Computer Science and Information Systems 22(4):1757–1775 https://doi.org/10.2298/CSIS241204069L

Formative Interviews for a User-Centered Design Study on Developing an Effective Gateway for Health Research Data Search – Towards a Sustainable Wellbeing Environment

Hsiu-An Lee[1,2], Tung Lin[3], Hsin-I Chen[1], Wei-Chen Liu[1], Yen-Ju Shen[1], Wen-Chang Tseng[1], and Chien-Yeh Hsu[2,4] and Yi-Hsin Yang[1,⋆]

[1] National Institute of Cancer Research, National Health Research Institutes
No.367, Sheng-Li Rd., North District, Tainan, 70456 Taiwan
100510@nhri.edu.tw
denise9306@nhri.edu.tw
q09855213@nhri.edu.tw
a0979251512@gmail.com
gdi89009@nhri.edu.tw
yhyang@nhri.edu.tw
[2] Standards and Interoperability Lab, Smart Healthcare Center of Excellence
No.365, Mingde Rd., Peitou Dist., Taipei City 112303, Taiwan
[3] Island Design Lab, F7, No.27, Ln. 66, Sec. 4, Heping E. Rd., Wenshan Dist., Taipei City, Taiwan
tunglin.sy@gmail.com
[4] Department of Information Management, National Taipei University of Nursing and Health Sciences, No.365, Mingde Rd., Peitou Dist., Taipei City 112303, Taiwan
cyhsu@ntunhs.edu.tw

Abstract. Despite the abundance of biomedical databases in Taiwan, there is currently no unified portal that effectively facilitates health research data searches to drive scientific discovery and promote a sustainable wellbeing environment. This study aims to design a user-centered gateway for health research data search, focusing on usability and ensuring that the platform supports the retrieval of fit-for-purpose datasets while maintaining data privacy, accessibility, and transparency.

A user-centered design approach was employed, involving personal interviews with domain experts. An initial set of questions, derived from literature reviews and expert consultations, explored various dimensions of health data usability. The interview results identified key criteria for assessing the effectiveness of health research data searches in supporting sustainable health outcomes.

Seven critical factors were identified for quick confirmation of search requirements: follow-up, publisher, purpose, source, time lag, data custodian, and specific requirements. The interviews also highlighted a lack of familiarity with dataset retrieval tools, emphasizing the need for cultivating user knowledge and habits to promote wider adoption and effective use of the gateway.

As dataset retrieval needs in Taiwan remain a relatively new area, understanding the characteristics of datasets and tailoring search patterns to meet user requirements are essential. This framework provides a foundation for improving health data accessibility. Future research should explore advanced methodologies for addressing

⋆ Corresponding author

diverse user needs, including intelligent recommendation systems to support a sustainable wellbeing environment.

Keywords: Metadata, Big Data, Real-world Data, User-Centered Design

1. Introduction

Digital health, as outlined in the World Health Organization's strategic plan, has the potential to revolutionize global healthcare [2]. The Covid-19 pandemic underscored the critical role of data and artificial intelligence in devising effective strategies to combat the virus. These technologies have significantly contributed to disease trend modeling, precise diagnosis, symptom categorization, result interpretation, vaccine development, therapy advancements, drug innovations, and forecasting medical demand hotspots [5].

As the volume and diversity of data generated and presented in various formats continue to grow, retrieving essential information efficiently has become increasingly challenging. Many large-scale research databases are constantly evolving, yet their content, usage guidelines, and application scopes are fragmented across platforms, hindering quick and accurate information retrieval for users. Addressing this challenge requires the development of a health data gateway adhering to FAIR principles (findable, accessible, interoperable, reusable) [16].

Several countries have taken proactive measures by establishing integrated platforms to facilitate data sharing, thereby enhancing accessibility and fostering collaboration. However, establishing a consistent data presentation framework remains a significant challenge due to diverse user backgrounds and needs. Notable examples such as BBMRI-ERIC in Europe, housing 100 million samples and delineating quality standards for European biobanks, and UK Health Data Research (UK HDR), a collaborative initiative enabling access and utilization of health-related data for research purposes, showcase distinct architectural designs and insights relevant to our proposed platform [3,13].

Taiwan with many large scale biomedical databases currently lacks a comparable integrated platform, motivating our research endeavor. This study seeks to contribute to the scientific community by addressing these challenges and conceptualizing a comprehensive platform that facilitates seamless data integration, promotes collaborative research, and nurtures a more accessible and impactful data ecosystem. Central to this approach is a user-centered design, offering filtering conditions for efficient data retrieval and assessing disparities between datasets. By constructing appropriate data gateways, data can be effectively utilized and an environment for sustainable development of medical technology can be created.

The primary objective of this study is to identify key factors in dataset screening frameworks and data availability criteria. A user-centered gateway plays a pivotal role in determining critical selection factors for data users, shaping data screening pathways, and defining dataset metadata essential for effective data navigation. Our study aims to achieve three main goals: (1) provide metadata for dataset definitions applicable to health data analysis research; (2) propose measures and tools for inclusion in future portals and metadata for research retrieval; and (3) identify areas requiring further research attention. Our researchers have balanced diverse stakeholder needs to design a prototype database search portal (dataset portal). While many national biomedical databases exist in Taiwan, there is a notable gap in developing user-centric platforms that provide intuitive access

and metadata-driven dataset screening mechanisms. This study directly addresses this gap through a structured user-centered design process, aligning with the special issue's focus on computational technologies for sustainable wellbeing environments.

Scientific Contributions of this Study are:

- 1. We introduce a structured metadata filtering framework derived from empirical insights of domain experts.
- 2. We operationalize user-centered design in the context of health dataset discovery, identifying seven metadata-driven screening criteria.
- 3. We develop a prototype interface ("Easy Search") informed by user needs, which bridges qualitative understanding with quantifiable utility indicators.
- 4. We enrich qualitative findings through operational design artifacts and metadata codification, laying groundwork for intelligent data gateway development in sustainable health environments.

2. Materials and Methods

This study employed an in-depth interview-based consensus approach to define critical criteria for dataset filtering. These criteria provide essential information about different dataset metadata and establish operational guidelines for future dataset selection. The researchers reviewed relevant literature to inform this study. Additionally, this research convened a panel of experts comprising individuals from diverse fields, including health information services, medical material testing research and development, drug development, auxiliary medical services, and academic research. These experts collectively discussed the current state of clinical dataset search.

The research focused on gathering insights from this diverse group to achieve its stated objectives. An interview panel was convened specifically to discuss the current state of clinical dataset search. This panel consisted of stakeholders representing various perspectives related to health data, data integration, research, and platform development. The in-depth interviews focused on the following critical points:

- Identifying Key Factors: The interviews aimed to identify essential factors for dataset screening and data availability identification. This involved understanding the criteria that researchers and data users consider important when selecting datasets.
- Designing an Effective Gateway: The panel's input helped in designing a gateway that aligns with identified selection factors, ensuring that the platform effectively meets the needs of data demanders.
- Developing Metadata: Collaboratively defining dataset metadata meaningful for health data analysis research, enabling users to better understand available datasets and their attributes.
- Identifying Research Needs: Recognizing gaps or areas requiring further research, such as understanding specific dataset requirements or addressing challenges related to data availability.
- Balancing Stakeholder Needs: The panel's insights contributed to balancing the needs and expectations of different stakeholders, ensuring that the database search portal prototype addresses various perspectives effectively.

The interview panel comprised experts from diverse backgrounds who collaboratively addressed the study's goals and objectives, ultimately contributing to the development of a comprehensive and impactful health data integration platform.

2.1. Designing Interview Interactions:

(1.) Interview Process and Practical Operation The research process(Fig. 1 shows the Interview and Practical Operation Process.) begins with Interview Design and Interviewee Selection, where the primary focus is on crafting appropriate questions that align with the study's objectives. At this stage, careful consideration is given to selecting the right interviewees whose experiences and insights can provide valuable contributions to the research.

Once the design is finalized, the next step is Interview Preparation. This phase involves refining the interview questions and ensuring that all necessary tools and materials for data collection are prepared. In this study, the interview was conducted online, and prior to the interview, an interview outline along with virtual cards (via a website link) were provided to the interviewees. This allowed them to better understand the purpose of the interview and prepare accordingly.

Following the preparation, the research enters the In-depth Interview phase. This is a key component of the study, where the interviewer engages with the selected individuals to explore their thoughts, experiences, and perspectives in great detail. During the interview, the researchers strictly followed the outline, ensuring the conversation stayed focused. If the interviewee needed to provide additional input or demonstrate practical operations, control of the screen was passed to them. In the third part of the interview, interviewees were asked to operate and explain processes based on their experience. If they lacked recent practical experience, they were encouraged to share their thoughts and needs regarding the database retrieval process.

Finally, the process concludes with the Interactive Interview phase. Unlike traditional one-sided interviews, this stage emphasizes a two-way exchange between the interviewer and the interviewee. The interaction allows for a more dynamic conversation, where both parties contribute to the dialogue, leading to the discovery of deeper insights and a fuller understanding of the subject matter. This interactive format was particularly useful in exploring practical demonstrations and conceptual understanding, further enriching the research findings.

(2.) Interviewee Selection: The researchers of this study employed a purposeful sampling methodology to ensure representation and diversity in the participant group. Experts from various fields were invited to participate in in-depth interviews to assess requirements. The participants were carefully selected to encompass diversity across several working fields, including health information service, medical material testing research and development, drug development, auxiliary medical services, and academic research.

An expert is defined as an individual who meets the following criteria: an active participant in health data analysis and research who can provide insights into the types of data needed, the challenges faced in accessing and using data, and requirements for a user-friendly platform. Also, an individual with expertise in the healthcare and medical fields who can provide insights into the practical applications of health data, the relevance

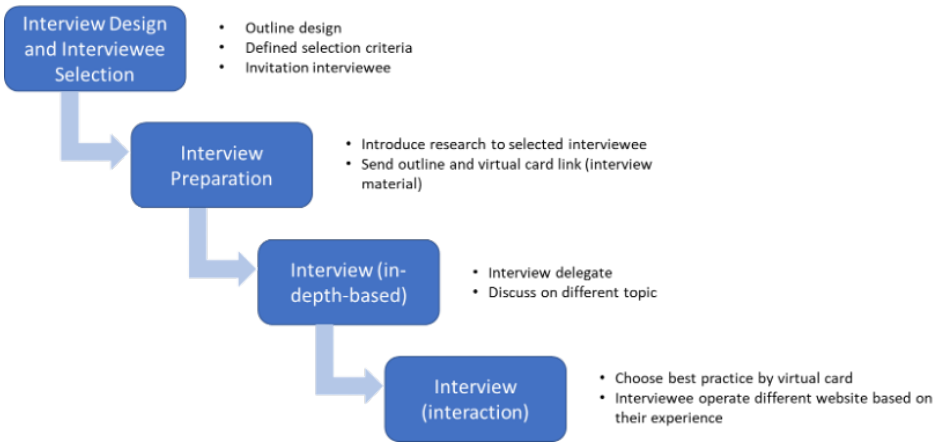

Fig. 1. Interview and Practical Operation Process

of data to medical practice, and the potential benefits of integrated platforms. Additionally, data analysts and data scientists, professionals with experience in data analysis, can provide perspective on technical aspects such as data integration, metadata creation, and the tools needed for effective data analysis.

A total of six participants were invited, and each of them was informed about the purpose of the study, with complete anonymity among experts. This study is primarily intended for the preliminary assessment of data gathering requirements and gateway design elements.

(3.) Key Factor (Factor Card Sorting) for Dataset Screening (Searching): This study summarized some important factors in designing virtual cards based on the initial survey results and referred to the screening field of UK HDR [8], and data explanations provided by UK Biobank [12]. During the interview, the interviewees referred to virtual cards, effectively highlighting the key factors and extending their explanation. The Fig. 2 shows virtual card we used during the interview.

1. Publisher
2. Phenotype
3. Coverage – spatial and follow-up
4. Provenance – purpose, source, collection situation, and time lag
5. Access – delivery lead time, jurisdiction
6. Format and standards – vocabulary, conforms, and language

(4.) Recording of interviews and the coding of questions

2.2. Interview Outline Design

This interview is divided into four sections: the interviewee's background and work experience, database search experience, gateway's functional requirements and expectations,

Fig. 2. Key factor virtual card during the interview

and the interviewee's perspective of the main values and principles of gateway. Due to the varying needs of experts (interviewees) in different fields, some questions were further asked after the survey answers were obtained. The interview guideline is as follows:

Background Survey – 10 mins Respondents introduced the interviewees' work background and current job content.

Dataset Screening Criteria – 40 mins The interviewees were asked about their experience with search databases. In addition to understanding data requirements, it is crucial to grasp the business logic and workflow to the fullest extent possible.

Additional point: If there are concerns related to censorship, inquire about the context of these needs, such as compliance with GDPR norms.

(a) Please share an example of a high-quality or particularly useful dataset, and explain its typical use.

(b) What aspects of data formats or standard models (e.g., the OMOP or the requirements for providing data via FHIR API) are critical to your work?

(c) The "number of items included in the database" is generally indicated for the coverage (rate) of the dataset. What are your needs (e.g., number of observations (value), observation points, etc.) in terms of quickly understanding the coverage (rate)?

(d) Lastly, what features of the dataset are you looking forward to?

Operational Requirements and Expectations of Database Search – 30 mins This section focused on asking the interviewees to share their experience in database searching and, if possible, open the website they use (any website or tool) and show the re-searchers how to use it. Additional point: If the interviewees use more than two portals (websites), please ask them to share the differences, advantages, and disadvantages of their user experience.

(a) Which database search portals (websites) have you used in the last three months?

(b) Please briefly introduce these portals (websites) and explain their significance in your work.

- How can you tell if a database site is useful for your work?
- What information can be used to determine whether a database site is useful? (e.g., other users' opinions, comments, and five-star reviews on the database)
- Have you ever used any portals (websites) that were difficult to understand, counter-intuitive, or otherwise required technical assistance?

(c) Please show an ideal database search portal (website) and explain why it is ideal. If none, what is lacking from the current portal (website) you use?

- What is your favorite feature or function of this site?
- How much time did you spend learning how to use this portal (website)? What learning strategy did you use?
- What was the most confusing or difficult aspect of using this portal (website)?
- Is there anything in the database's search portal (website) architecture that you of-ten find questionable or requires explanation (it cannot be self-evident)?

(d) As a user, what service functions (e.g., structure, metrics, scoring, etc.) do you think the database search portal (website) should include helping you make better use of the dataset?

- Please describe under what circumstances these features positively impact the utilization of datasets.

Gateway Values and Design Principles – 30 mins This section is designed to determine the value and design principles involved in developing a portal for database search.

(a) What principles are critical when developing and designing a database search portal to make it easier to find high-quality data? Additional point: This is an open exploration and is not limited to the content and reference factors of the interviewees' direct answers.

- Which filters are the most/least important for you to use when finding datasets?
- Additional point: This is a semi-structured exploration. If the answer aspect is related to the card, you can follow the opponent's context and ask additional questions. If not, provide a factor card that leads to the category discussion.
- (Provide the factor card.) Based on experience with the database search portal architecture, which factors are particularly useful/useless concerning the card? Why?
- Please try to find a dataset that you find useful on HDR UK.
- How would you rate the importance of the key factors you mentioned in understanding data utility based on your own data needs and experience?
- Are there any other filters you find useful?

(b) Please refer to the table for information about data utility. Which ones do you think are important or helpful to you? Please choose three to five options.

3. Results

Complex topics are deeply explored and analyzed through a series of in-depth one-on-one interviews with experts, allowing a detailed examination of various aspects relevant to the research objectives. Insights gained from these interviews provide valuable perspectives from experts in various fields, contributing to a comprehensive understanding of the subject. Interviews help gather rich qualitative data, allowing the identification of nuanced patterns, perspectives, and underlying themes. A total of 7 result categories were summarized based on the interview, including:

1. How Taiwanese users describe high-quality datasets
2. Experience sharing
3. User needs
4. Website design strategy
5. Metadata filter preference insights
6. Platform prototype development
7. Current data application process and difficulties

The research output incorporates multiple experts, ensuring a holistic view that encompasses different perspectives, disciplines and areas of expertise, thereby enhancing the robustness of the conclusions drawn. The in-depth interview and interactive process allows for detailed exploration of complex concepts, leading to a deeper understanding of complex interrelationships and factors.

1. Preliminary Screening: Initially, we conducted a preliminary screening of experts from various fields to ensure they possess relevant professional knowledge and experience. This may include reviewing their research background, work history, and relevant professional certifications.
2. Criteria Setting: Next, we set criteria for participants to ensure they meet specific conditions required for the study, such as actively engaging in clinical data analysis and research, having experience in gathering usable databases, familiarity with or usage of Taiwan databases, and working in the field of healthcare information and clinical research.
3. Invitation Selection: Based on the preliminary screening and set criteria, we invited experts from different fields to ensure diversity and representation among the interviewees.
4. Confirmation of Participation: We confirmed the willingness and availability of the invited participants to ensure they have sufficient time and resources to participate in the interview process.
5. Interview Conduct: We conducted in-depth interviews to gather insights and opinions from the interviewees to achieve the research objectives and goals.

Through the above selection process, we successfully invited a total of six interviewees who have diverse expertise and experience in different fields. These interviewees have experience in clinical data analysis, gathering usable databases, familiarity with or usage of Taiwan databases, and work in the field of healthcare information and clinical research, providing a range of perspectives and in-depth insights.

In-depth interviews yielded substantive and multifaceted insights. These insights provide a rich qualitative data set that comprehensively explores all dimensions of research

objectives, facilitating nuanced and informed analysis of topics. The background, experience and field of work of the experts are described in the Table. 1. A total of six experienced interviewees from different fields e participated in the in-depth interviews.

Table 1. Background statement of interviewees

No.	Work Field	Research Field	Affiliation	Data analysis experience (years)
1	Drug Developers	Genomics	Researcher	5-10
2	Health Information Service	Epidemic	Researcher	5-10
3	Academic Research	Pharmaceutical Management	Professor	More than 10
4	Medical Material Testing Research and Development	Biomarker	Senior Executive	More than 10
5	Medical Auxiliary Services	Clinical Trials	Researcher	5-10
6	Medical Auxiliary Services	Clinical Trials	Business Manager	More than 10

3.1. How Taiwanese users describe high-quality datasets

(1.) The data has high integrity Respondents expect the data in the target database to be complete and continuous. The target data is considered complete if it contains all the required data fields (such as the clinical data), and it is continuous if the target data has been valid for a period of time. Taiwanese researchers asserted that completeness and continuity ensured quality research. Respondent feedback:

"Whether the data is complete enough will also be affected by continuity. Continuity refers to whether the same question is asked every year in succession. Taiwan's National Health Interview Survey (NHIS) lacks continuity. Some questions were asked in the previous year and will not be asked the following year, causing an interruption. Meanwhile, NHIS in the United States is very continuous. If the continuity is not good, there will be no way to see the difference for several years." - 3

"One of the key factors in determining whether the database is easy to use is completeness. We check to see if the data has been collected at different times. In past experiences, there is a unit that provides a database of about 10,000 patients, but less than half of the people have complete information (such as kidney function and various clinical tests), so there is a gap in completeness." - 4

(2.) Conducive to multi-party cooperation and value-added application The definition of data fields is unified, and there are rules to follow, such as a unified format or filling method to facilitate data cleaning and effective serial file analysis, which is also beneficial for users in terms of collaborating with several stakeholders simultaneously. Respondent feedback:

"I would expect the data connection between different database systems to be easy, and be discussed not only with IT professionals or statisticians, but with clinicians who can participate in the discussion and explore the database. For example, I researched pre-end-stage renal disease (pre-ESRD). We obtained patient-related diagnoses, medication, treatment, examination, medical treatment history, and other relevant information from the case management data sheet (registry data). During analysis, we needed to combine these databases. For example, when analyzing medication, we had to do further analysis from the sorted drug file, which is integrated into many different forms." - 2

"We pay great attention to whether it can be compared to data from different cohorts. Once the data is in our hands, we need to perform data cleaning. I often encounter a

situation wherein the same target is measured, but the same field is stated to be different. Uniformity, maybe capitalization, whether there are spaces or brackets, and so on, or the units are not uniform. It would be best if this part could be unified." - 4

(3.) Good accessibility Users expect the barriers to accessing data to be as low as possible and in line with their research schedules. Good data accessibility involves a friendly application process, short waiting time, and expected frequency of data collection. Users mentioned that hospital-side databases, or databases organized by academic societies, have good accessibility and are more beneficial in achieving research output. Respondent feedback:

"Currently, the health insurance database requires entering the value-added center to access the data. If there is an issue following the analysis, it will be discovered a month later. If another issue is discovered later, there may be a constant need to present it in a monthly update." - 6

(4.) The data is mature enough This characteristic refers to whether the data collection period is sufficient to verify the research hypothesis, meet the research needs of the Taiwan region, and provide users with confidence in their ability to produce a certain level of analysis results. Simultaneously, the data covering a long-time span can avoid the inconvenience caused by time delay, suggesting there is no need to update and analyze the data repeatedly. In other words, the less affected by time, the more mature the material is. Respondent feedback:

"Maturity is a time-dependent outcome. Can I analyze, at least, something like median survival before my study is closed?" - 2

"At present, all hospitals assume that their data are all from patients who will go to them for a long time. If there is no long-term data, there will be no way to know whether these people are representative. Another possibility is that multiple rounds of analysis will be required." - 6

3.2. Experience-sharing

(1.) NHIS – easily accessible, with complete information Respondent feedback:

"The data collected by NHIS are sorted out every year and can be used directly after downloading. There are guidelines on how to obtain and merge the data. However, there are guidelines on SAS coding and file conversion. Therefore, data security and accuracy are very high. In NHIS, a person is an identification code, so data processing is fairly simple. Instead of merging additional files, one can just be pulled. Files are also very easy to obtain without any hindrance." - 5

(2.) Government open information platform - can be browsed quickly, grasps the information overview before applying The Taiwan government's open data platform (https://data.gov.tw/) contains several interpretations of data sets. Some even provide online viewing of demonstration data, which can obtain much information before data selection to ensure that the obtained data is suitable for analysis. Respondent feedback:

”On the open data platform, just press the sample data button to immediately see what the data looks like. The number of people and the amount in the sample data are good points to consider. Whether it is from a research or business perspective, knowing how much is enough can help professionals quickly determine whether to use it.” - 5

3.3. User needs

(1.) Interface design - Gradually develops the habit of using search engines Users in relevant fields in Taiwan have not yet used search engines to find data-bases. They usually identify and recognize usable databases based on referrals and by reading papers and materials.

”I know what topic I am looking for. Using too many keywords can be intrusive. But, if you want to explore how the database can be applied, keywords or filters are a good choice.” - 6

(2.) Information Design - Search results should display cross-domain key information Users expect that the search data set can briefly describe the data contained in the database, such as a list of data fields, data volume, complete transaction numbers, or missing data rate (missing rate). On the database introduction page, users can disclose complete information. At this time, the information can cover the different needs of various fields as completely as possible. Respondent feedback:

”Most databases should not have complete information. However, some research fields may only require some variables depending on the research field. In contrast, some research may require all of them. What the website platform can provide is the approximate missing rate of each database or each field, making it easier for researchers to evaluate the database effectiveness.” - 5

3.4. Website Design Strategy

(1.) Interface interaction - Open for self-downloading, simplifies the application process The interface presents a simple and clear call-to-action design on the search result page, lowers the threshold for data acquisition, and increases the data utilization rate. For example, in-depth cooperation users plan a concise application process for heavyweight or special data sets after logging in, design services for users’ common contact points (such as web pages, phone calls, or emails), and provide users with a painless application process. Respondent feedback:

”When I found the data I wanted to use, I noticed that I couldn’t download it. At this time, I went to find out how to apply. This part of the US TCGA function is deeply hidden. I can’t quickly identify which data should be logged in and cannot be obtained.” - 1

(2.) Quickly browses and grasps the information overview before applying Users can inspect the metadata of the data set for the search results, as well as data samples, to understand the cross-section of the data, give other users concise information in evaluating whether it is a desired high-quality file, and provide the users with the database application rate. Respondent feedback:

”I would like to know the experimental design method and materials involved in the data collection. If you see that the analysis method of this data is consistent with the

analysis method of my previous genetic data, you can save the original unorganized data and directly download the sorted vcf." - 1

(3.) Clear instruction documents to enhance the freedom of data To improve data usability, an explanatory document should be provided on the introduction page of the search database (collection), which should include the definition of data fields and the data connection method to ensure that users can freely utilize the data after downloading.

3.5. Metadata Filter Preference Insights

For user-centered design, this study intended to let users experience the actual operation process and leverage the "operation process" in enhancing the survey's effect. This study included interactions at the end of the interview.

Referring to their virtual cards during the interview, the interviewees effectively pointed out key factors and extended their explanation. Data quality management is the most important item according to statistics on each interviewee's preference insight, and data quality can assist demanders in ensuring that the data they found can be used. The second most important item is interpreting data integrity and compliance, which allows demanders to effectively evaluate whether the data is aligned with their intended use and rules. Meanwhile, the last item points out to data dictionary and semantic library, which can improve data usability and accuracy.

Data utility allowed the interviewees to choose virtual cards during the interviews. Data utility is primarily applicable to data set interpretation. It guides users in determining which indicators help understand whether a data set has good data utility (e.g., facilitating discussions with colleagues, supervisors, or other stakeholders or facilitating research business advancement).

The virtual cards, where each card represented a specific aspect and included possible options. The Utility items included: Documentation Completeness, Availability of documentation and support, Data Model, Data Dictionary, Provenance, Data Quality Management Process, DAMA Quality Dimensions, Pathway coverage, Length of follow-up, Allowable uses, Time Lag, Timeliness, Linkages, and Data Enrichment. During the question and answer process, users selected indicators from the virtual cards that were relevant to their usage scenario, and these selections were then compiled to identify the most frequently mentioned Data Utility items.

The Fig. 3 shows the statistical results of data utility selection. Over half of the interviewees identified four important items: documentation completeness, data quality management process, DAMA quality dimensions, and allowable uses.

(1.)Documentation completeness Documentation completeness refers to the availability of comprehensive documentation for clinical research datasets. This includes detailed information about data sources, data collection methods, variables, data formats, and any transformations or pre-processing steps used. Complete documentation is crucial for researchers to under-stand the datasets, replicate analyses, and interpret the results accurately. It contributes to clinical research transparency and reproducibility.

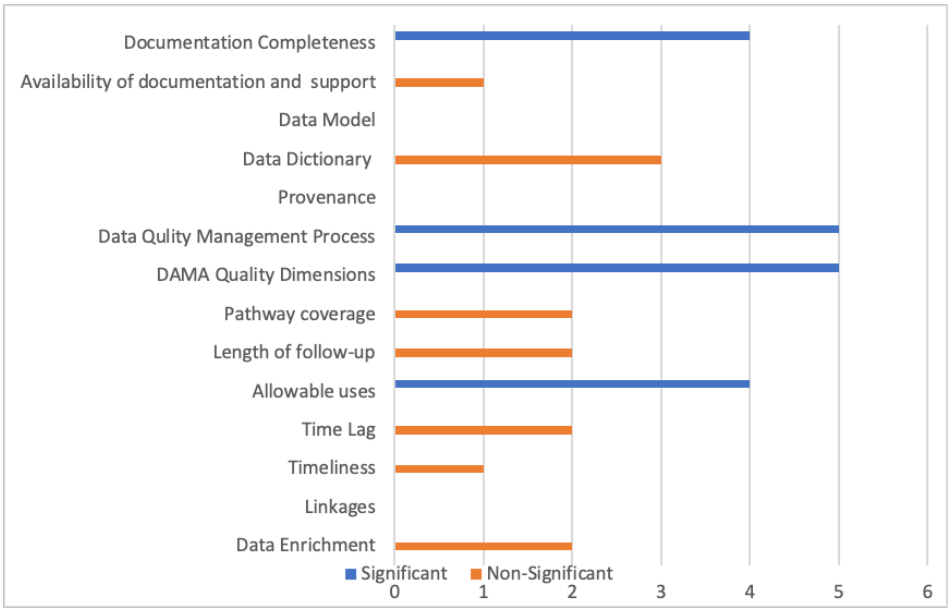

Fig. 3. Data utility selection results

(2.)Data quality management process Data quality management is critical in clinical research dataset searches. It entails the implementation of processes and procedures to monitor, assess, and improve data quality. This includes identifying and addressing data errors, inconsistencies, missing values, and outliers. A robust data quality management process assists researchers in ensuring the reliability and validity of the datasets they use, leading to more accurate and meaningful research outcomes.

(3.)Data quality dimensions Data quality dimensions provide a framework for evaluating and assessing data quality. These dimensions include several factors, such as accuracy, completeness, consistency, timeliness, uniqueness, and relevance. Evaluating these dimensions is crucial in clinical research dataset search as it assists researchers in understanding the strengths and limitations of the datasets they are working with. Addressing these dimensions helps researchers ensure that data fits the intended research purpose.

(4.)Allowable uses Due to privacy regulations, data-sharing agreements, and ethical considerations, clinical research datasets may have specific restrictions on their allowable uses. Therefore, understanding and adhering to these allowable uses is critical in clinical research dataset searches. Researchers must be aware of any limitations or constraints on dataset use to ensure compliance with legal and ethical requirements. This ensures the responsible and ethical use of the data while protecting patient privacy and maintaining data security.

Documentation completeness, data quality management processes, data quality dimensions, and adherence to allowable uses are all critical aspects of clinical research dataset

searches. These four issues contribute to clinical research data reliability, transparency, and ethical use, resulting in more robust and meaningful research outcomes. Based on the interview results and statistical summary, seven explicit factors that metadata could explain were defined. These factors include publisher, purpose, source (documentation completeness), data custodian (data quality management process), follow-up, time lag (data quality dimensions), and requirements (allowable uses).

3.6. Platform Prototype Development

The metadata framework of the dataset was designed based on the interview results, and seven important indicators were selected as the key processes involved in user-friendly use. During the dataset retrieval process, the dataset that meets the users' needs can be located using the "Easy Search," which selects the seven key factors. The prototype of Easy Search based on seven factors as shown in Fig. 4 4. The UI of Easy Search as shown in Fig. 5. The Data screening example as shown in Fig. 6.

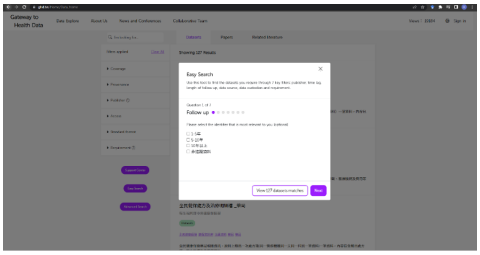

Fig. 4. Prototype of Easy Search based on seven factors

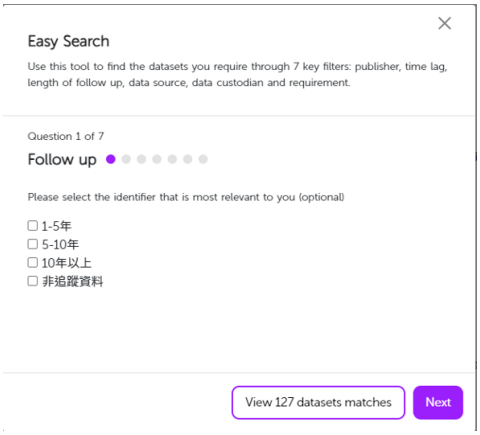

Fig. 5. UI of Easy Search

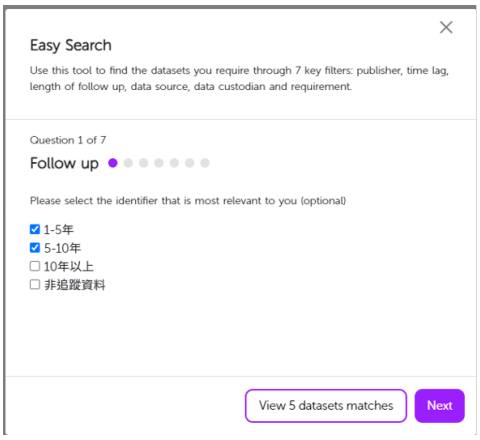

Fig. 6. Data screening example by Easy Search

For example, in the first indicator, ”Follow Up,” after selecting the options of 1-5 years and 5-10 years at the same time, the number of data sets is reduced from 127 to 5, indicating that screening such indicators can aid in the use of quick search.

3.7. Current Issues in the Data Analysis Process

In the data analysis research process, there are about 7 main steps, including ”finding data”, ”evaluating data use”, ”applying for data use”, ”obtaining”, ”data cleaning”, ”data analysis”, ”publishing research results or product development”, Most interviewees mention “applying” for the data use is currently the most difficult part of research process in Taiwan.

Calculated according to the burden points answered by the interviewees, the size of the red dot is shown in the Fig. 7. A larger red dot means more effort is required to perform the step. It is difficult to find the correct data, and the application after finding it is even more troublesome. It needs to go through many processes of review and waiting, and some data sets do not clearly explain how to apply.

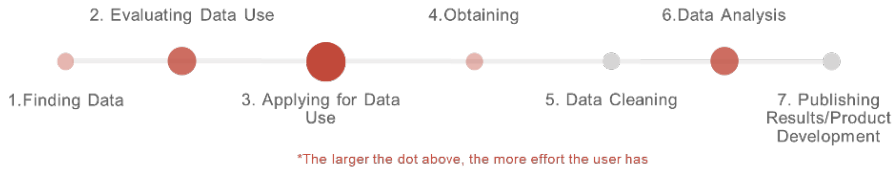

Fig. 7. Data analysis process and effort evaluate

Therefore, the design of the gateway should incorporate data and application-related information, and it is best to provide guidance and provide appropriate key reminders based on application specifications to effectively help demanders apply for information.

4. Discussion

This study performed in-depth interviews to identify some key points for assessing the availability and ease of use of public health data. The data custodian must improve public health data accessibility by ensuring it is readily available to relevant stakeholders [4]. Improving data accessibility involves addressing any barriers or restrictions that may hinder access, such as complex data formats, limited data sharing agreements, or outdated data systems.

The research results point out that High-Quality Datasets have several characteristics: 1) Data possesses high integrity and continuity, ensuring research quality; 2) Definition of data fields is standardized, with rules to facilitate collaboration and value-added applications; 3) Good data accessibility involves a user-friendly application process, minimal waiting time, and expected data collection frequency; 4) Data maturity entails a sufficient collection period to verify research hypotheses and meet regional research needs. Data quality assurance is a critical aspect of data content [9]. The gateway should establish mechanisms to assess the accuracy, reliability, and completeness of the public health data. This process involves implementing data validation processes, conducting regular audits, and promoting adherence to standardized data collection and reporting protocols. Standardizing data formats, coding systems, and terminologies across different sources is also essential for effective data integration and interoperability [7]. To facilitate data sharing and analysis, the gateway should support adopting common standards and ensure compatibility among disparate data sets.

For data management, The PIONEER Hub in UK [3] which is funded by UK HDR has a good practice reference. The PIONEER Hub is a data set and hub that includes primary, secondary, social care, and ambulance data, collects and curates acute care data from across the health economy. There are prudent, complete and clear requirements for data integration, management and release, including the use of international standard launched by International Organization for Standardization (ISO), including metadata catalog (ISO 11179)[11], data quality (ISO 8000)[10] and quality assurance (ISO 25012)[9]. Public health data often contains sensitive information, so it is crucial to prioritize privacy and security considerations[6]. There is a case that NIH TCGA (The Cancer Genome Atlas) is a large interdisciplinary initiative funded by the National Institutes of Health (NIH) in the United States, providing research scholars with access to de-identified data on specific research topics[15]. Its primary goal is to study the genomics alterations in various types of cancer, aiming to gain a deeper understanding of the molecular mechanisms of cancer and personalized treatment approaches. By analyzing genomics data from tumor samples, it reveals mutations, gene amplifications, gene deletions, and other variations present in different types of cancer. Allow researchers to obtain information honorably. TCGA secure data storage to protect individuals' privacy while enabling data access for authorized users. A well-defined governance framework is necessary to collect, store, share, and use public health data. This framework involves establishing clear roles and res possibilities, defining data ownership and stewardship, and adhering to ethical principles to ensure transparency, accountability, and responsible data management practices. At the same time, this type of data management architecture must be equipped with robust protocols for data anonymization, consent management, and encryption.

In terms of website design, the gateway to health data must have user-friendly interfaces to allow users to easily discover, access, and analyze the data[1]. This includes

designing intuitive search functionalities, providing data visualization tools, and offering user support and documentation for enhancing data usability and user experience. This is consistent with the research results. Research interviews pointed out that for a data search gateway, the interface design enables users to gradually adapt to using search engines to find datasets. Information design needs to display key information of across domains for search datasets. Website design strategy pointed out that users should be provided with information assistance for data application, which can effectively improve the difficulty of applying for data in the past, and help users understand data before applying through proper description of data and data content. Finally, the metadata of the data content of a dataset must have clear instruction documents enhance data freedom.

Taiwan's NHIS data contains complete information but only allows analysis in a controlled environment. Another example is the Taiwan government's open data platform allows for quick data browsing to grasp data overviews. The data custodian must emphasize the importance of comprehensive data documentation and metadata[14]. This process includes capturing relevant information about data sources, collection methods, variables, and any associated limitations or biases. Clear documentation helps users understand the context and quality of the data they are working with. Collaboration among various stakeholders, such as government agencies, research institutions, and the public, is vital for successful data utilization. The framework should encourage partnerships to promote data sharing, interdisciplinary research, and development of innovative solutions to public health challenges. The gateway should be designed in a way that it evolves over time and adapts to emerging technologies, changing data needs, and evolving best practices. Regular evaluations, feedback mechanisms, and a culture of continuous improvement are required to keep the gateway to health data remain relevant, effective, and updated. Promoting data literacy and providing user training opportunities is also crucial to maximizing public health data utilization. The gateway should support educational initiatives, capacity-building programs, and knowledge-sharing to equip users with the necessary skills to navigate, analyze, and interpret health data effectively.

5. Conclusion

Comprehensive interviews yielded crucial insights for designing an effective data gateway catering to researchers' needs. This study, involving six experienced interviewees from diverse fields, identified seven distinct categories of preferences and requirements among Taiwanese users regarding high-quality datasets.

Data integrity emerged as paramount, with an emphasis on completeness and continuity. Users stressed the importance of data being both complete, containing all necessary fields, and continuous over time. To address this, a robust data gateway enforcing stringent data standards is needed.

Collaboration and value-added applications were highlighted. Unified data field definitions and standardized formats were deemed essential for seamless data cleaning and analysis. Streamlining accessibility was another key consideration, emphasizing the need for a user-friendly application process with minimal waiting times. Data maturity, verified through extended data collection periods, was advocated to eliminate time-related inconveniences.

Experience-sharing underscored the importance of accessibility and comprehensibility, particularly through platforms like the Taiwan National Health Insurance System (NHIS) and the government's open data platform. User needs called for intuitive interfaces, keyword-based searches, and comprehensive dataset descriptions.

Metadata filter preferences, including data quality management and adherence to allowable uses, were identified as pivotal elements in dataset evaluation. The platform's prototype development focused on an "Easy Search" feature, streamlining dataset retrieval. Addressing challenges in data analysis, especially in the application process, was a key concern.

In summary, an effective data gateway aligned with these findings should prioritize data integrity, multi-party collaboration, streamlined accessibility, and dataset maturity. Additionally, it should cater to user needs through intuitive interfaces, comprehensive dataset descriptions, and efficient search mechanisms. Incorporating metadata filter preferences and addressing data analysis challenges will enhance the gateway's utility, bridging the gap between user expectations and high-quality dataset utilization for robust research outcomes.

Acknowledgments. This work was supported by the National Health Research Institutes, Taiwan. [grant number: No. CA 112-GP-09]

References

1. Aripiyanto, S., Agustin, F.E.M., Syakuro, A., Masruroh, S.U., Khairani, D., Sukmana, H.T.: User interface and user experience design using lean ux method on zakat ummat website. In: 2022 10th International Conference on Cyber and IT Service Management (CITSM). pp. 1–8
2. Director-General: Data and innovation: draft global strategy on digital health. Report, World Health Organization (23/12/2019 2019), https://apps.who.int/gb/ebwha/pdf_files/EB146/B146_26-en.pdf
3. Gallier, S., Price, G., Pandya, H., McCarmack, G., James, C., Ruane, B., Forty, L., Crosby, B.L., Atkin, C., Evans, R.: Infrastructure and operating processes of pioneer, the hdr-uk data hub in acute care and the workings of the data trust committee: a protocol paper. BMJ health & care informatics 28(1) (2021)
4. Hripcsak, G., Bloomrosen, M., FlatelyBrennan, P., Chute, C.G., Cimino, J., Detmer, D.E., Edmunds, M., Embi, P.J., Goldstein, M.M., Hammond, W.E.: Health data use, stewardship, and governance: ongoing gaps and challenges: a report from amia's 2012 health policy meeting. Journal of the American Medical Informatics Association 21(2), 204–211 (2014)
5. Intelligence, N.M.: Finding a role for ai in the pandemic. Nat Mach Intell 2, 291 (2020)
6. May, R., Denecke, K.: Security, privacy, and healthcare-related conversational agents: a scoping review. Informatics for Health and Social Care 47(2), 194–210 (2022)
7. de Mello, B.H., Rigo, S.J., da Costa, C.A., da Rosa Righi, R., Donida, B., Bez, M.R., Schunke, L.C.: Semantic interoperability in health records standards: a systematic literature review. Health and Technology 12(2), 255–272 (2022)
8. Sebire, N.J., Cake, C., Morris, A.D.: Hdr uk supporting mobilising computable biomedical knowledge in the uk. BMJ Health & Care Informatics 27(2) (2020)
9. Standardization, I.O.f.: Iso/iec 25012:2008 software engineering — software product quality requirements and evaluation (square) — data quality model (2008), https://www.iso.org/standard/35736.html
10. Standardization, I.O.f.: Iso 8000-1:2022 data quality — part 1: Overview (2022), https://www.iso.org/standard/81745.html

11. Standardization, I.O.f.: Iso/iec 11179-1:2023 information technology — metadata registries (mdr) — part 1: Framework (2023), https://www.iso.org/standard/78914.html
12. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M.: Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 12(3), e1001779 (2015)
13. UK, H.: Health data research uk (2023), https://www.hdruk.ac.uk/
14. Vardaki, M., Papageorgiou, H., Pentaris, F.: A statistical metadata model for clinical trials' data management. Computer Methods and Programs in Biomedicine 95(2), 129–145 (2009), https://www.sciencedirect.com/science/article/pii/S0169260709000601
15. Wang, Z., Jensen, M.A., Zenklusen, J.C.: A practical guide to the cancer genome atlas (tcga). Statistical Genomics: Methods and Protocols pp. 111–141 (2016)
16. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E.: The fair guiding principles for scientific data management and stewardship. Scientific data 3(1), 1–9 (2016)

Hsiu-An Lee – Senior researcher specializing in biomedical informatics and health data interoperability. Led the conceptualization and drafting of the study, focusing on user-centered design for health data search platforms.

Tung Lin – Researcher with expertise in information systems and usability engineering. Co-led the conceptualization and methodology design and contributed to the original draft.

Hsin-I Chen – UX designer and researcher experienced in human-computer interaction. Supported methodology development and iterative design testing.

Yi-Hsin Yang – Professor and expert in medical informatics and health data governance. Supervised the project, contributed to validation, and provided critical review and editing of the manuscript.

Wei-Chen Liu – Software engineer specializing in data platform development and integration. Participated in validation and system usability testing.

Yen-Ju Shen – Data analyst focused on health data quality and usability metrics. Contributed to validation and iterative improvement cycles.

Wen-Chang Tseng – System architect experienced in building scalable health information systems. Supported validation and technical review.

Chien-Yeh Hsu – Medical informatics researcher with expertise in health IT standards. Assisted in manuscript review and editing.

Received: December 04, 2024; Accepted: August 24, 2025.