Economics and Human Biology 17 (2015) 116–128

Contents lists available at ScienceDirect

Economics and Human Biology

journal homepage: http://www.elsevier.com/locate/ehb

Measuring obesity in the absence of a gold standard☆

Donal O'Neill *

Deptartment of Economics, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

ARTICLE INFO

Article history:
Received 25 August 2014
Received in revised form 4 February 2015
Accepted 9 February 2015
Available online 16 February 2015

Keywords:
Obesity
Multiple diagnostic tests
Latent class analysis

ABSTRACT

Reliable measures of body composition are essential to develop effective policies to tackle obesity. The lack of an acceptable gold-standard for measuring fatness has made it difficult to evaluate alternative measures of obesity. We use latent class analysis to characterise existing diagnostics. Using data on US adults we show that measures based on body mass index and bioelectrical impedance analysis misclassify large numbers of individuals. For example, 45% of obese White women are misclassified as non-obese using body mass index, while over 50% of non-obese White women are misclassified as being obese using bioelectrical impedance analysis. In contrast the misclassification rates are low when waist circumference is used to measure obesity. These results have important implications for our understanding of differences in obesity rates across time and groups, as well as posing challenges for the econometric analysis of obesity.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Obesity is defined as an excessive accumulation and storage of fat in the body and is a leading cause of morbidity, disability and premature death and increases the risk for a wide range of chronic diseases (WHO, 2009; Antonanzas and Rodriguez, 2010; Konnopka et al., 2011). Cawley and Meyerhoefer (2012) estimate that total medical costs of obesity for the full non-institutionalised population of adults aged 18 and older in the U.S. was $190.2 billion in 2005. In 2012 the American Medical Association put a resolution to its delegates asking that obesity be recognised as a disease in the hopes that doing so would change the way the medical community tackles this complex health issue. In the ensuing debate the Council on Science and Public Health (2012) published a report outlining the advantages and disadvantages of such a move. In particular they expressed concerns with existing diagnostic tests of obesity and noted that "if obesity is to be considered a disease, [then] a better measure of obesity than BMI is needed to diagnose individuals in clinical practice."[1] In this paper we draw on work from other areas of biostatistics to propose a new way of evaluating alternative tests for obesity and use our results to make recommendations for the diagnosis and management of obesity.

The traditional and most popular measure of obesity is based on an individual's body mass index (BMI). Despite its widespread use there is a body of research that argues that BMI is, at best, a noisy measure of fatness since it does not distinguish fat from muscle, bone and other lean body mass (Johansson et al., 2009; Burkhauser and Cawley,

☆ I would like to thank Olive Sweetman, Chris Ruhm and seminar participants at NUI Maynooth, the 2014 Irish Economic Association meetings, University of Limerick and the 2014 International Health Economics Association Annual Congress, Trinity College Dublin, as well as three anonymous referees for helpful comments on an earlier draft of this paper. The project uses only use publicly available anonymised data and as such no ethical approval was required.

* Tel.: +353 1 7083555.
E-mail address: donal.oneill@nuim.ie

[1] The motion to classify obesity as a disease was passed in June 2013, though there is still substantial debate and disagreement (see for example the exchange between Heshka and Allison, (2001) and Kopelman and Finer, (2001)).

http://dx.doi.org/10.1016/j.ehb.2015.02.002
1570-677X/© 2015 Elsevier B.V. All rights reserved.

2008; McCarthy et al., 2006; Smalley et al., 1990). Because of these shortcomings in BMI, a World Health Organisation expert consultation on Obesity drew attention to the need for other indicators to complement the measurement of BMI (WHO, 2000). Consequently, a number of alternative measures have been proposed. These include percent body fat estimated using bioelectrical impedance analysis (BIA), measures based on waist circumference (WC), Waist to Hip ratio (WHR) and the ABSI index of body shape[2]. Different approaches to measuring fatness not only yield different rates of obesity (Burkhauser and Cawley, 2008), have different impacts on outcomes (Johansson et al., 2009; Krakauer and Krakauer, 2012; WHO, 2011; Song et al., 2013), but also give rise to different trends in obesity over time (Elobeid et al., 2007; Burkhauser et al., 2009; Ford et al., 2014).

When evaluating alternative measures of fatness the tendency to date has been to settle on a specific, preferred measure as a gold-standard and use this measure to benchmark other diagnostic tests (Smalley et al., 1990; Mei et al., 2002; Burkhauser and Cawley, 2008). For example, using BIA measures as the gold standard Burkhauser and Cawley (2008) find that 61.25% of women classified as non-obese by BMI are false negatives, with no false positives, while for men 14.20% of those classified as obese by BMI are false positives and 33.5% classified as non-obese are false negatives.

In this paper we take a different approach to comparing the accuracy of alternative measures of obesity, motivated by the fact that, a-priori, there is no strong basis for choosing any single measure of obesity as a gold standard. In their survey of alternative measures of obesity Freedman and Perry (2000) note that "The lack of an acceptable gold-standard limits the assessment of the validity of field methods that can be used to estimate body fat." Hu (2008) provides a detailed discussion of the strengths and weaknesses of alternative approaches to the measurement of body composition. Recently developed high-tech imaging options, such as computed tomography and magnetic resonance imaging, offer excellent accuracy and allow researchers to distinguish between visceral and subcutaneous fat, a distinction that is important in helping understand the consequence of obesity. However, Hu (2008) notes that their cost, technical complexity and lack of portability prohibit their routine use in large scale studies. To date the use of these advanced approaches have been limited to small-scale studies[3].

Rather than specifying a gold-standard ex-ante we allow all measures to be potentially imperfect indicators of fatness. When one test is specified as a gold standard evaluating all other possible tests is straightforward. However, in the case where all of the tests are potentially imperfect the task of evaluating diagnostic tests is more difficult because the true underlying disease status of each individual in the study is unknown. However, by treating the true unknown disease status as a latent variable, it is possible to use latent class analysis (LCA) to estimate the true underlying prevalence of the disease along with the characteristics of each of the tests (Walter and Irwig, 1988; Biemer and Wiesen, 2002; Biemer, 2011)[4]. This approach has been used elsewhere in biostatistics, for example when comparing alternative skin tests for the presence of tuberculosis (Hiu and Walter, 1980), comparing diagnosis of myocardial infarction (Rindskopf and Rindskopf, 1986), evaluating diagnostic tests of autism (Szatmari et al., 1995) and malaria (Gonçalves et al., 2012). However, to my knowledge, LCA has not been used before to evaluate alternative measures of obesity.

Using data from a representative sample of US adults I show that that while obesity rates based on BMI and BIA misclassify large numbers of individuals, this is not the case for measures based on WC. The error rates for WC measures of obesity are of the order of 3% compared to error rates as high as 45–70% with BMI and BIA. In particular we show that BMI suffers badly from a high rate of false negative diagnoses while BIA suffers from a high rate of false positives. These results have important implications for differences in obesity rates across time and groups, as well as for traditional econometrics analysis of obesity.

In Section 2 of the paper we discuss latent class modelling in diagnostic testing, while Section 3 discusses the NHANES data used throughout the analysis. Section 4 presents and discusses my key results and Section 5 examines the robustness of these findings. Section 6 concludes.

2. Methods: Latent class models in diagnostic testing

To understand latent class models in diagnostic testing let C_i denote the unobserved or latent variable denoting true obesity status for person i and let T_1, T_2, and T_3 denote three alternative tests designed to measure outcome C. In our application C_i is a dichotomous indicator of the presence or otherwise of true underlying obesity, while T_{1i}, T_{2i}, and T_{3i} are the obesity classification of person i based on each of the three tests. Considering the cross-classification table for the variables C, T_1, T_2, and T_3, let (c, t_1, t_2, t_3) denote the cell associated with $C = c$, $T_1 = t_1$, $T_2 = t_2$, and $T_3 = t_3$ and let π_{c,t_1,t_2,t_3} denote the probability of an observation falling into this cell. Let $\pi_c = \Pr(C = c)$ for

[2] For an overview of these and other alternative approaches to measuring obesity see Hu (2008) and Madden and Smith (2014). The ABSI index was developed by Krakauer and Krakauer (2012) using residuals from a regression of waist circumference on height and weight as the basis for an adjusted WC measure.

[3] Furthermore, even these advanced approaches may suffer from limitations. For example dual energy X-ray absorpitometry cannot accurately distinguish between visceral and subcutaneous fat.

[4] As an alternative to LCA one could consider using factor analysis on the continuous measures. The use of the LCA analysis rather than factor analysis is in keeping with previous work in the area of disease diagnostics and has the advantage of directly evaluating existing and widely adopted thresholds for alternative obesity measures. Nevertheless the use of factor analysis could offer a useful complement to the work presented in this paper. With the LCA approach the thresholds are imposed on the individual measures prior to analysis, whereas with the factor analysis approach the determination of thresholds must be made on the estimated latent distribution after the initial analysis has been completed. We examine the robustness of our results to alternative thresholds later in the paper.

c = 1,0 and $\pi_{t_2|A} = \Pr(T_2 = t_2|A)$ for some event A. For example $\pi_{t_2|t_1,c} = Pr(T_2 = t_2|T_1 = t_1, C = c)$.

Using the law of conditional probabilities

$$\pi_{c,t_1,t_2,t_3} = P(C = c)P(T_1 = t_1|C = c)P(T_2 = t_2|T_1 = t_1, C = c)$$
$$P(T_3 = t_3|T_2 = t_2, T_1 = t_1, C = c) = \pi_c\pi_{t_1|c}\pi_{t_2|t_1,c}\pi_{t_3|t_2,t_1,c}$$

Therefore, the probability that an individual is classified into the cell ($T_1 = t_1$, $T_2 = t_2$, and $T_3 = t_3$) is given by

$$\pi_{t_1,t_2,t_3} = \sum_c \pi_c\pi_{t_1|c}\pi_{t_2|t_1,c}\pi_{t_3|t_2,t_1,c}$$

Let $Y = (Y_{111}, Y_{110}, Y_{100}, Y_{000}, Y_{011}, Y_{101}, Y_{001}, Y_{101})$ be the random vector representing the distribution of our sample across the eight possible cells. So for example Y_{111} is associated with all three tests returning a positive obesity reading. Y is governed by a multinomial distribution:

$$Y\ |\pi \sim \text{multinomial}\ (N,\ (\pi_{111}, \pi_{110}, \pi_{100}, \pi_{000}, \pi_{011}, \pi_{101}, \pi_{001}, \pi_{101}))$$

If we let y_{t_1,t_2,t_3} denote the realised number of observations in cell ($T_1 = t_1$, $T_2 = t_2$, and $T_3 = t_3$) then the kernel of the likelihood of observing the full table {T_1, T_2,T_3} is

$$L(T_1, T_2, T_3) = \prod_{t_1}\prod_{t_2}\prod_{t_3} \pi_{t_1,t_2,t_3}^{y_{t_1,t_2,t_3}}$$

In total we have eight possible cells, but since the number in each of the cells must add up to the total sample size N, we only have seven free pieces of information. Unfortunately in this model there are fifteen parameters to estimate:

$$\pi_1, \pi_{t_1=1|c=1}, \pi_{t_1=1|c=0}, \pi_{t_2=1|t_1=1,c=1}\pi_{t_2=1|t_1=1,c=0}\pi_{t_2=1|t_1=0,c=,0}\pi_{t_2=1|t_1=0,c=1}$$
$$\pi_{t_3=1|t_2=1,t_1=1,c=1}\pi_{t_3=1|t_1=1,t_2=1,c=0}\pi_{t_3=1|t_2=1,t_1=0,0}\pi_{t_3=1|t=0,t_1=0,c=0}$$
$$\pi_{t_3=1|t_2=0,t_1=1,c=1}\pi_{t_3=1|t_2=0,t_1=1,c=0}\pi_{t_3=1|t_2=0,t_1=0,c=1}\pi_{t_3=1|t_2=1,t_1=0,c=1}$$

In order to proceed we must impose some restrictions on the model. The standard identifying restrictions in this approach assumes that the three tests are independent conditional on true status. This is known as the local independence assumption (LIA) and specifies that the errors in the three tests are mutually independent. While LIA need not be true in general we will argue that it may be reasonable in the context of our analysis[5]. For example, one might be concerned that the dependence of both BMI and BIA on measured weight might violate local independence. However, this need not be the case. This would be a problem if the source of error in both BMI and BIA originated from mismeasured weight—as might be the case if the measures were based on self-reported weight. However, throughout our analysis, weight is determined by a trained expert, so we are less inclined to view mismeasured weight as a major source of error for these tests. Instead the major source of error in BMI is more likely to stem from its failure to distinguish fat and fat free mass, a feature which BIA directly addresses. In contrast the sources of error in BIA are likely to be associated with natural variation in body water content or incorrect placement of electrodes. Since the likely source of errors in the two measures is different, local independence may be a reasonable assumption, despite the common dependence of the two measures on weight.

LIA implies that $\pi_{t_2=j|t_1=1,c=1} = \pi_{t_2=j|t_1=0,c=1}$ and $\pi_{t_2=j|t_1=1,c=0} = \pi_{t_2=j|t_1=0,c=0}$ which eliminates two parameters and also that $\pi_{t_3=j|t_2=1,t_1=1,c=1} = \pi_{t_3=j|t_2=0,t_1=1,c=1} = \pi_{t_3=j|t_2=0,t_1=0,c=0} = \pi_{t_3=j|t_2=1,t_1=0,c=1}$ and $\pi_{t_3=j|t_2=1,t_1=1,c=0} = \pi_{t_3=j|t_2=1,t_1=0,c=0} = \pi_{t_3=j|t_2=0,t_1=0,c=0} = \pi_{t_3=j|t_2=0,t_1=1,c=0}$, which eliminates a further six parameters. These zero restrictions reduce the number of parameters to seven, which allows us to identify the remaining parameters. With these restrictions, $\pi_{t_1,t_2,t_3} = \sum_c \pi_c\pi_{t_1|c}\pi_{t_2|c}\pi_{t_3|c}$.

In epidemiology, the parameter $\pi_{tj=1|c=1}$ is known as the sensitivity of test j and is the probability that test j records a positive outcome when the individual truly has the latent characteristic. $\pi_{tj=0|c=0}$ is known as the specificity of test j and is the probability that test j records a negative outcome when the individual truly does not have the disease. The seven parameters to be estimated are the overall true prevalence π_1 and the sensitivity and specificity of each of the three tests[6].

With three or more tests there is no closed form solution for the maximum likelihood estimates (Hiu and Walter, 1980) but estimates can be obtained using a numerical algorithm such as Newton–Raphson or expectation maximisation. Alternatively Joseph et al. (1995) propose a Bayesian framework for estimation of this model, which allows additional information about the unknown parameters to be incorporated in the form of prior distributions, $\Pr(\pi)$. Branscum et al. (2005) provide a useful overview of Bayesian approaches in this context.

[5] Models that allow for conditional dependence between tests typically require results from at least four different tests for identification. Such models can be identified within a Bayesian context if one is able to impose strong priors on a sufficient number of the parameters (see for example Dendukuri and Joseph, 2001; Branscum et al., 2005). Such strong priors are not reasonable in our analysis.

[6] It is straightforward to show that the reparameterisation $\tilde{\pi}_{tj=1|c=1} = 1 - \pi_{tj=0|c=0}$ and $\tilde{\pi}_{tj=0|c=0} = 1 - \pi_{tj=1|c=1}$ yields the same value of likelihood function as the original parameterisation. To distinguish between these two parameters sets we impose the monotonicity condition, $\pi_{tj=1|c=1} + \pi_{tj=0|c=0} > 1$, $j = 1,2,3$. For a discussion of this condition in a related context see Hausman et al. (1998).

In the Bayesian approach to diagnostic evaluation uncertainty about the parameters is typically modelled using independent beta prior distributions:

$$\pi_c \sim \text{beta}(\alpha_{\pi_c}, \beta_{\pi_c})$$
$$\pi_{tj=1|c=1} \sim \text{beta}\left(\alpha_{1,j}, \beta_{1,j}\right), \quad j = 1,2,3$$
$$\pi_{tj=0|c=0} \sim \text{beta}\left(\alpha_{0,j}, \beta_{0,j}\right), \quad j = 1,2,3$$

The choice of the αs and βs determine the degree of prior information on each of the parameters. In the results below we set all αs and βs equal to .5 which can be interpreted in the context of Jeffrey's uninformative priors.

The posterior distributions of the parameters are given by $Pr(\pi|y) = \frac{Pr(y|\pi)Pr(\pi)}{Pr(y)}$. Any feature of the posterior distribution is legitimate for Bayesian analysis. However this typically involves taking posterior expectations of functions of π. Such integrals rarely have closed form solutions, so alternative approaches are required. Monte Carlo integration is one popular solution. If we can draw random samples from the posterior of interest then the population expectations can be estimated using sample means. Markov Chain Monte Carlo (MCMC) provides a means of sampling from the full posterior distribution given a likelihood and priors (Gilks et al. 1996).

The key to MCMC is finding a transition kernel, $Pr(\pi_{t+1}|\pi_t)$, such that the chain converges to the distribution of interest $Pr(\pi|y)$. The Metropolis–Hastings algorithm guarantees such a chain (Gilks et al. 1996). The Gibbs sampler used in this paper is a special case of the Metropolis–Hastings algorithm. At a given iteration, one simulates N random variables sequentially from N univariate conditional distributions, rather than a single N-dimensional vector in single pass from a joint distribution. This vector then serves as the conditioning vector at the next iteration. This process is repeated a large number of times, say T, and the first m, of these iterations are discarded. This burn-in period m, captures the period needed for the chain to have converged to its stationary distribution. The remaining T–m iterations in the chain are taken as random draws which can be used to evaluate the posterior distribution of the parameters[7].

3. Data

For this analysis we use the National Health and Nutrition Examination Survey (NHANES III). The NHANES III is a nationally representative survey of 33,994 individuals in the U.S. aged two months of age and older. The interviews were carried out over the period from 1988 to 1994. The NHANES data have been used in previous studies looking at the impact of obesity of labour market outcomes (e.g. Cawley, 2004). Burkhauser and Cawley (2008) describe the NHANES III as the "Rosetta Stone" for measures of fatness.

In this paper we focus on three alternative measures of fatness; BMI, WC and BIA[8]. In the NHANES survey all the health measurements were performed in specially designed and equipped mobile centres by a team of physicians and health technicians. BMI is the most widely-used measure of obesity and is defined as weight in kg/height in m^2. Individuals are classified as overweight if their BMI is between 25 and 30 and are classified as obese if their BMI exceeds 30 (WHO, 2000). Throughout our analysis we use clinically measured height and weight when determining BMI. This allows us to abstract from reporting errors typically associated with self-reported BMI (e.g. O'Neill and Sweetman, 2013; Biener et al., 2014). WC measures of obesity are based on a numerical measurement of the waist. According to the World Health Organisation's data gathering protocol, the waist circumference should be measured at the midpoint between the lower margin of the last palpable rib and the top of the iliac crest, using a stretch-resistant tape that provides a constant $100 \times g$ tension. Men are classified as being at "high risk" of obesity if their waist circumference exceeds 102 cm, while for women the threshold is 88 cm (Lean et al., 1995; NHLBI, 2000; Lear et al., 2010). Finally BIA determines the opposition to the flow of an electric current through body tissues which can then be used to estimate body fat. Fat-free mass contains mostly water, while fat contains very little water. Thus fat-free mass will have less resistance to an electrical current. By determining the resistance to the current one can estimate how much fat-free and fat is present. The Valhalla Scientific Body Composition Analyzer 1990 B is the instrument used for the measurement of whole body electrical resistance in NHANES. Electrodes were attached to the right wrist, hand, ankle and foot of the respondents and an electrical current is passed through the body. We follow the approach adopted in Burkhauser and Cawley (2008) to derive a measure of percent body fat (PBF) from the bio-electrical resistance data. This approach involves first rescaling

[7] For methods of sampling from full-conditional distributions see Gilks (1996). The WinBUGS software (Lunn et al., 2000) used in this paper uses a form of adaptive rejection sampling (Gilks and Wild, 1992).

[8] In principle one can extend the LCA approach to include other available measures of obesity such as skinfold thickness and Waist to Height (WHt) or Waist to hip (WHp) ratios (see for example Ashwell et al., 2012). However, we choose not to include these additional measures for a number of reasons. Although WC is a well-accepted measure of abdominal fat the biological meaning of WHp or Wht is less clear (Han et al., 1997; Hu, 2008). Furthermore the correlation between WC and height is very low in our data; of the order of .03 for white women (see also Johansson et al., 2009). This means that obtaining appropriate identifying variation in height may be difficult. In addition the inclusion of these alternative measures in addition to WC is likely to violate the local independence assumption required to identify the parameters. While skinfold thickness is associated with subcutaneous fat it is only weakly correlated with deep lying visceral fat (Despres et al., 1991). In addition skinfold measures are particularly difficult to measure, more prone to interobserver variations, and are less reproducible than other anthropometric methods (Ulijaszek and Kerr, 1999; Hu, 2008). When we re-estimated the three test LC model using BMI, WC and Skinfold, the results confirmed the superiorty of WC. The misclassification rates with BMI and skinfold measures were of the order of 26–34% compared to 6% with WC.

the NHANES BIA resistance scale in order to use the Fat Free Mass prediction equations developed by Sun et al. (2003)[9]. These equations allow us to predict Fat Free Mass (FFM) using BIA resistance measures. Once we have estimated FFM we can then calculate total body fat (TBF) as the difference between weight and fat free mass. Percent body fat is given by $\text{PBF} = \frac{\text{TBF}}{\text{Weight}} \times 100$. Men are typically classified as obese if their PBF exceeds 25%, while for women the threshold is 30% (NIDDK, 2001; Burkhauser and Cawley, 2008; Oreopoulos et al., 2011). We use these obesity thresholds throughout our analysis.

Each method of measuring body fat has its strengths and weaknesses (Freedman and Perry, 2000). BMI does not distinguish fat from fat free-mass such as muscle and bone, BIA readings are affected by a range of factors such as electrode placement, body position, dehydration, exercise and ambient temperature, while WC tells you the location of your body fat but not the absolute percentage of body fat and may be prone to measurement problems arising from incorrect placement of measuring tape and differences in subject posture during measurement. Despite the advances that have made in measuring fatness, there is little evidence that more recent measures of body fat are more accurate than simple combinations of height and weight (Freedman and Perry, 2000). Rather than taking one measure of obesity as a gold standard we treat all measures of fat as a-priori imperfect measures of underlying latent fatness and use the latent class approach outlined earlier to uncover the underlying characteristics of each of the tests, as well as a measure of latent obesity.

As noted in Section 2 estimation of the latent class model with 3 tests and one population requires identifying assumptions in the form of local independence. This assumption implies that the observed associations between the three tests are fully explained by the disease status. This assumption need not be valid in general and inappropriate specification of the dependence structure between tests may lead to invalid inferences (Albert and Dodd, 2004). For instance LIA may fail when two or more of the tests are based on the same biological basis or when different tests are subjected to a common source of contamination due to similar storage conditions. These factors are unlikely to be a problem in our context. For instance dehydration or body fluid near the electrodes may be a major source of error for BIA but less of a problem for measurement of waist circumference or BMI. Since all measurements were taken by the same physician it is possible that common physician error in reading tests or in calibrating the equipment could lead to dependent errors. However, while it may be possible that calibration errors may lead to misclassification in a given test, it is less likely that the calibration errors on very different pieces of equipments would lead to systematic errors across tests.

We carry out our analysis separately for six groups; White women, White men, Black women, Black men, Hispanic women and Hispanic men. We restrict attention to individuals aged between 18 and 64 and for women we exclude those who were pregnant at the time of the examination. When we exclude those with missing values on at least one of our three tests the final sample sizes are 2142 (White women), 1924 (White men), 1852 (Black women), 1628 (Black men), 1416 (Hispanic women) and 1662 (Hispanic men).

4. Results and discussion

Table 1 provides the prevalence rates for obesity for each of our groups using the three different diagnostics. There are clear and substantial differences in the prevalence rates using different measures[10]. The BMI measure tends to return the lowest obesity rate of all three tests, while BIA returns the highest rate. However, the difference between these two tests varies across groups, with the BIA prevalence being 3–4 times higher for women relative to that based on BMI, but approximately twice the rate for men. The relationship between obesity using WC and the other measures also show some differences. For White and Hispanic men and women and Black women the prevalence rate using WC lies between the BMI and BIA rates. However, for Black men, prevalence based on WC is lower than both the other measures.

To apply LCA we need to consider the joint distribution of the three tests. Table 2 provides the cross-classification of the three tests for each of our six groups. Looking down the rows in this table allows us to examine the level of agreement across the three tests. The level of agreement across the three tests (sum of first and last row) is 49.68% for White women, 63.64% for White men, 59.39% for Black women, 77.94% for Black men, 44.5% for Hispanic women and 59.9% for Hispanic men.

The data in Table 2 provide the raw input for our latent class analysis. Before looking at the results in detail Figs. 1 and 2 provide information on the history of the simulations to help assess convergence of the Markov chain. For each parameter we ran one long chain with 25,000 iterations in total. The first 5000 iterations were used for the burn-in period leaving us with 20,000 draws from the assumed stationary distribution. Fig. 1 provides a history trace of the simulations for every parameter, along with the median and the 95% credible interval[11]. These plots simply show the value of π_t chosen at each iteration t of the chain. The plots provide no evidence of drift and the mixing is good for each parameter. Furthermore, if the chain has converged to its stationary distribution then we would expect the distribution of draws to be the same over different ranges of the chain. Fig. 2 plots the density of the estimated parameters for the first 10,000 iterations and the second 10,000 iterations, along with the density based on the full chain. The similarity of all three distributions supports convergence of the chains[12].

[9] Although, no hispanics were included in the samples used by Sun et al. (2003), I use their prediction equations for all the race-ethnicity groups in our analysis.

[10] Burkhauser and Cawley (2008) report similar differences in raw reported obesity rates across different measures.

[11] Since the plots were similar for all demographic groups we only report the results for white females.

[12] We have also carried out a formal Geweke test for convergence. This test splits the chain into two parts and tests for equality of the means in the two subsamples. We follow previous work and compare the first 10% of the chain with the last 50%. In none of our analysis do we reject equality of the means.

Table 1
Obesity prevalence rates using alternative measures of body composition.

	BMI	Waist circumference	BIA
White women	23.30	42.16	72.50
White men	19.85	29.63	48.86
Black women	36.07	54.97	74.62
Black men	20.69	19.95	28.99
Hispanic women	31.21	54.17	85.73
Hispanic men	22.62	24.91	54.45

Table 3 reports the mean of the posterior distribution for each parameter, along with the 95% credible interval. A number of interesting features emerge from this analysis. Looking first at the characteristics of the three tests we see a number of important differences. The specificity rate of the BMI based test is relatively high for all six groups, implying that this test returns very few false positives; it is very unlikely that BMI records someone as obese when they are not truly obese. The false positive rate is higher for men than for women, which might be expected given that men tend to have more muscle and fat free mass than women. However, even for men the probability of a false positive is still only 1–2%. While the specificity rate of BMI is high, the same is not true of the estimated sensitivity rate. The rate is less than 70% for White men, White women and Black women, reaching a low of approximately 55% for White and Hispanic women. Only for Black and Hispanic men does the sensitivity rate exceed 80%. The problem with BMI therefore is not that it misclassifies non-obese people as obese but rather that it fails to truly detect obesity when it is present. The failure of BMI to detect true obesity is likely to be associated its failure to distinguish between muscle mass and fat. It is well established that aging is associated with substantial loss of muscle mass, a process known as sarcopenia (Hu, 2008). Thus as people age low body mass is more and more likely to reflect lower muscle mass and not necessarily low fat mass, leading to underestimation of obesity, particularly in the elderly. We explore this issue in more detail later in the paper.

The relatively high specificity rate and low sensitivity rate of BMI is consistent with previous work using different approaches. For example Smalley et al. (1990) report a sensitivity rate of 55.4% (44.3%) for all women(men) and a specificity rate of 98.2% (90.1%) using densitometric analysis based on underwater weighing as a reference point. Underwater weighting is generally perceived as one of the more accurate means of measuring body fat. However, it is not typically used nor is it widely accessible in publically available data sets. The similarity of our results with those of Smalley et al. (1990) is nevertheless encouraging.

It is also interesting to compare these estimated misclassification rates to those reported by Burkhauser and Cawley (2008). Like me, they report a false positive rate for BMI of zero for women and a false negative rate of approximately 33% for men. However, their estimated false negative rate for women (61.25%) is much higher than our estimates or those of Smalley et al. (1990). Part of the reason for this is that in contrast to the underwater weighing approach used by Smalley et al. (1990), Burkhauser and Cawley (2008) use BIA based measures of PBF as a gold-standard. However, as noted by Freedman and Perry (2000) "[BIA] has not consistently been found to provide more accurate estimates of adiposity than has anthropometry" pg. S41. This is a view shared by NHI who state that "Neither bioelectric impedance nor height–weight tables provide an advantage over BMI in the clinical management of all adult patients, regardless of gender." NHLBI (2000), p. 1. The specific problems associated with the BIA are evident in column three of Table 3. Although the estimated sensitivity of BIA is of the order of 90% or higher for all our groups, the specificity rate is much lower, particularly for women, where it is only of the order of 30–50%. This is in contrast to the 100% specificity rate assumed when BIA is used as a gold standard. In contrast to BMI measured obesity, the problem with BIA is the very high probability of a false positive. This can partly explain why the false negative rate reported by Burkhauser and Cawley (2008) for women seems so high; many of those classified as truly obese based on BIA may not in fact be obese. The relatively poor performance of BIA for women in our analysis is consistent with some previous work. Gleichauf and Roe (1989) and Dehghan and Merchant (2008) both discussed the impact of menopause and the menstrual cycle when using BIA to measure obesity. Dehghan and Merchant (2008) note that increased progesterone plasma levels after ovulation, along with the change in hydration status, can lead to the within-subject variability of impedance to be higher in women, while Gleichauf and Roe (1989) recommend the average of several BIA

Table 2
Cross-classification of BMI, WC and BIA tests.

Test outcome			White women (%)	White men (%)	Black women (%)	Black men (%)	Hispanic women (%)	Hispanic men (%)
BMI	WC	BIA						
+	+	+	22.7	16.94	35.15	13.3	30.3	17.8
+	+	–	0	1.5	.10	2.82	0	1.5
+	–	+	.51	.78	.81	2.70	.01	2.5
–	+	+	18.86	8.84	18.68	2.14	23.8	4.5
+	–	–	0	.57	0	1.84	0	.01
–	+	–	.51	2.28	1.0	1.66	0	1.1
–	–	+	30.3	22.29	19.88	10.8	30.7	29.7
–	–	–	26.98	46.7	24.24	64.64	14.2	42.1
			100	100	100	100	100	100

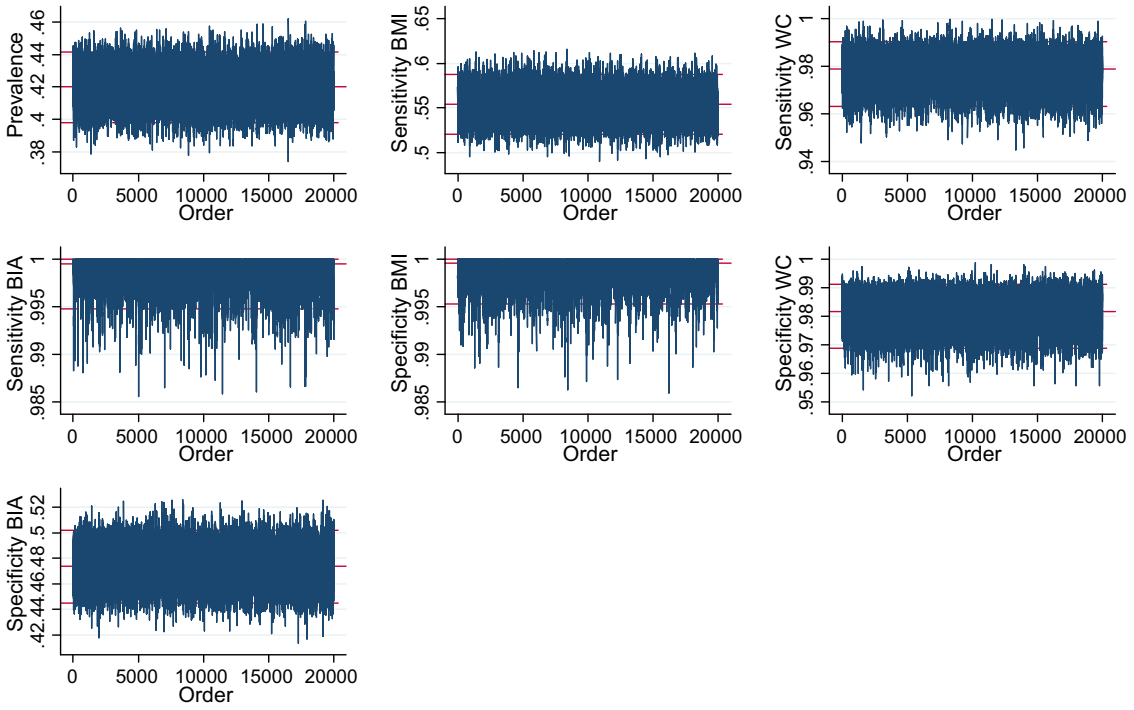

Fig. 1. History plot of Markov Chain Monte Carlo simulations: White women.

measures during a menstrual cycle be considered when estimating body composition.

These results have important consequences for traditional econometric analysis of obesity. It well known that measurement error arising from self-reported BMI can seriously bias standard estimators, causing researchers to draw misleading inferences concerning the relationship between obesity and outcomes such as health, employment and wages (O'Neill and Sweetman 2013). The results presented above show that having access to clinically

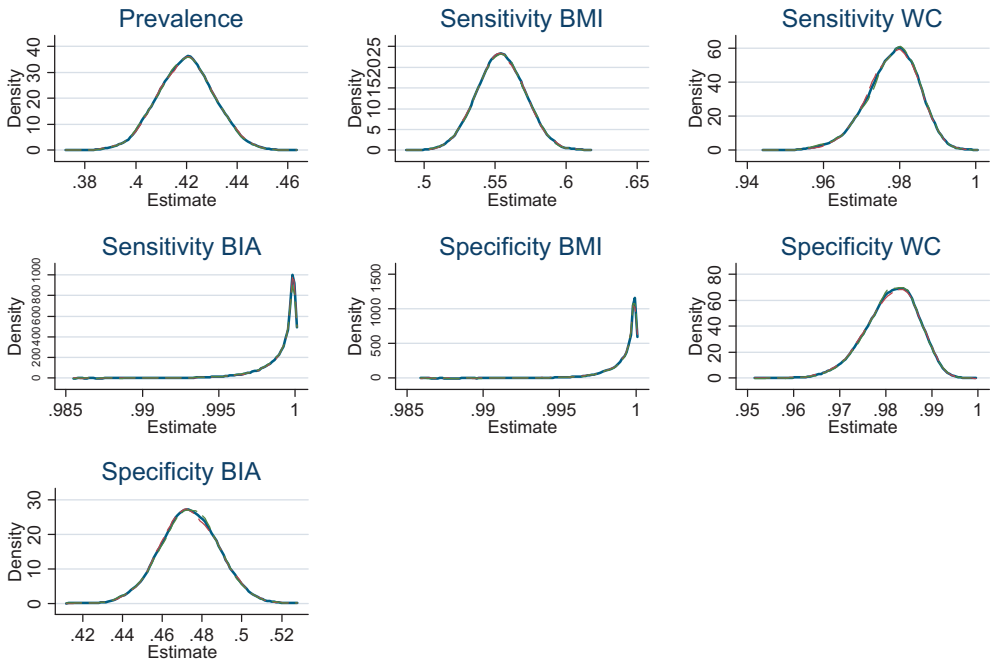

Fig. 2. Posterior density estimates for sections of the Markov chain. Dashed lines, densities for first and second half of the chain; solid line, density based on full chain: White women. Measuring obesity in the absence of a gold standard.

measured BMI (or BIA) is not sufficient to rid standard measures of obesity of measurement error. Although reporting error may no longer be an issue, one still has to contend with the false positive and false negative diagnoses associated with clinically measured BMI and BIA. Furthermore, our results show that such errors will tend to be non-classical and differ across different measures (such as BMI and BIA), causing even further problems. In short all the well-known problems typically associated with measurement error may still be present even with clinical measures of height and weight.

In contrast to the BMI and BIA measures, the results in Table 3 suggest that the classification of latent obesity based on waist circumference exhibits high degrees of accuracy both in terms of sensitivity and specificity. The probability of both false negatives and false positives is of the order of 3% for White men and for all the female subgroups. Only in the case of the sensitivity measure for Black and Hispanic men does the error rate exceed 5%. These results suggest that waist circumference may provide a cheap and effective measure of latent obesity. It is interesting to consider this finding in the light of recent work relating alternative measures of body composition to health and economic outcomes. In their study of obesity and labour market success in Finland, Johansson et al. (2009) found that only waist circumference had a negative association with wages for women, while Mosca (2013) found that among older Irish adults the negative employment elasticity associated with waist circumference is larger than the elasticity associated with BMI. Janssen et al. (2004) found that that WC outperformed BMI at predicting health risk associated with obesity. A recent review by Seidell (2010) noted that WC provided a better indicator of all-cause mortality than BMI and that waist alone could replace WHT and BMI as a single risk-factor for all cause mortality, while Chan et al. (2003) concluded that "WC is the anthropometric index that most uniformly predicts the distribution of adipose tissue. . .there apparently being little value in measuring WHT (Waist to hip ratio) or BMI."[13]

The fact that our latent class analysis identifies WC as an effective measure of fatness has important implications for our understanding of the growth in obesity over the last 20 years. The Centre for Disease Control and Prevention estimate that in 1990 obese adults made up less than 15% of the population in most U.S. states. By 2010, 36 states had obesity rates of 25% or higher and 12 of those had obesity rates of 30% or higher. These concerns about the increase in obesity have for the most part been based on increasing BMI. However, there is also evidence that the nature of excess body weight has been changing over this time. In particular a number of authors (e.g. Elobeid et al., 2007; Ford et al., 2014) have shown that over the last 50 years WC values have increased beyond those expected from BMI increases. Since our analysis suggests that the true prevalence rate is reflected in WC measures of fatness, this would suggest that the growth in obesity and associated costs may in fact be even more serious than that documented by rising BMI[14].

While our results support the use of WC based measures in determining obesity rates, it is important to note that our analysis is based on clinically measured BMI, BIA and WC. Although, our results establish the accuracy of WC in this setting, there is some evidence that measurement error in self-measured WC may be larger than that in self-measured BMI (Ulijaszek and Kerr, 1999 and Verweij et al., 2013). The intra class reliability of height, weight and WC were all high (above .97), while the interobserver reliability for WC was lower (of the order of .94). However, there is also evidence that proper guidance and the use of well-designed instruments can significantly reduce measurement error in self-measured WC. For example Han and Lean (1998) examined the consequences of using a special colour coded measurement tape along with step by step photographic instructions on proper measurement. Using clinical WC measurements as the reference point they found that the sensitivity and specificity rates of WC, with guidance, were over 95%, compared to sensitivity rates as low as 58% without guidance. We view proper guidance and assistance in the measurement and interpretation of WC as a crucial component of any policy initiative based on WC[15].

The last column of Table 3 reports our estimated true prevalence of latent obesity derived from LCA. From this we see that Black and Hispanic women have the highest estimated prevalence of obesity (of the order of 55%), while Black men have the lowest estimated obesity rate (22%). It is interesting to compare these estimates to the estimates based on other measures. In particular we follow Burkhauser and Cawley (2008) and examine Black–White racial differences in obesity rates. For reference we first consider the raw obesity rates in Table 2. The Black–White racial patterns we report using the raw data are consistent with the results reported in Burkhauser and Cawley (2008). When one defines obesity using BMI the obesity rate among Black women is about 12 percentage points higher than among White women, while there is less than a 1 percentage point difference in the rates between White men and Black men. However, the Black–White gap in obesity changes dramatically when one classifies people using PBF. While the female Black–White racial gap is significantly reduced substantial racial differences emerge for men. However, since both these measures appear to suffer from misclassification bias neither of these racial gaps reflect actual racial differences in obesity. To determine actual racial differences we turn to the

[13] For a recent review of this literature, see Huxley et al. (2010).

[14] If one is willing to assume that the sensitivity and specificity rates of BMI have remained constant over time one can combine these estimated rates with observed time-series on BMI to a create a time series for the true underlying prevalence of obesity using the fact that at any point in time the true underlying prevalence is given by $\pi_C = \frac{(\pi_{BMI} - (1 - \pi_{BMI=0|0}))}{(\pi_{BMI=1|1} - (1 - \pi_{BMI=0|0}))}$. This approach potentially allows us to recover true prevalence rates for time periods when only BMI is observed. See for example Komlos (1987), Carson (2009), Hiermeyer (2010) and Bodenhorn (2010) for an analysis of BMI in the US during the 19th century.

[15] For a discussion on the appropriate protocol for measuring waist circumference, see WHO (2011).

Table 3
Latent class analysis of obesity measures: mean of the posterior distribution with 95% credible interval in parentheses.

	Sensitivity BMI	Specificity BMI	Sensitivity WC	Specificity WC	Sensitivity BIA	Specificity BIA	Prevalence
White women	55.4 (52.1–58.7)	100 (99.5–100)	97.8 (96.3–99.0)	98.1 (96.9–99.1)	99.9 (99.5–100)	47.4 (44.5–50.2)	42 (39.8–44.2)
White men	67.5 (62.9–72)	98.8 (97.9–99.5)	97.0 (94.2–99.6)	96.8 (95.2–98.4)	91.4 (88.2–94.1)	67.9 (65.3–70.4)	28.2 (25.9–30.6)
Black women	66.2 (63.2–69.2)	99.9 (99.4–100)	97.7 (96.4–98.8)	96.1 (94.1–97.8)	99.6 (99–99.9)	55.3 (51.8–58.7)	54.4 (52–56.8)
Black men	87 (82.1–91.3)	98 (96.8–99.1)	84.0 (78.8–88.6)	98.1 (97–99.1)	82.3 (77.3–86.8)	86.0 (84.0–88)	22.0 (19.7–24.4)
Hispanic women	56.1 (52.5–59.6)	99.7 (98.6–100)	97.2 (95.4–99)	99.4 (97.9–100)	99.9 (99.5–100)	32.0 (28.33–35.71)	55.4 (52.7–58.2)
Hispanic men	81.6 (76.5–86.3)	98.4 (97.1–99.4)	89.5 (85.2–93.7)	98.1 (96.6–99.4)	91.9 (88.6–94.7)	58.9 (56.0–61.7)	26.3 (23.8–28.9)

estimated true prevalence rates reported in Table 3. Our estimated true prevalence rates imply a Black–White racial gap for women that is similar to the gap using BMI (of the order of 12 percentage points). However our estimated rates imply significantly lower obesity levels among Black men. Though statistically significant, the gap of 6 percentage points is smaller than that based on PBF (20 percentage points).

5. Sensitivity analysis

5.1. Alternative priors

In this section we examine the robustness of our results to alternative prior distributions in the estimation procedure. In particular, we consider robustness to the parameters of the beta prior distribution used in the previous section and robustness to the use of an alternative prior distribution. As noted earlier the Beta(.5,.5) distribution, used up to now, can be interpreted in the context of Jeffrey's uninformative priors. By varying the parameters of the prior distribution we can shift prior weight to either end of the parameter range. We consider two alternative paramterisations; Beta(.5,1) and Beta(1,.5). The first distribution is skewed to the right and places much heavier prior weight on parameter values close to zero, while the second is skewed to the left and places heavier weight on parameters values close to one. In addition we consider the consequences of using an uninformative uniform distribution U(0,1) rather than the Beta(.5,.5) distribution used in Section 4[16].

The results using these alternative priors are given in Table 4. As the findings for all groups are similar, we only report the results for White women. The first row, reproduces the results from our earlier analysis. The second, third and fourth rows show the results for each of the alternative prior distributions. It is clear from these results that our findings are robust to the choice of prior; the results are almost identical across rows. BMI suffers from a low sensitivity rate and BIA suffers from a low specificity rate, while the misclassification rate for WC is very low, irrespective of which prior distribution is used.

5.2. Simulations

The results from the analysis so far shows that both the error rates for obesity based on WC are low compared to those using BMI or BIA. It may be tempting to believe that the LCA approach used in this paper identifies WC as the most accurate measure simply because the observed obesity rate using WC falls between the other two measures. However, this is not the case. The obesity rate for any single measure is constructed using only the marginal distribution for that test. In contrast the error rates derived from the LCA analysis are identified from the joint distribution of the three tests. To illustrate this further I simulated a true underlying distribution of obesity by drawing 10,000 observations from a uniform distribution and assigning the top 20% of the distribution to the obese state. I then constructed three diagnostic tests. For simplicity I generate the tests such that all three have a specificity rate equal to 100%; that is there are no false positives. The three constructed tests differ only in terms of their sensitivity rates. The first test is chosen to have a sensitivity rate equal to 100%. This, combined with the assumption of a 100% specificity rate, implies that test 1 is a perfect predictor of obesity, with zero errors. In contrast test 2 is constructed to have a sensitivity rate of 70%; that is the results of test 2 were altered so that 30% of those who are truly obese had their result changed from a positive diagnosis of obesity to a negative one. Test 3 is constructed to have a sensitivity rate of .4[17]. As a result measured obesity using test 1 is 20%, 14% ($.7 \times .2$) using test 2 and 8% ($.4 \times .2$) using test 3. Thus the observed obesity rate for test 2 lies between the rates for test 1 and test 3 even though test 1 is the gold standard.

The estimated joint distribution of our three tests from our simulation is then used in conjunction with the LCA approach to estimate the parameters of each test, along with the true underlying obesity rate. Since we know the true DGP it is possible to compare the estimated parameters to the true parameters. The results are presented in Table 5. Looking at the last column we see that the LCA accurately estimates the true underlying obesity rate which is 20%. More importantly for the purposes of this section we see that the LCA approach also accurately estimates the error rates for all three tests. The specificity rates for each test is very close to the true rate of 100%, while the estimated sensitivity rates are close to their true values of 100%, 70% and 40% for test 1, test 2 and test 3, respectively. The LCA approach clearly identifies test 1 as the superior diagnostic despite its reported obesity rate being the highest of all three tests. These simulated

[16] The uniform[0,1] is equivalent to a Beta(1,1).

[17] LIA was imposed when misclassifying individuals in the simulated sample.

Table 4
Latent class analysis of obesity measures for white women with alternative prior distributions: mean of the posterior distribution with 95% credible interval in parentheses.

Prior distribution	Sensitivity BMI	Specificity BMI	Sensitivity WC	Specificity WC	Sensitivity BIA	Specificity BIA	Prevalence
Beta(.5,.5)	55.4 (52.1–58.7)	100 (99.5–100)	97.8 (96.3–99.0)	98.1 (96.9–99.1)	99.9 (99.5–100)	47.4 (44.5–50.2)	42 (39.8–44.2)
Beta(.5,1)	55.4 (52–58.8)	99.8 (99.4–100)	97.8 (96.3–99.1)	98.1 (96.8–99.1)	99.8 (99.3–100)	47.3 (44.6–50.2)	42 (39.7–44.2)
Beta(.1,.5)	55.5 (52.1–58.8)	99.9 (99.5–100)	97.8 (96.3–99.0)	98.1 (96.8–99.1)	99.9 (99.5–100)	47.4 (44.6–50.2)	42 (39.8–44.2)
Uniform(0,1)	55.4 (51.9–58.8)	99.8 (99.4–100)	97.8 (96.3–99.1)	98.1 (96.8–99.1)	99.8 (99.3–100)	47.3 (44.5–50.1)	42 (39.8–44.2)

Table 5
Simulated results: mean of the posterior distribution with 95% credible interval in parentheses (true parameters: prevalence = 20%, specificity of test 1 = specificity test 2 = specificity of test 3 = 100% sensitivity test 1 = 100%, sensitivity of test 2 = 70%, sensitivity of test 3 = 40%).

Sensitivity test 1	Specificity test 1	Sensitivity test 2	Specificity test 2	Sensitivity test 3	Specificity test 3	Prevalence
99.94 (99.25–100)	99.65 (98.39–100)	71.33 (66.44–76.19)	99.99 (99.86–100)	41.61 (37–46.57)	99.99 (99.85–100)	19.48 (17.46–21.30)

results clearly show that the low error rates estimated for WC in the previous section should not be attributed simply to the fact that its observed obesity rate falls between the obesity rate for BMI and that for BIA.

5.3. Age variation

We noted in Section 4 that the low sensitivity rate of BMI is likely to be associated with sarcopenia, the loss of muscle mass associated with aging. As the loss of muscle mass is more pronounced among the elderly one, would expect the sensitivity rate associated with BMI to be lower among older people. To examine the sensitivity of our results to variation in age we repeat the analysis in Section 4, this time splitting the population into two groups; those aged less than 50 and those aged 50 or over. The results are given in Table 6[18]. Although the relatively large credible intervals make it difficult to draw strong statistical conclusions across age groups age groups, the point estimates are consistent with expectations. The sensitivity rate of BMI is almost 10 percentage points lower for the older age group than it is for the younger group, reflecting the greater loss in muscle mass among the elderly. For both age groups the main findings of our paper are still evident: the BMI measure of obesity suffers from a low sensitivity rate, the BIA measure suffers from a low specificity rate, while the WC measure of obesity performs well in both dimensions.

5.4. Alternative thresholds

Finally in this section we consider the robustness of our findings to the use of alternative thresholds when defining obesity. The thresholds used in Section 4 for determining obesity are those recommended by leading health authorities (WHO, 2000; NHLBI, 2000; NIDDK, 2001) and have been used in previous studies evaluating tests for obesity (Burkhauser and Cawley, 2008). Given the nature of obesity, determining a universally accepted threshold is not straightforward. Clearly changing the threshold for a given test will involve a tradeoff between the two errors; raising the threshold will improve the specificity of a test but at the expense of a reduction in its sensitivity. However, it is not clear a-priori as to what the relative change in the magnitudes of the two errors will be. Therefore it is of interest to see to what extent changes in the thresholds for BIA or BMI affect our conclusions.

In Section 4 we classified a man as obese if his PBF exceeded 25% and a woman as obese if her PBF exceeds 30%. Lavie et al. (2010) propose PBF thresholds in the range 23–25% for men and 33–35% for women. The proposed thresholds for women are higher than the 30% threshold I used in Section 4. To examine the sensitivity of our findings to the use of a higher threshold for women we repeat the analysis in Section 4 for White women using a PBF cut off of 35% rather than 30%. The immediate impact of the higher threshold is to substantially reduce the BIA reported obesity rate from 72.50% to 51.01%, which is more in line with the 42.16% rate recorded using WC. However, of more interest here is the impact of the higher threshold on our estimated error rates. The results of the LCA using the higher PBF threshold are given in the first row of Table 7. As expected the specificity of the BIA measures improves using the higher threshold. The substantial improvement in specificity of BIA from 47.4% with the old threshold to 79.8% with the new threshold is achieved with only a modest reduction in the sensitivity of the test. However, despite the improvements in BIA, the use of the higher PBF threshold does little to alter our previous conclusion. Both the sensitivity and specificity of the WC are above 90%, with this measure being preferred to either the BMI or BIA measure.

Finally we consider the use of a lower threshold for BMI in an attempt to improve the sensitivity of this measure. To examine the sensitivity of our results to the BMI threshold we reduce the threshold from 30 to 27.5 (halfway towards the current cutoff for overweight). As expected this reduction increases the BMI based obesity rate from

[18] For brevity we only report the results for white women. The findings for the other groups are qualitatively similar.

Table 6
Latent class analysis of obesity measures: white women aged less than 50 and white women aged 50 or more. Mean of the posterior distribution with 95% credible interval in parentheses.

	Sensitivity BMI	Specificity BMI	Sensitivity WC	Specificity WC	Sensitivity BIA	Specificity BIA	Prevalence
White women less than 50	60.00 (52.1–58.7)	99.89 (98.25–100)	96.83 (91.78–100)	99.37 (96.62–100)	99.8 (97.4–100)	50.66 (43.06–56.7)	32.69 (28.09–37.8)
White women 50 or more	50.98 (39.91–61.57)	99.55 (95.44–100)	99.15 (93.45–100)	91.73 (76.68–100)	99.74 (96.18–100)	34.45 (23.22–47.13)	60.21 (51.11–68.98)

Table 7
Latent class analysis of obesity measures: alternative thresholds mean of the posterior distribution with 95% credible interval in parentheses.

	Sensitivity BMI	Specificity BMI	Sensitivity WC	Specificity WC	SensitivityBIA	Specificity BIA	Prevalence
White women PBF $\geq$ 35	59.23 (52.22–66.63)	99.86 (98.55–100)	97.92 (94.29–99.78)	93.84 (90.13–96.6)	98.66 (95.55–99.88)	79.8 (75.27–83.98)	39.2 (34.35–43.28)
White women BMI $\geq$ 27.5	73.64 (66.13–79.82)	99.91 (98.63–100)	93.67 (89.27–96.6)	98.35 (94.98–99.98)	99.64 (97.96–100)	48.96 (43.88–54.9)	44.05 (39.64–48.08)

23.30% to 32.5%. The impact of this new lower threshold on the estimated properties of the three tests is given in the second row of Table 7. Once again we see that the improvement in the sensitivity of the BMI test from 55.4% to 73.64% is achieved with only a small reduction in the specificity of the test. However the WC based test still dominates the other two with both high sensitivity and high specificity rates[19].

6. Conclusion

It is generally accepted that obesity rates have increased substantially over the last 40 years and that the costs of rising obesity can be significant. However, to date the lack of an acceptable gold-standard has limited the assessment of the validity of field methods used to measure obesity. When competing measures of obesity give conflicting results it is challenging to know how to reconcile these differences. In this paper we use latent class analysis to evaluate alternative measures of obesity in the absence of a gold standard. Using data from a representative sample of US adults we consider three popular measures of obesity; body mass index, bioelectrical impedance analysis and waist circumference. Rather than giving one of the measures ex-ante preference over another, we treat all three as potentially imperfect measures of underlying obesity and use class analysis to estimate the true underlying prevalence of the disease, along with measures of the sensitivity and specificity of each of the tests.

We show that while measures based on body mass index and bioelectrical impedance analysis misclassify large numbers of individuals, the classification of latent obesity based on waist circumference suffers from significantly less bias. The probability of both false negatives and false positives with this measure is of the order of 3% for White men, White, Black and Hispanic women.

While the results from our analysis reinforce earlier warnings regarding the use of BMI it is important to note that our findings are derived using weaker assumptions than those adopted in some previous studies. In addition our results emphasise the precise problem with BMI; namely its low sensitivity rate. Although this has been reported in a very small number of previous studies it does not seem to have been widely acknowledged, particularly in popular discussion. These discussions still tend to focus on the problems posed by BMI for groups such as athletes, who tend to have significant muscle mass. To the extent that high muscle mass poses a major problem for BMI it will manifest itself in low specificity rates, which is not the case. Arguing that BMI, although imperfect, is often the only measure available to researchers and therefore is the best that can be done is not a satisfactory response to the problems discussed in this paper. The measurement problems associated with both BMI and BIA and in particular the non-classical nature of these errors is likely to result in severe biases for a number of popular estimators used in econometric analysis. Researchers undertaking empirical work in obesity need to recognise this fact. For future research our findings suggest that the incorporation of a simple additional measurement of obesity, namely waist circumference, into future health studies will prove highly valuable.

Our findings also have important policy implications, both in terms of how we measure the growth in obesity over time and also in terms of how we evaluate racial gaps in obesity. Since our analysis suggests that measures based on WC accurately reflect the true prevalence of obesity and since WC measures of obesity have grown in excess of what would be predicted given the growth in BMI it is quite possible that we are underestimating the extent to which obesity has grown over time.

[19] I also carried out the analysis using both the new higher threshold for PBF and the lower threshold for BMI simultaneously. This sensitivity for BMI in this case is similar to that obtained using just the new BMI threshold and old PBF threshold and the specificity of BIA measure was similar to that using just the new PBF cut-off and old BIA measure. However, once again the sensitivity and specificity rates of WC were both over 95%, with the WC measure dominating both of the others.

Finally our analysis suggests that reliance on a measure such as BIA may divert attention away from what appears to be a serious obesity problem among Black and Hispanic women. A simple information campaign illustrating the appropriate procedure for measuring waist circumference could prove highly effective in helping us understand and combat the growth in obesity, particularly among the most vulnerable groups.

References

Albert, P.S., Dodd, L.E., 2004. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics 60, 427–435.

Antonanzas, F., Rodriguez, R., 2010. Feeding the economics of obesity in the EU in a healthy way. Eur. J. Health Econ. 11, 351–353.

Ashwell, M., Gunn, P., Gibson, S., 2012. Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis. Obes. Rev. 13, 275–286.

Biemer, P., 2011. Latent Class Analysis of Survey Error. Wiley and Sons, New Jersey.

Biemer, P., Wiesen, C., 2002. Measurement error evaluation of self-reported drug use: a latent class analysis of the U.S. National Household Survey on Drug Abuse. J. R. Stat. Assoc. 165 (Part 1), 97–119.

Biener, A., Meyerhoefer, C., Cawley, J., 2014. Estimating the Medical Care Costs of Youth Obesity in the Presence of Proxy Reporting Error ⟨http://www.lehigh.edu/∼aib210/aib_research.html⟩

Bodenhorn, H., 2010. Height and body mass index values of nineteenth century New York Legislators. Econ. Hum. Biol. 8, 121–127.

Branscum, A.J., Gardner, I.A., Johnson, W.O., 2005. Estimation of diagnostic-test sensitivity and specificity through Bayesian modelling. Prevent. Vet. Med. 68, 145–163.

Burkhauser, R.V., Cawley, J., 2008. Beyond BMI: the value of more accurate measures of fatness and obesity in social science research. J. Health Econ. 27, 519–529.

Burkhauser, R.V., Cawley, J., Schmeiser, M.D., 2009. The timing of the rise in U.S. obesity varies with measure of fatness. Econ. Hum. Biol. 7, 307–318.

Carson, S.A., 2009. Racial differences in body mass indices of men imprisoned in 19th century Texas. Econ. Hum. Biol. 7, 121–127.

Cawley, J., 2004. The impact of obesity on wages. J. Hum. Resour. 39, 451–474.

Cawley, J., Meyerhoefer, C., 2012. The medical care costs of obesity: an instrumental variables approach. J. Health Econ. 31, 219–230.

Chan, D.C., Watts, C.F., Barrett, P.H.R., Burke, V., 2003. Waist circumference, waist-to-hip ratio and body mass index as predictors of adipose tissue compartments in men. QJM 96 (Jun 6), 441–447, http://dx.doi.org/10.1093/qjmed/hcg069.

Council on Science and Public Health, 2012. Is obesity a disease? In: CSAPH Report 3-A-13Council on Science and Public Health.

Dehghan, M., Merchant, A.T., 2008. Is bioelectrical impedance accurate for use in large epidemiological studies?. Nutr. J. 7: 26. http://dx.doi.org/10.1186/1475-2891-7-26 PMC 2543039 . PMID 18778488.

Dendukuri, N., Joseph, L., 2001. Bayesian approaches to modelling the conditional dependence between multiple diagnostic tests. Biometrics 57, 158–167.

Despres, J., Prud'homme, D., Pouliot, M., Tremblay, A., Bouchard, C., 1991. Estimation of deep abdominal adipose-tissue accumulation from simple anthropometric measurements in men. Am. J. Clin. Nutr. 54, 471–477.

Elobeid, M.A., Desmond, R.A., Thomas, O., Keith, S.W., Allison, D.B., 2007. Waist circumference values are increasing beyond those expected from BMI increases. Obesity (Silver Spring, Md.) 15, 2380–2383.

Ford, E., Maynard, L., Li, C., 2014. Trends in mean waist circumference and abdominal obesity among US adults, 1999–2012. J. Am. Med. Assoc. 312, 1151–1153.

Freedman, D., Perry, G., 2000. Body composition and health status among children and adolescents. Prevent. Med. 31, S34–S53.

Gilks, W., Richardson, S., Spiegelhalter, D., 1996. Introducing Markov chain Monte Carlo. In: Gilks, W., Richardson, S., Spiegelhalter, D. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

Gilks, W., Wild, P., 1992. Adaptive rejection sampling for Gibbs sampling. J. R. Stat. Soc., Ser. C 41, 337–348.

Gleichauf, C.N., Roe, D.A., 1989. The menstrual cycle's effect on the reliability of bio impedance measurements for assessing body composition. Am. J. Clin. Nutr. 50, 903–907.

Gonçalves, L., Subtil, A., De Olivera, R., Do Rosario, V., Lee, P., Shaio, M.-F., 2012. Bayesian latent class models in malaria diagnosis. PLoS ONE 7, 1.

Han, T., Seidell, J., Currall, J., Morrison, C., Deurenberg, P., Lean, M., 1997. The influence of height and age on waist circumference as an index of adiposity in adults. Int. J. Obes. 21, 83–89.

Han, T.S., Lean, M.E.J., 1998. Self-reported waist circumference compared with the 'waist watcher' tape-measure to identify individuals at increased health risk through intra-abdominal fat accumulation. Br. J. Nutr. 80, 81–88.

Hausman, J., Abrevaya, J., Scott-Morton, F.M., 1998. Misclassification of the dependent variable in a discrete response model. J. Econometr. 87, 239–269.

Heshka, S., Allison, D., 2001. Is obesity a disease? Int. J. Obes. 25, 1401–1404.

Hiermeyer, M., 2010. The height and BMI values of West Point Cadets after the Civil War. Econ. Hum. Biol. 8, 127–133.

Hiu, S., Walter, S., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171.

Hu, F., 2008. Measurements of adiposity and body composition. In: Hu, F. (Ed.), Obesity Epidemiology. Oxford University Press, New York, pp. 53–83.

Huxley, R., Mendis, S., Zheleznyakov, E., Reddy, S., Chan, J., 2010. Body mass index, waist circumference and waist:hip ratio as predictors of cardiovascular risk—a review of the literature. Eur. J. Clin. Nutr. 64, 16–22.

Janssen, I., Katzmarzyk, P.T., Ross, R., 2004. Waist circumference and not body mass index explains obesity-related health risk. Am. J. Clin. Nutr. 79, 379–384.

Johansson, E., Bockerman, P., Kiiskinen, U., Heliovaara, M., 2009. Obesity and labour market success in Finland: the difference between having a high BMI and being fat. Econ. Hum. Biol. 7, 36–45.

Joseph, L., Gyorkos, T.W., Coupal, L., 1995. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am. J. Epidemiol. 141, 263–272.

Komlos, J., 1987. The height and weight of west point cadets: dietary change in antebellum America. J. Econ. Hist. 47, 897–927.

Konnopka, A., Bödemann, M., König, H.H., 2011. Health burden and costs of obesity and overweight in Germany. Eur. J. Health Econ. 12, 345–352.

Kopelman, P., Finer, N., 2001. Reply: is obesity a disease? Int. J. Obes. 25, 1405–1406.

Krakauer, N., Krakauer, J., 2012. A new body shape index predicts mortality hazard independently of body mass index. PLoS ONE 7 (7) , http://dx.doi.org/10.1371/journal.pone.0039504, Article Number: e39504.

Lavie, C., Milani, R., Ventura, H., De Schutter, A., 2010. Use of body fatness cutoffs—reply. Mayo Clin. Proc. 85, 1057–1058.

Lean, M.E.J., Han, T.S., Morrison, C.E., 1995. Waist circumference as a measure for indicating need for weight management. BMJ: Br. Med. J. 311, 158–161.

Lear, S.A., James, P.T., Ko, G.T., Kumanyika, S., 2010. Appropriateness of waist circumference and waist-to-hip ratio cutoffs for different ethnic groups. Eur. J. Clin. Nutr. 64, 42–61.

Lunn, D., Thomas, N., Best, N., Spiegelhalter, D., 2000. WinBUGS—a Bayesian modelling framework: concepts, structure and extensibility. Stat. Comput. 10, 325–337.

Madden, A.M., Smith, S., 2014. Body composition and morphological assessment of nutritional status in adults: a review of anthropometric variables. J. Hum. Nutr. Diet., http://dx.doi.org/10.1111/jhn.12278.

McCarthy, H.D., Cole, T., Fry, T., Jebb, S.A., Prentice, A.M., 2006. Body fat reference curves for children. Int. J. Obes. 30, 598–602.

Mei, Z., Grummer-Strawn, A., Pietrobelli, A., Goulding, M., Goran, M., Dietz, W., 2002. Diet validity of body mass index compared with other body composition screening indexes for the assessment of body fatness in children and adolescents. Am. J. Clin. Nutr. 75, 978–985.

Mosca, I., 2013. Body mass index, waist circumference and employment: evidence from older Irish adults. Econ. Hum. Biol. 11, 522–533.

National Heart Lung and Blood Institute (NHLBI), 2000. The Practical Guide: Identification, Evaluation and Treatment of Overweight and Obesity in Adults. NHLBI.

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), 2001. Understanding Adult Obesity. NIH.

O'Neill, D., Sweetman, O., 2013. The consequences of measurement error when estimating the impact of obesity on income. IZA J. Labor Econ. 2, 3.

Oreopoulos, A., Lavie, C., Snitker, S., Romero-Corral, A., 2011. More on body fat cutoff points—reply 1. Mayo Clin. Proc. 86, 584–585.

Rindskopf, D., Rindskopf, W., 1986. The value of latent class analysis in medical diagnosis. Stat. Med. 5, 21–27.
Seidell, J.C., 2010. Waist circumference and waist/hip ratio in relation to all-cause mortality, cancer and sleep apnoea. Eur. J. Clin. Nutr. 64, 35–41.
Smalley, K., Knerr, A., Colliver, A., Kendrick, Z., Owen, O., 1990. A reassessment of body mass indices. Am. J. Clin. Nutr. 52, 405–408.
Song, X., Jousilahti, P., Stehouwer, C.D.A., Soderberg, S., Onat, A., Laatikainen, T., Yudkin, J.S., Dankner, R., Morris, R., Tuomilehto, J., Qiao, Q., 2013. Comparison of various surrogate obesity indicators as predictors of cardiovascular mortality in four European populations. Eur. J. Clin. Nutr. 67, 1298–1302.
Sun, S.S., Chumlea, W., Heymsfield, S., Lukaski, H.C., Schoeller, D., Friedl, K., Kuczmarski, R., Flegal, K., Johnson, C., Hubbard, V., 2003. Development of bioelectrical impedance analysis prediction equations for body composition with the use of a multicomponent model for use in epidemiologic surveys. Am. J. Clin. Nutr. 77, 10.
Szatmari, P., Volkmar, F., Walter, S., 1995. Evaluation of diagnostic criteria for autism using latent class models. J. Am. Acad. Child Adolesc. Psychiatry 34, 216–222.
Ulijaszek, S.J., Kerr, D.A., 1999. Anthropometric measurement error and the assessment of nutritional status. Br. J. Nutr. 82, 165–177.
Verweij, L.M., Terwee, C.B., Proper, K.I., Hulshof, C.T.J., Van Mechelen, W., 2013. Measurement error of waist circumference: gaps in knowledge. Public Health Nutr. 16, 281–288.
Walter, S., Irwig, L., 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J. Clin. Epidemiol. 41, 923–937.
WHO, 2000. Obesity: Preventing and Managing the Global Epidemic. In: Technical Report Series 894WHO.
WHO, 2009. Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks. WHO Press.
WHO, 2011. Waist circumference and waist–hip ratio. In: Report of a WHO Expert ConsultationWHO, Geneva.