Department of Veterans Affairs

JRRD Journal of Rehabilitation Research & Development
Volume 47, Number 8, 2010
Pages 739–750

VHA Corporate Data Warehouse height and weight data: Opportunities and challenges for health services research

Polly Hitchcock Noël, PhD;[1–2]* Laurel A. Copeland, PhD;[1–2] Ruth A. Perrin, MA;[3–4] A. Elizabeth Lancaster, BS;[5] Mary Jo Pugh, RN, PhD;[1–2] Chen-Pin Wang, PhD;[1–2] Mary J. Bollinger, MPH;[1] Helen P. Hazuda, PhD[2]
[1]Department of Veterans Affairs (VA), South Texas Veterans Health Care System, San Antonio, TX; [2]University of Texas, Health Science Center at San Antonio, San Antonio, TX; [3]Center for Management of Complex Chronic Care, Hines, IL; [4]VA Information Resource Center, Hines, IL; [5]Veterans Health Administration Support Service Center, Washington, DC

Abstract—Within the Veterans Health Administration (VHA), anthropometric measurements entered into the electronic medical record are stored in local information systems, the national Corporate Data Warehouse (CDW), and in some regional data warehouses. This article describes efforts to examine the quality of weight and height data within the CDW and to compare CDW data with data from warehouses maintained by several of VHA's regional groupings of health care facilities (Veterans Integrated Service Networks [VISNs]). We found significantly fewer recorded heights than weights in both the CDW and VISN data sources. In spite of occasional anomalies, the concordance in the number and value of records in the CDW and the VISN warehouses was generally 97% to 99% or greater. Implausible variation in same-day and same-year heights and weights was noted, suggesting measurement or data-entry errors. Our work suggests that the CDW, over time and through validation, has become a generally reliable source of anthropometric data. Researchers should assess the reliability of data contained within any source and apply strategies to minimize the impact of data errors appropriate to their study population.

Key words: anthropometric measurements, body mass index, data error, electronic medical record, height, obesity, rehabilitation, secondary data, veterans, Veterans Health Administration, weight.

INTRODUCTION

Obesity is associated with significant morbidity and is a modifiable risk factor for a variety of chronic illnesses, including several leading causes of death, such as cardiovascular disease, diabetes, and some cancers [1–6]. Obesity also contributes to the disablement process and complicates rehabilitation [7–9]. In the Veterans Health Administration (VHA), overweight and obese veterans

Abbreviations: BMI = body mass index, CDW = Corporate Data Warehouse, DSS = Decision Support System, EMR = electronic medical record, FY = fiscal year, HDR = Health Data Repository, HL7 = Health Level Seven (International), HSR&D = Health Services Research and Development Service, IRB = institutional review board, SQL = structured query language, VA = Department of Veterans Affairs, VHA = Veterans Health Administration, VISN = Veterans Integrated Service Network, VistA = Veterans Health Information Systems and Technology Architecture, VSSC = VHA Support Service Center.

***Address all correspondence to Polly Hitchcock Noël, PhD; South Texas Veterans Health Care System, 7400 Merton Minter Blvd (11c6), San Antonio, TX 78229-4404; 210-617-5314; fax: 210-567-4423. Email: polly.noel@va.gov**
DOI:10.1682/JRRD.2009.08.0110

comprise the majority of the patient population and have high rates of disability and generally poor health [10]. Use of existing data collected during routine clinical encounters can potentially provide important information about obesity and its outcomes. Clinical data most relevant to investigations of obesity include ICD-9-CM (International Classification of Diseases-9th Revision, Clinical Modification) diagnoses and anthropometric measurements. Unfortunately, a number of recent studies indicate that obesity is underdiagnosed, especially among primary care patients [11–13]. Therefore, access to anthropometric measurements is critical for accurately identifying obese patients, the care they receive, and their outcomes. Weight data with contemporaneous height assessments permit calculation of a patient's body mass index (BMI). Although it is an indirect measure of body fat, BMI is easily assessed and has become the standard metric for obesity in routine clinical practice [1]. Most guidelines, however, also recommend assessing waist circumference because it more accurately assesses central adiposity and health-related risk [6].

The use of anthropometric data recorded in electronic medical records (EMRs) and subsequently made available in an analyzable database can help generate cost- and time-efficient evidence. Data derived from routine clinical encounters are generally assumed to contain more errors than information collected through carefully controlled and standardized assessments, such as those that occur during prospective epidemiological studies and clinical trials. If the errors are minimal and random, however, the cost and time efficiencies gained by researchers using these data may outweigh weaknesses in measurement or recording accuracy [14].

The overall goal of this work is to assess the usefulness of anthropometric data in VHA's national Corporate Data Warehouse (CDW). The specific aims were to (1) at the macro level, examine whether the data fields were populated as expected; (2) at the individual level, evaluate the completeness and accuracy of the data; and (3) explore implications of our findings for assessing obesity and its associated risk with BMI derived from heights and weights and waist circumference.

BACKGROUND

The CDW is a national repository comprising data from several VHA clinical and administrative systems [15]. The CDW's objective is to provide data and tools to support management decision making, performance measurement, and research. It contains data not previously available in VHA's other national databases. These data include anthropometric measures, such as weight, height, and waist circumference/girth; vital signs, including blood pressure, pulse oximetry, and temperature; and other measures, such as pain assessments. Historical data in the CDW go back to fiscal year (FY) 1999 (October 1998–September 1999), and current data are added nightly. While anthropometric and vital sign data are now available, additional domains such as laboratory, pharmacy, and inpatient diagnoses and procedures based on VHA's EMR will be added over the next several years. The CDW is currently the only source of nationwide VHA anthropometric and blood pressure data. In addition, in contrast to some of VHA's other national databases, such as the Medical SAS Data Sets and Decision Support System (DSS) National Data Extracts, it is a relational database rather than a set of discrete files separated by FY and data type [15].

In assessing the utility of anthropometric data available in the CDW, tracing the process by which clinical data are entered into the EMR and transferred into aggregated databases at the regional and national levels is useful. Anthropometric data (weight, height, and waist circumference) and vital sign data, such as blood pressure, are entered by clinical staff, stored in VHA facility EMR systems (Veterans Health Information Systems and Technology Architecture [VistA]) across the nation, and uploaded daily into VHA's national CDW. However, the CDW does not extract these data directly from the VistA files. VistA anthropometric and vital signs data are transmitted by HL7 (Health Level Seven) messages (HL7 International; Ann Arbor, Michigan) to Department of Veterans Affair's (VA's) Health Data Repository (HDR), from which the CDW extracts, transforms, and loads these data to its own structured query language (SQL) data fields. When the CDW database is updated (refreshed), changed data values are written over, not maintained. That is, the CDW is a regularly updated warehouse holding no stable reference files comparable to the VHA's Medical SAS Data Set or DSS National Data Extracts final FY files. Furthermore, while out of range values are cleaned from Medical SAS and DSS data, errors and out of range values in VistA data will be found in the CDW. Vital sign and anthropometric data appear in the CDW in both text and numeric form. The

text field presents data "as is" from its VistA extraction. The numerical result field presents the same result as a discrete numerical value based on a very conservative transformational algorithm [15].

Several of VHA's 21 regional groupings of healthcare facilities known as Veterans Integrated Service Networks (VISNs) have also developed data warehouses to support administrative and clinical decision making. These VISN data repositories draw from the same local VistA systems as the national CDW but typically through different technical processes (e.g., M Programming [Microsoft; Redmond, Washington] vs HL7 messaging). Moreover, the VISN warehouses extract vitals and anthropometric data either directly from VistA files or from a collector/feeder database on a SQL server connected to each VistA system but not from the HDR. Therefore, the "same" anthropometric data can exist at several levels within VHA, i.e., within the local VistA system, a VISN data warehouse, and the national CDW. Variations may arise, however, because some data could be lost in transmission or different filters used by different systems could result in the inclusion of slightly different subsets of the EMR data. Furthermore, VistA data constantly change with every new clinician entry, while warehouse updates, known in the data warehousing field as refreshing, are not done simultaneously by all the warehouses. Hence, small unavoidable differences will always exist in the data in one warehouse versus another.

Although anthropometric data have been maintained within VHA's distributed EMR system for more than a decade in local VistA systems, researchers have only been able to access national extracts of these data through the new CDW within the past 3 years. While the CDW is a potentially important and efficient source of heights and weights, assessing the quality of a novel database is imperative. Since most healthcare systems do not continue to maintain parallel electronic and paper medical charts, for VHA, the local information systems (VistA) are the gold standard, but access can be difficult because of the large number (approximately 130) of independent VistA systems nationwide. Therefore, comparison of the national CDW with regional VISN warehouses that also use the local VistA systems as their source can provide indirect evidence as to whether the national CDW accurately reflects data stored in local VistA systems. That is, derivative repositories drawing on the same source should reflect the source without variance. Variance, if found, could suggest data quality problems in one or both of the derivative repositories.

This article draws on two projects that assessed the consistency and quality of the data in the CDW. The first project was an administrative project directed by the VHA Support Service Center (VSSC). The VSSC monitors key indicators of the quality, quantity, and cost of VHA patient care, as well as compliance with clinical guidelines as part of VHA's ongoing examination of performance measures. Because the usefulness of VSSC's work depends on the accuracy of the underlying administrative and clinical data that contribute to VHA's quality and performance measures, the VSSC evaluated the CDW data when they became available. This included an examination of anthropometric data relevant to VHA's recent initiatives for preventing and managing obesity in primary care [16].

The second project was a retrospective cohort research study funded by the VA Health Services Research and Development Service (HSR&D) to examine obesity care practices within VHA. Data required for this study included height and weight to define a cohort of obese patients and track BMI outcomes [13]. Given the novelty of the CDW data, an assessment of data completeness and accuracy was also seen as essential to ensure the integrity of the research. The complementary nature of the goals and methods used in the two projects was recognized early, resulting in, first, a workshop presented to VA HSR&D researchers in 2008 and, subsequently, this article [17].

METHODS

Data Sources

We examined data completeness and quality at the macro level using anthropometric data from the CDW and inpatient and outpatient utilization data from VHA's national Medical SAS Data Sets for all VHA patients for FY2004 to FY2007. In addition, we examined anthropometric measures for patients in one VISN using data from the CDW, one VISN data warehouse, and one VistA system from January 1, 2004, to December 31, 2007.

Individual and record-level comparisons were based on anthropometric data from the CDW and four to six VISN warehouses. Analyses included data from all FY2002 primary care patients in the selected VISNs. Longitudinal data (FY2004, FY2006) were only collected

from those primary care patients who were identified as obese in FY2002 and who continued to receive VA care throughout the 4-year follow-up period.

Measures

Anthropometric measures included heights, weights, and waist circumferences. We obtained these from the numeric fields (not the text fields) in the CDW and VISN warehouses, derived from the VistA EMR files as described earlier. Each of these variables is recorded as a numerical value representing inches (for height, waist circumference) or pounds (for weight), unless otherwise noted. Prior to examination of these values, data were cleaned, eliminating records with any nonnumerical characters. For each measure, we also obtained the date, time, and facility where it was recorded. We also used Social Security numbers and scrambled identifiers to link files from the various sources to facilitate comparing data sources at the patient level. Utilization data consisted of the number of outpatient visits, inpatient bed days, and unique VHA patients from FYs 2004–2007.

"Biologically implausible" values for the adult veteran population were defined as heights <48 or >84 in. or weights <75 or >700 lb, as recommended by Das et al. [10]. To compare the number of measurements in the national (CDW) versus regional VISN databases, we identified the number of unique individuals who appeared in the two data sources (national and regional) for each FY by VISN. We then assigned each individual to one of four categories. Those who had—

1. The same number of measurements in both data sources.
2. No recorded measurements in either data source.
3. A greater number of measurements in the national database.
4. A greater number of measurements in the VISN database.

We defined matches as those individuals who either had an equal number of measurements in both data sources or no recorded measurements in either data source.

We calculated the differences between the minimum and maximum values recorded for individuals who had two or more heights or two or more weights recorded on the same day and within the same year. We assigned the height and weight differences into five categories each to estimate the extent to which any variation in the data reflected probable biological changes as opposed to ones that seemed more implausible or impossible. The height difference categories ranged from 0 to >10 in., while the weight difference categories ranged from 0 to >1,000 lb.

Analysis at Macro Level

To achieve our aim at the macro level, we examined (1) whether weight, height, and waist circumference data fields were populated as expected compared with one another in the CDW and compared with overall patient utilization and (2) the extent to which the CDW contained biologically implausible values. We first tallied the total number of height, weight, and waist circumference records in the national CDW for each FY from 2004 to 2007 to identify change in the number of records by data type and by year. We then compared changes in the number of measurements by type and year with overall VHA utilization data during the same years to identify general inconsistencies. Next, we examined the frequency distributions of height and weight values recorded in the CDW in FY2007 to identify the percentage of height or weight records that were biologically implausible or that made no clinical sense.

In addition, we compared height and weight data of 10 facilities within one VISN recorded in calendar years 2005 through 2007 using that VISN's warehouse and the CDW. The number of height and weight records, as well as the height and weight values, was compared. Differences between the two sources were examined by date, time, and measurement type for any patterns that could reveal systematic data quality issues. Furthermore, when inconsistencies were discovered, we consulted the EMR for these 10 facilities to confirm the actual data recorded and the potential cause of the anomaly.

Analysis at Individual or Record Level

To achieve our aim at the individual level, we evaluated the data's completeness and accuracy by (1) comparing the number and values of weights and heights recorded in the CDW and VISN warehouses and (2) examining implausible variation in repeated measurements recorded in the CDW on the same day or within the same year for the same individuals.

To compare the number of measurements, we identified the number of unique individuals who had matches (either the same number of measurements or no recorded measurements) in both the CDW and one of four VISN data warehouses for each of three FYs (2002, 2004, 2006) by data type. To compare measurement values, we identified for each of the three same FYs the number of occurrences in which either a weight or a height was recorded on the same date and at the same facility in both data sources (i.e., CDW and one of four data warehouses)

for the same individuals. For these, we calculated the percentages of weights and heights that were identical or had exactly the same value recorded, versus the percentage of weight or height values that were discrepant.

To examine implausible variation in repeated measurements, we only used national CDW data from the same three FYs. We calculated the differences between the minimum and maximum values recorded for individuals who had two or more heights or two or more weights recorded on the same day and within the same year and calculated the percentage that fell into each of the five maximum-minimum difference categories.

RESULTS

Table 1 displays counts of height, weight, and circumference/girth records within the CDW, FYs 2004 to 2007. The number of weight records increased slightly but with a decreasing rate over the target years, whereas the number of height records decreased at an increasing rate over the same period. This occurred in spite of an increased number of patients, increased outpatient utilization, and increased emphasis in VHA to monitor and reduce obesity. Although circumference/girth records increased by large percentages yearly, the field remained grossly underpopulated compared with height and weight. As shown in **Table 2**, <1 percent of height and weight values stored in the CDW in FY2007 fell into the biologically implausible ranges.

The comparison of data in the CDW and the VISN warehouse indicated that the CDW held 2,299,409 height and 4,302,285 weight records for the 10 facilities during the 3-year study period, while the VISN warehouse records included 2,301,615 heights and 4,306,573 weights. Throughout the analysis, these numbers varied slightly because of the regular refresh rates of the VISN and CDW repositories. Looking at monthly data, we found that the difference in the number of height records in the CDW versus the VISN warehouse was generally quite small at <1 percent in almost every month throughout the target period (**Figure**). Notably, starting in mid-2006 and coincident with initiation of a new CDW data extraction method, the CDW began showing consistently fewer records than the VISN warehouses. The same phenomenon

Table 1.
Patient utilization and number of anthropometric records in Corporate Data Warehouse and percent change from prior fiscal year.

Variable	2004 (*n*)	2005		2006		2007	
		n	% Change from Prior Yr	*n*	% Change from Prior Yr	*n*	% Change from Prior Yr
Anthropometric Measure							
Weight	14,764,754	15,258,657	3.35	15,490,210	1.52	15,497,385	0.05
Height	8,534,729	8,521,504	–0.15	7,435,312	–12.75	6,266,231	–15.72
Circumference/Girth	11,004	16,933	53.88	82,482	387.11	125,577	52.25
Utilization							
Outpatient Visits	49,966,268	53,342,682	4.76	53,381,153	1.98	55,704,314	4.35
Inpatient Bed Days	11,561,822	11,215,881	—	10,685,422	—	10,701,159	—
Unique Patients	4,976,773	5,094,494	2.37	5,188,836	1.85	5,230,452	0.80

Table 2.
Number and percentage of Corporate Data Warehouse weight and height records within, below, and above biologically plausible ranges in fiscal year 2007.

Anthropometric Measure	Total (*n*)	Records with Values Within Expected Range, *n* (%)	Records with Values Below Expected Range, *n* (%)	Records with Values Above Expected Range, *n* (%)
Weights	15,449,744	15,385,713 (99.6)	22,397 (0.1)	1,267 (0.0)
Heights	6,312,972	6,274,674 (99.4)	5,110 (0.1)	3,576 (0.1)

Note: Biologically plausible range is defined as weights 75–700 lb and heights 48–84 in. for adult veterans population; missing data not reported.

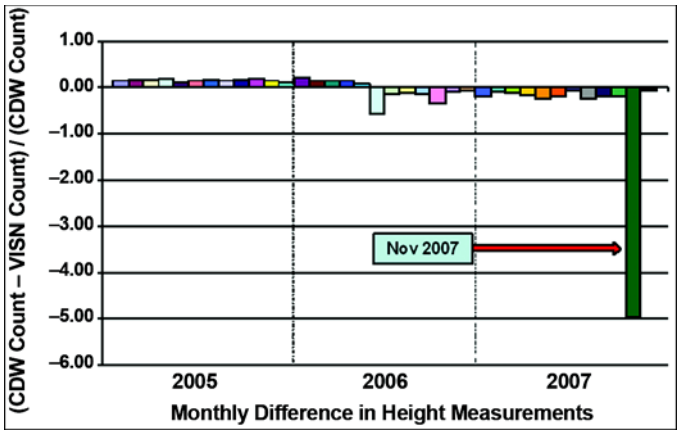

Figure.
Percent difference by month (distinguished by color), Corporate Data Warehouse (CDW) vs Veterans Integrated Service Network (VISN) data warehouse height measurement records, calendar year 2005–2007. Arrow indicates substantial decline in CDW records, Nov 2007.

was also observed for weight records (not shown). However, a significant anomaly appeared in November 2007, which resulted in nearly 5 percent fewer records for the month in the CDW because of a known transmission failure in multiple facilities. Although the percent difference was generally quite low between the two derivative data sources, we consulted the EMR to confirm the actual recording of the data and potential causes of the anomaly. For example, we discovered that measurements recorded in the last minute of a day in the EMR were date-stamped as the next day in the CDW; this resulted from a CDW process, since amended, which rounded up the seconds portion of the time stamp. Some of these differences could also have been due to a method of CDW data refreshment that caused some records to be dropped from or altered in the HDR and CDW. These and other issues, identified through the administrative analysis, have since been rectified.

For individuals found in both the national CDW and one of four VISN regional data sources, the percentage of those who had "matches" (equal numbers of weight records + none in both) in the CDW and the VISN warehouse ranged by VISN from 62.6 to 99.7 percent in FY2002; 98.6 to 99.5 percent in FY2004, and 97.6 to 98.6 percent in FY 2006 (**Table 3**). The percentage of those with matches in the number of heights ranged by VISN from 73.0 to 99.8 percent in FY2002, 98.6 to 99.6 percent in FY2004, and 98.7 to 99.5 percent in FY 2006

Table 3.
Comparison of number and percentage of weight records for unique patients who appeared in both Corporate Data Warehouse (CDW) and Veterans Integrated Service Network (VISN)-level data sources (A, C, E, and F) for fiscal years (FYs) 2002, 2004, and 2006.

Concordance	VISN A, *n* (%)	VISN C, *n* (%)	VISN E, *n* (%)	VISN F, *n* (%)
FY2002	***n* = 97,375**	***n* = 81,125**	***n* = 106,010**	***n* = 158,088**
Equal No. of Weights	96,033 (98.62)	50,563 (62.33)	105,236 (99.27)	157,365 (99.54)
None in Both	266 (0.27)	203 (0.25)	234 (0.22)	200 (0.13)
> in CDW Extract	1,074 (1.11)	30,351 (37.41)	531 (0.50)	355 (0.22)
> in VISN Extract	2 (0.00)	8 (0.01)	9 (0.01)	168 (0.11)
FY2004	***n* = 34,815**	***n* = 31,988**	***n* = 47,572**	***n* = 77,233**
Equal No. of Weights	34,208 (98.26)	31,496 (98.46)	47,187 (99.19)	76,766 (99.40)
None in Both	53 (0.15)	45 (0.14)	95 (0.20)	97 (0.13)
> in CDW Extract	553 (1.59)	431 (1.34)	285 (0.60)	363 (0.47)
> in VISN Extract	1 (0.00)	16 (0.05)	5 (0.01)	7 (0.01)
FY2006	***n* = 30,812**	***n* = 28,289**	***n* = 41,944**	***n* = 68,592**
Equal No. of Weights	30,046 (97.51)	27,579 (97.49)	41,275 (98.41)	67,373 (98.22)
None in Both	55 (0.18)	42 (0.15)	60 (0.14)	89 (0.13)
> in CDW Extract	539 (1.75)	297 (1.05)	315 (0.75)	298 (0.43)
> in VISN Extract	172 (0.55)	371 (1.31)	294 (0.70)	832 (1.21)

Note: FY2002 data based on primary care patients with one or more visits in FY2002. FY2004 and FY2006 data based on primary care patients identified as obese in FY2002.
No. = number.

(**Table 4**). The lowest percentage for both weights (62.6%) and heights (73.0%) occurred in the same VISN in FY2002. Of individuals from this VISN who were found in both data sources in FY2002, 37.4 percent (n = 30,315) had more weights and 27.0 percent (n = 21,934) had more heights recorded in the CDW, as compared with 0.01 percent (n = 8) and <0.01 percent (n = 3) who had more weights and heights, respectively, recorded in the VISN data warehouse. Further examination of the data indicated that the discrepancy resulted from missing data of five facilities extracted from this VISN's regional warehouse. We were not able to clarify, from the programmers who performed the extraction, whether the data were actually missing from the data warehouse or had been inadvertently omitted during creation of the data file for this study.

Among the weights and heights recorded the same day in the same facility for the same individuals in both data sources, the percentage of discordant values was <1 percent across years in all VISNs. In the VISN with the lowest concordance (99.7%) in FY2002, 1,260 of 387,138 weight or height records were not identical. In most of these cases (n = 1,230), the value recorded in the CDW was larger than the value recorded in the VISN data source. For 789, the CDW values were larger by ≤1 unit (inch or pound), suggesting rounding error. The remaining 441 cases of discordant values were larger by >1 in. or 1 lb. The value recorded in the VISN data source was larger than the value for the same person in the CDW in only 30 of the cases.

Among 105,425 occurrences recorded in the CDW in which patients had two or more weights recorded on the same day in FY2002 in six VISNs, the majority (55.6%) had identical values (**Table 5**). The remaining 44.4 percent were discrepant. Approximately 34.9 percent reflected differences between the minimum and maximum values of ≤10 lb, 8.1 percent had differences that ranged from >10 to 100 lb, and 1.4 percent had differences of >100 to ≥1,000 lb. Similar patterns were seen for FY2004 and FY2006. For all three FYs, >90 percent of the occurrences in which patients had two or more heights measured on the same day had differences of ≤1 in., while 4.7 to 5.9 percent of the occurrences had differences of 2 to ≥10 in.

A similar pattern was found for individuals who had two or more weights or two or more heights recorded in the same year (for FYs 2002, 2004, and 2006). Although

Table 4.
Comparison of number and percentage of height records for unique patients who appeared in both Corporate Data Warehouse (CDW) and Veterans Integrated Service Network (VISN)-level data sources (A, C, E, and F) for fiscal years (FYs) 2002, 2004, and 2006 by VISN.

Concordance	VISN A, *n* (%)	VISN C, *n* (%)	VISN E, *n* (%)	VISN F, *n* (%)
FY2002	***n* = 97,375**	***n* = 81,125**	***n* = 106,010**	***n* = 158,088**
Equal No. in Both	91,302 (93.76)	51,153 (63.05)	85,147 (80.32)	113,890 (72.04)
None in Both	5,138 (5.28)	8,035 (9.90)	20,564 (19.40)	43,869 (27.75)
> in CDW Extract	935 (0.96)	21,934 (27.04)	297 (0.28)	317 (0.20)
> in VISN Extract	0 (0.00)	3 (0.00)	2 (0.00)	12 (0.01)
FY2004	***n* = 34,815**	***n* = 31,988**	***n* = 47,572**	***n* = 77,233**
Equal No. in Both	32,347 (92.91)	25,660 (80.22)	30,482 (64.08)	43,630 (56.49)
None in Both	1,983 (5.70)	6,083 (19.02)	16,903 (35.53)	33,302 (43.12)
> in CDW Extract	483 (1.39)	242 (0.76)	186 (0.40)	298 (0.38)
> in VISN Extract	2 (0.01)	3 (0.01)	1 (0.00)	3 (0.00)
FY2006	***n* = 30,812**	***n* = 28,289**	***n* = 41,944**	***n* = 68,592**
Equal No. in Both	28,038 (91.00)	21,951 (77.60)	24,584 (58.61)	33,943 (49.49)
None in Both	2,294 (7.45)	5,968 (21.10)	17,152 (40.89)	34,187 (49.84)
> in CDW Extract	442 (1.44)	224 (0.80)	164 (0.39)	259 (0.38)
> in VistA Extract	38 (0.12)	146 (0.51)	44 (0.11)	203 (0.30)

Note: FY2002 data based on primary care patients with one or more visits in FY2002. FY2004 and FY2006 data based on primary care patients identified as obese in FY2002. In addition, cells not summing to 100% is due to rounding errors.
No. = number, VistA = Veterans Health Information Systems and Technology Architecture.

Table 5.
Frequency distribution of differences in minimum and maximum Corporate Data Warehouse weights (pounds) and heights (inches) and number and percent of occurrences in which individual patient had more than two recorded on same day by fiscal years (FYs) 2002, 2004, and 2006.

Range of Difference of Minimum & Maximum Values	Occurrences of ≥2 Weights Recorded, *n* (%)		
	FY2002, *n* = 105,425	FY2004, *n* = 87,532	FY2006, *n* = 107,015
Pounds			
Difference = 0	58,596 (55.58)	37,666 (43.03)	44,155 (41.26)
0 < Difference ≤ 10	36,775 (34.88)	39,613 (45.26)	50,916 (47.58)
10 < Difference ≤ 100	8,546 (8.11)	8,591 (9.81)	9,590 (8.96)
100 < Difference ≤ 1,000	1,465 (1.39)	1,615 (1.85)	2,314 (2.16)
Difference > 1,000	43 (0.04)	47 (0.05)	40 (0.04)
	Occurrences of ≥2 Heights Recorded, *n* (%)		
Inches	FY2002, *n* = 64,465*	FY2004, *n* = 34,359	FY2006, *n* = 33,424
Difference = 0	56,949 (88.34)	29,299 (85.27)	29,319 (87.72)
0 < Difference ≤ 1	3,198 (4.96)	2,147 (6.25)	1,771 (5.30)
1 < Difference ≤ 2	1,273 (1.97)	895 (2.62)	764 (2.28)
2 < Difference ≤ 10	2,115 (3.28)	1,374 (4.00)	1,207 (3.61)
Difference > 10	930 (1.44)	640 (1.86)	363 (1.09)

Note: Includes all primary care patients from six VISNs identified in FY2002 who continued to receive care in FY2004 and FY2006.
*Cells not summing to 100% is due to rounding.

the majority had values that did not differ or had different values that were within the realm of plausibility, approximately 1 to 2 percent had values that were suspect or clearly implausible. For example, in FY2006, 6,271 individuals had two or more weights recorded that differed by >100 to 1,000 lb, while 176 had weights that differed by >1,000 lb. Even more suspect variation was found among heights; e.g., in FY2006, 11,063 individuals had two or more recorded heights that differed by > 2 to 10 in., while an additional 2,712 individuals had heights that differed by >10 in.

DISCUSSION

CDW anthropometric data present opportunities and challenges for health services researchers. In spite of occasional anomalies, our work suggests that the national CDW generally appears to reflect weight and height data stored in VISN warehouses and thus, presumably, data stored in the VistA systems (gold standard). The concordance between the number and values of recorded heights and weights stored in both the CDW and the five different VISN data warehouses examined in the administrative and research projects was generally 97 to 99 percent. Moreover, several data anomalies identified by the administrative project have since been corrected, further enhancing data quality. Since the national data in the CDW and the regional data in the VISN warehouses are both drawn from the same local EMR VistA systems, these findings provide indirect support to suggest that the national CDW has become a reliable source of data in VHA's local information systems.

Use of the national CDW as a data source, even if only regional data are desired, will help avoid idiosyncrasies of local programming and extraction, which simplifies data cleaning and database development. Moreover, use of the national CDW will allow researchers to obtain all data using a single extraction, which usually helps avoid multiple institutional review board (IRB) applications. Despite these advantages, our work also illuminated challenges presented by the data in the CDW and the clinical practices it reflects, as well as challenges inherent in conducting obesity research with administrative data.

Challenges

The first challenge was that significantly more weights were recorded in the national and regional data sources than heights, a finding that has been noted elsewhere and probably reflects clinical practice [11–13]. Failure to record heights as frequently as weights could be due to a number of factors, including lack of proper equipment, low perceived importance, time constraints,

and competing clinical demands [11–12,18]. Heights may also be less likely to be measured and recorded for specific patient populations, such as amputees or those who are wheelchair-bound.

The failure to record heights as frequently as weights, however, may make some patients' BMI impossible to calculate within a specific time frame, especially for cross-sectional studies, or to accurately track BMI over time. Although height may be recorded less frequently than weight in adults because it is viewed as relatively stable, a systematic review of epidemiological studies of longitudinal height change suggests that after age 30, a person's rate of height loss increases with increasing age, such that by the age of 80 years, the average man will have lost approximately 5.0 cm from his maximum height and the average woman approximately 6.2 cm [19]. Therefore, failure to periodically record *both* height and weight may pose problems for certain types of research studies (e.g., osteoporosis in very elderly patients). In addition, anecdotal reports from our critical care medicine colleagues indicate that they sometimes must estimate height when it is missing from the EMR, because they use BMI to accurately dose some medications, such as anesthesia, and to calculate ventilation unit parameters in the intensive care unit. Although unlikely to be common, estimation of anthropometric measurements is another potential source of error that may be present in the data.

Furthermore, in spite of guideline suggestions to assess waist circumference, it was substantially less likely to be recorded than heights or weights. While having these data available would be helpful in examining central adiposity, prior studies have found that waist circumference tends to have more measurement error than other anthropometric measures [20]. Regardless, researchers should be aware of the relative incompleteness of height and waist circumference data in the CDW, which may limit its use for specific types of obesity research until clinical recording of these measurements improves.

A second challenge presented by CDW and/or VISN warehouse data was errors. First, we found examples of missing data for specific facilities or certain periods of time, such as occurred in November 2007. Second, we found some biologically implausible values, as well as some biologically improbable variation in heights and weights. Several reasons exist as to why such errors might occur. Routinely collected clinical information, including heights and weights, is typically assessed and directly hand-entered into the user interface of the EMR by clinical staff. These data may undergo further manipulation when subsequently transferred into a data repository and extracted for analysis. Errors may occur at several points during this process, including measurement, data entry, and transfer.

For example, measurement error may occur if equipment is incorrectly calibrated, if height or weight values are "rounded up," if patients are inconsistently measured with or without shoes over time, or if healthcare providers rely on self-reported weights or heights. More systematic measurement error can occur if clinicians are more likely to measure heights and weights in specific populations, such as the obese. Data-entry errors can occur if numbers are transposed or deleted, if extra numbers are added inadvertently, or if numbers on a keypad adjacent to the intended target are accidently keyed. During the data transfer or extraction process, numeric data can be redefined as character data by a data transfer program or by a warehouse. Rounding errors may arise when numeric data are stored as character variables with a fixed number of decimal values. Specific to VHA, the VISN warehouses and the CDW "refresh" themselves on different schedules and through different processes; therefore, cases may be temporarily in one warehouse and not the other.

While assessing the completeness and accuracy of secondary height and weight data is important, differentiating errors such as these from "true" biological variability that is inherent in repeated measurements of weights and heights can be challenging. An adult's weight and even height may vary slightly over a 24-hour period. Weight losses or gains may occur over weeks, months, and years because of changes in energy balance. Moderate to dramatic changes in weight and height can occur abruptly (e.g., due to surgery, traumatic injury) or gradually due to disease or the aging process. At least some of the more dramatic variations in height and weight values identified in our data during the same day, and even within the same year, for the same individuals suggested that some of the more improbable cases of variation were due to measurement or data-entry errors.

In smaller clinical or epidemiological studies, it might be feasible for researchers to examine individual data trends on a case-by-case basis to identify and eliminate obvious outliers or improbable data patterns [21]. However, this is usually not possible in a large database study. The massive volume of data that is typically available

limits the capacity to develop algorithms to eliminate errors. From our work, however, we identified strategies to control for, or minimize, the impact of the some of the identified challenges.

Strategies to Address Challenges in Future Research

To check for the possibility of errors introduced by data transmission failures or data extraction process, researchers need to test each data file supplied by a data repository for face validity. That is, do trends in the number of records over time make sense? Are there particular days or months for which no data are included? Do the monthly record counts vary widely by healthcare facility or by data element of interest? Do the values make clinical sense?

To reduce obvious errors, after deleting text entries in our height and weight data, we used the "trimming" procedure recommended by Das et al. in our research project to eliminate biologically implausible values that were recorded for this adult population of veterans (i.e., any weights ≤75 lb or ≥700 lb and any heights ≤48 in. or ≥84 in.) [10]. Although a few individuals may have heights and/or weights outside these ranges, only a small number of patients (n = 70) in our cohort were eliminated because they only had biologically implausible values [13]. Depending on the population being studied, however, researchers should select height and weight ranges that make sense with their populations (e.g., patients with cancer).

Based on our findings of extreme variance from repeated measurements obtained on the same day or within the same year, however, we know that some problematic data remained. To help control for this, we chose to divide each year of our 5-year observation period into quarters and to use the median weight within each quarter. Depending on the research question or population of interest, a researcher might choose instead to delete same-day measurements altogether or even choose to delete patient groups likely to have multiple measurements on the same day, such as patients on dialysis or with congestive heart failure or patients who would likely have a large weight loss or weight gain because of a medical procedure (e.g., leg amputation). Similarly, researchers may also choose to eliminate inpatient measurements, because weight fluctuations may be confounded by acute illness.

Possible strategies for addressing *missing* height or weight data vary, depending on whether a single "baseline" or point estimate of BMI as a control or covariate is needed or whether repeated measures of BMI are required. If a single baseline BMI is needed, a good chance exists that a height may not have been recorded on the same day as the first available weight or even within the same year, thus requiring the use of a height from a later time within the year of interest, or even an adjacent year. Since our retrospective cohort study required the use of repeated measures of BMI over time, we chose to use the modal, or the most frequently occurring, height across the 5-year study period for calculating not only our baseline BMI but also BMI across the yearly quarters. Unfortunately, 9,382 of potentially eligible patients in our cohort study had two or more modal heights across the 5-year period. We chose to average the modes if the differences between them were ≤3 in. Otherwise, cases (n = 1,750) in which the difference was ≥3 in. were deleted [13].

Technical or Policy Changes to Improve Data Completeness and Quality

Healthcare managers and policy makers can also have important roles in improving the completeness and quality of anthropometric data. As just noted, our data suggest that some of the height and weight values in the CDW reflect probable data-entry errors. Our data did not clearly show to what extent extreme weight or height values entered on the same day reflected explicit attempts by clinical staff to correct recognized data-entry errors. A "qualifier" field exists for free text comments associated with the anthropometric measures in the EMR, but it is not currently available in the CDW [15]. Although the data fields in VHA's EMR have "filters" or range checks to prevent the inadvertent entry of erroneous values, the ranges used for the weight and height data fields are so extreme that they still allow substantial room for errors. For example, the current range checks allow the entry of weights between 0 and 1,500 lb and of heights between 10 and 100 in. Modification of the range checks to exclude more implausible values or, at the very least, to program a query for the clinician to confirm extreme values, could help eliminate the most egregious errors.

Finally, the problem of relatively fewer heights and waist circumference values being recorded in the EMR may best be resolved by clinical reminders. VHA has been a leader in the use of performance measures and clinical reminders to foster evidence-based healthcare [22]. As part of its recent implementation of its new

nationwide program for managing obesity called MOVE! (Managing Overweight/Obesity in Veterans Everywhere!), VHA has introduced a "supporting" or pilot performance indicator for obesity screening and sample clinical reminders for use in local EMRs [16]. Local variability in their use will help identify best practices. Future implementation at the national level may help to better populate the data fields for anthropometric measurements.

CONCLUSIONS

Given recent decisions by Medicare and other healthcare organizations to define obesity as a disorder, reimburse for obesity-related treatments, and develop obesity-related performance measures, trends and variations in obesity-related practices are increasingly relevant to healthcare providers [23–24]. Thus, accuracy of administrative anthropometric data has important implications for not only obesity research but also reimbursement and assessment of quality of care. Our work suggests that the CDW has become an important and generally reliable source of weight and height data for VHA patients, and it can be effectively used for research within its recognized limitations.

ACKNOWLEDGMENTS

Author Contributions:
Study concept and design: P. H. Noël, A. E. Lancaster.
Acquisition of data: P. H. Noël, A. E. Lancaster, L. A. Copeland.
Analysis and interpretation of data: P. H. Noël, L. A. Copeland, M. J. Pugh, H. P. Hazuda, A. E. Lancaster, R. A. Perrin, C-P. Wang, M. J. Bollinger.
Drafting of manuscript: P. H. Noël, R. A. Perrin.
Critical revision of manuscript for important intellectual content: P. H. Noël, L. A. Copeland, R. A. Perrin, A. E. Lancaster, M. J. Pugh, H. P. Hazuda, M. J. Bollinger.
Statistical analysis: C-P. Wang, A. E. Lancaster, L. A. Copeland, M. J. Bollinger.
Obtained funding: P. H. Noël.
Administrative, technical, or material support: R. A. Perrin, M. J. Bollinger.
Study supervision: P. H. Noël, A. E. Lancaster.
Financial Disclosures: The authors have declared that no competing interests exist.
Funding/Support: This material was based on work supported by the VA, VHA, and HSR&D, grant IIR 05-121. Dr. Noël is a Research Psychologist at the South Texas Veterans Health Care System. The views expressed in this article are those of the authors and do not necessarily represent the views of the VA.
Additional Contributions: We would like to thank Jack Bates, CDW Manager, and Stephen Anderson, CDW Chief Architect, for generously sharing their expertise on the development and potential of VHA CDW.
Institutional Review: We obtained data for the research portion of this report after IRB approval; the administrative project was for quality assurance.

REFERENCES

1. NHLBI Obesity Education Initiative. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: The evidence report. Washington (DC): National Institutes of Health; 1998.
2. National Task Force on the Prevention and Treatment of Obesity. Overweight, obesity, and health risk. Arch Intern Med. 2000;160(7):898–904. [PMID: 10761953]
3. Nawaz H, Katz DL. American College of Preventive Medicine Practice Policy statement. Weight management counseling of overweight adults. Am J Prev Med. 2001;21(1):73–78. [PMID: 11418263] DOI:10.1016/S0749-3797(01)00317-8
4. Lyznicki JM, Young DC, Riggs JA, Davis RM; Council on Scientific Affairs, American Medical Association. Obesity: Assessment and management in primary care. Am Fam Physician. 2001;63(11):2185–96. [PMID: 11417771]
5. National Task Force on the Prevention and Treatment of Obesity. Medical care for obese patients: Advice for health care professionals. Am Fam Physician. 2002;65(1):81–88. [PMID: 11804445]
6. McTigue KM, Harris R, Hemphill B, Luxl L, Sutton S, Bunton AJ, Lohr KN. Screening and interventions for obesity in adults: Summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med. 2003;139(11):933–49. [PMID: 14644897]
7. Ells LJ, Langr, Shield JP, Wilkinson JR, Lidstone JS, Coulton S, Summerbell CD. Obesity and disability—A short review. Obes Rev. 2006;7(4):341–45. [PMID: 17038128] DOI:10.1111/j.1467-789X.2006.00233.x
8. Beck LA, Lovlien CA, Twedell DM. Care of morbidly obese people with spinal cord injury. J Trauma Nurs. 1996;3(4):98–107. [PMID: 9214978]
9. Goodell TT. The obese trauma patient: Treatment strategies. J Trauma Nurs. 1996;3(2):36–44. [PMID: 9025456]
10. Das SR, Kinsinger LS, Yancy WS Jr, Wang A, Ciesco E, Burdick M, Yevich SJ. Obesity prevalence among veterans at Veterans Affairs medical facilities. Am J Prev Med. 2005;28(3):291–94. [PMID: 15766618] DOI:10.1016/j.amepre.2004.12.007
11. Stafford RS, Farhat JH, Misra B, Schoenfeld DA. National patterns of physician activities related to obesity management.

Arch Fam Med. 2000;9(7):631–38. [PMID: 10910311] DOI:10.1001/archfami.9.7.631

12. McAlpine DD, Wilson AR. Trends in obesity-related counseling in primary care: 1995–2004. Med Care. 2007;45(4): 322–29. [PMID: 17496716] DOI:10.1097/01.mlr.0000254575.19543.01
13. Noël PH, Copeland LA, Pugh MJ, Kahwati L, Tsevat J, Nelson K, Wang CP, Bollinger MJ, Hazuda HP. Obesity diagnosis and care practices in the Veterans Health Administration. J Gen Intern Med. 2010;25(6):510–16. Epub 2010 Feb 24. [PMID: 20180155]
14. Lyratzopoulos G, Heller RF, Hanily M, Lewis PS. Risk factor measurement quality in primary care routine data was variable but nondifferential between individuals. J Clin Epidemiol. 2008;61(3):261–67. [PMID: 18226749] DOI:10.1016/j.jclinepi.2007.05.020
15. VHA Corporate Data Warehouse [Internet]. Washington (DC): Department of Veterans Affairs; 2009 [updated 2009 Sep 30; cited 2010 Jan 28]. Available from: http://www.virec.research.va.gov/DataSourcesName/CDW/CDW.htm/.
16. Kinsinger LS, Jones KR, Kahwati L, Harvey R, Burdick M, Zele V, Yevich SJ. Design and dissemination of the MOVE! Weight-Management Program for Veterans. Prev Chronic Dis. 2009;6(3):A98. [PMID: 19527600]
17. Perrin R, Bates J, Noel PH, Copeland LA, Lancaster B. National clinical data for VA Research: The VA Corporate Data Warehouse. HSR&D National Meeting; 2008 Feb 13–15; Baltimore, MD. Washington (DC): Department of Veterans Affairs; 2008.
18. Ruelaz AR, Diefenbach P, Simon B, Lanto A, Arterburn D, Shekelle PG. Perceived barriers to weight management in primary care—Perspectives of patients and providers. J Gen Intern Med. 2007;22(4):518–22. [PMID: 17372803] DOI:10.1007/s11606-007-0125-4
19. Sorkin JD, Muller DC, Andres R. Longitudinal change in the heights of men and women: Consequential effects on body mass index. Epidemiol Rev. 1999;21(2):247–60. [PMID: 10682261]
20. Ulijaszek SJ, Kerr DA. Anthropometric measurement error and the assessment of nutritional status. Br J Nutr. 1999; 82(3):165–77. [PMID: 10655963] DOI:10.1017/S0007114599001348
21. Harrist RB, Dai S. Analytic methods in Project HeartBeat! Am J Prev Med. 2009;37(1 Suppl):S17–24. [PMID: 19524151] DOI:10.1016/j.amepre.2009.04.004
22. Kupersmith J, Francis J, Kerr E, Krein S, Pogach L, Kolodner RM, Perlin JB. Advancing evidence-based care for diabetes: Lessons from the Veterans Health Administration. Health Aff (Millwood). 2007; 26(2):w156–68. Epub 2007 Jan 26. [PMID: 17259199] DOI:10.1377/hlthaff.26.2.w156
23. News Release. HSS announces revised Medicare obesity coverage policy [Internet]. Washington (DC): U.S. Department of Health & Human Services; 2004. Available from: http://archive.hhs.gov/news/press/2004pres/20040715.html/.
24. National Committee for Quality Assurance. Proposed new measures for HEDIS 2009: Body Mass Index (BMI) Assessment (BAA) BMI Percentile Assessment and Counseling for Nutrition and Physical Activity (BCA) [Internet]. NCQA 2008. Available from: http://www.ncqa.org/Portals/0/PublicComment/HEDIS2009/Obesity_Memo_PDF.pdf/.

Submitted for publication August 3, 2009. Accepted in revised form March 8, 2010.

This article and any supplementary material should be cited as follows:
Noël PH, Copeland LA, Perrin RA, Lancaster AE, Pugh MJ, Wang C-P, Bollinger MJ, Hazuda HP. VHA Corporate Data Warehouse height and weight data: Opportunities and challenges for health services research. J Rehabil Res Dev. 2010;47(8):793–50.
DOI:10.1682/JRRD.2009.08.0110