Original Article

Alliance chain-based simulation on a new clinical research data pricing model

Jing Li[1,2,3#^], Dejian Wang[4#], Guoqiang Qi[1,2,3^], Zheming Li[1,2,3^], Jian Huang[1,2,3^], Zhu Zhu[1,2,3^], Chen Shen[1,2,3^], Bo Lin[5,6^], Kexiong Dong[4^], Baolong Zhao[4], Qiang Shu[1,3], Jianwei Yin[5,6], Gang Yu[1,2,3,7^]

[1]Department of Data and Information, The Children's Hospital Zhejiang University School of Medicine, Hangzhou, China; [2]Department of Research, Sino-Finland Joint AI Laboratory for Child Health of Zhejiang Province, Hangzhou, China; [3]AI Lab, National Clinical Research Center for Child Health, Hangzhou, China; [4]Department of R&D, Hangzhou Healink Technology, Hangzhou, China; [5]College of Computer Science and Technology, Zhejiang University, Hangzhou, China; [6]Research Center of Domestic IT Innovation, Binjiang Institute of Zhejiang University, Hangzhou, China; [7]Polytechnic Institute, Zhejiang University, Hangzhou, China

Contributions: (I) Conception and design: J Li, D Wang; (II) Administrative support: G Yu, Q Shu, J Yin; (III) Provision of study materials or patients: Z Li, J Huang, Z Zhu; (IV) Collection and assembly of data: B Lin, G Qi; (V) Data analysis and interpretation: K Dong, C Shen, B Zhao; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Gang Yu; Qiang Shu. 3333 Binsheng Rd., Hangzhou 310052, China. Email: yugbme@zju.edu.cn; shuqiang@zju.edu.cn; Jianwei Yin. 866 Yuhangtang Rd., Hangzhou 310058, China. Email: zjuyjw@cs.zju.edu.cn.

Background: Multicenter clinical research faces many challenges, including how to quantitatively evaluate the data contribution of each research center. However, few data pricing model meets the requirements to the scenario. Thus, a suitable mechanism to measure the data value for clinical research is required.

Methods: Extensive documents were acquired and analyzed, including a rare disease list from the National Health Commission, data structures of the electronic medical records (EMR) system, diagnosis-related groups (DRGs) regulations from the Health Commission of Zhejiang Province, and the Clinical Service Price List of Zhejiang Province. Nine senior experts were invited as consultants from hospital and enterprises with professional field of clinical research, data governance, and health economics. After brainstorming and expert evaluation, seven data attributes were identified as the main factors affecting the value of medical data. Different weights were assigned for each attribute based on its influence on data value. Each attribute was quantized to an index based on proposed algorithms. The data value models for chronic diseases and other diseases were distinguished given the different sensitivity of data timeliness. A simulation system using blockchain and federated learning techniques was constructed to verify the data pricing model in the scenario of clinical research.

Results: A comprehensive clinical data pricing model is proposed and the simulation of three research centers with 50 million real clinical data entries was conducted to verify its effectiveness. It demonstrates that the proposed model can compute medical data value quantitatively.

Conclusions: Quantitative evaluation of the value of medical data for multicenter clinical research based on the proposed data pricing model works well in simulation. This model will be improved by real-world applications in the near future.

Keywords: Data value; data pricing; multicenter research; blockchain; federated learning

Submitted Jul 04, 2022. Accepted for publication Aug 02, 2022.
doi: 10.21037/atm-22-3671
View this article at: https://dx.doi.org/10.21037/atm-22-3671

^ ORCID: Jing Li, 0000-0002-3626-5815; Guoqiang Qi, 0000-0002-7863-6223; Zheming Li, 0000-0002-6640-9947; Jian Huang, 0000-0002-1955-4316; Zhu Zhu, 0000-0001-8868-2525; Chen Shen, 0000-0001-9187-9345; Bo Lin, 0000-0001-5682-2140; Kexiong Dong, 0000-0002-7899-9032; Gang Yu, 0000-0001-9935-9969.

Introduction

With the emergence of the big data industry, an increasing number of people have realized that data are valuable resources (1). There are more than 10 national data trading centers that are currently operating in China (2). Clinical data are precious resources (3) for multicenter research and medical organizations have developed various medical data platforms (4). However, distinguishing data contributions from different research centers is one of the biggest challenges facing the principal investigator (PI) in clinical research projects (5). Few researches focus on medical data pricing and current model can hardly meets the requirements of clinical researches. Therefore, an appropriate data pricing model is required to systematically estimate the value of clinical data (6).

Three typical pricing models are currently used, including the protocol pricing model based on game theory (7), the third-party pricing model based on data characteristics (8), and the tuple-based pricing model (9). As the most commonly used data pricing mechanism, the protocol pricing model based on game theory agrees on the data price by negotiating between data suppliers and users. Although it is simple to apply, it rarely evaluates the data value precisely due to the different understandings of the data between suppliers and users. For big data trading platforms, such as the Shanghai Data Exchange (10), the Guiyang Big Data Exchange (GBDEx) (11), Azure (12,13), and Datamarket, the third-party pricing model based on data characteristics is the most widely used mechanism. As the third party, data trading platforms calculate the data price according to their data quality evaluation indexes, including quantity, completeness, scarcity, and time span. However, the background of the data exchanges is usually too complex to guarantee their reliability. Furthermore, the prices are often set for datasets rather than tuples, which likely results in waste if the user only requires some tuples of data instead of a whole dataset. The tuple-based pricing model calculates the price for each tuple, given the information entropy, weights, cite index, and expense. However, in complex scenarios, pricing formulas with simple parameters seldom demonstrate the true data value.

Due to the complexity of medical service (14-16), building a related clinical data pricing model that is suitable for most medical organizations is challenging. Experts were invited for the consultation from the fields of clinical research, data governance, and health economics. After several rounds of discussion, the most important medical data attributes were selected by experts. With the selected data attributes, a prototype of a clinical research data pricing model is proposed. The simulation system was developed and the pricing model is simulated with real data.

Methods

Modeling

Extensive documents were acquired and researched, e.g., the electronic medical records (EMR) structure (17), the International Classification of Diseases (ICD) (18), a rare diseases list (19), a chronic disease list, the diagnosis-related groups (DRGs) regulations (20) from the Health Commission of Zhejiang Province, and the Clinical Service Price List of Zhejiang Province. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

As consultants, nine senior experts in different professional fields were invited from hospitals and enterprices, including clinical researchers, data experts, and health economists. All of them are familiar with medical industry and interested in the research area of medical data. Among the various factors in reflexing research value, seven attributes of medical data were selected and agreed on as the most important by senior experts from different professional fields, including clinical researchers, data experts, and financial experts. These attributes are as follows:

- Expense: fees charged for the medical services, drugs, or medical consumables in hospitals, which can be obtained from the EMR homepage. This is generally positively correlated to the value of medical service and related data.
- Scarcity: the rarer clinical case, the more value it likely has. This depends on the diagnosis acquired from the EMR and is negatively correlated to the morbidity of rare diseases and related data.
- Completeness: data quality is one of the most critical factors affecting research results. Data entry without critical information can rarely be used in clinical research. Therefore, the completeness of medical data is positively correlated to its value.
- Timeliness: for clinical research, data timeliness is insensitive to chronic diseases while being sensitive to others. Whether timeliness should be taken into account is dependent on the disease type.
- Hospital level: in general, higher-level hospitals can

provide better medical service and higher quality medical data, both of which are positively correlated to the value of clinical research.

- ❖ Surgery grade: surgery is one of the most important medical services in the hospital. A medical case with high-grade surgery is likely more valuable than those with low grades in clinical research.
- ❖ Doctor post: doctors with a high post can provide high-level medical service and data recorded by them is generally of high quality. The higher the doctor post it has, the more research value it likely has.

Table 1 Attributes and weights

Attributes	Weights (%)
Expense	30
Scarcity	21
Completeness	12
Timeliness	11
Hospital level	10
Surgery grade	9
Doctor post	7
Total	100

Table 2 Index of expense

Expense (RMB)	Index
>100,000	1
50,000–100,000	0.8
10,000–50,000	0.6
1,000–10,000	0.4
<1,000	0.2

The weights of these seven attributes are discussed and assigned in *Table 1*.

The normalized indexes were settled for each attribute as shown in *Tables 2-8*. Five index numbers were assigned for the attributes of expense, scarcity, hospital level, and surgery grade; 12 index numbers for completeness; six index numbers for timeliness; and four index numbers for doctor post.

Finally, data value for the i^{th} clinical research data entry was calculated as follows:

$$\begin{aligned} Value_i = {} & 30\% \times Index_{expense} + 21\% \times Index_{scarcity} \\ & + 12\% \times Index_{completeness} + 11\% \times Index_{timeliness} \\ & + 10\% \times Index_{hospital\ level} + 9\% \times Index_{surgery\ grade} \\ & + 7\% \times Index_{doctor\ post} \end{aligned} \quad [1]$$

For chronic disease, set $Index_{timeliness}$ as 1.

The total value of n data entries is calculated by the summation of each data entry.

$$Value_{total} = \sum_{i=1}^{n} Value_i \quad [2]$$

Simulation

A clinical research alliance with three hospitals was simulated, each of which has different medical datasets. Hospitals can share their data to earn "points" based on the data value model proposed above and need to pay "points" when they use the data of other hospitals.

Our application system is built on Arya privacy computing platform, which integrates with the most advanced technology, i.e., federated learning, multi-party computation, and blockchain, aiming to achieve data interconnection and collaborative computing in a secure

Table 3 Index of scarcity

Morbidity of adults' rare diseases	Morbidity of children's rare diseases	Index
Top 10 rare and strange diseases in the world	Top 10 rare and strange diseases in the world	1.7
<1/1,000,000	<1/500,000	1.6
1/1,000,000	1/500,000	1.5
<1/900,000	<1/400,000	1.4
<1/500,000	<1/10,000	1.0

Table 4 Index of completeness

Completeness	Index
100%	1.0
90–99%	0.9
80–89%	0.8
70–79%	0.7
60–69%	0.6
50–59%	0.5
40–49%	0.4
30–39%	0.3
20–29%	0.2
10–19%	0.1
1–10%	0.05
<1%	0

Table 5 Index of timeliness

Timeliness	Index
<1 year	1
2–4 years	0.8
5–7 years	0.6
7–9 years	0.4
9–11 years	0.2
11–13 years	0.1

Table 6 Index of hospital level

Hospital level	Index
Level 3 grade A	1.0
Level 3 grade B	0.8
Level 2 grade A	0.6
Level 2 grade B	0.4
Level 1	0.2

Table 7 Index of surgery grade

Surgery grade	Index
4	1.0
3	0.8
2	0.6
1	0.4
None	0.2

Table 8 Index of doctor post

Doctor post	Index
Attending	1
Associate attending	0.8
Fellow	0.6
Resident	0.4

way. All data pricing functions and modules can be quickly created with custom configuration (21). First, each hospital decides which data can be shared and uploads the hash number, rather than the medical data, onto the chain. Next, the hospital can check its points and data resource directory in the system. When researchers discover data of interest from other hospitals, they have to pay a certain amount of "points" before using them in clinical research.

Three layers were designed for the simulation system. On the Infrastructure As A Service (IAAS) layer, computing resources such as networking, storage, servers, data security, load balancing, and container were deployed. On the Platform As A Service (PAAS) layer, two function modules were developed: data pricing and alliance chain. In the data pricing module, seven data attributes were considered and their weights were designed to be configurable. In the alliance chain module, numerous functions were designed and developed. The smart contract function facilitates trustable agreements between peers on the chain without a third party, including smart contract uploading, smart contract maintenance, and smart contract details filling. The interface management function takes charge of interactions with other systems. The account book management function guarantees the accuracy of each peer's account book and updates the distributed account book system. The node management function monitors and controls the computing resource consumption of every peer on the chain, such as central processing unit (CPU), random access memory (RAM), storage, etc. On the Software As A Service (SAAS) layer, function modules for multicenter research were established. With the project management function, researchers can initiate research projects, fill out project information, terminate projects, and search for interesting projects. With the data dictionary function, system

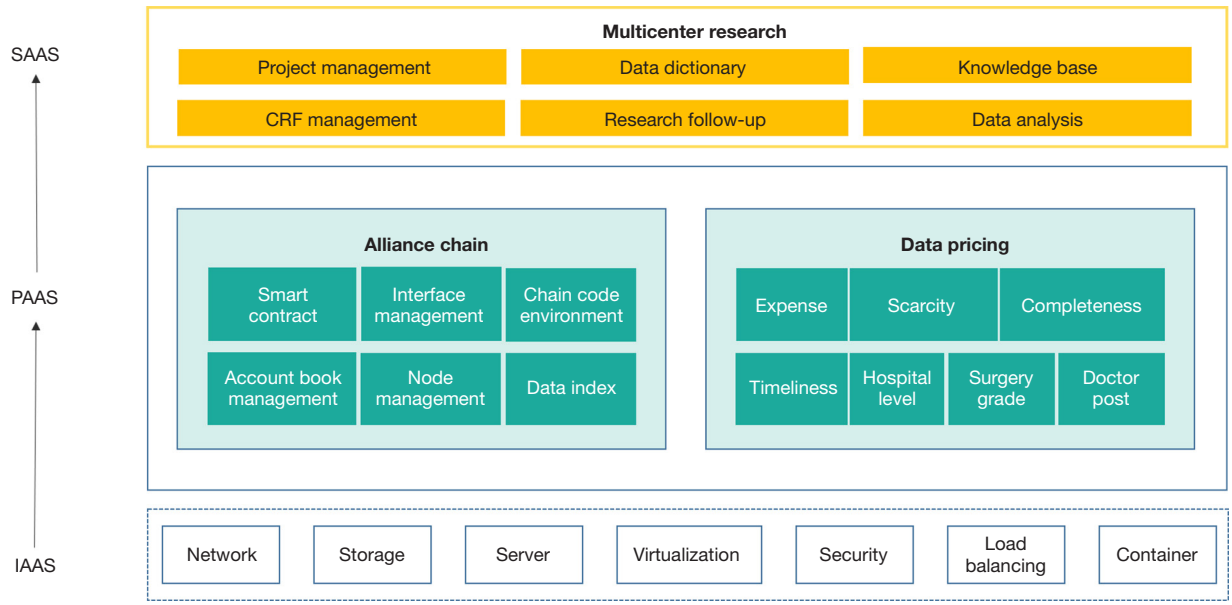

Figure 1 Architecture of the data pricing simulation system. SAAS, Software As A Service; PAAS, Platform As A Service; IAAS, Infrastructure As A Service; CRF, case report form.

administrators can import data dictionary trees, and display, search and edit dictionary elements. Researchers can create case report forms (CRFs) from scratch or use templates with the CRF management function. Using the data analysis function, researchers can perform classical statistics or build artificial intelligence (AI) models. The architecture of the data pricing simulation system is shown in *Figure 1*.

Instead of medical data, Hash numbers from each hospital are uploaded to the alliance chain to protect data security. The pricing model is then uploaded onto the chain. Points for medical data are computed based on the pricing model and the Hash numbers. After verification, the account book of each hospital is updated and distributed through the alliance chain. Finally, relative information is searchable on the chain, which includes the data directory of all alliance members, the pricing model configuration, the points on its account book, etc. The data process steps of the simulation system are shown in *Figure 2*.

Take one data entry of an inpatient's inspection as an example. The hospital level "Level 3 Grade A" is acquired from the hospital level data sheet, which refers to the index of the hospital level as "1.0". Based on the inspection data sheet, data completeness was computed to be 58% (42% of the data entry was blank), which refers to the index of completeness as "0.5". The time of this inspection "2018" (4 years from now) can also be acquired from the inspection data sheet, which refers to the index of timeliness as "0.8". The doctor's name can be obtained from the EMR homepage data sheet, through which the doctor post "fellow" is acquired from the staff data sheet. Surgery grade, disease diagnosis, and inspection expense are acquired from the EMR homepage data sheet, which refers to the related indexes as "0.4", "1.0", and "0.4", respectively. After entering the numbers of these seven indexes into the data pricing model "ESCTHSD", the points for this data entry can be computed. The computation process is shown in *Figure 3*.

Results

Following a 6-month discussion and improvement period, we propose the clinical research data pricing model "ESCTHSD", and its simulation system (based on alliance chain technique) has been developed. The screenshots of the simulation system are shown in Figure S1.

More than 50 million real-world medical data entries were desensitized and divided into three parts to simulate the datasets of the three hospitals. The number of data entries and points computed are listed in *Table 9*.

The simulation results demonstrate that the "ESCTHSD" data pricing model can properly evaluate the data value from simulated medical research centers. The seven attributes (expense, scarcity, completeness, timeliness, hospital level, surgery grade, and doctor post) of medical

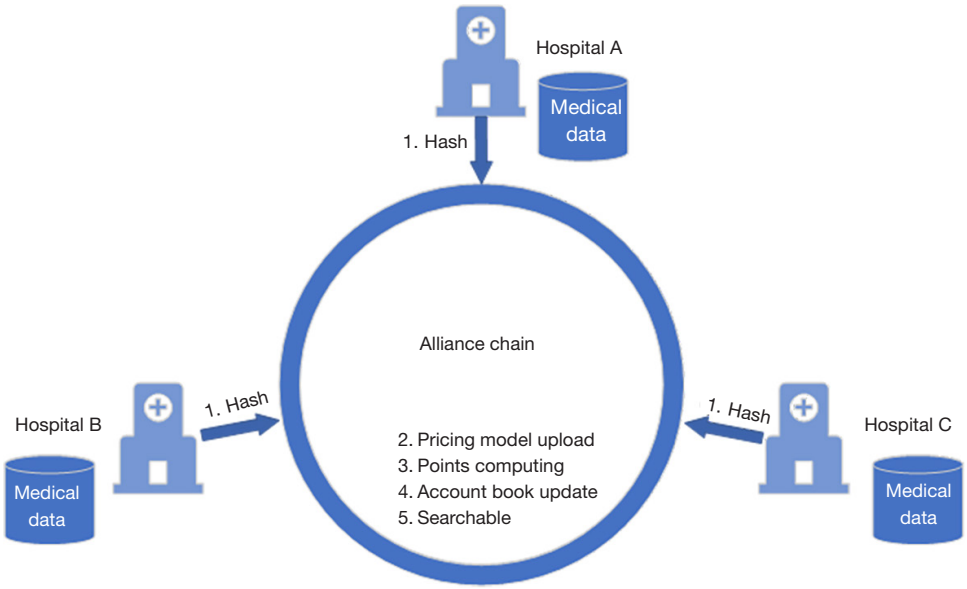

Figure 2 Data process steps of the simulation system.

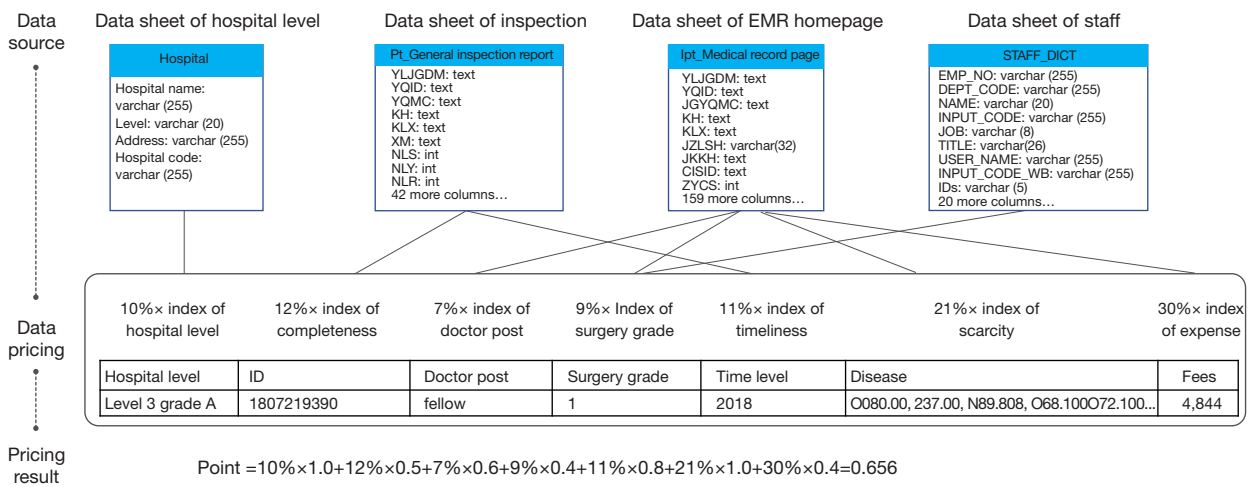

Figure 3 Value computation example of a single data entry. EMR, electronic medical records.

Table 9 Simulated data & points

Hospital	Data entries	Points
A	30,000,008	26,715,861.62
B	12,300,000	9,944,785.74
C	8,070,000	5,581,746.03
Total	50,370,008	42,242,393.39

data were well-considered and quantitatively computed.

It is also clear that the alliance chain is a promising technique for clinical research, which are especially data-driven. The simulation system effectively protects data privacy and performs robustly and smoothly with big data.

Discussion

This study proposes a new data pricing model to evaluate

the value of medical data for multicenter clinical research. The proposed model consists of seven data attributes, namely, expense, scarcity, completeness, timeliness, hospital level, surgery grade, and doctor post. Each of these attributes can be transformed into a numerical index according to a certain rule. The data value is computed for every data entry based on the weighted summation of the seven attributes. The simulation system, based on the alliance chain technique, was developed with the IAAS, PAAS, and SAAS layers. Over 50 million data entries were divided into three datasets to simulate clinical research in three hospitals. The data value of each data entry and the total value of each hospital were computed.

The "ESCTHSD" medical data pricing model and alliance chain-based simulation system work well in the lab environment for historical datasets. However, for real clinical research projects, the number of research centers and the volume of research data could be larger, and more complex cases may occur. The clinical data from different medical organizations in different regions will be simulated in the system to verify the robustness of the proposed "ESCTHSD" model.

In the future, the data pricing model will be assessed by the third party, tested in the real-world scenario of multicenter research and will be modified if necessary. More attributes of medical data can be considered besides the seven included in the "ESCTHSD" model. Instead of fixed numbers, dynamic weights of data attributes will be assigned for different attributes to increase the generalizability of the pricing model. Also, more complicated models will be researched and compared rather than linear functions, including higher-order functions and machine learning models. However, managing incomplete data presents another challenge. Natural language processing (NLP) and machine learning techniques will be adapted to increase the quality of clinical research data.

Moreover, the technical structure of the system will be optimized to increase system efficiency and decrease computing resource consumption. More functions for clinical research will also be developed in the system. Furthermore, the federated learning module will be developed to increase the data security level, which contains the functions of distributed computing engine, distributed storage engine, federated modeling process, visualized modeling, networking proxy, service proxy, service cooperation, trusted computing, task dashboard, and operation monitoring. Different research centers are able to share information and build data models without actually exchanging their local data. An AI and statistics modeling toolbox will be developed as well. Various AI and statistic algorithms will be embedded, e.g., isomorphic Poisson regression, isomorphic multilayer perceptron, isomorphic logistic regression, heterogeneous logistic regression, isomorphic multiple linear regression, mean value, variance, and standard deviation. The alliance chain module will be enhanced to orient real cases of data-driven research among clinical research centers. The Certificate Authority (CA) system will also be added to manage the authority of every node on the alliance chain.

Based on the system architecture, more application scenarios can be integrated in the future, such as clinical quality management (CQM), clinical decision support system (CDSS), multi-disciplinary treatment (MDT), special disease reporting, and pharmaceutical development.

Conclusions

In this study, a new data pricing model "ESCTHSD" is proposed for multicenter research. The simulation system of three virtual hospitals with 50 million data entries was designed and developed, which proved the feasibility of the quantitative evaluation of medical data value for multicenter clinical research. This provides a solid foundation for real-world applications in the near future.

Acknowledgments

Funding: This work was supported by the National Key R&D Program of China (No. 2019YFE0126200), the National Natural Science Foundation of China (No. 62076218), and the Zhejiang Province Research Project of Public Welfare Technology Application (Nos. LGF22H180004 and LGF22H110001).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-22-3671/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as

revised in 2013).

References

1. Pei J, Zhu F, Cong Z, et al. Data Pricing and Data Asset Governance in the AI Era. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021:4058-9.
2. Peng H, Zhou Y. Data Pricing Mechanism Status and Development Trends. Journal of Beijing University of Posts and Telecommunications 2019;42:120-5.
3. Telenti A, Jiang X. Treating medical data as a durable asset. Nat Genet 2020;52:1005-10.
4. Li J, Yu G, Ding W, et al. Data governance system of the National Clinical Research Center for Child Health in China. Transl Pediatr 2021;10:1905-13.
5. Topol EJ. Welcoming new guidelines for AI clinical research. Nat Med 2020;26:1318-20.
6. Shen Y, Guo B, Shen Y, et al. Personal big data pricing method based on differential privacy. Computers & Security 2022;113:102529.
7. Riahi Sfar A, Challal Y, Moyal P, et al. A Game Theoretic Approach for Privacy Preserving Model in IoT-Based Transportation. IEEE trans Intell Transp Syst 2019;20:4405-14.
8. Tan B, Anderson EG Jr, Parker GG. Platform Pricing and Investment to Drive Third-Party Value Creation in Two-Sided Networks. Information Systems Research 2020;31:217-39.
9. Pauliuk S, Heeren N, Hasan MM, et al. A general data model for socioeconomic metabolism and its implementation in an industrial ecology data commons prototype. J Ind Ecol 2019;23:1016-27.
10. Tang Q, Shao Z, Huang L, et al. Identifying Influencing Factors for Data Transactions: A Case Study from Shanghai Data Exchange. J Syst Sci Syst Eng 2020;29:697-708.
11. Li Q. Analysis of the Influence Path of Guiyang Big Data Expo on Regional Economic Developments. 2020 International Conference on New Energy Technology and Industrial Development (NETID 2020), 2021.
12. Verma A, Malla D, Choudhary AK, et al. A Detailed Study of Azure Platform & Its Cognitive Services. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019:129-34.
13. Copeland M, Soh J, Puca A, et al. Microsoft Azure: Planning, Deploying, and Managing Your Data Center in the Cloud. Berkeley, CA: Apress, 2015.
14. Cook SC, Desmarais AG, Berry JG. Room to Improve Prior Authorization in Children With Complex Medical Needs. Pediatrics 2022;149:e2021054843.
15. Klemm P, Müller-Ladner U, Lange U. Multimodal rheumatological complex treatment : A current inventory. Z Rheumatol 2022;81:369-75.
16. Lindskou TA, Pilgaard L, Søvsø MB, et al. Symptom, diagnosis and mortality among respiratory emergency medical service patients. PLoS One 2019;14:e0213145.
17. Setyawan R, Hidayanto AN, Sensuse DI, et al. Data Integration and Interoperability Problems of HL7 FHIR Implementation and Potential Solutions: A Systematic Literature Review. 2021 5th International Conference on Informatics and Computational Sciences (ICICoS), 2021:293-8.
18. Nicholas M, Vlaeyen JWS, Rief W, et al. The IASP classification of chronic pain for ICD-11: chronic primary pain. Pain 2019;160:28-37.
19. Ferreira CR. The burden of rare diseases. Am J Med Genet A 2019;179:885-92.
20. Meng Z, Hui W, Cai Y, et al. The effects of DRGs-based payment compared with cost-based payment on inpatient healthcare utilization: A systematic review and meta-analysis. Health Policy 2020;124:359-67.
21. Tian H, He J, Ding Y. Medical Data Management on Blockchain with Privacy. J Med Syst 2019;43:26.

(English Language Editor: A. Kassem)

Cite this article as: Li J, Wang D, Qi G, Li Z, Huang J, Zhu Z, Shen C, Lin B, Dong K, Zhao B, Shu Q, Yin J, Yu G. Alliance chain-based simulation on a new clinical research data pricing model. Ann Transl Med 2022;10(15):836. doi: 10.21037/atm-22-3671