Predicting Obesity Rate and Obesity-Related Healthcare Costs using Data Analytics

Article *in* Health Policy and Technology · February 2017

DOI: 10.1016/j.hlpt.2017.02.002

Sathish A.P Kumar
Cleveland State University

Ofir Ben-Assuli
Ono Academic College

Health Policy and Technology (▮▮▮▮) ▮, ▮▮▮-▮▮▮

Available online at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/hlpt

Predicting obesity rate and obesity-related healthcare costs using data analytics

Stephanie Revels[a], Sathish A.P. Kumar[a], Ofir Ben-Assuli[b,*]

[a]Coastal Carolina University, 119 Chanticleer Drive East, Conway, SC 29528, USA
[b]Faculty of Business Administration, Ono Academic College, 104 Zahal St., Kiryat Ono, Israel

KEYWORDS
Obesity;
Healthcare costs;
Data analytics;
Medical problems;
ARIMA model

Abstract
Objective: Obesity is a worldwide problem that has been linked to serious medical issues. Obesity-related conditions drain healthcare expenditures globally, and in particular in the U.S. This article suggests methods to forecast future costs associated with obesity-related healthcare in the next two decades.
Methods: An Auto Regressive Integrated Moving Average (ARIMA) time series analysis was implemented to model the data published by the Center for Disease Control and Prevention.
Results: The findings suggest that the proportion of individuals in the population defined as overweight will decline slowly in the next 20 years. However, the proportion of the population considered obese will increase substantially and could represent as much as 45% of the entire population by 2035. The proportion of morbidly obese will also increase considerably. These trends are likely to impact the actual costs of healthcare considerably.
Conclusions: Policy makers in the healthcare sector should be aware of this trend and prepare to deal with increasing numbers of medical problems related to obesity. Concrete recommendations for policy makers are put forward in the discussion as well as avenues for future research.

Introduction

Obesity is a worldwide problem that is known to entail a range of severe medical issues. Over the last five decades, the United States has been experiencing a rapidly growing health epidemic. From 1960 up to the most recent findings in 2012, the number of adults considered obese has increased approximately 260% [1] and the number of adults considered morbidly obese has increased about 733% as calculated from the data shown in Fig. 1. Though there are differences between genders in the prevalence of obesity [2], it has been declared a global epidemic [3]. Although debates have focused on whether these increases are due to high-caloric intake, sedentary lifestyles or some other

*Corresponding author.
E-mail addresses: slrevels@coastal.edu (S. Revels), skumar@coastal.edu (S.A.P. Kumar), ofir.benassuli@gmail.com (O. Ben-Assuli).

http://dx.doi.org/10.1016/j.hlpt.2017.02.002
2211-8837/

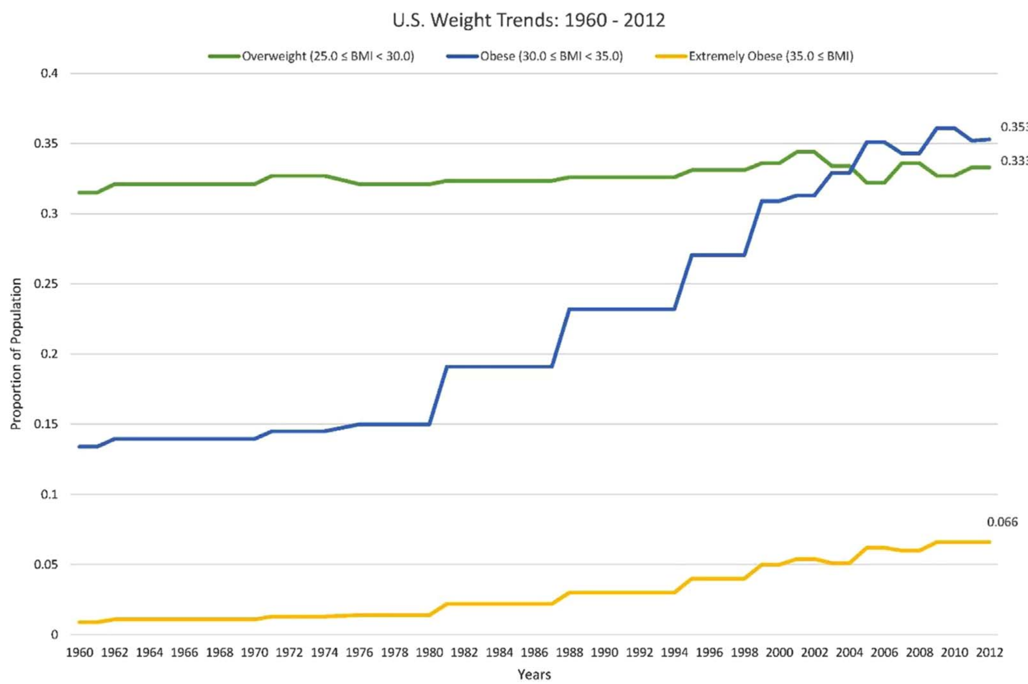

Fig. 1 Trends in overweight, obesity, and morbid obesity (Data obtained from the US National Health and Nutrition Examination Survey -NHANES).

factor, the impact has been clear- cut for healthcare expenditures in the U.S. and worldwide: healthcare costs will undoubtedly continue to rise [4]. Studies have shown that obesity is associated with a 36% increase in both inpatient and outpatient spending and a 77% increase in medications [5]. Total health care costs attributable to obesity may double every decade to reach 860.7 to 956.9 billion US dollars by 2030 [6]. Government intervention has been recommended as a way to stave off this costly epidemic [7].

Over 40% of all Americans are considered obese or morbidly obese. This trend is alarming, given the association between obesity and many chronic diseases including Type 2 diabetes, cardiovascular disease, several types of cancer (endometrial, postmenopausal breast, kidney, and colon cancers), musculoskeletal disorders, sleep apnea, and gall-bladder disease. The excess medical expenditures incurred from treating these obesity-related diseases are significant [8,9]. Currently, it is estimated that about 10% of all healthcare expenditures are spent on Body Mass Index (BMI)-related health issues [10], and as proven in the past, there is a direct correlation between BMI-related health issues and future healthcare costs [11]. These issues include hospital stays, doctors' visits, and prescription drugs associated with colon cancer, heart disease, high blood pressure and gall bladder disease, to name a few [12].

In particular, diabetes is considered a common direct cost-related outcome of obesity [13]. The increased prevalence of excessive visceral obesity and obesity-related cardiovascular risk factors is closely associated with the rising incidence of cardiovascular diseases and Type 2 diabetes mellitus. This clustering of vascular risk factors in (visceral) obesity is often referred to as the metabolic syndrome. There is a close relationship between an increased amount of visceral fat and metabolic disturbances, including low-grade inflammation [14]. Furthermore, obesity and diabetes significantly and independently increase the risk of Alzheimer's disease (AD). Though the level of risk is less than that with the apolipoprotein E4 (APOE4) allele, the high prevalence of these disorders may result in substantial increases in the future incidence of AD since physiological changes common to obesity and diabetes are thought to promote AD [15]. Similar to the recent increases in morbid obesity, diabetes has increased approximately 765% from 1960 to 2013 [16,17] and the medical costs associated with diabetes are also on the rise. In 2012, about $176 billion were spent on the direct costs of diabetes-related medical care [18]. In the past two decades, treatment of diabetes has improved tremendously such that there are fewer complications due to the disease, but the healthcare costs remain high because of its growing prevalence [19].

This study was designed to help forecast future costs associated with obesity-related healthcare by taking into account key variables such as the percentage of the adults in the U.S. who are overweight, obese, or morbidly obese, as well as national healthcare costs. In particular, it charts the relationship between obesity and rising healthcare costs through the use of data analytic methods.

There are also indirect non-medical costs associated with obesity-related illnesses. These include taking sick leave, lower wages, decreased productivity, and higher insurance premiums [20]. As these costs are difficult to measure and can vary considerably across individuals, indirect costs are not included in the analysis. The remainder of this article is organized as follows: Section 2 reviews the literature on obesity and details of the objectives of the study. Section 3 presents the methods and models. The results are presented in Section 4 and the conclusion and policy recommendations are discussed in Section 5. Section 6 covers limitations and future research directions.

Literature review

Allison et al. [21] examined whether the healthcare expenditures associated with obesity were counterbalanced by the increased mortality rate associated with obesity. The authors explored whether the medical costs of people who had returned to a normal-weight and lived longer differed from the medical costs of people who simply lived longer.

Allison et al. split the population into 5-year increments starting with the 20–24 age range. For each age range, the probability of obesity, a base mortality rate due to obesity, and the per capita healthcare costs were obtained and/or calculated. Then two sensitivity analyses were conducted: one that included mortality rates associated with obesity, and one that did not. The findings indicated that when including the mortality rates associated with obesity in the model, the healthcare costs were about 25% less than the projected healthcare costs for non-obese individuals.

However, their analyses had a number of limitations. First of all, the authors did not consider the costs associated with treating the obese such as prescriptions, operations, or support groups. Like the indirect costs associated with obesity, these costs are difficult to measure. These interventions might save expenditures on obesity-related medical conditions, but also might be costly to implement. They also did not include all possible obesity-related medical conditions which would be difficult to assess given the broad spectrum of possible health conditions. Furthermore, not all medical conditions associated with obesity are solely linked to weight. For example, normal-weight individuals can be born with diabetes, and individuals can contract heart disease after leading a perfectly healthy lifestyle. It is impractical to isolate which part of a medical condition often associated with obesity is actually due to obesity. The analyses also had relatively wide confidence intervals. While the confidence intervals suggested that healthcare costs were indeed lower when considering obesity mortality rates, it is uncertain whether these were marginally or significantly lower.

In another study, Thorpe et al. analyzed the ways in which obesity impacts healthcare spending [22]. The findings showed a significant difference between spending in 1987 and spending in 2001 on obesity, diabetes, and hyperlipidemia for those characterized as overweight or obese. Healthcare spending on the obese alone increased by 51%, compared to spending on any other category. Additionally, in 2001, spending on the obese was 37.2% greater than spending on individuals maintaining a normal-weight; this corresponded to an increase of 15.2% from 1987.

However, the most important limitation of their methodology is similar to that described for the Allison et al. study. Specifically, it was impossible to separate the costs spent on the major medical conditions based on a division into normal-weight, overweight and obese patients since these conditions, although more prevalent in overweight and obese patients, can affect normal-weight patients as well. Furthermore, the comorbidity list did not include every medical disorder associated with obesity. For example, gallbladder disease was not considered in this article although its likelihood increases significantly with rises in BMI [23].

Finkelstein et al. used data from the 1998 and 2006 Medical Expenditure Panel Surveys (MEPS) to assess the associations between obesity and rising healthcare costs [10]. This study found that the healthcare costs for an obese person in the sample were approximately 42% higher than those for a person with a normal-weight. Another important finding was that the steepest increase in healthcare expenditures attributed to obesity was the amount spent on prescription drugs: approximately $7 billion for non-institutionalized obese patients in 1998 alone.

A major limitation of this study was that MEPS allows respondents to self-report their heights and weights, which were then used to calculate BMI. Furthermore, MEPS only reports on individuals who are non-institutionalized, thus leading to a potential bias if the percentage of non-institutionalized obese patients differs from the overall percentage of obese patients.

Quesenbury et al. [24] found that medical costs associated with an individual in the 30.00 to 34.9 BMI range were 25% higher than an individual in the 20.00 to 24.9 BMI range, whereas an individual in the 35.00 and up BMI range had 44% more medical costs relative to the same group [24]. This study was designed to differentiate the costs associated with different healthcare issues between normal-weight individuals and the obese. The Quesenbury study tracked all healthcare costs for each patient for an entire year. This eliminated the need to include all possible healthcare issues associated with obesity and made it possible to isolate how many of those costs were linked to obese patients by relating each patient's medical costs to a BMI category.

While this study was able to accurately isolate costs associated with obesity, the heights and weights used to calculate BMI were self-reported. Initially, 610 observations had to be eliminated from the study due to missing or implausible data. Furthermore, all participants in the study were part of the Kaiser Permanente Health Care Plan. As can be expected, not all socioeconomic groups were equally represented. Since individuals had to be a part of the Kaiser Permanente Health Care Plan to be included in the study, those that could not afford healthcare or those that did not need healthcare were more likely to be underrepresented. The vast majority of literature dealing with BMI tends to conduct analyses within the traditional BMI ranges for normal-weight, overweight, and obese as we did as well. While this is not necessarily a limitation, it does make it challenging to use these results in combination with others.

Another study commissioned by the California Center for Public Health Advocacy investigated obesity in conjunction with physical inactivity [25]. The findings indicated that costs associated with obesity and physical inactivity were increasing. This may not seem surprising since obesity rates are likely highly to be correlated with physical inactivity, but not all physical inactivity leads to obesity. However, this study was unable to include indirect costs; hence medical situations unrelated to obesity were not measured. For example, a lingering medical condition that prevents an individual from being promoted and earning a higher salary

Please cite this article as: Revels S, et al. Predicting obesity rate and obesity-related healthcare costs using data analytics. Health Policy and Technology (2017), http://dx.doi.org/10.1016/j.hlpt.2017.02.002

could not be calculated. Although this study summarized obesity and healthcare costs for the entire state of California, not all regions were able to provide data and in these cases the data had to be estimated.

Formulating variables from the literature

This literature review helps pinpoint three key variables when forecasting future healthcare expenditures: the proportion of the population identified as overweight (BMI between 25.0 and 29.9), the proportion of the population identified as obese (BMI between 30.0 and 34.9) and the proportion of the population identified as morbidly obese (BMI greater than 35.0). Nevertheless, there is some overlap in the definitions of obese and morbidly obese across studies [26].

Using these figures, as well as the standard estimate that roughly 10% of United States healthcare costs are due to obesity, this study was designed to project the proportion of the population categorized as obese and the future national healthcare expenditures related to obesity. There is a heated debate over the proportion of healthcare costs is associated with obesity. Estimates range from as little as 5% to as much as 20% [10]. In an effort to be conservative, 10% was chosen here to estimate future healthcare costs.

Data sources and collection

Data for the following analyses were obtained from the Center for Disease Control and Prevention (CDC). From 1960 to 1990, data are only available for certain ranges. Thus, the missing data were imputed using the mean of the data surrounding these gaps for purposes of statistical analyses. The most recent results were released with 2012 update [27]. National healthcare expenditures were obtained from the Centers for Medicare & Medicaid Services [28].

Material and methods

ARIMA models

Analyzing trends in data over time is often referred to as time series analysis. The main goal of most time series analyses is to generate a model that predicts the future based on past data. One of the most common models in time series analysis is known as the autoregressive integrated moving average or ARIMA model [29]. While ARIMA models are commonly used for analyzing volatile time series data, these data do not necessarily need to fluctuate with a certain intensity. An ARIMA model is classified as ARIMA(p,d, q). This model is comprised of three separate components: an AR(p), an I(d), and an MA(q) process.

The first part is the autoregressive, or AR(p), process. This simply means that future predictions of the data will be based on p previous predictions of the data. For instance, an AR(1) model indicates that each successive observation is based on the previous observation, an AR(2) model indicates that each successive observation is based on the previous two observations, and so on. This kind of process is common in meteorology and finance where today's weather and prices often accurately predict tomorrow's weather and prices, respectively.

The second part of the ARIMA model is known as an integrated, or I(d), process. Time series data must be considered stationary in order to create accurate models. Time series are said to be stationary if the mean and variance do not vary over time. In the event that a time series is non-stationary, the time series data must be differenced before using the data to create a useful prediction model. This process of differencing is known as creating an integrated time series. For example, an I (1) model means that the time series needs to be differenced once. To difference the model once means to take the current observation and to subtract the previous observation for every observation in the dataset. Once a time series has been differenced, it is known as an integrated process.

The final part of the ARIMA model is the moving average, or MA(q), process. This method is used with the error terms in the series. If the error terms are not constant, an average must be taken of the current and the previous observation for every observation in the time series. For instance, for an MA(1) process, as each observation is added, a new average of the current and previous error term must be introduced into the model. For an MA(2) process, a new average of the current and the two previous error terms must be introduced into the model. Given that the average changes with each new observation included in the time series, including predictions, this process is aptly referred to as a moving average process.

Box-Jenkins method

In time series analyses, a helpful tool when building an ARIMA model is the Box-Jenkins method [30]. This method is made up of a four step process that begins with identifying the order of the model. This means determining the correct number of lags to include for each of the autoregressive, integrated, and moving average portions.

Software and packages

For this time series analysis, statistical software R, version 3.2.3 was used. The packages in R that were used for this analysis were forecast, ggplot2, timeSeries, and tseries.

Results

The first step in using the Box-Jenkins method is to determine the correct number of lags to include for each of the autoregressive, integrated, and moving average components. The second step in this process is to estimate the parameters of the model based on the previously determined order. The third step involves diagnostic checking with the model. This important step involves model validation with regression models. The fourth and final step is to forecast using the validated model.

Please cite this article as: Revels S, et al. Predicting obesity rate and obesity-related healthcare costs using data analytics. Health Policy and Technology (2017), http://dx.doi.org/10.1016/j.hlpt.2017.02.002

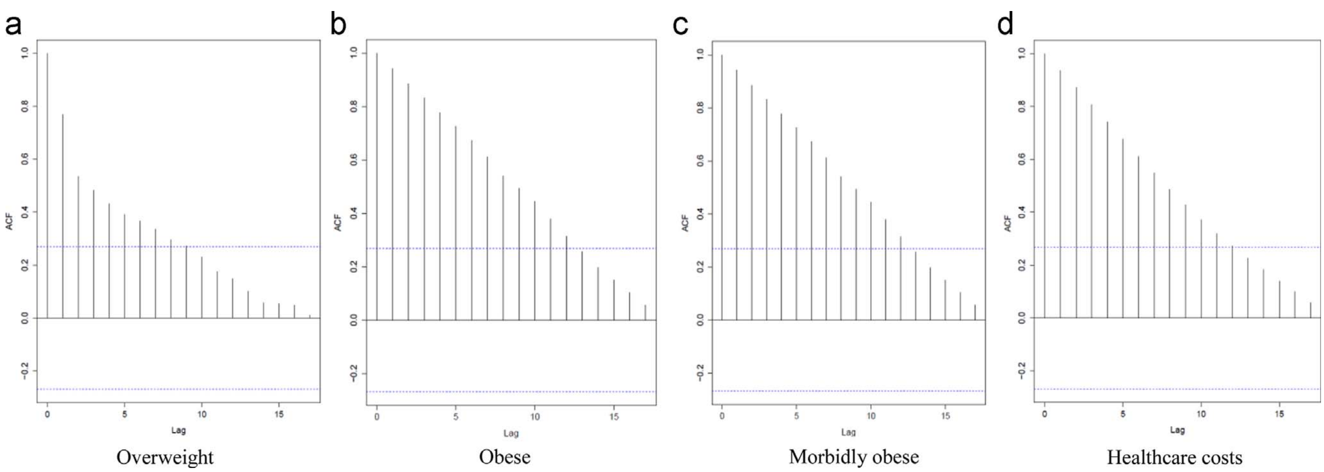

Fig. 2 ACF Plots. The dashed blue line indicates a significance threshold. When lags are consistently outside of the pair of dashed blue lines, as they are in the above figures, the trends are considered non-stationary. (a) Overweight. (b) Obese. (c) Morbidly obese. (d) Healthcare costs.

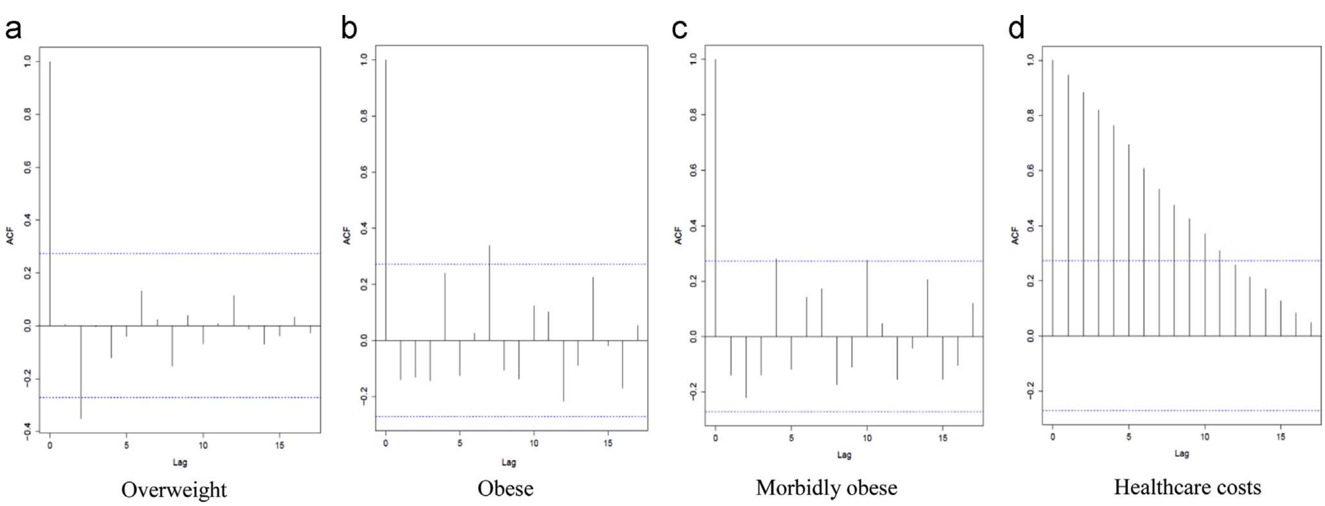

Fig. 3 ACF Plots of First Difference. (a) Overweight. (b) Obese. (c) Morbidly obese. (d) Healthcare costs.

Achieving stationarity

Before the Box-Jenkins method can be used, the model must be made stationary. Models that are non-stationary are not suitable for forecasting purposes. There are several methods to test for stationarity. One simple way to test for stationarity is to observe the Auto Correlation Function, or ACF plot [30]. The ACF plot displays lags. Lags show the correlation between successive observations. The ACF plot of a stationary time series will exhibit lags that drop to zero rapidly, whereas a non-stationary time series will experience a slower decline in lags. When using ACF plots, there is no specific threshold or rate of decay to determine the differencing needed [31]. As seen in Fig. 2a, the ACF of the overweight variable decays fairly quickly. This could indicate that the trend is stationary. If the process were non-stationary, the lags would stay well above zero and decay slowly. Examples of this decay can be seen in Figs. 2b, c, and d where the trends for obesity, morbid obesity and healthcare costs, respectively, drop slowly and thus appear to be non-stationary. Further analyses were performed here to definitively select the correct order of differencing. Specifically, the auto.arima function was applied to each of the time series in R. It was determined that the overweight, obese, and extremely obese trends would need a differencing of one, or I(1). The trend for healthcare costs trend needed a differencing of two, or I(2). It is important to note that a differencing of two is not the difference between the current observation and the previous observation lagged by two. Rather, it is the difference of the first set of differences.

Selecting the orders and estimating the parameters of the model

The next step is to tentatively select the orders of the AR (p) and MA(q) processes. The ACF and Partial Auto Correlation Function (PACF) plots are also useful for this, once any needed differencing has been applied. If a lag is significant

Please cite this article as: Revels S, et al. Predicting obesity rate and obesity-related healthcare costs using data analytics. Health Policy and Technology (2017), http://dx.doi.org/10.1016/j.hlpt.2017.02.002

it will remain above the top dashed line or below the bottom dashed line. It is important to note that for each increase in lag, all lags must be included. For instance, if only the 3rd lag seems significant and not the 1st or 2nd, then either all three lags or no lags must be included. It is impossible to include a certain lag in a time series model without including all prior lags. From the ACF in Fig. 3a, the first two AR lags appear significant whereas the PACF in Fig. 5a does not suggest that any MA lags are significant for the overweight trend. This suggests that an AR(1) an AR(2), or an AR(3), model could be used along with the previously identified differencing term. When including the differencing term, this would be an ARIMA(1,1,0), an ARIMA(2,1,0),

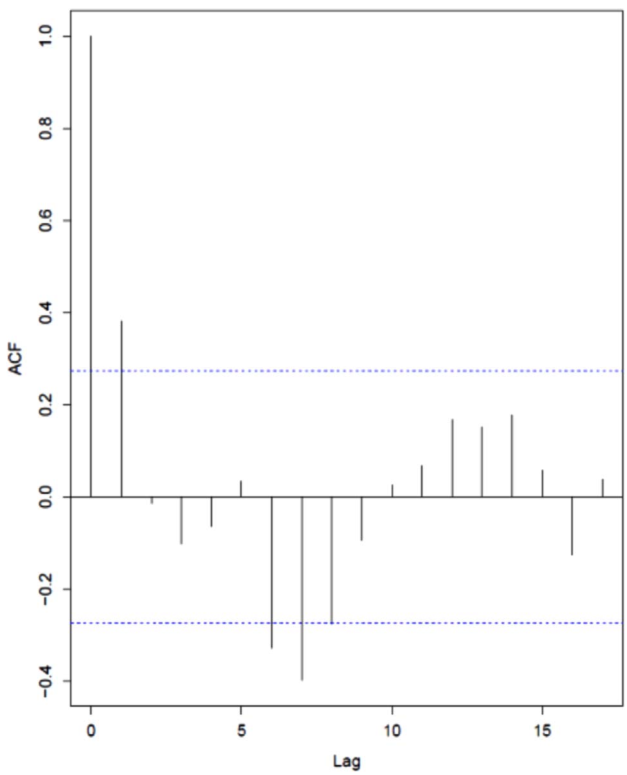

Fig. 4 ACF plot of second differences in healthcare costs.

or an ARIMA(3,1,0). From the ACF and PACF in Figs. 3b, 5b and 5c for the obese and morbidly obese trends, none of the AR or MA lags seem significant. This could suggest that these trends would be best fit with an ARIMA(0,1,0) where the model is simply differenced and no autoregressive terms and no moving average terms are included. An ARIMA(0,1,0) is also known as a random walk process [32].

As was found for the obesity and morbid obesity trends, the ACF of the healthcare cost data in Fig. 4 suggests that no AR terms are needed. However, in the PACF in Fig. 5d, the first lag appears to be significant. Thus, it is likely that these data should be modeled using an ARIMA(0,2,1).

After tentatively selecting the order of the models, the auto.arima function in R software was used to definitively select the orders as well as estimate the parameters. For the overweight time series, an ARIMA(2,1,0) best fit the data. The model is shown in Eq. (1) and the R code and its output results are shown in Appendix 1. The magnitude of the AR coefficients indicates the relative strength of the relationship to the current predicted value and the lags of the previous values. From Eq. (1), the second lag is expected to have a larger effect on the current predicted value than the first lag. It is often ill-advised to include large lags in any time series trend, but especially in a smaller data set such as the BMI data [33].

Using the same process as in R, models were created for the obesity, morbid obesity and healthcare cost time series. For obesity and morbid obesity, ARIMA(0,1,0) models produced the best fits. These models can be seen in Eqs. (2) and (3), respectively. These models forecast based only the previous value and a drift term, if applicable. Eqs. (2) and (3) have positive drift terms, as indicated by the constant terms in these equations. Positive drift terms suggest that both series are expected to experience an upward trend.

The healthcare data needed an ARIMA(0,2,1) which is displayed in Eq. (4). The first two terms in Eq. (4) account for the second difference, also known as the first difference of the first difference. This captures the change in the change in the value at time t. The last term indicates that forecasts will have about a two-year lag before reflecting current trends or turning points in the data. This two-year

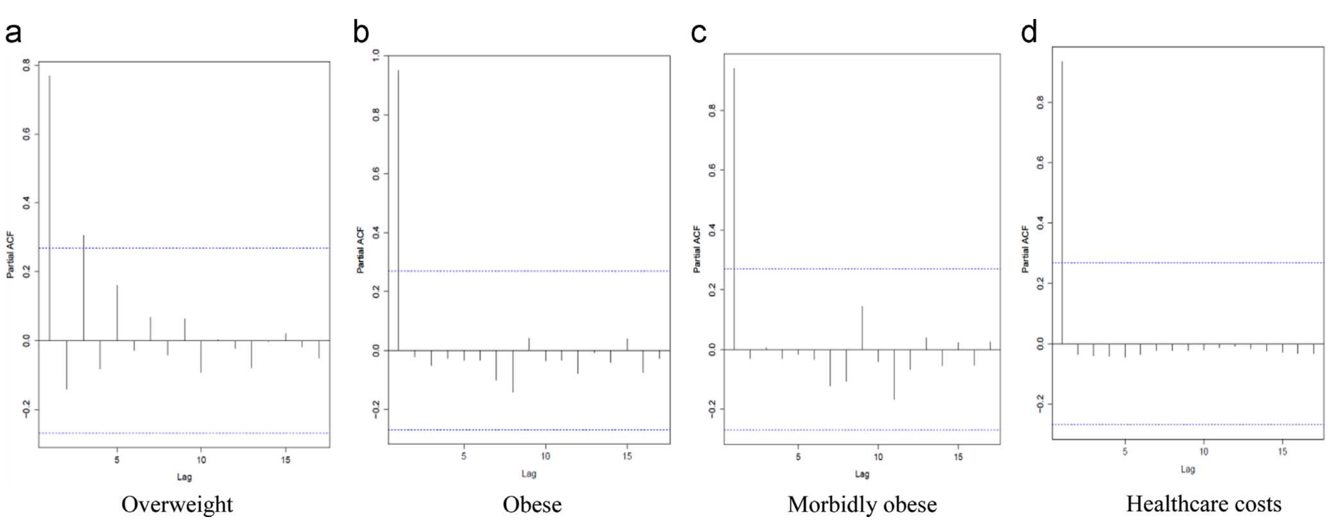

Fig. 5 PACF plots. (a) Overweight. (b) Obese. (c) Morbidly obese (d) Healthcare costs.

Please cite this article as: Revels S, et al. Predicting obesity rate and obesity-related healthcare costs using data analytics. Health Policy and Technology (2017), http://dx.doi.org/10.1016/j.hlpt.2017.02.002

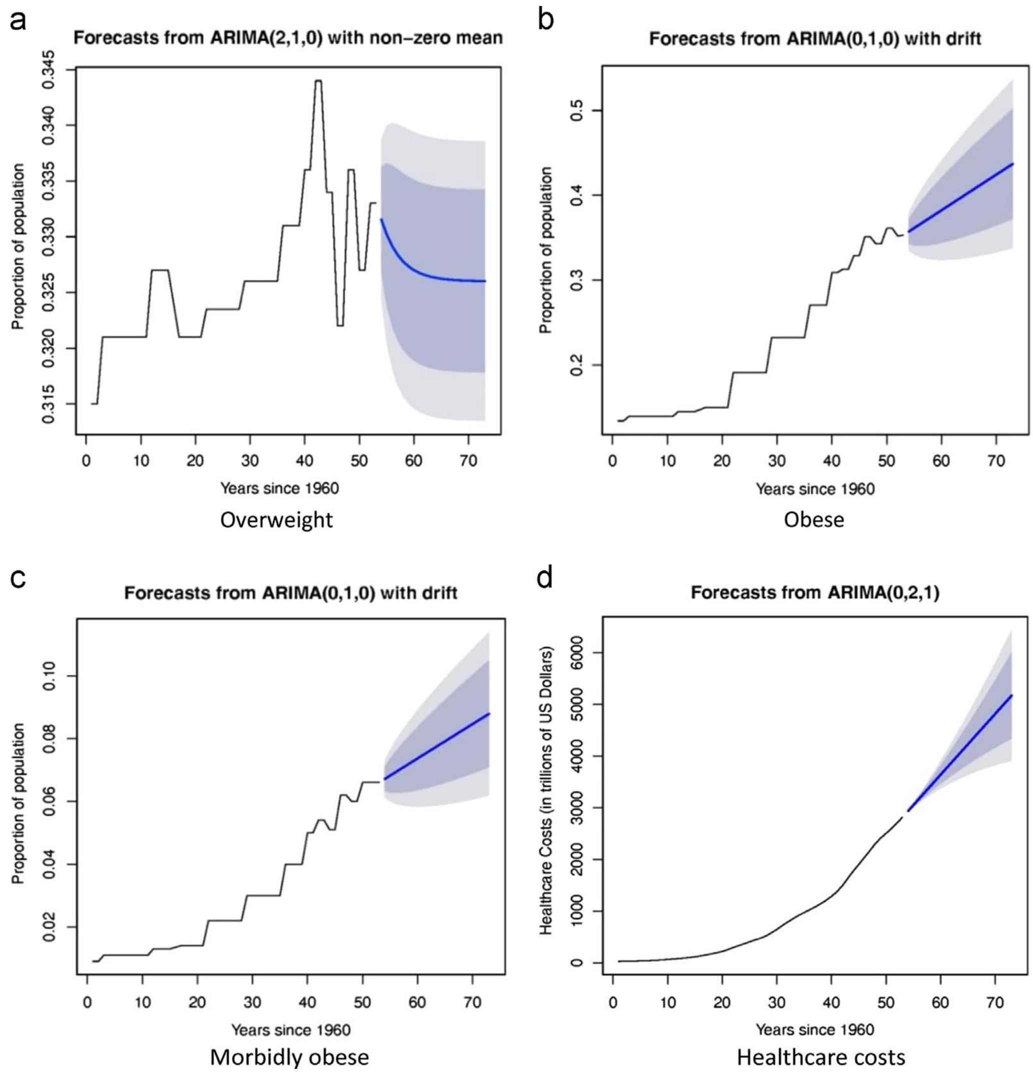

Fig. 6 Forecast using ARIMA models. (a) Overweight. (b) Obese. (c) Morbidly obese. (d) Healthcare costs.

lag is due to the results of the auto.arima function when applied to the overweight time series.

$$x(t) = x(t-1) + 0.0157 \cdot (x(t-1) - x(t-2)) - 0.3667 \cdot (x(t-2) - x(t-3)) \quad (1)$$

$$x(t) = 0.0042 + x(t-1) \quad (2)$$

$$x(t) = 0.0011 + x(t-1) \quad (3)$$

$$x(t) = 2 \cdot x(t-1) \text{-} x(t-2) - 0.5028 \cdot e^{(t-1)} \quad (4)$$

Diagnostic checking and validating the model

Before using these models to produce a forecast, the models need to be validated. For time series, the standard approach to model validation is to check the model for white noise. White noise means that the residuals are not auto-correlated [30]. In the presence of autocorrelations, a model is not sound enough to make accurate predictions. The Ljung-Box test has a null hypothesis that the residuals are white noise and the alternative is that at least one residual is auto-correlated. If the p-value of the Ljung-Box test is below 0.05, the null hypothesis of white noise will be rejected and the model will not be useful for forecasting. As shown in Appendix 2, each of the p-values of the Ljung-Box tests for the four-time series were greater than 0.05 in the R code. When applying the Ljung-Box test to each of the models created for the four time series, all the models appeared to exhibit white noise when the residuals were not auto-correlated. All four models were thus appropriate for forecasting. The R code and its output of the Ljung-Box test for the fit_overweight, fit_obese, fit_ext_obese and fit_costs are shown in Appendix 2.

Forecasting obesity using the validated model

From the ARIMA forecasts in Fig. 6, it appears that the proportion of the population that is overweight will drop slowly over the next two decades. This trend can probably be ascribed to the recent slight drops in the overweight population. Since 2008, the overweight population has dropped from 33.6% to somewhere between 32.7% and 33.3%. On the other hand, about 45% of the population

Table 1 Forecasts with 95% Confidence Intervals.

Year	Overweight	Obese	Morbidly Obesity	Healthcare Costs (in millions)
2016	0.33160 (0.31969,0.34350)	0.36985 (0.32534,0.41435)	0.07038 (0.05873,0.08204)	3288.51 (3170.20,3406.82)
2017	0.33130 (0.31821,0.34439)	0.37406 (0.32430,0.42382)	0.07148 (0.05845,0.08452)	3406.31 (3242.54,3570.08)
2018	0.33129 (0.31710,0.34548)	0.37827 (0.32376,0.43278)	0.07258 (0.05830,0.08686)	3524.11 (3310.27,3737.96)
2019	0.33140 (0.31631,0.34649)	0.38248 (0.32361,0.44136)	0.07367 (0.05825,0.08910)	3641.92 (3373.76,3910.07)
2020	0.33140 (0.31547,0.34733)	0.38669 (0.32375,0.44963)	0.07477 (0.05828,0.09126)	3759.72 (3433.34,4086.10)
2021	0.33136 (0.31459,0.34813)	0.39090 (0.32415,0.45766)	0.07587 (0.05838,0.09335)	3877.52 (3489.23,4265.81)
2022	0.33136 (0.31379,0.34893)	0.39512 (0.32475,0.46549)	0.07696 (0.05853,0.09540)	3995.32 (3541.66,4448.98)
2023	0.33137 (0.31305,0.34970)	0.39933 (0.32552,0.47313)	0.07806 (0.05872,0.09739)	4113.12 (3590.80,4635.45)
2024	0.33138 (0.31233,0.35043)	0.40354 (0.32645,0.48062)	0.07915 (0.05896,0.09935)	4230.93 (3636.79,4825.06)
2025	0.33137 (0.31162,0.35112)	0.40775 (0.32752,0.48798)	0.08025 (0.05923,0.10127)	4348.73 (3679.77,5017.69)
2026	0.33137 (0.31094,0.35180)	0.41196 (0.32870,0.49522)	0.08135 (0.05953,0.10316)	4466.53 (3719.85,5213.21)
2027	0.33137 (0.31029,0.35245)	0.41617 (0.32999,0.50236)	0.08244 (0.05986,0.10502)	4584.33 (3757.14,5411.52)
2028	0.33137 (0.30966,0.35309)	0.42038 (0.33137,0.50940)	0.08354 (0.06022,0.10686)	4702.14 (3791.73,5612.54)
2029	0.33137 (0.30904,0.35370)	0.42460 (0.33285,0.51635)	0.08463 (0.0606,0.108670)	4819.94 (3823.71,5816.16)
2030	0.33137 (0.30844,0.35430)	0.42881 (0.33440,0.52322)	0.08573 (0.06100,0.11046)	4937.74 (3853.16,6022.32)
2031	0.33137 (0.30785,0.35489)	0.43302 (0.33602,0.53002)	0.08683 (0.06142,0.11224)	5055.54 (3880.13,6230.95)
2032	0.33137 (0.30728,0.35546)	0.43723 (0.33771,0.53675)	0.08792 (0.06185,0.11399)	5173.34 (3904.71,6441.98)
2033	0.33137 (0.30673,0.35602)	0.44144 (0.33947,0.54342)	0.08902 (0.06230,0.11573)	5291.15 (3926.95,6655.35)
2034	0.33137 (0.30618,0.35656)	0.44565 (0.34128,0.55003)	0.09012 (0.06277,0.11746)	5408.95 (3946.90,6871.00)
2035	0.33137 (0.30565,0.35710)	0.44987 (0.34314,0.55659)	0.09121 (0.06325,0.11917)	5526.75 (3964.62,7088.88)

could be obese by 2035 (see Table 1). Combining this with the percentage of the population that is overweight, it is possible that approximately 3 out of 4 U.S. citizens will be considered overweight or obese by that date.

Discussion, conclusions and policy-making implications

Obesity is a worldwide problem that is known to be associated with serious medical problems [34]. It affects healthcare expenditures globally and in particular in the U.S. However, despite efforts to stem the tide, childhood obesity has reached epidemic proportions in the United States [35]. In addition, obesity is responsible for a growing proportion of healthcare expenditures, and healthcare costs will undoubtedly experience substantial increases [36].

In this paper we forecast future costs associated with obesity-related healthcare. For instance, obesity is known to lead to much greater incidence of diabetes [18].

We used one of the most common models in time series analysis, the autoregressive integrated moving average, or ARIMA, model [29] on the data obtained from the Center for Disease Control and Prevention.

Our results show that the proportion of the population that is overweight will drop slowly over the next two decades. However, the proportion of people who are obese will increase and could be as high as 45% by 2035. When combined with the population that is overweight, approximately 3 out of 4 U.S. citizens will be overweight or obese by that date. In addition, the proportion of the morbidly obese will increase substantially. Clearly the actual financial burden on healthcare will follow suit.

With national healthcare costs expected to be around $5.5 trillion in 2035, the costs associated with obesity could be around $550 million at the conservative estimated 10% (Table 1, the year 2035). Furthermore, if the national trends since 1960 continue, around 94% of Americans will be considered overweight or obese, with a little less than 10% of the population falling in the morbidly obese population. While it is difficult to estimate the increment in weight-related healthcare costs for a morbidly obese person compared to person at a normal-weight, there is little doubt that healthcare costs are generally higher for the former.

In addition to direct medical costs, there are indirect costs arising specifically from obesity. These costs include productivity loss originating in the labor market, including absences from work (caused directly from obesity), premature mortality and the loss of quality-adjusted life years (QALYs), higher rates of disability benefit payments and welfare loss in the health insurance market [37].

Obesity also has social costs [38]. Obesity can substantially decrease people's standard of living and can lead to a significantly shorter life span [39].

Although researchers can formulate increasingly better predictions of the obesity epidemic, forecasting will not solve the problem. More research needs to be done to understand how to best stem the obesity epidemic.

The traditional ways of preventing and treating people who are overweight or obese have almost invariably focused on changing their behavior, an approach that has proven woefully inadequate. Considering the many areas of American culture that promote obesity, from the proliferation of fast-food outlets to almost universal reliance on automobiles, reversing current trends will require a multifaceted public health policy approach as well as substantial funding. National leadership is needed to ensure the participation of health officials, researchers, educators, legislators, transportation experts, urban planners, businesses, and nonprofit groups in formulating a public health campaign with a better chance of success [40].

Please cite this article as: Revels S, et al. Predicting obesity rate and obesity-related healthcare costs using data analytics. Health Policy and Technology (2017), http://dx.doi.org/10.1016/j.hlpt.2017.02.002

Policy makers in the healthcare sector can take preventive measures such as public health interventions [41,42]. One possibility is to pass legislation requiring restaurants and school canteens to provide healthier alternatives. Gymnasium teachers can be encouraged to promote athletics as a key lifestyle in children and as a way of avoiding obesity in adulthood. Policy makers should consider defining national and international weight-related goals (in the short and long term). Insurance companies have already begun to charge more for life insurance for people with excess weight. Governments can also consider taxing food products that contribute to obesity such as the 'sugary drink' tax now under consideration. Finally, as has been done for nicotine products, governments can consider printing a warning on high-caloric foods that consumption can lead to obesity and are dangerous to health.

Limitations and future research

Some of the limitations affecting other studies [21,24] also apply here. First, we did not consider the costs associated with the successful treatment of obesity. Second, we did not include all possible obesity-related medical conditions. Although it is virtually impossible to isolate the exact cause of some common ailments that are typically associated with obesity, we took steps to separate such costs. Third, in this study we were unable to include the indirect costs of obesity problems. Fourth, regular time series analyses used for forecasting including ARIMA have the disadvantages of noise in the computations and a linearity assumption. We tried to resolve most of these drawbacks by using a complete process with the Box-Jenkins four-step process described in the methodology section. Finally, the analysis was not on the national level. These data are difficult to collect and many observations had to be imputed for data for certain years in the 1960s and 1970s. Currently there are now data at the state level but none prior to the 1990s.

With respect to future research, it would be interesting to incorporate some of the most frequent comorbidities associated with obesity and attempt to include these to forecast future healthcare expenditures. For example, including heart disease and diabetes in the model might make a better predictor of healthcare expenses. Another future research avenue would be to incorporate the costs associated with interventions. While most resources are currently devoted to projecting healthcare costs associated with obesity, it would be worthwhile to explore what benefits (or losses) interventions might have.

Funding

None.

Competing interests

None declared.

Ethical approval

Not required.

Acknowledgements

There are no declarations to make.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.hlpt.2017.02.002.

References

[1] Ogden CL, Carroll MD. Prevalence of overweight, obesity, and extreme obesity among adults: United States, trends 1960-1962 through 2007-2008. Centers for Disease Control and Prevention; 2010.

[2] Ogden CL, et al. Prevalence of overweight and obesity in the United States, 1999-2004. JAMA 2006;**295**(13):1549-55.

[3] WHO. Obesity: preventing and managing the global epidemic. World Health Organization, Geneva, Switzerland; 2000.

[4] Allen L. Rising obesity rates increase medicare expenditure. PharmacoEconomics Outcomes News, **728**; 2015.[28-23].

[5] Sturm R. The effects of obesity, smoking, and drinking on medical problems and costs. Health Aff 2002;**21**(2):245-53.

[6] Wang Y, et al. Will all Americans become overweight or obese? Estimating the progression and cost of the US obesity epidemic Obesity 2008;**16**(10):2323-30.

[7] McCormick B, Stone I. Economic costs of obesity and the case for government intervention. Obes Rev 2007;**8**(s1):161-4.

[8] Finkelstein EA, Fiebelkorn IC, Wang G. National medical spending attributable to overweight and obesity: how much, and who's paying? Health Aff 2003:W3.

[9] Kopelman PG. Obesity as a medical problem. Nature 2000;**404**(6778):635-43.

[10] Finkelstein EA, et al. Annual medical spending attributable to obesity: payer-and service-specific estimates. Health Aff 2009;**28**(5):w822-31.

[11] Thompson D, et al. Body mass index and future healthcare costs: a retrospective cohort study. Obes Res 2001;**9**(3):210-8.

[12] Frenk DJ. Economic costs: paying the price for those extra pounds. Havard School of Public Health, Boston, MA, USA; 2012.

[13] Stubbs A, Uzuner Ö. Annotating risk factors for heart disease in clinical narratives for diabetic patients. J Biomed Inform 2015;**58**:S78-91.

[14] Hajer GR, van Haeften TW, Visseren FL. Adipose tissue dysfunction in obesity, diabetes, and vascular diseases. Eur Heart J 2008;**29**(24):2959-71.

[15] Profenno LA, Porsteinsson AP, Faraone SV. Meta-analysis of Alzheimer's disease risk with obesity, diabetes, and related disorders. Biol Psychiatry 2010;**67**(6):505-12.

[16] Hata J, Kiyohara Y. Epidemiology of stroke and coronary artery disease in Asia. Circ J 2013;**77**(8):1923-32.

[17] Fox CS, et al. Trends in the Incidence of Type 2 diabetes mellitus from the 1970s to the 1990s The Framingham Heart Study. Circulation 2006;**113**(25):2914-8.

[18] Centers_for_Disease_Control_Prevention. National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014. Atlanta, GA: US Department of Health and Human Services; 2014.

[19] Gregg EW, et al. Changes in diabetes-related complications in the United States, 1990-2010. New Engl J Med 2014;**370**(16):1514-23.

[20] Trogdon JG, et al. Indirect costs of obesity: a review of the current literature. Obes Rev 2008;**9**(5):489-500.
[21] Allison DB, Zannolli R, Narayan KM. The direct health care costs of obesity in the United States. Am J Public Health 1999;**89**(8):1194–9.
[22] Thorpe KE, et al. The impact of obesity on rising medical spending. Health Aff 2004;**23**. [283-283].
[23] Yokum S, Ng J, Stice E. Relation of regional gray and white matter volumes to current BMI and future increases in BMI: a prospective MRI study. Int J Obes 2012;**36**(5):656-64.
[24] Quesenberry CP, Caan B, Jacobson A. Obesity, health services use, and health care costs among members of a health maintenance organization. Arch Intern Med 1998;**158** (5):466–72.
[25] English P. The economic costs of overweight, obesity and physical inactivity among california adults–2006. California: Center for Public Health Advocacy; 2009.
[26] Roberts K, et al. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform 2015;**58**:S111–9.
[27] Fryar CD, Ogden CL, Carroll MD. Prevalence of overweight, obesity, and extreme obesity among adults: united States, trends 1960–1962 through 2007–2008. Natl Cent Health Stat 2014;**6**(1):1–6.
[28] CMS.GOV, *Historical National Health Expenditure Data*. Centers for Medicare and Medicaid Services, 2016.
[29] Stock J, Watson M. Introduction to time series regression and fore-casting.2nd ed. Boston: Pear- son/Addison Wesley; 525-565. [Introduction to Econometrics].
[30] Gujarati DN, Porter T. Basic econometrics. *5th ed*. Boston: McGraw-Hill Education; 773-84.
[31] Trottier H, Philippe P, Roy R. Stochastic modeling of empirical time series of childhood infectious diseases data before and after mass vaccination. Emerg Themes Epidemiol 2006;**3**(1):1.
[32] Darrat AF, Zhong M. On testing the random-walk hypothesis: a model-comparison approach. Financ Rev 2000;**35**(3):105–24.
[33] Honaker J, King G. What to do about missing values in time-series cross-section data. Am J Polit Sci 2010;**54**(2):561–81.
[34] Shonibare T, et al. Medical problems in obesity. Endocrinology and diabetes.Springer; 191–9.
[35] Kaufman L, Karpati A. Understanding the sociocultural roots of childhood obesity: food practices among Latino families of Bushwick, Brooklyn. Social Sci Med 2007;**64**(11):2177–88.
[36] Fakhouri TH. Prevalence of obesity among older adults in the United States, 2007–2010. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics; 2012.
[37] Hammond RA, Levine R. The economic impact of obesity in the United States. Diabetes Metab Syndr Obes 2010;**3**(1):285–95.
[38] Bray GA. Medical consequences of obesity. J Clin Endocrinol Metab 2004;**89**(6):2583–9.
[39] Schilling R., *Obesity Shortens Life*; 2015.
[40] Nestle M, Jacobson MF. Halting the obesity epidemic: a public health policy approach. Public Health Rep 2000;**115**(1):12.
[41] Cuijpers P, De Graaf I, Bohlmeijer E. Adapting and disseminating effective public health interventions in another country: towards a systematic approach. Eur J Public Health 2005;**15** (2):166–9.
[42] Osrin D, et al. Ethical challenges in cluster randomized controlled trials: experiences from public health interventions in Africa and Asia. Bull World Health Organ 2009;**87**(10):772–9.

Please cite this article as: Revels S, et al. Predicting obesity rate and obesity-related healthcare costs using data analytics. Health Policy and Technology (2017), http://dx.doi.org/10.1016/j.hlpt.2017.02.002