Big Data Predictive Analytics:

How Smart Communities Become Healthy Communities Through Big Data Informed Public Policy Formulation and Implementation

Michael W. Popejoy, Ph.D., M.B.A., M.P.H., M.H.S.A.

Correspondence

dr_popejoy@hotmail.com

Paper/Presentation

American Society for Public Administration Conference

March 2017

Atlanta, Georgia

ABSTRACT

Toward applications by policy and program planners working within the nexus linking public administration and public health in supporting sustainable, built, complex adaptive healthy communities; predictive big data analytics is a series of emerging datascience methods by which critical connections are made and strengthened through the sharing and data mining of massive quantities of data located across diverse public datasets. The days of studying and working in any one discipline or niche are quite likely over; and, the polymath mind and related analytical techniques rules the process of future public policy planning in solving social problems that impact the health and welfare of communities. Big data predictive analytical tools provides communities with the power to make better informed decisions rather than relying on guesswork based on inadequate data access and analysis. Prediction from the massive amount of existing data is empowering, but, not perfect. However, any real time driven prediction remains more powerful and satisfying than merely relying on a public agency's best guess. The concept of big data reflects the reality today that massive amounts of data are stored in a variety of depositories; and, are awaiting download and analysis by public community planners and others. Big data is characterized by volume, velocity, variety, and vector. Volume is easy to understand. There is so much data stored it is characterized as big or massive and it is now measured in zettabytes (bytes with 20 zeros following). Velocity is also a characteristic since big data moves through the network with lightning speed. Further complicating how big data is downloaded and analyzed is the almost infinite Variety characterizing the type of data and its storage format as it arrives and is stored in various massive databases under widely differing categories which makes data mining complicated. The characteristic of Vector illustrates the direction of a social problem or pandemic disease such as obesity; and, tells policy makers whether it is getting better, worse, or is remaining relatively stable over time and geographic distance. These characteristics have driven the development of new statistical analysis tools capable of downloading massive amounts of data (Volume) at high rates of speed as new data arrives (Velocity) thus providing real-time updates, and mines critical information regardless of how it is stored (Variety) while not drilling down to individual identities thus ensuring privacy; and indicating (Vectors) and concentrations that can be mapped using GIS technology. This revolution in data management and analysis has created a new kind of professional; the data scientist who combines knowledge of computers with statistics and knowledge of the environment of smart, healthy communities in the 21st century. The benefits are readily available; but, the trajectory and speed of progress are accelerating in the direction of improved prediction in complex adaptive systems where once politically driven agency agenda specific best guesses based on acceptable data were the norm with potentially unacceptable failure rates and frequent misuse of scarce community resources invested in a less than optimal direction.

Things on a small scale behave nothing like a large scale at all. That's what makes physics hard, and interesting.

Richard Feynman; Lectures on Physics, Lecture 2, September 29, 1961

Introduction

Public Administration and Public Health community planning collaboration is essential in developing highly effective public policy targets toward effective program goals for smart healthy communities; however, to ensure accuracy in targeting the right public problems, while maximizing efficient use of scarce community resources and minimizing potential policy failure is the effective utilization of the emerging datascience field of big data predictive analytics (Popejoy and Akukwe ed., 2013). (Siegel, E. 2016). Health informatics data analysis has been long used as the foundation for policy planning (Knepper in Popejoy and Akukwe ed., 2013).

Too often, community public policy planners target what they perceive is a public problem and use best guesses based on unreliable or incomplete or politically acceptable data and/or templates adapted from other community programs to solve the problem they believe exists. Unfortunately, they too often invest scarce community resources targeting a problem not clearly well understood resulting in a far less than satisfactory solution; or worse, no solution at all. The critical key point missing in the policy and program planning process is the lack of adequate data to sufficiently inform the policy decision and the direction of community action; and the formal program review and evaluation process after implementation (Rainey, H. G. 2014), (Bamberger, M., Rugh, J., and Mabry, L. 2006).

Today, with the technological advances in datascience, public policy planners have the opportunity to better study a social problem for emerging patterns in their community and understand its scope and magnitude; whether or not it is getting bigger, remaining relatively stable, or is shrinking over time. Further, the immense total amount of data available for analysis provides the foundation for a wider range of policy options with an increase expectancy in prediction accuracy in selecting the optimum public policy and program design for intervention thus implementing the selected best policy most effectively with less chance of wasting scarce community resources (Kraft, M. E. and Furlong, S. R. 2013). (Siegel, E. 2016).

Most communities store huge amounts of data relevant to a social problem that unfortunately are not accessed or analyzed appropriately leaving public policy planners working with their best guesses from inadequate data in designing programs rather than from a firmer foundation of facts that lie dormant in massive databases that should be available to planners. Even if the selected public policy intervention is not a total failure based on best guesswork, it certainly may unlikely be the optimum choice for the best use of resources in ways that promotes healthy communities (McInntosh, E. Clarke, P. M., Frew, E. J.; and, L. J. J. ed. 2010)., (Siegel, E. 2016).

In preparation for this article, I selected as a demonstration, the emerging community social problem of homeless autistic adults. I chose this unique population since as potential homeless adults; they are characteristically unique from many; but, not all other categories of homeless adults. Their patterns of behavior related to ASD make them unique as a subset of the total population of homeless adults often requiring much more than just safe, modern housing.

Autism Spectrum Disorders (ASD) is a neurodevelopmental disorder characterized by impairments in communication and social skills, and marked by restricted and repetitive behavior disorders. In recent years, there has been an increasing growth in the prevalence of ASD with 1 in 68 children in the U.S. alone are diagnosed with the disorder (Valdez, A., ed. 2015). The preliminary reports nationally should inform community planners that this is a growing problem with consequences for the healthy community requiring smart communities to plan and implement a social program for sheltering and protecting adults diagnosed with ASD.

Autistic adults often need far more individual personal support and assistance from the community than just safe modern housing space. Many adults with Autism Spectrum Disorders (ASD) require assisted living facilities staffed with 24/7 supervision to provide for their safety and security, activities of daily living (ADLs), and assistance with personal financial management since most receive or will receive Supplemental Security Income (SSI) while not understanding how to manage a bank account or pay their living expenses; and, they could become victims of predators who would take advantage of their inability to manage their own affairs (Zachor, D.A. and Merrick, J. ed. 2015). It is important to note that one does not grow out of autism and there is no cure (Prizant, 2015).

This reality to communities everywhere means that the prevalence of children, adolescents, and young adults now in the community with ASD will be the same prevalence of mature adults with aging parents as caregivers who will eventually need to transfer that role to the community. It is important to know now how many people in the population suffer from ASD from point of diagnosis at approximately 3 to 5 years old, through the k-12 school system and then onto Social Security support through SSI recipients so that Demographic Vector Analysis can be applied to determine direction, geographical distribution, and age level concentrations that the community is experiencing; and, how fast the prevalence may be growing as more children are diagnosed each year.

I began field research with an interview with the Senior Vice President (SVP) and Chief Medical Officer (CMO) of Broward Health located in south Florida. This health district is one of the largest in the U.S. However, the SVP/CMO did not know how many autistic children and adults were currently in the Broward Health population capture area; further, he stated that he could not disclose that information if he knew it due to HIPPA law considerations and Broward Health Confidentiality Agreements which all principals must sign.

Consequently, apparently, this major urban health system has no concept of how big this problem is in their capture area or if it is getting bigger as time passes; nor are they certain of the legalities of population data sharing if they had it. And, if they don't know anything about their population with ASD; then, it can logically be extended that they also know little or nothing about their patient population with obesity, diabetes, heart disease or cancers; and, other chronic diseases. This apparent lack of current population health knowledge confounds any effort toward healthy community policy and program planning. The advantage of this new datascience is that population data can be assessed and analyzed without drilling down to individual identities. But, public agency policy can become a barrier even when Federal law is not prohibitive.

I then turned to the president of the Autistic Society for Greater Orlando (Florida); and, she also did not know the numbers of autistic people currently living in the Central Florida region. She suggested that I contact the Center for Autism & Related Disabilities (CARD) at the University of Central Florida also in Orlando, Florida. A thorough search of their website showed no relevant published data on the demographics of ASD in Central Florida. However, via email, their staff did provide limited data for this study on the existing known demographics of ASD in the Central Florida area. It is important to note that their data, by their own admission, was both inadequate and incomplete since they were not able to capture the full range of population data that is now stored in medical records, school records, and community records for people diagnosed with ASD living in the Central Florida population. Consequently, the full scope and magnitude of this public health issue is unknown to policy planners.

Based on the initial results of my field research, I found no easily obtainable data that could be used to inform health policy and public program planners of the emerging scope of this community social problem. Why should Autism be considered for policy attention as a special population of potential homeless adults in the community? First of all, most autistic children and young adults live with family as their caregivers; however, as the family ages along its life journey, it becomes increasingly difficult and ultimately impossible for family caregivers to continue in the caregiver role. Eventually, autistic adults will become the responsibility of the community. But, there is more to the problem than just homelessness. As stated previously, many Autistic adults must receive continuous assisted care with their activities of daily living (ADLs).They will not only need a 24/7 supervised safe place to live; but, also they will need varying degrees of professional assisted living care from trained care givers employed by the special housing facilities (Valdez, A. ed. 2015).

The imperative now for community planners in smart healthy communities is to prepare for a future when the currently unknown numbers of this special population will require adapted housing facilities with adequate professional staffing. The questions now become; how big is this population today and where are they now in their life journey; and, what is their geographical distribution? How soon will the community need to implement special housing facilities and staffing? Of course, it is important to know whether this social problem is getting bigger, remaining relatively stable or getting smaller. And, finally, where is the data located now to inform future healthy community capacity development policy? Certainly, newly diagnosed children with ASD will have initial and updated medical records providing early data on the numbers living in the area or relocating to the area. Then, once children enter the k-12 school system, they will generate school records for the years they are in school. And, finally, adults with ASD will likely receive Supplemental Security Income (SSI) and government records will define the size of this group in the adult population.

The power of the new datascience of Big Data Predictive Analytics is in its ability to combine relevant data from many separate diverse databases regardless of data storage format; and, then begin the process of analyses to accumulate critical demographic information to inform future public policy initiatives. Certainly, a preliminary argument focuses on legal access to the data since health records are protected under HIPPA law and school records are protected under FERPA law. It is important to note that data mining processes do not need to drill down to the level of individual identities to provide important population public health information to public policy planners.

It is far more critical for planners to know the numbers of autistic people at each stage in the life journey from initial diagnosis often at about the age of three, then into the school system, and; finally into adulthood with data inclusive on the aging of the autistic adult's parents or other family caregivers. Now, communities have data that can inform policy initiatives about the scope and magnitude of the problem; and, what will need to be done to protect autistic adults from becoming homeless and ultimately potentially being absorbed into department of corrections facilities (jails or detention facilities) or long term care facilities (nursing homes); both of which are far more expensive care options per capita and less desirable for the healthy community than planned autism assisted care facilities.

Big Data Predictive Analytics

Big data is characterized by Volume, Velocity, Variety, and Vector. Volume is easy to understand. There is so much data stored it is characterized as big or massive and it is now measured in zettabytes (bytes with 20 zeros following). Velocity is also a characteristic since big data moves through the network with lightning speed updating daily if not hourly. Further complicating how big data is downloaded and analyzed is the almost infinite variety characterizing the type of data and its storage format as it arrives and is stored in various massive databases under widely differing categories which makes data mining complicated (Anderson, A. and Semmelroth, D. 2015). Vector is a characteristic often of interest to public health planners and epidemiologists; and, refers to the direction of a problem (or disease burden) of interest to public policy development and monitoring. It also refers to the dynamics of a problem such as is it getting worse, better, or remaining relatively stable over time and geographic space (Popejoy, 2017).

These characteristics have driven the development of new statistical analysis tools capable of downloading massive amounts of data (Volume) at high rates of speed as new data arrives (Velocity), providing real-time updates, and mines critical information regardless of how it is stored (Variety) while not drilling down to individual identities thus protecting individual privacy under HIPPA (medical records) and FERPA (school records) laws; and, providing policy planners with information on Vector or direction over geographic space and time. This is a transformative revolution in data management and analysis that has created a new kind of professional; the data scientist who combines an extensive knowledge of computers with advanced multivariate statistical analysis and knowledge of the built environment of smart, healthy communities in the 21st century. (Siegel, E. 2016).

The benefits are already available; and, the trajectory and speed of progress in datascience are accelerating in the direction of improved prediction power in complex adaptive systems (smart healthy communities) where once politically driven best guesses were the norm with potentially unacceptable failure rates; and, a too often frequent misuse of scarce community resources invested in a less than optimal solution to a public problem.

Once the magnitude and direction of any public problem is well understood with confidence in the accuracy of the relevant data, then program planning processes can begin the task of solving specific community problems based on what a community already has in place and what additional built capacity will need to be developed to meet the expected needs in the future. Once this information is well established, then the budgeting process can project capital costs and ongoing operational expenses related to implementing a public program solution for the identified problem. Finally, it is then time for the community to confidently act, and implement the policy, and fund the program. Of course, each public program should undergo rigorous independent continuous program review and evaluation to ensure the effectiveness and efficiency of meeting the program's identified goals and objectives (Bamberger, M, Rugh, J, and Mabry, L 2006).

Smart Healthy Communities

It is critical to building community capacity in smart healthy communities through the public planning process that all available relevant data regardless of where it is stored or how it is stored can be accessed and analyzed using the best of modern datascience methods. The most substantial benefit that can be gained from big data predictive analytics is that prediction power is amplified by the analysis of all relevant data rather than from limited access databases. Although prediction is not always perfect, it is far more satisfying when it is prevalently assured in the public planning process during policy formulation and implementation and program review and evaluation. The expected impact of prediction power is to support planners in planning the right public program to address an accurately defined emerging social problem and to effectively implement the optimum solution in the most cost-effective manner possible. Finally, the program review and evaluation process is always data energetic in its successful application in the public sector (Bamberger, M, Rugh, Jim, and Mabry, Linda, 2006). (Morse, S.W. 2014).

Conclusion

The accountability of politicians and government agencies to the publics they serve remains a critical factor in a democracy (Borins, S., 2008). Any innovations in governing such as improved datascience methods applied to social problems is a strong move in the direction of enhanced accountability as public policy program planners can cite with confidence the evidence provided by massive quantities of relevant data accurately downloaded and appropriately analyzed demonstrating a need for and supporting a particular policy decision and implementation of a specific social intervention program.

Programs also benefit from comprehensive cost-benefit analyses that can enhance public support for program proposals as well as provide for efficient management of both program implementation and operation. Further, the program review and evaluation process is strengthened by the availability and analysis of all relevant data dynamically as it changes over time potentially changing the scope of the social program as originally implemented (Thorogood, M and Coombes,Y. ed. 2010).

Costly potential policy failures are minimized while invested community resources are efficiently allocated where they will do the most public good. It can also be expected that the publics of a healthy community will be satisfied with public programs when the supporting rationales used to design, fund and implement any particular public program aimed at intervening in an emerging social problem in the community are well understood by employing readily available datasets and applying appropriate analytical tools. It is not new that public policy community planners have relied on available data in the past in their work of selecting and formulating public policy goals and implementing public programs (Bryson, J. M. 2006).What is emerging as a new approach today is how complex adaptive smart and healthy communities can access massive inclusive datasets across different sectors and apply advanced analytical techniques to the existing data which allows for a richer stream of information imperative to knowing what to do, when, and how much it will cost now and in the future with the increased accuracy of predictive power.

It is vitally important today for governments and public agencies to adopt early appropriate innovations in governing and any related processes that bring to focus public programs that meet defined social need; and, do so with minimal chance for failure; and, provide a maximization of benefits at an optimum cost to the taxpayers is also likely to enjoy wider acceptance with far less criticism from the taxpayer public.

The illustration developed in this study to demonstrate the emerging role of big data predictive analytics as it relates to homelessness for adults diagnosed with ASD; however, this level of analysis can also be applied to other areas of interest to complex adaptive smart and healthy communities such as the prevalence of obesity in the community including the numbers of obese students in the k-12 school system. This then can be related to ongoing community population health problems indicative of the current U.S. epidemic of diabetes; and, the incidence of serious health comorbidities expected in the diabetic population such as cardiovascular disease and cerebral vascular disease; and of course, the prevalence of different forms of cancer in the community's population (Shi, L and Singh, D. A., 2011). The massive amounts of data gathered on population health can be uploaded into a geographical information system (GIS) software program to develop geographical maps of community site concentrations and analysis of population health issues specific to geographical locations.

GIS data analysis can provide community level information on health disparities population distribution by mapping disease prevalence concentrations and specific locations. Related to the illustration of homeless ASD adults, GIS analysis can help community planners understand where they live and in what concentrations. This could be useful information in planning where to construct and operate assisted living facilities for ASD adults.

It would also be informative in studying other chronic diseases such as those related to the health challenges of obesity and lived environments to better understand the composition of neighborhoods in terms of the existence of isolated food deserts, the number of fast food establishments (unhealthy choices), the number of work-out gyms, and the number of farmers' markets (healthy choices) that are geographically located within pre-established distances from where the chronically ill population lives. This level of population concentration health information would inform the development of aggressive disease-specific public health intervention education programs supporting the goals of healthy communities.

What social programs can be designed by the healthy community to provide interventions in any particular identified community health problem? It is imperative when designing the most cost effective intervention programs that the scope of the problem is understood accurately; and, is the problem worsening, remaining relatively stable or shrinking? What are the potential long term impacts on the community of a continuing population health problem? What are the potential solutions; and, how much will they cost to implement and how successful can we expect them to be over time as demonstrated in program review and evaluation?

Successful implementation of the new datascience of big data predictive analytics has the potential to provide complex adaptive communities with substantial gains in understanding more clearly the details of its population health status and in meeting emerging social needs using the most effective and efficient allocation of community resources possible. What is required immediately is the opening of various critical dataspaces for download and analysis by qualified public policy planners interested in population health and demographics and not drilling down to individual identities which are not needed in population health science thus preserving individual privacy under current laws.

There are many questions that a community needs to ask about its population and its health; and, the data is available but it needs to be accessed and analyzed to confer information accurately to confidently inform today's decisions about future necessary programs to ensure improved population health. Communities are collecting and storing this data anyway in various separate databases such as medical records and school records, etc.; so, it is up to community planners to acquire and analyze the existing data to improve community planning effectiveness and operating efficiency.

The rapidly developing analytical tools related to big data predictive analytics can provide communities with the power of prediction more quickly, comprehensively, and more accurately than ever before. It is imperative that community planners adopt and implement big data predictive analytics to assess all community level projects to ensure positive outcomes in terms of cost-effective benefits to the community; and, contribute to healthy communities and healthy populations ultimately lowering the demand on health care services based on the reduced health care needs of a healthy population; and, reducing the operating costs of hospitals and other health care delivery systems in the community.

References

Anderson, A. and Semmelroth, D. Statistics for Big Data (2015). John Wiley&Sons. Hoboken, NJ

Bamberger, M, Rugh, Jim, and Mabry, Linda. Real World Evaluation: Working Under Budget, Time, Data, and Political Constraints. (2006). Sage Publications. Thousand Oaks, CA.

Borins, Sandiford, ed. Innovations in Government: Research, Recognition, and Replication. (2008). Brookings Institution Press, Washington, D.C.

Bryson, John M. Strategic Planning For Public and Nonprofit Organizations: A Guide to Strengthening and Sustaining Organizational Achievement; 3rd ed. (2004). Jossey-Bass Publishers, San Francisco, CA

Estes, Carroll, L., Susan A. Chapman, Catherine Dodd, Brooke Hollister, and Charlene Harrington. Health Policy: Crisis and Reform 6th ed. (2013). Jones &Bartlett Learning. Sudbury, MA

Fallon, L.F. and Zgodzinski. Essentials of Public Health Management 3rd ed. (2012). Jones and Bartlett Learning. Sudbury, MA

Fitzgerald, M. The Mind of the Artist: Attention Deficit Hyperactivity Disorder, Autism, Asperger Syndrome, and Depression. (2015).

Knepper, Hillary; Exploring Health Informatics: A 21st Century Public Health Tool in Popejoy, M.W. and Akukwe, C. ed. Global Public Health Policy: Public Health in the 21st Century. (2013). Nova Science Publishers New York.

Kraft, Michael E. and Furlong, Scott R. Public Policy: Politics, Analysis, and Alternatives, 4th 3ed. (2013). Sage CQ Press, Los Angeles, CA

McInntosh, Emma, Clarke, Phillip M., Frew, Emma J.; and, Louviere Jordan, J. (ed). Applied Cost-Benefit Analysis in Health Care. (2010). Oxford Press, London, UK

Morse, Suzanne W. Smart Communities: How Citizens and Local Leaders Can Use Strategic Thinking to Build a Brighter Future.2nd ed. (2014). Jossey-Bass; San Francisco, CA.

Popejoy, M.W. Conference Paper Presentation. American Society for Public Administration national conference in Atlanta, GA March 21 (2017).

Popejoy, M.W. and Akukwe, C. ed. Global Public Health Policy: Public Health in the 21st Century. (2013). Nova Science Publishers. New York.

Preskill, Hallie and Russ-Eft, Darlene. Building Evaluation Capacity. (2005). Sage Publications, Thousand Oaks, CA

Prizant, B.M. (2015). Uniquely Human: A Different Way of Seeing Autism. Simon and Schuster, New York.

Thorogood, Margaret and Coombes, Yolande (ed). Evaluating Health Promotion: Practice and Methods, 3rd ed. (2010). Oxford Press, London, UK

Rainey, H. G. Understanding and Managing Public Organizations 5th ed. (2014). Josey-Bass Publishers. San Francisco, CA

Shi, Leiyu and Douglas A. Singh. (2011) The Nation's Health, 8th ed. Jones & Bartlett Learning. Sudbury, MA

Siegel, Eric. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. (2016). John Wiley & Sons., Hoboken, New Jersey.

Valdez, A. (ed.). Autism Spectrum Disorders: Early Signs, Intervention Options, and Family Impact. 2015.

Zachor, D.A. and Merrick, J. (ed.). Understanding Autism Spectrum Disorder: Current Research Aspects. (2013).