Open Research Issues and Emerging Research Directions in Data Quality for Public Health

ARUN SUNDARARAMAN, Accenture
SRINIVASAN VALADY RAMANATHAN, Accenture

The topic of Data Quality (DQ) in the field of Information Management has been extensively researched. Within this field, DQ relevant for Public Health Administration has been relatively less explored. DQ research has traditionally focused on a select set of DQ factors (e.g. timeliness or accuracy etc.), however, advances in the field of Information Management triggered by emerging technology trends such as Big Data has warranted that DQ research be approached with fresh outlook to consider DQ factors relevant to these new technology trends. Moreover, data for effective intervention in the administration of Public Health is dependent on a tighter integration of data from multiple data sets within the Healthcare ecosystem. Thus, in Public Health it is not necessarily true that individual sets of data with good DQ does necessarily mean a comprehensive data set with good DQ, due to integration issues. DQ Research needs to be revisited to address these issues. This Paper analyses the various DQ issues on the above lines, outlines the research issues and recommends research directions in DQ for Public Health to address the above problems.

Additional Key Words and Phrases: Public health, data quality, big data quality, information quality, data sources, research issues, data use, data silos, non-healthcare data, data quality dimensions, evaluation, assessment, personalization, information systems.

1. INTRODUCTION

Public health is "the science and art of preventing disease, prolonging life, and promoting physical health and efficiency through organized community efforts" [Winslow, C.E. 1920]. The ultimate goal of public health is to improve health at the population level, and this is achieved through the collective mechanisms and actions of public health authorities within the government context [Winslow, C.E. 1920 & Walker, R. 2008]. Three functions of public health agencies have been defined: assessment of health status and health needs, policy development to serve the public interest, and assurance that necessary services are provided [Walker, R. 2008]. Since data, information and knowledge underpin these three functions, public health is inherently a data-intensive domain [Institute of Medicine. 1988 & Andresen, E.; Bouldin, E.D. 2010]. High quality data are the prerequisite for better information, better decision-making and better population health [World Health Organization. 2008].

Public health data are generated from public health practice, with data sources being population-based and institution-based [World Health Organization. 2008, Australian Institute of Health and Welfare (AIHW). 2005]. Population-based data are collected through census, civil registrations, and population surveys. Institution-based data are obtained from individual health records and administrative records of health institutions [World Health Organization. 2008]. The levels and distribution of the determinants of health are measured in terms of biomedical, behavioral, socioeconomic and environmental risk factors. Data on public health interventions include prevention and health promotion activities, while those on system resources encompass material, funding, workforce, and other information [Australian Institute of Health and Welfare (AIHW). 2005].

With the huge volume of generated data, the fast velocity of arriving data, and the large variety of heterogeneous data, the quality of data is far from perfect. It has been estimated that erroneous data costs US businesses 600 billion dollars annually. There are tremendous opportunities for Big Data analytics to impact the productivity and quality of the health care sector. Researchers, analytics-vendors and software developers who create and deploy sophisticated infrastructure and organization-specific intelligence tools for health care decision and policy-making assume that the

data quality is assured by the data-supplying organization. Unlike sectors such as manufacturing, where market expectations drive the quality of a product, the market forces in the Big Data industry have not imposed a similar standard on data quality and this is particularly true for health care data [Sreenivas R. Sukumar et al. 2015].

Negative effects of poor data quality, however, have often been reported. For example, Australian researchers reported coding errors due to poor quality documentations in the clinical information systems. These errors had consequently led to inaccurate hospital performance measurement, inappropriate allocation of health funding, and failure in public health surveillance [Cheng, P et al. 2009].

Data quality in public health has different definitions from different perspectives. These include: "fit for use in the context of data users" [Canadian Institute for Health Information 2009], "timely and reliable data essential for public health core functions at all levels of government" [Institute of Medicine 2003], and "accurate, reliable, valid, and trusted data in integrated public health informatics networks" [Snee, N.L et al. 2004]. Whether the specific data quality requirements are met is usually measured along a certain number of data quality dimensions. A dimension of data quality represents or reflects an aspect or construct of data quality [Wang, R.Y et al. 1996].

This paper provides a consolidation of key issues and research interests of data quality in public health. The paper attempts to explain the need to change the data quality dimensions in the light of big data developments and in the context of data requirements for Public Health decision making. It also explains how data silos in public health affect the longitudinal view of entities thereby affecting insights. The data sources are explained in section 2 of the paper, the key issues in different data sources and its methods are listed in section 3 of the paper. The final section concludes by highlighting the key research issues and suggested future research directions of data quality in public health.

2. DATA SOURCES

Organizational data are increasingly distributed in heterogeneous resources and represented with different formats, even if they refer to the same organizational entities, ranging from almost unstructured, e.g. in file systems, document repositories and on the web, to highly structured, e.g. in database management systems. In the literature, data sources are classified depending on the level of structure that characterizes them. According to a widely accepted yet quite informal convention, the literature distinguishes between data sources in terms of

1. Structured data, if their formal schema (i.e., formats, types, constraints, relationships) is defined and explicit.
2. Semi structured data, (sometimes also called "self-describing" data) if data are something in between raw data and strictly typed data, i.e., when they have some structure, but it is not as rigidly structured as in databases. They are usually represented with XML markup language.
3. Unstructured data, if they are but sequences of symbols (at least at a human reader), e.g. a string in natural language, where no specific structure, type domains and formal constraints are defined [Batini Carlo et al. 2011].

The major data sources of public health data and its description is given in Table I

Table I. Data sources, description and use

Data sources	Description
Electronic Medical Records	An electronic health record (EHR), or electronic medical record (EMR), refers to the systematized collection of patient and population electronically-stored health information in a digital format. EMR includes demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information. EMR will include structured data and unstructured data like clinical, bed side notes etc.
Insurance claims	Claims data include medication and care providers bill claims
Clinical data like home care sensors etc.	The clinical data generated from bed side sensors, home care sensors and IOT devices to dynamically monitor the health of the population
Genomic data	Genomic data refers to the genome and DNA data of an organism. They are used in bioinformatics for collecting, storing and processing the genomes of living things. Genomic data generally require a large amount of storage and purpose-built software to analyze.
Hospital Utilization	The usage rate of a particular health care facility;a group of statistics referring to a population's use of hospital services
Pharmaceutical	Pharmaceutical events like population usage details
Social and Behavioral data	Lifestyle habits, Relationships and social circles, GPS tracking and IOT data on health parameter form the basis of social and behavior data.
Climate, Sanitation, Water resources, Environmental data	Epidemics would require climate, sanitation and environmental data to monitor and model them
Scientific (PubMed, SCOPUS, JAMA, Clinical trials, patents etc.)	The physicians and practitioners require access to current ongoing scientific data to help them in accurate diagnosis and treatment
News & Social Media	Epidemic, symptoms, precautions and step taken globally are now analyzed through social media across geographies. For examples there are tools to predict an epidemic from twitter feeds
Civic Registration, Census data	The Civic registration and census data complement the clinical data in entity resolution and in identifying population clusters etc.
Interventions data	These data constitute the interventions program planned and its corresponding outcomes

The method of data generation vary between manual, automated or machine generated. It also significantly varies in the volume generated, velocity of generation and variety. The data quality of manual data generated is better handled at the origination point. A data producer informed of the data use will generate high quality data fully aware of the lifecycle of the health data. Data quality assessment and improvement on consolidation will yield lesser value than the former.

3. KEY ISSUES AND DIMENSIONS

As information technology becomes an integral part of daily life, increasingly, people understand the world around them by turning to digital sources as opposed to directly interacting with objects in the physical world. This has resulted in explosion of data sources necessitating adoption of Big Data technologies and newer data sources to be considered in public health domain. With the data landscape thus changing, the scope of Information Quality (IQ) research is due to expand dramatically as the challenge becomes to capture the wealth and nuances of human experience.

In light of these developments, there exists immense potential for new research areas in study of IQ for Public Health: (1) expansion of the scope of traditional IQ dimensions (2) digital to physical mapping challenge, and (3) the increased need to manage content authenticity. UDI generates many novel questions and opportunities for the IQ research community [Lukyanenko et al. 2016].

An assessment of data quality in health care has to consider: (1) problems arising from errors and inaccuracies in the data itself; (2) the source(s) and the pedigree of the data; and (3) how the underlying purpose of data collection impact the analytic processing and knowledge expected to be derived [Sreenivas R. Sukumar et al. 2015].

There are some fundamental issues inherent to the above presented data sources and additional systemic issues in public health. Data quality dimensions or attributes provides a definition of the quality of the underlying data and a framework for the assessment of quality. In this context, quality issues along with the attributes relevant for measurement in public health are given in the below sections.

3.1 Key Concerns of Data Quality in Public Health

Public health data resides in different independent systems either in tangible or intangible form. The use of complete spectrum of the data is expected to provide enhanced quality of data and care. However, there are issues which hamper enhanced DQ.

Electronic Medical Records, Clinical, Claims data. A consideration of errors and inaccuracies in the data would include data entry errors, missing data field or entire records, and errors arising from transformations in the extracting and transforming process for analytics [Sreenivas R. Sukumar et al. 2015]. This frequently arises in manual data entry points like EMR, claims etc. There are about 8000 procedure codes and 16,000 diagnostic codes. Manual errors are naturally possible. The quality of health records is dependent on the health information provided by the authorized staff. Detailed and specific documentation is often considered a lesser priority. If the correct attributes and precise representation of data are not provided, health record linkage would become unreliable and of little value. Improper controls in data would lead to missing data, incorrect entry, duplication due to de-centralized data in separate care locations, redundant data in the same location, if the validation controls are not in place. Data comes in from a variety of sources, which means a variety of business rules, and improper understanding of these sources, data and granularity, would lead to anomalies and misses and sometimes incorrect representation during integration. In addition, data veracity issues can arise from attempts to preserve privacy and following Health Insurance Portability and Accountability Act (HIPPA) guidelines where de-identifying/disguising data is intentional. Also, data veracity is a function of how many sources contributed to the data collection process and their similarities and differences [Sreenivas R. Sukumar et al. 2015].

Big Data. The issue of DQ has been existent ever since data existed and the impact of poor DQ on decision making and organizational outcomes has always remained an area of concern and interest. With the advent of Big Data and ability to pool in large data sets (high volumes, rapid velocity and myriad variety) the nature DQ and the need for newer dimensions of DQ research are also changing rapidly – e.g. data has traversed from structured to unstructured patterns and the volume of data has been growing in leaps and bounds. In the public health domain, backed by

modern IT world and economies, for effective strategic and policy decision making, Public Health Authorities depend on a wide variety of external channels of information (e.g. Patient Health Records or privately held data such as wellness and fitness information), in addition to disparate internal source systems. These developments (in nature and sources of data required for policy decisions) have increased the complexity of DQ definition and its measurement techniques. In order for the research community to adequately respond to the changing landscape of DQ challenges, a unified framework for DQ research is needed [Sadiq et. al. 2011]. The structure of data changes from the traditional structured data for big data. This pushes the boundaries of data quality definitions as well. Completeness, an important attribute in structured data may not have the same significance in the realm of big data. Consistency and correctness may not have as high a significance as in structured data forms. In contrast, Redundancy may have a higher weightage.

Data Silos. Study of DQ is highly inter-disciplinary in nature. There are multiple factors that add to the complexity in measuring DQ such as evolution of Information Systems or forms of work in organizations or the environment in which we live. Each of these factors are highly relevant in the context of how these influence the definition of DQ in Public Health in today's world. For example, public health organizations are increasingly adopting Big Data technologies as part of their Information Systems Architecture. public health organizations are expanding their nature of work from intervention to use of information for prediction of disease break out and carry out preventive interventions or policy formulation aimed at measurable health outcomes for the population at large e.g. MDP goals; today's world is inter-connected from hand held devices to advanced computer networks. This has led to a scenario where data required for end-to-end analysis for public health research is contained in different islands of information as depicted below: [Sreenivas R. Sukumar et al. 2015]

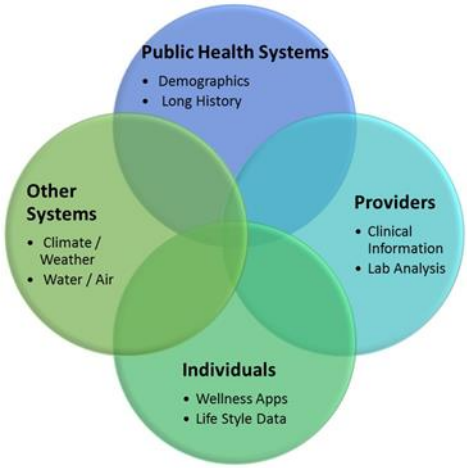

Fig. 1. Public Health Data.

Data in public health, even of high quality in individual pockets doesn't necessarily add value without longitudinal view of the entities. Entity resolution is one of the key issues in public health. Lifestyle and behavioural data even though exists is not

available for use completely in a timely fashion to deliver care. The integration of all the necessary data points for an entity in a timely fashion is often a challenge. Often datasets integrated from multiple sources are characterized by different levels of data quality. This can result in degradation of the overall quality of the integrated data to the lowest level of data quality of the contributing sources [Sreenivas R. Sukumar et al. 2015].

3.2 Attributes of Data Quality

As per the existing definitions of DQ, DQ attributes (interchangeably referred as DQ factors or DQ dimensions) are largely derived based on DQ management principles that seek to ensure syntactic correctness and semantic correctness [Wang et al.1995]. This traditional approach has led to several DQ frameworks classifying and categorizing DQ attributes under different logical groups. In their published work Knight and Burn, 2005 summarize and compare DQ frameworks and in the process present different sets of DQ factors considered in different frameworks.

As per Batini et. al., 2009 the DQ literature provides a thorough classification of data quality dimensions. However, there are a number of discrepancies in the definition of most dimensions due to the contextual nature of quality.

Besides identified appropriate DQ attributes (individual elements), it is also critical to identify the dependencies between each of these individual attributes. Amicis and Barone, 2006 argue that the analysis of the dependencies among DQ dimensions is extremely important in the area of information quality in order to improve the quality level of a data set, reconstruct the cause-effect patterns on data quality dimensions, select the most important improvement activities, and more generally increase knowledge on dimensions and their relationship.
A recent work (Information Quality Strategy - an Empirical Investigation of the Relationship between Information Quality Improvements and Organizational Outcomes. Ph.D. Thesis [Sundararaman, Arun. 2013]) establishes the need for 2 broad characteristics in the study of DQ i.e. context and comprehensiveness; scope exists for focused research in conducting empirical study in identifying the Decision categories in Public Health and definition / measurement of DQ from unstructured data in public health policy decision making process. The dimensions of context and comprehensiveness adds significant value for the measurement in case of data silos.

In a related study of methods and dimensions of electronic health record data quality assessment, five substantively different dimensions of data quality were derived from the literature. The dimensions are defined below.

Completeness. Is a truth about a patient present in the EHR? Terms in Literature: Accessibility, Accuracy, Availability, Missingness, Omission, Presence, Quality, Rate of recording, Sensitivity, Validity.

Correctness. Is an element that is present in the EHR true? Terms in Literature: Accuracy, Corrections made, Errors, Misleading, Positive predictive value, Quality, Validity.

Concordance. Is there agreement between elements in the EHR, or between the EHR and another data source? Terms in Literature: Agreement, Consistency, Reliability, Variation.

Plausibility. Does an element in the EHR makes sense in light of other knowledge about what that element is measuring? Terms in Literature: Accuracy, Believability, Trustworthiness, Validity.

Currency. Is an element in the EHR a relevant representation of the patient state at a given point in time? Terms in Literature: Recency, Timeliness [Nicole Gray Weiskopf et al. 2013]

In a popular work (A Review of Data Quality Assessment Methods for Public Health Information Systems) on Public health data quality across the literature by Hong Chen et al. 2014 completeness, accuracy, relevance, consistency, timeliness, accessibility and security are included as key dimensions of data quality.

In another study (Big Data Quality Challenges in the Context of Business Analytics) by Mirva Toivonen 2015, some of the differential quality attributes of are listed. The same is summarized below

Schema completeness describes the degree to which entities and attributes are not missing from the schema.

Minimality refers to avoiding undesired redundancy during the source integration Process.

Traceability refers to the fact that all kinds of requirements and decisions of users,

Designers, administrators and managers should be traceable in the data warehouse Schema.

Schema interpretability refers to how well the data model is explained, which makes querying easier.

Analyzability refers to the validation of each process and its ability to handle errors and self-report when errors occur.

Transactional availability refers to the time when information is not available due to update operations.

Scalability is represented by the amount of data being queried and the number of concurrent users simultaneously running the queries.

Credibility describes the trustability and believability of the source that provided the information.

Data interpretability measures the descriptions of data e.g. table description for relational databases, primary and foreign keys, aliases, defaults, domains, explanation of coded values, etc.

Accessibility refers to the extent to which data is available, or easily and quickly retrievable.

Viscosity refers to measuring the resistance to flow in the volume of data.

Virality refers to the quality of how quickly data is shared in a people-to-people (peer) network.

4. CONCLUSION AND RESEARCH TRENDS

In the finer cross section of DQ in public health domain considering big data based technology solutions, it is expected that hybrid approaches that define and measure DQ would emerge and this would remain to be an important research direction. Helfert and Foley, 2009 argue that the frameworks, quality indicators and measurement systems still have limitations. Despite the large amount of literature, agreed criteria lists or measurement approaches are still missing and that as of today no widely accepted IQ framework with generic, generally applicable measurements is available [Sundararaman, Arun 2013]. Woodall et. al., 2010 argue that no individual

existing technique for assessing DQ is wholly suitable to assess DQ for all types of requirements due to the varying nature of requirements over time and organizational needs; this finding assumes greater significance and relevant in today's scenario and the research directions in the new few years towards study of DQ for Public Health Administration. It was emphasized that DQ requirements may be different for every organization and even the same organization over time. It is in the context of this study that focused DQ research needs to evolve to meet the specific requirements of DQ requirements of Public Health Organizations. The current state of DQ research is such that while some of the DQ assessment techniques are geared towards specific application areas and are often not suitable in different applications, other techniques are more general and therefore do not always meet specific requirements [Woodall et al.. 2010].

A key summary of interest areas and research trends are summarized below

- Based on the principle of context of DQ, scope exists for research on multiple aspects around DQ dimensions
    - Identifying a revised set of DQ factors relevant for Public Health domain (e.g. DQ for unstructured data from clinical notes)
    - Identify a revised set of DQ factors relevant to Big Data sources (e.g. believability or reliability revisited in the context of Big Data sources)
    - Identify weightage factors for these revised DQ factors as they are applicable in a Public Health policy decision making context (e.g. importance of timeliness of data vs. completeness or significance of reliability over consistency etc.)
    - Identify weightage factors for these revised DQ factors as they are applicable in a Big Data platform (e.g. criticality of completeness Vs. redundancy or interpretability as it relates to unstructured data Vs. conciseness of unstructured data
    - Hybrid research approaches with a combination of above problem statements.
- Significant scope exists to study DQ Measurement models based on "inherited" DQ from underlying source systems. This proposed area of research would address specifically the issue related to data silos and their impact on overall DQ when information of derived by integrating these individual data sets.
- An extension to this proposed research would be an area research that would focus on the designing and development DQ forecasting models; DQ measured in individual data sets would be in the inputs for this model; however, the final resultant DQ score is not necessarily a sigma of constituent DQ scores, because of the integration issues discussed earlier in the paper. A research focused on developing capabilities to forecast a DQ given a base DQ of underlying data sets (e.g. DQ of clinical data set+ DQ of claims data set+ DQ of demographics data set playing to the overall DQ of an integrated public health data set) would be of immense value to the academics, Industry and custodians of IT Systems.
- Another proposed area of research would involve development of statistical models to be able to study the impact of non-health data (e.g. sanitation or climatic data) on DQ for public health policy decisions and decisions related to interventions in public health administration.
- Comprehensive view of the entities in a timely fashion and definition of measurement for combined DQ of public health data integrated from

different events to effect an action is a key interest area. Definition of data quality dimensions and methods to achieve the comprehension across structured to unstructured is one of future research focus areas.

- With changing landscape of data structures with big data applications and digitization more accessible from multitude of devices, redefining the boundaries and significance of DQ dimensions posts significant challenge in the future.
- The use and data quality measurement of non-health care which gives a comprehensive judgement needs further research.
- The applicability of DQ to further analytics or mining interests and outcome based quality measurement will add a future step.
- Further exploration of interdependence of quality dimensions will be beneficial.

In summary, several issues continue to remain open in the study of DQ as it pertains to Public Health and application of data from Big Data sources for decision making in Public Health. Study of these 2 subjects in a combined manner assumes significance based on the concepts of context and comprehensiveness as discussed in published literature. The research community has a huge potential to carry out focused research in various directions (listed above) with an objective to address these open issues.

REFERENCES

Winslow, C.E. 1920. The untilled fields of public health. Science 1920 51, 23–33

Walker, R. 2008. Health information and public health. Health Inf. Manag. J. 2008, 37, 4–5.

Institute of Medicine. 1988. *The Future of Public Health*; National Academies Press: Washington, DC, USA, 1988.

Andresen, E.; Bouldin, E.D. 2010. *Public Health Foundations: Concepts and Practices*; Jossey-Bass: Hoboken, NJ, USA, 2010.

World Health Organization. 2008. *Framework and Standards for Country Health Information Systems*; World Health Organization: Geneva, Switzerland, 2008.

Australian Institute of Health and Welfare (AIHW). 2005. *National Public Health Information Working Group.* National Public Health Information Plan 2005 .AIHW: Canberra, Australia, 2005.

Sadiq, S., Indulska, M. and Jayawardene, V. 2011. Research and industry synergies in data quality management, proceedings of the *16th International Conference on Information Quality*. 2011 pp. 314-326.

Nicole Gray Weiskopf and Chunhua Weng. 2013. *Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research*. 2012. J Am Med Inform Assoc 2013; 20: 144–151. Doi: 10.1136/amiajnl-2011-000681

Lukyanenko, Roman. 2016. *Information Quality Research Challenge: Information Quality in the Age of Ubiquitous Digital Intermediation*. Journal of Data and Information Quality, Vol. 7, 1-2, 3 pages. DOI=http://dx.doi.org/10.1145/2856038

Sreenivas R. Sukumar, Natarajan Ramachandran and Regina K. Ferrell. 2015. *Quality of Big Data in health care.* International Journal of Health Care Quality Assurance DOI: 10.1108/IJHCQA-07-2014-0080

Cheng, P.; Gilchrist, A.; Robinson, K.M.; Paul, L. 2009. *The risk and consequences of clinical miscoding due to inadequate medical documentation: A case study of the impact on health services funding*. Health Inf. Manag. J. 2009, 38, 35–46.

Canadian Institute for Health Information. 2009. *The CIHI Data Quality Framework*; CIHI: Ottawa, ON, Canada, 2009.

Institute of Medicine. 2003. *The Future of the Public's Health in the 21st Century*; The National Academies Press: Washington, DC, USA 2003.

Snee, N.L.; McCormick, K.A. 2004. *The case for integrating public health informatics networks*. *IEEE Eng. Med. Biol. Mag.* 2004, *23*, 81–88.

Wang, R.Y.; Strong. 1996. *D.M. Beyond accuracy: What data quality means to data consumers*. *J. Manag. Inf. Syst.* 1996, *12*, 5–33.

Batini Carlo, Barone Daniele, Cabitza Federico and Grega Simone. 2011. *A Data Quality Methodology for Heterogeneous Data*. International Journal of Database Management Systems (IJDMS), Vol.3, No.1, February 2011

Batini,C., Cappiello,C., Francalanci,C. and Maurino,A .2009*, —Methodologies for Data Quality Assessment and Improvement‖,* ACM Computing Surveys, Vol. 41, No. 3, Article 16.

Chung, W., Craig,F. and Wang, R. Y. 2005*, —Redefining the Scope and Focus of Information Quality Work‖, Information Quality*. New York: M.E.Sharpe, pp. 230-248.

Helfert, M. and Foley, O. 2009*, —A Context Aware Information Quality Framework‖,* Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, pp. 187-193.

Knight,S and Burn,J.2005. *Developing a framework for assessing Information Quality on the World Wide Web‖*. Informing Science Journal., Vol8, pp.159-172.

Amicis, Fabrizio De and Barone, D. 2006*, —An Analytical framework to analyze dependencies among data quality dimensions‖,* proceedings of the 11th International Conference on Information Quality (ICIQ), pp.369–383.

Mirva Toivonen. 2014. *Big Data Quality Challenges in the Context of Business Analytics*. PhD. Thesis

Hong Chen, David Hailey, Ning Wang and Ping Yu. 2014. *A Review of Data Quality Assessment Methods for Public Health Information Systems*. Int. J. Environ. Res. Public Health 2014, 11, 5170-5207; doi: 10.3390/ijerph110505170