A Study on the Trading Price Estimation Algorithm for Healthcare Transaction Data

Jong-Chil Son
Hankuk University of Foreign Studies
Si-Hyun Sung
Seoul Medical Information Intelligence Lab Inc
Eun-Jung Yang (enyang7@yuhs.ac)
Yonsei University Health System

Research

Keywords: Healthcare Data, Trading Prices, Fundamental Prices, Dynamic Prices, Algorithm

Posted Date: February 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-237190/v1

A Study on the Trading Price Estimation Algorithm for Healthcare Transaction Data

Jong-Chil Son[1], Si-Hyun Sung[2], Eun-Jung Yang[3]

1. Associate Professor, Division of Economics, Hankuk University of Foreign Studies, Seoul 02450, Korea; jcson@hufs.ac.kr
2. CEO, Seoul Medical Information Intelligence Lab Inc, Seoul 06060, Korea; shsung@Exdata.io
3. Clinical Associate Professor, Department of Plastic and Reconstructive Medicine, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea; enyang7@yuhs.ac

Corresponding Author

Eun-Jung Yang, MD, PhD

Clinical Associate Professor

Department of Plastic and Reconstructive Surgery

Yonsei University College of Medicine

50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea

Email: enyang7@yuhs.ac

Tel: +82-2-2228-222

Abstract

Background: While more attention has been paid of late to utilization plans for big data in the healthcare sector worldwide, few scholars have addressed the value estimation of healthcare data. Accordingly, this study aims to propose an idea of a reasonable price estimation algorithm that can be applied to bidirectional exchange in healthcare data platforms.

Methods: This study incorporates three methodologies for the data valuation, namely: cost-based, market-based, and impact-based approaches. The cost-based approach calculates the value of data based on the costs associated with data creation, management, and utilization. On the other hand, the market-based approach evaluates it by comparing the market price of a service similar to the data. Finally, the impact-based approach estimates the data value with an emphasis on improving future revenue generation and productivity as an effect of using the data.

Results: The trading prices of healthcare data are determined by the sum of two prices—the fundamental price and the dynamic price. Here, the fundamental price can be further subdivided into the beginning value, complexity value, and network value. The beginning value is determined in proportion to the physical file size of the data, and the fundamental price is estimated by adding the complexity value and network value that can reflect the qualitative value (within 20% of the beginning value) of the data to the beginning value. First, the complexity value can increase if more personal information, more relevant information to the national health insurance system, and more recent and long-term information are included in the dimensions of identification, material, and time information inherent in healthcare data. Second, the network value reflects whether the data can be well linked with data from, not only the healthcare sector, but also from other fields and sectors. The higher the match rate between the attribute value keyword of the data and the healthcare search keyword of journals of excellence and portal services, the higher value is given. Finally, dynamic price reflects real-time preferences for the data and changes in data supply and demand as the actual exchange proceeds through healthcare data trading. To this end, dynamic value is determined within the upper and lower 5% band of the previous month's trading

price based on the number of monthly views for the data, the number of downloads of summary data, and the number of actual purchases, and this is reflected in the next month's trading price.

Conclusions: If the algorithm for estimating the trading price of healthcare data proposed in this study is applied to actual data trades, it would expand the transactions of healthcare data from both supply and demand sides. Also, in the processes of actual data exchange and the accumulation of actual data trades, continuing studies on the weighting parameters are needed to better reflect reality; such studies would enable the assignment of additional values or penalties.

Keywords: Healthcare Data, Trading Prices, Fundamental Prices, Dynamic Prices, Algorithm

Contributions to the literature

- This study investigates a novel approach to estimating the valuation of healthcare data in a bidirectional data platform, a focus which is largely absent in existing literature.
- As part of its evaluation process, this study exploits not only the inherent data properties reflected in complexity and network values, but also the data market characteristics represented by dynamic prices through both the supply and demand sides of the data.
- Applying this study's algorithm to actual healthcare data trades can provide a relevant starting point for related future research in financial, healthcare, and other data.

Ⅰ. Background

With innovative advancements in mobile technology and the recent COVID-19 pandemic, the development of the big data industry has drawn attention from a wide range of fields, including financial, industrial, and healthcare applications. In this regard, discussions and research on utilization plans for big data in the healthcare and public health sectors have been actively and increasingly carried out worldwide [1]. In South Korea, research on data trading platforms related to the use of big data is in the early stages, mainly in the finance sector. In particular, there have been few studies on the trading platform for healthcare data and the related estimation of the optimal price.

For example, for the estimation standard for charge of data services provided by the Health Insurance Review and Assessment Service (HIRA), a simple cost standard based on public healthcare data is applied. In addition, the charge calculation standard of HIRA is based on the unidirectional service of data; therefore, it has limitations in application to the price estimation method for bidirectional data exchange of data trading platforms, which is expected to be increasingly implemented in the future. In the case of Financial Data Exchange, which was launched in May 2020, bidirectional data exchange is performed in the private sector, but it has yet to clarify the data price estimation method. That is, research on reasonable price estimation methods, which can be applied to bidirectional exchange of healthcare data in South Korea, is in the early stages of development.

Considering the above background, this study analyzed a reasonable price estimation method applicable to bidirectional exchange in the healthcare data platform to be developed in South Korea in the future. The research on value estimation of data is in the early stages across the world, and in particular, this is the first investigation in South Korea on the value estimation of healthcare data. Some scholars have already explored valuations of personal healthcare record systems and related technologies [2, 3, 4]. This study is the first to investigate how to evaluate the true values

themselves inherent in the healthcare data, to the best of our knowledge as of today. Accordingly, this study can be regarded as initiating further associated research, such as the utilization and optimal value estimation of healthcare data that will be developed at full scale in the future.

The main results of this study's analysis of the price estimation algorithm of healthcare data are as follows. First, the estimation of the trading price of healthcare data is equivalent to calculating the sum of the fundamental price and the dynamic price. The fundamental price is the sum of the beginning value, complexity value, and network value. The beginning value is set in proportion to the file size by referring to the data service charge of the National Health Insurance Service (NHIS). For example, the data service charge of the NHIS is set to 10,000 Korean won (roughly equivalent to 9 USD, as per the recent exchange rate) per 1GB (gigabyte). Based on the above beginning value, the fundamental price is determined by estimating the qualitative values of data, such as complexity value and network value, and adding these to the beginning value. The complexity value has a structure in which the value can be increased or, in some cases, reduced according to identification information, material information, and time information. More specifically, an additional value of 10%–20% can be added to the beginning value for each item: for identification information, the value is added with increasing proportion of personal information; for material information, the value is added with increasing proportion of items related with five major benefits in national health insurance; and for time information, the value is added with increasing proportion of more recent data and long-term time series.

The network value of healthcare data reflects how well the data can be linked and combined with data from, not only the healthcare sector, but also from other fields, such as engineering, psychology, and social sciences. To this end, the top 70 keywords of papers published in 71 academic journals of excellence selected by the National Research Foundation of Korea (NRF) for one year were derived, and the top 30 keywords related to healthcare for the past one year were derived from the internet portal services, and in this way, a total of 100 top keywords were

determined. When the match rate of the developed keywords and the keywords of the attribute value of the healthcare data exceeds 10%–20%, an additional value up to 10% can be added to the beginning value.

On the other hand, the dynamic price reflects the real-time preferences of data consumers and similar changes in the data supply and demand as the actual trading proceeds through data exchange after the fundamental price of healthcare data is estimated. In the process, based on several factors (views of the data, downloads of summary data, actual purchases), if any of these falls within the upper 25% of the distribution within a month, an additional value of 1%–5% is added to the previous month's price. On the contrary, if it falls within the lower 25% of the distribution, the price drops by 1%–5%, and if it falls within the medium 50% of the distribution, that is, 26%–74% of the distribution, the price of the previous month is reset as the price at the beginning of the next month.

II. Methods

According to previous studies, cost-based approaches, market-based approaches, and impact-based approaches have been mainly adopted as techniques for the evaluation of data values [5, 6, 7]. First, the cost-based approach is a method of calculating the value of data based on the costs associated with data creation, management, and utilization. On the other hand, the market-based approach evaluates the data value by comparing the price of a service similar to the data in a market where there are suppliers and consumers of the data. Finally, the impact-based approach estimates the data value with an emphasis on improving future revenue generation and productivity as an effect of using the data.

Since the trading market for healthcare data is not formed in South Korea, the data service charge of public institutions is set primarily by cost-based methods. For example, the cost estimation for

the HIRA was analyzed. According to the Korean Intellectual Property Office [8], there are 65 types of public data serviced by the HIRA, and they include both paid and free of charge data. Paid data is provided by the "Big Data Center" run by HIRA, which has both a fixed-term billing method (e.g., 50,000 Korean won for daily use or 300,000 Korean won for yearly use), and case-by-case billing method. This is based on the cost-based approach described above. In addition, a 50% discount is applied to the charges for healthcare institutions, public institutions, schools, and academic societies. However, there is insufficient information on the charge estimation method for these data.

Regarding the NHIS data [9], the charge was estimated based on the cost-based approach. The data were classified into a sample study database (DB) and a customized study DB. First, the sample study DB was classified into four types: health examination cohort DB, elderly cohort DB, infants and children screening cohort DB, and workplace women DB. The DB service charge is applied in proportion to the usage period: 25,000 Korean won per day, 112,500 Korean won per week, and 350,000 Korean won per month. In principle, when using a portable storage device a price of 10,000 Korean won per 1GB is charged. However, a discount rate of 50%–100% is applied, depending on the nature of academic research, policy research, and policy research projects by the NHIS and governmental departments. In addition, for a customized study DB, data viewing and analysis can be performed in the "data analysis room," a place in NHIs with PCs. The DB service charge in this case is set to 50,000 Korean won per day, which is higher than the DB charge for the sample study DB.

Currently, the financial industry is the sector with the most data exchange activities in South Korea, and in this regard, the Financial Data Exchange was launched in May 2020. According to the press release material from the Financial Security Institute [10], the number of participating institutions and registered data has increased sharply within 10 days of launch. However, a reasonable standard for data price estimation has not yet been established. For now, it is thought

that the Exchange accepts the initial quotation price from the data provider as it is.

While the service charge for healthcare data has been estimated based on cost by public service providers, this study incorporates all three methodologies for the data valuation, that is, it incorporates cost-based, market-based, and impact-based approaches. Accordingly, the transaction price estimation algorithm of healthcare data is analyzed by dividing it into two stages: the fundamental price and the dynamic price.

III. Structure of Healthcare Data in Korea

According to the National Data Map of DATA.GO.KR [11], the healthcare sector is classified into two categories: “health insurance” and “medical care.” The former is classified according to the proportion of serviced data, such as health insurance (7.12%), prescription drugs (3.31%), Korea Pharmaceutical Information Service (3.31%), healthcare resources (3.05%), medication covered by health insurance (2.8%), surgery (0.76%), and chronic disease patients (0.76%) among others. For medical care, the proportions of each category are as follows: hospitals (5.34%), medical treatment (2.31%), medical clinics (2.24%), healthcare benefits (0.31%), and public health (0.29%). In other words, data related to health insurance, medication, hospitals, and medical care account for a large part of healthcare data.

As seen in Table 1, according to HIRA [12], the healthcare big data of the relevant institution can be largely divided into six categories: healthcare treatment information, medication information, treatment materials information, healthcare resources information, healthcare service quality evaluation information, and non-benefit information. The data are primarily for insurance review, and data on healthcare treatment and medication, in direct relation to healthcare

charge, account for the main portion; a wide range of information, such as treatment materials and healthcare resources, is included in the big data.

[Table 1 around here]

In addition, the National Health Insurance Service [9] provides cohort DB services collected by year and province. It is divided into four types of DBs: health examination cohort, elderly cohort, infants and children screening cohort, and workplace women cohort. The DB is generally composed of information on qualifications, healthcare treatment, disease history and health examination, and long-term care facilities for the period.

As can be seen from the above, various healthcare data serviced by the public and private sectors have considerable differences in the structure and format of data. Furthermore, due to regulations such as personal information protection and institutional approval processes, there are difficulties in accessing to and using the data. To address these limitations, previous studies have investigated the utilization of distributed research network and common data model [13]. The distributed research network manages distributed data only, without sharing the original data between consumers, which reduces the difficulties from restrictions on the use of healthcare data. Therefore, for efficient utilization of data from multiple hospitals for research, the implementation of common data model, in which data are saved and managed in a shared format, is essential.

In this case, the healthcare data are highly standardized for utilization, facilitating links with other healthcare data, which adds higher values to the healthcare data. These properties of healthcare data are reflected in the "network value," which will be discussed in more detail in Chapter IV.

IV. Main Results: Trading Price Estimation Algorithm

The estimation of the trading price of healthcare data is largely divided into two stages: estimation

of the fundamental price and the dynamic price. In this study, an algorithm for estimating the optimal price was analyzed for the two stages. This algorithm encompasses the three approaches of data value discussed previously: a cost-based approach, a market-based approach, and an impact-based approach.

4.1 Fundamental price estimation

In this section, fundamental price estimation is first analyzed for the price estimation of healthcare data (database). The fundamental price consists of three components: the beginning value, complexity value, and network value. The estimation method for each value is presented below.

4.1.1 Beginning value

Healthcare data are likely to be in the form of an Excel or text file. It is reasonable that the beginning value for the fundamental price estimation of healthcare data is estimated in proportion to the simple size of the file. In the case of the NHIS data mentioned above, a price of 10,000 Korean won per 1GB is charged when using USB. Therefore, with reference to this charge, the beginning value can be applied as shown in Equation (1) below in proportion to the ratio that 1GB is equal to 10,000 Korean won:

$$BV = \alpha \times filesize$$

$$where \quad \alpha = \frac{10{,}000 \ \ Korean \ \ Won}{1{,}000 \ \ KB}, \tag{1}$$

where BV is the beginning value.

For example, if the size of the healthcare data is 4,150KB (kilobytes), the beginning value is set to 41,500 Korean won, roughly equivalent to 40 USD. However, it is necessary to consider that, because the beginning value is estimated based on the data service charge of a public institution

with a cost-based approach, the value can be lower than the price set by a private market. In addition, if there is a more reasonable case of reference for the beginning value estimation, it can be reflected and applied for estimation.

4.1.2 Complexity value

The second element of healthcare data fundamental price, the complexity value, evaluates the value of the complexity and importance of healthcare data, as shown in Equation (2) below. The value is set by estimating and reflecting the qualitative value of the data to the beginning value discussed above:

$$CV = \sum_{i=1}^{3} w_i BV, \tag{2}$$

where *CV* is the complexity value and w_i indicates the identification information dimension, material information (or depth) dimension, and time information (or width) dimension of the healthcare data, respectively. Examining this one by one, w_1 represents the aggregate level of identification information level, and a different weight is assigned for each case. The level of identification information is classified into four levels of person/county/province/nationwide, as can be seen in Equation (3):

$$w_1 = \sum_{j=1}^{4} s_j id_j = s_1 person + s_2 county + s_3 province + s_4 nationwide$$

$$s_1 + s_2 + s_3 + s_4 = 1$$

$$id_1 = person = 0.2, id_2 = county = 0.1,$$

$$id_3 = province = 0.05, id_4 = nationwide = 0, \tag{3}$$

where s_1, s_2, s_3, and s_4 represent the proportion of appliable identification information from the entire dataset. That is, $s_1 + s_2 + s_3 + s_4 = 1$. In addition, *persons, counties, provinces*, and *nationwide* are additional weights given to the data in units of person, county, province, and nation, respectively. Regarding personal level information, an additional weight of 20% was given. This is because the largest cost is required at the stage of the statistical process, where a personal identifier should be transformed into anonymity due to individual sensitive information. A weight of 10% is given for data in county units, 4% for province units, and no additional weight for data in nationwide units. This is because if personal level data are available, regional and higher-level aggregate data can be developed and constructed with ease.

For example, if all the data items are composed of personal information, the complexity value is set by adding an additional weight (*person* = 0.2) of 20% to the beginning value. For individual hospitals or clinics where identification information is not a person, an additional value can be given on a county basis. This is because, although not at the same level of sensitivity as personal information, the comparative analysis by county is likely to include more useful information compared to aggregate information in the unit of province. However, more empirical studies are needed for this type of additional weighting parameter system in the future. That is, the cost of de-identifying the identifier of personal information may be different for each statistical agency and for each type of statistic. In addition, the level of accessibility of users to aggregate data, such as county and province, can be very different for each case.

As shown in Equation (4), the content of material information is classified into four levels: five major benefits (benefit5), other benefits (benefit), and non-benefit/others,

$$w_2 = \sum_{k=1}^{4} s_k depth_k = s_1 benefit5 + s_2 benefit + s_3 nonbenefit + s_4 others$$

$$s_1 + s_2 + s_3 + s_4 = 1$$

$$depth_1 = benefit5 = 0.2, depth_2 = benefit = 0.1,$$

$$depth_3 = nonbenefit = 0.05, depth_4 = \text{others} = 0, \quad (4)$$

where w_2 represents the material information dimension of healthcare data (and a different weight is assigned to each case); s_1, s_2, s_3, and s_4 represent the proportion of appliable material information from the entire dataset ($s_1 + s_2 + s_3 + s_4 = 1$); and *benefit5, benefit, non-benefit,* and *others* represent items related to five major benefits in national health insurance, items related to other benefits, non-benefit-related items, and items other than national health insurance. First, the five major benefit-related items refer to the five diseases that account for major portions of the national health insurance payment, such as cancer. Other benefit-related items refer to the remaining benefit items, excluding the five major benefits items. Non-benefit-related items refer to those excluded from national health insurance benefits, such as cosmetic and dental procedures for esthetic purposes. Finally, all items not included in the above four categories were classified as "others". The healthcare resources and healthcare service quality evaluation information in <Table 1> can be classified into this category. In summary, this enables the provision of additional values to data attributes related to the actual prevalence of diseases and the associated costs for the public.

As can be seen from Equation (5), depending on whether the data are updated, weight or penalty is assigned. If data within the last three years are included, an additional value (10%) can be assigned owing to the update status of the data. However, if there is no data within the last three years, a penalty (-5%) can be given. This is because old data are less likely to be relevant in contemporary times.

$$w_3 = \sum_{l=1}^{3} I_l width_l = I_1 updated + I_2 period + I_3 frequency$$

$$I_1 = \begin{cases} 1: Incuding\ \ data\ \ within\ \ last\ \ 3\ \ years \\ -0.5: Not\ \ including\ \ data\ \ within\ \ last\ \ 3\ \ years \end{cases} \qquad updated = 0.1$$

$$I_2$$

$$= \begin{cases} 1: Cumulative\ data\ for\ 10\ years\ or\ longer \\ 0.5: Cumulative\ data\ for\ 5-9\ years \\ 0.3: Cumulative\ data\ for\ 2-4\ years \\ 0:\ Data\ for\ 1\ year\ or\ less \end{cases} \quad period$$

$$= 0.1$$

$$I_3$$

$$= \begin{cases} 1: 50\%\ or\ more\ of\ the\ data\ consisting\ of\ monthly\ or\ quarterly\ data \\ 0:\ 50\%\ or\ more\ of\ the\ data\ consisting\ of yearly\ data \end{cases}$$

$$frequency = 0.1 \quad (5)$$

where w_3 represents the time information dimension of healthcare data.

The time series span of the healthcare data is important. Because the cross-sectional information (depth) of the data has already been reflected in the material dimension of Equation (4), in this part, information in the time-series dimension (width) is considered and reflected as shown in Equation (5). That is, there is no additional weight for data within one year, and an additional weight is given in proportion to the time exceeding one year. The last factor for consideration in the dimension of time information of the data is the frequency information of the data. Because of the nature of healthcare data, annual data are mainly used for most cases, but there may be monthly or quarterly data depending on the occasion; when these data are included, an additional weight of 10% can be given. This is because if monthly/quarterly data are available, it is relatively easy to construct annual data, and in the case of monthly/quarterly data, it is easier to use in combination with other DBs.

4.1.3 Network value

The network value, the third and final component of healthcare data fundamental price, is a value that reflects whether the data can be well linked with data from other fields, such as engineering,

psychology, and social sciences. For example, even when there are two sets of healthcare data with the same beginning value and complexity value, if one set of data is more likely to flexibly connect with data from other research fields, a higher value can be assigned to this data compared to the other data set.

$$NV = \sum_{m=1}^{3} J_m link_m \, BV = (J_1 KCI_{HSS} + J_2 KCI_{ST} + J_3 Portal) BV$$

J_1

$$= \begin{cases} 1: 10\% \ or \ higher \ of \ keyword \ match \ rate \ of \ journals \ of \ excellence \\ 0.5: 5 - 9\% \ of \ keyword \ match \ rate \ of \ journals \ of \ excellence \\ 0: \ 4\% \ or \ lower \ of \ keyword \ match \ rate \ of \ journals \ of \ excellence \end{cases}$$

$$link_1 = KCI_{HSS} =$$

$0.1, where \ HSS \ standing \ for \ humanities \ and \ social \ science.$

J_2

$$= \begin{cases} 1: 20\% \ or \ higher \ of \ keyword \ match \ rate \ of \ journals \ of \ excellence \\ 0.5: 10 - 19\% \ of \ keyword \ match \ rate \ of \ journals \ of \ excellence \\ 0: \ 9\% \ or \ lower \ of \ keyword \ match \ rate \ of \ journals \ of \ excellence \end{cases}$$

$$link_2 = KCI_{ST} = 0.1, where \ ST \ standing \ for \ science \ and \ technology$$

J_3

$$= \begin{cases} 1: 20\% \ or \ higher \ of \ searching \ keyword \ match \ rate \ of \ healthcare \ in \ major \ portals \\ 0.5: 10 - 19\% \ of \ searching \ keyword \ match \ rate \ of \ healthcare \ in \ major \ portals \\ 0: \ 9\% \ or \ lower \ of \ searching \ keyword \ match \ rate \ of \ healthcare \ in \ major \ portals \end{cases}$$

$$link_3 = Portal = 0.1 \quad (6)$$

where *NV* is the network value, *BV* is the beginning value, and *KCI* is the Korea Citation Index.

In fact, it can be considered that the value of a certain capability of the network was present in

the material and time information of the data analyzed in complexity value described above. That is, if the proportion of the items related to the benefits of national health insurance is high in consideration of the importance of diseases such as cancer, academic interest from other fields such as psychology and social science is generally increased, and accordingly, it is highly likely that the expanded application of healthcare data will increase. In addition, this section also reflects the value of the network that healthcare data can have.

As can be seen from Equation (6), this section estimates the network value by reflecting the three search trends. First, from the current status of NRF-listed academic journals as of October 2020, there are 40 journals of excellence in the humanities and social sciences fields, and 31 journals of excellence in the science and technology fields. A pool of data can be developed by collecting keywords from all papers published in the journals during the past year. Duplicate or overlapping keywords were removed or organized to derive a list of the top 70 search keywords. In addition, the top 30 keywords related to healthcare were derived from portal service providers such as Naver and Daum over the past year, and 100 keywords were compiled in total. Similar and duplicate search keywords were treated as one word. Second, by examining the match rate of the keywords compiled in this way and the attribute value keywords of the applicable healthcare data, an appropriate additional value is given in proportion to the match rate.

In addition, the match rate can be applied differently for the humanities, social sciences, science and technology, and healthcare fields. For example, if the keyword match rate in the humanities and social field exceeds 10%, an additional weight of 10% is given, whereas in the case of science technology and healthcare, an additional weight of 10% is given only when the match rate exceeds 20%. However, as the actual application cases for the applicable healthcare data are accumulated, more realistic numbers and appropriate classification methods can be applied for these match rates.

The fundamental price of healthcare data, summarized by Equation (7), consists of three

components: beginning value, complexity value, and network value. Then, for complexity value and network value, through estimation of potential qualitative values that the data may have, these values can be added to the beginning value:

$$\begin{aligned} FP &= BV + CV + NV \\ &= BV + \sum_{i=1}^{3} w_i\, BV + \sum_{m=1}^{3} J_m\, link_m BV \\ &= (1 + \sum_{i=1}^{3} w_i + \sum_{m=1}^{3} J_m\, link_m) BV, \end{aligned} \tag{7}$$

where *FP* is fundamental price, *BV* is beginning value, *CV* is complexity value, and *NV* is network value.

4.2 Dynamic price estimation

Dynamic price is commonly known as a method of presenting differentiated prices to customers with different price sensitivities [14, 15, 16]. It has been reported that if prices are differentiated by reflecting changes in demand in real time based on techniques including AI algorithms, greater profits can be created than when a single price is set. The definition of dynamic price in this study is somewhat different from the definition above. That is, after the fundamental price of healthcare data is set, the dynamic price of this study refers to a price element that can change dynamically over time, as actual trades are made through data exchange.

Dynamic price is estimated by reflecting changes in the supply and demand of the market and preferences of data consumer groups, such as the intensity of real-time demand for the data with fundamental price value set, and variations in the supply of similar healthcare data. Due to the nature of healthcare data, the actual data preference expressed from the user side rather than the provider side is likely to lead to a dynamic price change. Accordingly, dynamic price is estimated for a specific period (e.g., one month) based on the number of views on the data, the number of

downloads of summary data, and the number of actual purchases, as shown in Equation (8):

$$DP = \sum_{n=1}^{3} D_n dynamic_m \, P = (D_1 View + D_2 Download + D_3 Purchase)P$$

$$D_1 = \begin{cases} 1: Upper\ 25\%\ or\ greater\ in\ distribution\ of\ number\ of\ views\ in\ one\ month \\ 0: Within\ 26-74\%\ in\ distribution\ of\ number\ of\ views\ in\ one\ month \\ -1:\ Lower\ 25\%\ or\ less\ in\ distribution\ of\ number\ of\ views\ in\ one\ month \end{cases}$$

$$dynamic_1 = View = 0.01$$

$$D_2 = \begin{cases} 1: Upper\ 25\%\ or\ greater\ in\ dist.of\ number\ of\ summary\ downloads\ in\ one\ month \\ 0: Within\ 26-74\%\ \ in\ dist.of\ number\ of\ summary\ downloads\ in\ one\ month \\ -1:\ Lower\ 25\%\ or\ less\ \ in\ dist.of\ number\ of\ summary\ downloads\ in\ one\ month \end{cases}$$

$$dynamic_2 = Download = 0.01$$

$$D_2 = \begin{cases} 1: Upper\ 25\%\ or\ greater\ in\ dist.of\ number\ of\ data\ purchase\ in\ one\ month \\ 0: Within\ 26-74\%\ \ in\ dist.of\ number\ of\ data\ purchase\ in\ one\ month \\ -1:\ Lower\ 25\%\ or\ less\ \ in\ dist.of\ number\ of\ data\ purchase\ in\ one\ month \end{cases}$$

$$dynamic_3 = Purchase = 0.01 \quad (8)$$

where *DP* is dynamic price.

As can be seen from Equation (8), an additional value of 1% is given to the number of data views and the number of summary data downloads, while an additional value of 3% is given to the number of actual data purchases. In addition, the value of the index function (D) was applied differently according to the relative distribution of the number of views, the number of downloads of summary data, and the number of actual purchases on a cumulative basis over a period of one

month. The distribution was divided into the upper 25%, medium 26%–74%, and the lower 25%, respectively, and the distribution of the medium 50% was set so that there was no change in the dynamic price. Furthermore, in the setting, for the upper 25%, price increase within 5% of the fundamental price, for the lower 25%, a price decrease within 5% of the fundamental price was allowed. Finally, a period of one month was set for variation, and price variations were set within the upper and lower 5% bands because the number of views, the number of downloads of summary data, and the number of actual purchases all have a high positive correlation; if the data demand is concentrated at once collectively, the value exceeds fundamental price too much for price determination and setting privations within this range reduces such risk. Furthermore, the number of views and the number of downloads of summary data may be exposed to the risk of human manipulation in some cases, so the dynamic price was set to vary within a certain band of the fundamental price.

When dynamic price is updated at the end of each month, the updated dynamic price is added to the already determined fundamental price and this is reflected in the next month's trading price. In addition, by updating the dynamic price again at the end of the month according to the distribution of the number of views and actual purchases, the trading price of the next month is reset. Of course, if the healthcare data are traded for the first time, the price is set based on the fundamental price in the month, and the dynamic price is set from the point when it is fully exposed for a one-month trading period.

4.3 Trading price

Summarizing the above analysis, the trading price of healthcare data can be represented as the sum of the fundamental price and the dynamic price at the end of the previous period, as shown in Equation (9),

$$P_t = FP_0 + DP_{t-1}$$
$$= BV_0 + \sum_{i=1}^{3} w_i BV_0 + \sum_{m=1}^{3} J_m \, link_m BV_0 + \sum_{n=1}^{3} D_n \, dynamic_n P_{t-1}$$
$$= (1 + \sum_{i=1}^{3} w_i + \sum_{m=1}^{3} J_m \, link_m) BV_0 + \sum_{n=1}^{3} D_n \, dynamic_n P_{t-1} \quad (9)$$

where *P* is the trading price, *FP* is the fundamental price, *DP* is the dynamic price, and *BV* is the beginning value.

That is, the fundamental price is determined at the time of data exchange registration (time 0), and the dynamic price based on the trading volume of the previous period (or previous month) varies, and the trading price of the next period (or next month) is updated. In addition, the fundamental price is determined by the values of three dimensions, namely, beginning value, complexity value, and network value, and dynamic price is determined by the relative distribution of the number of views, summary data downloads, and actual purchases for the month. This trading price decision process can be illustrated as an algorithm, as shown in <Figure 1>.

[Figure 1 around here]

4.4 Data value expiration and comparison with bond prices

According to Equation (8), when the value of the data falls within the lower 25% and the demand for the data continues to decrease thereafter, the probability of the data value remaining in the lower 25% may increase. In this case, the trading price may drop by approximately 5% of the last month's trading price every month. It is estimated to take about 7.5 years for the trading price to fall to 1% of the trading price, which is the price before the value is included in the lower 25%

distribution in terms of dynamic price estimation. If the upper and lower bands set at dynamic price estimation are increased, this time of 7.5 years will decrease, and when the bands are decreased, the time will increase. For example, if the upper and lower bands are decreased to 3%, it is estimated to take about 12.6 years for the data value to fall to 1%. This is similar to applying the principle of bond price estimation to the value estimation of healthcare data. That is, one of the most fundamental characteristics of a financial debt contract, such as bonds, is the maturity of liability. After 3, 5, or 10 years of maturity, the value of the bond expires. Until maturity, bond prices vary according to the principle of supply and demand in the bond's secondary market. When these characteristics are applied to healthcare data, after a paid period (e.g., 5, 7, or 10 years), on the basis of prior agreement with the data providers, the healthcare data can be provided free of charge.

The characteristic of the bond price structure is that when one reference bond price is determined, a certain liquidity premium or risk premium is applied in conjunction with the price, and the price of other similar bonds is automatically determined. For example, if the price of 3-year Korean Treasury Bond (KTB) is determined, it is highly possible that the price of 5-year and 10-year KTB, which have the same characteristics except maturity period, will be determined by adding only a certain liquidity premium to the 3-year KTB price. This is because in the case of KTB, there is little risk premium, regardless of whether the maturity is 3-year or 10-year, and only compensation for liquidity constraints is required.

When this principle is applied to the price estimation of healthcare, almost similar or identical healthcare data sets are traded every year or within a certain period, the trading price of the present healthcare data can serve as an important reference for the fundamental price of the newly participating data. For example, if the data are almost similar, only the premium of the remaining maturity is added to the fundamental price of the new data. However, factors such as how much similar or identical healthcare data can actually be traded, are matters for consideration that need to be further investigated, based on the trend of actual trades from a long-term perspective.

V. Discussion

There has been increasing interest across the globe on the value of big data utilization, along with the innovative advancement of mobile and platform technology. Research and discussions have been underway in South Korea to enhance the utilization of healthcare data in both supply and demand. Accordingly, this implies that it is advisable to establish a reasonable value estimation for the data because a reasonable price can contribute to the active trading of data while mediating the supply and demand of healthcare data. However, there are few studies on the value estimation of healthcare data in South Korea, except for this study. With this in mind, this study proposed an idea of a reasonable price estimation algorithm that can be applied to bidirectional exchange in the healthcare data platform that is expected to be developed in South Korea in the future.

The main results are as follows. First, the estimation of the trading price of healthcare data is done through first estimating the fundamental price and the dynamic price and then summing these prices. The fundamental price is the sum of the beginning value, complexity value, and network value. The beginning value is set in proportion to the file size by referring to the data service charge of the NHIS. Based on the beginning value, the fundamental price is determined by estimating the qualitative values of data such as the complexity value and the network value, and adding these values to the beginning value.

The complexity value has a structure in which the value can be added or, in some cases, a penalty can be imposed according to identification information, material information, and time information, the three dimensions inherent to the medical data. An additional value of 10%–20% can be added to beginning value for each item: for identification information, the value is added with increasing proportion of personal information; for material information, the value is added with increasing proportion of items related with five major benefits in national health insurance; and for time information, the value is added with increasing proportion of more recently updated

data and long-term time series.

The network value of healthcare data reflects how well the data can be linked and combined with data from, not only the healthcare sector, but also from other fields such as engineering, psychology, and social sciences. To this end, the top 70 keywords of papers published in 71 academic journals of excellence in the NRF for one year were derived, and the top 30 keywords related to healthcare for the past year were derived from the Internet portal service, in which a total of 100 top keywords were determined. When the match rate of the developed keywords and the keywords of the attribute value of the healthcare data exceeds 10%–20%, additional value up to 10% can be given to the beginning value.

On the other hand, dynamic price reflects the real-time preferences of data consumers and similar changes in the data supply and demand as the actual trading proceeds through data exchange after the fundamental price of healthcare data is estimated. In the process, based on the number of views of the data, the number of downloads of summary data, the number of actual purchases, if any of these falls within the upper 25% of the distribution within a month, an additional value of 1%–5% is given to the previous month's trading price. On the contrary, if it falls within the lower 25% of the distribution, the price drops by 1%–5% of the previous month's trading price, and if it falls within the medium 50% of the distribution, that is, 26%–74% of the distribution, the price of the previous month is reset as the trading price at the beginning of the next month.

In addition, after a certain time has elapsed, the demand for old healthcare data is bound to decrease continuously. When the equation of dynamic price estimation in this study is applied to this type of old healthcare data, it is highly probable that the price decreases monthly by about 5% of the price of the previous month. In this case, it is estimated to take about 7.5 years for the price of the data to fall to 1% of the fundamental price. When considering characteristics such as the expiration period of healthcare data value, after a certain paid period (e.g., 5, 7, or 10 years), on the basis of prior agreement with the data providers, the healthcare data can be provided free of

charge on a collective basis.

The algorithm for estimating the trading price of healthcare data proposed in this study is one of the ideas with possibilities of application. In the process of actual data exchange and accumulation of actual data trades in the future, further studies on the weighting parameters are needed to better reflect the reality, leading to the assignment of additional values or penalties on them. Finally, we expect that the discussion of this study will contribute to shedding light on future studies for the estimation of optimal values inherent to all types of data, including healthcare data.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Data in the manuscript, including equations and the figure (flowchart), were generated by the authors. The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

None.

Authors' contributions

JC developed the design of the methodology, created the models, and researched and wrote the original draft. SH conceptualized research subjects, provided resources, and reviewed and edited the manuscript. EJ edited the manuscript and supervised and managed the project. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

List of abbreviations

Database (DB)

Korean Treasury Bond (KTB)

Health Insurance Review and Assessment Service (HIRA)

National Health Insurance Service (NHIS)

National Research Foundation of Korea (NRF)

References

1. OECD, Data-Driven Innovation: Big Data for Growth and Well-Being, OECD Publishing, Paris, 2015.
2. Shah S, Kaelber D., Vincent A., Pan E., MD, Johnston D., Middleton B., A cost model for personal health records (PHRs). AMIA 2008 Symposium Proceedings. 2008.
3. Walker J., Pan E., Johnston D., Adler-Milstein J., Bates D., Middleton B. The value of health care information exchange and interoperability. Health Affairs. 2005.
4. Kaelber D., Pan E. The value of personal health record (PHR) systems. AMIA 2008 Symposium Proceedings. 2008.
5. Yoon KG, Ahn MO. Case Analysis and Implications for Data Value Creation, KISDI Publication 122, Autumn, 2020. (in Korean)
6. UNECE, Recommendations for promoting, measuring and communicating the value of official statistics. United Nations, 2018.
7. Li W., Nirei M., Yamana K. Value of data: There's no such thing as a free lunch in the digital economy. Discussion papers 19022, Research Institute of Economy, Trade and Industry. 2019.
8. Korean Intellectual Property Office, Study on calculating the cost of information provision service fee for industrial property rights, publication number 11-1430000-0016-565-01, 2017. (in Korean)
9. National Health Insurance Service. Applicable statistics and public data. 2020. https://nhiss.nhis.or.kr/bd/ag/bdaga001lv.do. Accessed October 25, 2020.
10. Financial Security Institute. Trend and upcoming plans for financial data exchange, press release. http://www.fsec.or.kr/user/bbs/fsec/41/18/bbsDataList.do. Accessed October 25, 2020. (in Korean)
11. DATA.GO.KR, National Data Map. 2020. https://www.data.go.kr/. Accessed October

25, 2020.

12. Health Insurance Review and Assessment Service, Healthcare big data used casebook, Big Data Center of Health Insurance Review and Assessment Service, 2020. (in Korean) http://repository.hira.or.kr/handle/2019.oak/2312.

13. Park RW, Yu SC. Cases of CDM application in healthcare sector and future utilization plan. Big Data Briefing Report, Health Insurance Review and Assessment Service, 2018;2(4):5–18. (in Korean)

14. Kim YH. Expanding Application of Dynamic Pricing. LG Economic Research Institute. 2017. (in Korean)

15. Lee HJ, Kim KB, Choi YK. Application of AI algorithms for business opportunities creation, Samjong KPMG ISSUE MONITOR, 2018;84. (in Korean)

16. Anderson C., Xie X. Dynamic pricing in hospitality: overview and opportunities. International Journal of Revenue Management, Inderscience Enterprises. 2016.

<Table 1> Composition of HIRA healthcare data

Classification	Description	Number of data as of December 2019
Healthcare treatment information	Billing statement, definition of healthcare treatment and patient classification, charge master, disease group and nursing hospital charge master, review criteria for each healthcare treatment, size of treatment by each category and others	8,893
Medication information	Benefits medication master, review criteria by benefits medication, medication production, manufacturer and distributor information, drug product distribution, benefits medication use, medication safety management	54,130
Treatment materials information	Treatment materials master, healthcare institutions purchase, use by treatment materials, information related to special materials	35,698
Healthcare resources information	Opening and closing up of long-term care facility, long-term care facility setup, health care personnel status and qualification, equipment possession status, usage information by equipment	94,865
Healthcare service quality evaluation information	Evaluation by long-term care facility, information on treatment results of acute diseases, chronic diseases and cancer, drugs, intensive care unit fixed-term charge, treatment volume and others	34[1)]
Non-benefit information	Non-benefit healthcare cost, non-benefit items (309 items), certifying documentation provision charge (31 items), minimum and maximum cost per healthcare institution and others	340

Note : 1) As of 2018

Figure caption

Fig. 1 Algorithm for trading price determination.

Figures

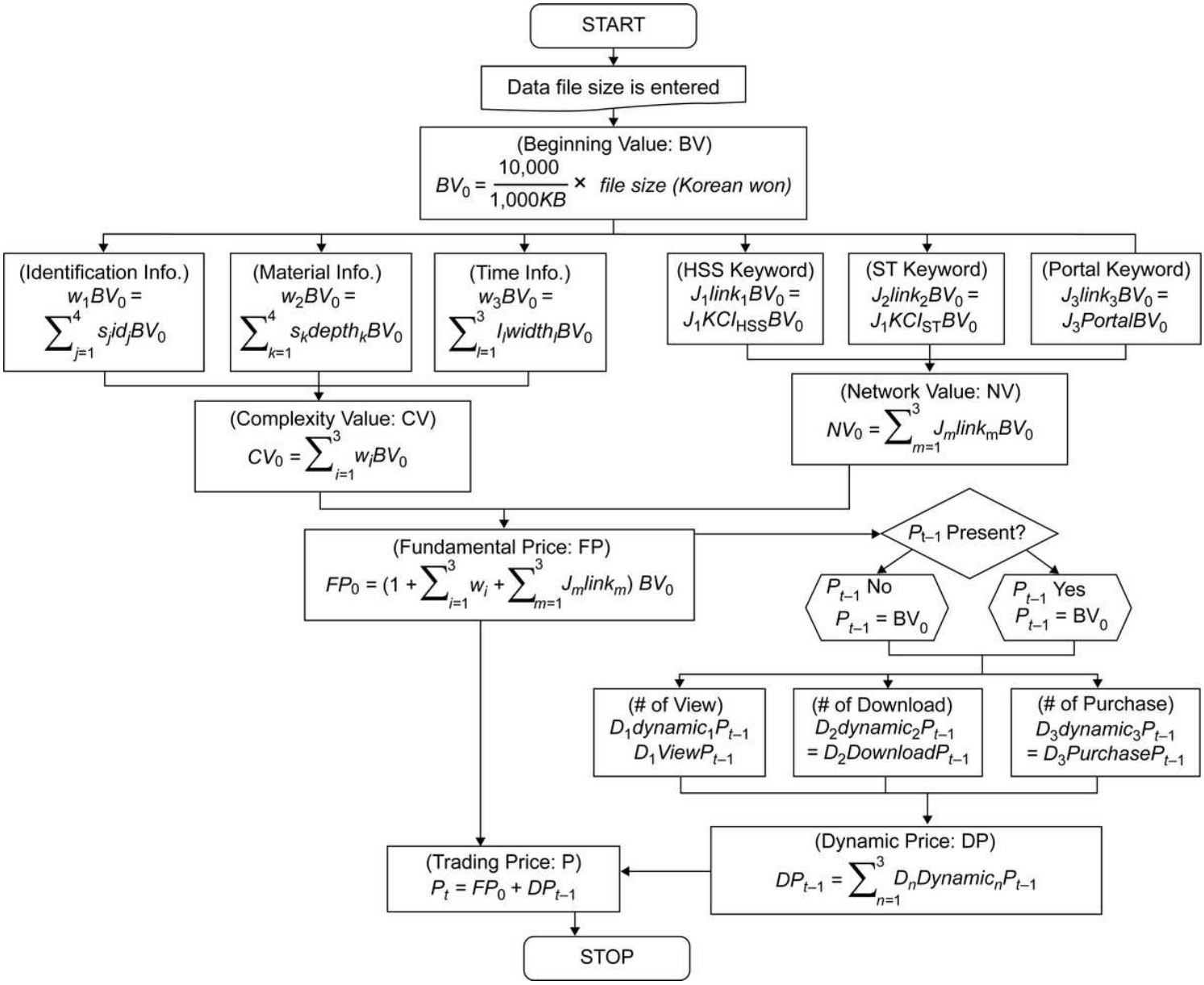

Figure 1

Algorithm for trading price determination.