Reducing Clinical Noise for Body Mass Index Measures Due to Unit and Transcription Errors in the Electronic Health Record

Robert Goodloe, MS[1], Eric Farber-Eger, BS[1], Jonathan Boston, BS[1], Dana C. Crawford, PhD[2], William S. Bush, PhD, MS[2]

[1]Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN, USA; [2]Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA.

Abstract

Body mass index (BMI) is an important outcome and covariate adjustment for many clinical association studies. Accurate assessment of BMI, therefore, is a critical part of many study designs. Electronic health records (EHRs) are a growing source of clinical data for research purposes, and have proven useful for identifying and replicating genetic associations. EHR-based data collected for clinical and billing purposes have several unique properties, including a high degree of heterogeneity or "clinical noise." In this work, we propose a new method for reducing the problems of transcription and recording error for height and weight and apply these methods to a subset of the Vanderbilt University Medical Center biorepository known as EAGLE BioVU (n=15,863). After processing, we show that the distribution of BMI from EAGLE BioVU closely matches population-based estimates from the National Health and Nutrition Examination Surveys (NHANES), and that our approach retains far more data points than traditional outlier detection methods.

Introduction

Genetic association studies are increasingly requiring large numbers of DNA samples linked to a multitude of phenotypes, traits and exposures to fully discover and describe the complex genetic architecture of human disease. In recognition of this need, concerted efforts are being made to amass the needed data through a variety of mechanisms including traditional epidemiologic designs or more contemporary biobanking approaches. In the United States, while large cardiovascular or cancer epidemiologic collections exist, there is no plan for the ascertainment of a larger US population-based cohort for genetic association studies given the enormous financial investment required[1]. Instead, resources and effort have been directed towards combining existing smaller studies[2] or partnering with health-care providers[3, 4]. The latter effort is receiving much support given the potential for practice-based biobanks to collect of large numbers of study samples linked to data collected in a clinical setting as part of patient care[5]. The advantage of this approach is that large numbers of clinically relevant DNA samples are readily available to investigators for genetic association studies.

The gold standard of study design is the prospective longitudinal cohort study, where individuals are ascertained from a population at a baseline start date for multiple measures (generally collected using a questionnaire) and are followed over time with updates at regular intervals. These cohort studies are very difficult to execute; recruitment is challenging, individuals drop out as the study progresses, the scope and data collection methods must be chosen and fixed at baseline, and large numbers of individuals are needed for statistical power. By comparison, practice-based biobanks at major metropolitan medical centers provide an attractive alternative; large portions of the population can be ascertained, most medically relevant data are collected in a loosely standardized way, individuals are tracked over time, and cost is reduced due to burden-sharing with health care providers. The major drawbacks however are non-regular intervals of information update and non-uniform data collection due to differences in clinical practice. These two issues could be jointly considered as a problem of "clinical noise."

While electronic health records (EHRs) are a rich source of phenotypic information, structured and free-text information from the EHR may require various degrees of processing to extract precise disease states. In the coarse sense, the presence of the same billing code from multiple distinct dates may be sufficient for phenotyping of some traits[6], but others may require refinement to eliminate confounding factors[7]. Continuous measures, such as laboratory values, may require extensive processing to remove confounding factors including medication use, comorbid conditions, and biases in sampling due to lab requisition protocols. Even critical measures such as vital signs can have high rates of missingness[8], and are subject to observational bias[9].

One research variable that best illustrates the "clinical noise" problem inherent in biobanks linked to EHRs is body mass index (BMI). BMI is a well-established risk factor for type 2 diabetes, hypertension, asthma[10], and various forms of cancer[11, 12]. BMI is a critical comorbidity for many clinical outcomes, and while this fact has been established by numerous epidemiological studies, the height and weight measurements that form the

basis for this measure are prone to transcriptional and conversion errors within EHR systems. The quality of BMI data has been previously examined from clinical records, and despite having an accurate protocol for measuring weight and height, only 35% of patient visits had data collected properly, typically because the patient's shoes were not removed prior to measurement[13]. However, measures were collected and recorded frequently (94.7% and 77.9% of the time for weight and height respectively). Wheelchair users are typically unable to stand for height measures with a stadiometer, forcing reliance on self-report[14] or other less accurate measures[15]. Furthermore, reliance on self-reporting for weight and height has well-established biases[13], and this bias has an racial/ethnic-dependent component[16], and varies with age though studies conflict on this effect[16, 17]. Even when height and weight are *measured* according to protocol, the results may not be recorded in consistent units across the clinic, and other studies using EHR data have required harmonization of units[18].

While it is known that clinical noise is especially problematic for assessing body-mass index from EHRs, there are few strategies proposed to address it. The most popular way to address this problem for BMI and other variables is manual curation. However, it is infeasible to extract and clean all height and weight data points manually given nearly every clinic visit has a recorded value resulting in a very large dataset (hundreds of thousands to millions of data points). Therefore, to enable the semi-automatic extraction of high quality height and weight data from EHRs to calculate BMI, we developed the Adjacency-based Longitudinal Outlier Extraction (ALOE) method and applied it to clinical records to a subset of the Vanderbilt University Medical Center's biorepository known as EAGLE BioVU (n=15,863)[19]. ALOE takes advantage of the longitudinal nature of the EHRs and the expectations of changes in weight and height over time for a given age range. Overall, we demonstrate that our data extraction method extracts high quality height and weight data with less data loss than standard outlier approaches.

Methods

Study Populations

BioVU is the Vanderbilt University Medical Center (VUMC) biorepository linked to de-identified EHRs. BioVU, including the ethical and legal considerations, has been previously described for the adult clinics[3] and pediatrics[20]. In brief, DNA is extracted from discarded blood samples drawn at VUMC outpatient clinics, and the DNA sample is linked to a de-identified version of the patient's EHR known as the Synthetic Derivative (SD). The VUMC SD contains approximately 20 years of clinical data representing ~2.1 million patients. To date, BioVU contains more than 200,000 DNA samples linked to de-identified EHRs. As part of the larger Population Architecture using Genomics and Epidemiology I (PAGE I) study[21], all DNA samples from minority (non-European descent) patients in BioVU as of 2011 were selected for study[19]. This subset of BioVU, referred to here as the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) BioVU, contains 15,863 DNA samples including DNA samples from African Americans (n=11,519), Hispanics (n=1,702), and 1,118 Asians (n=1,118). Race/ethnicity in BioVU is administratively assigned and has shown to be highly concordant with genetic ancestry among European Americans and African Americans[22] but less so for other groups such as Hispanics[23].

The National Health and Nutrition Examination Surveys (NHANES) are population-based cross-sectional surveys conducted by the National Center for Health Statistics at the Centers for Disease Control and Prevention. For each study participant, data on demographics, health, and lifestyle are collected. A physical exam is conducted by a CDC physician or health professional, and laboratory measures are assayed from blood and urine. Biospecimens for DNA extraction were collected beginning with phase 2 of NHANES III (between 1991 and 1994; n=7,159). DNA was also collected on consenting participants for NHANES 1999-2000 and 2001-2002 (n=7,839). NHANES is diverse and DNA samples were collected from self-described non-Hispanic whites (n=6,634), non-Hispanic blacks (n=3,458), Mexican Americans (n=3950), and others (n=956). For this study, CDC-measured height and weight were accessed for participants with DNA samples from NHANES III, NHANES 1999-2000, and NHANES 2001-2002 for a total of 14,734 samples.

All procedures were approved by the CDC Ethics Review Board and written informed consent was obtained from all participants. Because no identifying information was accessed by the investigators, Vanderbilt University's Institutional Review Board determined that this study met the criteria of "non-human subjects."

Adjacency-based Longitudinal Outlier Extraction (ALOE) Method

Step 1: Initial Outlier Detection and Characterization. We first examined the distributions of raw height and weight values to flag extreme unrealistic observations originating from transcription errors. Next, we divided the observations into obese and non-obese individuals. To identify obese individuals, clinical records were examined for International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) morbidity obesity codes (278.01) and/or mention of "obesity" in EHR clinical free text. Observations not having an obesity code or the obesity keyword were considered non-obese. This step was performed to disambiguate true distributional outliers from errors due to unit conversion or measure recording. Extreme outliers identified in non-obese individuals were manually investigated for validity and removed accordingly.

Step 2: Temporal Partitioning. Measurements of height and weight are typically recorded at regular intervals in the course of clinical care – often multiple times per year. If we assume that errors in transcription and unit conversion are distributed uniformly over all recorded measures, the distribution of an individual's height and weight measurements over a fixed time interval can be used to identify and correct errant measures. To evaluate observations across the longitudinal dataset, we generated a change-ratio distribution. A single index measurement was selected over a given year, and all subsequent measurements were divided by this index value to produce a range of $[0, \infty]$. This ratio is examined relative to established unit conversions [pounds to kilograms, inches to centimeters, feet to centimeters, meters to centimeters] to identify unit inconsistencies. If the value is approximately 1, both observations are recorded in the same unit, and deviations from 1 provide an approximation of the unit mismatch. For our dataset, we assumed that most measurements are recorded in centimeters (for height) and kilograms (for weight). The index observational value was defined by testing the following conditions per year:

- If total observations =2,
    - The first observation was divided by the second observation to generate the index value.
- If total observations =3,
    - The smaller of the first two observations were divided by the third observation. The value closest to 1 is chosen as the index value.
- If total observations > 3,
    - The two smallest values of first three observations were divided by the fourth observation. The value closest to 1 is chosen as the index value.

Step 3: Unit mismatch identification. Once the index value is assigned, all available measurements are divided by the index value to generate the change ratio distribution over this time interval and to identify spikes in the distribution indicative of unit mismatches.

- If the observed height index value was within a 0.10 standard deviation of 1 [0.9-1.1], the measurement is in centimeters.
- If the observed height index value was within a 0.003 standard deviation of 0.033 [0.029-0.036], the measurement is in feet.
- If the observed height index value was within a 0.04 standard deviation of 0.39 [0.35-0.43], the measurement is in inches.
- If the observed height index value was within a 0.001 standard deviation of 0.01 [0.008-0.011], the measurement is in meters.
- If the observed weight index value was within a 0.20 standard deviation of 1 [0.8-1.2], the measurement is in kilograms.
- If the observed weight index value was within a 0.45 standard deviation of 2.2 [1.75-2.65], the measurement is in pounds.
- If the observed weight index value was within a 0.10 standard deviation of 0.45 [0.35-0.55], the measurement was a kilogram measure assumed to be in pounds that was converted to kilograms.

Each of these conversions is then resolved to the base units of centimeters and kilograms (Figure 1). All conversion value ranges were given standard deviations to account for a 30 lb. change in weight and a 6-inch

deviation in height measurements over the course of a year. If this algorithm were applied over a shorter (months) or longer (5 year) time interval, these standard deviations would be adjusted to reflect natural changes in weight and height expected over that period. If values were outside the standard deviation, the corresponding
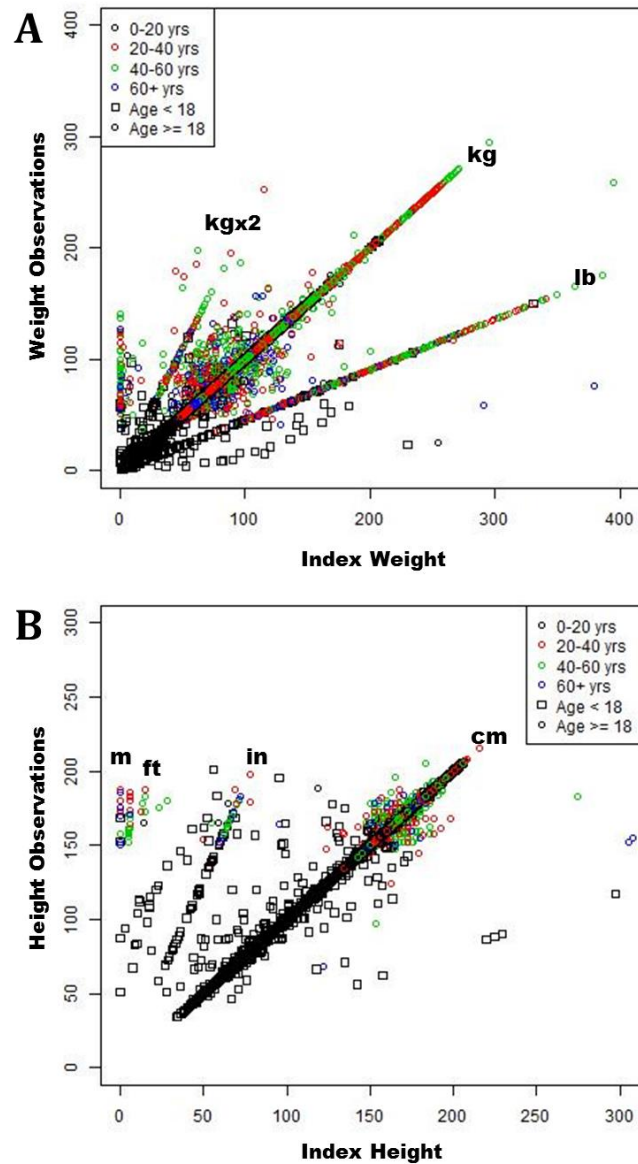

Figure 1. The null (diagonal) line represents near-identical index and observed weight (A) and height (B) values. Deviating lines represent original values that were recorded in pounds (lb), kilograms (kg), double-converted kilograms (kgx2), meters (m), feet (ft), inches (in), and centimeters (cm). Lone circles likely represent transcriptional errors. Squares represent pediatric measures which may represent true changes. Data points are color coded by age range. These plots were truncated to display an interpretable graph.

measurements were set to missing due to lack of validity. Manual editing and specific conditioning was used to preserve data in the case of clear transcription errors, such as the addition of a zero (e.g. a person gains 100 lbs. in one visit and lost 100 lbs. in the next visit).

Case Study

To demonstrate the ALOE method, we present here a single patient within EAGLE BioVU with 70 independent clinic visit dates spanning seven years in the clinical record. We plotted all measured weights available in the EHR for this patient and observed at least one dramatic weight difference between two clinic visit dates only one month apart (Figure 2). This observed difference suggests a dramatic weight loss of 163 lbs. (73.94 kgs) followed by a dramatic weight gain of almost the same amount of weight a month later. Similar observations were made for the last four clinic visit dates compared with the other weights immediately preceding them.

  To determine the nature of the transcription error, we divided the smallest two weights of the first three weights (121.28 and 126.1 kg, respectively) by the fourth weight (123.23 kg) to establish the weight index value. The smaller of the two weights (121.28 kg) was closest to 1 when divided by the fourth weight and declared the weight index value. We then divided all five suspect weights by the weight index value, and all five were within the range of 0.50 – 0.57 consistent with a transcription error where the original measurement was in kilograms assumed to be in pounds that was converted to kilograms (kgx2).
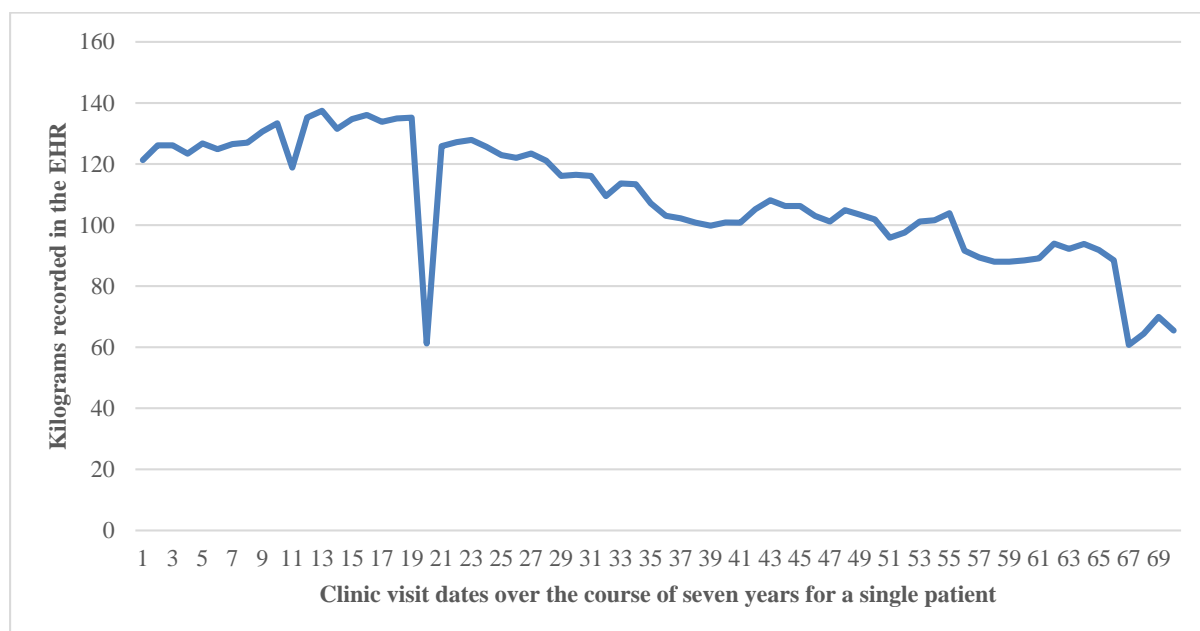

Figure 2. The distribution of weights (in kilograms) recorded in the electronic health record for a single patient over the course of seven years. X-axis represents independent clinic visits in order of visit and the y-axis represents the corresponding weights recorded and assumed to be in kilograms.

Residual Modeling Method

With this approach, we exploit the relationship between height, weight, and age. We regressed age onto height and weight respectively, creating a linear slope to predict values for each individual measure. Incorrect values exhibit a larger deviation from the predicted value and can be identified using a variety of statistical measures of influence, including Cook's distance, Leverage, DFfits, Studentized residuals, and Covariance Ratio. If the modeled data indicated at least three positive tests of any statistical measure, that individual data point was set to missing. This method was executed two different ways: generating a single model over all observations for an individual, and generating multiple models over all available observations iteratively. The iterative approach used the influence measures to identify significant outliers, excluded the identified data outliers, and repeated this procedure up to three times. All analyses were conducted using SAS v9.3.

Results

The majority of EAGLE BioVU individuals are African American (73%), followed by Hispanics (11%), and Asians (7%). The median age of individuals in EAGLE BioVU is 37 (20.46 standard deviation), with ~16% of individuals at least 55 years of age. As expected given EAGLE BioVU is drawn from a clinical population, the majority of individuals are female (63.35%). On a per patient basis, the number of clinic visits captured in EAGLE BioVU ranges from 1 to 1,456 with an average of 81.8 clinic visits per patient[19].

We extracted the height and weight values recorded in EAGLE BioVU for all clinic visits per individual and calculated BMI. The distribution of 225,903 per-visit BMI values for children (age < 18) and adults calculated from raw height and weight measurements (Figure 3). The effects of unit mismatches are clear, with impossible (-36) and extreme values (954) derived from improperly converted height and weight measures in the calculation. These errors also cause a wide standard deviation (14.88).
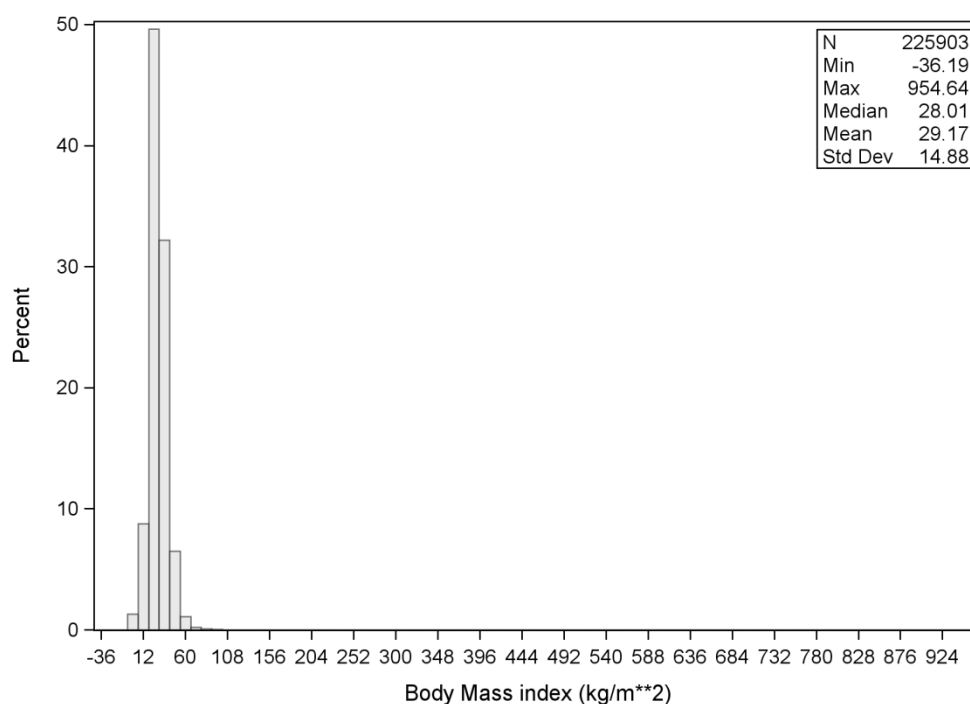

N	225903
Min	-36.19
Max	954.64
Median	28.01
Mean	29.17
Std Dev	14.88

Figure 3. Distribution of raw body mass index (BMI) values from EAGLE BioVU.

We then applied our ALOE method to the raw height and weight measures extracted from the clinic visits. Figure 1 illustrates the fundamental principle of the ALOE method. Once an index measurement is selected in step 2, observations effectively cluster (by slope) based on recorded units. Weight (Figure 1A) naturally fluctuates over the course of a year, shown by a cloud of points off the main diagonal. Height measurements (Figure 1B) have much less natural variability, and points off the diagonal for height likely represents measurement error, either due to recall bias or the impact of shoes on stadiometer measurements.
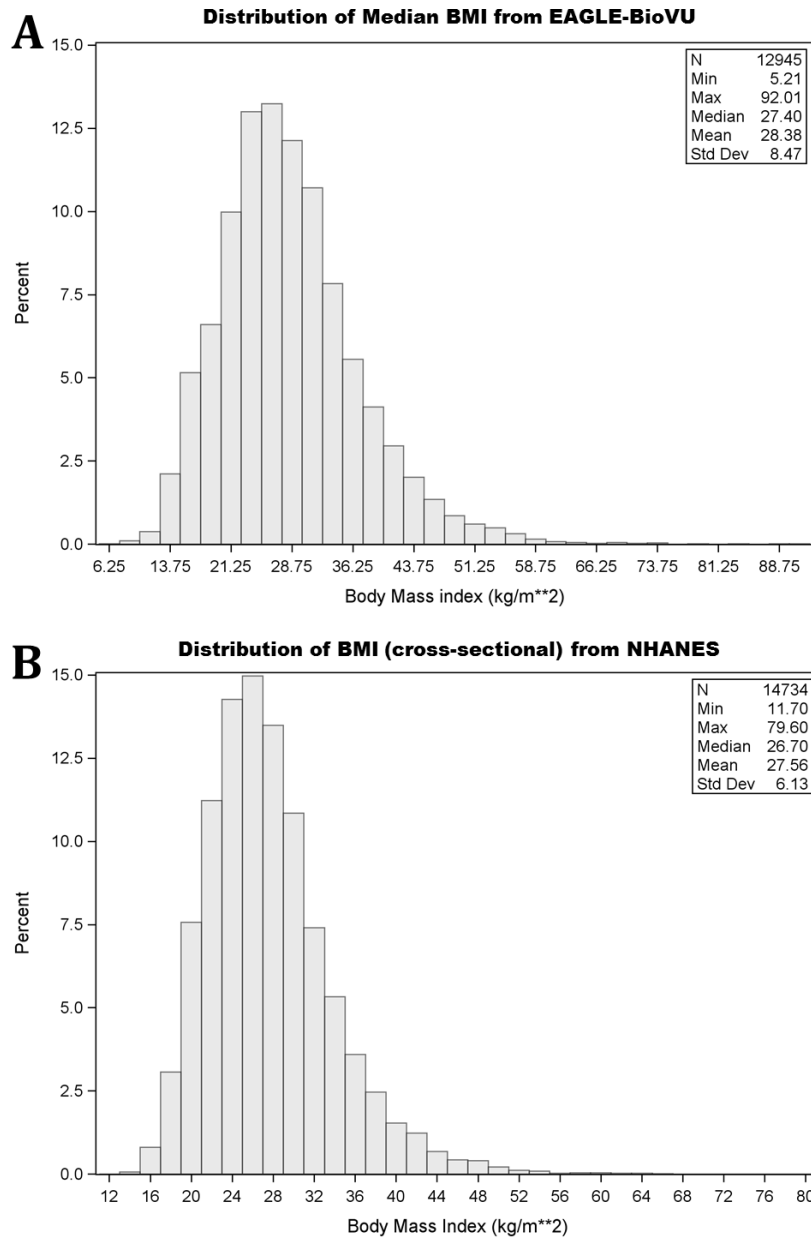
Figure 4. Comparison of median body mass index (BMI) values from EAGLE BioVU (A) to BMI values from the National Health and Nutrition Examination Surveys (NHANES) (B).

The distribution of median BMIs after processing by the ALOE method is shown in Figure 4a. Median BMI was selected per-individual and is plotted for comparison to baseline BMI measurements collected in epidemiological studies. The distribution of BMIs from the NHANES is shown in Figure 4b. After processing, our data show a very similar distribution to the population level estimate. There is a slight skew toward higher BMIs in EAGLE BioVU possibly reflecting both known geographical and racial/ethnic differences in BMI distributions in the United States[19, 24, 25].

We also examined the differences in dropped data points based on residual modeling and ALOE strategies. Table 1 illustrates that ALOE retains more data points than both versions of residual modeling. When performing residual modeling for outlier detection across the entire dataset, just over half of all observations are

considered usable after processing. Using an iterative approach (described in the methods section), we progressively eliminated outliers across the entire dataset which may have eliminated many true observations. Performing this modeling within each individual proved more successful, but still may not have detected outliers due to subtler unit conversions (inches to centimeters).

Table 1. Frequencies of all observations within EAGLE BioVU by processing method

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

Discussion

In this work, we have shown that height and weight measures extracted from BioVU, an EHR-based biorepository, follow distinct patterns representing problems in unit conversion. By exploiting the temporal nature of the EHR, and the fact that individuals often have multiple height and weight measurements over time, many errant entries of height and weight can be resolved into the correct units. The ALOE approach leverages expected changes in weight and height measurements over a fixed time period (1 year) to identify outlier observations which can be converted (in the case of unit error) or dropped (in the case of transcription error). Greater than 98% of all observations are retained from ALOE, and the resulting distribution of derived BMI measures closely matches those reported by the nationally representative NHANES.

The issue of clinical noise is due largely to the extreme heterogeneity that is typical of large clinical databases. Temporal heterogeneity is frequent, as some patient records have frequent visits and laboratory measures, where others have few or none. Various clinics use different laboratory panels, uneven collection of clinical measures, and may even record measures using inconsistent units. For example, while weight is typically consistently recorded as part of patient intake, height is not recorded as regularly. When it is recorded, it may be from self-report or direct measure via a stadiometer, and even then some clinics may record in metric versus US customary units. This is common when comparing pediatric or natal measures to adult measures. Self-report may result in transcription errors, such as the entry of 5 feet, 9 inches as 59 inches. All these issues are further compounded by the presence of true outliers in the clinical system – abnormal or out-of-range test values indicative of a clinical disorder.

The ALOE approach has limitations. The method relies on dense temporal data, with multiple measures over a fixed time period. In this study, we used a 1-year window, and while this could be expanded, larger time intervals allow for larger natural changes in weight that may reduce accuracy in clustering unit distributions. Also, as with any quality control process, a degree of manual editing and interaction with the data is still recommended to preserve some data points. That is, even when the ALOE approach is applied, corrections and removal of outliers is at the discretion of the individual investigator. The ALOE approach only offers solutions for research settings and does not address the cause of transcription errors in the actual clinical record. Nevertheless, despite the nearly ubiquitously measured height and weight values stored in clinical systems have systematic flaws that can be reasonably corrected in research settings with appropriate data processing techniques.

References

1.Collins FS. The case for a US prospective cohort study of genes and environment. Nature. 2004;429:475-7.

2.Willett WC, Blot WJ, Colditz GA, Folsom AR, Henderson BE, Stampfer MJ. Merging and emerging cohorts: Not worth the wait. Nature. 2007;445(7125):257-8.

3.Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clin Pharmacol Ther. 2008;84(3):362-9.

4.McCarty CA, Chapman-Stone D, Derfus T, Giampietro PF, Fost N. Community consultation and communication for a population-based DNA biobank: The Marshfield clinic personalized medicine research project. American Journal of Medical Genetics Part A. 2008;146A(23):3026-33.

5.Collins FS, Varmus H. A New Initiative on Precision Medicine. New England Journal of Medicine. 2015;372(9):793-5.

6.Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12(6):417-28.

7.Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. Journal of the American Medical Informatics Association. 2013;20(e2):e206-e11.

8.Gravel J, Opatrny L, Gouin S. High rate of missing vital signs data at triage in a paediatric emergency department. Paediatr Child Health. 2006;11(4):211-5.

9.Edmonds ZV, Mower WR, Lovato LM, Lomeli R. The reliability of vital sign measurements. Annals of Emergency Medicine. 2002;39(3):233-7.

10.Mokdad AH, Ford ES, Bowman BA. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. JAMA. 2003;289(1):76-9.

11.Morimoto LM, White E, Chen Z, Chlebowski RT, Hays J, Kuller L, et al. Obesity, body size, and risk of postmenopausal breast cancer: the Women's Health Inititative (United States). Cancer Causes Control. 2002;13(8):741-51.

12.Moore LL, Bradlee ML, Singer MR, Splansky GL, Proctor MH, Ellison RC, et al. BMI and waist circumference as predictors of lifetime colon cancer risk in Framingham Study adults. Int J Obes Relat Metab Disord. 2004;28(4):559-67.

13.Greenwood JLJ, Narus SP, Leiser J, Egger MJ. Measuring Body Mass Index According to Protocol: How are Height and Weight Obtained? Journal for Healthcare Quality. 2011;33(3):28-36.

14.Froehlich-Grobe K, Nary DE, VanSciver A, Washburn RA, Aaronson L. Truth Be Told: Evidence of Wheelchair Users' Accuracy in Reporting Their Height and Weight. Archives of Physical Medicine and Rehabilitation. 2012;93(11):2055-61.

15.Froehlich-Grobe K, Nary DE, Van Sciver A, Lee J, Little TD. Measuring height without a stadiometer: empirical investigation of four height estimates among wheelchair users. Am J Phys Med Rehabil. 2011;90(8):658-66.

16.Fillenbaum GG, Kuchibhatla MN, Whitson HE, Batch BC, Svetkey LP, Pieper CF, et al. Accuracy of Self-reported Height and Weight in a Community-Based Sample of Older African Americans and Whites. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 2010;65A(10):1123-9.

17.Kuczmarski M, Kuczmarski R, Najjar M. Effects of Age on Validity of Self-Reported Height, Weight, and Body Mass Index: Findings from the Third National Health and Nutrition Examination Survey, 1988-1994. Journal of the American Dietetic Association. 2001;101(1):28-34.

18.Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. Journal of the American Medical Informatics Association. 2010;17(5):568-74.

19.Crawford DC, Goodloe R, Farber-Eger E, Boston J, Pendergrass SA, Haines JL, et al. Leveraging epidemiologic and clinical collections for genomic studies of complex traits. Human Heredity. 2015;79(3-4):137-46.

20.McGregor TL, Van Driest SL, Brothers KB, Bowton EA, Muglia LJ, Roden DM. Inclusion of Pediatric Samples in an Opt-Out Biorepository Linking DNA to De-Identified Medical Records: Pediatric BioVU. Clin Pharmacol Ther. 2013;93(2):204-11.

21.Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, et al. The Next PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. American Journal of Epidemiology. 2011;174(7):849-59.

22.Dumitrescu L, Ritchie MD, Brown-Gentry K, Pulley JM, Basford M, Denny JC, et al. Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. Genet Med. 2010;12(10):648-50.

23.Hall JB, Dumitrescu L, Dilks HH, Crawford DC, Bush WS. Accuracy of Administratively-Assigned Ancestry for Diverse Populations in an Electronic Medical Record-Linked Biobank. PLoS ONE. 2014;9(6):e99161.

24.Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of childhood and adult obesity in the United States, 2011-2012. JAMA. 2014;311(8):806-14.

25.Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System Survey Data. Atlanta, GA.