OPEN ACCESS

Digital Medicine REPORT

The Youyi Kepler Biobank (YKB): A hospital-based multimodal longitudinal data resource

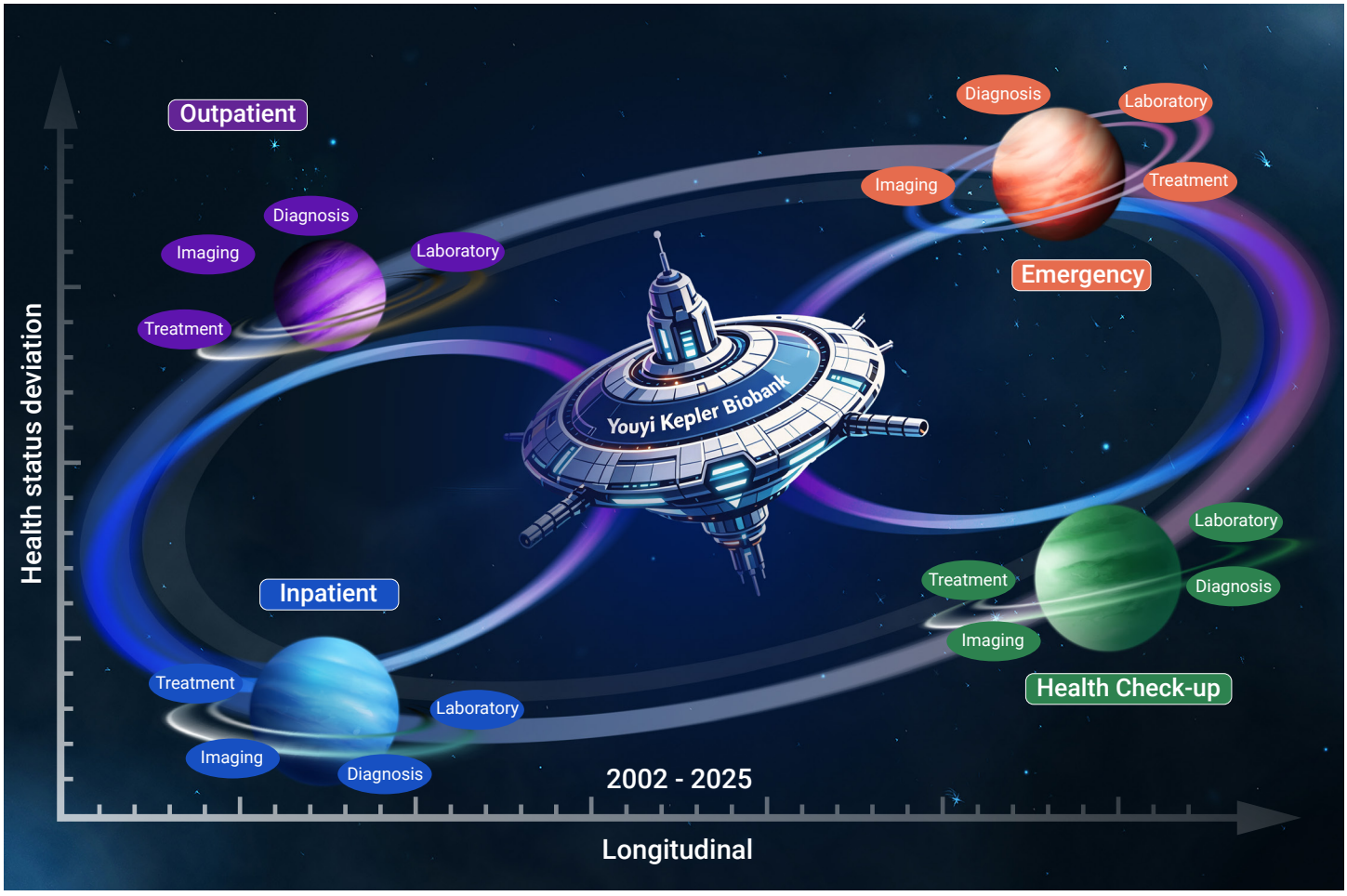
Yanbo Liang,[1,5] Xiaoqing Yuan,[1,5] Ruilin Sun,[2,5] Zhiming Lu,[3] Qi Sun,[3] Guangliang Zhang,[4] Qing Liu,[4] Xinyuan Li,[4] Yabing Hou,[2] Zhenchang Wang,[1,*] Jia Li,[1,*] and Han Lv[1,2,3,4,*]

*Correspondence: cjr.wzhch@vip.163.com (Z.W.); Currylee92@163.com (J.L.); chrislvhan@126.com (H.L.)

Received: July 20, 2026; Accepted: October 10, 2026; Published Online: October 14, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100066

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- The Youyi Kepler Biobank integrates long-term health check-up and clinical data from 297,127 individuals.
- It links repeated health measurements with subsequent outpatient, emergency, inpatient, and imaging records.
- The resource enables investigation of health-to-disease trajectories and early disease warning signals.

REPORT Digital Medicine

OPEN ACCESS

The Youyi Kepler Biobank (YKB): A hospital-based multimodal longitudinal data resource

Yanbo Liang,[1,5] Xiaoqing Yuan,[1,5] Ruilin Sun,[2,5] Zhiming Lu,[3] Qi Sun,[3] Guangliang Zhang,[4] Qing Liu,[4] Xinyuan Li,[4] Yabing Hou,[2] Zhenchang Wang,[1,*] Jia Li,[1,*] and Han Lv[1,2,3,4,*]

[1]Department of Radiology, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China
[2]Department of Health Data Science, School of Medical Technology, Capital Medical University, Beijing 101300, China
[3]Precision and Intelligence Medical Imaging Lab, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China
[4]Health Information Management Department, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China
[5]These authors contributed equally
*Correspondence: cjr.wzhch@vip.163.com (Z.W.); Currylee92@163.com (J.L.); chrislvhan@126.com (H.L.)

Received: July 20, 2026; Accepted: October 10, 2026; Published Online: October 14, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100066

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Liang Y., Yuan X., Sun R., et al. (2026). The Youyi Kepler Biobank (YKB): A hospital-based multimodal longitudinal data resource. *The Innovation Informatics* **2**:100066.

Health status is inherently multidimensional and dynamic. The transition from health and subhealth to disease is a long-term process that requires the integration of health screening data with subsequent clinical records. We established the Youyi Kepler Biobank (YKB), a hospital-based multimodal longitudinal data resource derived from the health check-up population of Beijing Friendship Hospital. Health check-ups were linked with subsequent outpatient, emergency, and inpatient encounters. The database incorporates demographic characteristics, anthropometric measurements, laboratory tests, diagnoses, treatments, imaging examinations, neuropsychological assessments, and other examinations. A total of 297,127 individuals who underwent health check-ups between June 2002 and August 2025 were included, with a mean age of 41.2 ± 14.1 years. Among them, 50.5% were male. 24.3% of participants had 3-9 health check-ups. Among the clinical records, 57.3% of participants had at least one outpatient encounter, 13.6% had an emergency encounter, and 3.7% had an inpatient encounter. The database captured age-related distributions of anthropometric and laboratory indicators. Hyperlipidemia and hypertension were the most frequently documented chronic diseases. Imaging data were also available, including digital radiography, computed tomography, magnetic resonance imaging, and nuclear medicine examinations. The YKB provides a longitudinal framework for linking health check-ups with subsequent disease. The YKB also serves as a data infrastructure for artificial intelligence (AI), providing longitudinal, multimodal, and reusable medical data to support AI for science applications in health trajectory modeling and disease risk prediction. Its longitudinal and multimodal structure may further provide a data foundation for future medical world models that represent evolving patient states and learn their temporal dynamics.

INTRODUCTION

The acceleration of global population aging and increasing burden of chronic diseases pose major challenges to public health systems.[1,2] The modern medical model is undergoing a shift from disease-centered treatment to health-centered prevention.[3] This transformation emphasizes that disease prevention must be advanced to the preclinical or even the fully healthy stage. Therefore, comprehensively and systematically delineating the continuous dynamic trajectory from healthy status to disease occurrence, progression, and outcome, and identifying the risk factors driving disease evolution, are of critical public health importance for optimizing resource allocation, reducing healthcare costs, and improving population health.

Population-based research has increasingly emphasized concepts such as health indicators and health management,[4-6] yet still faces challenges that hinder systematic analysis of the disease continuum. On one hand, traditional studies mostly rely on cross-sectional data from clinically diagnosed patients.[7,8] Cross-sectional data capture health status only at a single time point, failing to reflect dynamic trajectories of health indicators over time and reveal evolutionary patterns of preclinical states preceding disease onset. Clinical diagnosis and treatment data focus on patients who have already developed disease, with disease-specific bias, making it difficult to cover the critical transition from health to disease and to construct a complete disease continuum.

On the other hand, even though existing longitudinal studies have explored dynamic changes in health indicators,[9-11] few databases simultaneously include long-term health check-up records and real-world clinical visits. Existing databases are commonly derived either from general health check-up or community-based populations, or from clinically diagnosed patient cohorts. The former can capture repeated measurements, but often lack systematic linkage to subsequent real-world medical visits and clinical diagnoses. The latter contain detailed clinical information on diagnosis, treatment, medication, and prognosis, but usually provide limited insight into long-term health trajectories before clinical recognition. Consequently, health check-up data and clinical visit data are frequently separated, making it difficult to determine how healthy or subhealthy individuals progress toward clinically recognized disease.[12]

Health status is inherently multidimensional and dynamic. Just as the astronomer Johannes Kepler deciphered the celestial laws governing planetary motion, we propose that an individual's health can be viewed as a dynamic position within a vast, multi-organ state space. By aligning longitudinal data across this Health Galaxy, we can reconstruct unique health trajectories, mapping the direction and velocity of movement to reveal the hidden patterns of transition before and after disease recognition.[13] Complex chronic diseases, such as cardiovascular and cerebrovascular diseases, metabolic disorders, respiratory diseases, and malignant tumors, often evolve over years and may present with multi-system, multi-organ, and multi-morbidity patterns, bridging the continuum from health to disease.[14,15] Therefore, a database that originates from a health check-up population and navigates through real-world clinical visits serves as a critical vessel, linking subtle physiological shifts observed in health to the definitive onset of disease. This database not only preserves the advantages of long-term health surveillance, but also enables the stages of health, subhealth, preclinical states, disease, and disease progression to be closely connected within the same data framework.[16,17]

Artificial intelligence (AI) for science provides powerful approaches for extracting patterns from long-term, dynamic, large-scale, and heterogeneous biomedical data.[18] However, the performance and reliability of these approaches depend heavily on the availability of high-quality longitudinal and multimodal datasets.[19] This has increased attention to data for AI perspective, which emphasizes the systematic linkage, standardization, curation, and reuse of clinical data.[20,21] Therefore, data for AI provides the data foundation for AI for science, forming a continuous cycle in which high-quality data enable scientific discovery, while newly generated knowledge further improves data organization and model development.[22] An emerging extension of longitudinal medical AI is the concept of medical world models. Originating from world-model research in artificial intelligence, these models aim to learn representations of system states and the temporal dynamics governing transitions between states. In medicine, this framework can be extended to represent evolving patient states and, ultimately, to simulate future clinical trajectories under different interventions. Such development depends on longitudinal, multimodal data that capture repeated physiological measurements, clinical events, treatments, and imaging information over time.

In this context, we established the Youyi Kepler Biobank (YKB), a hospital-

based multimodal longitudinal data resource based on the health check-up population of Beijing Friendship Hospital. Functioning as a central command module in the cosmos of healthcare, the YKB framework acts as a gravitational nexus. Acting as a vessel, it seamlessly docks disparate data streams, including diagnoses, laboratory tests, imaging, outpatient/inpatient/emergency records, and health check-up results, into a unified constellation. By bridging these isolated information silos, the framework transforms fragmented signals into a coherent, longitudinal narrative of the patient journey. This database aims to depict the long-term dynamic trajectories of health indicators, identify early warning signals before disease occurrence, and establish dynamic risk prediction models based on multimodal data fusion. Ultimately, this resource will provide a scientific basis for developing precision prevention strategies and evaluating early screening efficacy based on real-world evidence, offering robust data support for promoting the transformation from passive disease treatment to active disease prevention.

MATERIALS AND METHODS

Study design

The YKB is a hospital-based, multimodal, longitudinal real-world database established at Beijing Friendship Hospital affiliated to Capital Medical University, a large tertiary general hospital in Beijing, China. The database was designed to characterize the continuum from health and subhealth to clinically diagnosed disease by linking health check-up records with subsequent clinical diagnosis and treatment records within the same hospital information environment.

Health check-up records were used as the entry point of the database, with the first eligible health check-up defined as the baseline visit. Through patient-level and visit-level identifiers, these records were linked with subsequent outpatient, emergency, and inpatient encounters, as well as diagnoses, laboratory measurements, clinical assessments, treatments, medications, and imaging examinations arranged along the individual timeline. This design enables patient-level longitudinal tracking and visit-level alignment of clinical events.

The first eligible health check-up was defined as the index health check-up. Variables available at the index examination were extracted from that visit. When a specific variable was unavailable at the index health check-up, the earliest valid measurement of that variable from subsequent health check-up records was used for descriptive characterization. Consequently, measurements of different variables were not necessarily obtained on the same date.

Study population

The participants included individuals who underwent health check-up at the Health Management Center of Beijing Friendship Hospital between June 2002 and August 2025. The inclusion criterion was individuals with at least one complete health check-up record. Individuals were excluded if key baseline data were missing or if the records could not be anonymized.

Data integration and multimodal database construction

Integration of historical records. The hospital data lake was established in 2018, whereas clinical information systems had undergone several upgrades before that time. Consequently, part of the historical data was stored in different systems. To overcome the interoperability barriers across database versions, InterSystems IRIS and global mapping technology were used to integrate records from both historical and current systems. The archived medical records were further used as reference sources to validate and consolidate medical information across different periods, thereby enabling unified longitudinal linkage of health check-up and clinical visit data.

Integration of health check-up and clinical databases. The individuals with health check-ups may have multiple subsequent visits to the hospital, including outpatient, emergency, and inpatient encounters. Using the patient master index as the core linkage mechanism, the multimodal clinical data from the hospital information system (HIS), laboratory test systems, pathology systems, psychological assessment platforms, and other information systems were integrated. Data sources included health check-up, outpatient, emergency and inpatient records. Through this linkage, health check-up records were used to characterize health or pre-disease states, whereas clinical data were used to capture disease states and subsequent medical outcomes, thereby enabling continuous tracking of the transition from health to disease.

Construction of the multimodal database. Based on clinical order information, imaging data of the included individuals were further retrieved from the Picture Archiving and Communication System (PACS), including digital radiography (DR), computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), nuclear medicine examinations, fundus imaging, and so on. Visit indices were used to match and integrate image files with corresponding clinical records, thereby expanding the database from structured clinical and laboratory text data to a multimodal health research resource. In addition, an application programming interface (API)-based platform was developed to connect the database with multimodal data systems, enabling cross-platform retrieval, linkage, and access to structured clinical data, free-text reports, and imaging data.

The resulting database was organized into 7 primary modules: visit information, general health check-up conclusions, detailed health check-up items, clinical examinations, laboratory tests, medical orders, and diagnoses.

Data phenotypes. The YKB contains multidimensional phenotypic measurements derived from routine health check-ups and subsequent clinical visits. The database comprises physical and anthropometric examinations, specialist assessments, laboratory tests, imaging examinations, electrocardiography and other functional examinations. Because the database was constructed from real-world hospital-based health examination records, the specific items available for each participant varied according to the selected health check-up package, clinical indication, and subsequent medical encounters.

Physical examination and anthropometric examinations. Physical examination was performed by trained doctors according to routine clinical procedures. Anthropometric measurements included height, body weight, body mass index (BMI), and waist circumference (WC). Height and weight were measured with participants wearing light clothing and without heavy accessories. BMI was calculated as body weight in kilograms divided by height in meters squared. WC was measured at midpoint between the lower rib margin and the iliac crest during minimal respiration.

Specialist assessments. Internal medicine assessments generally included cardiopulmonary auscultation, abdominal examination, and evaluation of common chronic disease-related symptoms or signs. Surgical assessments included inspection and palpation of the skin, superficial lymph nodes, spine, extremities, thyroid region, breast, and other regions according to the selected examination package. Otorhinolaryngology, ophthalmology, stomatology, and other specialist assessments were also available for participants.

Laboratory tests. Laboratory tests were performed by the hospital clinical laboratory according to routine standardized procedures. The laboratory domain covered major test categories, including routine blood tests, urine and stool tests, liver and kidney function, glucose and lipid metabolism, electrolytes and minerals, coagulation-related indicators, tumor markers, inflammatory and immune markers, endocrine indicators, infectious disease tests, and other biochemical assays. Blood samples were generally collected in the morning after overnight fasting, and urine or stool samples were collected when included in the examination package.

Imaging examinations. The imaging examinations included DR, CT, MRI, US, fundus imaging, nuclear medicine, and other examination types. Common CT scans include the head, chest, abdomen, spine, or limbs according to the clinical protocol. MRI examinations could involve the brain, spine, joints, or abdomen depending on clinical indications. US examinations included the liver, gallbladder, spleen, kidneys, thyroid, carotid arteries, breast, prostate, uterus, adnexa, and superficial lymph nodes. Nuclear medicine examinations included positron emission tomography/computed tomography (PET/CT), single-photon emission computed tomography/computed tomography (SPECT/CT), bone scintigraphy, myocardial perfusion imaging, thyroid imaging, renal dynamic imaging, and other radionuclide-based functional examinations. Ophthalmic imaging and fundus examinations were also recorded when performed.

Neuropsychological assessments and other functional examinations. Neuropsychological and mental health assessments were available according to health check-up packages and clinical indications. Cognitive and

Table 1. Basic characteristics of participants.

Characteristics	Overall	Male	Female
Age,[a] years, mean ± SD	41.2 ± 14.1	42.0 ± 14.0	40.5 ± 14.1
Sex, No. (%)			
Male	150,147 (50.5%)	150,147 (100.0%)	/
Female	146,801 (49.4%)	/	146,801 (100.0%)
Unknown	179 (0.1%)	/	/
Health check-up times, No. (%)			
1-2 times	210,703 (70.9%)	107,759 (71.8%)	102,777 (70.0%)
3-9 times	72,308 (24.3%)	36,357 (24.2%)	35,939 (24.5%)
≥10 times	14,116 (4.8%)	6,031 (4.0%)	8,085 (5.5%)
Period of first health check-up, No. (%)			
2002-2010	31,378 (10.6%)	17,134 (11.4%)	14,149 (9.6%)
2011-2015	83,065 (28.0%)	43,092 (28.7%)	39,891 (27.2%)
2016-2020	98,063 (33.0%)	48,246 (32.1%)	49,816 (33.9%)
2021-2025	84,621 (28.5%)	41,675 (27.8%)	42,945 (29.3%)
Systolic blood pressure,[b] No. (%)			
<120 mmHg	98,873 (33.3%)	38,128 (25.4%)	60,745 (41.4%)
120-140 mmHg	77,369 (26.0%)	48,006 (32.0%)	29,362 (20.0%)
>140 mmHg	22,632 (7.6%)	13,192 (8.8%)	9,440 (6.4%)
Unknown	98,253 (33.1%)	50,821 (33.8%)	47,254 (32.2%)
Diastolic blood pressure,[b] No. (%)			
<60 mmHg	11,797 (4.0%)	3,003 (2.0%)	8,794 (6.0%)
60-90 mmHg	170,999 (57.6%)	84,387 (56.2%)	86,611 (59.0%)
>90 mmHg	16,084 (5.4%)	11,940 (8.0%)	4,144 (2.8%)
Unknown	98,247 (33.1%)	50,817 (33.8%)	47,252 (32.2%)
Heart rate,[b] No. (%)			
<60 beats/min	6,317 (2.1%)	4,207 (2.8%)	2,110 (1.4%)
60-100 beats/min	190,952 (64.3%)	94,367 (62.8%)	96,584 (65.8%)
>100 beats/min	1,597 (0.5%)	759 (0.5%)	838 (0.6%)
Unknown	98,261 (33.1%)	50,814 (33.8%)	47,269 (32.2%)
Body mass index,[b] No. (%)			
<18.5 kg/m^2	8,456 (2.8%)	1,297 (0.9%)	7,159 (4.9%)
18.5-23.9 kg/m^2	89,235 (30.0%)	30,126 (20.1%)	59,108 (40.3%)
24.0-27.9 kg/m^2	68,082 (22.9%)	44,610 (29.7%)	23,472 (16.0%)
≥28.0 kg/m^2	31,210 (10.5%)	22,257 (14.8%)	8,953 (6.1%)
Unknown	100,144 (33.7%)	51,857 (34.5%)	48,109 (32.8%)
Smoking status,[b] No. (%)			
Never smokers	87,511 (29.5%)	38,108 (25.4%)	49,386 (33.6%)
Current smokers	22,379 (7.5%)	21,251 (14.2%)	1,119 (0.8%)
Past smokers	2,563 (0.9%)	2,481 (1.7%)	82 (0.1%)
Unknown	184,674 (62.2%)	88,307 (58.8%)	96,214 (65.5%)
Drinking status,[b] No. (%)			
Never drinkers	76,132 (25.6%)	29,631 (19.7%)	46,485 (31.7%)

Table 1. *(Continued)*

Characteristics	Overall	Male	Female
Current drinkers	35,660 (12.0%)	31,565 (21.0%)	4,086 (2.8%)
Past drinkers	681 (0.2%)	657 (0.4%)	24 (0.0%)
Unknown	184,654 (62.1%)	88,294 (58.8%)	96,206 (65.5%)

Abbreviations: SD, standard deviation. [a]These data were taken from the first health check-up record. [b]For each variable, the earliest available valid measurement in the health check-up records was used when the value was unavailable at the index health check-up.

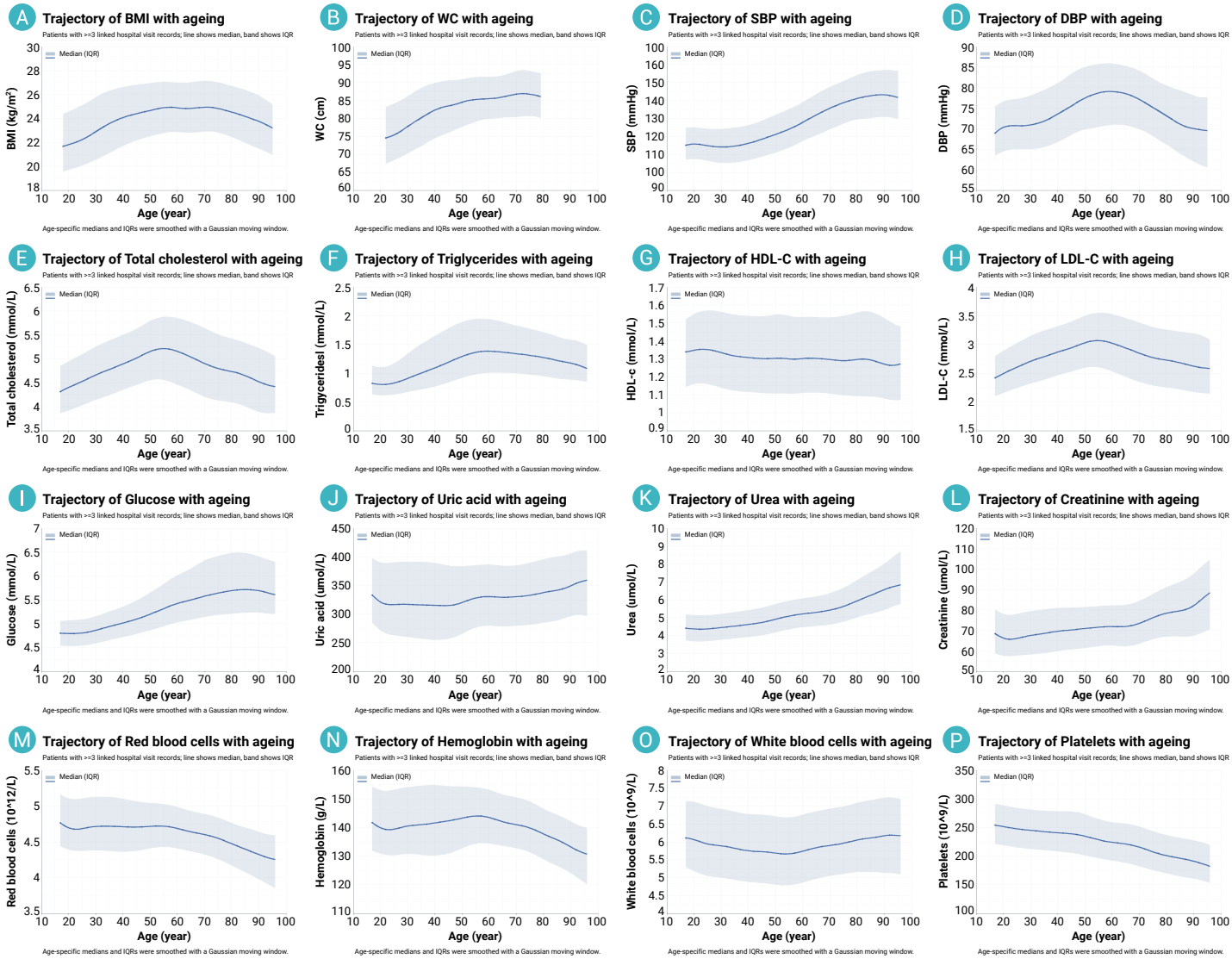


Figure 1. Age-related distributions of anthropometric and laboratory indicators The solid curves represent the median values, with the shaded areas indicating the interquartile ranges within each age group. Abbreviations: BMI, body mass index; WC, waist circumference; SBP, systolic blood pressure; DBP, diastolic blood pressure; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; IQR, interquartile range.

dementia-related assessments included the Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA). Anxiety was assessed using the Hamilton Anxiety Rating Scale (HAMA) and the anxiety subscale of the Symptom Checklist-90 (SCL-90), whereas depressive symptoms were evaluated using the Hamilton Depression Rating Scale (HAMD) and the depression subscale of the SCL-90. Sleep quality was assessed using the Pittsburgh Sleep Quality Index (PSQI). These assessments were administered by trained healthcare professionals according to the corresponding standardized procedures. Other functional and auxiliary examinations were also available, including pulmonary function tests, bone density measurements, body composition analysis, and vascular function assessments.

RESULTS

Basic characteristics of participants

The basic characteristics of the health check-up individuals are shown in Table 1. A total of 297,127 individuals were included in the health check-up population, including 150,147 males (50.5%), 146,801 females (49.4%), and 179 individuals with unknown sex (0.1%). The mean age at the first health check-up was 41.2 ± 14.1 years, with mean ages of 42.0 ± 14.0 years for males and 40.5 ± 14.1 years for females.

Repeated health check-ups were available for a substantial proportion of participants. Overall, 210,703 individuals (70.9%) had 1-2 health check-ups,

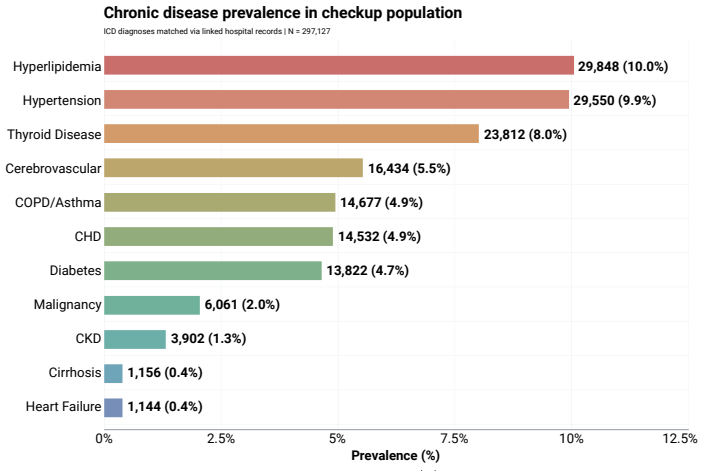


Figure 2. Spectrum of chronic disease diagnoses The bars show the number and proportion of participants with at least one corresponding documented diagnosis during the available observation period. Individual participants could contribute to more than one disease category. Abbreviations: COPD, chronic obstructive pulmonary disease; CHD, coronary heart disease; CKD, chronic kidney disease.

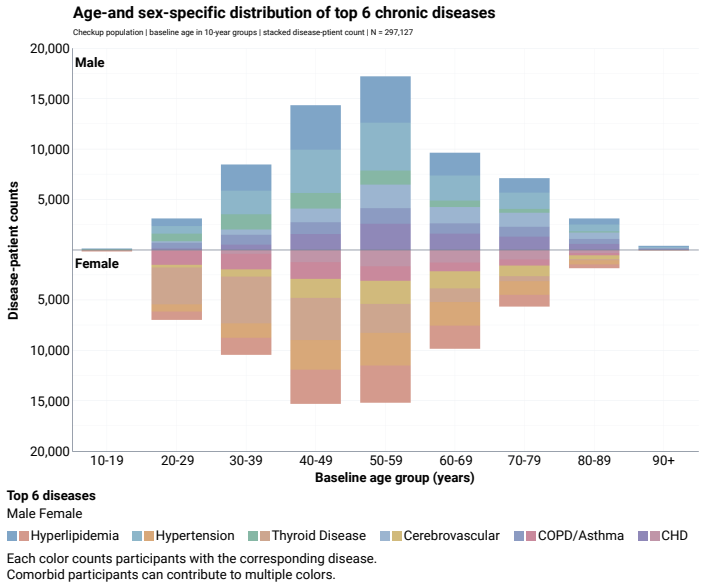


Figure 3. Age- and sex-specific distributions of common chronic diseases The mirrored stacked bar chart shows the numbers of male and female participants with the six most frequently documented chronic diseases across baseline age groups. Bars above the horizontal axis represent male participants, whereas bars below the horizontal axis represent female participants. Different colors indicate hyperlipidemia, hypertension, thyroid disease, cerebrovascular disease, chronic obstructive pulmonary disease or asthma, and coronary heart disease. Age was defined at the first eligible health check-up. Abbreviations: COPD, chronic obstructive pulmonary disease; CHD, coronary heart disease.

72,308 (24.3%) had 3-9 health check-ups, and 14,116 (4.8%) had 10 or more health check-ups.

The median follow-up duration from the index health check-up to the last available dated record was 1.3 years [interquartile range (IQR), 0.0-5.0 years]. Overall, 74,345 participants (25.0%) had a follow-up duration of ≥5 years, and 29,262 (9.8%) had a follow-up duration of ≥10 years. The median number of health check-up time points was 1 (IQR, 1-3), and 131,065 participants (44.1%) had at least two health check-up dates. In addition, 109,742 participants (36.9%) had at least two serial clinical encounters, and 85,765 (28.9%) had at least two serial imaging examination dates.

The database includes participants enrolled across multiple periods, with the largest proportion first examined between 2016 and 2020. By period of first health check-up, 98,063 participants (33.0%) entered the database during 2016-2020.

Visit types were defined according to whether participants had at least one corresponding encounter type in the linked hospital records. All participants had at least one health check-up record. In addition, 40,375 participants (13.6%) had at least one emergency encounter, 11,056 participants (3.7%) had at least one inpatient encounter, and 170,314 participants (57.3%) had at least one outpatient encounter. Emergency encounters were recorded in 15.5% of females and 11.8% of males, inpatient encounters were recorded in 4.9% of females and 2.6% of males, while outpatient encounters were recorded in 59.8% of females and 54.9% of males.

For each health indicator, the value recorded at the index health check-up was used when available; otherwise, the earliest subsequent valid measurement from the health check-up records was used. For systolic blood pressure (SBP), 98,873 participants (33.3%) had values <120 mmHg, 77,369 (26.0%) had values between 120 and 140 mmHg, and 22,632 (7.6%) had values >140 mmHg. For diastolic blood pressure (DBP), 11,797 participants (4.0%) had values <60 mmHg, 170,999 (57.6%) had values between 60 and 90 mmHg, and 16,084 (5.4%) had values >90 mmHg. Heart rate was within 60-100 beats per minute in 190,952 participants (64.3%), below 60 beats per minute in 6,317 (2.1%), and above 100 beats per minute in 1,597 (0.5%). Overall, 8,456 participants (2.8%) had BMI <18.5 kg/m^2, 89,235 (30.0%) had BMI between 18.5 and 23.9 kg/m^2, 68,082 (22.9%) had BMI between 24.0 and 27.9 kg/m^2, and 31,210 (10.5%) had BMI ≥28.0 kg/m^2. Compared with females, males had higher proportions of overweight and obesity.

Lifestyle information showed marked sex differences. Current smoking was reported in 22,379 participants (7.5%), with a much higher proportion in males than females (14.2% vs. 0.8%). Current drinking was reported in 35,660 participants (12.0%), also more frequently in males than females (21.0% vs. 2.8%).

Age-related distributions of anthropometric and laboratory indicators

Age-related distributions of anthropometric and laboratory indicators are shown in Figure 1. Overall, these indicators exhibited heterogeneous age-related patterns. Several indicators, such as WC and SBP, increased from early adulthood to middle or older age, whereas BMI, DBP, and lipid levels tended to stabilize or decline in the oldest age groups. Renal markers generally increased at advanced ages, while red blood cell count, hemoglobin, and platelet count showed gradual reductions among older participants.

Clinical diagnosis spectrum

The clinical diagnosis spectrum based on International Classification of Diseases (ICD)-coded diagnosis records is shown in Figure 2. Hyperlipidemia and hypertension were the most frequently documented chronic conditions, involving 29,848 individuals (10.0%) and 29,550 individuals (9.9%), respectively. Thyroid diseases were documented in 23,812 individuals (8.0%), followed by cerebrovascular diseases in 16,434 (5.5%), chronic obstructive pulmonary disease or asthma in 14,677 (4.9%), coronary heart disease in 14,532 (4.9%), and diabetes in 13,822 (4.7%).

Age- and sex-specific distributions of common chronic diseases

The age- and sex-specific distributions of the six most frequently documented chronic diseases are presented in Figure 3. The number of participants with documented chronic diseases was relatively low among individuals younger than 30 years and increased substantially from middle adulthood. For both males and females, the largest numbers of diagnoses were observed mainly between 40 and 69 years of age.

Among males, the number of chronic conditions reached its highest level in the 50-59 years age group and declined in older age groups. Among females, high numbers of diagnoses were observed in the 40-49 and 50-59 year age groups. Hyperlipidemia and hypertension contributed a large proportion of the diagnoses across most middle-aged and older groups, while thyroid disease, cerebrovascular disease, chronic obstructive pulmonary disease or asthma, and coronary heart disease showed different age and sex distributions.

Imaging data

A total of 1,339,906 imaging records were classified according to imaging modality (Figure 4A). CT was the most frequently represented modality, accounting for 677,264 report records (50.55%), followed by DR (378,536 records, 28.25%), MRI (169,117 records, 12.62%), and US (110,490 records, 8.25%).

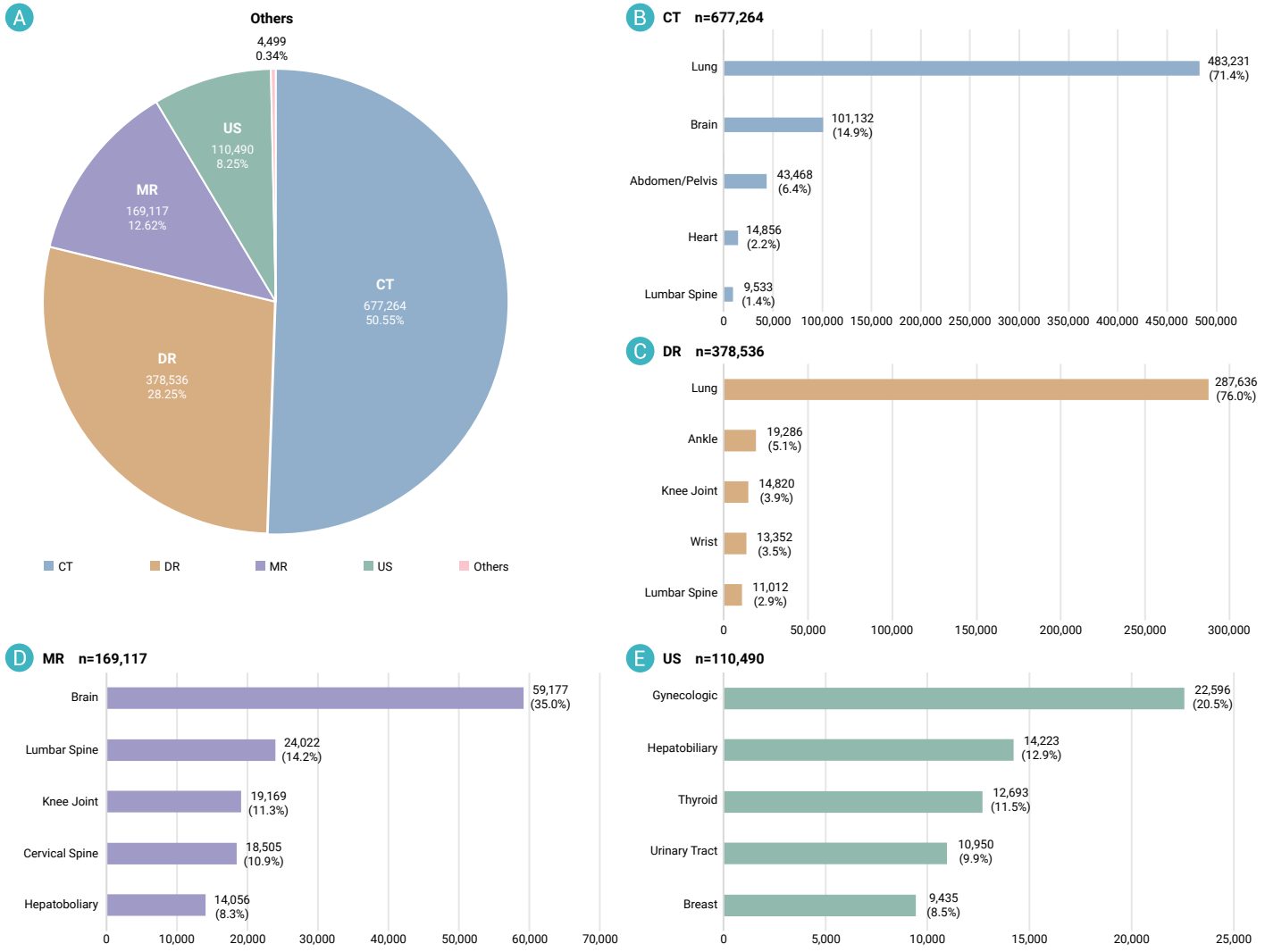


Figure 4. Distribution of imaging records by modality and anatomical site (A) Distribution of imaging records across major imaging modalities. (B-E) The five most frequently recorded anatomical sites within CT, DR, MR, and US, respectively. Abbreviations: CT, computed tomography; DR, digital radiography; MR, magnetic resonance; US, ultrasonography.

The distribution of anatomical sites varied across modalities (Figures 4B-E). Lung examinations predominated in both CT and DR, accounting for 483,231 CT reports (71.4%) and 287,636 DR reports (76.0%). Brain imaging was the second most common CT examination site. For MRI, brain examinations were most frequent, followed by examinations of the lumbar spine, knee joint, cervical spine, and hepatobiliary system. US reports showed a broader distribution across anatomical sites, with gynecologic examinations being the most common (22,596 reports, 20.5%), followed by hepatobiliary, thyroid, urinary tract, and breast examinations.

Moreover, imaging files retrieved from PACS are stored in Digital Imaging and Communications in Medicine (DICOM) format and classified by modality, examination site, and study number to support future image-based quantitative analyses.

DISCUSSION

Principal findings

This data profile describes the design, construction, and preliminary content of the YKB. The database integrates health check-up data and clinical diagnosis and treatment records from a large tertiary hospital over a 23-year period. Its feature is the linkage of health check-up data, which reflect health or preclinical states, with subsequent outpatient, emergency, and inpatient records, which reflect disease states and clinical management. The YKB is established to map the intricate trajectories of human health. In this vast Health Galaxy, where an individual's physiological state is akin to a star navigating the cosmos, the YKB serves as the observatory. It is dedicated to uncovering the hidden patterns and dynamic laws that dictate the journey from health to disease, transforming the complexity of biological data into a navigable map for precision evaluation of health. This architecture enables research on the full continuum from health to subhealth and disease.

Comparison with released cohort databases

Several large-scale cohorts and health databases have been established worldwide. International cohorts are mostly based on non-Chinese populations, and differences in genetic background, lifestyle, diet, environment, and healthcare systems may limit their direct generalizability to Chinese populations. [23-26] In China, the China Kadoorie Biobank (CKB) enrolled approximately 513,000 adults from 10 regions and collected questionnaire, physical examination, and biospecimen data. [27-29] However, access to this database is relatively restricted. The China Health and Retirement Longitudinal Study (CHARLS) is a longitudinal survey of Chinese adults aged 45 years and older, with repeated face-to-face interviews and periodic physical and blood sample measurements.[30,31] Nevertheless, it focuses mainly on middle-aged and older adults and lacks detailed hospital diagnostic, medication, laboratory, and imaging records. The Taizhou Longitudinal Study is a large population cohort, with approximately 200,000 participants and long-term follow-up, mainly focusing on genetic and environmental determinants of chronic diseases,[32] but routine hospital clinical records are not comprehensively captured. The Kailuan Study is an ongoing prospective cohort in Tangshan, enrolling approximately 170,000 adults since 2006 and conducting health check-ups every two years. The cohort was derived from a specific commu-

nity-based population, which may limit generalizability to populations from other regions.[33]

The Health Elements framework suggests that health is shaped by the joint influence of biological, behavioral, social, environmental, and technological factors, and that digital data and AI can help study how these factors change and interact over time.[34] The YKB mainly provides rich biological and clinical information through health check-ups, laboratory tests, imaging examinations, diagnoses, and medical orders. It also provides the basis for data linkage, text structuring, imaging analysis, and risk prediction. This design enables more continuous characterization of the transition from health or pre-disease status to clinically diagnosed disease, and provides a real-world longitudinal data infrastructure for identifying early warning signals, modeling health trajectories, and developing individualized risk prediction models for Chinese populations.

Strengths

The database has several strengths. First, it is established from a hospital health check-up population and links routine health check-up data with subsequent outpatient, emergency, and inpatient clinical encounters. This enables the database to characterize not only baseline status, but also subsequent disease states. Second, the database covers a long-time span from 2002 to 2025, allowing the reconstruction of long-term health trajectories. Third, the database integrates multiple modalities, including structured clinical indicators, laboratory tests, diagnostic records, medical orders, and classified DICOM imaging files for future image-based studies. Fourth, the database is designed as a reusable research infrastructure rather than a disease-specific cohort, allowing serial investigations across chronic disease domains and across different stages of the health-to-disease continuum.

Limitations

Several limitations should be acknowledged. First, the database was constructed from a hospital-based health examination population at a single tertiary hospital in Beijing. Individuals who undergo regular health check-ups may be more health-conscious, have better healthcare access, or be more likely to be employed in organizations providing routine examinations. Therefore, selection bias is possible, and the findings may not be fully generalizable to the entire Chinese population, particularly rural residents, socioeconomically disadvantaged groups, or individuals with limited access to tertiary healthcare. Second, because this is a real-world database derived from routine clinical information systems rather than a prospectively designed research cohort, the availability and frequency of measurements varied across individuals, examination packages, and clinical indications. Third, although the database links health check-up records with outpatient, emergency, inpatient visits within the same institution, medical encounters occurring outside Beijing Friendship Hospital may not be fully captured. Fourth, some important lifestyle, socioeconomic, environmental, and genetic variables are incompletely recorded.

Future directions and potential applications

The YKB provides longitudinal, multimodal real-world data for future studies on health-to-disease transitions. By linking repeated health check-up measurements with subsequent outpatient, emergency, and inpatient clinical encounters, the database can be used to construct long-term trajectories of anthropometric, laboratory, imaging, and functional indicators, and to identify early warning signals preceding disease.

Future studies are being planned. Cohorts for common chronic diseases, including cerebrovascular, respiratory, heart diseases, and diabetes will be established. Different diseases will use different data modules. This strategy follows the idea that complex health questions should be studied with a clear scope and clinically meaningful variables. It can also help keep prediction models understandable and easier to apply in practice.[35]

The database may also support AI-assisted studies. AI methods can be used to detect patterns in repeated indicators, structure free-text reports, extract imaging phenotypes, and build disease-risk models. Future prediction model studies will require external evaluation, transparent reporting, and assessment of model performance and fairness across clinically relevant age and sex subgroups.[36,37]

Beyond conventional risk prediction, the longitudinal and multimodal structure of the YKB may also support future research on medical world models. Repeated health check-ups, laboratory measurements, diagnoses, treatments, clinical encounters, and imaging data can be integrated to construct longitudinal representations of patient states and to model their transitions over time. With further methodological development, such models may eventually enable simulation of disease trajectories under different clinical conditions or interventions.

Future work will also focus on turning this resource into practical outputs. These may include regular reports on major chronic diseases, long-term curves of key health indicators, imaging phenotype datasets, disease-risk dashboards, and clear procedures for collaborative data use. These outputs may help hospitals monitor population health, improve early screening, follow chronic disease trends, and generate real-world evidence.

With continued refinement of data standardization, natural language processing, and imaging analysis, this database is expected to serve as a practical platform for precision prevention and proactive health management.

CONCLUSION

Electronic health records have become an essential component of contemporary healthcare systems, and the longitudinal, multimodal hospital data provide an important opportunity to better understand health-to-disease transitions. The YKB was established to comprehensively integrate and utilize such real-world information by linking health check-up records with subsequent outpatient, emergency, inpatient, laboratory, diagnostic, and imaging data within the same hospital information environment. Formal empirical validation of all potential applications of this database is beyond the scope of the present data profile. Nevertheless, based on its broad coverage of repeated health check-ups and clinical visits, this database has the potential to support both validation of known disease patterns and identification of novel, previously under-recognized associations across functional aspects of medicine and healthcare. Future studies using this resource may identify early warning signals before clinical disease recognition, and develop dynamic risk prediction models to support precision prevention and proactive health management.

REFERENCES

1. Kemoun P., Ader I., Planat-Benard V., et al. (2021). A gerophysiology perspective on healthy ageing. *Ageing Res. Rev.* **73**:101537. DOI:10.1016/j.arr.2021.101537
2. Murray C.J.L. (2024). Findings from the Global Burden of Disease Study 2021. *Lancet* **403**:2259–2262. DOI:10.1016/S0140-6736(24)00769-4
3. Liu J., Li W., Yao H., et al. (2022). Proactive health: An imperative to achieve the goal of Healthy China. *China CDC Wkly.* **4**:799–801. DOI:10.46234/ccdcw2022.156
4. van Sloten T.T., Valentin E., Climie R.E., et al. (2023). Association of cardiovascular health with risk of clinically relevant depressive symptoms. *JAMA Psychiatry* **80**:342–349. DOI:10.1001/jamapsychiatry.2022.5056
5. Aris I.M., Rifas-Shiman S.L., Perng W., et al. (2025). Trajectory of cardiovascular health across childhood and adolescence. *JAMA Cardiol.* **10**:417–426. DOI:10.1001/jamacardio.2024.4022
6. Daskalopoulou C., Koukounari A., Wu Y.-T., et al. (2019). Healthy ageing trajectories and lifestyle behaviour: The Mexican Health and Aging Study. *Sci. Rep.* **9**:11041. DOI:10.1038/s41598-019-47238-w
7. Zhang Y., Liu D., Cai X., et al. (2024). Associations of life's essential 8 with extent of multi-territorial atherosclerotic plaques and stenosis: A cross-sectional study. *BMC Geriatr.* **24**:503. DOI:10.1186/s12877-024-05119-6
8. Zhang Y., Xia Z., Cai X., et al. (2023). Association of metabolic dysfunction-associated fatty liver disease with systemic atherosclerosis: A community-based cross-sectional study. *Cardiovasc. Diabetol.* **22**:342. DOI:10.1186/s12933-023-02083-0
9. Ikram M.A., Brusselle G., Ghanbari M., et al. (2020). Objectives, design and main findings until 2020 from the Rotterdam Study. *Eur. J. Epidemiol.* **35**:483–517. DOI:10.1007/s10654-020-00640-5
10. Bookheimer S.Y., Salat D.H., Terpstra M., et al. (2018). The Lifespan Human Connectome Project in Aging: An overview. *Neuroimage* **185**:335–348. DOI:10.1016/j.neuroimage.2018.10.009
11. Pan Y., Jing J., Cai X., et al. (2020). PolyvasculaR Evaluation for Cognitive Impairment and vaScular Events (PRECISE)-A population-based prospective cohort study: Rationale, design and baseline participant characteristics. *Stroke Vasc. Neurol.* **6**:145–151. DOI:10.1136/svn-2020-000411
12. Lin Y., Chen L., Huang J., et al. (2025). Breadth versus depth: Balancing variables, sample size, and quality in Chinese cohort studies. *BMJ* **391**:e082566. DOI:10.1136/bmj-2024-082566

13. Lv H., Li J., Li D., et al. (2026). Health Galaxy: A multi-organ trajectory-computing framework in type 2 diabetes. *Research (Wash D C)* **9**:1378. DOI:10.34133/research.1378
14. Nature Aging (2024). Multiorgan biological age shows that no organ system is an island. *Nat. Aging* **4**:1182-1183. DOI:10.1038/s43587-024-00690-4
15. Devesa A., Delgado V., Valkovic L., et al. (2025). Multiorgan imaging for interorgan crosstalk in cardiometabolic diseases. *Circ. Res.* **136**:1454–1475. DOI:10.1161/CIRCRESAHA.125.325517
16. Gong W. and Cheng K.K. (2022). Challenges in screening and general health checks in China. *Lancet Public Health* **7**:e989–e990. DOI:10.1016/S2468-2667(22)00207-9
17. Jeong C.-U., Leiby J.S., Kim D., et al. (2025). Artificial intelligence-driven biological age prediction model using comprehensive health checkup data: Development and validation study. *JMIR Aging* **8**:e64473. DOI:10.2196/64473
18. Wang H., Fu T., Du Y., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature* **620**:47–60. DOI:10.1038/s41586-023-06221-2
19. Schwabe D., Becker K., Seyferth M., et al. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: A systematic review. *NPJ Digit. Med.* **7**:203. DOI:10.1038/s41746-024-01196-4
20. Wilkinson M.D., Dumontier M., Aalbersberg I.J.J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**:160018. DOI:10.1038/sdata.2016.18
21. Whang S.E., Roh Y., Song H., et al. (2023). Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal* **32**:791–813. DOI:10.1007/s00778-022-00775-9
22. Tang A.S., Woldemariam S.R., Miramontes S., et al. (2024). Harnessing EHR data for health research. *Nat. Med.* **30**:1847–1855. DOI:10.1038/s41591-024-03074-8
23. Slimani N., Kaaks R., Ferrari P., et al. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study: Rationale, design and population characteristics. *Public Health Nutr.* **5**:1125–1145. DOI:10.1079/PHN2002394
24. Sudlow C., Gallacher J., Allen N., et al. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**:e1001779. DOI:10.1371/journal.pmed.1001779
25. Cook M.B., Sanderson S.C., Deanfield J.E., et al. (2025). Our Future Health: A unique global resource for discovery and translational research. *Nat. Med.* **31**:728–730. DOI:10.1038/s41591-024-03438-0
26. Johnson C.L., Dohrmann S.M., Burt V.L., et al. (2014). National health and nutrition examination survey: Sample design, 2011-2014. *Vital Health Stat.* **2**:1–33. PMID:25569458
27. Chen Z., Chen J., Collins R., et al. (2011). China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**:1652-1666. DOI:10.1093/ije/dyr120.
28. Walters R.G., Millwood I.Y., Lin K., et al. (2023). Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genom.* **3**:100361. DOI:10.1016/j.xgen.2023.100361
29. Li L.-M., Lv J., Guo Y., et al. (2012). The China Kadoorie Biobank: Related methodology and baseline characteristics of the participants. *Chin. J. Epidemiol.* **33**:249–255. DOI:10.3760/cma.j.issn.0254-6450.2012.03.001
30. Zhao Y., Hu Y., Smith J.P., et al. (2012). Cohort profile: The China Health and Retirement Longitudinal Study (CHARLS). *Int. J. Epidemiol.* **43**:61–68. DOI:10.1093/ije/dys203
31. Chen X., Crimmins E., Hu P.P., et al. (2019). Venous blood-based biomarkers in the China Health and Retirement Longitudinal Study: Rationale, design, and results from the 2015 wave. *Am. J. Epidemiol.* **188**:1871–1877. DOI:10.1093/aje/kwz170
32. Wang X., Lu M., Qian J., et al. (2009). Rationales, design and recruitment of the Taizhou Longitudinal Study. *BMC Public Health* **9**:223. DOI:10.1186/1471-2458-9-223
33. Sun J., Hui Y., Li J., et al. (2023). Protocol for Multi-modality MEdical imaging sTudy bAsed on KaiLuan Study (META-KLS): Rationale, design and database building. *BMJ Open* **13**:e067283. DOI:10.1136/bmjopen-2022-067283
34. Li P., Zhou Y., Li Y., et al. (2026). Digital and AI-empowered health elements: An integrated pathway to advancing health. *Health Data Sci.* **6**:0468. DOI:10.34133/hds.0468
35. Cox E.G.M., Meijs D.A.M., Wynants L., et al. (2024). The definition of predictor and outcome variables in mortality prediction models: A scoping review and quality of reporting study. *J. Clin. Epidemiol.* **178**:111605. DOI:10.1016/j.jclinepi.2024.111605
36. Collins G.S., Moons K.G.M., Dhiman P., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**:q902. DOI:10.1136/bmj.q902
37. Arora A., Alderman J.E., Palmer J., et al. (2023). The value of standards for health datasets in artificial intelligence-based applications. *Nat. Med.* **29**:2929–2938. DOI:10.1038/s41591-023-02608-w

FUNDING AND ACKNOWLEDGMENTS

This work was supported by Grant No. 62522119 (Han Lv) from the National Natural Science Foundation of China, Grant No. G202522110 (Han Lv) from the Beijing High-Level Innovation and Entrepreneurship Talent Support Program, young top talent projects, Grant No. L242024 (Han Lv) from Beijing Natural Science Foundation, Grant No. B2408 (Han Lv) from Capital Medical University, and Grant No. 7254539 (Jia Li) from Beijing Natural Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Y.L.: Conceptualization, Visualization, Writing-original draft. X.Y.: Conceptualization, Visualization, Writing-original draft. R.S.: Conceptualization, Methodology. Z.L.: Conceptualization, Visualization. Q.S.: Conceptualization, Methodology. G.Z.: Conceptualization, Methodology. Q.L.: Conceptualization, Methodology. X.L.: Conceptualization, Methodology. Y.H.: Conceptualization, Methodology. Z.W.: Conceptualization, Supervision. J.L.: Conceptualization, Supervision. H.L.: Conceptualization, Investigation, Supervision, Writing-review & editing. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

The database was approved by the Ethics Committee of Beijing Friendship Hospital, Capital Medical University (approval number: 2025-P2-337-01), with a waiver of informed consent. All the procedures were performed in accordance with the principles of the Declaration of Helsinki. The study was registered in the Medical Research Registration Information System (registration No. MR-11-25-071036) and Clinical Trial Registry (registration No. NCT07375303).

DATA AND CODE AVAILABILITY

Data are available from the corresponding author upon reasonable request.