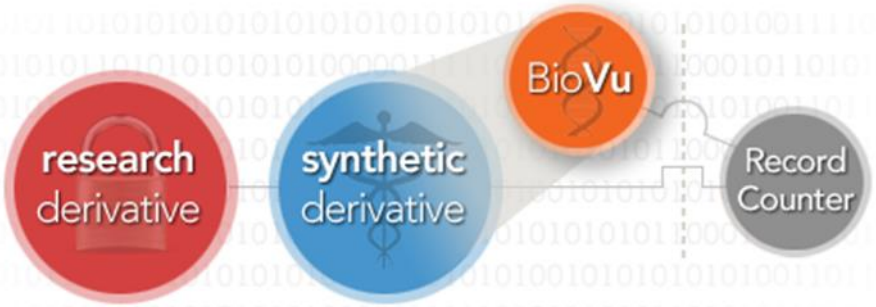
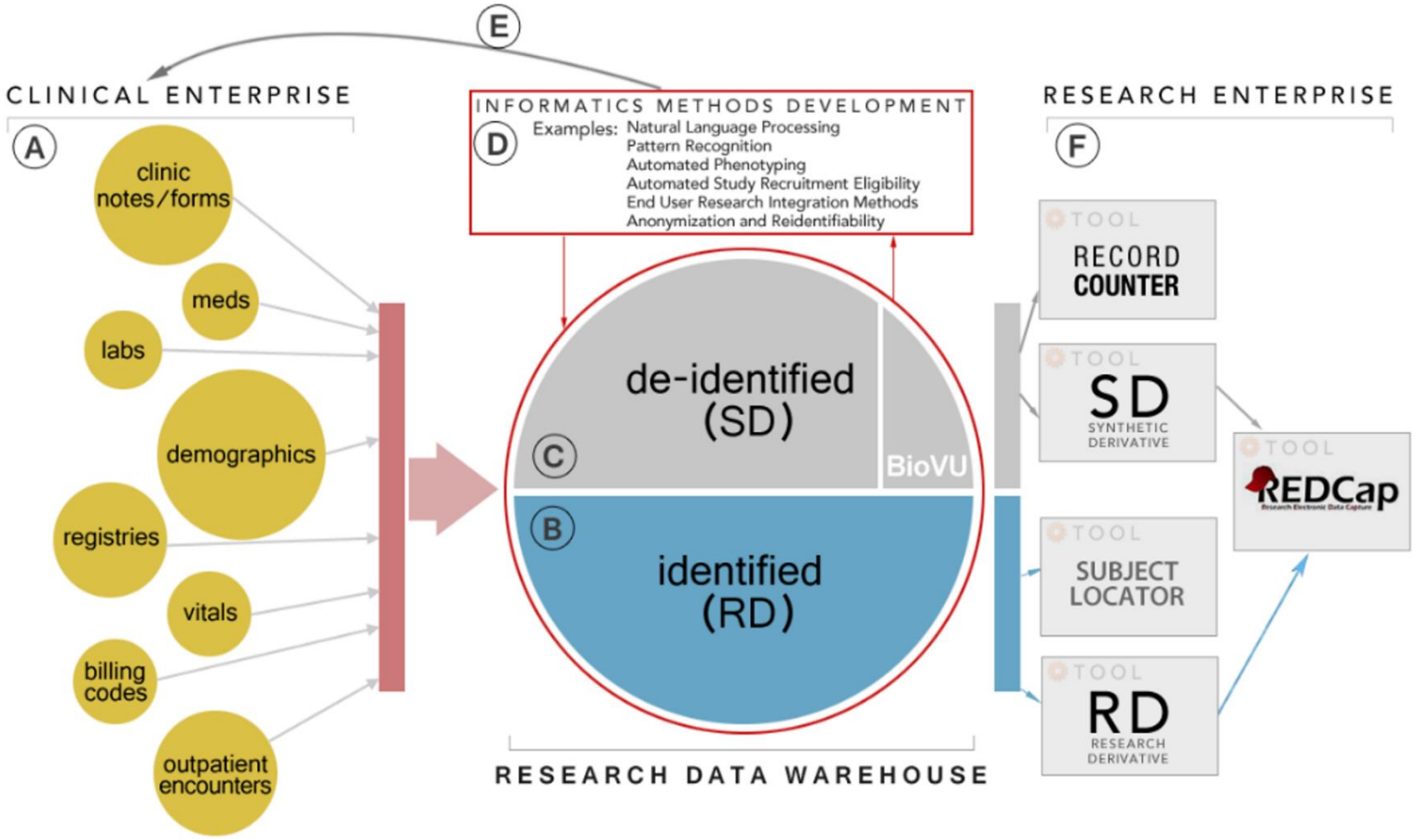


Systematic and longitudinal approach to height, weight, and body mass index data cleaning for efficient reuse of EHR data for research

Yi Jiang, MS; Melissa Basford, MBA; Jacqueline Kirby, MS; Xiaoming Wang, MS;
Paul Harris, PhD; Josh Denny, PhD MD
Vanderbilt University, Nashville, TN

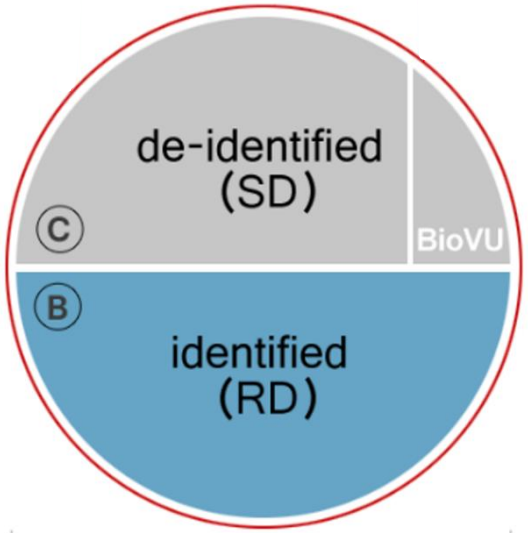
VANDERBILT UNIVERSITY
MEDICAL CENTER

Secondary Use of Clinical Data at Vanderbilt



VANDERBILT UNIVERSITY MEDICAL CENTER

Research Data Warehouse



- Synthetic Derivative (SD)
  - De-identified clinical & demographic data
  - BioVU - Vanderbilt DNA repository
    - Collect DNA from leftover clinical blood samples
- Research Derivative (RD)
  - Repository of identified clinical data
  - Updates regularly and is typically about 4 weeks behind the present date
- Data Access
  - User Interfaces
  - Programming services

Secondary Use of Clinical Data

- Benefit:
  - Patients and clinicians
    - Improvement in patient medical outcomes
  - Clinical and translational research
    - Hypotheses generation
    - Rapid cohort identification
  - Healthcare system
    - Public health surveillance for emerging threats
    - Healthcare quality measurement and improvement
- Data quality issues:
  - Incompleteness
  - Inconsistency
  - Inaccuracy

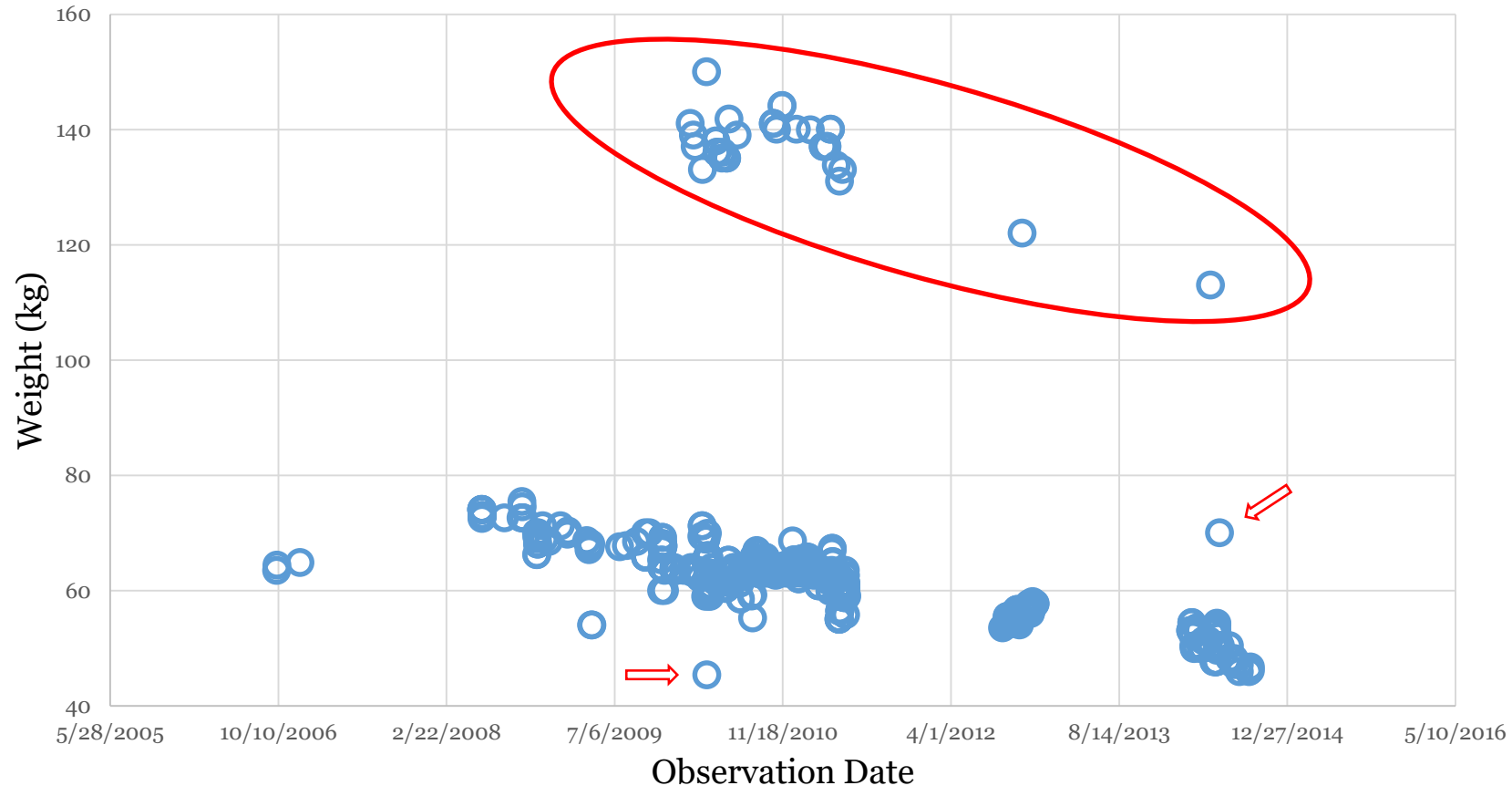
VANDERBILT UNIVERSITY MEDICAL CENTER

Weight, Height and BMI

- Characteristics:
  - numerous per patient record
  - span long time periods (20+ years for some)
  - multiple developmental stages (pediatric, adult, elderly)
  - life events (pregnant, non-pregnant, amputations)
  - medical conditions (nutritional disorders, dwarfism)

- Inaccuracies in the data are of two types
  - implausible values caused by measurement or data-entry errors
  - improbable measurements due to the longitudinal nature of the repository

Examples of Inaccuracies



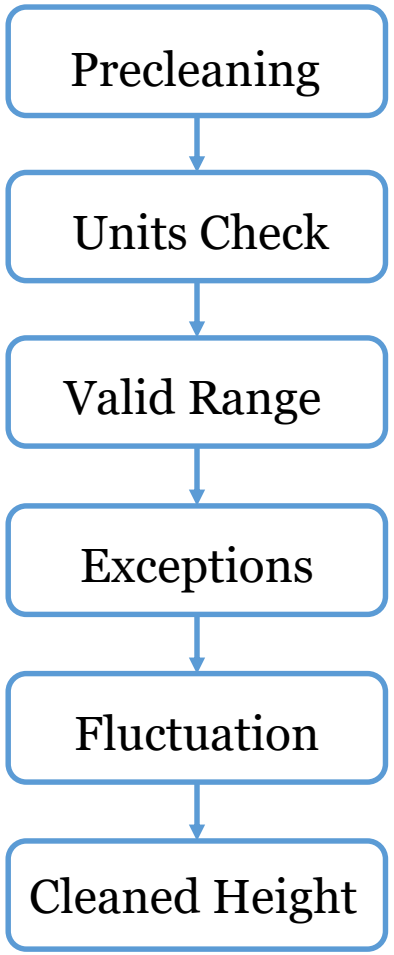
VANDERBILT UNIVERSITY MEDICAL CENTER

Our approach

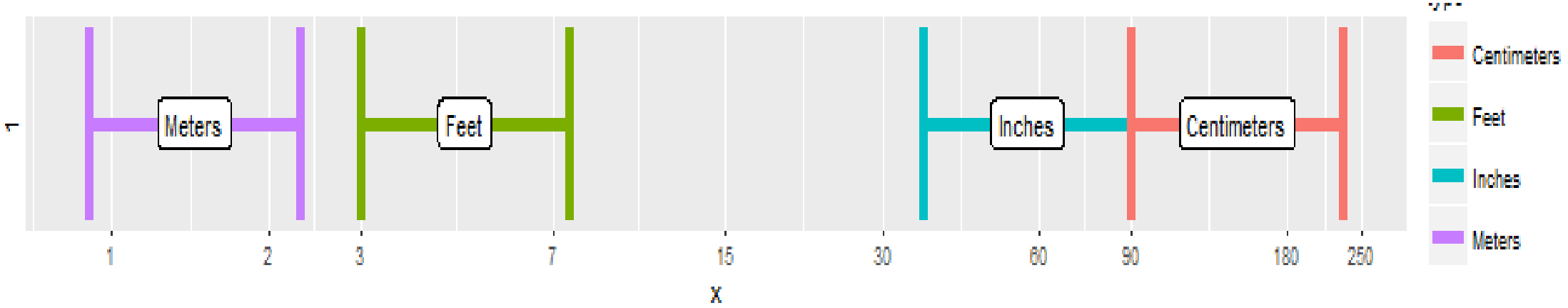
- Clean data in the entire all-patient research repository
- Clean for each subject individually
- Use a longitudinal perspective to address issues related to both implausible and improbable values
- Values measured at age >= 18 years, and not during pregnancy

Methods – Height Cleaning

- Correct values in wrong units
  - 36.0-89.9: in inches
  - 3.0-7.5: in feet
  - 0.9-2.3: in meters
- Valid value range: 90 – 230 cm
- Cases that allow values out of the range:
  - osteoporosis
  - spinal stenosis
  - arthroplasty
  - amputation of a lower limb
  - is wheel-chair bound
- Abnormal fluctuation
  - differ more than 3% from the median

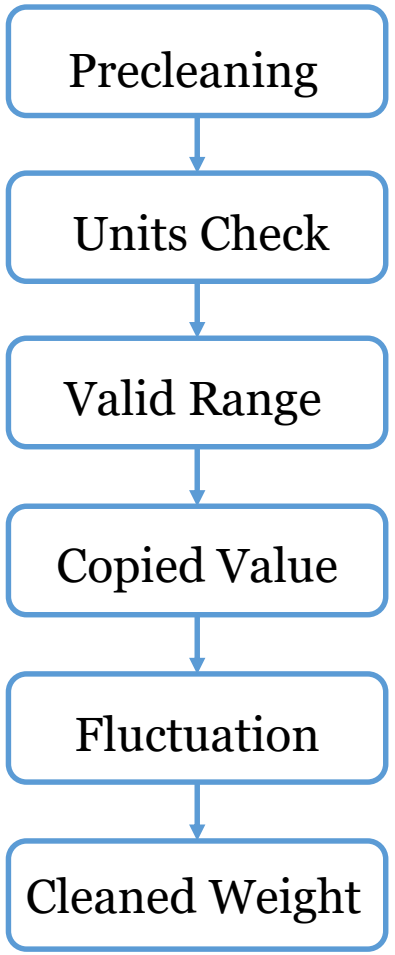


Methods – Height Cleaning



Methods – Weight Cleaning

- Correct values in wrong units
  - >1.5 × Median: in pounds
- Valid value range: 30 – 250 kg
  - dwarfism, anorexia or extreme weight loss: 20 – 250 kg
  - extreme obesity or extreme weight gain: 30 – 450 kg
  - has three or more weights that are > 250 kg: 30 – 450 kg
- When there are multiple weights on the same day, flag those that are equal to weights in previous measurement on a different day
- Abnormal fluctuation
  - differ more than 33% from the median within two years
  - differ more than 20% from the median within 60 days
  - differ more than 14% from the median within 30 days
  - differ more than 12% from the median within 21 days

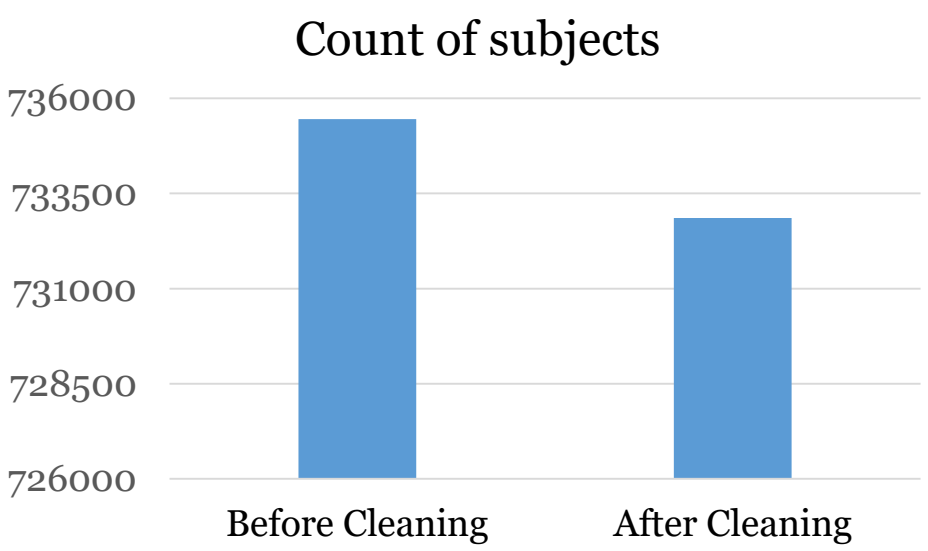


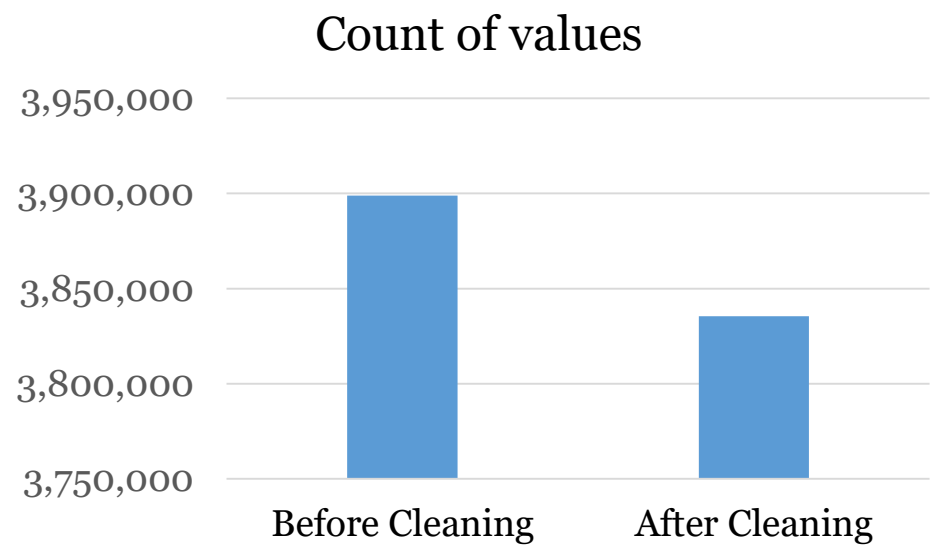
Methods – Cleaned BMI

- Use cleaned weights and heights

- If heights are not measured at the same time when weights are measured:
  - If the weight is measured at an age < 20, a height measured on the closest date within one year will be used to pair with that weight
  - If the weight is measured at an age ≥ 20, a closest height within five years will be used.
  - If no proper heights are found, the BMI will not be calculated

Results – Height





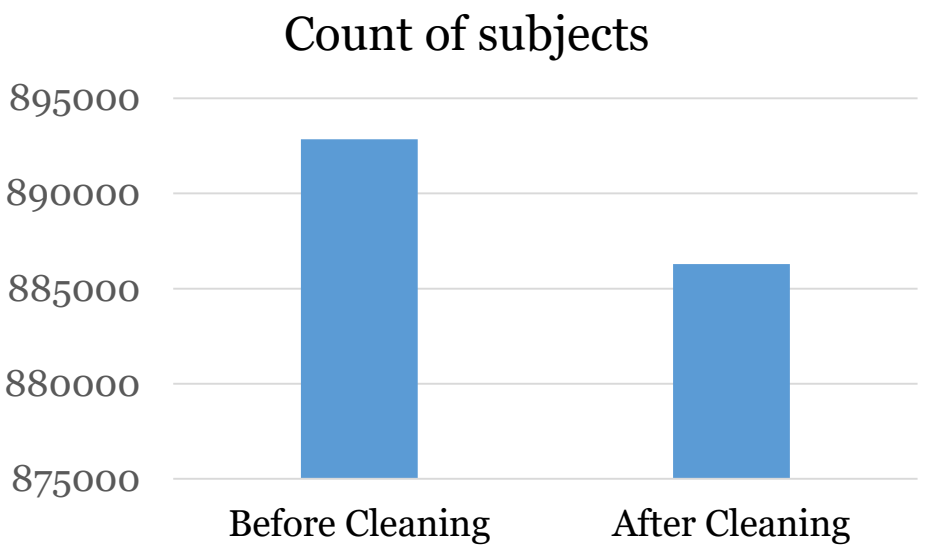
	Before Cleaning	After Cleaning
Range (min, max)		
Corrected Values		

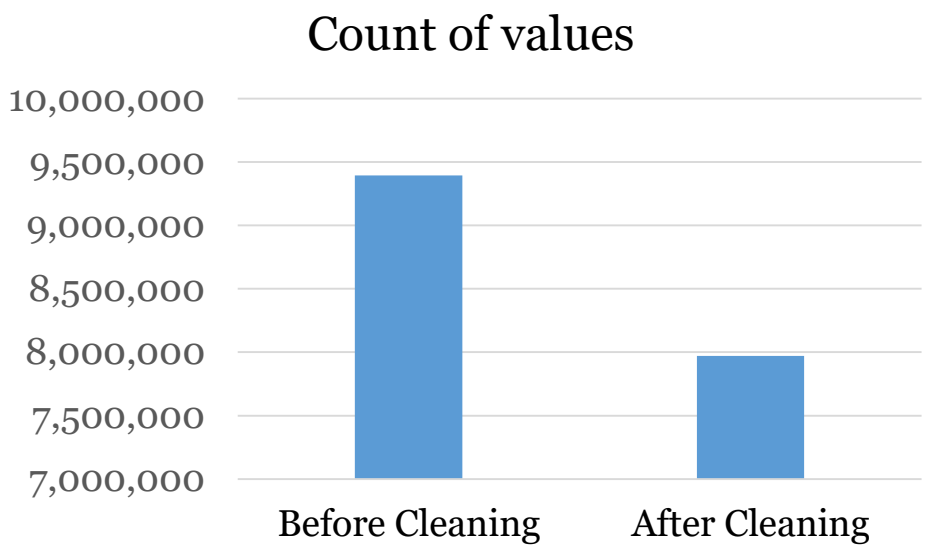
VANDERBILT UNIVERSITY MEDICAL CENTER

Results – Examples of invalid height values

	Before cleaning	After cleaning
Extremely low	0.165	Invalid
Extremely high	216329.26	Invalid
Wrong unit		
Values in inches (36.0-89.9)	68.5	173.99
Values in feet (3.0-7.5)	6	182.88
Values in meters (0.9-2.3)	1.62	162
Abnormal Fluctuation (median, diff%)	175.26 (167.64, 4.5%)	Invalid

Results – Weight



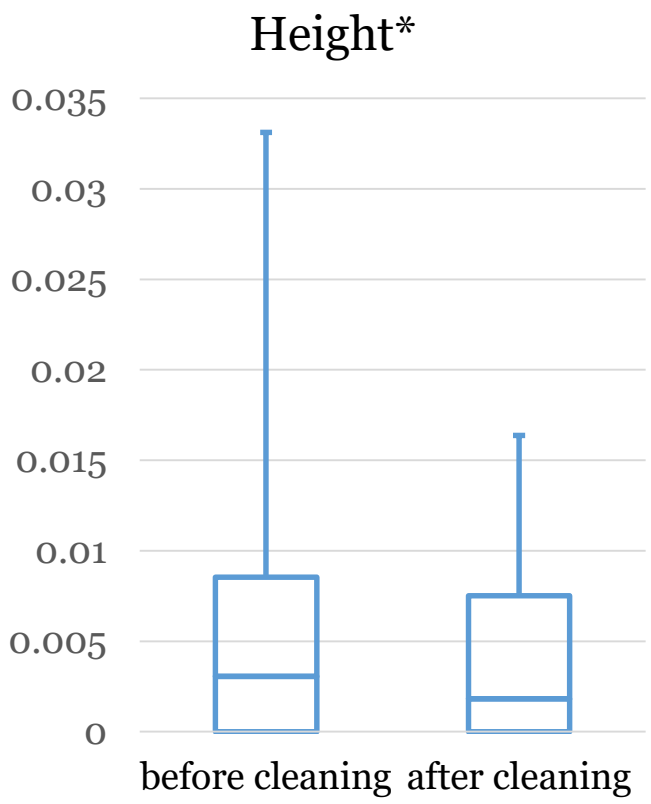


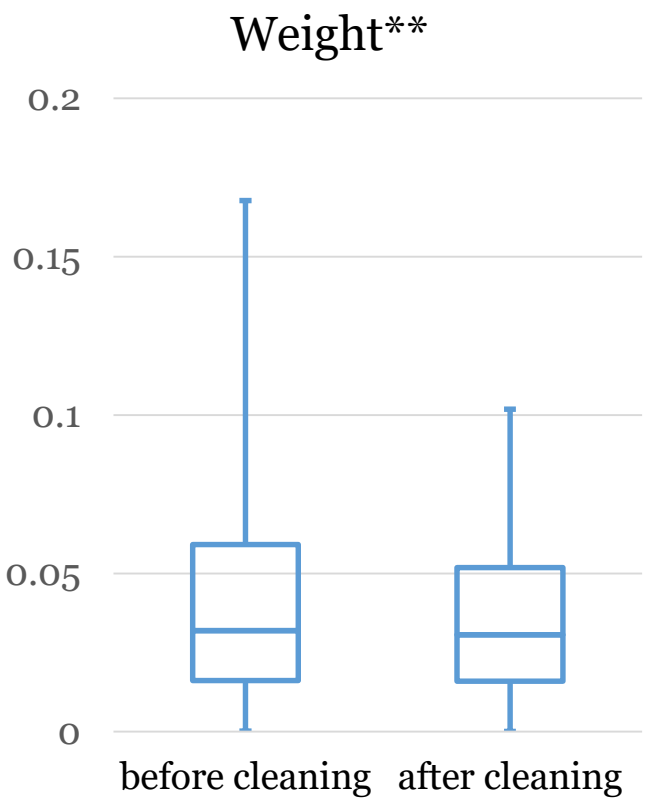
	Before Cleaning	After Cleaning
Range (min, max)		
Corrected Values		

Results – Examples of invalid weight values

	Before cleaning	After cleaning
Extreme low	0.06	Invalid
Extreme high	12037.43	Invalid
Values in lbs (median, diff%)	178.00 (80.29, 120%)	80.74
Abnormal Fluctuation (median, diff%)		
>= 33% within two years	63.73 (105.25, 39.4%)	Invalid
>= 20% within 60 days	59.0 (48.14, 22.5%)	Invalid
>= 14% within 30 days	92.9 (109.9, 15.5%)	Invalid
>= 12% within 21 days	100.7 (89.25, 12.8%)	Invalid

Results – Coefficient of Variance

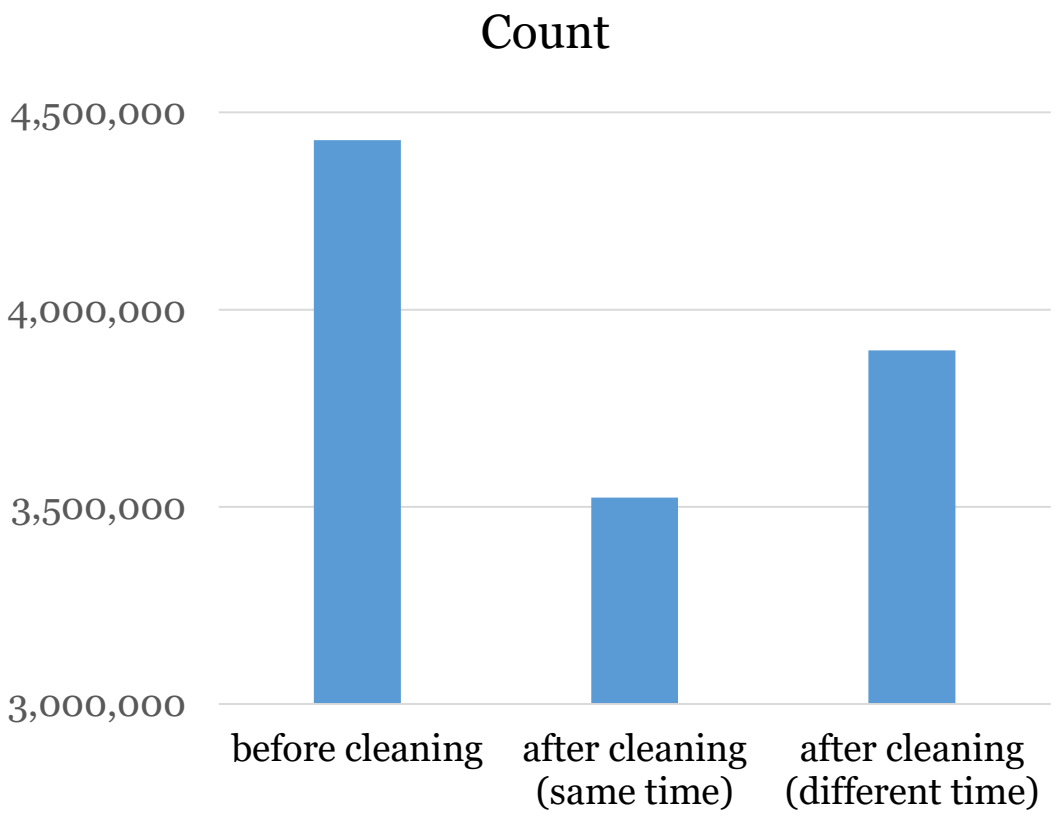


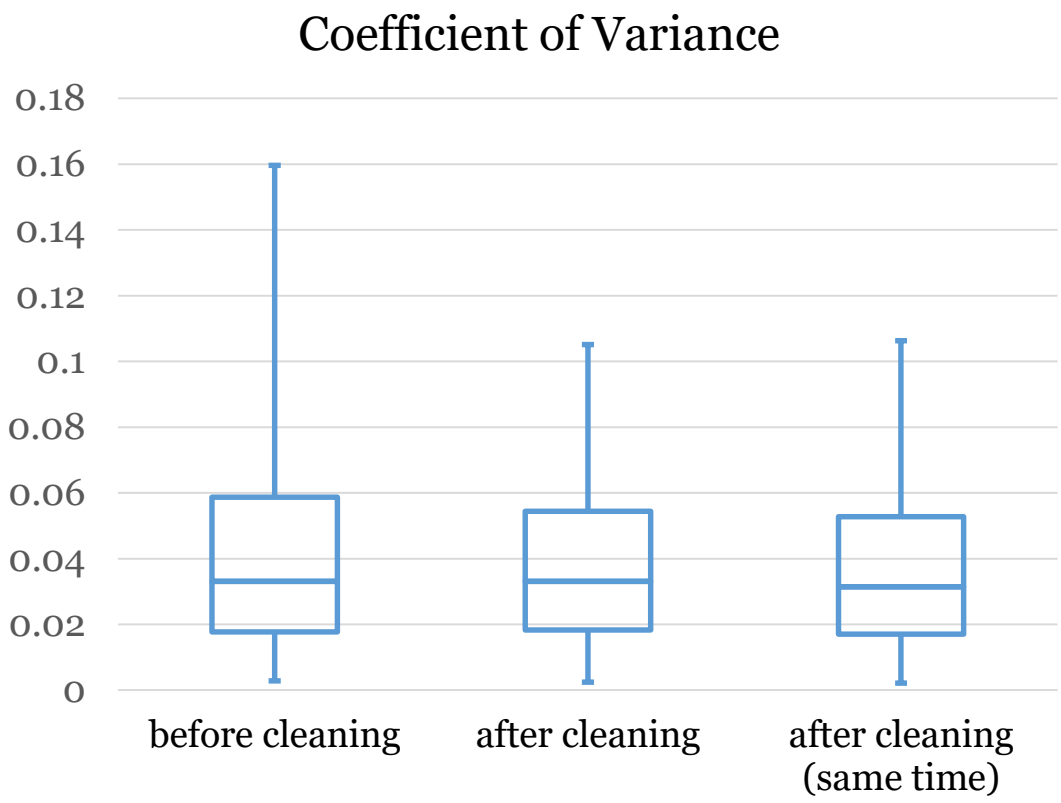


*,**: p-value < 2.2×10^{-16}

VANDERBILT UNIVERSITY MEDICAL CENTER

Results – BMI

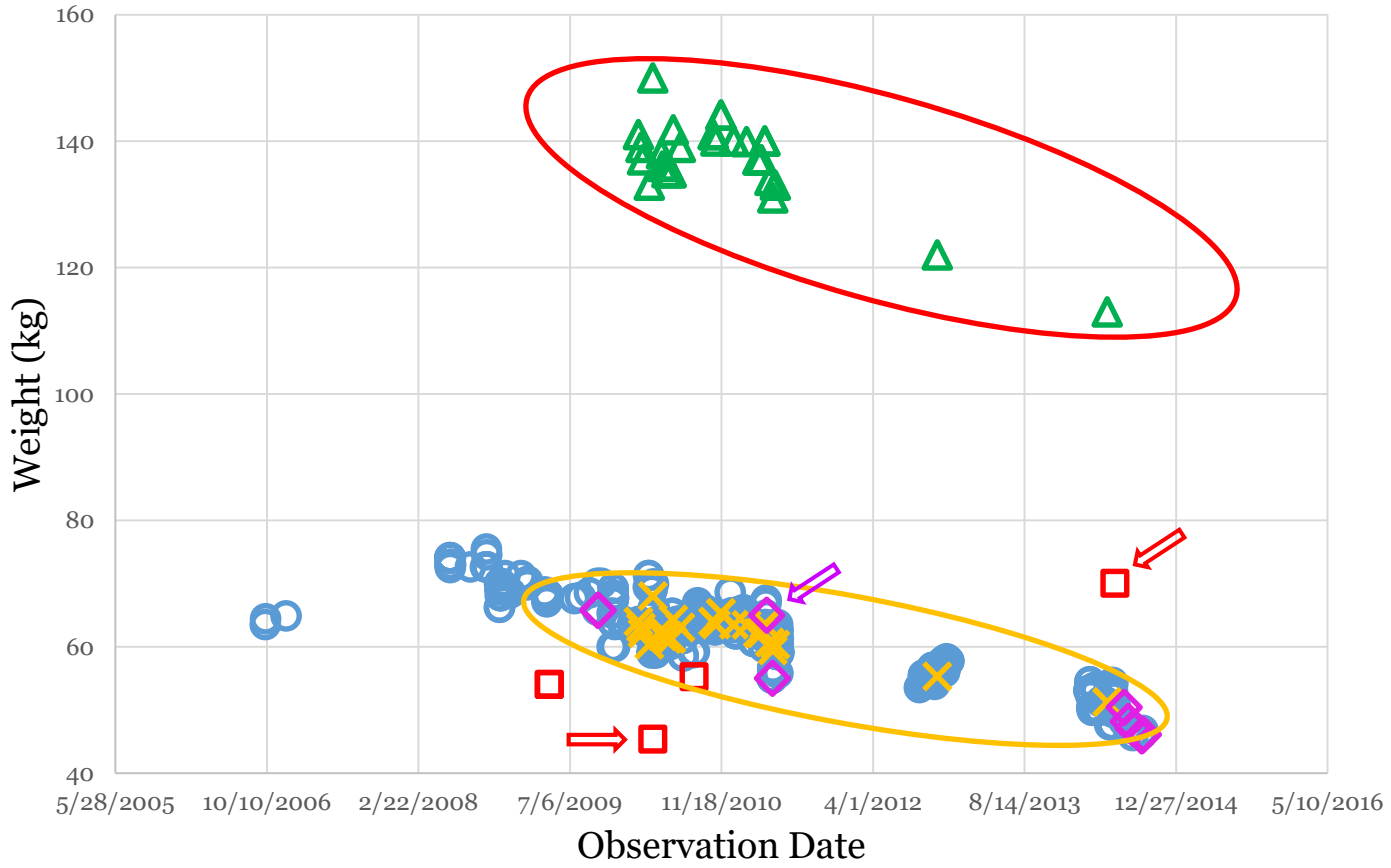




* p-value < 2.2×10^{-16}; ** p-value < 2.2×10^{-16}

VANDERBILT UNIVERSITY MEDICAL CENTER

Results – Example of single subject



- valid
- invalid
- uncorrected
- corrected
- copied values

VANDERBILT UNIVERSITY MEDICAL CENTER

Results – Manual review

- 20 subject's weights with 719 weight values
  - PPV = 97%
  - NPV = 91%
  - Sensitivity = 98%
  - Specificity = 88%

Results – Use case

- Type 2 diabetes and BMI

Sample Size (50% cases and 50% controls)	Regression Coefficient of BMI (*p*-value)	
	Before Cleaning	After Cleaning
100 Subjects	0.0314 (9.8×10^{-19})	0.372 (2.0×10^{-22})
500 Subjects	-0.0000357 (0.43)	0.870 (<0.001)
1000 Subjects	0.0000295 (0.27)	0.767 (<0.001)

Results – Performance

- Computation intensive
  - Whole repository: 13M values and 900K subjects
  - Updated daily:
    - Weight: 10000 new values for 5000 subjects.
    - Height: 3000 new values for 3000 subjects.
- Parallel processing
  - leverage the IBM® PureData™ System
- 7 minutes

Conclusion

- Longitudinal data are continually updated
- Data cleaning approaches require methods to account for changes that occur over time.
- Our evolving model changes assessment based on new information introduced.
- Upon our implementation, we have found at the subject level improved representations of BMI over time and a more accurate view of changes as they occur.

Acknowledgements

Melissa Basford, MBA
Josh Denny, PhD MD
Robert Carroll, PhD
Paul Harris, PhD
Jacqueline Kirby, MS
Jonathan Shildcrout, PhD
Xiaoming Wang, MS