SCIENCE FORUM (JOURNAL OF PURE AND APPLIED SCIENCES) 22 (2022) 444 – 455

http://dx.doi.org/10.5455/sf.XXXXX

Science Forum (Journal of Pure and Applied Sciences)

journal homepage: www.atbuscienceforum.com

Modeling effects of diabetes risk factors using modified logistics regression

Alhaji Abdullahi Gwani[1,2,3*], Ali B. Adamu[1], Abbas Umar Farouk[1,4], Sek Siok Kun[3]

[1] *Department of Mathematical, Abubakar Tafawa Balewa Bauchi, Bauchi University, Bauchi, Nigeria.*

[2] *Department of Mathematical, Sciences Bauchi State University Gadau, Bauchi, Nigeria.*

[3] *Department of Statistics, School of Mathematical Sciences, Universiti Sains Malaysia, Penang, Malaysia.*

[4] *Department of Mathematics and Statistics, Federal Polytechnic Bauchi, Bauchi, Nigeria.*

ABSTRACT

This paper model the effects diabetes risk factors using modified logistic regression model (MLRM) since ordinary least square model (OLSM) <AQ> is deficient, especially when dealing with irreversible variables. A real data was sampled from medical out-patient department (MOPD) unit, Federal Medical Centre Azare using systematic sampling technique. A total sampled of 407 patients' data were used to test and compare the MLRM and the other competitive model. Gnu Regression, Econometrics and Time-series Library (GRETL) and Statistical Package for Social & Sciences (SPSS) programming bundles were utilized during the information analysis. The result reveals that MLRM is more efficient than LRM. Chi-square was employed to test for dependency between diabetes status (DBStatus) and the other suspected variables; it was shown that there is association between DBStatus and virtually all the variables incorporated in the MLRM model. The MLRM model can be used to predict the chance of an individual to be diabetic given data as input variables. Based on this model, it was observed that the reduced MLRM Model M_2 is the suitable model than M_1 that is the Model with all the variables. This paper also revealed that blood pressure (BP) systolic, weight, height, BP diastolic, age, and sugar level are statistically significant, while the other remaining variables are not statistically significant. We recommend for further investigation using the machine learning approach.

ARTICLE INFO

Article history:
Received 10 December 2021
Received in revised form
11 December 2021
Accepted 11 December 2021
Published 27 April 2022
Available online 27 April 2022

KEYWORDS

Model
Diabetes
Status
Patient
Data
Blood Pressure (BP)
MLE
OLS
Probablity

1. Introduction

The technique of modifying and building models is well-known in the literature. Modeling and evaluating lifespan data is critical in many practical fields, including medical, engineering, and finance, biostatistics (Shayib 2018) to name a few. The diabetes risk factors were modelled using ordinary least square (OLS). The quality of statistical analysis techniques is greatly influenced by the probability model or distributions assumed. Based on this, OLS has some limitations, in that it returns real numbers, which can be difficult to interpret when a variable becomes negative, and its redundancy value can only be determined by the value of the R2 coefficient of determination (McArdle and Anderson 2001), which can be biased in the absence of certain assumptions, such as serial correlation (Hsieh and Pan 2018) and heteroscedasticity. Because the satisfying model assumptions enhances precision and statistical power (Frank and Harrell 2018); However, numerous fundamental issues must be resolved before OLS may be used with real-world data.

* **Corresponding author** Alhaji Abdullahi Gwani ✉ aagwani1@gmail.com ✉ Department of Mathematical, Abubakar Tafawa Balewa Bauchi, Bauchi University, Bauchi, Nigeria

The main purpose of statistical learning, according to Berk (2017), is to "analyze the link between inputs and outcomes." While the best procedure is uncertain currently, there are various alternatives. Further information diabetes, according to Alwan (2010) and Zaoui (2007), is a severe danger to global development. Its size is growing at an alarming rate. Among the most recent publications is diabetes which can cause major issues with the eyes, kidneys, nerves, and even stroke, but heart disease (Polat and Salih, 2007; Stratton et al., 2019). Long term follow-up, according to Davy et al. (2016), is required to properly record the development of more distal consequences such as heart disease, medical costs, and early mortality. Furthermore, Connor et al. (2015) predicted that by 2030, at least 400 million people will have diabetes worldwide, with many of those affected being relatively young and living in low- or middle-income countries, including Nigeria. They likewise expressed that the diabetes is connected to an essentially expanded danger of stroke, is currently the main source of visual impairment, and is a main source of end-stage kidney infection. Likewise, this review will be founded on the suggestions of certain creators, for example, Davy et al. (2016), Andrew et al. (2019), just as crafted by Connor et al. (2015), who announced that by 2030, there will be something like 400 million individuals with type 2 diabetes around the world, with a significant number of those impacted being moderately youthful and living in low or center pay nations. This is congruent with the findings of Cho et al. (2018). According to Cho et al. (2018), there were 451 million diabetes globally. These numbers are expected to rise to 693 million by the year 2045. According to Berk (2017), the "main objective of statistical learning is to examine the link between inputs and outputs." While the method for doing so is uncertain, there are several feasible options. Furthermore, Smokovski et al. (2018) advocated for more research into the risk factors related with growing diabetes prevalence. The goal of this study would be to reduce the diabetes causalities linked with the projected hazard variables. That this review labels as "Demonstrating the impact of Diabetes Risk Factors Using Multiple Logistics Regression: A Case Study of Federal Medical Center Azare" with the desire to be diminished or limited the causalities of τ diabetes and thought of a reasonable model among the serious models.

2. Material and Methods

The materials used for the purpose of this work were the patient folders, patient record in the folders [age, blood pressure (BP) systolic, BP diastolic, height, weight, location, sugar level, and employment status], as well as computer programming bundles Gnu Regression, Econometrics and Time-series Library (GRETL) and Statistical Package for Social & Sciences (SPSS). These materials were used in double sample to obtain data from out-patient department record unit at FMC Azare, Katagum local Government, Bauchi State, Nigeria. The methods used for this paper are as follows:

2.1. The modified logistic regression model (MLRM)

Multiple logistic regressions are statistical techniques use in dealing with dichotomous responds in relations to two or more variables (Hosmer et al., 2000). Let us consider the general case of logistic regression model in conjunction with various explanatory variables $(X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_k)$ for binary response Y. We use $\boldsymbol{\tau(x)}$ to correspond to the probability (Borovkov 2013) that $Y = 1$ $[P(Y = 1)]$ for success and $1-\boldsymbol{\tau(x)}$, to signify the probability that $Y = 0$ $[P(Y = 0)]$ (Badi 2017; Bonamente 2017; Forsyth 2018).

These probabilities are written in the following form (Berkson 1994):

$$\boldsymbol{\tau(x)} = P(Y = 1/X_1, X_2 \ldots\ldots\ldots X_k 1 - \boldsymbol{\tau(x)} = P(Y = 0/X_1, X_2 \ldots\ldots\ldots X_k$$

The log odds model is:

$$\textbf{logit}\ (\boldsymbol{\tau(x)}) = \frac{P(Y=1/X_1, X_2 \ldots\ldots\ldots X_k)}{P(Y=0/X_1, X_2 \ldots\ldots\ldots X_k)}$$

$$in\left(\frac{\boldsymbol{\tau(x)}}{1-\boldsymbol{\tau(x)}}\right) = \beta_0 + \beta_1 X_1 + X_2\beta_2 \ldots\ldots \beta_k X_k + \varepsilon \quad (1)$$

$$in\left(\frac{\boldsymbol{\tau(x)}}{1-\boldsymbol{\tau(x)}}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon \quad (2)$$

From Equations (1) and (2) we can deduce Equation (3) as follows:

$$\boldsymbol{\tau(x)} = P(Y = 1/X_1 \ldots\ldots\ldots X_k) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon}} \quad (3)$$

The constraint βj describes to the impact of Xj on the log odds that $Y = 1$, regulating the other predictor variables. For example, Exp (β_j) is the multiplicative impact on the odds of a one-unit increase in Xj at fixed levels of other variables in the model. Equation (3) is the MLRM. Demidenko (2012) reported that odds

ratio, also known as the exponentiated logistic regression coefficient, is a widely used measure of association. Typically, in logistic regression, the confidence interval (CI) for the odds ratio (OR) is calculated.

2.2. MLE estimation

The primary point of numerous calculated relapses is to assess the $K + 1$ unknown parameters $\beta(\beta\_0,\beta\_1,\beta\_2....\beta\_k)$. This should be possible with the most extreme probability assessment which involves seeing as the arrangement of boundaries for which the likelihood of the noticed information is the most prominent. The most extreme probability condition is gotten from the binomial conveyance of the reaction variable. For a bunch of perceptions in the data $(x_i;y_i)$, the contribution to the likelihood function is $\boldsymbol{\tau}(x_i)$ where $y_i = 1$, and = 0 (Borovkov 2013). The following equation results for the contribution [call it (x_i)] to the likelihood function for the observations $(x_i;y_i)$:

$$\boldsymbol{\Phi}(x_i) = \boldsymbol{\tau}(\boldsymbol{x})^{y_i} 1 - \boldsymbol{\tau}(\boldsymbol{x})^{1-y_i} \tag{4}$$

The Equation (4) accounts for only one set of observations. Wakefield (2013) and Pathirana and Varathan (2018) reported that the observations were assumed to be independent of each other so that we can multiply their likelihood contributions to obtain the complete likelihood function (Shiraishi et al., 2018).

$$l(\beta) = \prod_{i=1}^{k} \boldsymbol{f}\ (x_i) \tag{5}$$

By subtitling Equation (4) in Equation (5), we obtain

$$l(\beta) = \prod_{i=1}^{k} \boldsymbol{\tau}(\boldsymbol{x})^{y_i}[1 - \boldsymbol{\tau}(\boldsymbol{x})]^{1-y_i}$$

$$l(\beta) = \boldsymbol{\tau}(\boldsymbol{x})^{\sum_{i=1}^{k} y_i}[1 - \boldsymbol{\tau}(\boldsymbol{x})]^{k-\sum_{i=1}^{k} y_i}$$

$$l(\beta) = \boldsymbol{\tau}(\boldsymbol{x})^{\sum_{i=1}^{k} y_i}[1 - \boldsymbol{\tau}(\boldsymbol{x})]^{k}[1 - \boldsymbol{\tau}(\boldsymbol{x})]^{\sum_{i=1}^{k} y_i}$$

This takes the lead us to write down Equation (5) as follows:

$$l(\beta) = \left(\frac{\boldsymbol{\tau}(\boldsymbol{x})}{1-\boldsymbol{\tau}(\boldsymbol{x})}\right)^{\sum_{i=1}^{k} y_i} [1 - \boldsymbol{\tau}(\boldsymbol{x})]^{k} \tag{6}$$

From Equations (2) and (3), (Arulampalam et al. 2004; Cramer 2002)

$$\left(\frac{\boldsymbol{\tau}(\boldsymbol{x})}{1 - \boldsymbol{\tau}(\boldsymbol{x})}\right) = e^{\beta_0+\sum_{j=1}^{k} \beta_j X_j+\varepsilon}$$

$$\boldsymbol{\tau}(\boldsymbol{x}) = \frac{e^{\beta_0+\sum_{j=1}^{k} \beta_j X_j+\varepsilon}}{1 + e^{\beta_0+\sum_{j=1}^{k} \beta_j X_j+\varepsilon}}$$

We can write

$$l(\beta) = \left(e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}\right)^{\sum_{i=1}^{k} y_i} \left(1-\frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}}\right)^{k}$$

$$l(\beta) = \left(e^{\beta \sum_{i=1}^{k} y_{i_0}+\sum_{i=1}^{k} y_i \sum_{j=1}^{k} \beta_j X_j+\varepsilon}\right) \left(1 + e^{\beta_0+\sum_{j=1}^{k} \beta_j X_j+\varepsilon}\right)^{-k} \tag{7}$$

β is the collection of parameters $(\beta_0,\beta_1,\beta_2....\beta_k)$ and $l(\beta)$ is the likelihood function of β. The maximum likelihood approximates $(\beta_0,\beta_1,\beta_2....\beta_k)$ can be acquired by calculating which β maximizes. But we can make simpler in the mathematics, by taking the logarithm of Equation (3). As shown in Equation (4),

$$\Phi(x_i) = \tau(x)^{y_i}\, 1-\tau(x)^{1-y_i}$$

Let $L(\beta)$ denotes the log likelihood expression.

$$L(\beta) = inl(\beta) = in\left(e^{\beta\sum_{i=1}^{k} y_{i0}+\sum_{i=1}^{k} y_i \sum_{j=1}^{k} \beta_j X_j+\varepsilon}\right)\left(1+e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}\right)^{-k}$$

$$L(\beta) = inl(\beta) = \left(\beta\sum_{i=1}^{k} y_{i0} + \sum_{i=1}^{k} y_i \sum_{j=1}^{k} \beta_j X_j + \varepsilon\right) - Kin\left(1+e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}\right) \tag{8}$$

The critical points of a task take place when the first derivative of the equation and equate it to 0. But the second derivative evaluated at that point is less than 0, then the critical point is said to be maximum if not, its minima. Thus, finding the maximum likelihood estimates necessitates computation the first derivative of the log likelihood function $L(\beta)$.

Therefore, by taking the derivative of Equation (7) with respect to β_0 we got

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{k} y_i - K\left(\frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}}\right) \tag{9}$$

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{k} y_i - K(\boldsymbol{\tau}(\boldsymbol{x})) \tag{10}$$

The maximum likelihood estimates for β_j can be obtained by setting for each one of the equations,

respectively, Equations (9) and (10) equal to zero and solve for each β_j, this would yield the MLEs of the j parameters.

2.3. Sample, sample size and preanalysis

In this section, we provide the work systematically sampled. Let the sample size be n (Barreiro and Albanzo 2001) n = 407 patients' folders from medical out-patient department unit that can be used to model the probability factors of emerging diabetes, since large sample suited logistic model as proposed by Nad and Kanuch (2018) The risk factors were identified from the patient folders which include the following: BP diastolic, BP systolic (Emdin et al., 2015; Yu et al., 2018), weight (Ike and Rodney 2015), height age (Kusurkar et al., 2010), gender (Wang et al., 2018), and employment status and sugar level (Whiting et al., 2011). The outline of the data obtained is in Table 1.

Table 1 presents an outline of the example information, including the classification of each component, the quantity of diabetic patients in every classification, the quantity of non-diabetic patients in every classification, and their rates. In general, 67 patients were diabetic, representing 16.46% of the all-out examined, while the excess 340 patients were non-diabetic, representing 83.54% of the all-out tested. This data could be utilized to work on the model. The Chi-square trial of freedom was utilized to assess the relationship among diabetes and every one of the likely factors. The connection among diabetes and the variables was found to be BP systolic, BP diastolic, age, weight, stature, and glucose level. Sex, occupation, and geology, then again, had a measurably immaterial connection with diabetes. The decision of strategic relapse model and not set in stone after testing, considering the non-direct connection among DBStatus and factors. The choice to utilize a calculated relapse model and change it to meet this undertaking was unavoidable after a few examinations. The result of the test shows that BP systolic (p-value = 0.0008 < 0.05), sugar level (p-value = 0.0260 < 0.05), and height (p-value = 0.0041 < 0.05) are huge indicators. The variables that are not measurably critical are sex (p-value = 0.7270 > 0.05), BP diastolic (p-value = 06072. > 0.05), age (p-value = 0.7477 > 0.05), area (p-value = 0.7043 > 0.05), weight (p-value = 0.0526 > 0.05), and business status

Table 1. Summary of the sampled patients' data (n = 407).

Factor	Category	ND	*D*	TOTAL	ND%	*D*%
EmployStatus	Employed	175	33	208	84.13	15.87
	Non-employed	165	34	199	82.91	17.09
SugarLevel	High (5 and above)	6	61	67	8.96	91.04
	Normal (<5)	331	9	340	97.35	2.65
BPDiastolic	High (>80)	335	62	397	84.38	15.62
	Normal (≤80)	5	5	10	50.00	50.00
Height	<1.65 M	315	58	373	84.45	15.55
	1.65 M and above	25	9	34	73.53	26.47
Location	Within state	318	61	379	83.91	16.09
	Outside state	22	6	28	78.57	21.43
Weight	31–79	335	62	397	84.38	15.62
	80 and above	5	5	10	50.00	50.00
Gender	Male	187	33	220	85.00	15.00
	Female	153	34	187	81.82	18.18
Age	12–39	77	19	96	80.21	19.79
	40–59	168	21	189	88.89	11.11
	60 and above	95	27	122	77.87	22.13
BPSystolic	High	6	5	11	54.55	45.45
	Normal	69	15	84	82.14	17.86
	Low	265	47	312	84.94	15.06
Diabetes status (DBStatus)	Diabetic	0	67	67	0.00	100.00
	Non-diabetic	340	0	340	100.00	0.00

(0.0070 > 0.05). As per the review's indication of coefficients examination, being utilized, getting more fit, and having a high BP diastolic expands the danger of diabetes, yet being jobless, shedding pounds, and having a high BP diastolic lower the danger. The chances proportion information show that patients with high systolic pulse, patients with high sugar levels, more established age, and patients who live inside the state are all at high danger for diabetes. A model with just critical indicator factors was fitted notwithstanding the model with every free factor. At the point, when we looked at the two models, we found that the diminished and altered model, which has a normal region under the receiver operating characteristic (ROC) bend of 0.66 and a by and large illustrative strength of 86.75%, fits the information better than the model with all indicator factors, which has a normal region under the ROC bend of 0.58 and a generally logical strength of 84.08%

2.4. Checking for non-linearity between factors in the model

The calculated relapse expected non-direct connection between the wards and the free movers regularly conveyed mistake terms are not accepted. Table 2 is the result of SPSS 16.0 on the linearity between the reliant variable (DBStatus) and the free factors (BP diastolic, BP systolic, weight, age, height, sex, employment status, and location).

From the Table 2, it can be examined that most of the correlation figures assumed very small values except for one variable which is sugar level and DB status that attained 0.741 but beside these all-other variables has smaller or negligible correlations. This is evidence of non-linearity between independent variables and the dependent variable. Figure 1 below shows line graph of these relationships.

3. Results and Discussion

In this section, we presented the results obtained together with some discussions.

3.1. The fitting model with predictors' coefficients

The number of variables used were nine which are: BP systolic, employment status, gender, age, sugar level, BP diastolic, location, height, and weight. The dependent variable quantity continues the identical: the DBStatus. From the general model, Equation (11) as follows :

$$in\left(\frac{\tau(x)}{1-\tau(x)}\right) = \beta_0 + \sum_{j=1}^{k}\beta_j X_j + \varepsilon \tag{11}$$

$$\boldsymbol{\tau(x)} = P(Y = 1/X_1 \ldots\ldots\ldots X_k) = \frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}} \tag{12}$$

The model to fit is as follows: Let Y_i be a binary outcome of diabetes (diabetic =1 or non-diabetic = 0), for individual, i $Y_i \sim$ Bernulli ($\boldsymbol{\tau}_i$). Let X_j be a vector, (X_1,X_2...X_9) = (EmployStatus, SugarLevel, BPSystolic, Height, Location, Weight, Gender, Age, and BPDiastolic).

$$\boldsymbol{\tau(x)} = P\left(Y = 1/X_1 \ldots\ldots\ldots X_k\right)$$

$$= \frac{e^{\beta_0+\beta_1X_1+X_2\beta_2+X_3\beta_3+X_4\beta_4+X_5\beta_5+X_6\beta_6+X_7\beta_7+X_8\beta_8+X_9\beta_9+\varepsilon}}{1+e^{\beta_0+\beta_1X_1+X_2\beta_2+X_3\beta_3+X_4\beta_4+X_5\beta_5+X_6\beta_6+X_7\beta_7+X_8\beta_8+X_9\beta_9+\varepsilon}} \tag{13}$$

GRETL and SPSS programming bundles were utilized to deal with the information. The greatest probability strategy is utilized to gauge the coefficients

Table 2 . Correlations between the variables in the model.

No.	Control variables	1	2	3	4	5	6	7	8	9	10
1	Age	1.000	0.030	0.207	0.088	0.390	0.168	0.096	0.067	0.048	0.110
2	Gender		1.000	0.048	−0.411	0.050	0.036	0.050	0.044	−0.528	0.036
3	Weight			1.000	0.231	0.120	0.010	0.046	0.045	0.072	0.007
4	Height				1.000	−0.046	−0.037	−0.007	0.030	0.330	0.009
5	BPDiastolic					1.000	0.614	0.044	0.051	−0.060	0.004
6	BPSystolic					*	1.000	0.066	0.097	−0.046	−0.037
7	SugarLevel							1.000	0.741	−0.020	0.023
8	DBStatus								1.000	−0.018	−0.036
9	EmployStatus		*							1.000	−0.052
10	Location										1.000

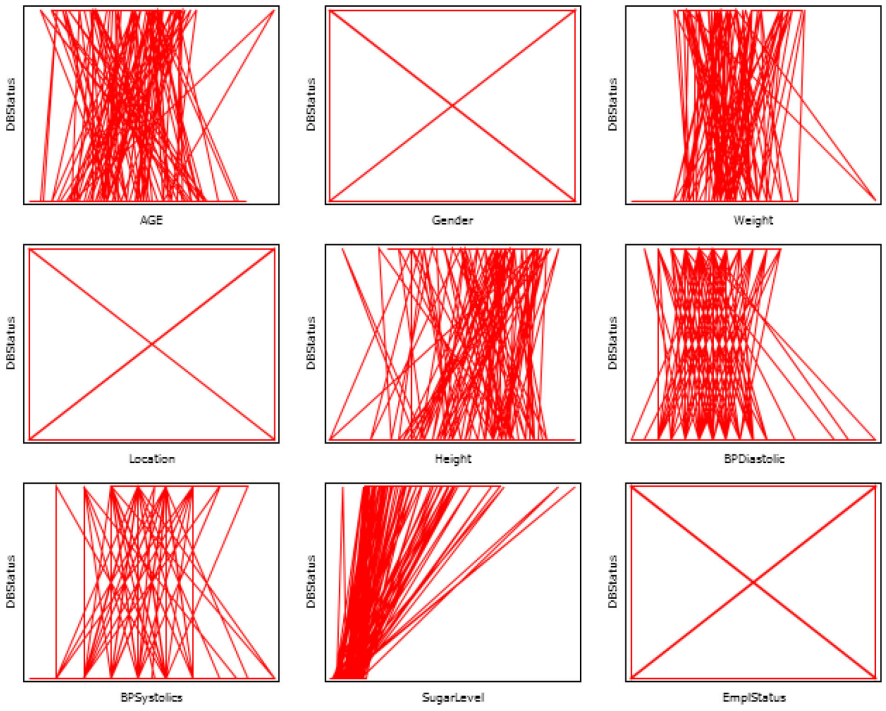

Figure 1. Line graph between DB status and other risk factors.

Table 3. Dependent variable: DBStatus. QML standard errors GRETL Model 1 (M_1): logistic, using observations 1–407.

No.	Risk factors	Coefficient (β)	Standard error	z	p-value
1	Const	−51.7402	17.0781	−3.030	0.0024
2	AGE	0.00676319	0.0210226	0.3217	0.7477
3	Gender	0.275988	0.790575	0.3491	0.7270
4	Weight	−0.0711377	0.0367099	−1.938	0.0526
5	Location	−0.461025	1.21482	−0.3795	0.7043
6	Height	7.39090	2.57223	2.873	0.0041
7	BPDiastolic	−0.00734258	0.0142846	−0.5140	0.6072
8	BPSystolics	0.0924917	0.0274945	3.364	0.0008
9	SugarLevel	7.06717	3.17550	2.226	0.0260
10	EmplStatus	0.00316502	0.851346	0.003718	0.9970

and its standard blunder. Table 3 shows yield and the assessed coefficients β_j, their SE, and Wald test for the full model and the result of GRETL bundle investigation.

The values of the output estimated coefficients, for the two packages are closed only four decimal places were considered in the output of GRETL. We decided to use GRETL outputs since it considered four places after decimal unlike SPSS with three places.

Therefore, applying the assessments of the parameters in Table 3, we can formulate this model the full model as follows:

$$\tau(x)=\frac{e^{-51.646-0.003X_1+7.0672X_2-0.0073X_3+7.3909X_4+0.4610X_5-0.0711X_6-0.2760X_7+0.0068X_8+0.0925X_9+\varepsilon}}{1+e^{-51.646-0.003X_1+7.0672X_2-0.0073X_3+7.3909X_4+0.4610X_5-0.0711X_6-0.2760X_7+0.0068X_8+0.0925X_9+\varepsilon}} \quad (14)$$

3.2. Testing for the significance of the individual parameters in the model

According to Shayib (2018) to test the hypothesis,

$$H_0 : \beta_j = 0$$

Versus

$$H_1 : \beta_j \neq 0$$

Table 4. Odds ratios and 95% CIs for covariates.

			Asymptotic 95% CI	
Variable	**Wald**	**Exp(*B*)**	**Lower bound**	**Upper bound**
Weight	0.000	0.931	0.442	0.597
Age	17.709	1.007	0.484	0.620
BPSystolic	3.901	1.097	0.507	0.652
BPDiastolic	1.472	0.993	0.485	0.631
EmployStatus	0.085	0.997	0.413	0.565
SugarLevel	1.911	1.173E3	0.964	1.011
Height	0.049	1.621E3	0.458	0.623
Location	0.046	1.586	0.411	0.564
Gender	0.073	0.795	0.453	0.604

From Wald and Sig. segments of Table 4 over, the line position of individual element is given by: the table uncovers that the huge indicators are: BP Systolic (p-esteem = 0.0008 < 0.05) and sugar level (p-esteem = 0.0260 < 0.05) height (p = 0.0041 < 0.05). On the opposite side, the indicators which are not measurably critical are: gender (p-esteem = 0.7270 > 0.05), BP diastolic (p-esteem = 06072 > 0.05), age (p-esteem = 0.7477 > 0.05), location (p = 0.7043 > 0.05), weight (p = 0.0526 > 0.05), and employment status (0.0070 > 0.05).

3.3. The odds ratio results

From Table 4, Exp(B) column holds the exponential of parameter estimations in the model. These figures represent odds ratios for the corresponding predictor variables in the same model. Furthermore, same Table 4 indicates the 95% Wald confidence interval (CI) for the odds ratio of each the models' predictors (Ratsimalahelo 2017).

From Table 4 it is evident that patients with high blood pressure systolic, patients who have high sugar level, older age, and located within the state have a highly susceptible for diabetes occurrence based on the decision suggested by Cleophas and Zwinderman (2011). Mohammed et al. (2015) concluded on the danger of co-existence of hypertension with diabetes increases the risk of cognitive impairment, while contradicted (Kpozehouen et al., 2015) result who concluded that gender and age are not determining the risk factors of diabetes.

3.4. Signs of coefficients analysis

The sign of the coefficients of the significant factors based on the estimates of logistic function in Table 4 explains the variables applied, as given in Table 5.

3.5. Likelihood ration test

To assess the full fitted model, likelihood ratio was use using the SPSS output below: Table 6 presents the Likelihood Ratio test. The −2log likelihood for the null model achieves by fitting the constant only model is 47.207; and the −2 log likelihood for the overall model with 41.751. Therefore, the value of the likelihood ratio test is as follows:

$G^2 = 47.751 - 41.207 = 5.863$

For large sample, n = 407 a very small value of G is a signal that the model fit the data good.

The null hypothesis is as follows:w

$$H_0 : \beta_j = 0 \; for \; j = 1,2,3.....n$$

Versus

$$H_1 : \beta_j \neq 0 \; foratleastonej$$

Table 6 indicates small p-values = 0.000, which is less than the conventional value 0.05. This would lead us to reject the null hypothesis meaning that at least one and conceivably all beta's coefficient is not the same from zero.

3.6. ROC curve for the model with all the variable

Contrasting the two models (model with all indicators and the diminished model), normal region under the ROC bend has turned into an especially significant measure for assessing models' presentation since it is the normal affectability over every single imaginable particularity. The bigger the region, the better the model performs. Table 7 shows the region for every factor.

The area under the curve for each factor is shown in Figure 2.

Figures 2 and 3 can be summarized in the ROC figure below with different colours as indicate in the titled source of the curve.

The typical twist locale for the full model is 0.58. By use of the ROC twist in Figure 4 for the portrayal precision, it is seen that the locale under the ROC twist, which goes from 0 to 1 gives the extent of the model's ability to isolate between those subjects who

Table 5. Variables' sign analysis

No.	Factors	*B*	Sign	Explanation
1	EmployStatus (X_1)	-0.003	Positive	Employed decreases the probability of having diabetes
2	SugarLevel (X_2)	7.067	Positive	High sugar level increases the probability of having diabetes
3	BPSystolic (X_3)	0.092	Positive	High BP systolic increases the probability of having diabetes
4	Height (X_4)	7.391	Positive	High height increases the probability of having diabetes
5	Location (X_5)	0.461	Positive	Resident of the state increases the probability of having diabetes
6	Weight (X_6)	−0.071	Negative	Low weight decreases the probability of having diabetes
7	Gender (X_7)	−0.276	Negative	Male decreases the probability of having diabetes
8	Age (X_8)	0.007	Positive	Higher age increases the probability of having diabetes
9	BPDiastolic (X_9)	−0.007	Negative	Low BP Diastolic increases the probability of having diabetes
10	EmployStatus (X_1)	−0.003	Positive	Employed decreases the probability of having diabetes

Table 6. Model selection.

Step	–2 Log likelihood	Cox & Snell *R* square	Nagelkerke *R* square	Sig
1	47.207[a]	0.541	0.915	0.000
2	41.751[b]	0.547	0.925	0.000

Table 7. ROC area under curve.

		Asymptotic 95% CI	
Variable	Area	Lower bound	Upper bound
Weight	0.519	0.442	0.597
Age	0.552	0.484	0.620
BPSystolic	0.579	0.507	0.652
BPDiastolic	0.558	0.485	0.631
EmployStatus	0.489	0.413	0.565
SugarLevel	0.987	0.964	1.011
Height	0.540	0.458	0.623
Location	0.488	0.411	0.564
Gender	0.529	0.453	0.604

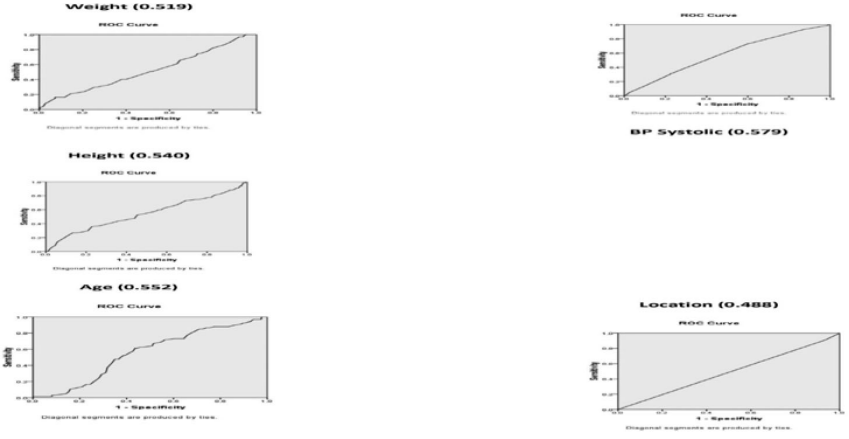

Figure 2. ROC area curve for each variable in the model.

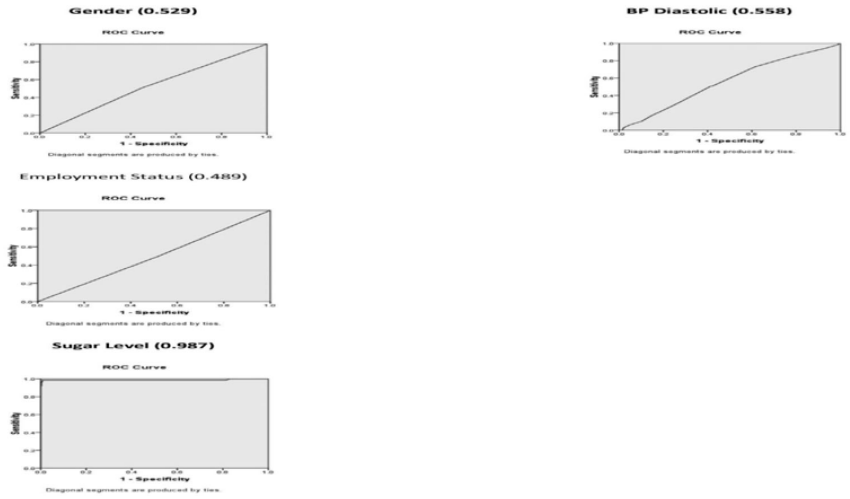

Figure 3. ROC area curve for each variable in the model.

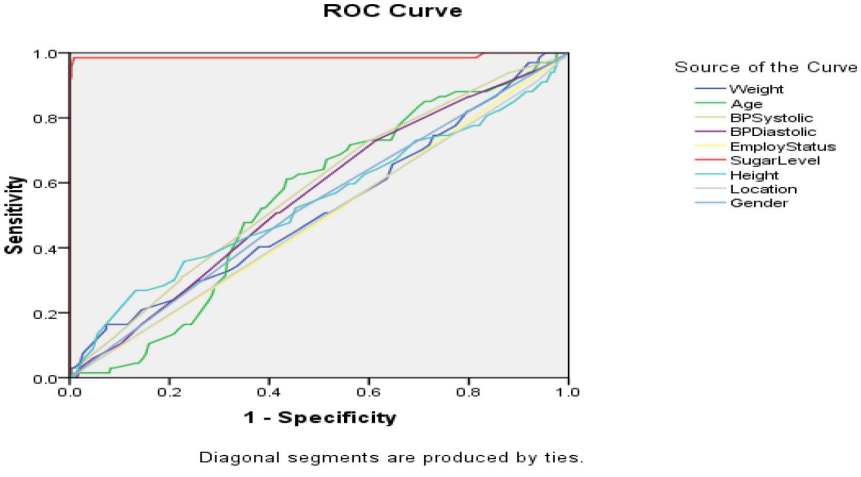

Figure 4. ROC curve for the full model.

experience the response of interest versus individuals who do not. The locale under the ROC twist for the full model is 0.925 which may be considered as reasonable partition.

3.7. ROC curve for reduced model

This is the lessened model with basic limits assessment using GRETL. The basic limits in the full model are sugar level, BP systolic, height, and weight; the result of the assessment with the possibility of only four tremendous limits is given in Table 8.

Table 8 reveals only two significant factors which are sugar level with p = 0.0001 and the BP systolic with p = 0.0199. The ROC curves the model with the significant parameters is shown in Figure 5.

ROC curve has turned into a particularly significant measure for assessing models' performance presentation since it is the normal affectability over every conceivable explicitness. The bigger the area under the region, the better the model performs. Model M_2 showed an average ROC area under curve of 0.58 comparing it with full model (model with all predictors), Model M_1 showed an average ROC area under curve of 0.66. This indicate that M_2 performed better than M_1.

3.8. Modified logistic regression full model (M_1)

Table 3 shows the GRETL output of the fitted full model (M_1) that is the model with full suspected variables which is mathematically represented in Equations (4) and (5).

3.9. The reduced MLRM (M_2)

Two models were proposed in this study, the full model fitted with all the suspected variable and the modified reduced model fitted with only significant variables.

The proposed modified reduced model is given as follows

Table 8. Model 2: logistics, using observations (n = 407).

No.	Factors	Coefficient	Std. error	z	p-value	
1	Const	−51.4262	14.1918	−3.624	0.0003	***
2	Weight	−0.0655577	0.0432534	−1.516	0.1296	
3	Height	6.82919	5.09405	1.341	0.1800	
4	BPSystolics	0.0833049	0.0357701	2.329	0.0199	**
5	SugarLevel	7.15305	1.65199	4.330	0.0001	***

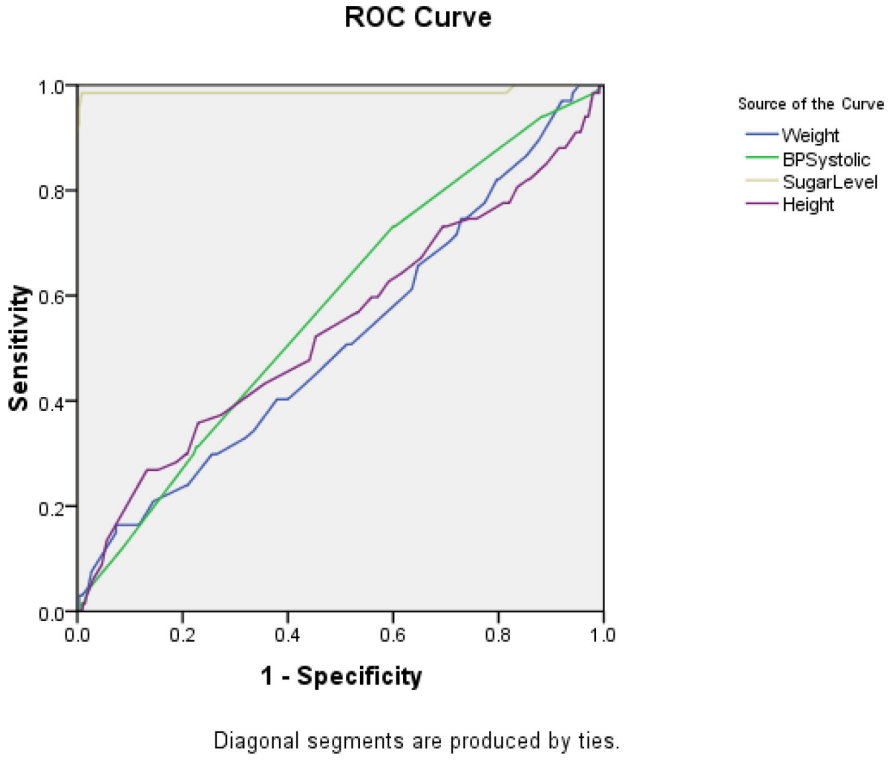

Figure 5. ROC curve for the reduced model.

$$\tau(x) = \frac{e^{\beta_0+X_2\beta_2+X_4\beta_4+X_6\beta_6+X_9\beta_9+X_2+\varepsilon}}{1+e^{\beta_0+X_2\beta_2+X_4\beta_4+X_6\beta_6+X_9\beta_9+X_2+\varepsilon}} \quad (15)$$

From Tables 4 and 5, we have enough evidence to propose the reduced model as in Equation (15) and the modified reduced model with the estimated parameters using the data from the center is represented by Equation (7)

$$\tau(x) = \frac{e^{-51.646+7.0672X_2+7.3909X_4-0.0711X_6+0.0925X_9+\varepsilon}}{1+e^{-51.646+7.0672X_2+7.3909X_4-0.0711X_6+0.0925X_9+\varepsilon}} \quad (16)$$

The above model can be used to predict the probability of a person given the significant data $X_i = (X_2, X_4, X_6, X_9)$,

and traditional interpretation of probability can be used in interpretation of the model results.

3.10. Model selection [full MRLM (M_1) vs. reduced MRLM (M_2)]

The model assortment is brought out using goodness-of-fit measures log-likelihood.

Table 9 reveals the two model which were level M_1 and M_2, and there is evidence that the most suitable model is M_2 (reduced model) with adjusted R square of 86.75 and average ROC area of 0.66. Therefore, the

Table 9. Comparing full model and modified reduced model.

No.	Measure	Full MRLM (M_1)	Reduced MRLM (M_2)
1	Log-likelihood	−18.97108	−19.11508
2	Adjusted R-squared	0.840845 (84.08%)	0.867522(86.75%)
3	Average ROC Area	0.58	0.66

model proposed is M_2 which represented mathematically in Equation (7) above.

4. Conclusion

In this paper, we modeled the diabetes risk factors based on logistic regression model. The risk factors considered were: weight, height BP diastolic, BP systolic, age, gender, employment status, and sugar level. First, the Chi-square test of association between diabetes and all the predictor variables showed that BP systolic, weight, height BP diastolic, age, and sugar level are statistically significant. Logistic regression model was considered as the suitable model since most important assumptions of the model were tested. The significance testing for the logistic coefficients using Wald test prove that factors like weight, height, sugar level, and BP systolic are statistically significant in predicting variables of diabetes. We compared the results of the two models and found that the modified model is more efficient that the full model with all the variables although the model showed that getting diabetes does not statistically depend on age, employment status, gender, BP diastolic, and height of the patients the modified model which is more efficient based in this study can be used to estimate the probability of an individual having diabetes given the significant data of that individual. To assess the fitness of the model, the maximum likelihood test was used, and the results revealed that the reduced modified model called RMLRM is a good model. We hope that the this model may be a focus for wider applications in survival analysis. Furture research using machine learnig approach is recommend with the same data.

References

Alwan A. Raising the priority accorded to diabetes in global health and development: a promising response. Int J Diabet Mell 2010; 3(2): 2139–140.

Andrew B, Broadley MM, Bishop T, White MJ. The relationship between attentional bias to food and disordered eating in females with type 1 diabetes. Appetite 2019; 140:269–76.

Arulampalam W, Naylor RA, Smith JP. A hazard model of the probability of medical school dropout in the United Kingdom. J Royal Stat Soc 2004; 167:157–78; doi:10.1046/j.0964-1998.2003.00717.x

Badi NHS. Asymptomatic distribution of goodness-of- fit tests in logistic regression model. Open J Stat 2017; 7:434–45; doi:10.4236/ojs.2017.73031

Barreiro PL, Albanzo JP. Population and sample, sampling teachnique, management mathematics for European schools, University of Seville, 13 p, 2001. Available via http://www.mathematik.uni-kl.de/~mamaeuschs

Berk RA. Statistical learning from the regression perspective springer series in statistics. Springer International Publishing AG Swaziland, pp 173–9, 2017 Available via www.springer.com

Berkson J. Applications of the logistic function to bioassay. J Am Stat Assoc 1944; 9:357–65.

Bonamente M. Statistics and analysis of scientific data. Springer Science+Business Media LLC, University of Alabama, Huntsville, AL, pp 29–38, 2017.

Borovkov AA. Probability theory. Springer-Verlag, London, UK, pp 614–5, 2013. ISBN: 978-1-4471-5201-9

Cho N, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, et al. IDF diabetes Atlas: global estimate of diabetes prevalence for 2017 and projection for 2045. Diabet Res Clin Pract 2018; 138:271–81; doi:10.1016/j.diabres.2018.02.023

Cleophas TJ, Zwinderman AH. Statistical analysis on clinical data on pocket calculator. Springer Science+Business Media, Springer, Dordrecht, Netherlands; Heidelberg, Germany; London, UK; New York, NY, pp 25–37, 2011. e-ISBN: 978-94-007-1211-9

Connor AE, Kazem RD, Bruce N, Thomas C, Vlado P, Anushka P. Blood pressure lowering in type 2 diabetes a systematic review and meta-analysis, American Medical Association. JAMA 2015; 313(6):603–15; doi:10.1001/jama.2014.18574

Cramer JS. Origin of logistic regression. Tinbergen institute discussion paper TI2002-119/4 1-15, Cambridge University Press, Cambridge, UK, 2002.

Davy V, Christoph UC, Britta G, Michel P, Marc DH, Philip BW, et al. Diabetes mellitus in people with schizophrenia, bipolar disorder and major depressive disorder: a systematic review and large-scale meta-analysis. World Psych 2016; 15(2):166–74.

Demidenko E. The shortest width confidence interval for odds ratio in logistic regression. Open J Stat 2012; 2(3):305–8; doi:10.4236/ojs.2012.23037.

Emdin CA, Rahimi K, Neal B, Calendar T, Perkovic V, Patel A. Blood pressure lowering in type 2 diabetes; a systematic review and meta-analysis American Medical Association. JAMA 2015; 313(6):603–15; doi:10.1001/jama.2014.18574

Forsyth D. Probability and statistics for computer science. University of Illinois Urbana, Champaign, IL, pp 116–37, 2018.

Frank E, Harrell J. Regression models, logistic and ordinary regression, and survival analysis, springer series in statistics. Springer International Publishing AG Swaziland pp 5–79, 2015. Available via www.springer.com

Hosmer WD, Stanley L, Rodney XS. Applied logistic regression. Wiley, New York, NY, 2000; doi:10.1002/0471722146

Hsieh JJ, Pan CC. Linear Regression analysis for symbolic interval data. Open J Stat 2018; 8:885–901; doi:10.4236/ojs.2018.86059

Ike SO, Rodney L. Prediabetes awareness, healthcare provider's advice, and lifestyle changes in American adults. Int J Diabet Mell 2015; 3:11–18.

Kusurkar R, Kruitwagen C, Ten Cate O, Croiset G. Effects of age, gender and educational background on strength of motivation for medical school. Adv Health Sci Educ 2010; 15:303–13; doi:10.1007/s10459-009-9198-7

Kpozehouen A, Djrolo F, Sossa C, Gbary A, Chouehanou Y, Fambo, D, et al. Prevalence and associated factors of diabetes mellitus in Benin. Open J Epidemiol 2015; 5:163–71; doi:10.4236/ojepi.2015.53021.

McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. J Ecol Soc Am 2021; 82(1), 290–297.

Mohammed A, Hannan H, Barra RR, Nazia U, Bhaskar P, Santhosh K. Cognitive impairment in type 2 diabetes mellitus. Int J Diabet Mell 2015; 3:19–24.

Nad OL, Kanuch P. Why Sampling ration matters: logistic regression and study of habitat use. PLoS One 2018; 13(7):e0200742.

Pathirana HPRR, Varathan N. Statistical analysis of variables influencing type of birth in Sri Lanka: a logistic regression approach. Open J Stat 2018; 8(02):317.

Polat K, Salih G. An expert system approach based on principal component analysis and adaptive neuro-fussy inference system to diagnosis of diabetes disease. Selcuk University, Konya, Turkey, pp 702-710, vol. 17, 2007; doi:10.1016/j.dsp.2006.09.005

Ratsimalahelo Z. Generalised Wald type test of nonlinear restrictions. Open Access Lib J 2017; 4:1–8; doi:10.4236/oalib.1103923

Shayib MA. Inferential statistics—the basic for biostatistics. Prairie View A&M University, Houston, TX, pp 11–6, vol. II, 2018. ISBN: 978-87-403-2127-2.

Smokovski I, Milenkovic T, Nam HC. First stratified diabetes prevalence data for Republic of Macedonia derived from the National eHealth System. J Diabet Res Clin Pract 2018; 143:179–83.

Shiraishi H, Taniguchi M, Yamashita T. Higher-order asymptotic theory of shrinkage estimate for general statistical models. J Multivar Anal 2018; 166:198–211.

Stratton AD, Matus C, Sweitzer K. Diabetes management in resource poor countries. J Diabet Mell 2019; 9:24–30; doi:10.4236/jdm.2019.91003

Yu D, Cai Y, Graffy J, Holman D, Zhao Z, Simmons D. Derivation and external validation of risk algorithms for cerebrovascular (Re) hospitalisation in patients with type 2 diabetes: two cohorts study. Diabet Res Clin Pract 2018; 144:74–81; doi:10.1016/j.diabres.2018.08.006

Wakefield J. Bayesian and frequentist regression methods. Springer Science+Business Media, Springer, New York, NY, pp 75–321, 2013. e-ISBN: 978-1-4419-0925-1

Wang Z, Zou Z, Yang Z, Dong Y, Song J, Dong B, et al. The association between fetal-stage exposure to the China famine and risk of diabetes mellitus in adulthood: results from the China health and retirement longitudinal study. BMC Public Health 2018; 18(1):1–7; doi:10.1186/s12889-018-6134-x

Whiting D, Guariguata L, Weil C, Shaw J. IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. Diabet Res Clin Pract 2011; 94:311–21.

Zaoui S, Biemont C, Meguenni K. Epidemiology of diabetes in urban and rural regions of Tlemcen (Western Algeria). Sante 2007; 17:15–21.