DISCUSSION PAPER SERIES

IZA DP No. 7893

Measuring Obesity in the Absence of a Gold Standard

Donal O'Neill

January 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Measuring Obesity in the Absence of a Gold Standard

Donal O'Neill
National University of Ireland Maynooth
and IZA

Discussion Paper No. 7893
January 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 7893
January 2014

ABSTRACT

Measuring Obesity in the Absence of a Gold Standard*

Reliable measures of body composition are essential in order to develop effective policies to tackle the costs of obesity. To date the lack of an acceptable gold-standard for measuring fatness has made it difficult to evaluate alternative measures of obesity. In this paper we draw on work in other areas of epidemiology and 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 show that while measures based on Body Mass Index and Bioelectrical Impedance Analysis appear to misclassify large numbers of individuals, this is not the case for classification based on waist circumference. The error rates associated with waist circumference are of the order of 3% for most of our samples compared to error rates as high as 40-50% with the other measures. These results have implications for racial differences in obesity. Our estimated true prevalence rates imply that the obesity rate among black women is substantially higher than among white women. However, the opposite is true for men, with the black men having a significantly lower obesity rate among black men. The fact that neither the BMI nor the BIA based measures of obesity are capable of capturing both these features highlights the dangers associated with measuring obesity and the potential costly policy mistakes that may arise from arbitrarily adopting a single measure as a gold standard.

JEL Classification: I18, C38

Keywords: obesity, multiple diagnostic tests, latent class analysis

Corresponding author:

Donal O'Neill
Department of Economics
National University of Ireland Maynooth
Maynooth, Co. Kildare
Ireland
E-mail: donal.oneill@nuim.ie

* I would like to thank Olive Sweetman for helpful comments on an earlier draft of this paper.

1. Introduction

Obesity is an important 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). In June 2013, the the American Medical Association voted to classify obesity as a disease in the hopes that by recognizing obesity as a disease it will help change the way the medical community tackles this complex health issue. However, the decision to classify obesity as a disease raises fresh concerns as to how best to measure and diagnose obesity. The traditional and most popular measure of obesity is based on an individual's body mass index (BMI), defined as weight in kg/height in m^2. Despite its widespread use there is a body of research arguing that BMI is, at best, a noisy measure of fatness since it does not distinguish fat from muscle, bone and other lean body mass. (for example Johansson et al. 2009, Burkhauser and Cawley 2008, McCarthy et al. 2006, Smalley et al. 1990). Consequently, a number of alternative measures of fatness have been proposed. These include percent body fat estimated using Bioelectrical Impedance Analysis (BIA) and measures based on Waist Circumference and Waist to Hip ratio. In the obesity literature to date researchers have settled on a specific, preferred measure as a gold-standard and used this measure to benchmark the other diagnostic tests. For example, Burkhauser and Cawley (2008) use obesity status defined on the basis of BIA to estimate the misclassification rates associated with BMI based measures. They 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. These estimates are based on the assumption that the misclassification rates with the BIA methods are zero.

In this paper we take a different approach to comparing the accuracy of alternative measures of obesity which is 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." Rather than specifying a gold-standard ex-ante we allow all measures to be potentially imperfect measures 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 the 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 to estimate the true underlying prevalence of the disease along with measures of the sensitivity and specificity of each of the tests (see for example Walter and Irwig 1988, Biemer and Wiesen 2002 and Biemer 2011).[1] 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 2012). However, to our knowledge latent class analysis has not been used to evaluate alternative measures of obesity.

Using data from a representative sample of US adults we show that that while obesity rates based on Body Mass Index and Bioelectrical Impedance Analysis misclassify large numbers of individuals, this is not the case for measures based on Waist Circumference. The error rates for Waist Circumference measures of obesity are of the order of 3% compared to error rates as high as 45-50% with the BMI and BIA approaches. This has important implications for the measurement and classification of obesity and suggests that Waist Circumference measures may provide a cheap effective means of classifying obesity. Furthermore the latent class approach allows us to compare estimated true prevalence rates of obesity across racial groups. The estimated true racial gap in obesity for women is similar to that based on BMI, both of which in turn are significantly higher than that gap suggested by the BIA method. In contrast however, the BMI approach suggests no difference in the obesity rate between black and white men, while our estimated true rates imply a significantly lower obesity rate for black men, which is in keeping with the findings from the BIA analysis. The fact that neither the BMI nor BIA based measures of obesity are capable of consistently measuring the racial gap for both men and women highlights the dangers associated with measuring obesity and the potential costly policy mistakes that may arise from arbitrarily adopting a single measure as a gold standard.

[1] Discrepant Analysis (DA) and Composite Reference Standards (CRS) have been proposed as alternatives to latent class analysis when assessing the accuracy of diagnostic tests in the absence of a gold standard (see for example Alonzo and Pepe 1999). The DA approach may be biased even when carried out under ideal conditions (Miller 1998). The CRS approach requires initial judgements about the characteristics of the existing tests in order to form the composite reference standard. Such prior information may not be available.

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, while section 4 presents our key results. Section 5 concludes.

2. Methods: 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_{1i}, T_{2i}, and T_{3i} denote three alternative tests designed to measure outcome C. In our application C_i is a dichotomous variables indicating the presence or otherwise of true underlying obesity, while T_{1i}, T_{2i}, and T_{3i} are dichotomous indicators of C. 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. Also let $\pi_{c,t1,t2,t3}$ denote the probability of an observation falling in this cell. Likewise π_c = Pr(C=c) for c=1,0 and$\pi_{t2|A}$=Pr(T2= t2|A). So for example $\pi_{t2|t1,c}$=Pr(T_2= t_2|T_1=t_1, C=c).

Using the law of conditional probabilities

$$\pi_{c,t1,t2,t3} = P(C = c)P(T_1 = t_1|C = c)P(T_2 = t_2|T_1 = t_1, C = c)\ P(T_3 = t_3|T_2 = t_2, T_1 = t_1, C = c) = \pi_c\pi_{t1|c}\pi_{t2|t1,c}\pi_{t3|t2,t1,c}$$

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

$$\pi_{t1,t2,t3} = \sum_c \pi_c\pi_{t1|c}\pi_{t2|t1,c}\pi_{t3|t2,t1,c}$$

This is a mixture model with unobserved regimes determined by π_c.

Let $n_{t1,t2,t3}$ denote the number of observations in cell (T_1=t_1, T_2=t_2, and T_3=t_3) and assume that the cell counts are distributed as a set of multinomial random variables. Then the kernel of the likelihood of observing the full table {T_1, T_2,T_3} is

$$L(T_1, T_2, T_3) = \prod_{t1}\prod_{t2}\prod_{t3} \pi_{t1,t2,t3}^{\,n_{t1,t2,t3}}$$

In total we have 2^3=8 possible cells, but since the probabilities must sum to 1 we only have 2^3-1=7 degrees of freedom. Unfortunately in this model there are 15 parameters to estimate:

$$\pi_1, \pi_{t1=1|1,}\pi_{t1=1|0,}\pi_{t2=1|1,1}\pi_{t2=1|1,0}\pi_{t2=1|0,0}\pi_{t2=1|0,1}$$

$$\pi_{t3=1|1,1,1}\pi_{t3=1|1,1,0}\pi_{t3=1|1,0,0}\pi_{t3=1|0,0,0}\pi_{t3=1|0,1,1}\pi_{t3=1|0,1,0}\pi_{t3=1|0,0,1}\pi_{t3=1|1,0,1}$$

Therefore in order to proceed we must impose some restrictions on the model. The standard identifying restrictions in this approach is to assume that the three tests are independent conditional on true status. This is known as local independence assumption (LIA) and specifies that the errors in the three tests are mutually independent. Models that allow for conditional dependence between tests typically require results from at least four different tests in order to be identified.[2] While LIA need not be true in general, in Section 3 we will argue that it may be reasonable in the context of our analysis.

LIA implies that $\pi_{t2=j|1,1} = \pi_{t2=j|0,1}$ and $\pi_{t2=j|1,0} = \pi_{t2=j|0,0}$ which eliminates two parameters and also $\pi_{t3=j|1,1,1} = \pi_{t3=j|0,1,1} = \pi_{t3=j|0,0,1} = \pi_{t3=j|1,0,1}$ and $\pi_{t3=j|1,1,0} = \pi_{t3=j|1,0,0} = \pi_{t3=j|0,0,0} = \pi_{t3=j|0,1,0}$ which eliminates a further six parameters. Therefore the restrictions imposed by LIA reduce the number of parameters to 7 allowing us to identify the remaining parameters.

Letting $\boldsymbol{y}$ denote the data vector of joint test results; $\boldsymbol{y}$=(y_{111}, y_{110}, y_{100}, y_{000}, y_{011}, y_{101},y_{001},y_{101}) and $\boldsymbol{\pi}$ denote the (7x1) vector of parameters specified above we write the data generating process for our model Pr($\boldsymbol{y}|\boldsymbol{\pi}$) as

$$\boldsymbol{y}|\boldsymbol{\pi} \sim \text{multinomial}\left(n, (\pi_{111}, \pi_{110}, \pi_{100}, \pi_{000}, \pi_{011}, \pi_{101}, \pi_{001}, \pi_{101})\right)$$

where $\pi_{t1,t2,t2} = \sum_c \pi_c \pi_{t1|c} \pi_{t2|c} \pi_{t3|c}$.

For example

$$\pi_{111} = \pi_1 \pi_{t1=1|1} \pi_{t2=1|1} \pi_{t3=1|1} + (1 - \pi_1) \pi_{t1=1|0} \pi_{t2=1|0} \pi_{t3=1|0}$$

[2] 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.

$$= \pi_1 \pi_{t1=1|1} \pi_{t2=1|1} \pi_{t3=1|1} + (1-\pi_1)\left(1-\pi_{t1=0|0}\right)\left(1-\pi_{t2=0|0}\right)(1-\pi_{t3=0|0})$$

$\pi_{tj=1|1}$ 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. $\pi_{tj=0|0}$ 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.

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 the EM algorithm. 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($\boldsymbol{\pi}$). Branscum et al. (2005) provide a useful overview of Bayesian approaches to estimation of the sensitivity and specificity of diagnostic tests.

In particular uncertainty about the parameters is typically modeled using independent beta prior distributions:

$$\pi \sim \text{beta}\left(a_\pi, b_\pi\right)$$
$$\pi_{tj=1|1} \sim \text{beta}\left(\alpha_{1,j}, \beta_{1,j}\right), j = 1,2,3$$
$$\pi_{tj=0|0} \sim \text{beta}\left(\alpha_{0,j}, \beta_{0,j}\right), j = 1,2,3$$

The choice of the as and bs determine the degree of prior information on each of the parameters and imply probabilistic restrictions on the parameter vector $\boldsymbol{\pi}$. In results below we set all as and bs equal to 0.5 which corresponds to Jeffrey's uninformative priors.

The posterior distributions of the parameters are given by $\Pr(\pi|y) = \frac{\Pr(y|\pi)\Pr(\pi)}{\Pr(y)}$. However, evaluation of this distribution is difficult since it requires solving for the probability of the data over all possible parameter values. However, we note that the posterior distribution is proportional to the product of the likelihood function and the prior:$\Pr(\pi|y) \propto \Pr(y|\pi)\Pr(\pi)$. Markov Chain Monte Carlo (MCMC) provides a mean of sampling from the full posterior distribution given the above likelihood and priors. MCMC is a popular technique for generating random

draws from posterior distributions that may be only known up to a constant of normalization as it overcomes the need to evaluate the probability of the data (Gilks et al. 1996). Once we have generated sufficient draws from the posterior distribution using MCMC then a range of summary statistics, such as the median, mode and 95% credible interval can be computed to summarise the posterior distribution of each of the parameters.

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. For the Metropolis-Hastings algorithm, one starts off with an arbitrary value π_0 and then samples a candidate π_1 from some proposal distribution $q(.|\pi_0)$. For example $q(.|\pi_0)$ might be a multivariate normal distribution with mean π_0 and fixed covariance matrix. The candidate point π_1 is then accepted as the next iteration in the chain with probability equal $\min\left(1, \frac{\Pr(y|\pi_1)\Pr(\pi_1)q(\pi_0|\pi_1)}{\Pr(y|\pi_0)\Pr(\pi_0)q(\pi_1|\pi_0)}\right)$. If accepted the candidate point π_1 becomes the next iteration in the chain, if not the chain does not move and π_0 is used again to make the draw at the next iteration. This process is competed 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. The key feature of the Metropolis-Hastings algorithm is that the proposal distribution can have any form and the chain will converge to the required stationary posterior distribution (Gilks et al. 1996).[3]

3. **Data**

[3] While any proposal distribution will ultimately deliver a sequence of draws from the target distribution, the convergence of the chain to this target distribution will depend on choice of the proposal distribution. Therefore it is important to check convergence of the chain when using MCMC. We discuss this later in the paper. The Gibbs sampler is a special case of the Metropolis-Hastings algorithm wherein the random draw is always accepted. The key to the Gibbs sampler is that it only considers univariate conditional proposal distributions – only one element of the vector is sampled at a time with the remaining elements remaining fixed. Thus 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. 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)).

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 1988-1994. 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 many measures of fatness, in that it includes a range of alternative measures of body composition.

In this paper we focus on three alternative measures of fatness Body Mass Index (BMI), Waist Circumference (WC) and Bioelectrical Impedance Analysis (BIA). 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. Waist circumference measures of obesity are based on a numerical measurement of your 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 g tension. Men are classified as being at "high risk" of obesity if their waist circumference exceeds 102cm, while for women the threshold is 88cm. 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 of a current running through your body, theoretically we could get an estimate of how much fat-free and fat mass you have. The Valhalla Scientific Body Composition Analyzer 1990 B is the instrument used for the measurement of whole body electrical resistance (Bio-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. The National Institute of Health

(NIH) classifies a man if his PBF exceeds 25 percent and a woman as obese if her PBF exceeds 30 percent. 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 waist-circumference tells you the location of your body fat but not the absolute percentage of body fat and may be prone to standard measurement problems. 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)). Thus rather than taking one measure as a gold standard we treat all measures of fat available as a-priori imperfect measures of underlying latent fatness and use the latent class approach outlined in the previous section 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, which requires that observed associations between the three tests is fully explained by the disease status (errors in the three tests are independent). 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 while dehydration may be a major source of error for BIA, this is unlikely to 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 we believe 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 error across tests.

We carry out our analysis separately for four groups; white women, white men, black women and black men. We restrict attention to individuals aged between 18 and 64 and for women we excluded those women who were pregnant at the time of the examination. Excluding those with missing values on at least one of our three tests the final sample sizes were 2142 (white women), 1924 (white men), 1852 (black women) and 1629 (black men).

4. **Results**

Table 1 provides the prevalence rates for "at risk of obesity" for each of these groups using our three different diagnostics. There are clear and substantial differences in the prevalence rates using different measures. The BMI measure tends to return the lowest obesity rate of all three tests, while BIA returns the highest rate for all groups. 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 men, white 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 latent class analysis we need to consider the joint distribution of the three tests. There are eight different combinations of tests outcomes to consider when using three dichotomous tests. Table 2 provides the cross-classification of the three tests for each of our four groups. Looking down the rows in this table allows us to examine the level of agreement across the three tests. There is substantial variation in the consistency of the tests across the four groups. The level of agreement across the three tests (sum of first and last row) was 49.68% for white women, 63.64% for white men, 59.39% for black women and 77.94% for black men.

The data in Table 2 provide the raw input for our latent class analysis. Before looking at the results in detail Figures 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 and discarded from the analysis, leaving us with 20,000

draws from the assumed stationary distribution. Figure 1 provides a history trace of the simulations for every parameter, along with the median and the 95% credible interval. 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. 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. Figure 2 plots the density of the chains 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.[4]

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 across tests. The specificity rate of the BMI based test is relatively high for all four groups, implying that this test returns very few false positives. Therefore it is very unlikely that this test diagnosis someone as obese when in fact they are truly not 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 then the probability of a false positive is still only 1.5% for men. 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 and women and for black women, reaching a low of 55% for white women. Only for black men does the sensitivity rate exceed 80%. Thus the problem with BMI is not that it misclassifies non-obese people as obese but rather its failure to truly detect obesity when it is present. 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 weighting 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.

[4] We have also carried out formal Geweke test for convergence. This test splits the sample 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%. For none of our parameters or groups can we reject equality of the means.

It is also interesting to compare these estimated misclassification rates to those reported by Burkhauser and Cawley (2008). Like us 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 either our estimates or those of Smalley et al (2009). Part of the reason for this is that in contrast to the densiometric gold-standard used by Smalley et al (2009), Burkhauser and Cawley use PBF based on BIA as the gold-standard. However, as noted by Freedman and Perry (2000) while BIA can prove useful because of its low interobserver error, moderate costs and simplicity it "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." pg NHLBI (2000) pg1. The specific problems associated with the BIA are evident in column three of Table 3. Although the sensitivity of BIA is estimated to be 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 40-50%. This is in contrast to the 100% specificity rate assumed by Burkhauser and Cawley (2008). In contrast to BMI measured obesity, the probability of a false negative with BIA is very low but the probability of a false positive is high, suggesting that BIA overestimates true obesity rates. This can partly explain why the false negative rate reported by Burkhauser and Cawley for women seems so high; many of those classified as truly obese by Burkhauser and Cawley based on BIA are not in fact obese. Consequently the BMI classification is not a false negative but in fact a correct diagnosis. 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 measures during a menstrual cycle be considered when estimating body composition.

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 women and black women. Only in the case of sensitivity measure for black 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 Johannson et al. (2009) found that only waist circumference had a negative association with wages for women. Also Janssen et al (2004) and Wang et al (2005) found that that WC outperformed BMI at predicting health risk associated with obesity. Wang et al (2005) 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."

Finally the last column of Table 3 reports our estimated true prevalence of latent obesity derived from LCA. It is interesting to compare these estimates to the estimates based on other measures. In particular we follow Burkhauser and Cawley (2008) and examine racial differences in obesity rates. We first consider the raw obesity rates in Table 2. The 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% points higher than among white women, while there is less than 1% 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. The female racial gap is significantly reduced while the PBF measure implies a substantially higher obesity rate among white men. However, since both these measures appear to suffer from misclassification bias neither of these racial gaps need reflect actual racial differences in obesity. To determine actual racial differences we turn to the estimated true prevalence rates reported in Table 3. Our estimated true prevalence rates imply a racial gap for women that is similar to the gap using BMI (of the order of 12% points). However, while there is no male racial gap in BMI based obesity measures our estimated true rates imply a significantly lower obesity rate among black men though the gap of 6% points is smaller than that based on PBF (20% points). These

findings highlight the danger of relying on single measures such as BMI and BIA when comparing obesity rates.

5. 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 appear to 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 and women and black women. This has important policy implications since Waist Circumference is a very simple and cheap procedure. The fact that all our measurements were taken by trained physicians clearly limits the chance of misclassification, however the results for WC do suggest that if properly implemented this approach can be effective in classifying obesity. With this in mind a simple information campaign illustrating the appropriate procedure for measuring waist circumference could prove highly effective in the fight against obesity.

The importance of having accurate measures of obesity is evident in our findings on racial-obesity gaps. Our estimated true prevalence rates imply a racial gap for women, with black women being significantly more obese than white women.

However, the opposite is true men; the estimated true prevalence rates imply a significantly lower obesity rate among black men. The fact that neither the BMI nor the BIA based measures of obesity are capable of capturing both these features highlights the dangers associated with measuring obesity and the potential costly policy mistakes that may arise from arbitrarily adopting a single measure as a gold standard.

References

Albert P, Dodd L. A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error without a Gold Standard, *Biometrics* 2004; 60:427-435.

Alonzo, T. and M. Pepe (1999), "Using a Combination of Reference Tests to Assess the Accuracy of a New Diagnostic Test," *Statistics in Medicine,* 18, 2987-3003.

Antonanzas, F., Rodriguez R. feeding the Economics of Obesity in the EU in a Healthy Way. *European Journal of Health Economics* 2010; 11:351-353.

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

Biemer P, Wiesen C. Measurement Error evaluation of self-reported drug use: a latent class analysis of the US National Household Survey on Drug Abuse. *Journal of the Royal Statistical Association A* 2002; 165, Part 1:97-119.

Black M, Craig B. Estimating Disease Prevelance in the Absence of a Gold Standard. *Statistics in Medicine* 2002; 21: 2653-2669.

Branscum AJ, Gardner IA, Johnson WO. Estimation of Diagnostic-Test Sensitivity and Specificity through Bayesian Modelling. *Preventitive Veterinary Medicine* 2005; 68: 145-163.

Burkhauser R, Cawley J. Beyond BMI: The Value of More Accurate Measures of Fatness and Obesity in Social Science Research. *Journal of Health Economics* 2008; 27:519-529.

Cawley J. The Impact of Obesity on Wages. *Journal of Human Resources* 2004; 39: 451-474.

Dendukuri, N. and L. Joseph (2001), "Bayesian approaches to Modelling the Conditional Dependence between Multiple Diagnostic Tests," *Biometrics*, 57, pp. 158-167.

Freedman,D. and G.Perry (2000) "Body Composition and Health Status among Children and Adolescents," *Preventive Medicine* 31(2), S34-S53.

Gilks, W.R and P. Wild (1992), "Adaptive Rejection sampling for Gibbs Sampling," *Journal of the Royal Statistical Society: Series C*, vol. 41, No. 2, pp. 337-348.

Gilks, W. R (1996), Full conditional Distributions, in Markov Chain Monte Carlo in Practice (eds) Gilks, W. R, S.Richardson and D.J. Spiegelhalter, London, Chapman & Hall.

Gilks, W. R, S.Richardson and D.J. Spiegelhalter (1996), Markov Chain Monte Carlo in Practice, London, Chapman & Hall.

Gleichauf CN, Roe DA: (1989) "The menstrual cycle's effect on the reliability of bioimpedance measurements for assessing body composition," Am J Clin Nutr, 50:903-907.

Goodman L. Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika* 1974; 61(2):215-231.

Hadgu, A. and W. Miller (2001), "Letter to the Editor: Using a Combination of Reference Tests to Assess the Accuracy of a New Diagnostic Test," *Statistics in Medicine,* 20, 656-660.

Hiu S, Walter S. Estimating the Error Rates of Diagnostic Tests. *Biometrics* 1980; 36: 167-171.

Janssen I, Katzmarzyk PT, Ross R (2004). "Waist circumference and not body mass index explains obesity-related health risk". Am. J. Clin. Nutr. 79 (3): 379–84. PMID 14985210.

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. Economics and Human Biology 7, 36-45.

Johnson W, Gastwirth J, Pearson L. Screening without a "Gold Standard": The Hui Walter Paradigm Revisited. *American Journal of Epidemiology* 2001; 153(9): 921-924.

Joseph L, Gyorokos TW, Coupal L. Bayesian Estimation of Disease Prevelance and Parameters for Diagnostic Tests in the Absence of a Gold Standard. *American Journal of Epidemiology* 1995; 141: 263-72.

Konnopka A, Bodemann M, Konig H. Health Burden and Costs of Obesity and Overweight in Germany. *European Journal of Health Economics* 2011; 12:345-352.

Krul A, Daanen H, Choi J. Self-Reported Weight, Height and Body Mass index (BMI) in Italy, the Netherlands and North America. *European Journal of Public Health* 2010; 21:414-419.

Lunn, D., A. Thomas, N. Best and D. Spiegelhalter (2000), "WinBUGS – A Bayesian modelling framework: concepts, structure and extensibility," *Statistics and Computing,* Vol. 10, pp. 325-337.

Luzia Gonçalves, Ana Subtil,M. Rosário de Oliveira, Virgílio do Rosário, Pei-Wen Lee Men-Fang Shaio (2012) " Bayesian Latent Class Models in Malaria Diagnosis" PLoS ONE; Vol. 7 Issue 7, p1

McCarthy, H.D., Cole, T., Fry, T., Jebb, S.A. and Prentice A.M. (2006), "Body Fat Reference Curves for Children," International Journal of Obesity, 30, pp. 598-602.

Miller, W. (1998), "Bias in Discrepant Analysis: When Two Wrongs Don't Make a Right," Journal of Clinical Epidemiology, 51(3), 219-231.

National Task Force on Obesity. *Obesity: The Policy Challenges* http://www.dohc.ie/publications/pdf/report_taskforce_on_obesity.pdf, 2005.

NHLBI (2000) "The Practical Guide Identification, Evaluation, and Treatment of Overweight and Obesity in Adults,"

Rindskopf D, Rindskopf W. The Value of Latent Class analysis in Medical Diagnosis. *Statistics in Medicine* 1986; 5(1): 21-27.

Smalley, K., Knerr, A, Kendrick, Z., Colliver A, Owen O (1990), "A Reassessment of Body Mass Indices," American Journal of Clinical Nutrition, 52, pp. 405-408.ObesityAdults NHLBI Obesity Education

Wada, R., Tekin, E., 2010. Body Composition and Wages. Economics and Human Biology 8, 242-254.

Walter, S. and L. Irwig, (1988), "Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review," *Journal of Clinical Epidemiology*, 41(9), 923-937.

WHO. Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks. World Health Organisation 2009; WHO Press.

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

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

Test Outcome			White women	White Men	Black women	Black Men
BMI	WC	BIA	%	%	%	%
+	+	+	22.7	16.94	35.15	13.3
+	+	-	0	1.5	.10	2.82
+	-	+	0.51	.78	.81	2.70
-	+	+	18.86	8.84	18.68	2.14
+	-	-	0	.57	0	1.84
-	+	-	0.51	2.28	1.0	1.66
-	-	+	30.3	22.29	19.88	10.8
-	-	-	26.98	46.7	24.24	64.64
			100	100	100	100

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)

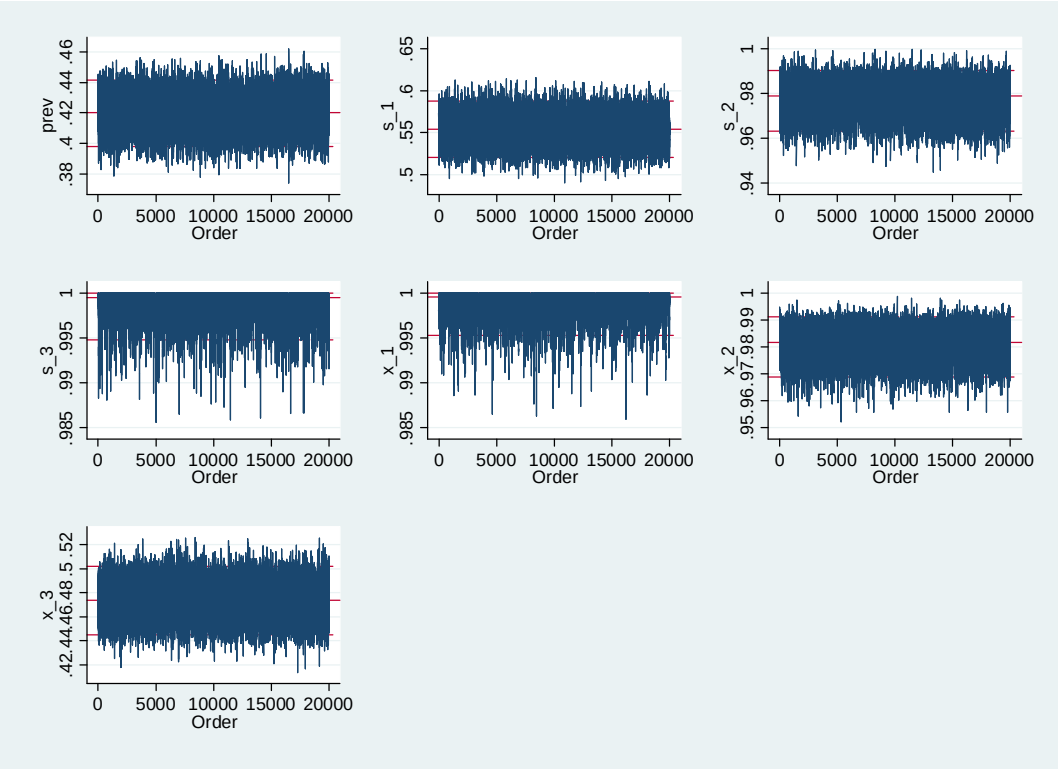

Figure 1a: History Plot of MCMC simulations: White Women

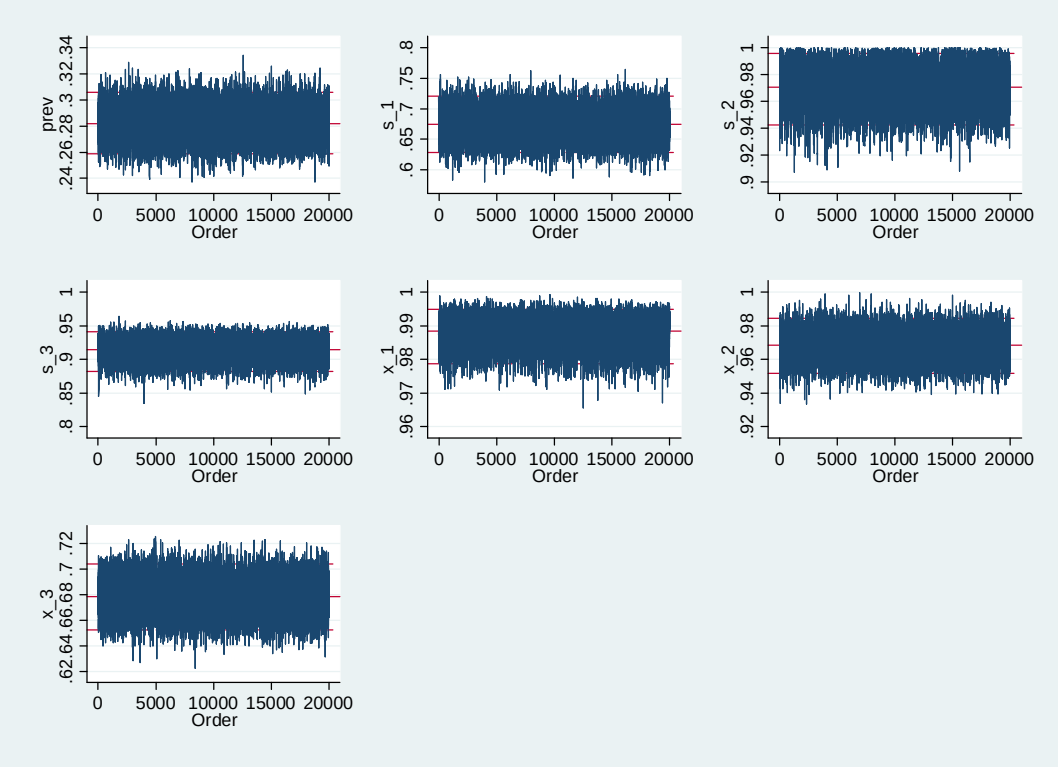

Figure 1b: History Plot of MCMC simulations: White Men

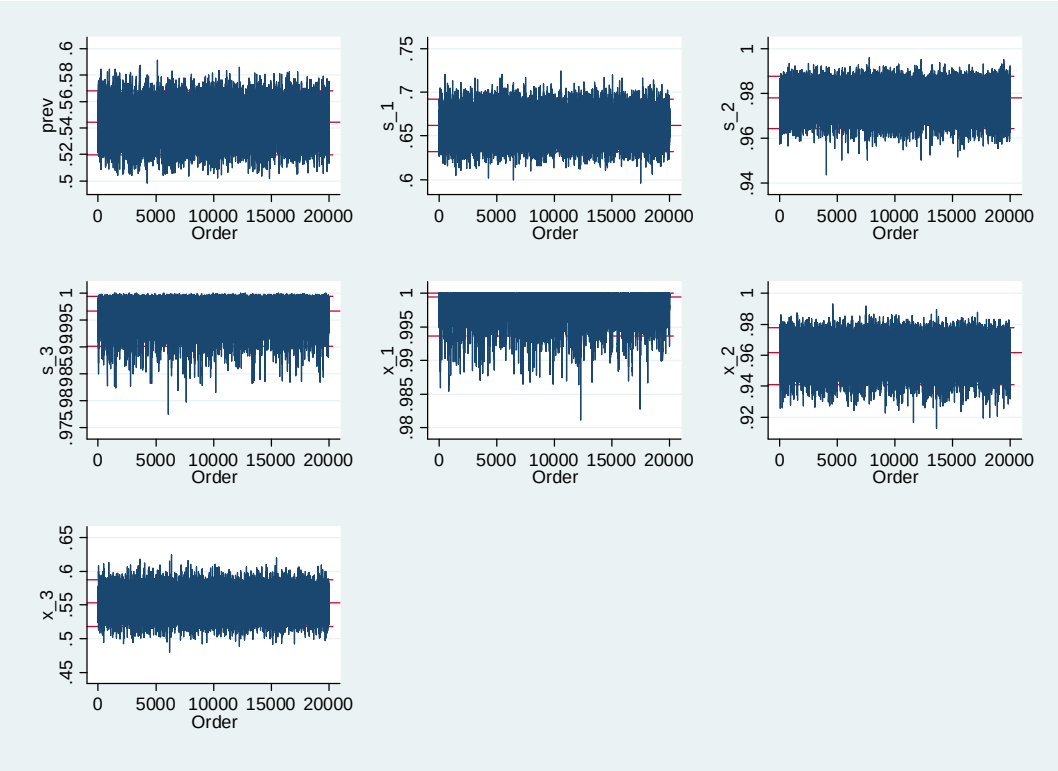

Figure 1c: History Plot of MCMC simulations: Black women

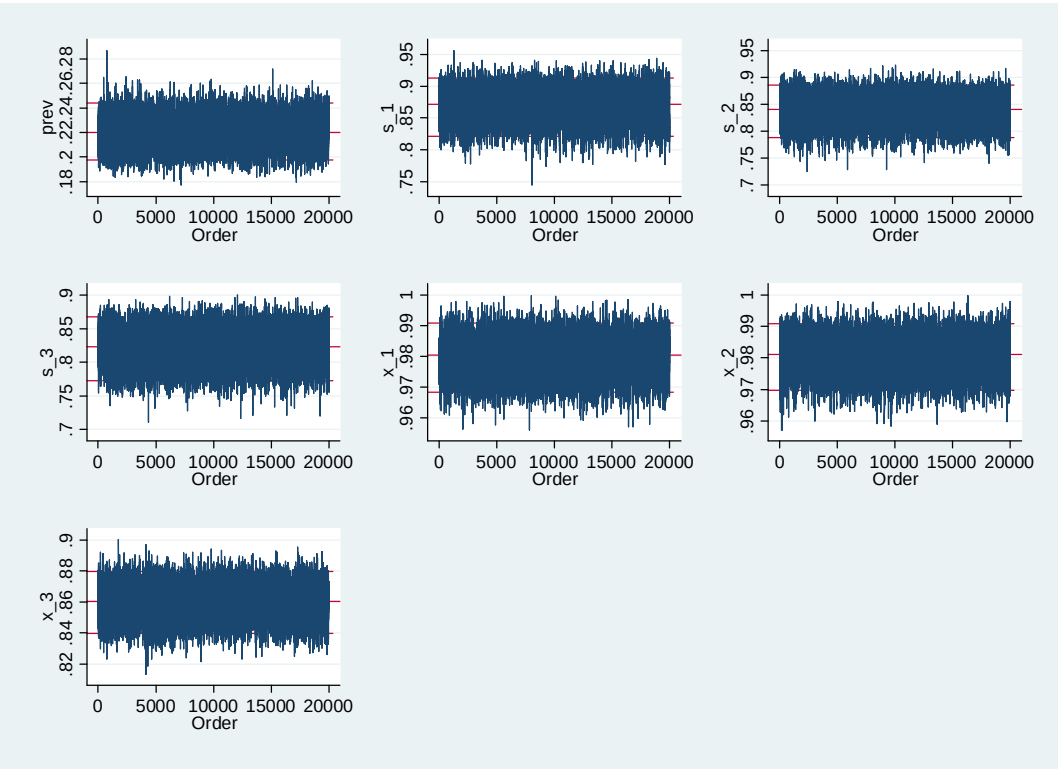

Figure 1d: History Plot of MCMC simulations: Black Men

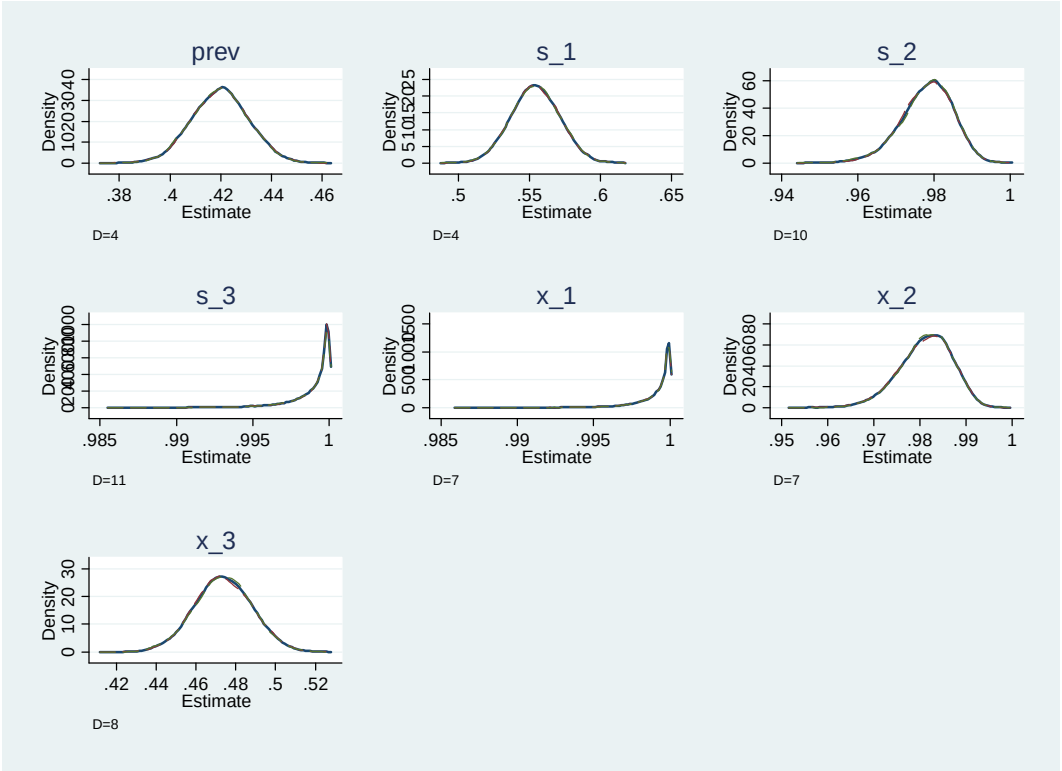

Figure 2a: Posterior density estimates for sections of the Markov chain. Dashed lines, densities for first and second half of the chain; solid line, density based on full chain: White Women

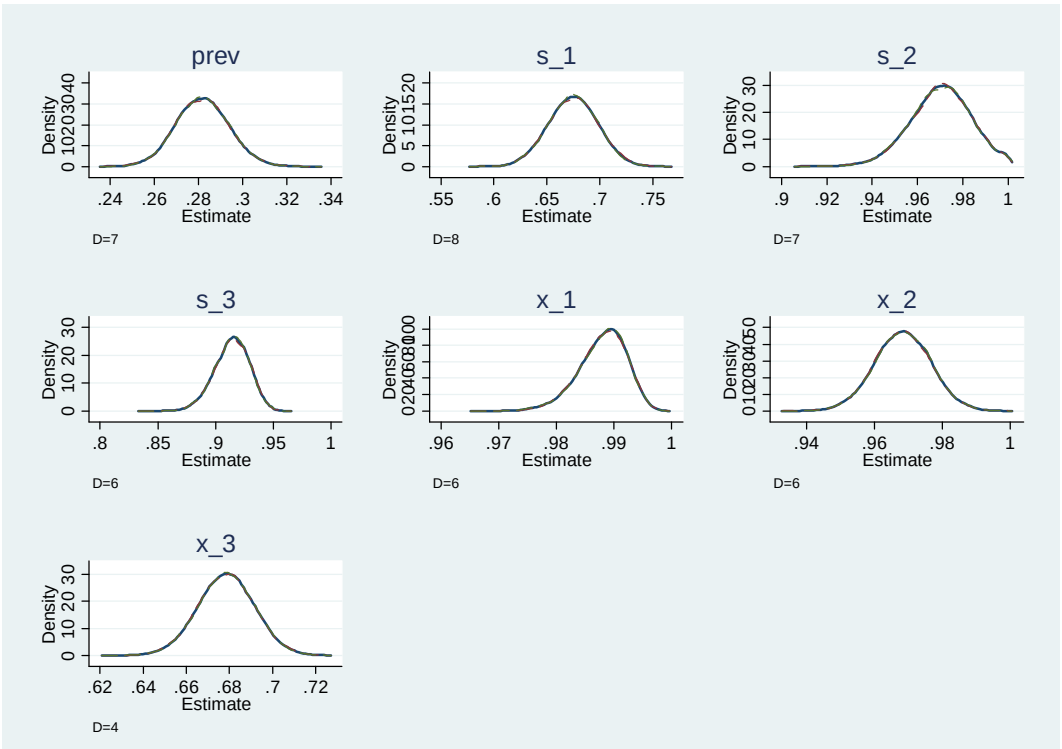

Figure 2b: Posterior density estimates for sections of the Markov chain. Dashed lines, densities for first and second half of the chain; solid line, density based on full chain: White Men

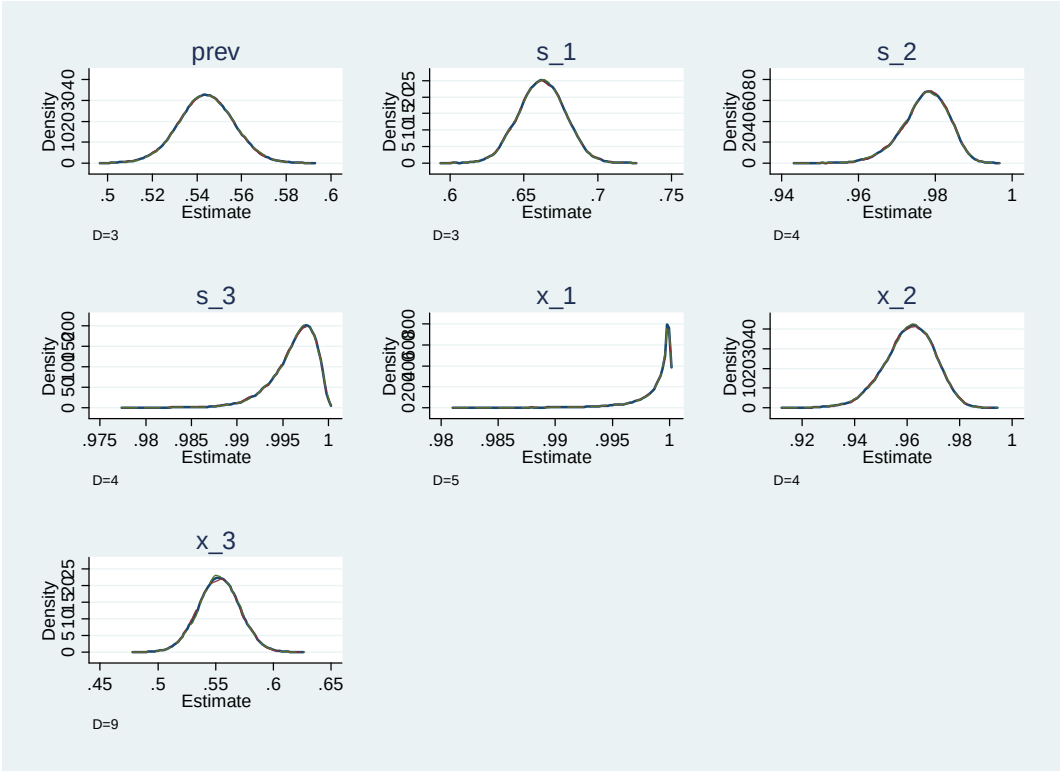

Figure 2c: Posterior density estimates for sections of the Markov chain. Dashed lines, densities for first and second half of the chain; solid line, density based on full chain: Black Women

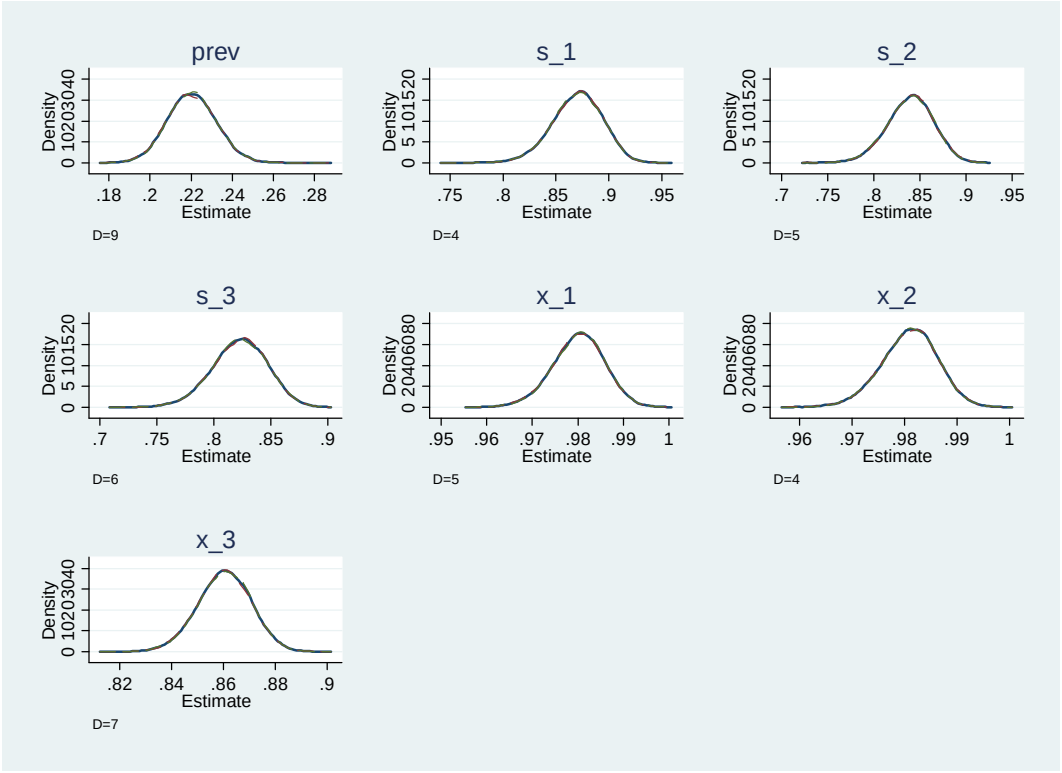

Figure 2d: Posterior density estimates for sections of the Markov chain. Dashed lines, densities for first and second half of the chain; solid line, density based on full chain: Black Men