Is working good for your health?

Chris Keitel
Big Data Analytics
New York University
ck1456@nyu.edu

George Dagher
Big Data Analytics
New York University
gd793@nyu.edu

Khen Price
Big Data Analytics
New York University
cp1425@nyu.edu

Abstract—

We explore whether there is a correlation between employment and changes in health by looking at employment statistics and in-patient hospital data for New York State. Rather than tracking individuals in a longitudinal study of direct cause and effect, this project studied the larger impact of unemployment as a macro-economic effect on social outcomes in geographic areas. Using comprehensive data gathered by central agencies across many years, we show that the macro-economic goals of the government are aligned with the well-being of the population in aggregate.

Keywords—unemployment, health, jobs, analytics

I. INTRODUCTION

Using New York state employment statistics by county, and in-patient hospital data by zip code, we explore a correlation between unemployment and changes in health. All data is de-identified (anonymized) so we have no way of knowing if a particular person lost their job and then spent more time in the hospital as a result, or whether they got a job, and experienced a workplace injury, for example. However, we were able to obtain and analyze relevant metrics on a per-county basis for each of New York's 62 counties for the years 2009 through 2012. Based on the depth of data and the authority by which it was gathered, this serves as a compelling proxy for public health and employment analysis.

The purpose of this study is to illustrate an objective trend that can be used to frame the discussion of policy relating to unemployment initiatives and attitudes towards public health and health care at the state and municipal level. Critically, because of the recent political pressure to alleviate the overwhelming national dependence on employer-sponsored health care, the results can also be interpreted by individuals when evaluating their employment situation, health status, and health care needs.

II. MOTIVATION

The relationship between jobs and health is inextricably bound up with the two most important cultural events of the last decade in this country. The first was the global recession beginning in 2008 which started in the U.S. and has left an indelible impact on the economy and the workplace as many companies failed and millions of people were laid off. During the recession national and regional unemployment was as high as 10.0% [5] and a major policy initiative of federal and state governments has been to reduce unemployment ever since.

The second major social event is the far-reaching reform of the health care system through the Affordable Care Act passed in 2010. In part because there has been such a long-standing tradition of employer-sponsored health care, losing one's job as a result of the recession inevitably had a dramatic impact on one's access to health care. This study attempts to determine whether this impact goes further and actually affects health.

Does having a job encourage good health because you have access to health care, are more likely to seek preventative care, and are surrounded by a structured environment where health is correlated with productivity and is heavily incentivized? Or, on the contrary, does working many long hours promote stress and the onset of diseases that would otherwise be avoided in lifestyle with a more reasonable pace? By definition workplaces injuries happen at work, so if people were not at work they would not have been injured. The nearly 3 million workplace injuries reported in 2012 [6] are certainly a direct negative impact of work on health.

In a time when it is commonplace for workers to complain that "my job is killing me", it is particularly important to understand whether or not that is true. There is ubiquitous anecdotal evidence that as the economy has recovered, industrial productivity has increased much faster than the job creation rate would justify. This implies that the workers who didn't lose their job several years ago are now managing to fill their coworkers' roles by working longer and harder. As employees routinely have to compete with ever-increasing technological productivity benchmarks, it is conceivable that jobs really are causing deterioration in mental and physical health. In the report we do not attempt to identify the precise mechanisms that are speculated above and result in better or worse health among individuals. Furthermore, this study does not consider mental well-being, but solely quantitative metrics of injury and illness based on being admitted to a hospital.

While the authors do not expect anyone to quit their job in the name of staying healthy, it should be vitally important for political and industry decision makers to understand that the national unemployment number should not be the sole target metric for supporting the national well-being of individuals through policy decisions.

III. Related Work

In a related study of a Swedish population spanning 16 years, several findings corroborate links between Socio-Economic Status (SES) and health status [7]. The researchers point out the reciprocal manners in which this relationship exists: namely, better SES can under certain circumstances be a predictor of maintaining good health status. The latter, however, can also be viewed as a selector into an occupation class. In other words, better health as an initial condition may raise the probability of one attaining a more desired occupation over time. As per SES indicators, class of origin (occupational class of the parents), occupational position, education, and income are used. This serves in showing that while occupation is usually thought to affect health status due to variance in income, in actuality income levels influence health but only indirectly. It can be argued that income influences education, which in turn influences health. Additionally, in this research, occupational classes are subdivided internally (for example lower white collar and higher white collar positions), which enabled more insight in the analysis stage.

This research used subjective health condition reports as the main health status indicator. Our study uses objectively calculated institutional metrics in an attempt to mitigate self-reporting bias which could be implicated in much of the existing work in this field.

According to a recent working paper by the National Bureau of Economic Research [8] using a longitudinal study of U.S. adult males, there is a correlation between job classes and changes in health status over time. Namely, men with blue-collar jobs, or more physically oriented occupations, reported a greater probability of transitioning from good health status to bad, compared to white-collar jobs and service jobs. However, there was no indication of blue-collar jobs being correlated with a lesser chance of transitioning back to good health. Again, the health indicators here were reported by the subjects themselves.

In order to justify the use of in-patient hospital discharge reports as a proxy for public health, our analysis depends on an association between health status and health care utilization. A comprehensive report produced by the CDC [11] identifies reasons and indicators of why people get medical care with specific emphasis on the forces that affect utilization of hospital services. The report identifies enabling factors such as health insurance coverage and ability to pay as important reasons why health care utilization goes up. But it also indicates that a considerable factor is whether people realize that they need care in the first place. This can be informed by social factors including a culture of health or social cues that would encourage a person to identify themselves as unhealthy and seek treatment. One such cultural environment is the workplace, where employers have an incentive to make sure workers are healthy and productive and colleagues in close settings are likely to point out a noticeable change in your health.

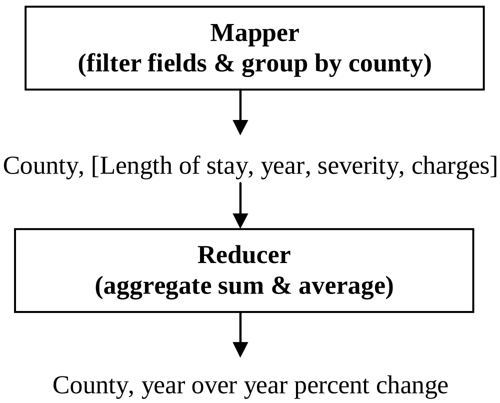
A methodological challenge in determining whether or not there exists a correlation between employment status and overall health arises from the ambiguity of whether employment leads to better health or if better health leads to employment. In other words, it will not suffice to test whether an unemployed person is unhealthy, because their unemployment status might be a direct result of their health. To overcome this challenge, Ross and Mirowsky [9] ran their experiment on the same group of working individuals at different times. Their results suggest that the individuals who lost their job after the initial test showed signs of deteriorated health. It must be noted that specifically, involuntary loss of employment led to the deterioration of health. Individuals who willingly left their jobs did not display the same decline in health. Another interesting fact is that adjustments in pay accounted for a small part of the effect on health. An individual's health was measured in two ways: the first was how the individual perceived his/her own sense of health (self-reporting); the second was to use a standard index of health that was fixed across all subjects.

Jin, Shah, and Svoboda [10] came to the same conclusion as above in that an involuntary loss of one's job results in a decline in health. Their research, however, focuses specifically on individuals who died after having lost their jobs and analyzes the reasons that caused them to die. They found that the main two reasons that caused the passing of individuals in the period after their termination were cardiovascular disease and suicide. It is presumed that the losing of one's job results in more stress which accelerates cardiovascular disease. For suicides, the evidence varies amongst the different countries. Some show a strong correlation while other data suggest a very small amount. In addition, they demonstrated that there is a correlation between one losing his/her job and an increase in higher amounts of drinking alcoholic beverages. The data suggests that this decline was present amongst men and women and across varying ethnic backgrounds.

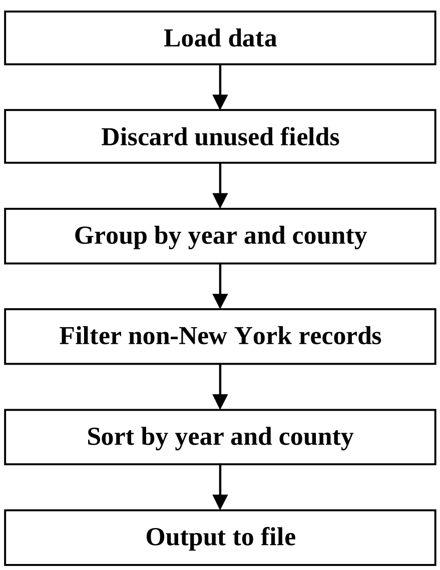
In a paper aimed specifically at public health policy recommendations, Adler and Newman forcefully state the positive correlation between individuals' socioeconomic status (SES) and health based on summary results and recommendations from the UK and empirical data from the U.S. [12] In the author's view, the primary measures of socioeconomic status are education, income, and occupation. Thus unemployment appears to be a uniquely positioned metric to capture a useful correlation because it encompasses both income and occupation, particularly for people in the margin between high SES and low SES (because their income is almost entirely dependent on occupation, where that might not be the case for higher SES). However, as we had also realized, this study highlights a major challenge in determining the time-lag between a decline in SES and a decline in health, or the reverse. Adler and Newman even indicate a possibility that there could be a decline in health in *advance* of a rise in unemployment due to forecast job cuts and anxiety about job security. While our work attempts to prove a similar correlation, the timeframe for a change in health associated with a change in employment is normalized across the population and specific instances as described do not adversely affect the ultimate findings.

IV. Design

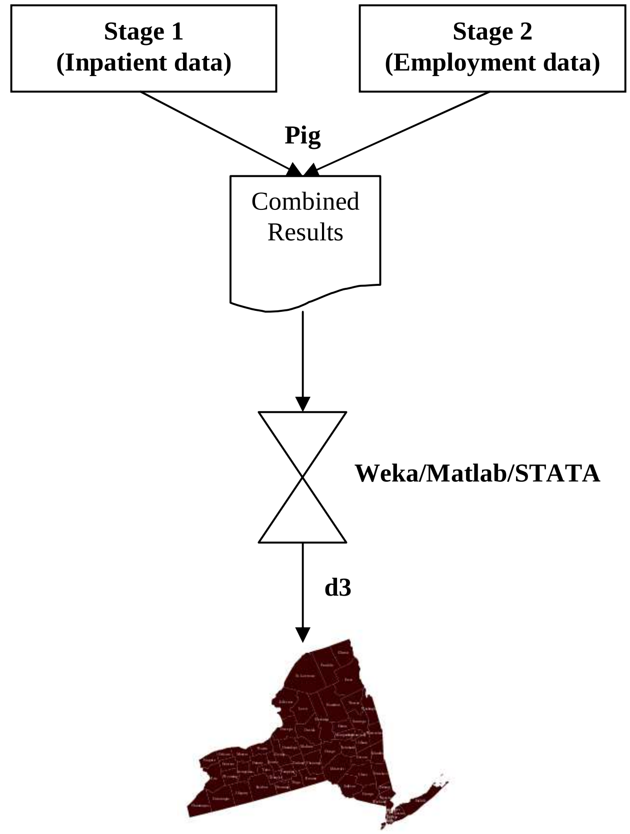
In order to analyze our data sets we used a combination of Hadoop MapReduce and Pig as well as several statistical analytics tools (including Weka, STATA, Matlab) in a multi-stage processing the pipeline. The first stage takes inpatient hospital stay data and produces a summary report for each county (4GB → 500KB):



In the second stage, Pig is used to efficiently prune and reorganize unemployment statistics for the entire country and limit to only counties of interest (1MB+ → 50KB):



In stage 3, the results of stage 1 and 2 are joined and the two types of values are compared on a per county basis using Pig. The resulting correlations are visualized using d3:



Preprocessing of the data included removal of outliers and normalization. Additionally, to use certain algorithms, some attributes would need to be converted from numeric to nominal values. For example, instead of using the actual unemployment rates percentage, we calculate annual changes. We then group the changes into three – increase, no change, and decrease. Finally, we tabulate the data in several ways to seek possible correlation.

With the combined results, we used several tools to explore possible correlations. One track used a machine learning tool called Weka (http://www.cs.waikato.ac.nz/ml/weka/) to discern correlation coefficients for functional dependencies across all available attributes in the data.

Another approach used Matlab to perform direct cross-correlation analysis for select attributes that takes into account variation in metrics over time for each county. This analysis fed directly into d3 visualizations that give the highest level snapshot of the combined data. This proved to be an effective way to easily get an intuitive at-a-glance view of the data that is completely obscured by the volume of the data in its original format.

The last route that we pursued was a detailed statistical analysis in STATA with a simplified model of how unemployment actually affects individuals by county. This led to the most statistically relevant and surprising results that we obtained.

V. Results

In an attempt to find broad correlations between unemployment and health, we tried to build a model that could predict unemployment rates based on health data.

Since the data we summarized spans only 4 years of measurements for 58 out of 62 counties in NY state (data for 4 counties was missing from the NY open data dataset), we disregarded the time factor when building the model, and form one column for each attribute, per county. It should be noted that this step was taken after an unsuccessful attempt to build a model with the time information taken into account. **Table 1** shows nominalized data that captures the change for a select set of counties between 2009 and 2010.

AgvStay_Change	AgvSev_Change	Unemployment_Change
-0.003198694528	3.229114917	0
0.09561654493	1.262250522	1
0.04643041335	3.596058411	1
0.162756353	2.970314355	1
0.4929466343	2.575361812	1
-0.2188391935	2.757300732	0
0.03536426083	3.351784987	-1
0.06531031416	2.199749927	0
-0.07144239977	3.281937023	1
0.346279395	2.473422592	0
-0.08691772861	2.585461701	0

Table 1 Average stay days in the hospital, average severity, and unemployment. For all three, in this instance, changes across years are examined instead of the absolute values.

Several methods were tested using Weka, such as a J48 decision tree, logistic regression, and multilayer perceptron. However, none resulted in a usable model which could predict a change in unemployment rates with high correctness and without large error. This was the case with or without consideration of the time data. This result was disappointing because it initially seemed to nullify our intuition about the interaction between employment and health in the real world.

Regardless, our continued work with the datasets led us to focus on fewer variables that could capture a meaningful real-world effect:

- Average Stay Days
- Average Severity
- Total Charges
- Unemployment Percent

We proceeded to look more closely at the data per county, by visualizing selected attributes on a map. This approach let us identify possible relationships between data attributes that are difficult to notice otherwise. We looked at correlations between pairs of the above attributes. **Fig. 2 and Fig. 3** show such visualizations that indicate an otherwise hidden insight.

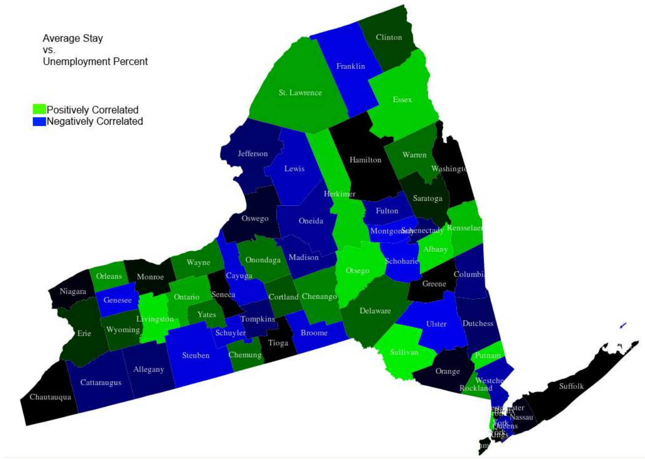


Figure 2 Correlation across time between average length of stay in the hospital vs. county unemployment rate. Within green counties, the two attributes co-vary. Within blue counties, the attributes vary in opposite directions.

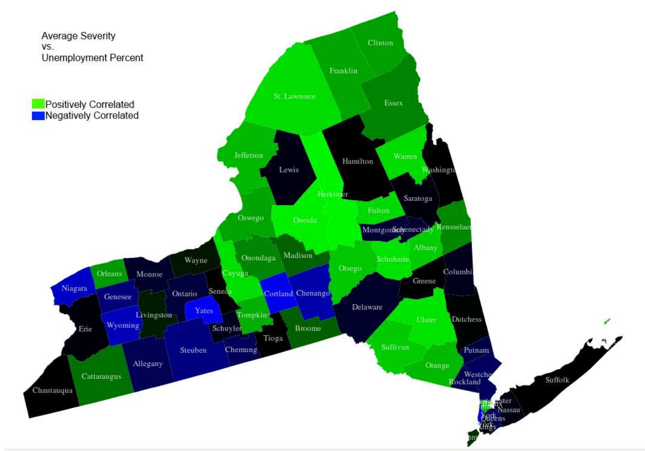


Figure 3 Correlation across time between average severity of inpatients' health condition vs. county unemployment rate. Within green counties, the two attributes co-vary. Within blue counties, the attributes vary in opposite directions.

At first glance, it seems that there is no consistent correlation between length of stay and unemployment (**Fig. 2**). However, in comparison to **Fig. 3**, there appears to be a somewhat stronger correlation between average severity and unemployment. This isn't the case statewide, but in particular counties this might serve as a valuable insight for policy officials to understand how their region is affected by employment trends.

Due to the limited range of years that we have comprehensive data for, we cannot objectively report that there is a stable correlation across time, but we believe that these trends are worth noting and possibly investigating further. We therefore choose to proceed by analyzing the data within each year separately, while verifying the statistical significance of our data.

For each year, we divide the counties into 2 groups by computing the median unemployment rate. For each attribute that we chose as a dependant variable, we calculate the average for the variable within that group (group 1 is less than the median, group 2 is higher). In some cases, the dependant variable was influenced by the independent variable – the unemployment rate being below or over the median value – with statistical significance ($p < .05$). **Table 2** shows the inspection of the correlation between unemployment rates and health attributes for each year.

2009		Low Unemployment	High Unemployment	P-value	Statistically significant?
Avg stay days per patient	n	28	28	0.0038	yes, 99% level
	Average	5.04	4.49		
	STDEV	0.58	0.76		
Avg. severity per patient	n	28	28	0.0347	yes, 95% level
	Average	1.95	1.89		
	STDEV	0.09	0.12		
Average cost per patient	n	28	28	0.0016	yes, 99% level
	Average	16,350.64	10,026.86		
	STDEV	9,077.17	3,565.54		

2010		Low Unemployment	High Unemployment	P-value	Statistically significant?
Avg stay days per patient	n	29	27	<0.0001	Yes, 99.5% level
	Average	4.93	4.54		
	STDEV	0.66	0.72		
Avg. severity per patient	n	29	27	0.0685	no but trending
	Average	1.99	1.95		
	STDEV	0.09	0.08		
Average cost per patient	n	29	27	0.0279	yes, 95% level
	Average	16,074.54	11,476.18		
	STDEV	9,371.97	5,363.05		

2011		Low Unemployment	High Unemployment	P-value	Statistically significant?
Avg stay days per patient	n	32	24	0.0026	yes, 99% level
	Average	5.00	4.41		
	STDEV	0.63	0.76		
Avg. severity per patient	n	32	24	0.2179	no
	Average	2.02	1.99		
	STDEV	0.09	0.10		
Average cost per patient	n	32	24	0.0028	yes, 99% level
	Average	17,385.54	11,146.77		
	STDEV	9,898.43	4,487.10		

2012		Low Unemployment	High Unemployment	P-value	Statistically significant?
Avg stay days per patient	n	26	26	0.0607	no but highly trending
	Average	4.88	4.49		
	STDEV	0.74	0.73		
Avg. severity per patient	n	26	26	0.5386	no
	Average	2.03	2.01		
	STDEV	0.08	0.11		
Average cost per patient	n	26	26	0.0795	no but highly trending
	Average	17,496.71	13,268.87		
	STDEV	10,776.53	5,371.67		

Table 2 Correlation between unemployment rates and health attributes by year.

We noticed, for example, that for length of stay metrics, the argument that higher unemployment is associated with shorter stays does hold for years 2009-2011 but not for 2012. For severity metrics, it seems that the existing data cannot support a clear conclusion. However, it is worth mentioning that the severity metric seems to be rising over time in both groups. For average cost of stay per inpatient, we noticed lower charges for the high unemployment group for years 2009-2011 and a similar trend for 2012.

It is interesting to point out that the difference between the number of days spent between the low unemployment group and the high unemployment group was approximately 0.5 days for all four years. The cost difference, however, for those with extended stays increased between $4000-$5000. This suggests an exponential increase in costs with the longer the stay. Data for more years would help confirm all of these trends.

VI. Conclusion

While the tabulated data generally shows lower severity for the higher unemployment group, the map visualization suggests that for many counties severity metrics rise with an increase in unemployment. We concluded that the time series data is worth monitoring further and this should be done on a per-county basis, as it seems reasonable that different counties exhibit different trends. Our current analysis could be used by state or insurance officials to target their attention appropriately. For example, our data suggests that when unemployment goes up, state Medicare aid could be allocated more heavily toward Albany, Franklin, and Oneida counties which experience a positive correlation between unemployment and negative health indicators, while such allocation might be underutilized in Essex, Lewis, and Cayuga counties which experienced a negative correlation.

The time series metric implies the more one stays in the hospital the stronger the indicator of declined health. On the other hand, looking at each year individually, the statistics suggest that the higher the unemployment the less that people are utilizing the services of a hospital. This in turn, can lead to even a greater decline in health. One conclusion is that unemployed individuals are more likely to not seek the medical attention they need because of the high costs associated with it. Naturally, this would contribute to greater risks to the health of those individuals, and ultimately more severe conditions if and when they finally do go to the hospital.

Our analysis has definitely raised important questions in regards to how one can define a correlation between employment and health. We believe that ultimately employment and health are related but possibly through more complex mechanisms than easily discerned through pairwise correlations.

VII. Future Work

Our research focused on regions within New York state only. The data that was used was taken from freely available datasets provided online by the Department of Labor and New York state. Given this focal point, and the need to be able to compare the health and labor data, we were limited to use data gathered from 2009-2012 only. This limitation will likely be easier to overcome once the NY open data site, which is a fairly young entity, will have aggregated data for several more years. A longer time series would help in verifying the suggested trends we have found in the current datasets.

Once more data is gathered, we intend to compare subsequent findings with occupational class data. We have gathered this information using similar methods as mentioned above. Several previous research attempts in this field were reliant on subjective reportings of individuals, when it came to their health conditions. We argue that as government authorities continue to collect and publish quantative data regarding unemployment, health, and other metrics, researchers can use tools such as those described here to analyze this data and reach more objective and large scale conclusions than before.

Furthermore, we have laid the ground work for a much more comprehensive study that would easily scale to all counties across the United States and potentially other geographic regions depending on the availability of data. The tools leverage infrastructure designed to scale easily with the data using the same MapReduce and Pig code that we implemented here. In that sense, this study presents a procedure by which finely grained data (such as inpatient data) is summarized with the help of current technology that is geared towards handling large datasets. The visualization of this data summary may illuminate certain correlation or trends that could guide the research in new and unexpected directions.

Another relationship work investigating would be that between employment and medication intake. Since our results suggest the higher the unemployment rate the less the utilization of hospitals, it would be interesting to see if people are finding alternatives to coping with their health needs. Incorporating other types of datasets from fields related to the health industry would shed more light on the true nature of the relationship between employment and health.

References

[1] T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.

[2] A. Gates. Programming Pig. O'Reilly Media Inc.,Sebastopol, CA, October 2011.

[3] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implemenation, 2004.

[4] S. Ghemawat, H. Gobioff, S. T. Leung. The Google File System. In Proceedings of the nineteenth ACM Symposium on Operating Systems Principles – SOSP '03, 2003.

[5] Bureau of Labor Statistics. http://data.bls.gov/timeseries/LNS14000000

[6] Bureau of Labor Statistics. http://data.bls.gov/timeseries/IIU00000000061100

[7] B. Hallerod and J. Justafsson. A longitudinal analysis of the realtionship between changes in socio-economic status and changes in health. In Social Science & Medicine, 2010, Vol. 72 No. 1 pp 116-23.

[8] G. Brant Morefield, David C. Ribar, and Christopher J. Ruhm. Occupational and Health Transitions. National Bureau of Economic Research, Cambridge, MA, February 2011.

[9] C. Ross and J.Mirowsky. Does Employment Affect Health? Journal of Health and Social Behavior 1995, Vol. 36 (September) pp 230-43.

[10] R. L. Jin, C. P. Shah, and T. J. Svoboda. The Impact of Unemployment on Health: A Review of the Evidence. Journal of Public Health Policy, 1993, Vol. 18 No. 3, pp 275-301.

[11] Bernstein AB, Hing E, Moss AJ, Allen KF, Siller AB, Tiggle RB. Health care in America: Trends in utilization. Hyattsville, Maryland: National Center for Health Statistics. 2003.

[12] N. Adler and K. Newman. Socioeconomic Disparities In Health: Pathways and Policies. Health Affairs, March, 2002, Vol. 21 No. 2 pp 60-76.

[13] Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/