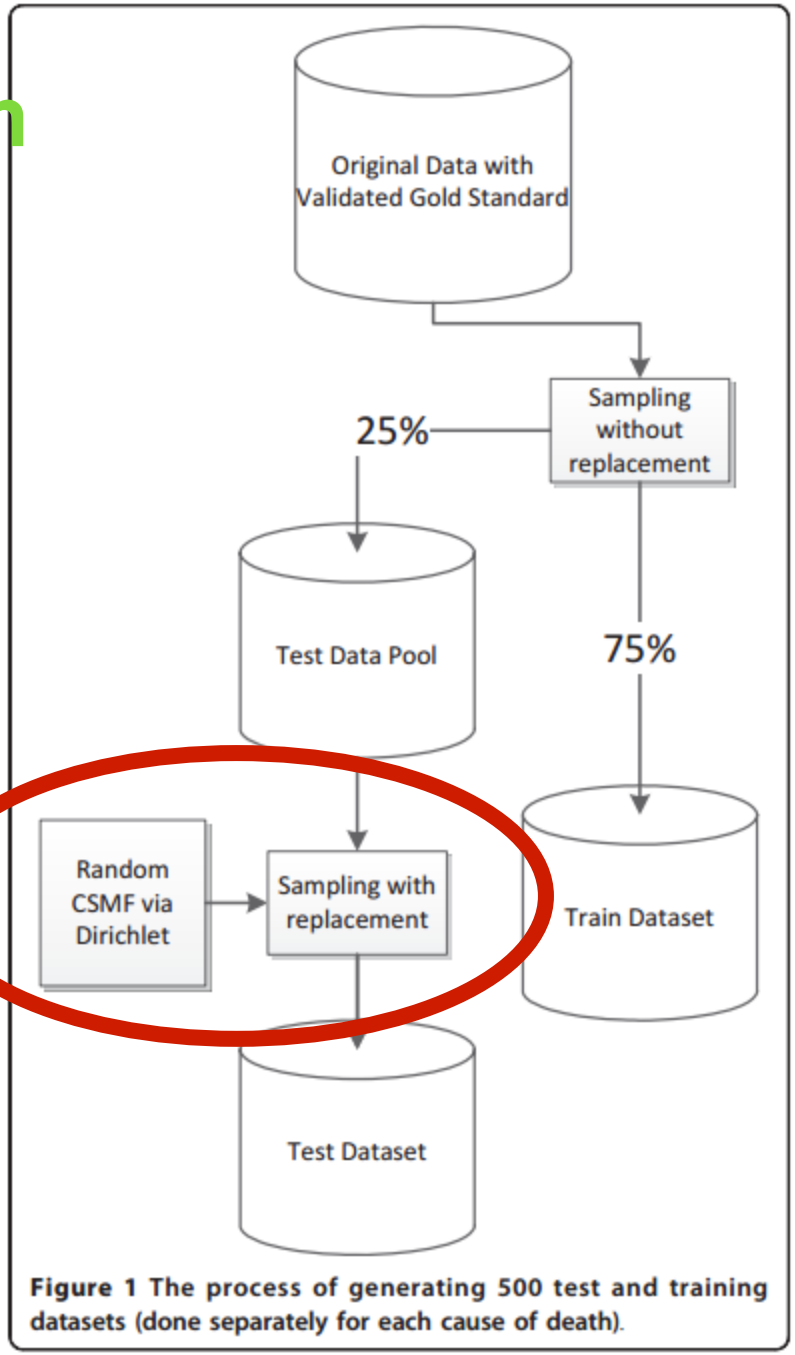
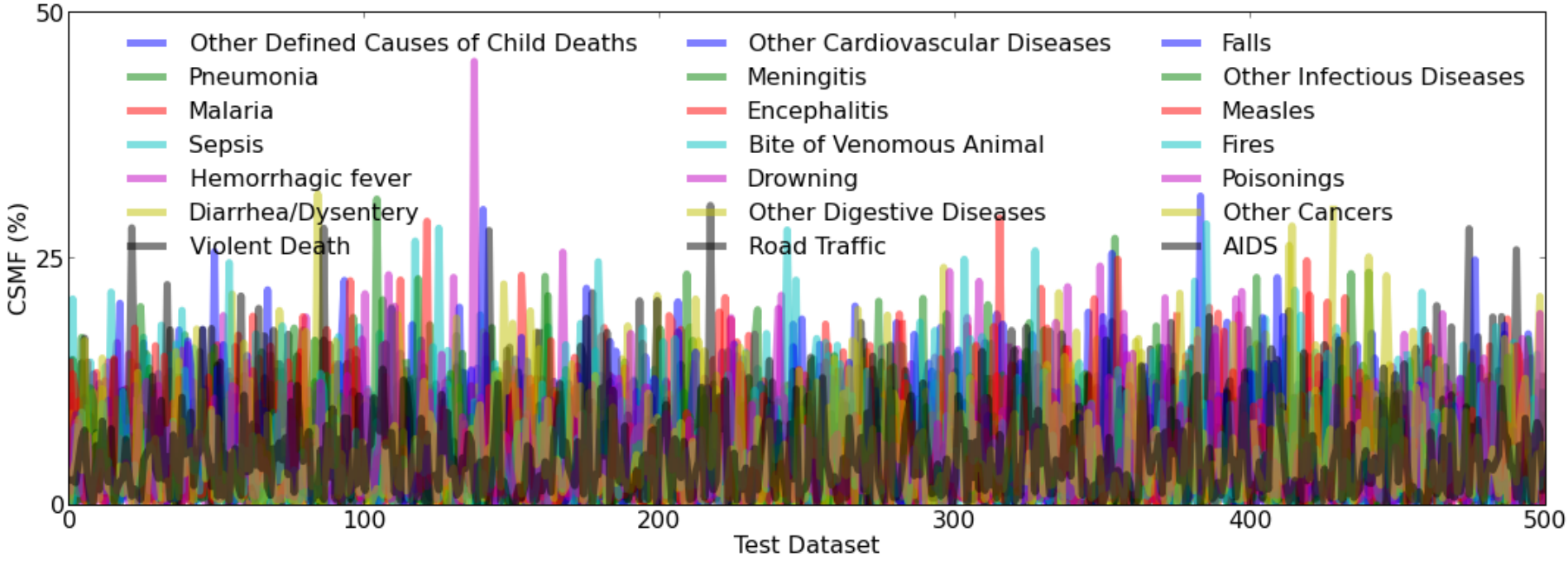


Figure 1 The process of generating 500 test and training datasets (done separately for each cause of death).

Out-of-sample validation

Resample test set to have random CSMFs



Chance-corrected CSMF Accuracy

Flaxman *et al. Population Health Metrics* (2015) 13:28
DOI 10.1186/s12963-015-0061-1

POPULATION HEALTH METRICS

RESEARCH Open Access

Measuring causes of death in populations: a new metric that corrects cause-specific mortality fractions for chance

Abraham D. Flaxman[1*], Peter T. Serina[1], Bernardo Hernandez[1], Christopher J. L. Murray[1], Ian Riley[2] and Alan D. Lopez[3]

Outline for this half of minitutorial

Verbal autopsy

Machine learning for predicting cause of death

Tools of the trade: python, sklearn, github, binder, software carpentry, …

Outline of Minitutorial

Overview of Disease Burden Measurement

Deep Dive into Cause-of-Death Estimation

The Challenges to Come

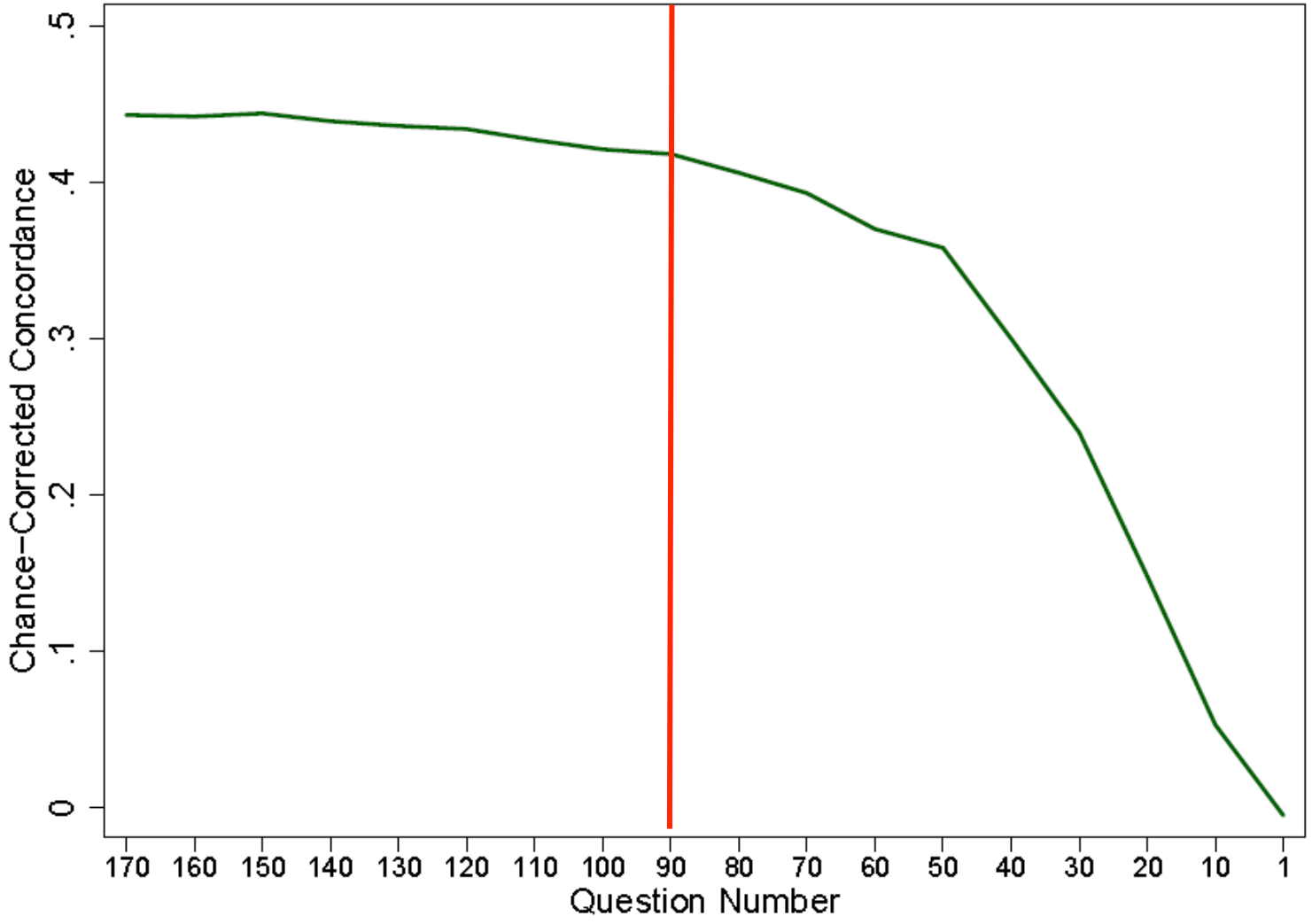
- Explaining why
- Item reduction
- Quality Assurance for Translation

Explaining why

Important for understanding errors, building trust

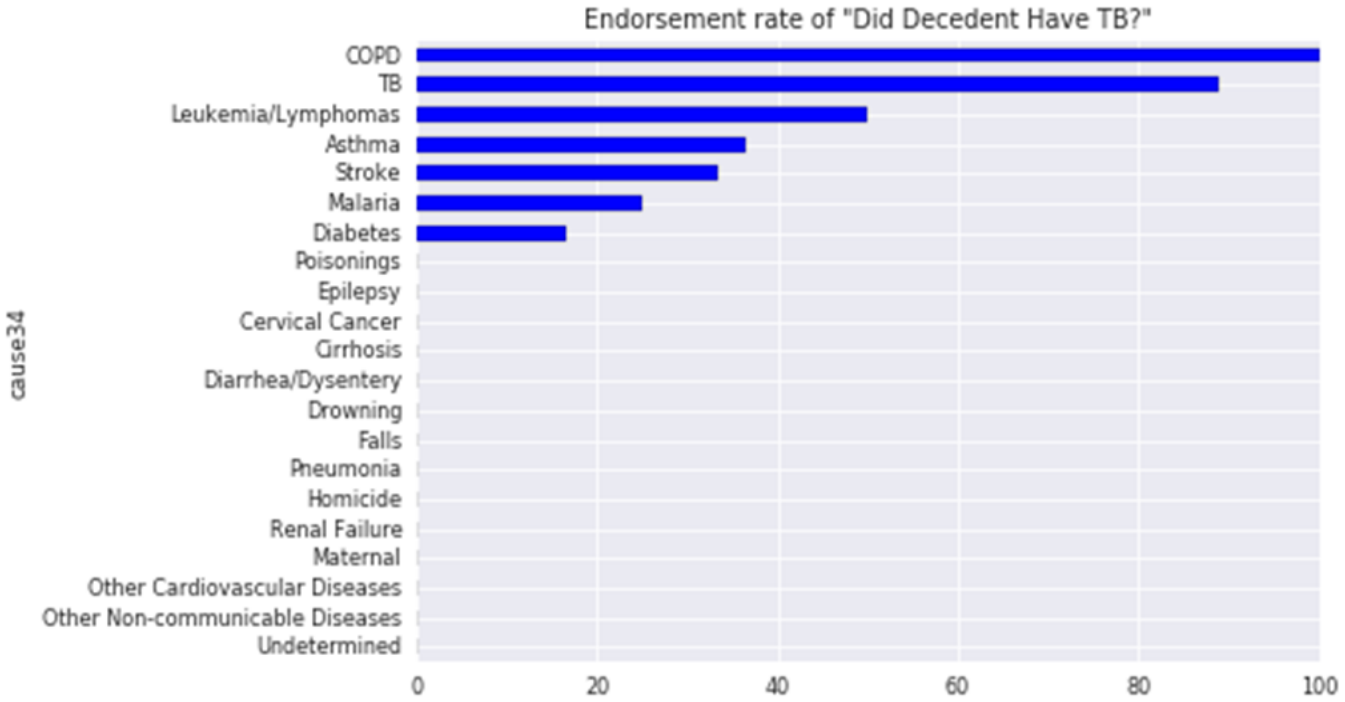
	B	C	D	E	F	G
1	cause_of_deatl	symptom_	symptom_str	tariff	endorsement_rate	n_subj
122	Stroke	s110	Paralyzed upper part of body	25.5	2.5%	630
262	Stroke	s107	Paralyzed on one side (arm and leg)	8.5	37.3%	630
292	Stroke	s105	Was [name] in any way paralyzed?	7.5	51.1%	630
326	Stroke	s8	Did Decedent Have Epilepsy?	6.5	15.4%	630
381	Stroke	s12	Did Decedent Have Stroke?	5.5	25.9%	630
383	Stroke	s116	Paralyzed other	5.5	5.7%	630
567	Stroke	s95	Sudden loss of consciousness	3.5	50.8%	630
568	Stroke	s106	For how long before death did [name] have paralysis? [days]	3.5	16.3%	630
570	Stroke	s152	Decedent suffered fall	3.5	8.9%	630
571	Stroke	s112	Paralyzed one arm only	3.5	1.9%	630
656	Stroke	s97	Did it continue until death?	3	60.3%	630
657	Stroke	s113	Paralyzed whole body	3	5.4%	630
767	Stroke	s140	Type of tobacco used: pipe	2.5	3.8%	630
768	Stroke	s94	Did [name] experience a period of loss of consciousness?	2.5	66.7%	630
772	Stroke	s10	Did Decedent Have Hypertension?	2.5	68.6%	630
			For how long did the period of loss of			

Data-driven Item Reduction



Quality Assurance for VA Translation

Figure to follow up on possibility of mistaking copd and asthma for tb:



Quite possible that majority of predicted COPD predicted Asthma deaths were actually caused by TB.

Conclusion of Minitutorial

Overview of Disease Burden Measurement

Deep Dive into Cause-of-Death Estimation

The Challenges to Come

- Explaining why
- Item reduction
- Quality Assurance for Translation

Thank You!