Review Article/บทฟื้นฟูวิชาการ

Big Data Analytics in Healthcare

Ratchainant Thammasudjarit[1], Anuchate Pattanateepon[1], Oraluck Pattanaprateep[1]

[1] *Section for Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand*

Abstract

Nowadays, healthcare data, has been recorded more in electronic formats, both structured and unstructured forms. To utilize huge and complex data, it needs skills in big data analytics. However, most medical staffs are unfamiliar with big data. This article is therefore aimed to introduce type of data analytics, machine learning, statistics, and data mining which contribute to healthcare big data analysis. The article should be useful for healthcare professionals who are new to this field.

Keywords: Big data analytic, Healthcare, Machine learning, Data mining

Corresponding Author: Oraluck Pattanaprateep

Section for Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, 270 Rama VI Road, Ratchathewi, Bangkok 10400, Thailand.

Telephone: +66 2201 1820 Fax: +66 2201 1284 E-mail: oraluck.pat@mahidol.ac.th

Introduction

Healthcare and other business domains are generating tremendous amounts of data every day (eg, email communications, financial transactions, social media). Likewise, in health care, medical records, prescriptions, laboratory, medical databases, health care system are electronically stored, rapidly growth and resulted in availability in large amounts of healthcare data. Such extremely large amounts of data are named as *big data* which characterize by 4 Vs; volume refers to enormous amount of data, variety refers to heterogeneous data sources, velocity refers to the speed of changing over time, and veracity refers to the biases, noises, and abnormalities in data. Utilizing the big data is very challenge to healthcare researchers to make a benefit from using the big data for improving quality of healthcare and also healthcare practice. The article is therefore aims to introduce type of data analytics and analytic methods that have been used to deal with big data including machine learning, statistical methods, and data mining for general readers who are unfamiliar with such fields using healthcare scenarios.

Types of Data Analytics

There are four types of data analytics[3] as shown in Figure 1. First, descriptive analytics is mainly aimed to only describe characteristics of data in term of what, whom, when and how. Diagnostic analytics is aimed to assess association between risk or prognostic factors (known as features in big data) and clinical or health care outcomes. Predictive analytics is aimed to predict how more or less likely disease will occur given some characteristics or conditions. Finally, the prescriptive analytics is aimed at finding what intervention or treatment programs should be applied/implemented given patients' conditions.

We use a scenario of heart failure patients who admitted in a hospital due to several conditions for demonstrating how each type of data analytics is applied. First, descriptive analytic is applied to explore historical data of these patients to find out what occurrence, to whom, and when. The key value of this analytics is to understand the current condition of heart failure by exploring data. Technically, exploring such data could be done with simple queries and data visualization, which was less complex but could lead to identify what occurrence, occurred to whom, with what condition. For instance, exploring historical data of these patients found that number of admission from heart failure increased during winter. In addition, asthma was a common comorbidity of these heart failure patients. This might be explained that lower temperature might stimulate airway hyper-responsiveness and thus induced heart failure.

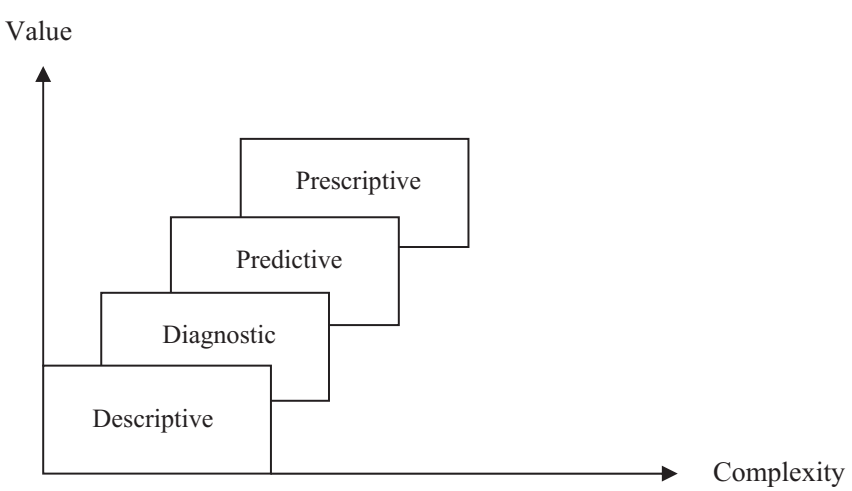


Figure 1 *Types of Data Analytics3*

Second, a diagnostic analytic is applied to further assess what medical and environmental factors could induce and thus associate with heart failure occurrence. In addition, what triggers the heart failure admission increased in winter should also be explored. Usually, it requires content knowledge about risk factors or triggers of heart failure. Therefore, diagnostic analysis is more complexity than the descriptive analysis because it needs integration of analytics skill with specific context.

Third, having known risk factors will lead to the next step of predictive analytics in order to develop a prediction model of heart failure that simultaneously considering medical risk factors and also environmental factor from weather analytics. The predictive model will lead to predict whether it is more or less likely to develop heart failure given a set of risk factors. The predictive analytics uses machine learning, statistics, and data mining to learn from the data and predict it in the future. Therefore, it is more complexity than diagnostic analytics.

Finally, prescriptive analytics is the process of implementation of intervention or treatment program/ strategies using technology such as chat bot (an automatic conversation program) to improve healthcare quality. Patients' and weather data can be gathered to predict risk of heart failure occurrence. Patients will be suggested how to look after themselves and see their physician when needed. It is a process that extends from findings of diagnostic and predictive analytics to inform organization healthcare best practice to improve services. The intelligent chat bot adds business value to the hospital in terms of innovation as well as good patient impression to the hospital services. Therefore, the value and complexity of prescriptive analysis are more than predictive analytics.

Machine Learning, Statistics, and Data Mining

Healthcare big data analytics are empowered by machine learning, statistics, and data mining which incorporate with content knowledge from experts. What are they about? This section aims to clarify the similarities and differences between the fields of study in machine learning, statistics, and data mining. Table 1 summarizes the similarities and differences between machine learning, statistics, and data mining based on two dimensions (ie, objectives and emphases).

For the objective, machine learning and statistics attempt to learn model's parameters from data whereas data mining makes use of such parameters to discover knowledge. Statistics' role, such as applying logistic regression, is focusing on predicting the probability of chronic heart disease using electrocardiogram status (ECG), age, and catecholamine level (CAT). Machine learning's role emphasizes on getting the best out of model performances indicated by accuracy, precision, recall, etc. In comparison, statistics firstly emphasizes on statistical inference of parameters such as odds ratios of factors (eg, age and ECG, CAT) included in logistic regression model. In addition, results of statistic inference can also lead to develop predictive model by simultaneously consider risk factors to predict occurrence of chronic heart disease using the logistic regression model. Machine learning focuses on how well these risk factors (eg, age, ECG, CAT) can predict occurrence of chronic heart disease regardless whether they are statistical associations.

Table 1 Examples of Learning Tasks

Dimensions	Machine Learning	Statistics	Data Mining
Objective	Learn from data	Learn from data	Knowledge discovery
Principal focus	Model performance	Statistical Inference Model performance	Database and data warehouse environment

In a real practice, data scientist mostly uses both machine learning and statistics because of they have similar objectives. However, terminology uses are different because machine learning is originally from the school of computer science whereas statistics is originally from the school of mathematics. These terminologies will be clarified by Figure 2.

In summary, the common aspect between machine learning and statistics is how to learn from data. However, machine learning focuses on how to learn well from data whereas statistics focuses on what the model tells us in the formal statistical inference. *Data mining is strikingly different, which focuses to work in database and data warehouse environments whereas machine learning and statistics work on files in analytic environment to develop a model.* Data mining employs machine learning and statistics to discover knowledge in databases which can be considered as practical machine learning.[4, 5]

Learning from Data

This section emphasizes on how a machine can learn from data. By definition, machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to learn from data.[8] This concept is applicable for both machine learning and statistics approach. A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance *P* if its performance *P* at tasks *T* can improve with experience *E*.[6] The tasks *T* assume that patterns do exist, but their mathematical formalizations are very difficult to achieve whereas data is available for training experience. Therefore, computers learn from experience E which is a known as training data. As example describe in Table 2, for any given learning problem, the class of tasks, performance measures, and the sources of experiences as training data must be identified.

Learning from data involves programming but it is different from traditional programming as shown in Figure 3. For the traditional programming, a computer takes data and runs a program of some deterministic functions in order to determine the output (eg, finding the square root of a given number). The machine learning works differently by taking data and expected output or class labels that learned from training experience into account and it is expected to produce a program, which is a machine learning model. As display in Figure 2, the machine learning model is trained by training data and is evaluated by test data. If the model performance is satisfactory, it will be used to predict diabetes.

Terminologies: (ML, Statistics)

(Features, Independent Variables)

Training Data

(Target Class, Dependent Variables)

Blood Pressure	Sex	Age	Diabetes
Normal	Male	25	No
Low	Female	20	No
High	Male	60	Yes
Normal	Female	65	Yes

(Feature Vector, Observation) → (Class label, Outcome)

Test Data

Blood Pressure	Sex	Age	Diabetes
Normal	Male	40	?
Low	Female	35	?
High	Male	55	?

Figure 2 *Terminologies in Machine Learning and Statistics*

Table 2 Examples of Learning Tasks

Learning Problems	Tasks	Performance Measures	Training Experience
Checkers competition	Playing checkers	Percent of games won against opponents	Games played (against itself or player-player)
Disease diagnosis	Classifying patient's illness	Amount* of correct diagnosis	Medical records
Handwriting recognition	Classifying handwritten characters within images	Amount* of correct recognition	A collection of hand written character images
Self-driving car	Driving on public highway by controlling directions, velocity, and acceleration	Amount* of non-error actions+	A sequence of images from cameras on vehicles

* There are several indicators which will be discussed in the model evaluation.

+ There are several non-error actions (eg, non-violate the traffic laws, safety from accidents).

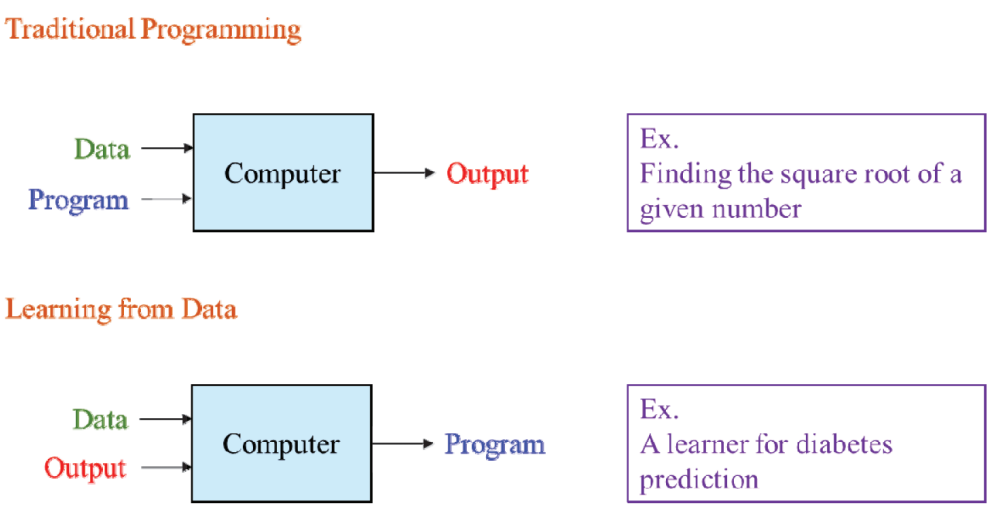


Figure 3 *Traditional Programming vs Learning from Data*

Detail on how a machine is learning from data is illustrated in Figure 4.[7] Suppose we would like to generate the model for prediction of diabetes. In practice, it is relatively impossible to learn from populations who are at risk for diabetic occurrence by collecting data from the whole population. Predictive features of such data is defined as population $\boldsymbol{x}$ whereas the targeted diabetic variable is defined as $\boldsymbol{y}$. There is an ideal function f: $\boldsymbol{x} \rightarrow \boldsymbol{y}$ that maps the input $\boldsymbol{x}$ to the output $\boldsymbol{y}$, but such a function is unknown. Instead of using population data, sample data can be sampling from population data, and learn from this sample data. As describe in Table 3, features including age, sex, body mass index [weight (kg)/height^2 (m)], and blood pressure can be used for predict diabetes occurrence.

Training health data can be storage in several forms (eg, text, image, signal, video) which are known as unstructured data. Machine learning programs cannot directly learn from unstructured data. The procedure called data transformation is needed to convert unstructured data to structured data, which is required data preprocessing techniques such as natural language processing (NLP) for text data, image processing for image data and video which is a sequence of images, and signal processing for signal data.

The structured data is usually in the tabular format as shown in Table 3, which is an example of patient data extracted from healthcare databases. Each row represents the unit of analysis as a feature vector which mostly refers to individual patient or subject. Each column represents a variable or a feature. Depending on the task, such as predictive analysis, a particular feature can be marked as the targeted class where values are called class labels. A training sample is defined as a pair of feature vector and class label. For example, the first patient as a training sample shown in Table 3 is ($\langle$Normal, Male, 180$\rangle$, No).

Formally, let the feature vector $\mathbf{x}_i \in \mathcal{X}$ where i = 1, 2, …, N; and the class label $y_i \in \mathcal{y}$ where i = 1, 2, …, N. The training sample is defined as $(\mathbf{x}_i, y_i)$. The training data is a collection of training samples and is defined as $(\mathcal{X}, \mathcal{y})$.

From the Table 3, we can specify training samples as follows:

$\mathbf{x}_1 = \langle$Normal, Male, 180$\rangle$ y_1 = No

$\mathbf{x}_2 = \langle$Low, Female, 165$\rangle$ y_2 = No

$\mathbf{x}_3 = \langle$High, Male, 175$\rangle$ y_3 = Yes

$\mathbf{x}_4 = \langle$Normal, Female, 160$\rangle$ y_4 = Yes

These training samples are used for training the models with a set of hypothesis $\mathcal{H}$. For example, there are many models that can predict diabetes such as logistics regression, decision tree, etc. Solving such a learning problem requires choosing the model from hypothesis set. Each model has its specific method to learn from data called learning algorithm $\mathcal{a}$. The learned model g is an approximation of f. The closer training samples represent population the closer the approximation to f by g.

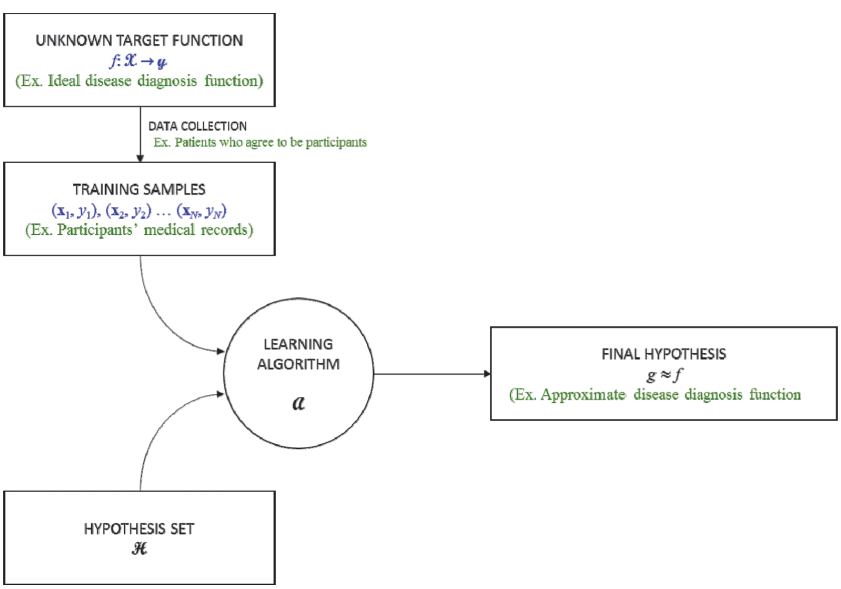


Figure 4 *Learning from Data*

Table 3 Training Data

Blood Pressure	Sex	Height, cm	Weight, kg	Age, y	Diabetes
Normal	Male	180	80	25	No
Low	Female	165	55	20	No
High	Male	175	85	60	Yes
Normal	Female	160	45	65	Yes
Normal	Male	185	88	40	No
Low	Female	170	65	35	No
High	Male	170	65	55	Yes

Conclusions

This article has introduced brief concepts of type of big data analytics, machine learning, statistics, and data mining using healthcare scenarios. Four types of data analytics have been described within the context of heart failure analytics. Since machine learning, statistics, and data mining are important in contributions to healthcare big data analytics, the similarities and differences between them have been discussed using the context of diabetes prediction. General concept to learn from data has also been discussed. These concepts are useful for readers and researchers to move to further how to perform big data analytics using machine learning, statistics, and data mining.

References

1. Wang X, Sontag D, Wang F. Unsupervised learning of disease progression models. In: KDD 2014 - Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. *Association for Computing Machinery.* 2014;85-94. doi:10.1145/2623330.2623754.
2. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes using claims data and analysis of risk factors. *Big Data.* 2015;3(4):277-287. doi:10.1089/big.2015.0020.
3. Maydon T. Four types of data analytics. https://www.kdnuggets.com/2017/07/4-types-data-analytics.html. Accessed May 7, 2018.
4. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Magazine.* 1996;17(3):37-54.
5. Mitchell TM. *Machine Learning.* New York, NY: McGraw-Hill; 1997.
6. Witten IH, Frank E, Hall MA. *Data mining: Practical Machine Learning Tools and Techniques.* 3rd ed. Burlington, MA: Elsevier; 2011.
7. Abu-Mostafa Y.S, Magdon-Ismail M, Lin H-T. *Learning from Data*. New York, NY: AML Book; 2012.
8. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Develop*. 1959: 3(3):211-229.

Review Article/บทฟื้นฟูวิชาการ

การวิเคราะห์ข้อมูลขนาดใหญ่ด้านการดูแลสุขภาพ

รัตน์ชัยนันท์ ธรรมสุจริต[1], อนุตเชษฐ์ พัฒนธีร์ปพน[1], อรลักษณ์ พัฒนาประทีป[1]

[1] กลุ่มสาขาระบาดวิทยาคลินิกและชีวสถิติ คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล

บทคัดย่อ

ปัจจุบันข้อมูลด้านสุขภาพมีการจัดเก็บข้อมูลในรูปแบบอิเล็กทรอนิกส์ทั้งที่มีรูปแบบโครงสร้างข้อมูลที่ชัดเจนและไม่ชัดเจน ซึ่งการใช้ข้อมูลขนาดใหญ่ที่มีความซับซ้อนจำเป็นต้องใช้ทักษะในการวิเคราะห์ข้อมูลขนาดใหญ่ อย่างไรก็ตามบุคลากรทางการแพทย์ส่วนใหญ่ไม่มีความคุ้นเคยกับศาสตร์ด้านนี้ บทความนี้ได้รวบรวมความแตกต่างและแนะนำการใช้เทคนิคการเรียนรู้โดยใช้เครื่องมือ ข้อมูลสถิติ และการทำเหมืองข้อมูล ซึ่งจะนำไปสู่การวิเคราะห์ข้อมูลขนาดใหญ่ด้านสุขภาพ เหมาะสำหรับบุคลากรทางการแพทย์ที่เป็นผู้อ่านที่ไม่ต้องอาศัยเทคนิคทางด้านนี้มากนัก

คำสำคัญ: การวิเคราะห์ข้อมูลขนาดใหญ่ การดูแลสุขภาพ การเรียนรู้ของเครื่อง เหมืองข้อมูล

Corresponding Author: อรลักษณ์ พัฒนาประทีป
กลุ่มสาขาระบาดวิทยาคลินิกและชีวสถิติ คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี มหาวิทยาลัยมหิดล
270 ถนนพระรามที่ 6 แขวงทุ่งพญาไท เขตราชเทวี กรุงเทพฯ 10400
โทรศัพท์ +66 2201 1820 โทรสาร +66 2201 1284 อีเมล oraluck.pat@mahidol.ac.th