Cloud Computing and Data Science
https://ojs.wiserpub.com/index.php/CCDS/

UNIVERSAL WISER PUBLISHER

Research Article

Applying Latest Data Science Technology in Cancer Screening Programs

Lian Wen[1*], Wuqi Qiu[2], Kedian Mu[3]

[1]Institute for Integrated and Intelligent Systems, School of Information and Communication Technology, Griffith University, Brisbane, Australia
[2]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China
[3]School of Mathematical Sciences, Peking University, Beijing, China
E-mail: wen@griffith.edu.au

Received: 27 May 2020; **Revised:** 30 June 2020; **Accepted:** 30 June 2020

Abstract: Cancer screening programs have been implemented in many different countries for many years to collect information of the fatal diseases, to provide early diagnosis, to support medical research, and to help governments making policies. However, few of those programs have utilized latest data science technologies, therefore not be able to generate the maximum benefits from those programs. To overcome this problem and improve the quality of cancer screening programs, this paper firstly (i) reviews the typical architecture and IT technologies used in current screening programs and recognizes their limitations; then (ii) introduces recent developments in data science that could be implemented in screening programs; finally (iii) proposes the structure of general medical screening framework (GMSF), which could be developed to host future cancer screening programs that will advance data coverage, data accuracy, data usage and lower in the costs. The structure of GMSF and its key elements are described in this paper and some practical approaches to build GMSF are discussed. This work might initialize a series or research to bring the latest IT technologies, particularly data science technologies, into cancer screening programs, and significantly increase the efficiency and reduce the cost of future cancer screening programs.

Keywords: cancer screening program, data science, big data, artificial intelligence, machine learning, medical informatics, ontology

1. Introduction

Due to the high mortality of cancer [1] (16.0/million in urban China [2]), many cancer screening programs have been implemented in different regions through different medical organizations [3-5]. For example, the Cancer Screening Program in Urban China (CanSPUC) has been implemented in 14 provinces to provide cancer screening for lung, breast, colorectal, esophageal, gastric and liver cancers [6] from 2012. The benefits of those cancer screening programs have been recognized, and the cost-utility analysis has been conducted on some programs to justify the value and help governments to make decisions to provide most cost-effective screening programs [7].

However, many of the cancer screening programs have been designed and implemented without considering using latest development of Information and Communications Technology (ICT), particularly big data [8] and machine learning [9-10] even though people realize that big data and machine learning are closely relevant to modern health care [11-12]. The lack of using latest ICT technologies might not only affect the quality of those programs, increase their

DOI: https://doi.org/10.37256/ccds.112020445

cost, but also, most importantly limit their values.

In the last few decades, ICT has made significant progress. Big data, cloud computing and artificial intelligence are among the pioneers in ICT to change the world [13]. Because cancer screening programs are very complex activities; they involve huge amount of data collection, processing and analysis, therefore, it would be one of the most suitable areas to implement latest technologies of data science and machine learning.

Nevertheless, to adapt latest ICT technologies into cancer screening programs is not an easy task, because it crosses disciplines and involves many kinds of stakeholders including governments, legislation bodies, medicine institutes and ICT service providers. In order to address this issue, a suitable framework to reach a common understanding and recognition among different stakeholders is essential.

In order to achieve that, this paper firstly reviews the architecture and techniques used in many current cancer screening programs and shows their problems and limitations, and then briefly introduces some latest development in data science and explains their possible applications on cancer screening programs. After that, this paper proposes a General Medical Screening Framework (GMSF) with the core of general medical screening system (GMSS) that would arguably help to design and implement new cancer screening programs with higher quality and more economic and scientific benefits. Some practical issues related to GMSF have also been addressed.

The rest of this paper is structured in the following way: Section 2 reviews the structure of current screening programs and discusses its limitation. Section 3 introduces some latest data science technologies that could be used in cancer screening programs. Section 4 presents the proposed GMSF, and Section 5 discusses some practical issues related to GMSF. Finally, a conclusion and future research is given in Section 6.

2. A typical structure of existing screening programs and its limitations

Successful implementation of cancer screening programs is challenging and requires significant clinical effort for as-yet uncertain patient benefi [14]. According to a survey [10], few cancer screening programs have utilized advanced ICT technologies. Even though each individual screening program may have its specialties, a simplified structure and process model as shown in Figure 1 is fit to many of existing cancer screening programs [15].

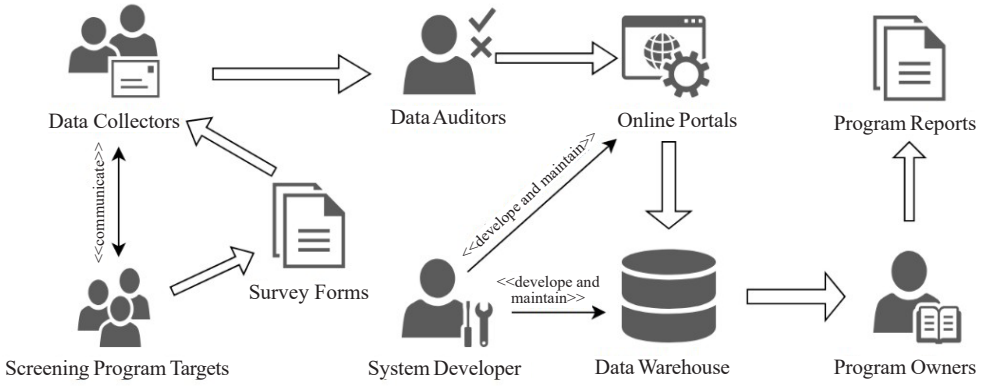

Figure 1. A typical structure for existing screening programs

Figure 1 shows the structure and process of typical cancer screening programs. The figure contains five types of stakeholders, while each type includes a group of people; some individual persons could belong to different types of stakeholders.

• **Program owners:** Who are the designers and owners of this program. They are responsible to determine the target screening population and the survey questions. They are also responsible to process and interpret the survey data and create final survey reports.

• **System developers:** Who are Information technology (IT) professionals and they are responsible to develop and

maintain an information system to store the survey data. The information system may contain a backend data warehouse and a frontend user interface. All the survey data will be stored in the backend data warehouse. The frontend user interface may have an online portal, through which all the row data are inputted.

• **Data auditors:** Who are responsible to check the quality of the raw data, filter out incomplete and invalidate data, and input the valid data into the system.

• **Data collectors:** The number of data collectors is determined by the scale of the screening program. It could be huge, for example, from 2012 to 2016, the CanSPUC covered 3.5 million urban residents, which has involved about 1000 hospitals and community health centres to collect data [6].

• **Screening program targets:** Who are the target population of the screening program.

Figure 1 also shows the main process of a screen program. The broad arrows display the flow of the survey data. Firstly, the data collectors communicate with the target people and send the survey forms to them. After those survey forms are filled they are sent back to the data collectors. The data collectors perform an initial checking and then submit the forms to the data auditors. The auditors perform a more thorough checking of the data and then input the valid data into the system. After all the survey data have been inputted into the system, the program owners evaluate the data and use them to create the final reports.

Of course, Figure 1 shows only a simplified structure of typical screening programs. A real screening program might be different in some details, but the overall designs are quite similar. Even though this structure has been implemented in many screening programs and is proved to be feasible, but it shows following shortcomings:

Firstly, the value of the data can hardly be fully utilized due to the limitation of their accessibility. The most important asset generated from a screening program is the data collected through the program. However, due to the sensitivity of the data, the data are usually stored privately that can only be accessed by the program owners. Even for the program owners, after they have completed their due reports of the program, without a continuous maintenance, they might find difficult to access those data after some time. Third-party access to the raw data is extremely difficult due to no suitable policy and regulation and the lack of technique support. Therefore, after the program owners have completed their program reports, the raw data will be buried in some corners that can hardly be used by any other researchers.

Secondly, the safety and security of the data can be an issue. As cancer screening programs are developed and implemented by different levels and types of medical research institutes. Each of them may have their own IT support teams. However, the technique used by their IT teams and the quality of their infrastructure are diverse, it is impossible to guarantee a consistent level of data integrity, confidentiality and availability.

Thirdly, the collected data may not be in high quality. Data quality of a screening program is determined by many factors including the design of the questionnaires in the survey form, the sample size, the identification of the relevant target population, the completeness and accuracy of the survey forms completed by the target people, the amount of data input mistakes etc. If there are no national or international guidelines to regulate the entire process with the consideration of long-term usages of the collected data, the understanding of the capacity of latest data science technologies, and the utilizing those technologies, the collected data could be in poor quality.

Generally, various cancer screening programs have been proved to be useful to monitor the health status of target people, provide strong support for scientific research, and improve the life quality and health for large populations [3, 5]. However, the lack of understanding and using of latest IT techniques will significantly reduce the values and impacts of those screening programs [10].

3. Latest development of data science technologies and software engineering

With the quick development of hardware infrastructure and computer science in the last few decades, data science and software engineering have immense impacts on nearly all aspects on human life. This section briefly introduces a few technologies that could be adapted and utilized in future cancer screening programs.

3.1 *Data volume*

While traditional database technology can usually handle tables about hundreds of columns and thousands of rows, current software systems are capable to process millions or even billions of rows and millions of columns [16].

The dramatic increase of data volume capacity in software systems enable new screening programs to collect data from real-time wearable devices and use unstructured data such as images and videos. It also improve the programs' capability to handle much larger target population, therefore to increase the quality of the programs.

3.2 *Data structure*

Traditional surveys usually store the collect data such as age, weight, health ratings in structured forms like relational databases or spreadsheets, and textural data with discernible patterns such as the description of a patient's symptom in semi-structured forms. While latest development in data science introduce techniques and tools to process quasi-structured data, that means textual data with erratic data, and even unstructured data including text document, PDF files, images and videos [16].

With the introduction of quasi-structured and unstructured data in healthcare [17], a screen program would be able to collect much broader range of data, to provide a more accurate and comprehensive picture of the medical situation, and eventually to produce better support for later medical research and policy making.

3.3 *Data lifecycle*

In traditional screening programs, the life of the raw data is usually finished after the program owners complete their reports. No other people will use the data, and they may be deleted after a period. It could be a great loss as historical data could be critical for future researches. Therefore, it is of great importance to increase the data lifespan. In the traditional approach, the budget may only cover the development of the system, while it may not have the extra budget to maintain the hardware infrastructure and the supporting technicians when the project is closed. Development of big data dramatically reduces the cost of keeping large amounts of data alive [15], therefore it becomes feasible to extend the lifespan of the previously collected data practically forever.

3.4 *Data collection*

Most current cancer screening programs still use the traditional approaches to collect data. Those approaches include asking survey targets to fill in survey forms, collect their medical samples such as blood, urine etc., and/or invite them to undertake some medical examinations. However, with the quick development of wearable sensors [18], it is possible to collect more accurate and objective information [19].

For example, to collect the data of a patient's sleep quality, the traditional approach could be to ask them to answer a questionnaire or to rate their sleep quality. Data collected in this approach could be subjective and false due to individual's interpretation of the questions. However, some wearable device can monitor and record a patient's sleep pattern, and provide much more accurate and objective data. This kind of technology would significantly increase the amount and accuracy of data collected from the targets and improve the data quality.

3.5 *Data storage*

The traditional way to store structured data is to save data in relational databases which are usually hosted in dedicated hardware servers. The safety and security of the data is affected by the hardware infrastructure and software settings. Now with the increasing data volume and data security requirements, data could be stored in distributed computing and cloud environments with blockchain provided by international service providers [20]. They regularly update their hardware and software settings to make the data safe and secure. They also provide routine data backup service, so the data safety and security can be maintained in a higher and more consistent level compared to traditional approach while the medical institutes have to setup their own data storage environments.

3.6 *Data access*

Traditional centralized data warehouse approach limits the usage of the data to only the data owners. Other parties have neither the privilege nor the essential technical details of the database design to access the data. Therefore, it

reduces the potential values of the collected data.

However open data approach can improve healthcare [21]. New development in data science including open data and analytic sandbox [22] make it possible for the collected data to be used by other medical research institutes. Of course, regulations are required to decide which data can be released to public.

3.7 *Data verification and reasoning*

Data collected in traditional screening programs is only processed through some basic statistics tools. The information retrieved from data is usually limited as percentage, average, and some correlation relationships.

Data quality is also essential for screening programs. To reach correct results, data auditors need to make sure that the data are robust and reliable [23], but the traditional data verification is usually based on human inspection and some simple data rules that are very time consuming and might not capture subtle issues in the data.

Recent data science development has been heavily affected by the technique in artificial intelligence including ontology [24], machine learning [25], and data mining [26] etc. The new technique will help to process more data and retrieve information which is impossible to be discovered through traditional methods. The new technologies can also verify the reliability and validity of the data, integrate and compare data from different sources, therefore provide much richer services than before.

3.8 *Software engineering and software tools*

To develop complex software intensive systems, best practices in Software Engineering must be applied. Latest software engineering technology like Behavior Engineering [27] could be implemented to develop large and complex software intensive system with high quality and low cost [28-30]. At the same time, system development standards [31] are encouraged to be used to guild the system development process. To process huge volumes of data, suitable software tools and environments are necessary. A software environment is a working environment that contains several relevant software tools being integrated together and configured properly so that certain types of complex tasks can be performed efficiently.

There are many tools [32] used for big data projects that could be explored and used for future cancer screening programs. Those tools include Hadoop, Alpha Miner, OpenRefine, Data Wrangler for data preparation, R, SQL Analysis services, SPSS Modeler, STATISTICA, Octave and SAS/ACCESS for data modelling.

3.9 *New roles*

The traditional screening programs may contain only one general IT technician role that will take the responsibility in developing the software system, design the database, and help to process the data to get some statistics results. However, with the quick development of data science, it may require people with more specified expertise for big data projects. The general IT support role could be split into three more specific roles:

The first role for technology and data enablers as they are closer to the traditional IT support people, who are responsible to developing and maintaining the software system and databases. The second role for data savvy professionals who have basic knowledge of statics or machine learning can define key questions that can be answered using advanced analytics. The third role for deep analytical talent as those people have strong analytical skills, they are capable to handle raw unstructured data and apply complex analytical techniques at very large scales [33].

As new roles with strong technique skills are involved, general medicine institutes may need to cooperate with IT companies in data science to produce most efficient and cost-effective screening programs.

4. General medical screening framework (GMSF)

Considering current situation of cancer screening programs and the potential development in the future, this paper proposes a general medical screen framework that aims to be an embryonic concept for future cancer screening programs. The structure of GMSF is shown in Figure 2.

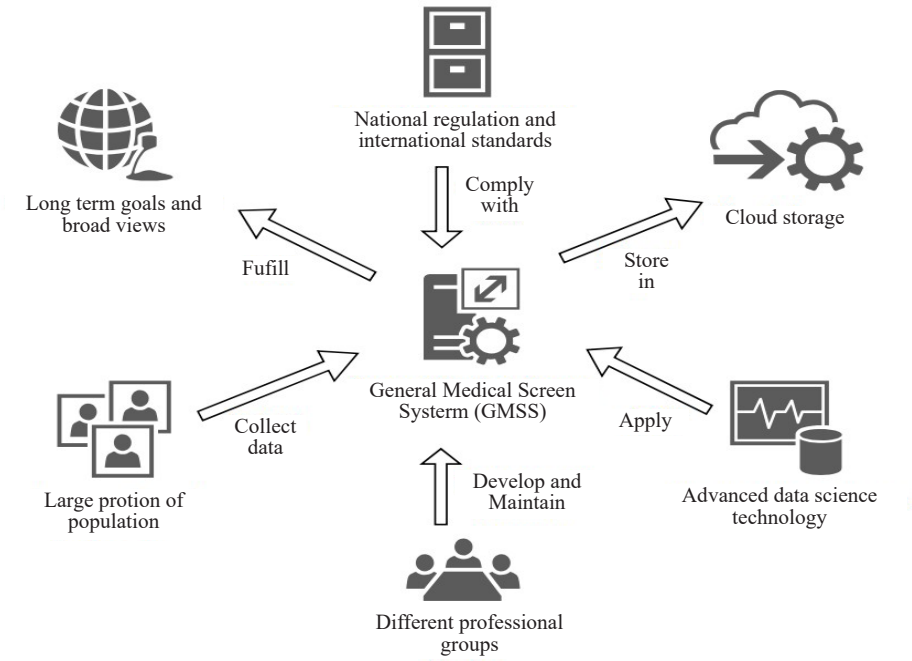

Figure 2. The structure of GMSF

The core of GMSF is a system called general medical screen system (GMSS) that needs to be designed and implemented in the future for different cancer screening programs. However, rather than to dig into the detailed internal design of the core, which will be further explored in future research, this paper focuses on its relationships with other key elements in the framework.

Besides GMSS, GMSF contains other six external elements, which are directly connected to GMSS. We will discuss those six elements and their relationship with the core one by one in the rest of this section.

4.1 *Vision and goals*

In order to make cancer screening programs successful and to harvest the most fruitful results, the authorities who are going to conduct a program should aim high and have broader visions and long-term goals, rather than only considering collecting data for a one-off usage.

They should consider that each individual screening program to be an important block of a massive health informatics building which contains all the data and knowledge of cancers, and it provides an ultimate foundation for human to eventually fully conquer this kind of diseases. They should not limit their concerns only to a small scale, but to have long-term goals in mind; that means the data collected from one single screening program can be integrated with other programs to generate a comprehensive data network that will help medical research to solve more complex and difficult problems.

For one single screening program, the possible extensions have many dimensions such as time dimension: combing data with historical data to cross long period, geography dimension, sharing and combing data in national wide and/or international wide, and domain dimension, combing data of different cancers and/or even other relevant data. To have the right version and goals will help to align the aim of GMSS development.

4.2 *Policy, regulation and international standards*

As a proven practice to improve people's general health and increase cancer patients' life quality, cancer screening programs have been implemented in many different countries and some of them have already made national level guidelines or regulations [2-3]. However, more detailed reginal and national policy and regulations are required to have a comprehensive coverage of all cancers, especially with the trend of globalization. As cancer research are cross country

activities, it is expected that in the future, to maximize the benefits of cancer screening programs, international standards could be developed to provide guidelines for all cancer screening programs. That will make every piece of collected data to be a useful part for global cancer data network. This practice is common in other disciplines, for example, Software Engineering. Many international standards [31] have been published to guild the development of software and system development.

In the future, individual cancer screening programs are still organized by different levels of institutes, but with the guideline of national regulations and international standards, different programs will reach the same quality level and can be integrated seamlessly.

4.3 *Data storage*

One essential element of all cancer screening programs is the data storage. In order to handle the demand of the future programs, the data storage must be safe, so the data can be kept without being lost and damaged; secure, so the sensitive section of the data will not be released to unauthorized users; accessible, so the legal users can retrieve and use the data easily; extensible, so it can store data that rapidly increases in volume and varies in format and structure.

Those challenging demands of the data storage make it very costly for individual medical institutes to setup their own data storage infrastructure. A feasible solution is to use leading commercial service providers who have implemented latest technologies including cloud computing and blockchains.

4.4 *Data science technology*

Apart from traditional statistics techniques, latest development of data science provides much more powerful methods to discover the secrets from collected data. Particularly the machine learning has potential that is far beyond the capacity of traditional statistics.

The GMSF makes the data science technology as a separate element independent to the data. That means with the emergence of new data analytics technology, the data collected could be revisited again to support new research beyond the purpose of the original screening programs.

4.5 *Professional groups*

Current cancer screening program teams may contain roles related to IT support and data analytics, but their tasks may not be specified in detail. However, to apply advanced data analytics technology and utilize latest infrastructure and software tools, the future screening programs may require more complex project team with specialized professional roles including business intelligence analyst, data engineer and data scientist etc.

4.6 *Program coverage*

Due to the high cost of data collection and challenges of data management and processing, the traditional cancer screening programs usually cover relatively small portion of population in a district and might not be able to identify the same individuals among different programs. It is usually difficult to connect a target person's data collected from a screening program with this person's other medical records. This limitation will affect the usefulness of the screening program.

For example, if we have identified that through a screening program, people who developed cancer A usually demonstrate physical abnormity B. However, due to the lack of long-term continuous medical records, it would be difficult to conclude whether cancer A causes abnormity B, or abnormity B causes cancer A.

In the future, the cancer screening programs will cover much larger population; and eventually the entire population will be covered by different screening programs. Furthermore, as the connection of the entire medical data network, it is possible to retrieve an individual's life-time medical records from different screening programs, other medical examinations, and even relevant data such as daily exercises and sleep patterns recorded by wearable device, therefore, providing a much more comprehensive service for society.

4.7 *Summary*

The proposed GMSF is a complex system which involves many different elements. The core element GMSS requires the support from all the other external elements. Among them the advanced technologies in data science and cloud computing are already available. The most urgent tasks at current stage would be to form broad views and set up long-term goals. With the correct views and goals, the next step would be to establish national regulations and international standards. While those elements are gradually becoming mature, other elements will emerge correspondingly.

5. Discussion

Previous section has introduced the proposed GMSF which aims to setup a framework for future cancer screening programs that will have more complete coverage and provide more powerful and comprehensive services than current programs. This section discusses some practical issues.

5.1 *Feasibility*

From the structure of GMSF in Figure 2, we can see that apart from the core element GMSS, only two out of the six external elements are in the technology sector, while the rest four external elements are in the human and management sector.

The elements in the technology sector are designed based on recent developments in data science and the services provided by leading cloud computing companies. They are already available, therefore, from the technique point of view, GMSF is feasible.

Compared to the technique aspects, the human aspects could impose more challenges. However, due to the common recognition of the severity of cancers and their impacts on human life, the challenges to tackle the worst human killers, more research fund and international cooperation have been established. Also, with the inevitable trend of globalization, it is possible that national and international bodies are formed to promote the cooperation and unified research of cancers and to build GMSF for the benefit of the entire human race.

5.2 *Risks*

The risk analysis can be classified in three categories: technique risks, project risks, and security risks. From engineering's point of view, with the help of modern software engineering, many systems with similar scale and complexity have already been developed. That does not mean there are no technique risks and project risks for GMSF, but we would say that the similar kinds of risks have already been well studied and understood, so with the careful implementation of latest risk management practices, those risks could be controlled.

Compared to traditional cancer screening programs, GMSF may create special security risks. Because the data collected in traditional screening programs is only accessible to the program owners, the data is more secure compared to the screening data stored in GMSF that could be accessible by other research groups. Of course, all the legitimate accesses to any part of the data are through defined processes, under proper authority and privilege, and assented by all the relevant parties based on the regulation. The risk issues are illegal access and data leak. However, compared to financial and military data, cancer screening data is less sensitive and the leak of some piece of data will cause much less damage. Therefore, compared to those projects, the security risks for GMSF are controllable.

5.3 *Practical approaches*

GMSF is a huge project and it requires large-scale international cooperation to complete. This paper only proposes some of the key concepts. Apart from continuous academic research to finalize the details of the framework, in order to speed up the achievement of the final goal, there are two practical approaches.

The first approach is bottom up. That means when people start to design new screening programs, rather than repeat

what they have done in the previous programs, they could gradually improve their designs based on GMSF, therefore, to make the new screening programs more compatible to GMSF. There are following recommendations that can be applied to improve the programs:

- To change the data storage from centralized local data warehouses to cloud storage.
- To include unstructured data in the data collection.
- To recruit data scientists and software engineers in their project team.
- To classify the data and assign different levels of access.
- To add open data access interfaces.
- To integrate new data processing algorithms particularly machine learning algorithms in the screening program to enhance the data quality and data usage.
- To comply with relevant regulations and standards in industry. Above is a list of recommendations, but of course it is not exhaustive.

The second approach is top down. That means high level authorities could start to legislate and regulate screening programs based on advice from data scientists and promote cooperation across regions and countries. This approach requires national level or even international level collaborations.

5.4 *Legacy systems*

With the realization of GMSF, new screening programs will generate more useful information and be more helpful for medical research. They will even be more influential in helping governments to make medical related decisions. However, the values of existing screening programs are not replaceable. Therefore, some projects will be setup to migrate the data collected from old screening programs into GMSF.

6. Conclusions and further research topics

Cancer screening is a powerful tool to fight cancers. However, current practices in cancer screening programs haven't implemented latest technologies derived from recent development in data science. It greatly limits the capability of the data collected from those programs. This paper reviews latest technologies in data science that could be integrated in cancer screening programs to increase their quality, efficiency and reduce cost at the same time. Furthermore, it proposes GMSF, a general medical screening framework as a guideline to direct the research and development of future screening programs. It also discusses some practical issues of GMSF.

This paper has introduced the overall structure of GMSF and the functions of its key elements. The future research would focus on details of individual elements, particularly the core element of GMSS.

Declarations

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Griffith University and Peking University Collaborative Research Scheme 2017 and Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (2017-I2M-1-006).

Conflict of interest

The authors declare no competing financial interest.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN

estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. 2018; 68(6): 394-424.
[2] National Health and Family Planning Commission. *China's Health and Family Planning Statistical Yearbook*. Peking Union Medical College Publishing; 2017.
[3] Smith RA, Andrews KS, Brooks D, Fedewa SA, Manassaram-Baptiste D, Saslow D, et al. Cancer screening in the United States, 2017: A review of current American Cancer Society guidelines and current issues in cancer screening. *CA: A Cancer Journal for Clinicians*. 2017; 67: 100-121.
[4] He E, Lew JB, Egger S, Banks E, Ward RL, Beral V, et al. Factors associated with participation in colorectal cancer screening in Australia: Results from the 45 and Up Study cohort. *Preventive Medicine*. 2018; 106: 185-193.
[5] Warnakulasuriya S, Fennell N, Diz P, Seoane J, Rapidis A. LDV Lifelong Learning Programme. An appraisal of oral cancer and pre-cancer screening programmes in Europe: A systematic review. *Journal of Oral Pathology & Medicine*. 2015; 44(8): 559-570.
[6] Dai M, Shi JF, Li N. Design and expected goals of the cancer screening program in urban China. *Chinese Preventive Medicine*. 2013; 47(2): 179-182.
[7] Zong X, Wu J. Basic knowledge of QALYs. *China Journal of Pharmaceutical Economics*. 2009; 5: 77-85.
[8] Yang C, Huang Q, Li Z, Liu K, Hu F. Big data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*. 2017; 10(1): 13-53.
[9] Alpaydin E. *Introduction to Machine Learning*. 4th ed. MIT Publishing; 2020.
[10] Sun N. *Informed Aging: Information and Communication Technology Usage and Cancer Screening Beliefs*. Oxford, Ohio; 2018.
[11] Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018; 319(13): 1317-1318.
[12] Kaur P, Sharma M, Mittal M. Big data and machine learning based secure healthcare framework. *Procedia Computer Science*. 2018; 132: 1049-1059.
[13] Hankel A, Heimeriks G, Lago P. A systematic literature review of the factors of influence on the environmental impact of ICT. *Technologies*. 2018; 6(3): 85.
[14] Wood DE. National comprehensive cancer network (NCCN) clinical practice guidelines for lung cancer screening. *Thoracic Surgery Clinics*. 2015; 25(2): 185-197.
[15] UR Rehman MH, Chang V, Batool A, Wah TY. Big data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management*. 2016; 36(6): 917-928.
[16] Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, et al. Big data and its technical challenges. *Communications of the ACM*. 2014; 57(7): 86-94.
[17] Raghupathi W, Raghupathi V. Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*. 2014; 2: 3.
[18] Billinghurst M, Starner T. Wearable devices: New ways to manage information. *Computer*. 1999; 32(1): 57-64.
[19] Zheng YL, Ding XR, Poon CCY, Lo BPL, Zhang H, Zhou XL, et al. Unobtrusive sensing and wearable devices for health informatics. *IEEE Transactions on Biomedical Engineering*. 2014; 61(5): 1538-1554.
[20] Xia QI, Sifah EB, Asamoah KO, Gao J, Du X, Guizani M. MeDShare: Trust-less medical data sharing among cloud service providers via blockchain. *IEEE Access*. 2017; 5: 14757-14767.
[21] Kostkova P, Brewer H, De Lusignan S, Fottrell E, Goldacre B, Hart G, et al. Who owns the data? Open data for healthcare. *Frontiers in Public Health*. 2016; 4: 7.
[22] Phillips-Wren G, Iyer L, Kulkarni U, Ariyachandra T. Business analytics in the context of big data: A roadmap for research. *Communications of the AIS*. 2015; 37: 23.
[23] Van der Werf LR, Voeten SC, Van Loe CMM, Karthaus EG, Wouters MWJM, Prins HA. Data verification of nationwide clinical quality registries. *BJS Open*. 2019; 3(6): 857-864.
[24] Gawich, M, Alfonse M, Aref MM, Salem AM. Developing a system for medical ontology evolution. *Egyptian Computer Science Journal.* 2017; 41(2): 53-62.
[25] Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. *Radiographics*. 2017; 37(2): 505-515.
[26] Arslan AK, Colak C, Sarihan ME. Different medical data mining approaches based prediction of ischemic stroke. *Comput Methods Programs Biomed*. 2016; 130: 87-92.
[27] Wen L. *Behavior Engineering World*. 2012. Available from: http://www.beworld.org/BE/home/be-resources/ [Accessed 25 June 2020].
[28] Wen L, Rout T. Using composition trees to validate an entry profile of software engineering lifecycle profiles for very small entities (VSES). In: Mas A, Mesquida A, Rout T, O'Connor RV, Dorling A. (eds.) *Software Process*

Improvement and Capability Determination. Berlin, Heidelberg: Springer; 2012. p.38-50.
[29] Wen L, Tuffley D, Dromey RG. Formalizing the transition from requirements' change to design change using an evolutionary traceability model. *Innovations in Systems and Software Engineering*. 2014; 10: 181-202.
[30] Wen L, Tuffley D, Rout T. Using composition trees to model and compare software process. In: O'Connor RV, Rout T, McCaffery F, Dorling A. (eds.) *Software Process Improvement and Capability Determination*. Berlin, Heidelberg: Springer; 2011. p.1-15.
[31] ISO/IEC 15288. *Information Technology-System Engineering-System Life Cycle Process*. 2015.
[32] Van der Aalst W. *Data Science in Action, in Process Mining.* Berlin, Heidelberg: Springer; 2016. p.3-23.
[33] EMC Education Service. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. New York: John Wiley & Sons; 2015.