F-001

Knowledge-based Data Validation for Efficient Secondary Data Utilizations

Xinjie Zhang†, Mayuko Ozawa†, Mika Takata†

Abstract

The demand for secondary data is increasing due to its wide utilization for purposes beyond its original collection in various industries. Particularly in healthcare, electronic clinical data including electronic health records and health insurance claims, has significantly advanced secondary data utilization in healthcare, supporting large-scale medical research for comprehensive insights. However, the utilization of secondary data in healthcare presents challenges, particularly in data quality, which can compromise the accuracy and safety of healthcare outcomes. To address this data quality issue, we have developed a knowledge-based data validation method that improves the efficiency of assessing statistical representativeness of datasets by automating the comparison of significant variables. This approach was evaluated in scenarios predicting future health conditions from medical records, showing a significant reduction in the number of variables needed for validation, thus enhancing process efficiency by over 99.1%. Our findings highlight the potential of targeted data validation to improve health outcomes, underscoring a shift toward more precise, data-driven healthcare solutions that can be effectively scaled in public health and medical services.

1. Introduction

The increase of electronic clinical data, such as electronic health records (EHRs) and health insurance claims, has catalyzed the expansion of secondary data utilization in healthcare, notably enhancing public health research and the development of personalized medical services. Surge in secondary data utilization is supported by legislative initiatives that aim to harness these vast data reserves for generating new economic and social value. Despite the benefits, the use of large-scale secondary data is fraught with challenges concerning data quality that could compromise the safety and efficacy of healthcare outcomes. Skewed statistical representativeness of an analytical dataset can lead to inaccurate analysis insights. The accurate statistical representativeness of these analytical datasets is critical, necessitating robust validation processes to ensure their reliability and applicability in healthcare contexts.

To tackle this data quality challenge, we propose a knowledge-based data validation approach that significantly enhances the efficiency of statistical representativeness assessments by automating comparisons of significant variables. The effectiveness of this proposed approach is evaluated in three distinct cases, predicting the future health conditions of residents based on their medical records. The rest of this paper is structured as follows: Section 2 explains current secondary data utilization and its data quality issue in healthcare domain. Section 3 discusses previous research work on improving data quality of secondary data. Section 4 proposes the knowledge-based data validation approach that enhances efficiency of data validation on secondary data. Section 5 demonstrates the experiment and result of evaluating knowledge-based data validation in comparison with baseline approach, the naïve data validation. Section 6 concludes the effectiveness of the proposed knowledge-based data validation approach and future work.

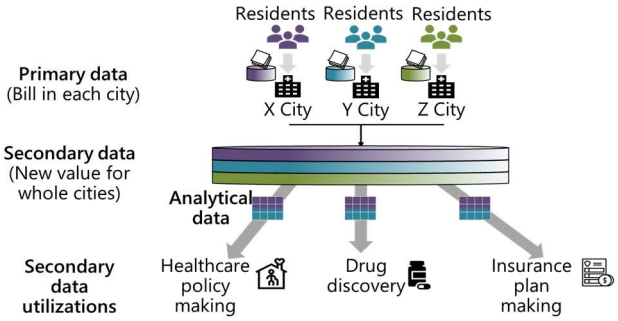

Figure 1. Secondary data utilization in healthcare

2. Current Secondary Data Utilizations

The expanding utilization of electronic clinical data such as electronic health records (EHRs) and electronic health insurance claims is driving the rapid growth of secondary data utilization in healthcare, fostering advances in public health research and personalized medical services. As illustrated in Figure 1, secondary data utilization involves utilizing data for purposes beyond its original collection [1][2], thereby facilitating large-scale research on public health and accelerating commercial activities such as novel medication development. Since 2018, Japan's Next Generation Medical Infrastructure Act [3] has promoted the secondary data utilization, including health insurance claims and EHRs collected from institutions, in order to create novel economic and social value. Thus, by harnessing the expanded availability of large-scale healthcare data from secondary sources, organizations including governments can conduct comprehensive public health studies and provide personalized healthcare services [5][6], such as personalized health risk prediction using statistical tools.

† 日立製作所　Hitachi Ltd.

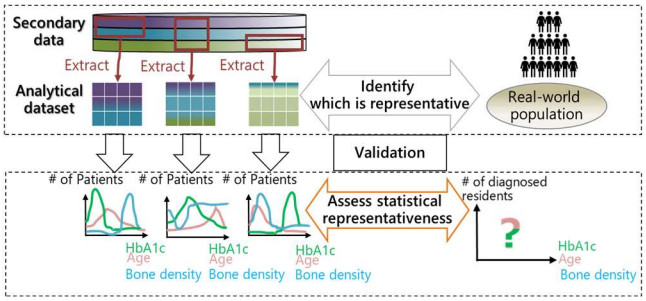

Figure 2. Skewed analytical datasets

Utilization of large-scale secondary data for statistical analysis faces challenges related to data quality, impacting its effectiveness in domains where precision is crucial for safety, such as healthcare. Figure 2 depicts the process of extracting and evaluating subsets extracted from secondary data, such as electronic medical records of patients diagnosed with diabetes, to ensure they accurately represent the broader real-world population. Analytical dataset refers to the extracted subset for analysis, while representative analytical dataset refers to the one that accurately reflects the general population's characteristics, such as demographics and health conditions. Statistical representativeness is used to assess how closely these analytical datasets align with actual population. When an analytical dataset is extracted for an analytical target, the skewness resulting from discrepancies in data extraction or processing can lead to biased or inaccurate representations of the real-world population. Utilizing secondary data to obtain accurate statistical analysis results in healthcare services requires high-standard data validation, particularly its accurate representation of the real-world population as the analytical target, for ensuring the applicability of healthcare insights derived from this data [4][11].

3. Related Work

Data validation on secondary data for accurate statistical representativeness is crucial for leveraging secondary data in statistical analysis in healthcare, requiring time-consuming and labor-intensive validation processes. Various approaches for evaluating data quality have been proposed [14][15][16]. However, these approaches require tailored adaptation because these approaches prioritize excluding intrinsic data errors over meeting target-specific needs [7][8][10]. Secondary data utilization in healthcare reuses clinical data that are not primarily intended for research or secondary analysis, of which unsystematically collecting process can have negative impacts on findings generated from the data [12]. Assessing the quality of descriptive clinical data elements is proposed [13], such patient race, by leveraging intra-institutional databases with varying data quality standards. Manual refinement of queries, particularly through Structured Query Language (SQL), is adopted to tailor extracted secondary data for specific analytical targets, thereby enhancing the accuracy of outcomes in both healthcare research and services. Botsis et al. [4] demonstrated that effective detection of issues, such as incorrect statistical representativeness in data, requires not only data manipulation skills but also medical expertise.

The quality of secondary data can be improved by having medical experts manually review the statistical representations on analytical target-related variables. For example, Kahn et al. [9] designed a data quality assessment framework that provides a structured approach to evaluate data, tailored to the specific requirements of the healthcare scenario in question. However, the reliance on knowledge-intensive operations poses a significant challenge to data validation, requiring medical experts assess the quality of secondary data by comparing its statistical representations on specific variables with statistics about real world population, which is a laborious and time-consuming process. Distinct from these approaches, we proposed a knowledge-based data validation approach that takes actual analytical targets into consideration and reduces both labor and time costs of data validation process.

4. Knowledge-based Data Validation

Knowledge-based data validation enhances the efficiency of assessing the statistical representativeness of secondary data by automating the comparison statistical representations on significant variables. The time required for validation involving naïve comparisons heavily depends on the number of variables, which can be substantial considering the number of check-up items in the case of EHRs. As shown in Figure 3, knowledge-based data validation further refines such naïve comparison by selectively comparing statistical representations on a few significant variables, using reliable text information about the analytical target population. Knowledge-based data validation approach can ensure the validation results while significantly reducing both time and labor costs through optimizing the mechanical comparison of statistical representations by reducing the number of variables to compare and selecting highly authoritative information.

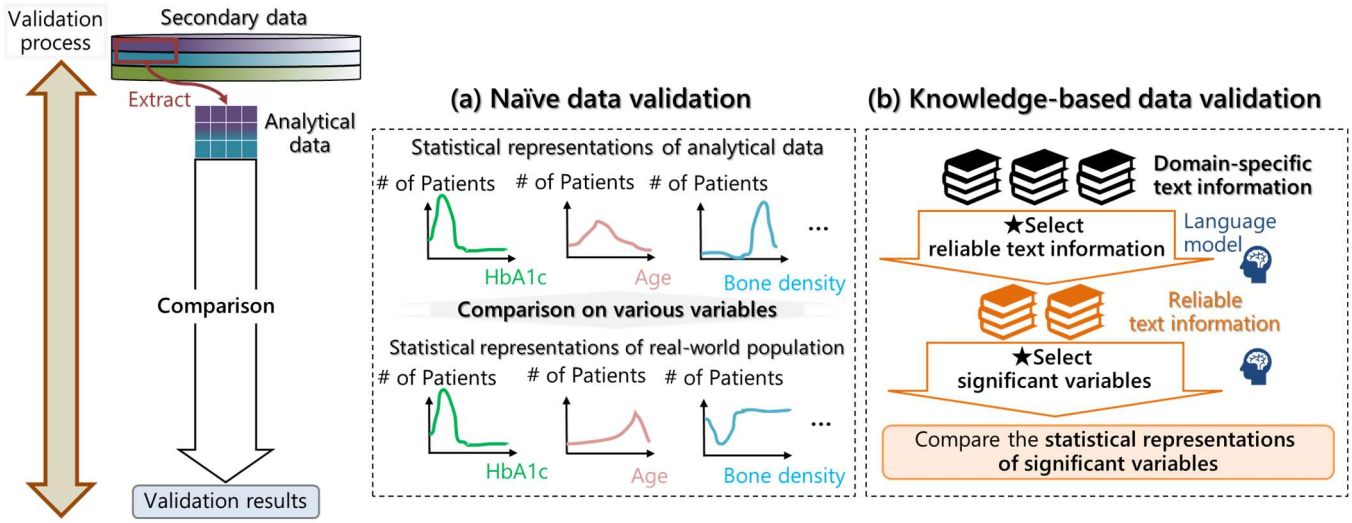

Figure 3. Comparison of (a) naïve data validation and (b) knowledge-based data validation

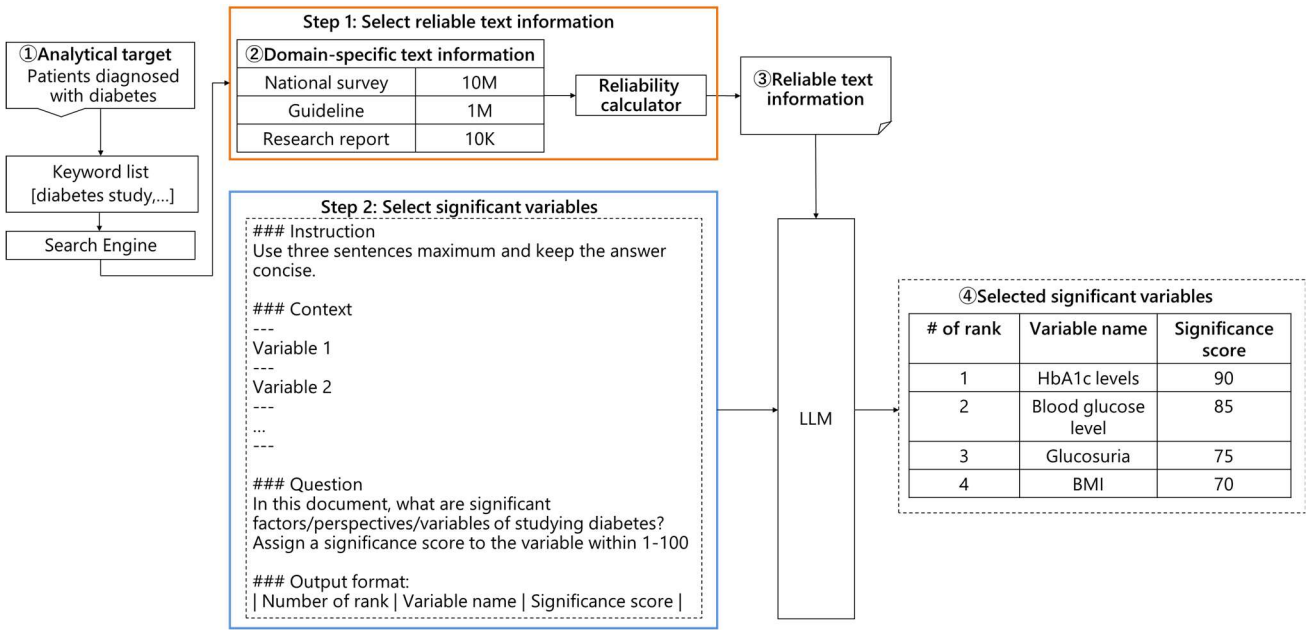

Figure 4. Implementation of knowledge-based data validation approach

4.1 Reliable text information selection

In our knowledge-based data validation approach, gathering authoritative domain-specific text information related to analytical targets is fundamental to ensuring the validity of secondary data assessment. For example, shown in Figure 4, the collection of text information begins from a wide search of domain-specific sources related to the analytical target through search engines like Google Scholar. A keyword list tailored to the analytical target is crafted to gather various text information sources, including government surveys and scientific reports, encompassing relevant aspects to the analysis, such as specific medical conditions or demographic factors. Such search retrieves extensive access to analytical target relevant text information and their citations, for the further selection on reliable information.

The reliability of text information in the proposed approach is quantified by a normalized citation count, ensuring that sources with higher citations are prioritized for their authority. The citation counts serve as a proxy for the academic impact and credibility of the text information source. A predefined numerical range, with a lower and upper limit, establishes the reliability score for a text information source. The reliability score is calculated by scaling the number of citations between the maximum (the most cited source) and the minimum citation counts (the least cited source) within this range. The span of citation counts for gathered text information is determined by the difference between the maximum and minimum citation counts. Based on this span, the relative position of each text information source's citation count is quantified by calculating its normalized citation count. The final reliability score of a text information

	①Analytical Target	②Domain-specific text information	④Significant variables
Diabetes	Predicting future diabetes risk for patients without prior diagnosis	Nation Survey	HbA1c
		Association guidelines	Blood glucose level
		Medical institute research	Glucosuria
		Medical paper	BMI
Hyper-tension	Predicting future hyper-tension risk for patients without prior diagnosis	Association research report	SBP
		Association research report	DBP
		Association research report	Gender Ratio
		Association research report	BMI
Hyper-dyslipidemia	Predicting future hyper-dyslipidemia risk for patients without prior diagnosis	Medical paper	LDL-C
		Medical paper	HDL-C
		Medical paper	TG

Figure 5. Selected variables for three analytical targets

source is calculated by adding the span of the normalized citation reliability range to the minimum reliability threshold. The score is used for quantifying authority of text information sources and prioritizing those with higher citations as more reliable. This reliable text information selection based on quantified reliability scores ensures that variables selected for comparison in the next step are derived from highly credible and relevant sources. It not only narrows the scope of data validation but also enhances the accuracy of validation results, thereby reducing the labor costs associated with manual evaluations.

4.2 Significant variable selection

Selecting variables that reveal significant statistical patterns of the analytical target population is the crucial step in enhancing data validation process, as the number of variables directly impacts the time for comparing statistical representations. With the selected reliable text information in the previous step, variables to compare statistical representations are further refined based on their semantic significance related to the analytical target.

To assess the significance of variables in a dataset for a specific analytical target, we leverage a large language model (LLM) to evaluate the semantic distance between the textual representations of these variables and the target. Leveraging a large language model (LLM) to evaluate the semantic distance between variables and analytical targets optimizes variable selection, substantially streamlining the data validation process. As illustrated in Figure 4, for a patient group with diabetes, the given variables (e.g., check-up items) can be ranged from direct check-up items such as HbA1c levels to less direct, yet relevant factors such as BMI or exercise frequency. The LLM processes the input text containing variables in the dataset and computes a relevance score by measuring the semantic distances of the given variables and the analytical target in the pre-trained embeddings. It then ranks these variables according to their relevance scores, clarifying which are significant to their analytical target. This prioritization of significant variables ensures that only most relevant are used to validate the dataset, thereby enhancing the

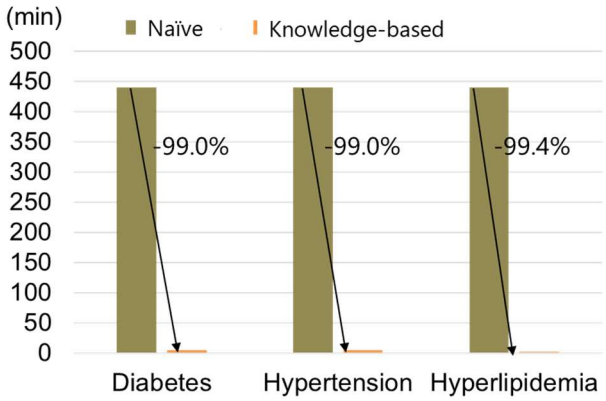

Figure 6. Comparison of validation time between naïve and knowledge-based data validation approach

validation efficiency by substantially reducing the number of variables used to compare statistical representations.

5. Experiment

This section demonstrates the efficiency of the proposed knowledge-based data validation approach is evaluated by reducing the data validation time using medical simulation data.

5.1 Experiment Setup

This experiment targets to validate the dataset for predicting three life-habit diseases. An analytical dataset targets the prediction of diagnosis risk for a disease within a specified time period is extracted, containing health check-up items, values, and health insurance claims of a group of residents.

The experiments include evaluation across three distinct analytical targets: (a) An analytical target is to predict whether a group of residents who have not received any prescription on diabetes medication in a certain period would be diagnosed with diabetes in a future period; (b) An analytical target is to predict whether a group of residents who have not received any prescription on hypertension medication in a certain period would be diagnosed with hypertension in a future period; (c) An analytical target is to predict whether a group of residents who have not received any prescription on hyper-dyslipidemia medication in a certain period would be diagnosed with hyper-dyslipidemia in a future period.

This paper evaluates the efficiency of the proposed knowledge-based data validation approach by measuring the time required for comparisons on statistical representations. We measure the query execution time using naïve data validation as baseline, involving checking the statistical representations on all check-up items— all attributes in the dataset. For each check-up item, we measure the query execution time needed to aggregate records, such as retrieving records within an age range or comparing gender ratios of a group with a diagnosis flag on a disease.

5.2 Result

This section demonstrates the efficiency of the proposed knowledge-based data validation approach is evaluated by reducing the data validation time using medical simulation data.

Figure 6 shows that the proposed knowledge-based data validation approach significantly reduces the numbers of variables for comparisons by 99.1% from 440 variables on average, where the selected variables are shown in Figure 5. In contrast to the baseline experiment, where a naïve approach involving an exhaustive statistical comparison across all possible attributes without selection, the proposed knowledge-based data validation significantly reduces the time required for comparing statistical representations, as depicted in Figure 6. The knowledge-based data validation approach streamlines data validation by selectively focusing on key variables, thus significantly reducing time needed compared to exhaustive naïve approach and enhancing efficiency in a secondary data utilization use case in healthcare.

6. Conclusion

The increasing demand for secondary data across various industries, especially in healthcare, underscores its critical role in advancing large-scale medical research and personalized medical services. Utilization of electronic health records and health insurance claims has greatly enhanced secondary data utilization, but challenges related to data quality continue to pose risks to healthcare outcomes. To mitigate these challenges, we have proposed a knowledge-based data validation method that streamlines the assessment of data statistical representativeness by automating the comparison of key variables. Our evaluation demonstrated that the efficiency of data validation was enhanced through reducing the number of variables required for comparisons by over 99.1%, streamlining the process significantly compared to naïve data validation approaches in scenarios predicting future health conditions from medical records. These results underscore the effectiveness of targeted data validation in improving health outcomes, paving the way for more precise, data-driven healthcare solutions that are scalable across public health and medical services. Further research should explore the effectiveness of the proposed approach on various diseases, such as rare diseases with fewer existing studies.

References

[1] Meystre, S.M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., & Lehmann, C.U. (2017). Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. *Yearbook of Medical Informatics, 26*, 38 - 52.

[2] Rizer, M.K. (2011). Health Information Technology for Economic and Clinical Health Act. *NASN School Nurse, 26*, 48 - 48.

[3] 改正次世代医療基盤法について https://www.mhlw.go.jp/content/10808000/001166476.pdf

[4] Botsis, T., Hartvigsen, G., Chen, F., & Weng, C. (2010). Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translational Bioinformatics, 2010*, 1 - 5.

[5] Cowie, M.R., Blomster, J., Curtis, L.H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M., Michel, A., Ong, S., Pell, J.P., Southworth, M.R., Stough, W.G., Thoenes, M., Zannad, F., & Zalewski, A. (2016). Electronic health records to facilitate clinical research. *Clinical Research in Cardiology, 106*, 1 - 9.

[6] Penrod, N.M., Okeh, C., Edwards, D.V., Barnhart, K.T., Senapati, S., & Verma, S.S. (2023). Leveraging electronic health record data for endometriosis research. *Frontiers in Digital Health, 5*.

[7] Diaz-Garelli, J., Bernstam, E.V., Lee, M., Hwang, K.O., Rahbar, M.H., & Johnson, T.R. (2019). DataGauge: A Practical Process for Systematically Designing and Implementing Quality Assessments of Repurposed Clinical Data. *eGEMs, 7*.

[8] Bastarache, L.A., Brown, J.S., Cimino, J.J., Dorr, D.A., Embí, P.J., Payne, P.R., Wilcox, A.B., & Weiner, M.G. (2021). Developing real-world evidence from real-world data: Transforming raw data into analytical datasets. *Learning Health Systems, 6*.

[9] Kahn, M.G., Callahan, T.J., Barnard, J.G., Bauck, A., Brown, J., Davidson, B.N., Estiri, H., Goerg, C., Holve, E., Johnson, S.G., Liaw, S., Hamilton-Lopez, M., Meeker, D., Ong, T.C., Ryan, P.B., Shang, N., Weiskopf, N.G., Weng, C., Zozus, M.N., & Schilling, L.M. (2016). A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs, 4*.

[10] Cohen, J.A., Trojano, M., Mowry, E.M., Uitdehaag, B.M., Reingold, S.C., & Marrie, R.A. (2019). Leveraging real-world data to investigate multiple sclerosis disease behavior, prognosis, and treatment. *Multiple Sclerosis (Houndmills, Basingstoke, England), 26*, 23 - 37.

[11] Rudolph, J.E., Zhong, Y., Duggal, P., Mehta, S.H., & Lau, B. (2023). Defining representativeness of study samples in medical and population health research. *BMJ Medicine, 2*.

[12] Brennan, Patricia Flatley and William W. Stead. “Assessing Data Quality: From Concordance, through Correctness and Completeness, to Valid Manipulatable Representations.” *Journal of the American Medical Informatics Association : JAMIA* 7 1 (2000): 106-7 .

[13] Kahn, M.G., Eliason, B.B., & Bathurst, J.E. (2010). Quantifying clinical data quality using relative gold standards. *AMIA ... Annual Symposium proceedings. AMIA Symposium, 2010*, 356-60 .

[14] Diaz-Garelli, J., Bernstam, E.V., Lee, M., Hwang, K.O., Rahbar, M.H., & Johnson, T.R. (2019). DataGauge: A Practical Process for Systematically Designing and Implementing Quality Assessments of Repurposed Clinical Data. *eGEMs, 7*.

[15] Weiskopf, Nicole Gray and Chunhua Weng. “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.” *Journal of the American Medical Informatics Association: JAMIA* 20 (2013): 144 - 151.

[16] Cai, Li and Yangyong Zhu. “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.” *Data Sci. J.* 14 (2015): 2.

[17] Ramanathan, T.R., Schmit, C.D., Menon, A.N., & Fox, C. (2015). The Role of Law in Supporting Secondary Uses of Electronic Health Information. *The Journal of Law, Medicine & Ethics, 43*, 48 - 51.