International Journal of Computational and Theoretical Statistics
ISSN (2210-1519)
Int. J. Comp. Theo. Stat. 6, No. 2 (Nov-2019)

http://dx.doi.org/10.12785/ijcts/060201

Statistical Issues in Small and Large Sample: Need of Optimum Upper Bound for the Sample Size

Subramanian Chandrasekharan[1], Jayadevan Sreedharan[2] and Aji Gopakumar[3]

[1] *Department of Statistics, Annamalai University, Chidhambaram, Tamilnadu, India*
[2] *College of Medicine, Department of Community Medicine, Gulf Medical University (GMU), Ajman, UAE*
[3] *Research Scholar, Annamalai University, Annamalai Nagar, Tamilnadu, India*

Received October 28, 2018, Revised February 21, 2019, Accepted March 17, 2019, Published November 1, 2019

Abstract: As fewer samples are meaningless and lead to fallacious conclusions, researchers are used to calculate minimum sample size before the conduct of any study. Although the larger samples can yield more accurate results, an extent for maximum sample size is not fixed. Though large samples are able to give précised and accurate estimates, the studies that collect more samples than the minimum required, may lead to fallacious conclusions. Generally, the test statistics are increasing functions of sample size and limit of the p value (as 'n' tents to infinity) results the statistical significance. The current paper investigated the pattern of changes in the estimates and testing results for varying sample sizes. The assessment of this type of patterns in the data and an extended study on this topic will help to find an interval for the sample size. Study concluded with a finding that larger sample does not make differences on the values of descriptive statistics, but has significant impact on the values of inferential statistics and therefore an upper bound for the sample size needs to be fixed. Hence this article gives relevant information about the need of finding adequate sample size interval (n_1, n_2) within which valid statistical conclusions can be derived, that assures significance of real difference.

Keywords: Small sample size, Large sample size, Sample size interval, Upper bound of sample size, Issues at large sample

1. INTRODUCTION

Statistical methods have greater application in conduct of a research from the stage of planning through designing, collecting, analyzing and interpretation of data. A researcher should be familiar with basic concepts of statistical methods which are applied at various stages of a research [1]. One of the important steps in a research is the determination of sample size [2]. Generalization is not possible if the sample size is not adequate [3]. Number of factors needed to be considered while estimating sample size; different formulas are derived to calculate minimum sample size based on the study objective. But there is no modest solution for how large should be a sample for any research. Upper limit of larger samples depends upon the accuracy and the precision required.

The purpose of calculating sample size in an estimation study is to estimate the parameter value. Large data is essential to get more précised estimate as precision increases, marginal of error decreases. Precision can be measured from the confidence interval and the confidence level. A large sample size gives narrowed confidence interval that indicates higher precision. Parameter estimation with 99% confidence required large sample compared to parameter estimation at 95% confidence level. Sample size can also be calculated in accordance with the statistical test planned to apply [4]. Calculation of sample size in studies related to testing of hypothesis is to achieve a desired power at a fixed level of significance to detect clinically or practically valid difference [5].

Though there are many advantages for large sample studies, a number of errors are also associated with large samples. Sampling error occurs due to selection of unrepresentative sample and it is the difference of sample statistic from the actual value of the parameter [6-8]. There are also some issues in large sample associated with inferential results, specifically in p values. It is to be verified that difference in the treatment outcome is only due to the effect of large sample by quantifying the magnitude of the effect size and degree of the effect [9-12]. A large sample study lead to higher power that detects small and subtle effects as statistically significant and therefore decisions taken based on the p value alone, will lead to wrong conclusions [13-15].
Since a large sample is considered as more representative of the population, an adequate large sample is necessary to produce valid results [16]. It is also assumed that the statistic does not lead to normality unless the sample size is adequately large and hence researchers are interested to take larger samples to apply more powerful tests to prove their hypothesis. There are studies which reported the requirement of larger sample if there is comparison among the variables with multiple categories. When multivariate analysis is included in the research, size of the sample can be at least 10 times larger than the number of variables under study. In the

E-mail address: manistat@gmail.com, drjayadevans@gmail.com, ajitha_gopan@yahoo.co.in

case of a sample of 500, sample size assures the sample error which will not exceed 10% of the standard deviation. In a multiple regression analysis, sample size is recommended to select 15 to 20 per predictor variables. A study using factor analysis required 200 samples if there are 10 items [17-21].

Study with a smaller sample provides results on a worthless trial which will be insufficiently powered to detect a real difference [22]. In inferential analysis, the test statistic value gets decrease due to increase in standard error, which may occur by the lesser number of samples. Therefore chance of acceptance of null hypothesis increases and thereby it is unable detect the significance of the effect. In an experiment with larger sample size, unwanted number of study subjects may expose to either a beneficial or risky treatment [23]. Since large sample studies use more resources and hence raise economic concerns. Sample size is a key concern pointing ethical issues in a research involving human or animal subjects.

The main factors considered while calculating sample size are Type I error (α), Type II error (β), Power (1-β), Confidence level, Standard Deviation and the effect size. Various formulas are derived for determining sample size in accordance with the study design (experimental & non-experimental studies) and objectives of the study such as estimation of population parameter or test of various hypotheses. Since larger sample gives more précised results, researchers have growing interest in studies with large sample. This research is important as it studies the need of fixing a maximum limit for the sample size. The study takes interest in finding trend in the statistical results for varying samples from small to large.

2. OBJECTIVE

Though larger sample makes insignificant results significant, lesser margin of error and narrowed confidence interval (indicator of higher precision) can be derived with a large sample data. The solution for an accurate and précised estimate is to include sufficiently large sample size, but no suitable criteria are sorted out that explain maximum adequate level. This paper aimed to identify the need for calculating an adequate upper bound for the sample size (n), also to identify the pattern of changes in the statistical results and conclusions for small to large sample size.

3. METHODOLOGY

This research is conducted on a secondary data which is collected from one of the teaching Hospital of UAE. The data includes patient's lipid profile and clinical parameters which are classified among patients with Diabetes and non-Diabetes. The data is retrieved from hospital online system during Jan-Sept. 2016. Data includes a total of 10406 patient details, of which 6026 are identified with Diabetes (HbA1c≥6.5%) and 4380 were non-Diabetics (HbA1c<6.5%). Few clinical parameters are considered for the current study to identify the pattern of changes in the statistical results for varying sample size. The selected variables are 'Total serum cholesterol, HDL, Hemoglobin, Serum Creatinine and HbA1c' of non-diabetic group. Among these variables, missing observations are also observed except in HbA1c and therefore valid sample size (valid n) is presented in all the tables for more clarity. A small sample of size 50, 100, 200 to a large sample of more than 4000 is selected randomly from the mentioned secondary data.

Various statistical methods such as Descriptive statistics (Mean, Median, Mode, SD, Minimum, Maximum, Standard Error of Mean, Standard Error of Skewness, Standard Error of Kurtosis) and inferential techniques (Kolmogorov Smirno test, Shapiro-Wilk test, Independent single sample test, Chi-square test & Binary logistic regression) are applied to find the trend in statistical results for small to large sample. Measures of central tendency are calculated to identify the pattern of changes in the central point to which the observations of the selected variables are clustered around. Standard Deviation are also presented to show the scatteredness or closeness of the observations with regards to the measure of central tendency. Since the skewness gives the degree and direction of departure from the symmetry, changes in measure of skewness for different sample sizes are also figured in the below tables to identify the presence of lack of symmetry (skeweness=0). Kurtosis is presented to identify the presence of extreme outliers and compared kurtosis of variable distribution to the kurtosis of normal distribution (kurtosis=3). Histogram and Normal Probability plot for varying sample size is also drawn to present the gap between test results and descriptive results. $P<0.05$ is considered as statistically significant. SPSS -23 version is used for the statistical data analysis.

4. RESULTS & DISCUSSION

In the primary section of analysis, Descriptive Statistics are determined on Non-Diabetic group (4380) to show the trend against different samples sizes. In order find pattern of changes in statistics and statistical test results, tables are arranged for sample size n≤300 and n>300.

I. DESCRIPTIVE STATISTICS & INFERENTIAL RESULTS FOR VARYING SAMPLE SIZE

Slight differences are observed in descriptive statistics at different samples, which are not clinically significant difference. In order to check for normality assumption, equality of mean, median & mode are observed and found averages are slightly varied at samples 'n' ≤300 (Table 1). At larger samples (n>300), three averages are found to be almost equal (Table 2). But normality test at smaller samples shows that the variables follow normal distribution. Normality assumption is violated for larger samples, by test of normality (Table 5 & 6). This is due to the increased power of the test at larger samples at which small effects detected as significant. Since normality tests are sensitive at larger samples, normal Q-Q plots, Histograms, skewness and kurtosis are also presented for

verifying the assumptions of normality. Though normality test resulted as not normal for more than 740 samples (valid 'n'), histogram and Q-Q plot of the variable 'cholesterol' shows approximate normality at large and extremely large samples (figure 1-8).

Skewness starts decreasing in all the variables for larger samples. Increasing trend in skewness for large samples is identified only in one variable 'Cholesterol'. As a whole, skewness is not far from zero; it is identified in an acceptable range of ±2 for normality [24-26]. Though the kurtosis is 3 for normally distributed variables, low kurtosis in the selected variables of this study indicates the data is free from outliers (Table 3).

Standard error in the means is having slight and steady decrease for increased samples size. Standard error of the skewness closes to zero for large samples and Standard error of kurtosis also starts declining when size of the sample getting large. Overall, smaller standard errors are identified at larger samples (Table 1 & 4). Slight variations in SD are observed at samples below 300. But at larger samples, SD is almost same and consistent in all the variables. When sample size increases, minimum value gets lower and lower, maximum gets slightly higher. Therefore range increases for increasing sample size, but later gets consistent. Interquartile range is also almost same for larger samples. In small samples, continuous shift in statistics can be seen. When samples get larger, frequency of shift is comparatively less.

II. NORMALITY TEST AT SMALL AND LARGE SAMPLES

Normal model is the most important probability model in statistics. Normal distribution has wider application in all the areas. Entire small sample tests (t, F, chi-square etc) is based on the fundamental assumption that its parent population from which the samples taken follow normal distribution. Many statistical tests rely on the assumption of population distribution is normal. The important properties of Normal Distribution are symmetry and equality of mean, median & mode. In normally distributed data, standard deviation (SD) helps to determine the width of the normal curve. With regards to Normality concept, application of few statistical techniques over various samples is presented in the below section.

Statistics at various Samples: Tables & Figures

Table 1. "Measures of Central Tendency & Dispersion" at samples 'n' ≤300

Statistics at n=50	Valid n	Mean	Median	Mode	SD	Minimum	Maximum	Std. Error of Mean	Std. Error of Skewness	Std. Error of Kurtosis
Cholesterol	25	189.32	182.00	211.00	49.18	119.00	299.00	8.83	0.42	0.82
HDL	23	41.21	40.00	33.00[a]	9.59	26.00	66.00	1.81	0.44	0.86
Hemoglobin	30	14.49	14.60	14.10	1.54	11.00	17.70	0.27	0.41	0.81
Creatinine	25	0.88	0.82	0.64[a]	0.18	0.64	1.34	0.03	0.43	0.83
HbA1c	50	5.83	5.80	5.80	0.37	4.90	6.40	0.05	0.34	0.66
Statistics at n=100										
Cholesterol	46	188.95	191.00	143.00[a]	41.38	119.00	299.00	5.25	0.30	0.60
HDL	54	41.26	39.50	33.00[a]	9.42	26.00	66.00	1.28	0.32	0.64
Hemoglobin	49	14.62	14.70	14.10	1.77	9.80	19.00	0.25	0.34	0.67
Creatinine	54	0.88	0.84	0.79	0.17	0.64	1.34	0.02	0.32	0.64
HbA1c	100	5.88	5.90	5.80	0.38	4.70	6.40	0.04	0.24	0.48
Statistics at n=200										
Cholesterol	112	191.49	190.00	190.00	40.63	109.00	299.00	3.84	0.23	0.45
HDL	99	43.91	42.00	48.00	11.87	26.00	97.00	1.19	0.24	0.48
Hemoglobin	100	14.37	14.55	14.10[a]	1.76	9.80	19.00	0.18	0.24	0.48
Creatinine	99	0.86	0.83	0.72[a]	0.19	0.30	1.34	0.02	0.24	0.48
HbA1c	200	5.84	5.90	5.80	0.41	4.40	6.40	0.03	0.17	0.34

Statistics at n=300										
Cholesterol	158	189.39	188.50	183.00[a]	42.85	109.00	300.00	3.41	0.19	0.38
HDL	144	45.14	43.00	48.00	12.31	25.00	97.00	1.03	0.20	0.40
Hemoglobin	159	14.27	14.60	15.40	1.84	8.40	19.00	0.15	0.19	0.38
Creatinine	162	0.83	0.81	0.81	0.19	0.30	1.41	0.01	0.19	0.38
HbA1c	300	5.82	5.90	5.80	0.42	4.40	6.40	0.02	0.14	0.28

[a]Multiple modes exist. The smallest value is presented in the table

Valid 'n' is the no. of samples excluding missing observations, at which all statistics are calculated

Statistics at Large Samples:

Table 2. "Measures of Central Tendency" at various samples (at n=400 to 4380)

Samples (n)	Average	Cholestrole	HDL	Hemoglobin	Creatinine	HbA1c
n=400	Mean	191.65	45.24	14.15	0.83	5.81
	Median	190.00	43.00	14.50	0.81	5.90
	Mode	188.00	48.00	15.40	0.81	6.10
	valid n	212	194	211	208	400
n=800	Mean	190.97	45.22	13.82	0.82	5.75
	Median	188.00	43.00	14.10	0.81	5.80
	Mode	188.00	39.00	15.40	0.81	5.80
	valid n	393	360	403	365	800
n=1200	Mean	190.54	45.72	13.65	0.80	5.73
	Median	188.00	44.00	13.90	0.79	5.80
	Mode	188.00	41.00	12.70[a]	0.81	5.80
	valid n	557	513	584	512	1200
n=1600	Mean	190.90	45.55	13.64	0.80	5.72
	Median	188.00	44.00	13.80	0.79	5.80
	Mode	188.00	39.00[a]	12.70	0.79	5.80
	valid n	740	679	767	674	1600
n=2000	Mean	190.50	45.31	13.65	0.80	5.73
	Median	188.00	44.00	13.80	0.80	5.80
	Mode	188.00	39.00	15.70	0.79	5.80
	valid n	922	842	948	820	2000
n=2400	Mean	191.48	44.91	13.66	0.81	5.73
	Median	188.00	43.00	13.90	0.80	5.80
	Mode	188.00	39.00	15.70	0.83	5.80
	valid n	1119	1021	1144	982	2400
n=2800	Mean	190.23	44.50	13.72	0.81	5.73
	Median	188.00	43.00	14.00	0.81	5.80
	Mode	188.00	39.00	15.40	0.87	5.80
	valid n	1308	1195	1332	1160	2800
n=3200	Mean	190.35	44.53	13.74	0.82	5.74
	Mode	188.00	39.00	15.40	0.87	5.80
	Median	187.00	43.00	14.00	0.81	5.80
	valid n	1528	1396	1521	1335	3200
n=3600	Mean	190.28	44.60	13.71	0.81	5.74
	Median	187.00	43.00	13.90	0.80	5.80
	Mode	186.00	39.00	14.10	0.87	5.80
	valid n	1713	1567	1710	1478	3600

n=4000	Mean	190.68	44.58	13.71	0.81	5.74
	Median	188.00	43.00	13.90	0.81	5.80
	Mode	186.00	39.00	15.40	0.87	5.80
	valid n	1904	1741	1885	1628	4000
n=4380	Mean	190.73	44.50	13.75	0.81	5.74
	Median	188.00	43.00	14.00	0.81	5.80
	Mode	186.00	39.00	15.40	0.87	5.80
	valid n	2097	1913	2044	1790	4380

Table 3. Measures of Dispersion at various sample size (at n=400 to 4380)

Descriptive Statistics		**n=400**	**n=800**	**n=1200**	**n=1600**	**n=2000**	**n=2400**	**n=2800**	**n=3200**	**n=3600**	**n=4000**	**n=4380**
Cholesterol	Std. Deviation	42.56	41.92	42.20	41.10	39.85	40.52	40.78	41.14	40.64	40.71	40.81
	Minimum	112.00	103.00	88.00	88.00	88.00	88.00	78.00	78.00	78.00	78.00	78.00
	Maximum	299.00	299.00	302.00	302.00	302.00	329.00	329.00	329.00	329.00	332.00	332.00
	Skewness	0.21	0.26	0.21	0.27	0.26	0.37	0.40	0.46	0.47	0.52	0.51
	Kurtosis	-0.48	-0.37	-0.36	-0.33	-0.27	0.13	0.17	0.26	0.28	0.43	0.46
HDL	Std. Deviation	12.70	12.32	12.44	12.90	12.92	12.62	12.57	12.36	12.39	12.17	12.15
	Minimum	19.00	19.00	19.00	19.00	19.00	19.00	19.00	19.00	19.00	19.00	19.00
	Maximum	99.00	99.00	99.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	Skewness	1.25	1.22	1.13	1.09	1.14	1.10	1.06	1.02	1.01	0.99	0.98
	Kurtosis	3.05	2.41	1.85	1.93	2.04	1.99	1.89	1.77	1.63	1.65	1.58
Hemoglobin	Std. Deviation	1.49	1.61	1.68	1.64	1.67	1.68	1.66	1.64	1.65	1.67	1.65
	Minimum	9.40	8.90	8.90	8.90	8.90	7.90	7.90	7.90	7.90	7.90	7.90
	Maximum	17.80	17.80	17.80	17.80	17.80	17.80	17.90	17.90	17.90	17.90	17.90
	Skewness	-0.66	-0.77	-0.65	-0.60	-0.64	-0.68	-0.67	-0.69	-0.65	-0.65	-0.67
	Kurtosis	1.31	0.71	0.30	0.22	0.18	0.36	0.38	0.49	0.34	0.34	0.38
Creatinine	Std. Deviation	0.20	0.20	0.20	0.21	0.20	0.20	0.20	0.20	0.20	0.20	0.20
	Minimum	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
	Maximum	1.48	1.48	1.48	1.71	1.71	1.71	1.71	1.71	1.71	1.71	1.71
	Skewness	0.48	0.47	0.44	0.77	0.71	0.63	0.52	0.48	0.56	0.53	0.58
	Kurtosis	1.42	1.09	0.78	1.90	1.73	1.54	1.39	1.17	1.40	1.36	1.48
HbA1c	Std. Deviation	0.37	0.40	0.40	0.39	0.39	0.40	0.39	0.39	0.39	0.39	0.39
	Minimum	4.80	4.10	4.10	4.10	4.10	4.10	4.10	4.10	4.10	4.10	4.10
	Maximum	6.40	6.40	6.40	6.40	6.40	6.40	6.40	6.40	6.40	6.40	6.40
	Skewness	-0.56	-0.78	-0.73	-0.77	-0.72	-0.69	-0.68	-0.66	-0.66	-0.64	-0.67
	Kurtosis	-0.03	1.44	0.90	0.88	0.61	0.40	0.29	0.25	0.23	0.19	0.32

Table 4. Standard Error at various sample size (at n=400 to 4380)

Sample Size	**Descriptive Statistics**	**Cholesterol**	**HDL**	**Hemoglobin**	**Creatinine**	**HbA1c**
n=400	Std. Error of Mean	2.91	0.91	0.13	0.01	0.02
	Std. Error of Skewness	0.17	0.18	0.17	0.17	0.12
	Std. Error of Kurtosis	0.33	0.35	0.33	0.34	0.24
n=800	Std. Error of Mean	2.12	0.63	0.10	0.01	0.02
	Std. Error of Skewness	0.12	0.13	0.12	0.13	0.09
	Std. Error of Kurtosis	0.25	0.26	0.24	0.26	0.17

n=1200	Std. Error of Mean	1.78	0.53	0.08	0.01	0.01
	Std. Error of Skewness	0.10	0.11	0.10	0.11	0.07
	Std. Error of Kurtosis	0.21	0.22	0.20	0.22	0.14
n=1600	Std. Error of Mean	1.52	0.46	0.07	0.01	0.01
	Std. Error of Skewness	0.09	0.09	0.09	0.09	0.06
	Std. Error of Kurtosis	0.18	0.19	0.18	0.19	0.12
n=2000	Std. Error of Mean	1.32	0.42	0.06	0.01	0.01
	Std. Error of Skewness	0.08	0.08	0.08	0.09	0.05
	Std. Error of Kurtosis	0.16	0.17	0.16	0.17	0.11
n=2400	Std. Error of Mean	1.24	0.37	0.06	0.01	0.01
	Std. Error of Skewness	0.07	0.08	0.07	0.08	0.05
	Std. Error of Kurtosis	0.15	0.15	0.15	0.16	0.10
n=2800	Std. Error of Mean	1.16	0.34	0.05	0.01	0.01
	Std. Error of Skewness	0.07	0.07	0.07	0.07	0.05
	Std. Error of Kurtosis	0.14	0.14	0.13	0.14	0.09
n=3200	Std. Error of Mean	1.08	0.32	0.05	0.01	0.01
	Std. Error of Skewness	0.06	0.07	0.06	0.07	0.04
	Std. Error of Kurtosis	0.13	0.13	0.13	0.13	0.09
n=3600	Std. Error of Mean	1.00	0.30	0.05	0.01	0.01
	Std. Error of Skewness	0.06	0.06	0.06	0.06	0.04
	Std. Error of Kurtosis	0.12	0.12	0.12	0.13	0.08
n=4000	Std. Error of Mean	0.95	0.29	0.04	0.01	0.01
	Std. Error of Skewness	0.06	0.06	0.06	0.06	0.04
	Std. Error of Kurtosis	0.11	0.12	0.11	0.12	0.08
n=4380	Std. Error of Mean	0.90	0.27	0.04	0.00	0.01
	Std. Error of Skewness	0.05	0.06	0.05	0.06	0.04
	Std. Error of Kurtosis	0.11	0.11	0.11	0.12	0.07

Though there is only slight variation observed in mean, median & mode, non-normality is observed when size of the sample starts increased. Confidence interval of the mean is getting less wider when sample size increases. Since Kolmogorov Smirnov test (KS test) is generally used for large samples and Shapiro Wilk test (SW test) is suitable for small samples, variables were tested for normality by both the tests at 5% level of significance. Test results at samples around 100 and 200 showed that variables are normally distributed in the population (table 5). This often happens due to KS & SW tests are less powerful at smaller samples in rejecting null hypothesis [27, 28]. But when sample size increased, normality assumption is violated for the selected variables since KS & SW tests are sensitive for larger sample. KS test of normality resulted with a significant difference from the normal distribution with an evidence of decreasing p values for large samples (Table 6). Though larger samples ensure normality by Central Limit Theorem, statistical test produces non-normality as they are more powerful at larger samples in detecting insignificant small deviations as significant. This is ascertained by plotting histogram and Q-Q plots (figure 1-8).

In the plots, the variable 'Cholesterol' is normally distributed at small samples, but normality assumption is lost at a sample size above 1600 (valid n 740). But normal plots not match with the normal test results of significant departure from normality. Plots on the variable 'Cholesterol' results 'approximate normality' among small to large samples and satisfied the properties of normal distribution, though the tail of the distribution starts slightly skewed to the right from n=2400 (valid n 1119) onwards. But normality is not supported by KS test at larger samples specifically at number of samples around 1600 (valid n 740, p=0.03) and above.

Table 5. Test of Normality for samples below 300

P values at various samples		**n=100**		**n=200**		**n=300**	
		KS test	Shapiro-Wilk test	KS test	Shapiro-Wilk test	KS test	Shapiro-Wilk test
Cholesterol	p value	NS	NS	NS	NS	NS	NS
	Valid n	46		112		158	
HDL	p value	NS	NS	NS	NS	p<0.05	p<0.01
	Valid n	54		99		144	
Hemoglobin	p value	NS	NS	NS	NS	p<0.05	p<0.05
	Valid n	49		100		159	

Creatinine	p value	NS	NS	NS	NS	NS	NS
	Valid n	54		99		162	
HbA1c	p value	NS	p<0.001	p≤0.001	p<0.001	p<0.001	p<0.001
	Valid n	100		200		300	

NS- Not significant (p>0.05)

Table 6. Test of Normality for samples 400-4380

p values at various samples		n=400	n=800	n=1200	n=1600	n=2000	n=2400	n=2800	n=3200	n=3600	n=4000	n=4380
Cholesterol	p value	NS	NS	NS	P<0.05	P<0.05	P<0.05	P<0.05	P<0.01	P<0.001	P<0.001	P<0.001
	Valid n	212	393	557	740	922	1119	1308	1528	1713	1904	2097
HDL	p value	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001
	Valid n	194	360	513	679	842	1021	1195	1396	1567	1741	1913
Hemoglobin	p value	P<0.05	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001
	Valid n	211	403	584	767	948	1144	1332	1521	1710	1885	2044
Creatinine	p value	P<0.01	P<0.01	P<0.05	P<0.01	P<0.01	P<0.01	P<0.01	P<0.01	P<0.01	P<0.01	P<0.001
	Valid n	208	365	512	674	820	982	1160	1335	1478	1628	1790
HbA1c	p value	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001	P<0.001
	Valid n	400	800	1200	1600	2000	2400	2800	3200	3600	4000	4380

NS- Not significant (p>0.05)

Figures – Histograms and Q-Q plots on the variable cholesterol at various samples

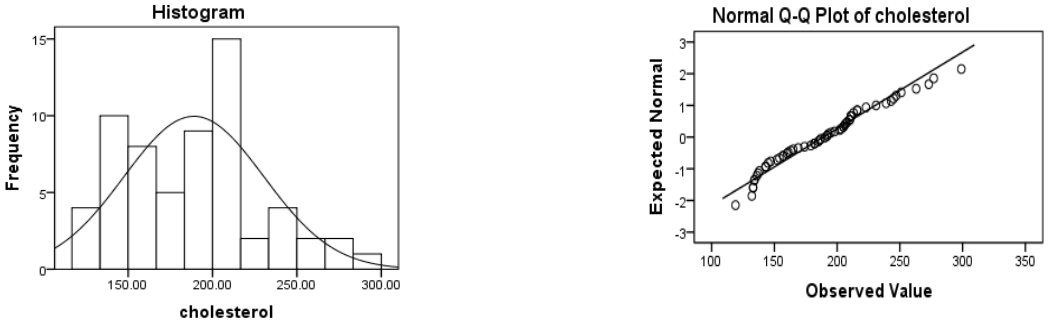

Figure 1. Histogram and Q-Q plot at n=100 (valid n=46)

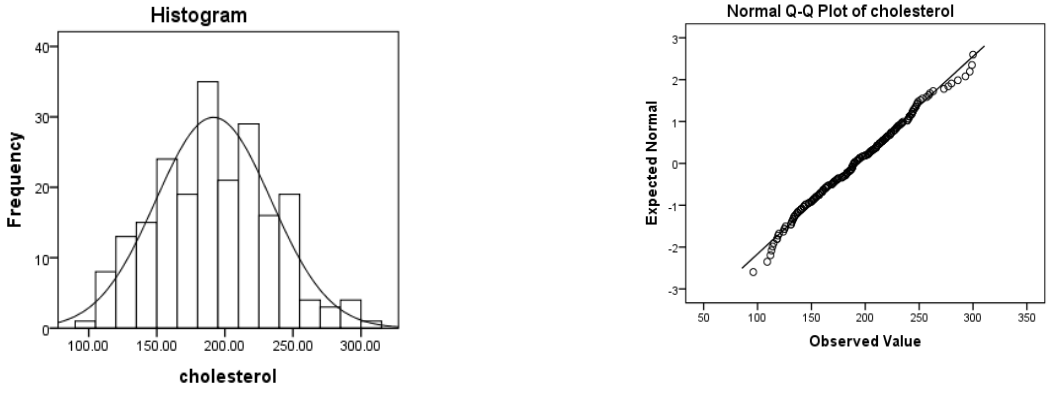

Figure 2. Histogram and Q-Q plot at n=400 (valid n=212)

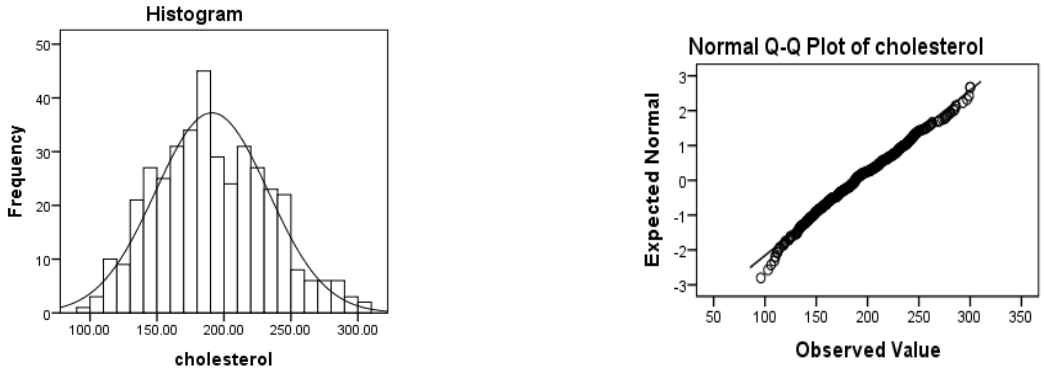

Figure 3. Histogram and Q-Q plot at n=800 (valid n=393)

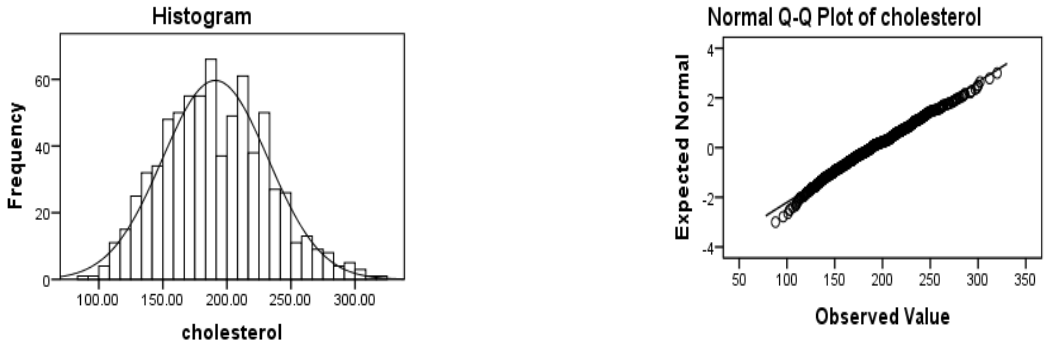

Figure 4. Histogram and Q-Q plot at n=1600 (valid n=740)

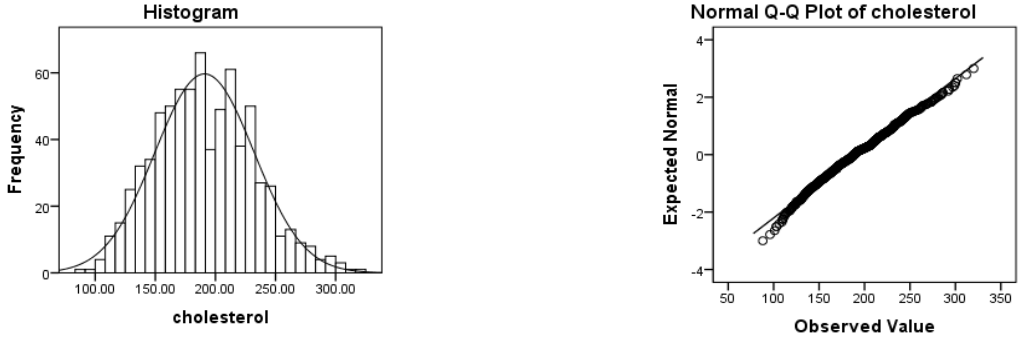

Figure 5. Histogram and Q-Q plot at n=2000 (valid n=922)

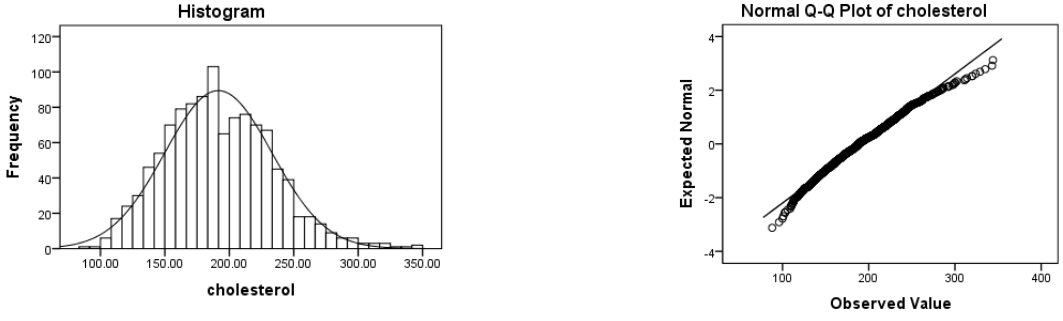

Figure 6. Histogram and Q-Q plot at n=2400 (valid n=1119)

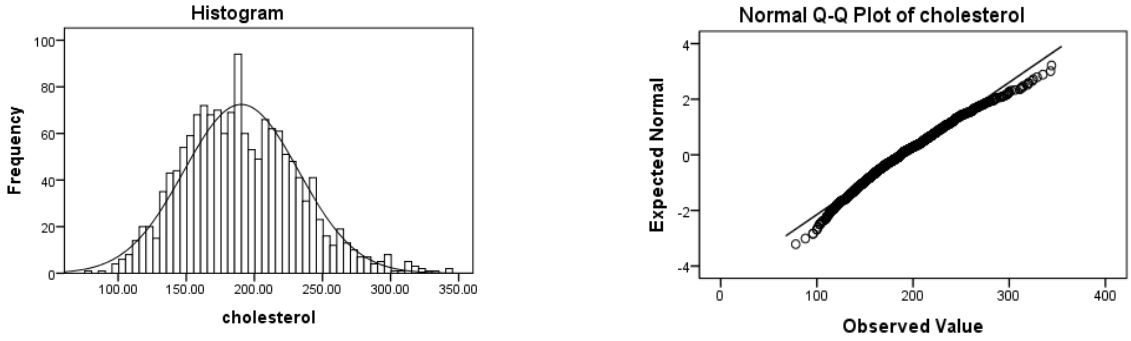

Figure 7. Histogram and Q-Q plot at n=3200 (valid n=1528)

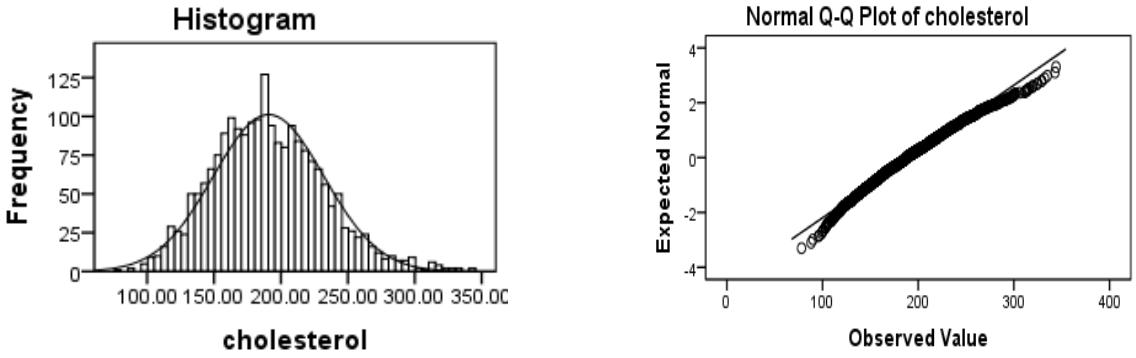

Figure 8. Histogram and Q-Q plot at n=4380 (valid n=2097)

Overall, measures of central tendency and dispersion resulted from the data not depicts a serious departure from symmetry and normality. While tested for normality of the selected continuous variables using Kolmogorov Smirnov (KS) test, variables were found to be non-normal for large samples size. Similarly, other statistical tests resulted with smaller p values for increased samples which indicate differences/effects are statistically significant. But conclusions derived based on statistics are not giving such notions. Implies, after a point of sample size, values of statistics become consistent and same. Therefore studies do not need to go behind larger samples since the value of descriptive statistics gets approximately same and test results sensitive at large 'n'.

Descriptive and inferential techniques are performed on the data after removing probable outliers which are decided based on clinically possible upper limits of the selected parameters and existing statistical concepts. Results are presented in the tables 1-9 at various sizes of samples in a difference of 100 and 400.

III. TEST OF MEAN

Consider the variable 'cholesterol' which is normally distributed at a sample size below 1600 (valid n 740). Therefore a parametric single sample test (Table 7) is applied to find whether the mean cholesterol (μ_0=191) is significantly differ across various samples (testing H_0:μ=191). The test results show a decreasing p values from small to larger samples indicating probable statistical significance at extreme large samples. This also proves that smaller p values are the result of statistical tests at large samples. Narrowed Confidence Interval for the effect size can also be seen for increasing sample size. Since KS test proved non-normality for the variable 'Cholesterol' beyond 2000 samples (valid n 922), parametric test is not performed later on.

Table 7. Parametric Single sample Test

Independent sample test to find whether the difference in the Mean Cholesterol is statistically significant from 191 (μ_0=191, identified at sample size=4380)							95% Confidence Interval of the Difference	
Non-Diabetic Samples (n)	**Valid n**	**Mean**	**Std. Deviation**	**Std. Error Mean**	**p value (NS)**	**Mean Difference**	**Lower**	**Upper**
100	46	190.75	47.5	7.1	0.97	-0.2	-14.5	14.0
200	112	191.49	40.6	3.8	0.89	0.5	-7.1	8.1
300	158	190.32	43.1	3.5	0.84	-0.7	-7.6	6.3
400	212	190.18	41.9	3.1	0.79	-0.8	-6.9	5.3
800	393	190.97	42.1	2.1	0.73	-0.7	-4.9	3.5
1200	557	190.54	41.9	1.8	0.52	-1.2	-4.6	2.3
1600	740	190.13	40.5	1.5	0.56	-0.9	-3.8	2.1

NS- Not significant (p>0.05)

IV. TEST OF CORRELATION

Correlation between variables 'HbA1c and Cholesterol' is depicted in Table 8. Non-parametric Spearman Rank correlation method is applied to find the degree of correlation since HbA1c is not a normally distributed variable. At various samples, a negative weak correlation is observed between two variables. p value is found to be large for smaller samples (p value >0.05). Statistical significance at larger samples resulted due to decrease in p values for increasing samples size.

Table 8. Correlation between HbA1c and Serum Cholesterol

Correlation Test	**n=100**	**n=200**	**n=300**	**n=400**	**n=800**	**n=1200**	**n=1600**	**n=2000**	**n=2400**	**n=2800**	**n=3200**	**n=3600**	**n=4000**	**n=4380**
Correlation Coefficient, 'r'	-0.157	-0.067	-0.074	-0.115	-0.098	-0.092	-0.051	-0.048	-0.037	-0.052	-0.051	-0.050	-0.044	-0.045
p value	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	p<0.05	p<0.05	p<0.05	p<0.05
Valid n	46	112	158	212	393	557	740	922	1119	1308	1528	1713	1904	2097

NS- Not significant (p>0.05)

V. CHI-SQUARE TEST OF ASSOCIATION AT SMALL AND LARGE SAMPLES

In the study of size 4380, 2703 patients were males. An association test is performed to identify the effect of gender on patient's HbA1c level and pattern of p values is studied in Table 9.

Table 9. Degree of Association between patient's Gender and HbA1c Level for varying sample size 100≤n≤4380 (By Pearson's Chi-square test and Binary Logistic Regression)

Sample Size (valid n)	**Gender**	**Normal (HbA1c <5.7)**		**Pre-Diabetic (HbA1c 5.7-6.4)**		**P value**	**OR**	**CI for OR**	**% of Prediction**
		No.	**%**	**No.**	**%**				
n=100	Female	19	46.3	22	53.7	NS (p=0.075)	1	--	64
	Male	17	28.8	42	71.2		2.1	0.93-4.91	
n=200	Female	41	54.7	34	45.3	P<0.005	1	--	64
	Male	43	35.0	80	65.0		2.2	1.24-4.03	
n=300	Female	69	55.6	55	44.4	P≤0.001	1	--	60
	Male	64	36.4	112	63.6		2.2	1.37-3.51	
n=400	Female	81	53.6	70	46.4	P<0.001	1	--	62
	Male	80	32.1	169	67.9		2.4	1.61-3.71	
n=800	Female	138	51.5	130	48.5	P<0.001	1	--	63
	Male	167	31.4	365	68.6		2.3	1.72-3.14	

n=1200	Female	242	49.6	246	50.4	P<0.001	1	--	61
	Male	224	31.5	488	68.5		2.1	1.69-2.72	
n=1600	Female	324	51.0	311	49.0	P<0.001	1	--	61
	Male	313	32.4	652	67.6		2.2	1.77-2.67	
n=2000	Female	392	49.9	394	50.1	P<0.001	1	--	61
	Male	396	32.6	818	67.4		2.1	1.71-2.47	
n=2400	Female	472	50.9	456	49.1	P<0.001	1	--	61
	Male	480	32.6	992	67.4		2.1	1.81-2.53	
n=2800	Female	517	50.7	503	49.3	P<0.001	1	--	61
	Male	580	32.6	1200	67.4		2.1	1.82-2.49	
n=3200	Female	588	50.2	583	49.8	P<0.001	1	--	61
	Male	653	32.2	1376	67.8		2.1	1.83-2.46	
n=3600	Female	690	50.4	680	49.6	P<0.001	1	--	62
	Male	701	31.4	1529	68.6		2.2	1.93-2.54	
n=4000	Female	779	50.4	766	49.6	P<0.001	1	--	62
	Male	764	31.1	1691	68.9		2.3	1.97-2.57	
n=4380	Female	835	49.8	842	50.2	P<0.001	1	--	62
	Male	843	31.2	1860	68.8		2.2	1.93-2.48	

HbA1c level is categorized into normal (HbA1c<5.7) and pre-diabetic (HbA1c 5.7-6.4) group since the data does not include diabetic patients. Frequency distribution of HbA1c across males and females is studied and degree of association is presented with Odds Ratio (OR) by the method of Binary Logistic Regression. Confidence Interval of OR is also provided. Percentage of prediction is also presented to show the validity of regression model. From small to large sample, statistical significant association between gender and HbA1c level is observed with decreasing p values. Though degree of association (OR) is same at various samples, association is statistically significant for larger samples.

5. CONCLUSION

Values of statistics (Mean, Median, Mode, Skewness, Kurtosis, Standard deviation and Standard Error) are almost equal at different number of samples (small/large). Though the values of descriptive statistics shows slight variations at small samples, it is almost same and consistent at larger samples. Implies, larger size of the sample does not make differences on the values of descriptive statistics. But, variation in sample size has significant impact on the results of inferential techniques which helps to conclude about the population characteristics. When sample is large, the results of the statistical tests are not completely reliable as smaller effects turned to be statistically significant. This study showed that there is a need of identifying an upper bound for the sample size so that the conclusion of the study will be more reliable within that specific sample size interval. An extended research is in progress for fixing an adequate upper bound for the sample size (n).

6. ACKNOWLEDGMENT:

The authors would like to acknowledge Annamalai University, India and Gulf Medical University, UAE for the support provided to conduct the present research.

CONFLICT OF INTEREST: Nil

FUNDING SOURCE: Nil

REFERENCES:

1. W.W. Daniel, "Biostatistics: A Foundation for Analysis in Health Sciences," 7th ed., Singapore, Asia: John Wiley and Sons Pte. Ltd, 2004.
2. V. S. Binu, S. S. Mayya, and M. Dhar M, "Some basic aspects of statistical methods and sample size determination in health science research," An International Quarterly Journal of Research in Ayurveda, 35(2), pp. 119-123, 2014.
3. A. Delice, "The Sampling Issues in Quantitative Research," Educational Sciences: Theory & Practice, Autumn, 10(4), pp. 2001-2018, 2010.
4. T. Dahiru, A. Aliyu, and T.S. Kene, "Statistics in Medical Research: Misuse of Sampling and Sample Size Determination," Annals of African Medicine, 5(3), pp. 158–161, 2006.

5. C. S. Chow, H. Wang, and J. Shao, "Sample Size Calculation in Clinical Research," USA, Chapman and Hall/CRC Press, 2003.N. Burns, and S. K. Grove, "The Practice of Nursing research: Appraisal, Synthesis, and Generation of evidence," 6th ed., St. Louis, MO: Saunders Elsevier, 2009.
6. N. Burns, and S. K. Grove, "The Practice of Nursing research: Appraisal, Synthesis, and Generation of evidence," 6th ed., St. Louis, MO: Saunders Elsevier, 2009.
7. F. Scheuren, "What is a Margin of Error?," Springer-Verlag, Chapter 10, 2005.
8. R. M. Kaplan, D. A. Chambers, and R. E. Glasgow, "Big data and large sample size: A cautionary note on the potential for bias," Clinical and Translational Science, 7(4), pp. 342-6, 2014.
9. J. Khalilzadeh, and T. Asli, 2017. "Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research," Tourism Management, 62, pp. 89-96, 2017
10. GM. Sullivan, and R. Feinn, "Using Effect Size-or Why the P Value Is Not Enough," The Journal of Graduate Medical Education, 4(3), pp. 279-82, 2012.
11. S. Nakagawa, and I. C. Cuthill, "Effect size, confidence interval and statistical significance: a practical guide for biologists," Biological reviews of the Cambridge Philosophical Society, 82(4), pp. 591-605, 2007.
12. P. Mirmiran, Z. Bahadoran, M. Golzarand, H. Zojaji, and F. Azizi, "A comparative study of broccoli sprouts powder and standard triple therapy on cardiovascular risk factors following H.pylori eradication: a randomized clinical trial in patients with type 2 diabetes," Journal of Diabetes and Metabolic Disorders, 13(1), 64, 2014.
13. J. Faber, and L. M. Fonseca, "How sample size influences research outcomes," Dental press journal of orthodontic, 19(4), pp. 27-9, 2014.
14. P. Runkel, "Large Samples: Too Much of a Good Thing?," The Minitab Blog, 2012.
15. M. Lin, H.C. Lucas, G. Shmueli, "Too Big to Fail: Large Samples and the p-value Problem," Information Systems Research, Articles in Advance, pp. 1–12, 2013.
16. M. Patel, V. Doku, and L. Tennakoon, "Challenges in recruitment of research participants," Advances in Psychiatric Treatment, 9(1), pp. 229-238, 2003.
17. T. Lumley, P. Diehr, S. Emerson, and L. Chen, "The importance of the normality assumption in large public health data sets," The Annual Review of Public Health, 23, pp. 151-69, 2002.
18. K. Siddiqui, "Heuristics for Sample Size Determination in Multivariate Statistical Techniques," World Applied Sciences Journal, 27 (2), pp. 285-287, 2013.
19. P. Kadam, and S. Bhalerao, "Sample size calculation," International Journal of Ayurveda Research, 1(1), pp. 55-57, 2010.
20. J. Charan, and T. Biswas, "How to Calculate Sample Size for Different Study Designs in Medical Research?" Indian Journal of Psychological Medicine, 35(2), pp. 121-126, 2013.
21. B. K. Nayak, "Understanding the relevance of sample size calculation," Indian Journal of Ophthalmology, 58(6), pp. 469-470, 2010.
22. P. Bacchetti, C. E. McCulloch, and M. R. Segal, "Simple, defensible sample sizes based on cost efficiency," Biometrics, 64(2), pp. 577-85, 2008.
23. R. Lenth, "Some Practical Guidelines for Effective Sample-Size Determination," The American Statistician, 55(3), pp. 187-193, 2001.
24. H.Y. Kim, "Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis," Restorative Dentistry & Endodontics, 38(1), pp. 52-4, 2013.
25. W. M. Trochim, and J. P Donnelly, "The research methods knowledge base," 3rd ed., Cincinnati, OH:Atomic Dog, 2006.
26. F. Gravetter, and L. Wallnau, "Essentials of statistics for the behavioral sciences" 8th ed., Belmont, CA: Wadsworth, 2014.
27. D. Oztuna, A. H. Elhan, and E. Tuccar, "Investigation of four different normality tests in terms of type 1 error rate and power under different distributions," Turkish Journal of Medical Sciences, 36(3), pp. 171-6, 2006.
28. A. Ghasemi, and S. Zahediasl, "Normality Tests for Statistical Analysis: A Guide for Non-Statisticians," International Journal of Endocrinology and Metabolism 10(2), pp. 486-489, 2012.