Webappendix: Extended Methodology

1 Data

For each of the five surveys, stratified data were extracted by age, sex, population group and province. The information includes mean BMI and proportions of overweight and obese as well as samples sizes, sampling variability and confidence intervals. These were all calculated using the corresponding survey design. The cluster, strata and weight variables take into account the complex design of each survey sampling scheme.

In total there are 216 possible age (6) x population group (4) x province (9) combinations for which data could be extracted. However, there were zero or one observations for a number of these subpopulations. Obviously no means can be calculated if there are no observations, and standard errors could not calculated for factor combinations only one observation. Since these are used to fit the bayesian model, they were removed from the study. Furthermore, to insure comparability with the data used for the logistic prevalence models, the proportions without standard error estimates were also removed for that part of the study.

Samples for the minority population groups are thin as a whole, and are lacking entirely for some combinations of province and age group. The majority of missing data are in older age categories.

There are more data for females than males. All provinces were sampled, and Eastern Cape and KwaZulu Natal had the largest sample sizes in the study. The latter was oversampled to provide sufficient data for the Asian/India population group. Sampling occurred for both urban and rural areas.

The data for all surveys were in STATA format, and survey calculations were made using the *svy* command set. Output were in the form of tables which were

extracted to Excel and formatted for use in R and OpenBUGS. There are two reasons for doing this. First, the surveys include observations for different individuals in different years, which means the five datasets cannot be merged into one STATA file consisting of observations over the five different survey years, without losing the sampling design information, which is also different for each survey.

The statistics were calculated with each data file before extraction to the Excel spreadsheets, where data were allocated by year, age, population group and province. Therefore, there are six spreadsheets in total - mean BMI, proportion overweight and proportion obese for both sexes. At this stage, additional national BMI and proportion descriptive statistics were calculated. The national and subpopulation data were then read into R.

The national data were used to create figures in R for BMI distribution, means and prevalences throughout the period with error bars, and age distribution of BMI. The figures are included either in the main text or appendix A.

Variables had to be prepared for use since some were either of the incorrect type or in the wrong format. For instance, the standard errors calculated with the *svy* command are exported to Excel as strings, and had to be converted to numbers. Other necessary changes included converting age category factors to an integer variable with ages taken as midpoints of age groups, for fitting age as a continuous variable.

The preparation of variables was done separately for each of the three models fitted. A Bayesian regression model was used for mean BMI, and linear regression models were fitted with an appropriate transformation of response for overweight and obese.

2 Bayesian Model

In total, five different bayesian models were fitted and the model with the lowest deviance was chosen. Table 2.1 shows the models and their deviances. Model five is the model that was chosen, which includes linear time trend, age as a continuous variable with cubic splines, fixed effects for province and population group, and an additional variance term for each study. Table 2.2 describes the other models.

Table 2.1: Model selection for bayesian analysis of mean BMI

	Model	Deviance	DIC
Male			
	1	11 600	11 610
	2	6 925	6 939
	3	6 646	6 662
	4	6 551	6 569
	5	2 526	2 560
Female			
	5	3 335	3 358

Table 2.2: Model description

1	Age as a categorical variable
2	Age as a continuous variable
3	Age with a single cubic spline at age 40
4	Age with two cubic splines at ages 40 and 60
5	Added the additional variance τ_i terms.

The posterior distribution of the chosen bayesian model is given by uninform-

ative prior distributions for all parameters and the following likelihood, where t_i and z_j are the years and midpoint of survey i and age group j respectively.:

$$y_{i,j,k,h} \sim N(\beta t_i + \gamma(z_j) + \delta(k) + \zeta(l),\ s.e._{i,j,k,h}^2 + \frac{1}{\tau_i})$$

$$\begin{aligned} \text{survey } i &= 1, 2, ..., 5 \\ \text{age group } j &= 1, 2, ..., 6 \\ \text{population group } k &= 1, 2, ..., 4 \\ \text{province } l &= 1, 2, ..., 9 \end{aligned}$$

The mean of the likelihood is dependent on the linear equation of the effects of baseline mean and the time, age, population group and province association. The effects and their symbols are defined in sections 2.1 to 2.5. The variance component of the likelihood is related to the standard errors of each extracted mean and an additional term. Each parameter has an uninformative normal prior distribution, except the additional variability terms which have an uninformative gamma prior.

2.1 Fixed Effects of Covariates

For population group and province effects, a fixed effects model was used. The indicator variables have to be hard-coded for use in OpenBUGS, which does not accept factor variables from R.

The baseline for the fixed effects is Africans in the Eastern Cape. The averages for populations in other provinces can be calculated with the estimated coefficients of the following equations (the $I_{(\text{factor})}$ is an indicator variable for the levels of each factor):

$$\delta(k) = \delta_1 I_{\text{Eastern Cape/African}} + \delta_2 I_{\text{Free State}} + \delta_3 I_{\text{Gauteng}} + \delta_4 I_{\text{Natal}} + \delta_5 I_{\text{Limpopo}}$$
$$+ \delta_1 I_{\text{Mpumalanga}} + \delta_1 I_{\text{North-West}} + \delta_1 I_{\text{Northern Cape}} + \delta_1 I_{\text{Western Cape}}$$
$$\zeta(l) = \zeta_2 I_{\text{Asian/Indian}} + \zeta_3 I_{\text{Coloured}} + \zeta_4 I_{\text{White}}$$

$$\delta_k \sim N(0, \tau_\delta)$$
$$\zeta_l \sim N(0, \tau_\zeta)$$

There are no interaction effects included for the possible difference in mean BMI trends among different populations in different provinces. The would be included by including extra coefficients on each factor level multiplying the factor level by time.

The reason for not including the coefficients are both statistical and computational. The model could become over-parameterised when there is such little data. In other words, one can estimate a high number of coefficients relative to the number of data points, which would fit the data more closely but resemble the nuances of the specific sample data and not actual associations.

Furthermore, the more parameters that are included in the model, the more complex the form of the posterior. This becomes computationally onerous when implementing Monte Carlo Markov Chain algorithms like Gibbs Sampling and Metropolis-Hastings, especially when handled in a general manner as in the case of OpenBUGS. In other words, the model implementation procedure could be improved by manually simplifying the conditional posterior functions used in the algorithms. Hence, with a longer time frame and higher computational capacity, including these and other extra terms could improve the model.

2.2 Linear time components

The model includes a single parameter for estimating the trend of mean BMI.

$$\beta \sim N(0, \tau_\beta)$$

Each survey i corresponds to a survey year t_i. Survey years are subtracted by 1998 so that zero is the baseline year, and t_i is an integer between zero and 14. The trend coefficient β accounts for increases in BMI through time above the estimated fixed effects of population group and province.

The extracted means indicated that there was a sharper rise in BMI, overweight and obesity in the latter half of the period. However, the final model does not include nonlinear changes in time. Nonlinear trends were fitted and were not found to be significant, due to the sparsity of data in time.

2.3 Age group

Mean BMI is non-linearly associated with age. The model includes power-truncated splines, with knots for cubic powers at ages 40 and 60. These knots were initially chosen by observing the curved relationship of mean BMI and age. They were also evaluated against splines at other ages, and therefore chosen for their best improvement in the statistical model.

Baseline age was set to the age group three (35-44) by subtracting 40 from all midpoints. The following equation relates to the age term in likelihood function of the model.

$$\gamma(z_j) = \gamma_1 z_j + \gamma_2 z_j^2 + \gamma_3 z_j^3 + \gamma_4 (z_j - 40)^3 + \gamma_5 (z_j - 60)^3$$

$$\gamma_{(1,..,5)} \sim N(0, \tau_\gamma)$$

Although it is possible that age association differs across surveys and populations in different provinces, it is assumed to be the same. Extra variation for this is included to compensate.

2.4 Residual Variability

An extra term for each survey was included in the variance component of the likelihood. It accounts for the additional variability in age patterns of subpopulations that do not follow the general age pattern for the country. This can be seen as a simple alternative to including interaction terms for age with population group and age with province, which does not over-parameterise the model.

The equation below describes how the relevant term for study i is included in the model.

$$\tau_i = \tau_1 I_{i=1} + \tau_2 I_{i=2} + \tau_3 I_{i=3} + \tau_4 I_{i=4} + \tau_5 I_{i=5}$$

$$\tau_{1,..,5} \sim Gamma(0.1, 0.1)$$

2.5 MCMC

Models are specified using text files in the OpenBUGS syntax. The package R2OpenBUGS is available to run models straight from R. The data, initialised parameters and specified burn-in are passed to OpenBUGS when calling a model with the *bugs* function.

Each chain was run a minimum of 20, 000 times to ensure convergence. The first 60% of all draws for the posterior were discarded as part of the burn-in period. Plots of the chain for each parameter are shown below. The level of

convergence can be evaluated by deterring the range of values, after the burn-in, used to estimate posterior means. All of the models were fitted with enough iterations to conclude convergence was reached. The standard deviation and the credibility intervals of each parameter were also evaluated for plausibility before the model was validated. All these figures and estimates used are available in the R2OpenBUGS output, example from the final model shown below as well.

2.6 Goodness of fit

Goodness of fit is measured by the discrepancy between data and the model. A standard measure is the deviance.

$$deviance = D(y, \theta) = -2logL(y, \theta)$$

This depends on the values for parameters θ and the values of y. To get an estimate of the model deviance, one can average the deviance over the samples from the posterior simulation.

$$\hat{D}_{avg}(y) = \frac{1}{L} \sum_{l=1}^{L} D(y, \theta_l)$$

The lowest deviance represents the least distance or discrepancy and is preferred. However, the statistic does not take into account the number of parameters or model complexity. Another statistics is needed to compare different models on the same data.

Gelman *et al.* (2013) recommend the deviance information criteria (DIC) criterion for model fit based of predictive power. The aim of this model is to estimate future trend in BMI growth and, hence, the DIC is used as the selection criterion. It is defined as:

$$\hat{D}_{avg}^{pred}(y) = 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

where

$$D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y))$$

2.7 OpenBUGS and computation

As stated, the more complex the model, the longer it takes for the chain algorithm to run. The first few models took minutes to run, while for the final model, the first chain took over 20 hours to complete. The time depends on where the chain is initialised (starting values used for each parameter). If the chain is initialised within the poster distribution, it has essentially already converged and takes much quicker to run.

The OpenBUGS flexibility does also create two shortfalls. Firstly, any model can be specified regardless of whether it makes sense. Hence, a model may be correctly specified and results estimated without knowing they are meaningless. Secondly, the flexibility requires that the algorithms are sufficiently general. This means BUGS runs slower than coded models.

3 Linear regression models

Normal linear regression models were fitted to the proportion data. Typically, modelling proportions is done with logistic regression. This requires individual-level data for response indicators. However, it is difficult to use individual level data in R when considering the different survey designs and the need to take them into account to accurately reflect population characteristics. The proportions extracted from STATA, which account for design, do not have individual-level data and so an alternative fitting procedure is required.

The linear equation used by Stevens *et al.* (2012) for predicting overweight and obese prevalences is applied in this study, with the exception of not using any socio-economic variables, and including population group. In addition to the set of predictors used in the bayesian model, prevalence is modelled against BMI and higher power splines.

The proportions are transformed using a logit function so that the response has the same range as the linear function of parameters.

A number of models were fitted to determine the effectiveness of age and BMI splines. Table 3.1 and 3.2 describe the exact parameters included in these different models. Quadratic and cubic splines are fitted to try improve models, based on observing relationships of logit proportions and age and mean BMI. Based on insignificant increase in R^2 and reduction in deviance, fitted cubic terms and splines to age and BMI did not improve both overweight and obesity models, and were excluded in the final model.

The final model equation:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 \text{year} + \beta_2 m + \beta_2 m^2 + \beta_4 \text{age} + \beta_5 \text{age}^2 + \beta_6.\text{population group} + \beta_7.\text{province}$$

where m is mean BMI of the subpopulation that has proportion p (overweight or obese). Of these parameters, the

Residual analysis was used for model validation.

Table 3.1: Models of proportion overweight

	Model Number					
	Male			Female		
Model Components	1	2	3	1	2	3
Year	X	X	X	X	X	X
Province	X	X	X	X	X	X
Population Group	X	X	X	X	X	X
Age	X	X	X	X	X	X
Age^2	X	X	X	X	X	X
$(\text{Age-40})_+^2$		X			X	
Age^3			X			X
$(\text{Age-40})_+^3$			X			X
BMI	X	X	X	X	X	X
BMI^2	X	X	X	X	X	X
$(\text{BMI-23})_+^2$		X			X	
$(\text{BMI-28})_+^2$		X			X	
BMI^3			X			X
$(\text{BMI-23})_+^3$			X			X
$(\text{BMI-28})_+^3$			X			X
R^2	0.7409	0.7422	0.7423	0.6825	0.6852	0.6884
Deviance	187.40	186.48	186.36	204.21	202.48	200.44

Table 3.2: Models of proportion obese

	Model Number					
	Male			Female		
Model Components	1	2	3	1	2	3
Year	X	X	X	X	X	X
Province	X	X	X	X	X	X
Population Group	X	X	X	X	X	X
Age	X	X	X	X	X	X
Age^2	X	X	X	X	X	X
$(Age\text{-}40)^2_+$		X			X	
Age^3			X			X
$(Age\text{-}40)^3_+$			X			X
BMI	X	X	X	X	X	X
BMI^2	X	X	X	X	X	X
$(BMI\text{-}23)^2_+$		X			X	
$(BMI\text{-}28)^2_+$		X			X	
BMI^3			X			X
$(BMI\text{-}23)^3_+$			X			X
$(BMI\text{-}28)^3_+$			X			X
R^2	0.7101	0.7124	0.7133	0.7193	0.7195	0.7215
Deviance	214.00	212.28	211.58	171.80	171.67	170.44

References

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian data analysis*, CRC press.

Stevens, G. A., Singh, G. M., Lu, Y., Danaei, G., Lin, J. K., Finucane, M. M., Bahalim, A. N., McIntire, R. K., Gutierrez, H. R., Cowan, M., Paciorek, C. J., Farzadfar, F., Riley, L. and Ezzati, M. (2012). National, regional, and global trends in adult overweight and obesity prevalences, *Population Health Metrics* **10**(1): 22–22.
URL: *http://search.ebscohost.com/login.aspx?direct=truedb=cmedmAN=23167948site=ehost-live*

4 Code