Reducing clinical noise for body mass index measures due to unit and transcription errors in the electronic health record

INSTITUTE FOR COMPUTATIONAL BIOLOGY

March 28, 2017

Dana C. Crawford, PhD
Associate Professor
Epidemiology and Biostatistics
Institute for Computational Biology

Body Mass Index, the Variable

Calculated using height and weight:

$$\text{weight} / (\text{height})^2$$

$$\text{kg/m}^2$$

Body Mass Index, the variable

BMI is important

Height and weight measured* at most clinic visits

Stadiometers

Scales

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Body Mass Index and EHR

There are problems with EHR-calculated BMI

- Units (lb v kg; in or ft v m)
- Just plain wrong numbers
- Age disassociated from weight
- Weight but missing heights
- Etc.

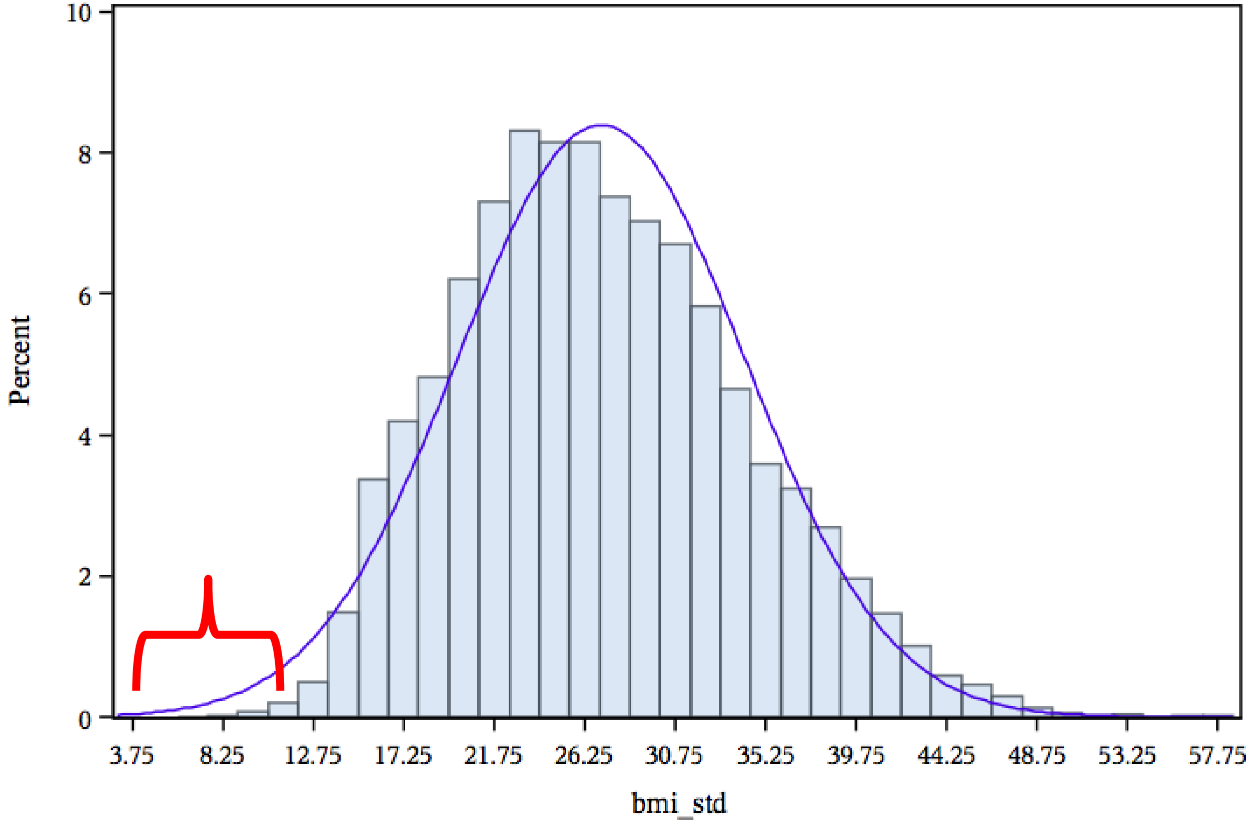
INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Body Mass Index and EHR

- In clinical practice, records and values are examined individually
  - Errors are easy to spot and ignore
  - Aggregate data (on a single subject) is synthesized by experts in the context of a patient encounter
- In research, records and values are examined in aggregate over many thousands of individuals
  - Errors are not so easy to identify
  - Data points are not considered in the context of the patient

Body Mass Index and EHR

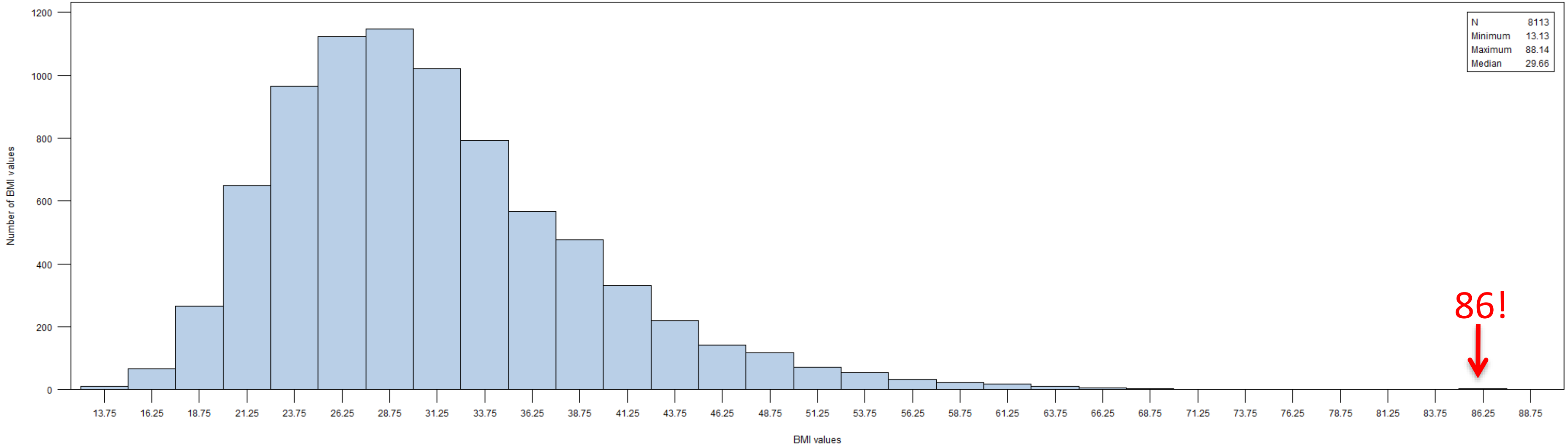
Implausible low BMIs



INSTITUTE FOR COMPUTATIONAL BIOLOGY

Body Mass Index and EHR

Questionable high BMIs

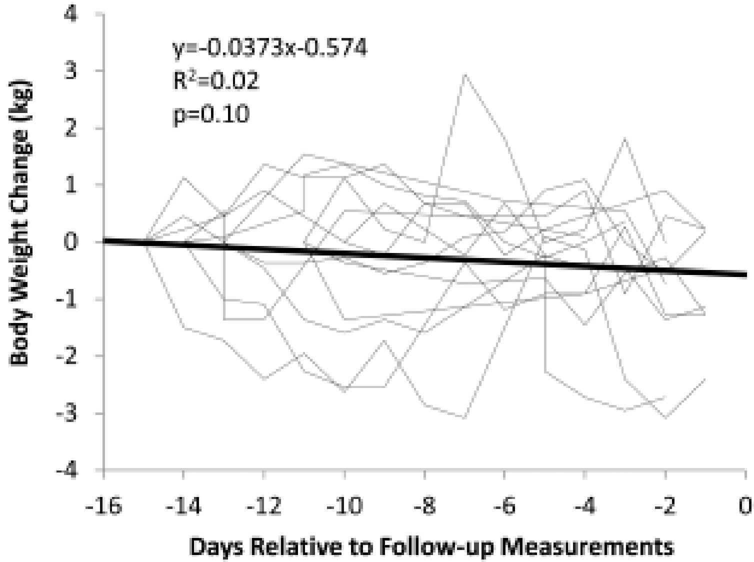


Body Mass Index and EHR

Can we use multiple measures per patient to identify these errors?

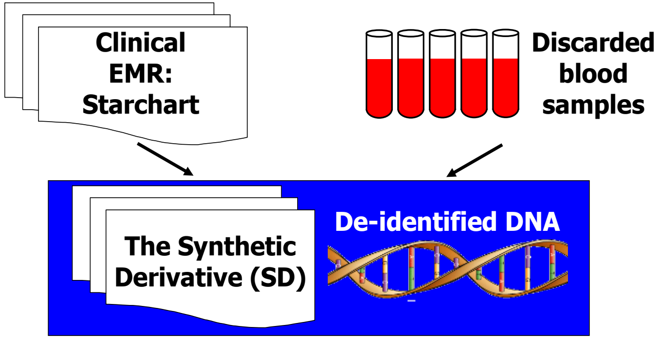
Biggest Loser contestants

Fothergill et al (2016) *Obesity* 24(8):1612-1619



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

VUMC BioVU



- Opt-out model (2007-2015)
- DNA collected from discarded blood after routine clinical testing has been completed
  - Matched with clinical and demographic data within de-identified EHR ("Synthetic Derivative" database)
  - >225,000 DNA samples

VUMC EHR

VUMC in Nashville, TN

- StarChart
  - Designed, built, and maintained by faculty-led teams
  - Being replaced with EPIC (2018)
- >2 million records, including order entry data on inpatients since 1994
- A document-centric architecture
  - Structured (ex. ICD-9-CM and ICD-10-CM codes)
  - Unstructured (ex. clinical notes)

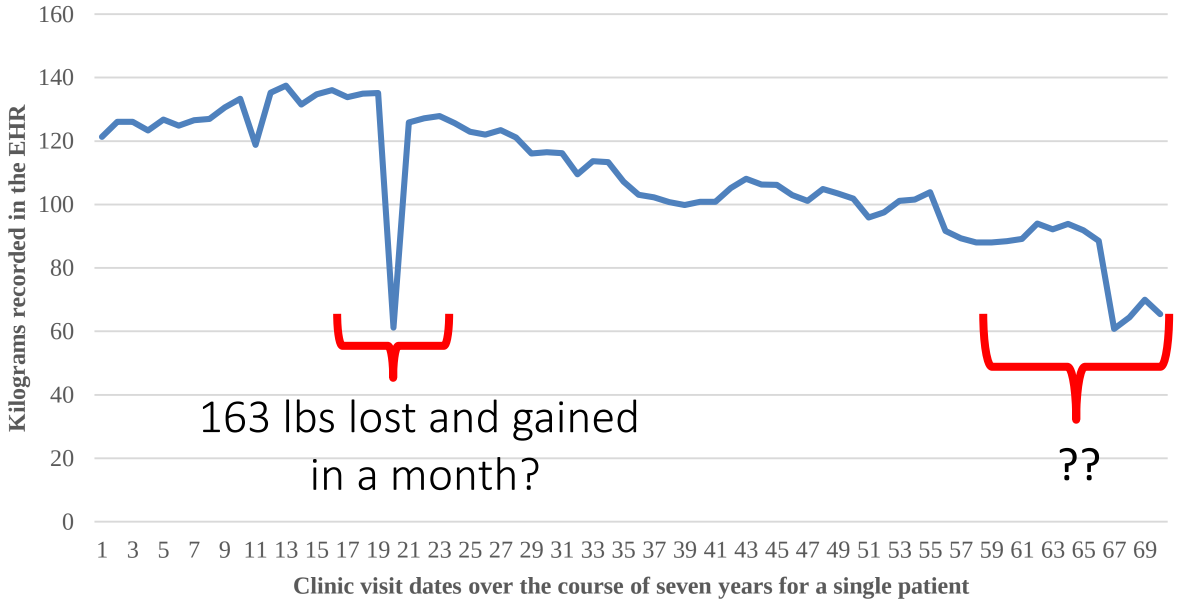
EAGLE BioVU

	n=15,863
Female (%)	10,050 (63.35)
Median (SD) age	37 (20.46)
African American (%)	11,521 (73.06)
Hispanic (%)	1,714 (10.87)
Asian (%)	1,122 (7.12)
Other (%)	1,412 (8.95)

EAGLE BioVU case study

Single patient

70 clinic visit dates



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Adjacency-based Longitudinal Outlier Extraction (ALOE)

- Step 1: Initial Outlier Detection and Characterization

Calculate and plot raw BMI distributions

Stratify obese from non-obese using ICD-9-CM or ICD-10-CM codes

Extreme outliers in non-obese individuals can be manually inspected

Adjacency-based Longitudinal Outlier Extraction (ALOE)

- Step 2: Temporal Partitioning

Divide the smallest two weights of the first three values by the fourth weight

121.28 kg

126.10 kg

123.23 kg

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Adjacency-based Longitudinal Outlier Extraction (ALOE)

- Step 2: Temporal Partitioning

Establish the “weight index value” closest to 1

$$\boxed{121.28\text{ kg} / 123.23\text{ kg} = 0.984}$$

$$126.10\text{ kg} / 123.23\text{ kg} = 1.023$$

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Adjacency-based Longitudinal Outlier Extraction (ALOE)

- Step 3: Unit mismatch identification

Divide all measurements with “weight index value”

Generate change ratio distribution

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Adjacency-based Longitudinal Outlier Extraction (ALOE)

- Step 3: Unit mismatch identification

For weight, if observed value is

within 0.20 SD, value is in kilograms

within 0.45 SD, value is in pounds

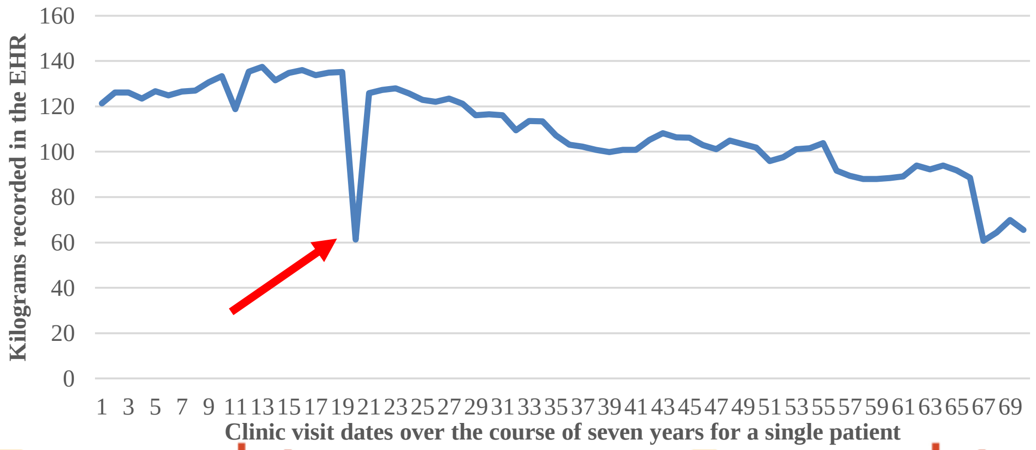
within .10 of 0.45, value is in kilograms but assumed pounds and converted to kilograms (kgx2)

Adjacency-based Longitudinal Outlier Extraction (ALOE)

- Step 3: Unit mismatch identification

61.23 / 121.28 = 0.50

Kg thought to be in lbs and converted to kg (kgx2)



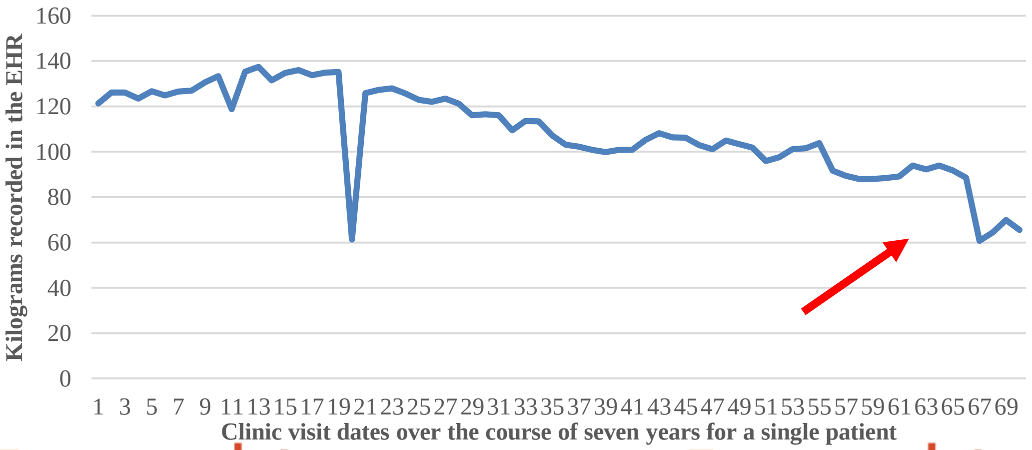
INSTITUTE FOR COMPUTATIONAL BIOLOGY

Adjacency-based Longitudinal Outlier Extraction (ALOE)

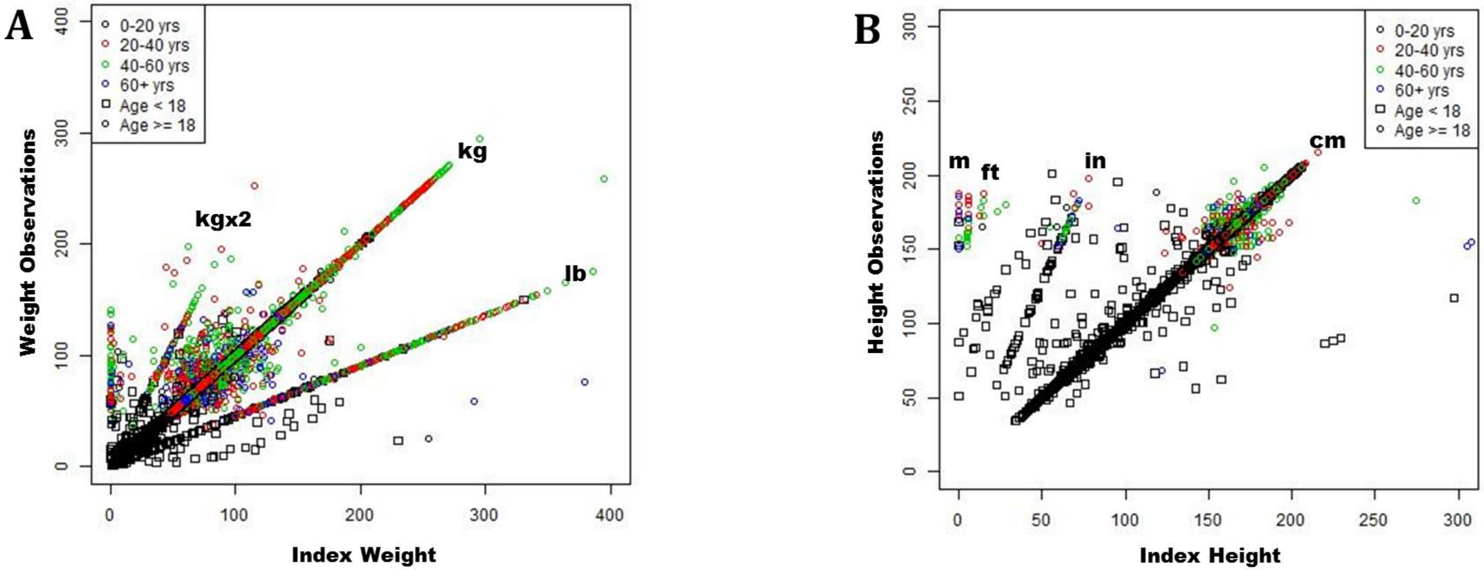
- Step 3: Unit mismatch identification

60.78 / 121.28 = 0.50
64.41 / 121.28 = 0.53
69.94/ 121.28 = 0.58
65.49 / 121.28 = 0.54

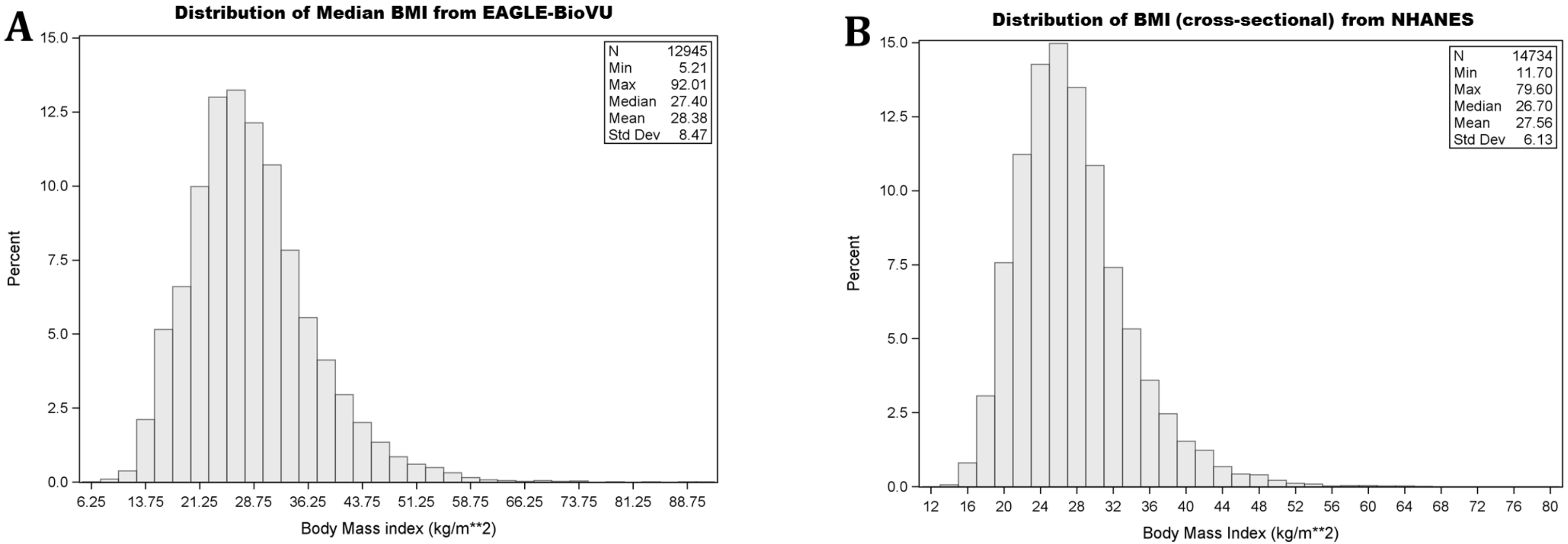
Kg thought to be in lbs and converted to kg (kgx2)



ALOE applied to EAGLE BioVU



Comparison of EAGLE BioVU with NHANES



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Residual Modeling

- Exploits relationship between height, weight, and age
  - Regress age onto height and weight, respectively
  - Calculate deviation from predicted value (Cook's Distance, Leverage, DFfits, Studentized residuals, Covariance Ratio)
  - If modeled data has three positive tests, data set to missing
- Model executed two different ways
  - A single model over all observations for an individual
  - Multiple models over all observations iteratively

ALOE versus Residual Modeling

- ALOE retains more data

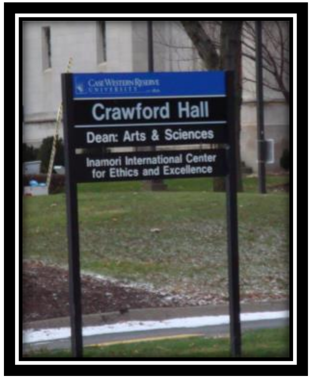
	RM (all)	RM (individual)	ALOE	Raw Data Total
Weight	155,781 (66%)	226,685 (96%)	230,701 (98%)	235,624
Height	57,707 (51%)	106,424 (94%)	111,536 (99%)	112,862

Conclusions and Directions

- ALOE cleaned BMI data and retained data points
- But relies on dense temporal data with multiple measures
- Manual corrections still necessary and are discretionary

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Crawford Lab and Collaborators



Robert Goodloe, MS

Williams S. Bush, PhD, MS

Dana C. Crawford, PhD
Brittany Hollister, PhD candidate

NIH/NHGRI HG004798 (EAGLE)
NCATS 2 UL1 TR000445-06 (BioVU)

Eric Farber-Eger Jonathan Boston

CWRU and Cleveland

Biomedical Data Science open rank position available!

http://epbiwww.case.edu/

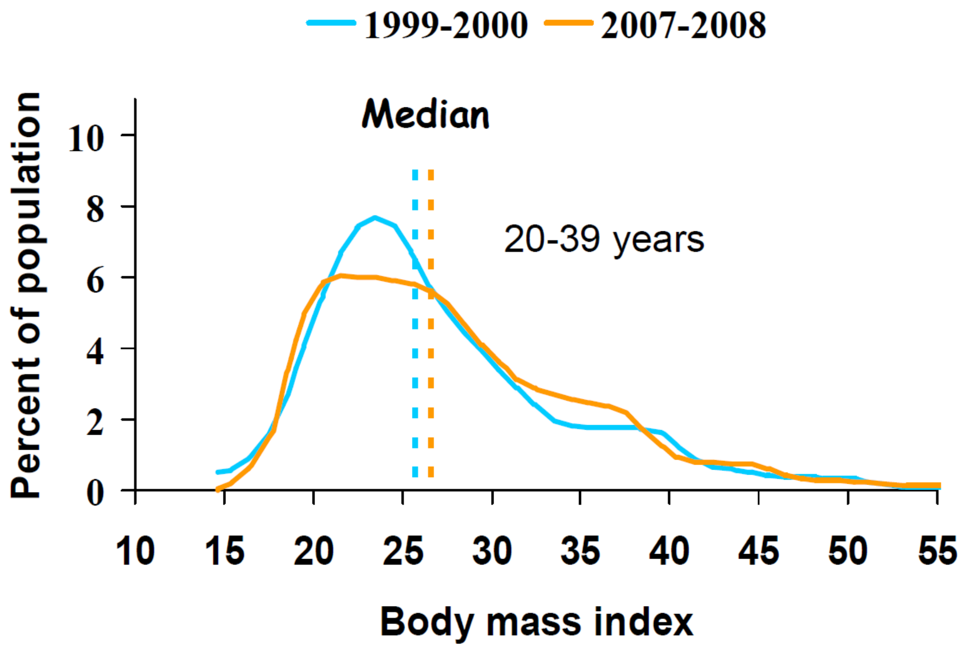
INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Body Mass Index, the Variable

BMI is variable in human populations

US women from NHANES

Flegal et al (2010) *JAMA* 303(3):235-241

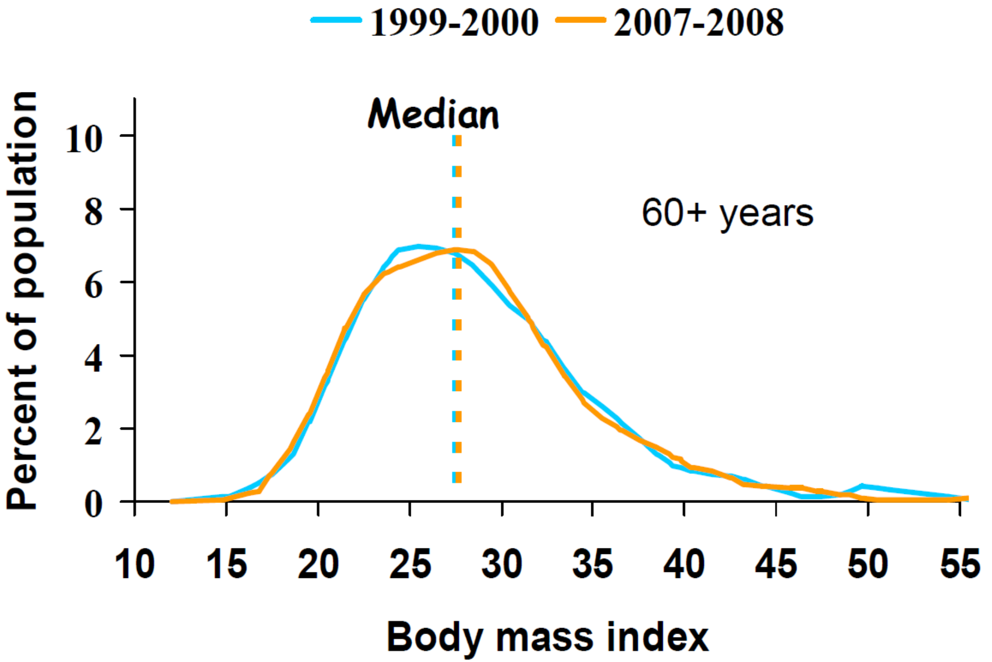


Body Mass Index, the variable

BMI changes over the lifecourse

US women from NHANES

Flegal et al (2010) *JAMA* 303(3):235-241



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Body Mass Index, the variable

BMI is an important health variable

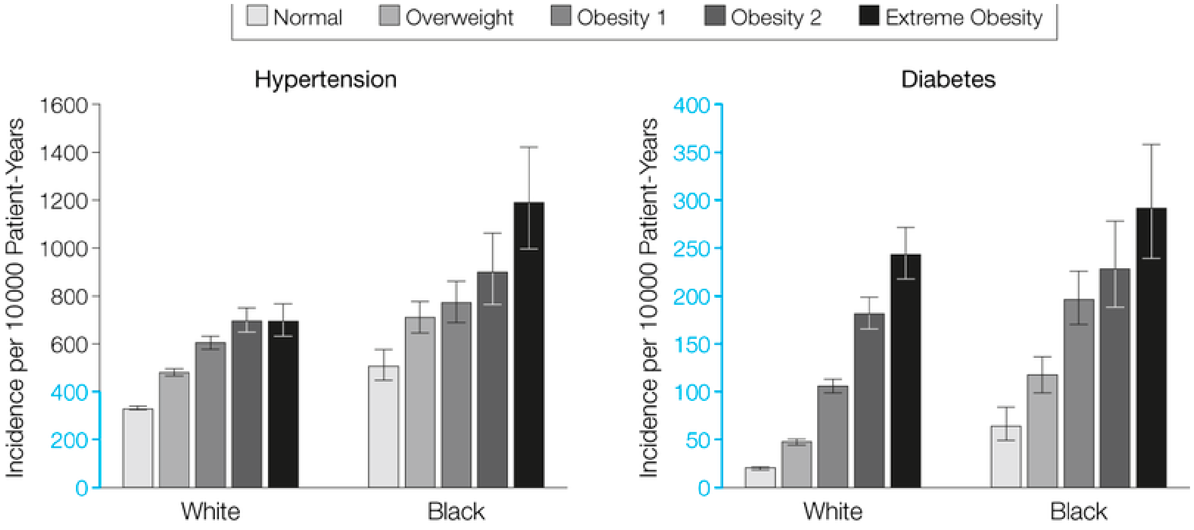
BMI	WHO classification
<18.5	underweight
18.5-24.9	normal weight
25.0-29.9	overweight
30.0-34.9	class I obesity
35.0-39.9	class II obesity
≥40.0	class III obesity

Body Mass Index, the variable

BMI is associated with health outcomes

WHI Observational Study

McTigue et al (2006) *JAMA* 296(1):79-86



Error bars represent 95% confidence intervals.

INSTITUTE FOR COMPUTATIONAL BIOLOGY

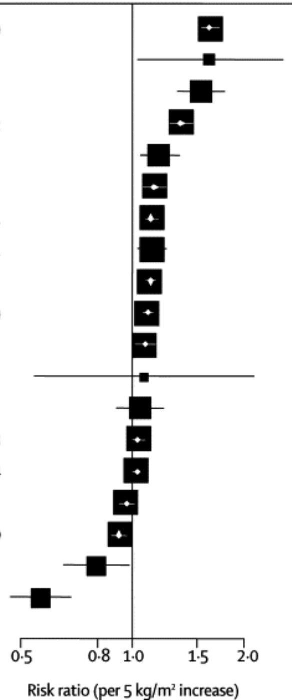
Body Mass Index, the variable

BMI is associated with health outcomes

Meta-analysis of observational studies

Renehan et al (2008) *Lancet* 371(9612):16-22

Cancer site and type	Number of studies	RR (95% CI)	p	I^2
Endometrium	19	1·59 (1·50–1·68)	<0·0001	77%
Gallbladder	2	1·59 (1·02–2·47)	0·04	67%
Oesophageal adenocarcinoma	3	1·51 (1·31–1·74)	<0·0001	0%
Renal	12	1·34 (1·25–1·43)	<0·0001	45%
Leukaemia	7	1·17 (1·04–1·32)	0·01	80%
Thyroid	3	1·14 (1·06–1·23)	0·001	5%
Postmenopausal breast	31	1·12 (1·08–1·16)	<0·0001	64%
Pancreas	11	1·12 (1·02–1·22)	0·01	43%
Multiple myeloma	6	1·11 (1·07–1·15)	<0·0001	0%
Colon	19	1·09 (1·05–1·13)	<0·0001	39%
Non-Hodgkin lymphoma	7	1·07 (1·00–1·14)	0·05	47%
Liver	1	1·07 (0·55–2·08)		
Gastric	5	1·04 (0·90–1·20)	0·56	4%
Ovarian	13	1·03 (0·99–1·08)	0·30	55%
Rectum	14	1·02 (1·00–1·05)	0·26	0%
Malignant melanoma	5	0·96 (0·92–1·01)	0·05	0%
Premenopausal breast	20	0·92 (0·88–0·97)	0·001	39%
Lung	6	0·80 (0·66–0·97)	0·03	84%
Oesophageal squamous	2	0·57 (0·47–0·69)	<0·00 01	60%



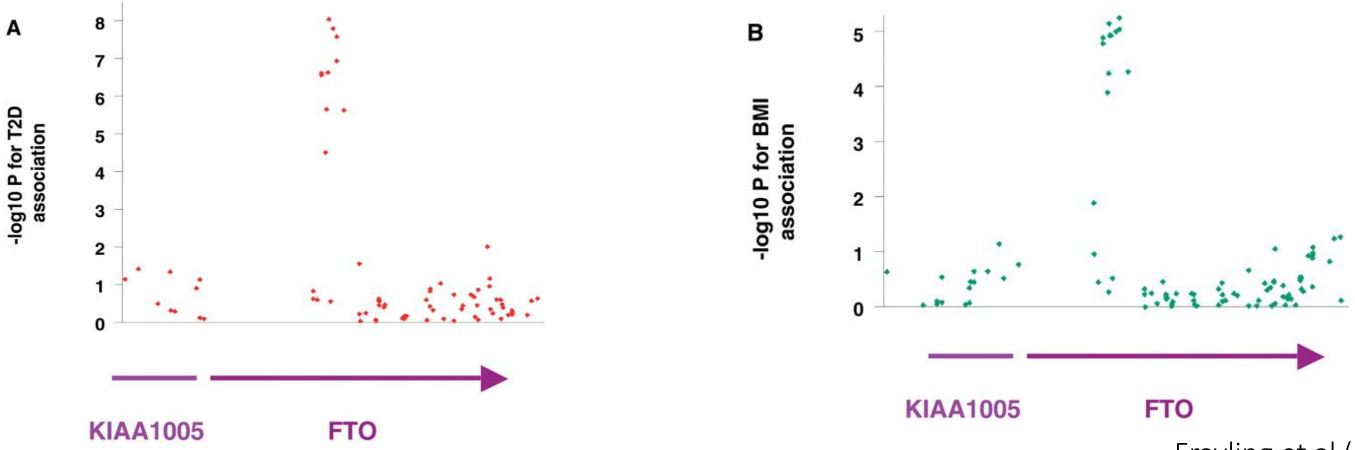
INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Body Mass Index, the Variable

BMI is a known mediator

FTO associated with T2D

FTO associated with BMI



Frayling et al (2007) *Science* 316(5826):89-94

VUMC BioVU is clinic-based

	Davidson County (n=626,684)	BioVU (n=162,716)
% female	51.55	51.93
% adults 18-64 years	68.06	57.66
% adults ≥65 years	10.23	24.83
% European American	60.48	81.07
% African American	28.43	8.65
% Hispanic	10.04	1.32
% Asian	3.10	0.83

Over-represents European-descent and elderly

Crawford et al (2015) *Hum Hered* 79(3-4):137-46

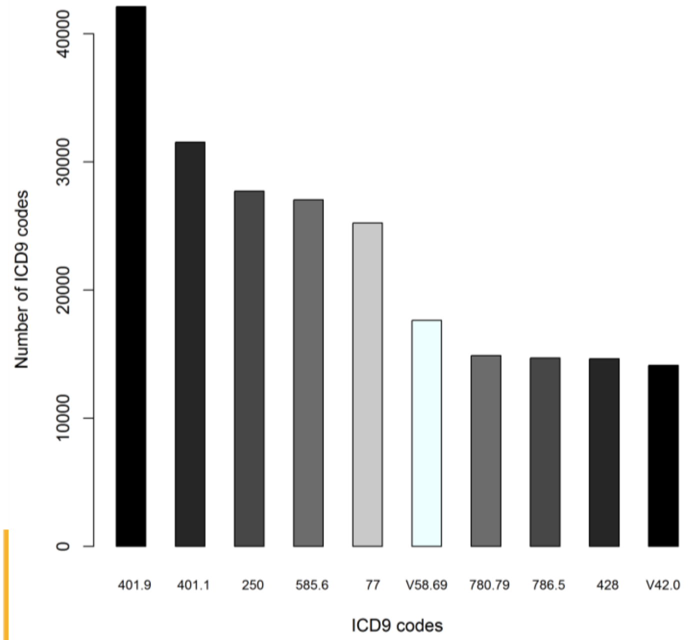
EAGLE BioVU Common Codes

Top 10 codes for African American adults

Hypertension (401.9, 401.1)

Diabetes Mellitus (250)

End-stage renal disease 9585.6)



Crawford et al (2015) *Hum Hered* 79(3-4):137-46

INSTITUTE FOR COMPUTATIONAL BIOLOGY

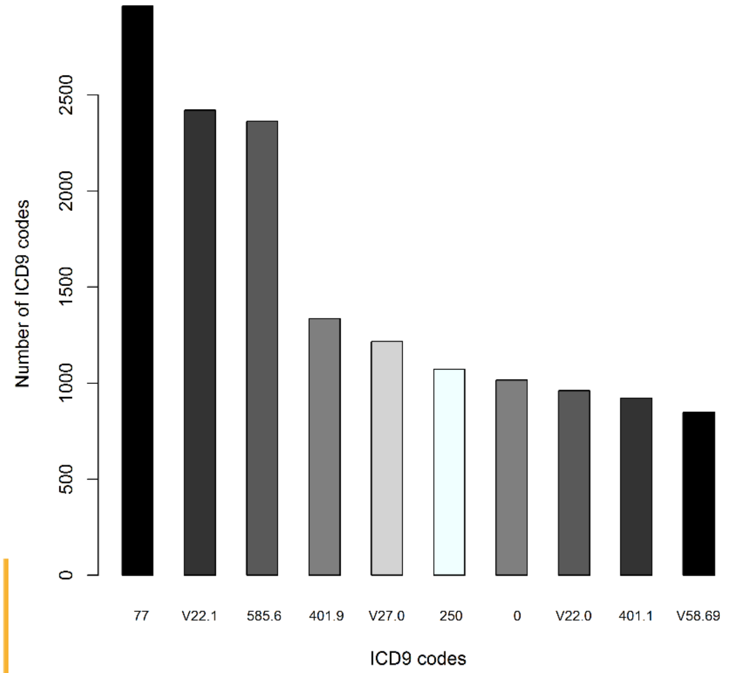
EAGLE BioVU Common Codes

Top 10 codes for Mexican American adults

Sequestrectomy (77)

Supervision of other normal pregnancy (v22.1)

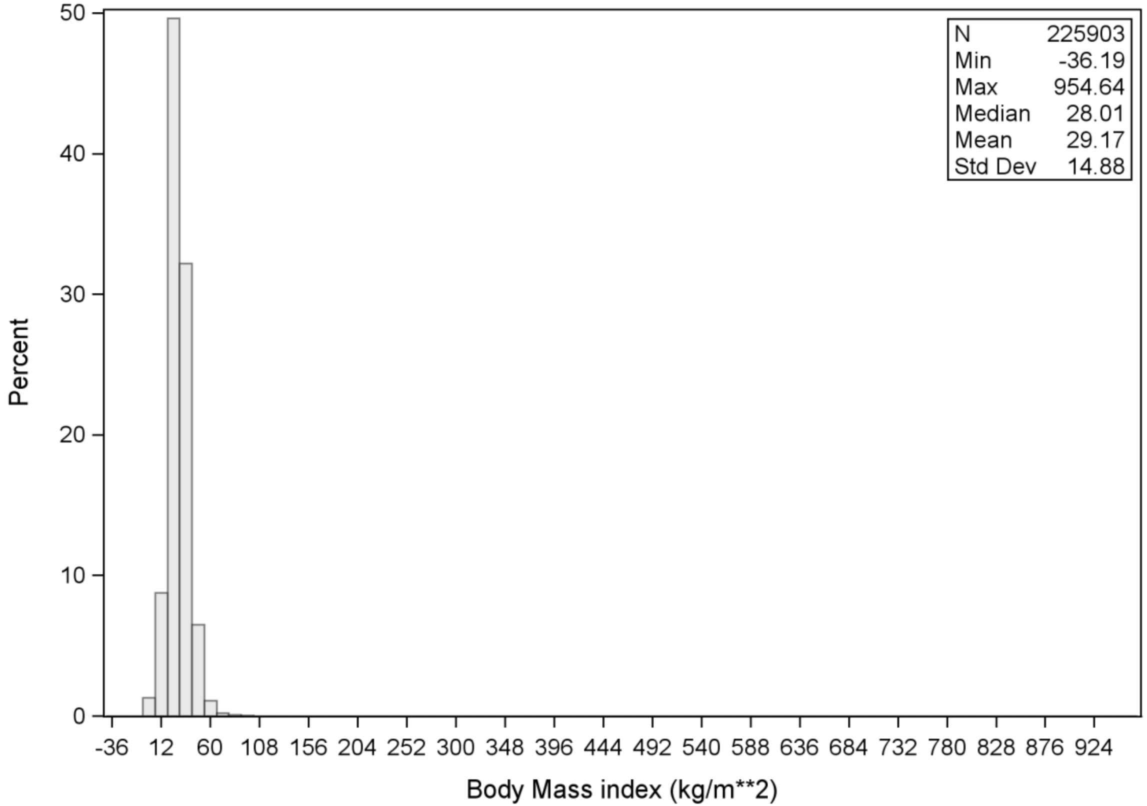
End-stage renal disease (585.6)



Crawford et al (2015) *Hum Hered* 79(3-4):137-46

INSTITUTE FOR COMPUTATIONAL BIOLOGY

Raw EAGLE BioVU BMI



INSTITUTE FOR
COMPUTATIONAL BIOLOGY