

# Development of digital health management systems in longitudinal study: The Malaysian cohort experience

DIGITAL HEALTH  
Volume 10: 1–13  
© The Author(s) 2024  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20552076241277481  
journals.sagepub.com/home/dhj



Noraidatulakma Abdullah<sup>1</sup> , Nurul Faeizah Husin<sup>1</sup>, Ying-Xian Goh<sup>1</sup>,  
Mohd Arman Kamaruddin<sup>1</sup>, Mohd Shaharom Abdullah<sup>1</sup>, Aiman Fitri Yusri<sup>1</sup>,  
Azwa Shawani Kamalul Arifin<sup>1</sup> and Rahman Jamal<sup>1</sup>

## Abstract

**Background:** The management of extensive longitudinal data in cohort studies presents significant challenges, particularly in middle-income countries like Malaysia where technological resources may be limited. These challenges include ensuring data integrity, security, and scalability of storage solutions over extended periods.

**Objective:** This article outlines innovative methods developed and implemented by The Malaysian Cohort project to effectively manage and maintain large-scale databases from project inception through the follow-up phase, ensuring robust data privacy and security.

**Methods:** We describe the comprehensive strategies employed to develop and sustain the database infrastructure necessary for handling large volumes of data collected during the study. This includes the integration of advanced information management systems and adherence to stringent data security protocols.

**Outcomes:** Key achievements include the establishment of a scalable database architecture and an effective data privacy framework that together support the dynamic requirements of longitudinal healthcare research. The solutions implemented serve as a model for similar cohort studies in resource-limited settings. The article also explores the broader implications of these methodologies for public health and personalized medicine, addressing both the challenges posed by big data in healthcare and the opportunities it offers for enhancing disease prevention and management strategies.

**Conclusion:** By sharing these insights, we aim to contribute to the global discourse on improving data management practices in cohort studies and to assist other researchers in overcoming the complexities associated with longitudinal health data.

## Keywords

Data management, database, information system, the Malaysian Cohort, cohort studies

Submission date: 12 February 2024; Acceptance date: 7 August 2024

## Introduction

The healthcare system continuously generates enormous amount of data from patients' medical records, administrative data, human resource data, laboratory data, drugs and prescription data, costs, and insurance claims data, and many more.<sup>1</sup> It is advantageous if the healthcare system,

<sup>1</sup>UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia  
NA and NFH contributed equally to this work.

### Corresponding author:

Rahman Jamal, UKM Medical Molecular Biology Institute (UMBI), Jalan Yaacob Latif, Bandar Tun Razak, Cheras, 56000 Kuala Lumpur, Malaysia.  
Email: rahmanj@ppukm.ukm.edu.my



Creative Commons NonCommercial-NoDeriv CC BY-NC-ND: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDeriv 4.0 License (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits any use, reproduction and distribution of the work as published without adaptation or alteration, provided the original work is attributed as specified on the SAGE and Open Access page (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

including the healthcare facilities, is digitalized, and integrated at the national level, like in most developed countries, such as the NHS in the United Kingdom and Medicare in Australia. In many low- and middle-income countries, like Malaysia, patients' records are mostly still in manual written form and not fully integrated into the laboratory, radiological, or pharmacy information systems within the same hospital. These manual patient records are not easy to manage, laborious to store and retrieve, and prone to be lost or misplaced. Electronic medical records (EMRs) have emerged with the development of Web 3.0 and Healthcare 3.0, enabling tailoring and optimization of EMRs for better, and customizable utilization.<sup>2,3</sup> Thus, the use of EMRs is critical to generate and gather information in the era of big data technology to ensure functional exploitation of this enormous data to improve healthcare delivery systems. The concept of digitized health information should also be extended and applied to large-scale population-based cohort studies.

Since its establishment in 2006, The Malaysian Cohort (TMC) project has recruited more than 120,000 volunteers aged 35 and older to participate in the prospective population-based cohort study.<sup>4</sup> To date, it has successfully resurveyed at least 40% of its baseline participants.<sup>5</sup> TMC is a platform for the storage of both big data and biospecimens. The biobank processes and stores biospecimens (such as blood and urine samples), while the database stores various datasets, including demographic data, healthcare and medical data, exposure data, data on physical activity and diet, data on biophysical measurements, laboratory data, follow-up data and mortality data which are all digitized and could be used for future research, especially for identification of risk factors and biomarker discovery research.<sup>6-8</sup> TMC is a deeply phenotyped population with representation from the three major ethnic groups, namely Malay, Chinese, and Indian. TMC has also performed genotyping on about 10% of the participants, while there are also whole genome and whole exome sequencing data on a limited number of cases. Combining both genotypic and phenotypic data and linking this to outcomes and mortality will allow researchers to explore associations, predictive algorithms, biomarker panels for early detection of diseases and modelling of disease burden with time.<sup>1</sup>

The data of prospective studies like TMC will continue to increase over time; hence it is crucial to develop a database system that will be sustainable for the future.<sup>9</sup> The database created should be functional, secure, flexible, robust and able to adapt according to additional data from the participants, new surveys, and biophysical measurements added to increase the value of the cohort study. Developing this kind of database requires extensive planning and technical expertise, which are usually costly. Over the past 18 years from the time of establishment, TMC has consistently developed and improved the TMC data and information management systems to cater the needs of data capacity, and technology advancement. Since baseline recruitment

through the follow-up phase, we have developed several in-house databases and systems such as listed in Figure 1.

The first database that developed in-house was Cohort Information Management System (CIMS) that manages the critical aspects of the study, especially with the data collected.<sup>4</sup> The CIMS consists of four parts, namely the electronic Cohort Information Management System (eCIMS), Diet Information Management System (DIMS), Health Diary Information Management System (HeDIMS), and Tube and Sample Information Management System (TSIMS). The information systems were designed by the information and communication technology (ICT) team using the specifications requested by the management team of TMC. The TMC ICT team collaborated with a local ICT vendor, and this smart partnership helped to substantially reduce the cost as compared to using commercially available information management systems. The source code is shared by the ICT vendor with the TMC ICT team, allowing complete and independent maintenance by the local team. The development of CIMS is not only for data collection, processing, and storage but also for business intelligence, dashboarding, and reporting to management.

## Methods and outcomes

### Database management system

TMC is a prospective health study that has been collecting health-related data from healthy Malaysian citizens aged 35 to 70 across the entire country, including both Peninsular Malaysia and Borneo, since its initiation in 2006.<sup>4</sup> The project has been ongoing for approximately 18 years and continues to accumulate significant volumes of health-related information, necessitating large databases for data archival. As medical data is diverse in nature, our cohort data consists of different types of data, ranging from administrative health information to biophysical and laboratory measurements. Managing and retaining this data can be tedious due to the diversity and complexity of the data types. Additionally, environmental conditions and the availability of electricity supply and connections also need to be considered when collecting data in the field. Hence, our primary concerns when selecting an appropriate Database Management System (DBMS) in 2007 were stability, security, system compatibility, and high performance. Other minor considerations included feasibility, performance, scalability, operation, and efficiency.

To store the data in the DBMS, TMC used Michael Widenius's Structured Query Language (MySQL) server. A relational database that uses SQL was chosen for its ability to create meaningful information by joining information from various sources. Moreover, physical conditions like a shortage of electricity and no internet connection in the field made a staging SQL server with Ethernet is the appropriate choice for temporarily storing the data. SQL servers were used to avoid data incompatibility when transferring data from the staging server to the main server.

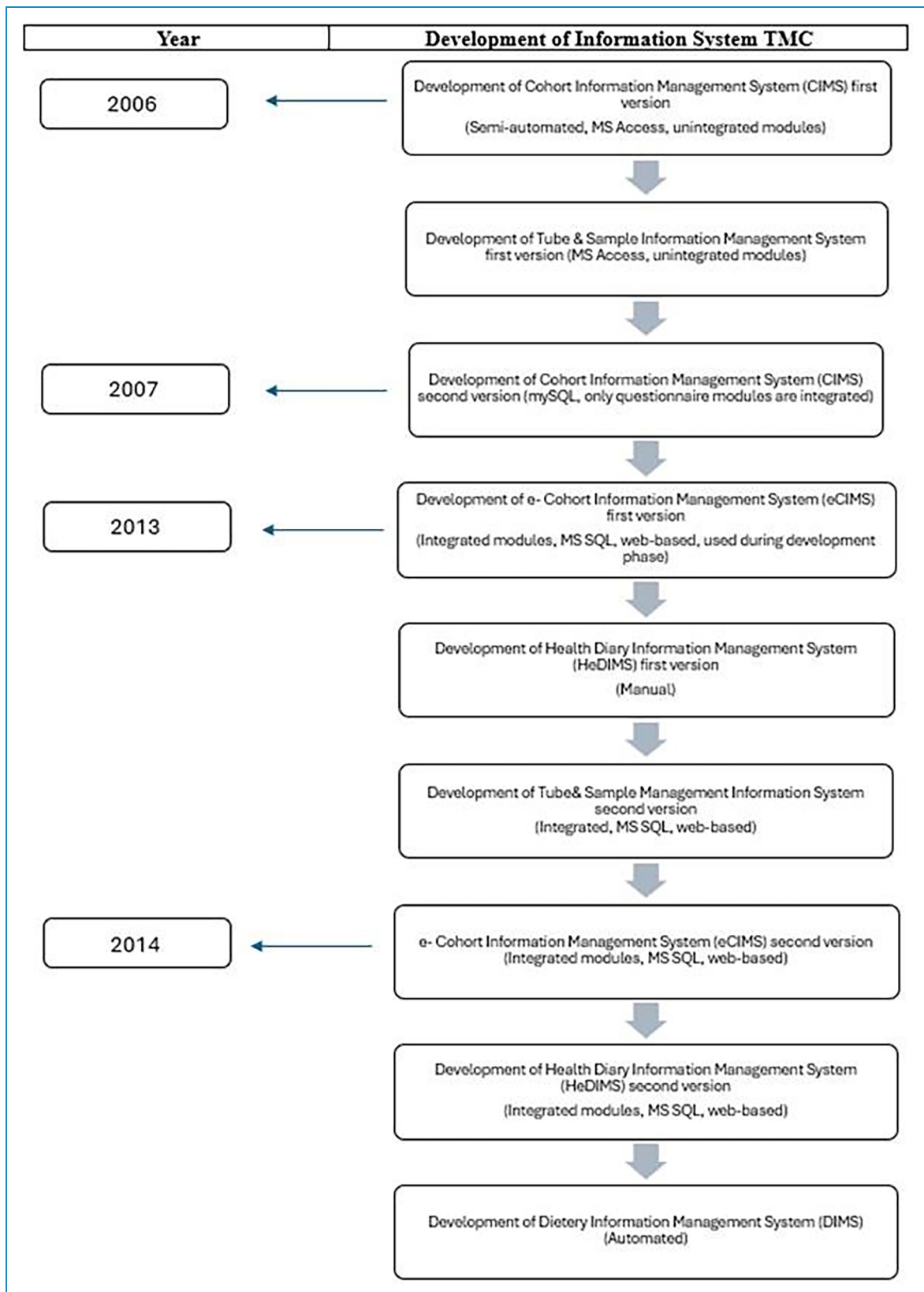


Figure 1. The development of information system TMC. TMC: The Malaysian Cohort.

As the data grew and required higher scalability to perform complicated queries, TMC upgraded from MySQL to Microsoft SQL (MS SQL) servers. MS SQL is more adaptive to data sharing and enables data distribution from one table to multiple instances and machines. Furthermore, MS SQL does not block the database while backing up data, making it easier for users to back up and restore vast amounts of data.

Migrating from MySQL to MS SQL server offers a range of compelling advantages. Upgrading Cohort Management System to the new features use ASP.NET language. MS SQL Server seamlessly integrates with the .NET ecosystem, including Visual Studio which offering smoother collaboration compared to MySQL. Moreover, SQL Server's T-SQL boasts powerful imperative programming features enabling users to accomplish complex tasks efficiently. MS SQL Server's superior replication support makes it ideal for scaling databases beyond basic configurations while its granular locking mechanisms minimize access disruptions compared to MySQL. In short, the use of appropriate DBMS and staging servers has allowed TMC to manage its data efficiently, ensuring data integrity and security.

### Data security and confidentiality

Maintaining privacy and confidentiality is crucial to any ecosystem that collects personal and health-related data. In terms of data security and confidentiality, an independent infrastructure was developed to limit access to the databases to specifically authorized personnel. Having this control over the data is crucial, as highlighted in other studies.<sup>10,11</sup> Each authorized personnel are given a unique password based on their access level.

Since TMC is parked at the Universiti Kebangsaan Malaysia (UKM) Medical Molecular Biology Institute (UMBI), all systems are protected within the university network infrastructure and cannot be accessed beyond the university network. There are five layers of security to protect the data: physical security, storage security, access management security, network security, and internal security. The layers of security were meant to address the operational control of data and protect it from loss or unauthorized disclosure, including cyber-attacks (Figure 2).

1	The internet protocol (IP) address was protected by the university's network security and firewall. It cannot be accessed, unless using university network.
2	The servers are located at specific Data Centre that are equipped with security system.
3	Each server has unique IP address, username, and password.
4	All systems have specific IP address, username and password with restricted access. Audit trail and Recovery system.
5	Personal information are encrypted.

Figure 2. Five security layers.

Sensitive data, including name, MyKad number (the unique identification number given to each citizen of Malaysia), and contact number, will be encrypted. Cohort Information Management System (CIMS) has implemented the AES 128-bit encryption techniques to safeguard and encrypt the sensitive data. Personal data is legally collected with informed consent and protected from abuse and exploitation, as well as respecting the rights of data owners. In addition, written informed consent, including consent to publish the study results, was obtained from all subjects prior to their participation in the study. AES 128-bit encryption scrambles data into blocks applies multiple rounds of substitutions and mixing operations using a 128-bit key. This process transforms the data iteratively, producing ciphertext that extremely difficult to decipher without the corresponding decryption key. AES 128-bit achieves a harmonious equilibrium between security and efficiency, rendering it well-suited for numerous practical applications requiring robust encryption while mitigating computational burdens compare to AES 256-bit which greater computational resources and time. Those who have access to the data will need to sign and keep a current Privacy Protocol agreement and an Oath of Secrecy upon their appointment of work in TMC. Audit trails are also built into the system to track users and any processing and downloading of data.

### Data back-up

Maintaining reliable backups is a critical component of data management. Consistent backup procedures help to mitigate potential damage or loss resulting from hardware failure, software or media faults, virus or hacking attacks, power outages, or other human errors. Details on backup locations, dates, and times are electronically recorded by automated backup procedure. If the data was backup manually, the backup's file contains the address of the file, the system's name, data type, backup date, and location.

Due to the large amount of data in TMC, full daily backups (both manual and scheduled) are executed, accompanied by the maintenance of write-ahead logs to facilitate restoration to any desired point within the retention period after working hours, when the server is at its low-usage time. Manual backup enables us to select a specific collection of files, file groups, or transaction logs from MS SQL Server Graphical User Interface (GUI), while the scheduled backup is planned at midnight each day. In addition, the unstructured data backups are performed weekly by the team after all the files are completed and reviewed.

To ensure comprehensive data preservation and enhance data security, it is essential to have a backup and recovery strategy in place. This strategy should be determined based on the Data Management Plan, as outlined in the Standard Operating Procedure (SOP) for Backup Plan Management.

These backup data are securely stored in discipline-specific repositories located within different local network server storage with authentication or DVDs (during recruitment at field) and kept in a safety locker in a room with a security door that only authorized personnel can access. Additionally, an off-site backup (two physical backups with authentication stored in different locations) was also created to ensure redundancy and disaster recovery. This practice not only enhances the efficiency of data backups but also offers a cost-effective option.

### Cohort information management system

**Electronic cohort information management system.** The eCIMS contains 16 modules to collect various information from the participants. The modules have a total of 310 digital data parameters. The eCIMS generates a unique subject identifier (SID) for every participant and manages the various types of information archived to ensure easy reference and retrieval of data during the long-term follow up of participants. It also serves as the primary system that manages data such as contact details and participation status, as well as health-related information, including demographic data, questionnaire data, biophysical data, biochemical tests, biobank, and follow-up data. All the data from the participants were digitalized except the consent form, which remained in the written format for ethical purposes. The eCIMS is integrated all modules and upgraded CIMS that is web-based design compatible with PC browser and mobile view for better user experience. In addition, it provides a solution for CIMS to easily access real-time data and ensure data reliability by using a Personal Area Network to access the system and connect all recruitment PC via the network.

**Subject identifiers.** The subject identifier (SID) is generated automatically by the eCIMS for each participant at registration and before the interview and health screening. This SID serves as the only reference for individual data information.

**Table 1.** The nine-digit ALSWH ID number format, ABC-DEFGH-I as follows.

Item	Code	Classification*
A	1, 2, 3	Age group
B	1, 2, 3, 4, 5, 6, 7, 8	State
C -	1, 2, 3	Area
DEFGH-	0-9	5-digit counting code
I	0-9	0-9 Check digit

ALSWH: Australian Longitudinal Study on Women's Health.

Usually, this SID is comprehensive yet simple enough to be applied practically. Unlike the identifier in Australian Longitudinal Study on Women's Health (ALSWH),<sup>9</sup> our 10-digit SID number is relatively more straightforward. The ALSWH ID format includes nine digits, with the first digit denoting age cohort, the second digit representing state of residence and the third digit indicating urban, rural, or remote area of residence within Australia. The middle five reflect the sequential number of invitations sent to each age cohort. A check digit ensures the validity of the ID (Table 1). The basic information about participant recruitment, such as the recruitment date and location, is incorporated into the SID. This allows us to identify the basic information about the participants for follow-up, contact tracing, and answering queries, enabling us to retrieve the records quickly (Table 2). The format of the SID is arranged as "D-D-M-M-Y-Y-L-L-T-T," where the first six digits of the SID represent the recruitment date in the day-month-year format: D, M, and Y indicate for "day," "month," and "year," respectively, for the date of recruitment. As for L (-\_-), it indicates the location

**Table 2.** Basic information of the recruitment location.

Code	Description	Classification*
0_-	Federal Land Development Authority (FELDA)	Rural
1_-	Hospital Universiti Kebangsaan Malaysia (headquarter)	Urban
2_-	Village	Rural
3_-	Town	Urban
4_-	Aboriginal/Indigenous areas	Rural
5_-	Sabah (East Malaysia)	Urban
6_-	Sarawak (East Malaysia)	Urban
7_-	Rural areas in Sarawak (East Malaysia)	Rural
8_-	Recruitment from home**	Urban
A-Z	Respective new collaboration project	Not applicable
_0	Red team	Not applicable
_1	Blue team	Not applicable
_2	Green team	Not applicable

\*The recruitment location generally can be classified into two main groups (rural and urban) with some additional information as described above.

\*\* Serves for the participants who unable to return for a follow-up screening due to various reasons as described in Abdullah et al.<sup>5</sup>

information that being described in Table 2: the first digit of L indicates the recruitment location, while the second code indicates the recruitment team. During recruitment at field, three separated teams were assigned to recruit study subjects simultaneously at different locations. The T indicates for the “turn number” during the recruitment day. For instance, if the SID is 2702231001, we will know that the participant was recruited on 27 February 2023 (2702311001) at the headquarter (Hospital Universiti Kebangsaan Malaysia) by the red team (2702311001), and he/she was the first participant of that day (2702231001).

Each participant of TMC has two identification numbers, the SID, and the Malaysian national identity card (MyKad)

**Table 3.** Participation follow-up status categorization.

Category	Description
Participating	Participants recruited during baseline agreed and are actively participating in subsequent follow-up recruitment.
Pending	Participants who did not answer the follow-up invitation call or those who are not sure to further participate in the project due to: <ul style="list-style-type: none"> <li>• Traveling, outstation, or currently not available at the recruitment site;</li> <li>• Transportation difficulties / Depending on others to deliver them to the recruitment site; and</li> <li>• Too busy or not free during the recruitment period</li> </ul> which will be noted in the subfield of our databases.
Withdrawn	Participants who would like to withdraw from the study are unwilling to contribute in the future. In this category, we further note their reason for withdrawal in the subfield, which includes: <ul style="list-style-type: none"> <li>• No reason was given;</li> <li>• No longer interested;</li> <li>• Unsatisfied with the service provided by TMC; and</li> <li>• The participant is undergoing treatment at other places.</li> </ul>
Lost to contact	Participants who are uncontactable include: <ul style="list-style-type: none"> <li>• Wrong contact number/Participant has changed their contact number and</li> <li>• Contact number not in service or could not be reached.</li> </ul>
Deceased	Participant has died, and the death information is cross-validated with National Registration Department twice annually.

TMC: The Malaysian Cohort.

number. Although the SID was created for each participant, the MyKad number is used to confirm the participant’s identity, especially during follow up. These unique forms of identification, such as MyKad number, could help eliminate substitutions of participants in large-scale population studies.<sup>9</sup> Besides, a new SID will be assigned to the participants if they return for follow-up recruitment, which is linked to the original SID in the system. Since the SID did not contain information regarding the number of times a participant took part in the study, especially for those involved in the follow-up phase, the MyKad number plays its essential role here in linking all the recruitment information together. Also, the MyKad numbers are linked to the National Registration Department (NRD), which registers all the deaths in the nation. We send all the MyKad numbers every six months to the NRD, and it then provides us with the mortality data and the cause of death. For security reasons, the MyKad numbers will be encrypted in our system and can only be decrypted for valid reasons such as data checking and confirmation.

*Contact details and participation status.* Each participant’s name, as per MyKad, is recorded during the baseline recruitment. Our participants seldom change names by marriage or for other reasons; however, if there are any, we will record them accordingly.

Contact details of participants and their next of kin, such as their residential and workplace telephone numbers are recorded for follow-up purposes. The follow-up is primarily to obtain information regarding the health-related events of the participants such as morbidity, hospital admissions, visits to the clinics, updates on medication and surgical procedures as well as mortality status.<sup>12</sup> The contact number of a close relative or a proxy contact number, usually offspring or younger close relatives of the participant, plays a crucial role as a communicator between the elderly participants and us. Also, using a proxy will help us to give information on follow-up data<sup>13</sup> if the participant is uncontactable or has language barriers or difficulties in communication.

Our database also records the mailing addresses of each participant. Each address has a postal code that serves as a unique identifier to help us locate the participant and gives an indication of whether the participant comes from a rural or urban area. The mailing address is also crucial for us to send out the health screening report to the participants. Furthermore, in cases where the participants are uncontactable via phone call for a follow-up visit, we will send the invitation and reminder based on the address to update their latest contact information and health status. In a particular situation where the participants have difficulties coming to us for the follow-up visit, our mobile health screening team will visit them based on their addresses.<sup>5</sup> Although this information is encrypted in the database for security and privacy, it can be referred to and decrypted for authorized personnel only during the follow-up session

and data checking. The data is not shared with any third party to ensure the privacy and confidentiality of the participants.

This project's follow-up status is also regularly recorded and updated in our database. The participation status at follow up is divided into five categories: participating, pending, withdrawn, lost to contact, and deceased (Table 3).

**Medical history and lifestyle details.** A piece of important information in a longitudinal study is the health and lifestyle data such as medical and surgical histories, physical activity, living environment, exposure to alcohol and tobacco, and dietary patterns. These are rigorously obtained at baseline and follow up, and this information is maintained in the database and updated during follow up.

**Quality control.** To conduct effective analysis of large datasets, it is imperative to implement quality control measures such as data standardization and calibration. These measures help eliminate data noises that can impede analytics. In TMC, a quality control system was developed in every section to minimize errors and ensure data quality. A quality control assessment was conducted by trained personnel on each data collected before importing data into the system. It is known that self-reported questionnaires collected in longitudinal studies are prone to error<sup>14</sup> as they depend highly on the participant's memory. Moreover, there is also a frequent recall and reporting bias. Thus, to minimize the error, each administered interview was recorded on an MP3 player and was played backed during quality control session by another independent enumerator. This procedure will allow the enumerators to rectify the errors if there is any.<sup>4</sup>

As for our data generated from the laboratory test, the quality of laboratory assays and readings were audited every month according to the Royal College of Pathologists of Australasia (RCPA) external quality assurance (EQA) programs.<sup>15,16</sup> In addition, the laboratory was accredited with the MS ISO 15189:2014 for Medical Testing for Pathology and Hematology by the Department of Standards Malaysia.<sup>15,16</sup>

There are also quality control procedures covering data storage and data management. Our biostatisticians performed a semi-annual audit to ensure that the data is up-to-date and ready for analysis. Any issues identified during the audit are reviewed by our database officer to ensure data integrity. This rigorous process guarantees high-quality and reliable data for accurate analysis.

### *Diet information management system*

The DIMS is a system to manage diet information from both the 24-hour food recall and the data from the Food Frequency Questionnaires (FFQ). The development and validation of TMC dietary intake based on both dietary tools have been published elsewhere.<sup>17</sup> Based on our

knowledge, the DIMS from TMC is the largest food database in Malaysia that has collected more than 4000 kinds of food to date, comprising the diverse and rich Malaysia's multi-ethnic and multi-cultural dishes. In addition, this system has been developed to calculate everyone's nutrition and calorie intake automatically. Both eCIMS and DIMS are linked together to ensure the feasibility of data retrieval and extraction on an individual basis.

### *Health diary information management system*

Another system developed under CIMS was the HeDIMS to collect information on participants' health status. Previously, a Health Dairy booklet was given to the participants during the baseline to fill in any information on every health check-up, hospital admission, and medication update. They had to return it via post or self-delivery to our center. However, due to the low return rate, the HeDIMS was developed to replace the Health Dairy booklet and used during the follow-up call or visit.

### *Tube and sample information management system*

TMC biobank is the largest in Southeast Asia that consists of 42 units of  $-80^{\circ}\text{C}$  freezers and 40 liquid nitrogen tanks, and stores over 9 million tubes of biospecimens, including blood, urine, and stool. Due to the large volume of samples, we cannot catalog the biospecimen manually. Hence, the TSIMS was developed to manage the inventory of the biobank. This system allows us to capture the information and location of biospecimens effectively.

Biospecimens are kept in the cryogenic tubes on 96-well tube racks for biobanking. Each tube has a two-dimensional barcode at the bottom of the tube. The sample-containing tubes are scanned using the Tracxer Code Reader (Micronic, USA) after processing them into various sample types and fractions. The information of each sample (such as SID) is keyed in accordingly to allow us to identify the owner and origin of the biospecimen. A total of 54 tubes consists of 44 tubes of blood samples (plasma, serum, whole blood, red blood cell, and white blood cell) and 10 tubes of urine samples (whole urine and urine pallet) are generated and stored for each participant during the baseline phase. While in the follow-up phase, we reduced the production to 20 tubes, which include 16 tubes of blood samples and 4 tubes of urine samples. The details of the samples fraction can be referred in Table 4. In addition, there is also a barcode embedded on the tube rack, the so-called rack ID, for identification purposes.

When each 96-well tube rack is filled up, an in-house 8-digit alphanumeric code will be assigned to the rack. The code contains information on the location of the rack in the biobank. For example, L34G0511 indicates that the rack was stored at liquid nitrogen tank number 34 (L34), compartment G (G), rack number five (05), and at level 11 (11), counted from below, of the rack. Depending on

its arrangement, each nitrogen tank has six to nine compartments in either a “pie” or “rectangular” pattern. A liquid nitrogen tank can store 96 tube racks, with 15 to 16 layers for each rack.

The rack ID and the eight-digit alphanumeric code serve as vital information for sample retrieval in TSIMS. The SID of the sample will be used as the entered query to search in TSIMS and link to the eight-digit alphanumeric code for the SID and its related information. Our trained research assistants will retrieve the samples according to the location indicated by the eight-digit alphanumeric code.

### Cohort Mobile Application

Big data, together with digitalization, also has the potential to engage and empower patients’ health.<sup>18</sup> By leveraging big data analytics and linking the data with our databases, our participants can access their personalized health information at their fingertips, allowing them to monitor their health using the Cohort Mobile Application (app) installed in their smart phone.<sup>19–21</sup>

The app’s goal is to simplify the data collection process and integrate multiple users into a user-friendly system

**Table 4.** The fraction of samples generated from The Malaysian Cohort.

Number	Sample type	Fractions	Baseline	Follow up	
1	ACD (6.0 ml) × 1	Whole Blood	2	4	
		DMSO			
		Plasma	4		
		MNC-ACD	2		
		DNA extraction		2	
2	EDTA (10 ml) × 1	Plasma	16	4	
		Buffy coat	4		1
		Red blood cells	4		2
3	SST (8.5 ml) × 1	Serum	12	3	
4	Urine (25 ml)	Whole urine	8	2	
		Pellet urine	2		2
<b>Total aliquots</b>			<b>54</b>	<b>20</b>	

DMSO: dimethylsulfoxide; EDTA: ethylenediaminetetraacetic acid; MNC: multinational corporation.

(Figure 3). It also improves on the limitations of the current information management system, such as being limited to on-premises servers and restrictions on changes to the questionnaires’ structure. The mobile app developed may save time and streamline the communication between the cohort participants, researchers, and the technical management team, making the processes of data collection and management more user-friendly and efficient (Figure 3).

The mobile app was designed to include various features to enhance its functionality for the cohort participants. These features include an early registration option, EMR management, generation of health reports and scores, tracking of the participant’s dietary patterns, health diagnosis, administering of questionnaires, and a health calculator (to calculate metrics such as body mass index, calorie intake, and body fat percentage). Besides, the design of the mobile app took into consideration the need to improve participants’ engagement and reduce cohort attrition. The inclusion of a reward points system, user support frequently asked questions (FAQ) section, notifications for health campaigns, and links to reliable health information are all features aimed at enhancing participants’ experience and ensuring their continued use of the app.

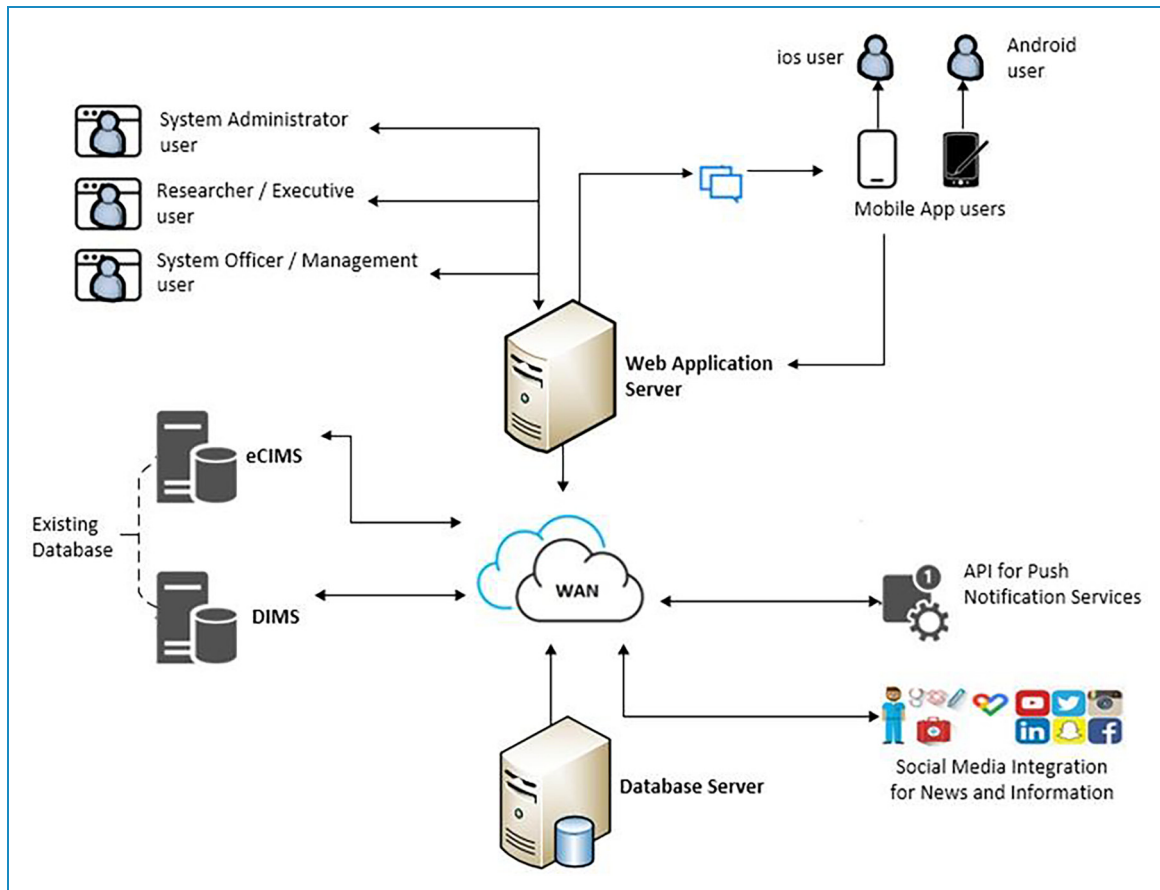
The reward points system, for instance, serves as a motivational tool to encourage participants to stay engaged with the app and maintains healthy habits. The user support FAQ section provides a resource for participants to seek answers to common questions and resolve issues they may encounter while using the app. Notifications for health campaigns and links to reliable health information also serve to educate participants about healthy practices and provide them with the resources, they need to maintain good health.

Furthermore, to stay up to date with current technology trends, the app will incorporate artificial intelligence technology to enable diagnosis based on reported symptoms. This will provide medical personnel with a better understanding of the participant’s condition before they attend and receive treatment at a healthcare center.

Apart from that, the Cohort Mobile Application serves as the prototype of a smartphone-based EMR, where participants can access their data exclusively based on the unique ID provided. The app also enables general comparisons between the participant’s current health status, the overall TMC participants, and the general healthy population, allowing the participants to know their real-time health status compared to the others. Through this intervention, the healthcare providers can tailor patient education and engagement strategies according to the evidence, needs, preferences, and goals of respective patients.<sup>21,22</sup> Eventually, this approach will improve patient satisfaction and outcomes, while also promoting patient-centered care.

Although this mobile app is still under development, this can help healthcare providers identify gaps in patient engagement in the near future. These gaps can be later





**Figure 3.** The Cohort Mobile Application aims to streamline and automate the data collection and management process by integrating multiple parties such as system administrators, researchers, and cohort participants into a web application server. The server will be synchronized with our main database server using a wide area network to ensure that our main database is always up to date. The mobile app features a user-friendly interface that enhances the user experience. Furthermore, it will also incorporate social media platforms to deliver credible news and information related to health.

addressed by developing targeted interventions that improve patient activation and involvement in their own health.<sup>18</sup> Through these efforts, healthcare delivery and outcomes would be improved in the long run.

### *Big data analysis using machine learning and artificial intelligence*

Machine learning (ML) analysis has become increasingly popular in medical health studies in recent years, as it can help identify patterns and relationships that may not be apparent using traditional statistical methods. Thus, we have embarked on ML to predict diseases using TMC's data, such as cardiovascular disease (CVD) and colorectal cancer.

We have conducted a discovery study on 60 participants from TMC to predict the risk of CVD in Malaysian individuals using ML techniques.<sup>23</sup> This study analyzed digital electrocardiogram (ECG) data, blood pressure, and cholesterol levels. Six different ML algorithms were evaluated for

the most accurate CVD risk prediction, and the artificial neural network (ANN) method had the highest prediction performance, with a 90% accuracy rate.<sup>23</sup> Later, we validated this study in larger dataset consist of 244 participants.<sup>24</sup> In this study, an automated peak detection algorithm was used to detect certain fiducial points of ECG and extracted 48 different features.<sup>24</sup> Out of 48, 6 features related to the T wave were important in identifying individuals with CVD.<sup>24</sup> In addition, five different ML, namely ANN, k-nearest neighbors (kNN), support vector machine (SVM), discriminant analysis (DA), and decision tree (DT), were used to distinguish individuals with CVD.<sup>24</sup> Similar to previous study, they found that ANN was the best-performing method in distinguishing CVD, with a high level of accuracy and specificity.<sup>23,24</sup>

In addition to cardiovascular disease, ML has also been used to predict the risk of colorectal cancer among Malaysian cohort participants.<sup>25,26</sup> In this predictive model, 25 trace elements (TE) and their interactions with environmental risk factors were used to predict the risk of

developing colorectal cancer. The prediction model was developed based on three ML algorithms, logistic regression, SVM, and ANN with good accuracy during the discovery phase.<sup>26</sup> Further modeling using a 24-TE panel resulted in 100% accuracy in predicting colorectal cancer, followed by the TE-environmental risk combination (86.5%) and environmental risk factors alone (67.3%).<sup>26</sup> In short, the study revealed a positive interaction between red meat intake of  $\geq 50$  g/day and cobalt levels of  $\geq 4.77$   $\mu\text{g/L}$ , which doubled the risk of colorectal cancer.<sup>26</sup>

## Discussion

### *Challenges of big data in public health and precision medicine*

Big data has transformed the field of public health and precision medicine by providing unprecedented access to vast amounts of health-related data. However, it also poses several challenges that need to be addressed to make the most of the data available.

Medical data are unique compared to other types of big data.<sup>27</sup> It includes administrative health information, biomarker data, clinical measurements, biometrics, and images that originate from various sources like electronic health records, laboratory testing, clinical registries, governmental databases, biobanks, and the self-report of the patient.<sup>28</sup> Due to the complicated and diversified nature of medical data, managing and streamlining these data from disparate sources can be extremely challenging.<sup>29</sup> This is because different sources of data often use different formats and standards, making it difficult to integrate. In addition, ensuring data quality will be a significant challenge when we have multiple data sources.<sup>29,30</sup> Multiple data sources of varying levels of quality may impact the accuracy, completeness and consistency of data that are critical for effective analysis and decision making.

Apart from that, large-scale datasets contain sensitive health information. Hence, strict data security and privacy measures are required to protect participant confidentiality. However, maintaining privacy and security while also making data accessible for analysis is a complex challenge.<sup>31</sup> At TMC, we use a unique identifier called SID, to keep the participant confidentiality.<sup>4</sup> Personal data such as name, MyKad number, and contact details were encrypted and kept separated from other data. In short, maintaining privacy and security is crucial to ensuring public trust in the use of big data in healthcare.<sup>31,32</sup>

The complexity of big data in healthcare requires multidisciplinary expertise who can comprehend data science, computer science, public health, and epidemiology as well as medical sciences.<sup>33</sup> Analyzing large-scale datasets requires advanced analytical capabilities, including machine learning, natural language processing, and data visualization.<sup>29</sup> Advanced infrastructures and settings would also be needed for the data analytics. For instance, recent application of

neural networks in constructing disease prediction models relies heavily on large-scale datasets and high-end computing infrastructure.<sup>34</sup> Nevertheless, developing and deploying these capabilities in the fields of public health and precision medicine can be a significant challenge in a low- and middle-income countries (LMICs),<sup>35</sup> like Malaysia as it requires significant investment in technology and expertise.

In addition, due to the vast amount of data (130 TB), maintaining and managing the server would be very expensive in the long run. Thus, TMC is exploring cloud storage for more feasible, cost effective, and secure data storage in the near future.

### *Opportunities of big data in public health and precision medicine*

The potential for big data to revolutionize healthcare delivery and research in public health and precision medicine is undeniable. With the help of big data analytics, early disease detection and effective treatment as well as prevention strategies can be implemented.<sup>30,33</sup> Predictive algorithms and analytics can be used to identify the patterns and risk factor associated with certain diseases, enabling healthcare providers to intervene early and prevent the progression of complication.<sup>29</sup> Furthermore, personalized treatment plans can be tailored based on individual characteristics such as genetic, environmental, lifestyle, and risk factors. This could ultimately lead to more effective and targeted treatments with fewer side effects.

Apart from that, big data analytics could advance clinical research by revolutionizing the workflow of drug discovery, including drug repurposing.<sup>36,37</sup> With the ability to analyze vast amounts of data, new drug targets can be identified, thus improving clinical trial design. Consequently, this will result in more effective treatments, speeding up the drug development processes.<sup>29</sup> In fact, similar workflow and pipeline have been applied in accelerating drug repositioning during the COVID-19 pandemic.<sup>38–40</sup> By integrating artificial intelligence, deep learning, and network medicine, potential drug structures and candidates to treat COVID-19 can be predicted in a cheaper, faster, and effective approach, with reduced possibility of adverse effect in future clinical trials.<sup>38,39</sup> Additionally, researchers and stakeholders can make better decisions on the potential drug candidates to pursue by leveraging big data analytics. This effort can optimize patient selection criteria<sup>20</sup> and identify unrealized safety concerns beforehand in the development process.<sup>29</sup> This approach will significantly reduce the costs and time-to-market of a medication,<sup>36</sup> ultimately leading to more efficient and affordable healthcare.<sup>18–22</sup>

### *Potential application of blockchain-based database in healthcare*

The future of healthcare database development could rely on blockchain technology, which ensures safety, security,

and scalability when sharing data.<sup>3</sup> Unlike the Healthcare 3.0 era, where data management was more centralized, blockchain technology provides a platform to decentralize data management and disrupt traditional practices.<sup>3</sup> Blockchain technology plays a critical role in Healthcare 4.0, facilitating smarter and more interconnected healthcare systems.<sup>41,42</sup>

With blockchain, patient data and EMRs can be stored across different blocks of healthcare providers (such as clinics and hospitals) and databases (such as EMRs and radiographic images). These data are chained only by the patients' identity, forming a single network focused on that patient.<sup>3</sup> Decentralized data storage and management also enables real-time data auditing and tracing for all operations. The technology also allows for vast data collection in real-time, reducing the time required for disease detection and diagnosis.

Furthermore, data stored under blockchain technology can be left unsupervised, autonomously audited, and controlled by embedded algorithms, which could eventually reduce the manpower required to oversee and curate data. Besides, data stored is secured by shared cryptography, where users are assigned a unique username and password that is encrypted,<sup>3</sup> providing increased security, privacy, and exclusivity for the respective patient.

### *The future is now: Application of natural language processing and large language models in precision medicine*

The integration of natural language processing (NLP) and large language models (LLMs) has revolutionized precision medicine, offering unprecedented capabilities in data management, data analysis, information extraction, and predictive modelling. NLP techniques enable researchers to extract valuable insights from unstructured textual data sources such as EMRs, clinical notes, and biomedical literature.<sup>43</sup> By automatically analyzing and synthesizing vast amounts of textual data, NLP facilitates the identification of disease patterns, risk factors,<sup>44</sup> treatment outcomes,<sup>45,46</sup> and adverse events with greater efficiency and accuracy than conventional approaches, such as prognostic scores.<sup>47</sup> Besides, NLP will enable clinicians to design personalized treatment plans based on the patient's unique genetic profile and their medical history.<sup>48,49</sup> Additionally, NLP facilitates the integration of diverse data modalities,<sup>50</sup> including genomics,<sup>51,52</sup> proteomics,<sup>53,54</sup> metabolomics,<sup>55</sup> and clinical imaging,<sup>56</sup> to provide a comprehensive view of disease pathology and treatment outcomes.

The advent of LLMs such as GPT (Generative Pre-trained Transformer) has further advanced the capabilities of NLP in precision medicine.<sup>57</sup> LLMs leverage vast amounts of pre-existing textual data to learn intricate language patterns, semantic relationships, and contextual nuances, enabling them to generate coherent text,

summarize information, and answer complex queries.<sup>57</sup> In the future, LLMs may assist clinicians in interpreting genomic data, generating patient-specific clinical summaries, and predicting treatment outcomes based on historical data and medical literature.

Despite the considerable potential of NLP and LLMs in precision medicine, numerous hurdles persist. These encompass safeguarding patient data privacy and security<sup>58</sup> and rectifying biases and inaccuracies in algorithmic forecasts.<sup>59</sup> Notably, a recent study has shown that the performance sensitivity to prompt format and result instability across LLM (i.e. GPT-4) versions pose significant limitations to its clinical diagnostic utility.<sup>60</sup> Further efforts are warranted to enhance (1) the evolution of clinical NLP methods from mere extraction to comprehension; (2) the identification of relationships among entities rather than isolated entities; (3) temporal extraction for comprehensive understanding of past, present, and future clinical events; (4) utilization of alternative sources of clinical knowledge; and (5) accessibility of large-scale, de-identified clinical datasets.<sup>61</sup> To surmount these challenges, interdisciplinary collaboration between data scientists, clinicians, and domain experts is imperative. Nonetheless, the ongoing advancement and integration of NLP and LLMs hold immense promise in advancing our comprehension of disease mechanisms, enhancing patient care, and fundamentally reshaping medical practice.

### **Conclusion**

This article has described the development of TMC data and information management systems over the past 18 years from the time of establishment, baseline recruitment through the follow up of the participants. The systems and databases that have been developed are still functioning, flexible, durable and sustainable over time. The systems have allowed us to manage and maintain our data cost-effectively and keep track of our participants' records. We have generated and published many publications leveraging the deeply phenotype data and genotype data obtained from the participants. As in other longitudinal studies, the value of TMC will undoubtedly increase with time. In addition, this article also discussed on the challenges and potentials of big data to transform public health and precision medicine by providing new insights into various aspects.

**Acknowledgements:** The authors would like to thank all The Malaysian Cohort staff members and research assistants for collecting the data. The voluntary participation of all participants is greatly appreciated.

**Contributorship:** NA, NFH, and YG were involved in conceptualization; NFH and MSA in software; AFY and ASKA in data curation; NA and NFH in writing—original draft


preparation; YG, and RJ in writing—review and editing; RJ in supervision and funding acquisition; and MAK in project administration. All authors have read and agreed to the published version of the manuscript.

**Declaration of conflicting interests:** The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Ethics approval:** This study was conducted according to the guidelines in the Declaration of Helsinki. All procedures involving research study participants were approved by the Ethics Committee of Universiti Kebangsaan Malaysia (Project Code: FF-205-2007). Written informed consent was also received from the subjects prior to their participation in the study.

**Funding:** The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The Malaysian Cohort was funded by a top-down grant from the Ministry of Education Malaysia, PDE48. Additional funding, including infrastructure and utilities, was provided by Universiti Kebangsaan Malaysia.

**Guarantor:** RJ.

**ORCID iD:** Noraidatulakma Abdullah  <https://orcid.org/0000-0002-7186-2455>

**Supplemental material:** Supplemental material for this article is available online.

## References

- Marx V. The big challenges of big data. *Nature* 2013; 498: 255–260.
- Tiago MTB, Tiago F, Amaral FEB, et al. Healthy 3.0: Healthcare digital dimensions. In: Dwivedi A (ed.) *Reshaping medical practice and care with health information systems*. Hershey, PA: IGI Global, 2016, pp.287–322.
- Singh S, Kumar Sharma S, Mehrotra P, et al. Blockchain technology for efficient data management in healthcare system: opportunity, challenges and future perspectives. *Mater Today Proc* 2022; 62: 5042–5046.
- Jamal R, Syed Zakaria SZ, Kamaruddin MA, et al. Cohort profile: the Malaysian Cohort (TMC) project: a prospective study of non-communicable diseases in a multi-ethnic population. *Int J Epidemiol* 2015; 44: 423–431.
- Abdullah N, Kamaruddin MA, Goh Y-X, et al. Participants attrition in a longitudinal study: the Malaysian Cohort study experience. *Int J Environ Res Public Health* 2021; 18: 7216.
- Australian Government National Health and Medical Research Council (NHMRC). *Biobanks information paper 2010*. Canberra: National Health and Medical Research Council. <https://www.nhmrc.gov.au/about-us/publications/biobanks-information-paper> (2010, accessed 8 May 2024).
- Organisation for Economic Co-Operation and Development (OECD). A framework for biotechnology statistics, <https://web.archive.oecd.org/2012-06-14/76870-aframeworkforbiotechnologystatistics.htm> (2005, accessed 8 May 2024).
- Chalmers D, Nicol D, Kaye J, et al. Has the biobank bubble burst? Withstanding the challenges for sustainable biobanking in the digital era. *BMC Med Ethics* 2016; 17: 39.
- Adamson L and Graves A. Cohort management: developing and maintaining participant databases in longitudinal studies. *Int J Mult Res Approaches* 2007; 1: 147–155.
- Mendy M, Caboux E, Lawlor RT, et al. *Common minimum technical standards and protocols for biobanks dedicated to cancer research*. IARC Technical Report, No. 44. SECTION 3, Recommendations for biobanks. Lyon: International Agency for Research on Cancer, 2017.
- Chojenta C, Mooney R and Warner-Smith P. Accessing and disseminating longitudinal data: protocols and policies. *Int J Mult Res Approaches* 2007; 1: 104–113.
- Adamson L and Chojenta C. Developing relationships and retaining participants in a longitudinal study. *Int J Mult Res Approaches* 2007; 1: 137–146.
- Hunt JR and White E. Retaining and tracking cohort study members. *Epidemiol Rev* 1998; 20: 57–70.
- Spicker D and Wallace MP. Measurement error and precision medicine: error-prone tailoring covariates in dynamic treatment regimes. *Stat Med* 2020; 39: 3732–3755.
- Abdullah N, Ismail N, Abd Jalal N, et al. Prevalence of anaemia and associated risk factors amongst the Malaysian Cohort participants. *Ann Hematol* 2020; 99: 2521–2527.
- Abdullah N, Goh Y, Othman R, et al. Stability of glycated haemoglobin (HbA1c) measurements from whole blood samples kept at  $-196^{\circ}\text{C}$  for seven to eight years in the Malaysian Cohort study. *J Clin Lab Anal* 2023; 37: 1–7.
- Shahar S, Shahril MR, Abdullah N, et al. Development and relative validity of a semiquantitative food frequency questionnaire to estimate dietary intake among a multi-ethnic population in the Malaysian Cohort project. *Nutrients* 2021; 13: 1163.
- Genes N, Violante S, Cetrangol C, et al. From smartphone to EHR: a case report on integrating patient-generated health data. *NPJ Digit Med* 2018; 1: 23.
- Batko K and Slezak A. The use of big data analytics in healthcare. *J Big Data* 2022; 9: 3.
- Raghupathi W and Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014; 2: 3.
- Lopez-Olivo MA and Suarez-Almazor ME. Digital patient education and decision aids. *Rheum Dis Clin N Am* 2019; 45: 245–256.
- Krist AH, Tong ST, Aycock RA, et al. Engaging patients in decision-making and behavior change to promote prevention. *Stud Health Technol Inform* 2017; 240: 284–302.
- Nayan NA, Hamid A, Suboh H, et al. Cardiovascular disease prediction from electrocardiogram by using machine learning. *Int J Online Biomed Eng (iJOE)* 2020; 16: 34.
- Suboh MZ, Nayan NA, Abdullah N, et al. Cardiovascular disease prediction among the Malaysian Cohort participants using electrocardiogram. *Comput Mater Continua* 2022; 71: 1111–1132.
- Mohammed Nawi A, Chin S-F and Jamal R. Simultaneous analysis of 25 trace elements in micro volume of human

- serum by inductively coupled plasma mass spectrometry (ICP-MS). *Pract Lab Med* 2020; 18: e00142.
26. Mohammed Nawi A, Chin SF, Mazlan L, et al. Delineating colorectal cancer distribution, interaction, and risk prediction by environmental risk factors and serum trace elements. *Sci Rep* 2020; 10: 18670.
  27. Cios KJ and William Moore G. Uniqueness of medical data mining. *Artif Intell Med* 2002; 26: 1–24.
  28. Rumsfeld JS, Joynt KE and Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol* 2016; 13: 350–359.
  29. Shilo S, Rossman H and Segal E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat Med* 2020; 26: 29–38.
  30. Cirillo D and Valencia A. Big data analytics for personalized medicine. *Curr Opin Biotechnol* 2019; 58: 161–167.
  31. Price WN and Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019; 25: 37–43.
  32. van Staa T-P, Goldacre B, Buchan I, et al. Big health data: the need to earn public trust. *BMJ* 2016; 354: i3636.
  33. Alyass A, Turcotte M and Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics* 2015; 8: 33.
  34. LeCun Y, Bengio Y and Hinton G. Deep learning. *Nature* 2015; 521: 436–444.
  35. Haque M, Islam T, Sartelli M, et al. Prospects and challenges of precision medicine in lower- and middle-income countries: a brief overview. *Bangladesh J Medical Sci* 2019; 19: 32–47.
  36. Park K. The use of real-world data in drug repurposing. *Transl Clin Pharmacol* 2021; 29: 117.
  37. Glicksberg BS, Li L, Chen R, et al. Leveraging big data to transform drug discovery. *Methods Mol Biol* 2019; 1939: 91–118.
  38. Mohanty S, Harun AI, Rashid M, et al. Application of artificial intelligence in COVID-19 drug repurposing. *Diab Metabol Syndr Clin Res Rev* 2020; 14: 1027–1031.
  39. Zhou Y, Wang F, Tang J, et al. Artificial intelligence in COVID-19 drug repurposing. *Lancet Digit Health* 2020; 2: e667–e676.
  40. Sonawane AR, Weiss ST, Glass K, et al. Network medicine in the age of biomedical big data. *Front Genet* 2019; 10: 445334.
  41. Li J and Carayon P. Health care 4.0: a vision for smart and connected health care. *IIEE Trans Healthc Syst Eng* 2021; 11: 1–10.
  42. Al-Jaroodi J, Mohamed N and Abukhousa E. Health 4.0: on the way to realizing the healthcare of the future. *IEEE Access* 2020; 8: 211189–211210.
  43. Juhn Y and Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol* 2020; 145: 463–469.
  44. Houssein EH, Mohamed RE and Ali AA. Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. *Sci Rep* 2023; 13: 7173.
  45. Nunez J-J, Leung B, Ho C, et al. Predicting the survival of patients with cancer from their initial oncology consultation document using natural language processing. *JAMA Netw Open* 2023; 6: e230813.
  46. Lee RY, Kross EK, Torrence J, et al. Assessment of natural language processing of electronic health records to measure goals-of-care discussions as a clinical trial outcome. *JAMA Netw Open* 2023; 6: e231204.
  47. Sung S, Chen C, Pan R, et al. Natural language processing enhances prediction of functional outcome after acute ischemic stroke. *J Am Heart Assoc* 2021; 10: e023486.
  48. Lee RY, Brumback LC, Lober WB, et al. Identifying goals of care conversations in the electronic health record using natural language processing and machine learning. *J Pain Symptom Manage* 2021; 61: 136–142.e2.
  49. Kingsmore SF, Cakici JA, Clark MM, et al. A randomized, controlled trial of the analytic and diagnostic performance of singleton and trio, rapid genome and exome sequencing in ill infants. *Am J Human Genet* 2019; 105: 719–733.
  50. Kumar G, Basri S, Imam AA, et al. Data harmonization for heterogeneous datasets in big data—a conceptual model. In: *Software engineering perspectives in intelligent systems*. Cham: Springer, 2020, pp.723–734.
  51. James KN, Clark MM, Camp B, et al. Partially automated whole-genome sequencing reanalysis of previously undiagnosed pediatric patients can efficiently yield new diagnoses. *NPJ Genom Med* 2020; 5: 33.
  52. Chafai N, Bonizzi L, Botti S, et al. Emerging applications of machine learning in genomic medicine and healthcare. *Crit Rev Clin Lab Sci* 2024; 61: 140–163.
  53. Le NQK. Leveraging transformers-based language models in proteome bioinformatics. *Proteomics* 2023; 23: 2300011.
  54. Wen B, Zeng W, Liao Y, et al. Deep learning in proteomics. *Proteomics* 2020; 20: 1900335.
  55. Majumder EL-W, Billings EM, Benton HP, et al. Cognitive analysis of metabolomics data for systems biology. *Nat Protoc* 2021; 16: 1376–1418.
  56. Cai T, Giannopoulos AA, Yu S, et al. Natural language processing technologies in radiology research and clinical applications. *RadioGraphics* 2016; 36: 176–191.
  57. Poon H, Naumann T, Zhang S, et al. Precision health in the age of large language models. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM, 2023, pp.5825–5826.
  58. Mahendran D, Luo C and McInnes BT. Review: privacy-preservation in the context of natural language processing. *IEEE Access* 2021; 9: 147600–147612.
  59. Garrido-Muñoz I, Montejó-Ráez A, Martínez-Santiago F, et al. A survey on bias in deep NLP. *Appl Sci* 2021; 11: 3184.
  60. Reese JT, Danis D, Caufield JH, et al. On the limitations of large language models in clinical diagnosis. *medRxiv* [Preprint]. 2024: 1–12.
  61. Sheikhalishahi S, Miotto R, Dudley JT, et al. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019; 7: e12239.