CHAPTER 13

DATA COLLECTION APPROACHES FOR ARTIFICIAL INTELLIGENCE APPLICATIONS IN HEALTHCARE

Murat GEZER*, Çiğdem SELÇUKCAN EROL**

*Dr, İstanbul University, Informatics Department,İstanbul, Turkey
e-mail: murat.gezer@istanbul.edu.tr

**Assoc. Prof. Dr., İstanbul University, Informatics Department, İstanbul, Turkey
e-mail: cigdems@istanbul.edu.tr

DOI: 10.26650/B/ET06.2020.011.13

Abstract

As in all other fields, research in the field of artificial intelligence is rapidly continuing in the field of health. As a result of this research, the importance of data comes to the fore. In this study, which includes data collection approaches in the field of health, we aim to emphasize the importance of data in this field and to contribute to the more conscious handling of the data to be used in artificial intelligence applications at every stage. For this purpose, the definition of data and how to distinguish information and knowledge are mentioned. The characteristics of data and data collection methods are also mentioned, and an attempt is made to emphasize the importance of health data collection in artificial intelligence research.

As a result of this study, we believe that all personnel working in data-related departments and the health field, where the moment is vital, must receive training on collecting, storing, sharing data, and data security in particular. In our study we emphasize that especially the people who produce and consume data must have the awareness and morality for every step of data collection and handling, and that this issue should be prioritized in the field of health.

Keywords: Artificial intelligence, Data collection approaches, Data, Healthcare, Machine learning

Introduction

1. Data, Information, and Knowledge

In today's information age, almost every sector has realized the value of data. However, there is much competition in converting data into value (information/knowledge) faster than the competitors. Speed tests are performed at every stage of the process of transforming data into knowledge. No matter how fast you are, you can only go as far as your data. Even if you use the best, most up-to-date algorithms, computers, and technology and work with the best experts, you can only go as far as your data. Therefore, we decided to focus on data on this chapter.

"Data are symbols that represent the properties of objects and events" (Ackoff, 1999). These symbols are the smallest building blocks of knowledge. They are raw information such as numbers, letters, sounds, images, videos that we use frequently and are familiar with. Farmers do not collect raw fruit; they know that this has no or very low value. They wait for it to maturate. After the fruit matures, they collect and sell it as soon as possible. At this point, with many internal and external factors, time is a very critical factor. On the way from data to knowledge, time is also critical. Collecting early or belatedly may not be useful. In order to let data mature, it must be processed. The process of processing data occurs according to an algorithm consisting of a series of steps. We implement algorithms through computers and work with experts; in other words, there are people at every stage of this process. Although it may seem quite simple, it is a complex process, just like the maturing of fruit. So how do we distinguish data from knowledge? Unfortunately, it is not understood simply by biting, as in the case of fruit, and does not turn into knowledge directly. There is another step that we call the information step. This process includes three stages: Data - Information - Knowledge. In sum, data shows that something abstract or concrete exists. But we don't know what that is. When we get the answer to the question "what is this," in other words, when we give the meaning of data, we transform data into information. For example, when we ask "what is 3", it can be answered that 3 is a number. In this case, while 3 represents the data, the number is not information. The fact that 3 is a number characterizes the data itself, whereas we expect it to give meaning to the events or objects it represents. When the answers to the question "what is 3" are like the following examples, data transforms into information;

1. The number of patients waiting in the line
2. A period expressing how many days left for surgery

3. The amount expressing how much medicine I should take
4. The amount of new x-ray devices purchased for the hospital

3 represents different elements such as human, time, medicine, and device, respectively, but this symbol can be the answer to all or even more questions. Information gives us the meaning of the data; in other words, 3. It gives an identity. Three as a number is not just a number anymore. For example, 3 is a person in the first answer! So, is this information valuable? Yes, of course, it is. We often consume this information quickly. But there is one more thing that is more valuable and less available than information: Knowledge.

Knowledge includes the answer to the question "how" and allows us to understand the relationship between multiple information. It contains a process and a result. Let's suppose we have information about a regional power outage in three days. This information, when combined with the information that the generator in a hospital in the same area is out of order, turns into a decision that would delay the operation to be performed in three days at that hospital. And this decision creates a correlation between the knowledge about the process and the information about how a surgical operation can be performed. And as a result, it is decided that surgery should be postponed. Thus, the information about how a surgical decision was taken includes the relationship of information related to the operation process, such as the suitability of the patient to the operation, the availability of an empty operating room, and the availability of the physician and assistant health personnel at that time.

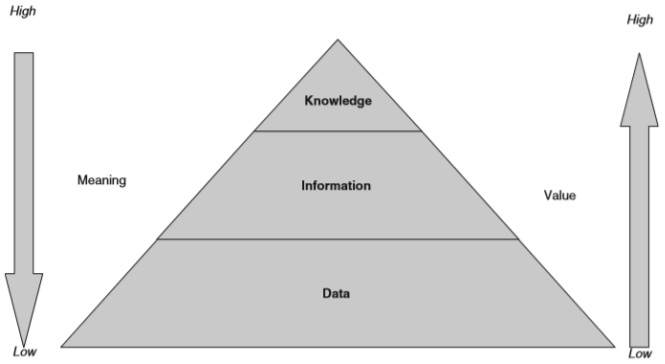


Figure 1: Data, information, and knowledge- Meaning and Value (Chaffey and Wood, 2005 in Rowley, 2007).

As shown in Figure 1, there is a value relationship between the data-information-information pyramid. Knowledge is more valuable than information, and information is more

valuable than data. The more knowledge you access, the more valuable it gets. On the other hand, while the knowledge has the most meaning, meaning decreases when stepping down to data. According to the information you obtain in this chapter, examine the relationship between "3" and "postponing the surgery" again.

1.1. Data sources

Returning to our main subject, we process data to transform it into more valuable information and knowledge. But before processing it, we need to collect and store the data. So why should I take the trouble? Because I have a purpose or a problem. For example, I aim to improve the quality of service in the hospital, or I have encountered a problem in my area of expertise, and I want to solve this problem. First of all, I need to collect data for my purpose or problem. Where can I collect this data when it comes to health and what are my data sources?

1- Patient and patient's relative: A great variety of data can be obtained from the patient and patient's relative such as radiography - image data, blood gas analysis - numeric data, voice recording during anamnesis - audio data, x-ray report - text data, using diabetes pump - sensor data, genetic testing data. We may even include social media data. Although it does not belong to a single patient or a relative, it is possible to obtain data for a drug or disease from thousands of patients and/or their relatives through social media.

2- Electronic health records: Records kept in hospital information systems.

3- Internet: Databases, articles, and similar resources that were previously developed and open to the internet for your purpose or problem.

We need to store the data after collecting it according to our purposes. However, there are a few things to be aware of: Is this data related to my purpose? Is it accurate, reliable, and up-to-date? Is there any legal obstacle for me to share this data? Does the owner of the data give his/her consent?

1.2. Data Types

We want to express the process of storing data using the metaphor of tidying up a room. Your room is very messy, and you start to tidy it up. If you take everything in the room and throw it in your wardrobe, does that equate to tidying up your room? Yes, your room will certainly be tidied up. But what happens to the things you threw into the wardrobe, and how long does it take you to find something when you need it? Or will you even be able to find it?

The process of collecting and storing data is similar to that. If you classify your items one by one and place them where they belong, it will be easier to access them when you need them. Even if you put them into boxes, it will be a waste of time searching which item is in which box as time goes by. Therefore, it is also useful to label the boxes. We made a metaphor about your item, which stands for your data and the box you place it. When we substitute the data for the item again, you will have data about your data, and we call it **metadata**.

We often use databases to store data. We often use organized, ordered, i.e., structured data in databases (NoSQL databases have been used in recent years to store unstructured data). We group our data into structured, semi-structured and unstructured data. For example, an x-ray report of a patient or an e-mail from the head physician is called unstructured data.

(A) UNSTRUCTURED DATA

Dear colleagues,
Your 37-year-old patient with ID number 1 has a body mass index of 25.6, and the mean blood sugar value for three months is 6.1%. Your 25-year-old patient with ID number 2 has a body mass index of 27.8 and a hemoglobin a1c value of 6.7%. It is submitted for consideration.
Kind regards

(B) STRUCTURED DATA

ID	Age	BMI	Hb1ac (%)
1	37	25.6	6.1
2	25	27.8	6.7

Figure 2: Datatypes; Structured and unstructured data

The unstructured data (A) in Figure 2 is converted to structured data (B) and stored in the databases. Apart from these, there is a semi-structured data type that is usually used on websites. This type of data, as its name signifies, has a format between structured and unstructured data. It is stated as labels (Figure 3).

```
<Hospital>
        <Patient ID=1>
                <Age> 37 </Age>
                <BMI> 25.6 </BMI>
                < Hb1ac> 6.1 </ Hb1ac>
        </Patient>
        <Patient ID=2>
                <Age> 25 </Age>
                <BMI> 27.8 </BMI>
                < Hb1ac> 6.7 </ Hb1ac>
        </Patient>
</Hospital>
```

Figure 3: Semi-structured data

Also, the data types are divided into two according to the purpose of collection: primary data type and secondary data type. The data that the researcher collects for the first time for a particular purpose is the primary data, while the secondary data is the data formed by converting it and making it ready for use again. Figure 4 presents a comparison of the two data types.

Primary data	**Secondary Data**
Data collected for research	Data collected in the past
The source of the data is certain	The source of the data is uncertain
Helps us find the solution to the problem	Supports finding the solution to the problem
Data is collected on demand; therefore, it can be structured according to the needs.	
The cost of data collection can be high	It has more relevant costs

Figure 4: Primary data vs. secondary data

2. Health Data Collection Methods

Although the issue of data collection in each sector has country-specific regulatory rules and management challenges, the complexity of data in the health sector and the strict regulations make collecting data difficult (WHO,2003). Since it is related to the privacy of the patient, health data is considered as personal data in the "sensitive data" group (Dülger, 2015). In addition, some of the health data can be categorized as trade secret data. In this respect, the method of collecting and storing health data is important. Health data and a wide range of health indicators for a community are used to assess the costs of measurable clinical health services. Also, scientific studies provide clinical comparisons and can be used to identify the measures needed. Collected data should include all findings related to the patient's condition. These include diagnosis, treatment, and other events that have occurred.

Despite all the difficulties, how can we make the data collected from the data sources usable? Then how can we store this data securely? Nowadays, it is important that the data collected for artificial intelligence research should be usable. Usability is considered an issue that needs to be solved in data collection approaches. Since health data is increasingly used in the field of artificial intelligence, the accuracy and reliability of the collected data is important for the production of useful information from that data.

The choice of data collection methods depends on the purpose of the study (evaluation), the questions to be answered, and the sources from which the data will be collected. In the field of health, data is generated and collected by doctors, nurses, health technicians, health officers, patients, and technological devices. Data collection methods are divided into two

categories, namely *the primary data collection method* and *secondary data collection method,* depending on the type of data (primary and secondary) (Hox 2005).

Primary data collection methods are used for data which is designed, collected, and analyzed to answer a specific research question in the research and these are divided into two groups: *qualitative data collection* and *quantitative data collection*.

In health care, data is collected by devices (Laboratory instruments, Scanning devices, Electroencephalography, etc.), as a written document (surveys, forms, anamnesis, etc.) or as computer inputs (barcode scans, data typing, audio input, text input, etc.). Today, data is collected mostly through digital channels and with the help of numerous applications available on the market. In their study, Sarkies et al. (2015) stated that data collection in the field of health care is done by *observational data*, *retrospective data extraction*, and *retrospective review* methods.

Observational data can be defined as the collection of information that corresponds to the effects of a patient's laboratory values, behavior and life changes, demographic characteristics such as malignancy, and being ill. *Retrospective data extraction* is described as extracting knowledge from previous studies that correspond to a certain impact in a health information system. Retrospective analysis can be defined as scanning all written documents and saving them into an electronic environment after the patient is discharged (Sarkies 2015).

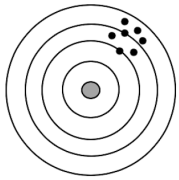
The advantages of the primary data collection method are that the data is collected for the intended purpose, the quality of the collected data is already under the inspection of the research team, and additional data collection is possible when needed. The issue of disadvantages in the primary data collection process is one of the main problems that the researcher has to solve. These disadvantages can be explained as the time required for data collection, ethics committee permission requirements, and costs. The advantages of the secondary data collection are shorter data collection time and lower cost. However, additional data needed for the research can hardly be obtained, and data quality cannot be monitored during data collection. These can be considered as the disadvantages of secondary data collection.

3. Data Quality and Usage

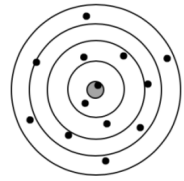
The development of artificial intelligence and the performance of machine learning algorithms depend on the usage of large data sets. Understanding whether the collected data can be used in health research, and artificial intelligence research in particular, concerns data

quality and information governance. The criteria for data quality are validity and reliability (Otieno-Odawa, 2014).

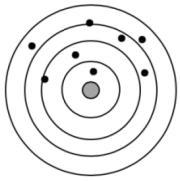
Validity acknowledges the true accuracy of a piece of data. Validity is defined as the concordance of the data to be collected with different instruments, i.e., the data does not affect the results. In other words, there should be no uncertainty. Any deviation in the accuracy of the data will cause the results to be inaccurate. Reliability is not just a feature of the measurement tool. It is a feature of the measurement tool and the results of the tool. If the collected data is obtained in the same way in repeated cases, the data collection method or medium can be considered reliable. Also, reliable data must be understandable. Figure 5 shows the meaning and relationship between the reliability and validity of the data generated as a result of a certain number of shots on a target board (Troachim, 2006). The collected data in Figure 5a is assembled with the same reliability each time, but incorrect measurement of the data is performed systematically. In Figure 5b, the collected data were randomly spread on the target. It was rarely appropriately collected, but in one case, the correct answer was received. Figure 5c shows that the collected data cannot be adequately obtained under any circumstances, and the measurement tools are defective, and finally, Figure 5d shows that the data collection is accurate and that the measurement tools work correctly.



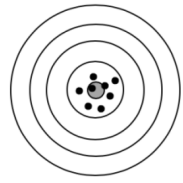
a) (Reliable Not Valid)



b) (Valid Not Reliable)



c) (Neither Reliable Nor Valid)



d) (Both Reliable And Valid)

Figure 5: Relationship between Validity and Reliability (Trochim, 2006).

In addition to the above two criteria, the accuracy, completeness, legibility, relevance, timeliness, and accessibility features of the data are examined in the field of health during the measurement of the quality of data for data processing (WHO 2003, Kirch 2008).

Accuracy: This is defined as the state of data when it meets the gold standard criteria and is measured with reliability and validity (Kirch 2008).

Completeness: This indicates whether the data that should be included in the dataset is missing or not. It is especially important for the accurate decision making of artificial intelligence systems. That is why it should contain all records related to activities about the collected health data/health records, and it should be documented completely and properly (WHO 2003, Kirch 2008).

Legibility: As the data collected during patient registration is **legible**, it becomes possible for data users to comment on the subject. The handwriting to be used in the records should be legible, no undefined coding should be used, and abbreviations should be in accordance with standards (WHO 2003, Kirch 2008).

Relevance: This can be defined as the relevance of an attribute collected in the data content to the subject to be studied. This needs to be taken into account, especially when collecting data for artificial intelligence algorithms, as unnecessary attributes will affect performance (WHO 2003, Kirch 2008).

Timeliness: This is defined as the retention of data in such a way as to include time-dependent changes (WHO 2003, Kirch 2008). Clinical information should be documented as soon as actions are performed. Each activity should be recorded during treatment. Postponement of data entry may cause the skipping of information and errors.

Accessibility: This indicates that the data should be accessible by authorized bodies and persons when needed (WHO 2003, Kirch 2008). If data is not available when it is required, information loses its value.

Data security and information management are also essential factors since data contains sensitive personal information. Therefore, it is necessary to create a very precise balance in data quality and information management (AoRMC, 2019). These balance factors can be listed as clinical considerations, ethical concerns and practical issues arising from the processing of the data (AoRMC 2019). For example, in clinical examinations, the usage of artificial intelligence algorithms in a doctor-patient relationship will involve potential third parties. Confidentiality and security of the data obtained during the treatment will become

questionable. In this case, ethical concerns may arise about who is the owner of the data. Who can be the owner of the data - the patient (i.e., source), the collector (i.e., doctor), the processor (i.e., system), system manager or system owner? Therefore, regulations such as GDPR (General Data Protection Regulations) allow the deletion of personal data. However, in this case, it should be examined whether it is practically possible to remove this data from the artificial intelligence algorithm.

4. Conclusion

In artificial intelligence studies, machines learn from any data such as audio, image, text, signal, etc.. The data we collect and store from different sources for our purpose goes through a very intensive process that we call data preprocessing to use it in artificial intelligence applications. Collecting the data deliberatively in accordance with specific standards provides faster and more efficient execution of preprocessing; hence, artificial intelligence applications.

In this study, we discussed the data collection approaches that are critical in data preprocessing. In addition to these approaches, we believe that following certain standards while collecting and storing data is very important, especially in the field of health, in which there is so much data, and every second is crucial. We believe that all personnel working in data-related departments must receive training on collecting, storing, sharing data and data security in particular. In artificial intelligence research, while we mainly focus on machines, data, algorithms and applications, we ignore the human factor which produces and consumes it. We want to draw attention to the fact that health data, in particular, must be used by conscious and ethically-aware producers/consumers.

References

AoRMC, (2019) Academy of Royal Medicine Colleges Report: Artificial Intelligence in Colleges.

Ackoff, R. L. (1999) Ackoff's Best. New York: John Wiley & Sons, pp 170 – 172.

Ackoff, R.L. (1989) From data to wisdom, Journal of Applied Systems Analysis 16 3–9.

Chaffey, D., Wood, S. (2005). Business Information Management: Improving Performance using Information Systems (FT Prentice Hall, Harlow).

Doğan G.,(2019). Veri toplama araçları [Data Collection Tools]. https://acikveri.ulakbim.gov.tr/acik-veri-acik-bilim/bolum-2-arastirma-verisi-hazirlama-sureci/2-5-veri-toplama-araclari/. (erişim tarihi: 01.8.2019)

Dülger M. V. (2015), Sağlık Hukukunda Kişisel Verilerin Korunması ve Hasta Mahremiyeti [Protection of Personal Data In Health Law And Patient Privacy], İstanbul Medipol Üniversitesi Hukuk Fakültesi Dergisi 1)20, 43-80.

Jox j.j., Boeije H. R., (2005) Data Collection Primary vs. Secondary, Encyclopedia of Social Measurement, Volum 1, Elsevier.

Kirch W,(2008) Encyclopedia of Public Health, Springer.

Lim, C., Kim, K. H., Kim, M. J., Heo, J. Y., Kim, K. J., & Maglio, P. P. (2018). From data to value: A nine-factor framework for data-based value creation in information-intensive services. International Journal of Information Management, 39, 121-135.

Rowley, J., (2007). The wisdom hierarchy: representations of the DIKW hierarchy. Journal of information science, 33(2), 163-180.

Trochim, W. M., (2006). The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: <http://www.socialresearchmethods.net/kb/> (version current as of October 20, 2006).

Sarkies, M. N., Bowles, K. A., Skinner, E. H., Mitchell, D., Haas, R., Ho, M., … Haines, T. P. (2015). Data collection methods in health services research: hospital length of stay and discharge destination. Applied clinical informatics, 6(1), 96–109. DOI:10.4338/ACI-2014-10-RA-0097

Otieno-Odawa, C. F., & Kaseje, D. O. (2014). Validity and reliability of data collected by community health workers in rural and peri-urban contexts in Kenya. BMC health services research, 14 Suppl 1(Suppl 1), S5. DOI:10.1186/1472-6963-14-S1-S5

WHO Library Cataloguing in Publication Data (2003) Improving data quality: a guide for developing countries.

World Health Organization, Geneva