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.
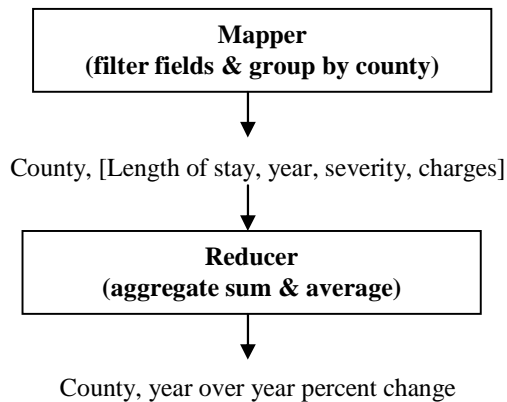
Jin, Shah, and Svoboda [10] came to the same conclusion as above in that an involuntary loss of one's job results in a decline in health. Their research, however, focuses specifically on individuals who died after having lost their jobs and analyzes the reasons that caused them to die. They found that 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.
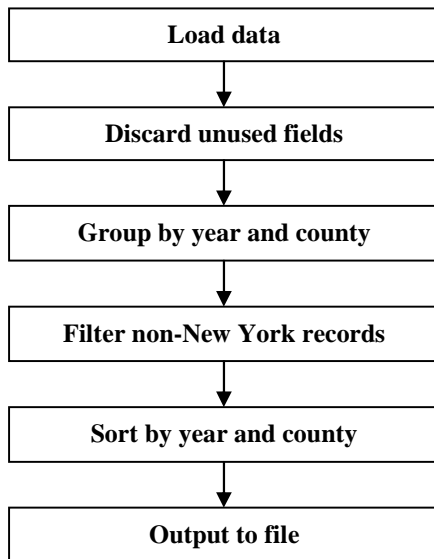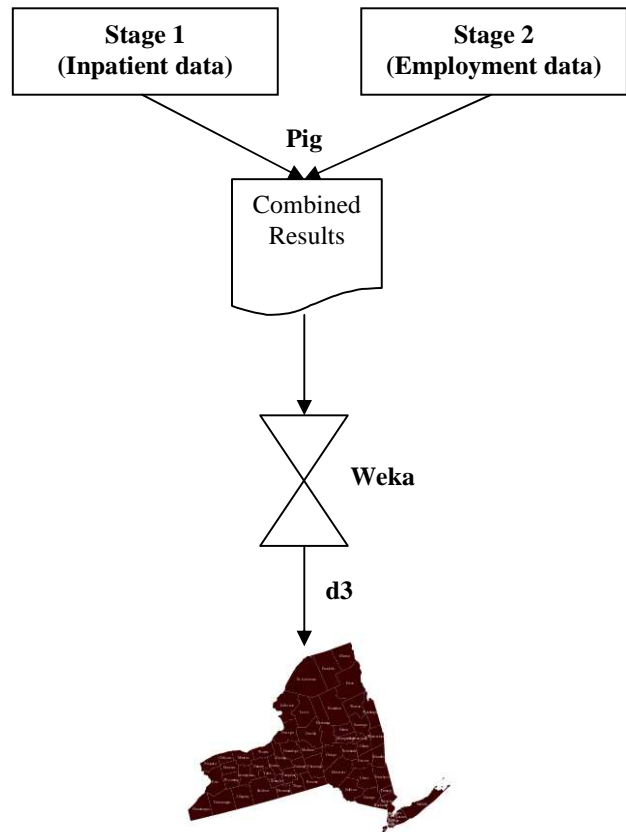
IV. DESIGN

In order to analyze our data sets we used a combination of Hadoop MapReduce and Pig in a three-stage processing the pipeline. The first stage takes inpatient hospital stay data and produces a summary report for each county (4GB → 500KB):

```
┌─────────────────────────────────────┐
│           Mapper                     │
│  (filter fields & group by county)   │
└─────────────────────────────────────┘
              │
              ▼
County, [Length of stay, year, severity, charges]
              │
              ▼
┌─────────────────────────────────────┐
│           Reducer                    │
│    (aggregate sum & average)         │
└─────────────────────────────────────┘
              │
              ▼
County, year over year percent change
```

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:

```
┌─────────────────────────────┐
│        Load data            │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│   Discard unused fields     │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│  Group by year and county   │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│ Filter non-New York records │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│   Sort by year and county   │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│        Output to file       │
└─────────────────────────────┘
```

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:

```
┌─────────────────────┐    ┌─────────────────────┐
│      Stage 1        │    │      Stage 2        │
│  (Inpatient data)   │    │  (Employment data)  │
└─────────────────────┘    └─────────────────────┘
              \      Pig      /
               \            /
              ┌──────────────┐
              │  Combined    │
              │  Results     │
              └──────────────┘
                    │
                    ▼
                   ╲╱        Weka
                   ╱╲
                    │  d3
                    ▼
          [map of New York State]
```

With the combined results, we use a machine learning tool called Weka (http://www.cs.waikato.ac.nz/ml/weka/) to discern correlation coefficients for functional dependencies in the data and then use those values to visualize the strength of the correlation on a map of New York.

V. RESULTS

VI. FUTURE WORK

VII. CONCLUSION

(Future… One or two paragraphs about the value/accuracy/goodness of your analytic.)

ACKNOWLEDGMENT

REFERENCES

(Add references for all of the papers/texts that you refer to in your paper. You will probably want to include the papers you read that were related to your project. You may have websites to reference, the Hadoop book, the MapReduce paper, the Pig Latin paper, etc. Some references are added below as an example.)

[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/