MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220020

Coverage of Clinical Research Data Retrieved from Standardized Structured Medical Information eXchange Storage

Masaharu Nakayama[a,b], Feng Hui[a], and Ryusuke Inoue[b]

[a] *Department of Medical Informatics, Tohoku University Graduate School of Medicine, Miyagi, JAPAN*
[b] *Medical IT center, Tohoku University Hospital, Miyagi, JAPAN*

Abstract

Clinical researchers hold high expectations for the utility of health data sourced from hospital information systems. In Japan, the standardized structured medical information eXchange version 2 (SS-MIX2) storage is a common resource for obtaining clinical data from different medical databases. However, little is known about the coverage of the data types derived from the SS-MIX2 storage. In this regard, we calculated the proportions of a dataset that could be extracted via SS-MIX2 for various clinical study categories listed in various articles published in the New England Journal of Medicine. In the 95 articles reviewed, the proportions varied from 13.3% ± 13.3% (mean ± SD) for dementia to 61.8% ± 13.7% for diabetes. For cardiology, the proportion of data accessed in a unique format (SEAMAT) increased significantly. We further noted that there was room for improvement in the coverage of SS-MIX2 data.

Keywords:

Health information interoperability, Information systems

Introduction

In the era of big data, the sufficient utilization and analysis of large amounts of clinical data is important [1–3]. In Japan, the standardized structured medical information eXchange version 2 (SS-MIX2)—authorized by the Ministry of Health, Labour, and Welfare of the Japanese government in 2006—is commonly used as a standard data storage medium to share clinical data from various vendor-derived hospital information systems (HISs) [4]. Several projects—such as databases, data repositories, and regional health-information exchanges—using SS-MIX2 have been launched for storage and access of medical data [5–7]. In fact, the number of hospitals with the SS-MIX2 uploader was 1554 in March 2020 [8].

SS-MIX2 storage is divided into two categories. One is "standardized storage" (HL7 v2.5), which comprises standardized clinical data such as basic patient information, allergies, encounters, diagnoses, hospitalizations, prescriptions, and laboratory data. The other is "extension storage" comprising data that is not stored in standard storage. The data in the standard storage have been mainly used for clinical and research purposes because they are structured and have standard codes. However, comprehensive clinical research requires a variety of patient data. Since no studies have reported the extent to which SS-MIX2 data covers typical clinical research, this study aims to clarify the utility of the data archived in SS-MIX2.

In addition, we highlighted the impact of specific data stored in the extension storage of SS-MIX2; such data conforms to the standard export data format (SEAMAT) established in 2015 by the Japanese Circulation Society in conjunction with related cardiological associations and includes electrocardiograms (ECGs), echocardiograms or ultrasonic cardiograms (UCGs), and cardiac catheterization (CATH) data [9]. Through physical-examination and catheter report systems based on SEAMAT, specific cardiological data such as ECGs, UCGs, and cardiac CATHs can be transferred to the SS-MIX2 extension storage, resulting in efficient secondary use of these data for research purposes.

Methods

We evaluated the coverage of certain items in SS-MIX2 pertaining to patient characteristics in clinical research. We selected 140 original articles published in 2018 in the New England Journal of Medicine—from 378 vol 1 to 379 vol 10—as models of clinical research. The articles were categorized according to corresponding ailments. These categories included cancer, cardiology, stroke, infectious diseases, dementia, diabetes, respiratory illnesses, and blood disorders. Datasets were determined based on patient characteristics, as shown in Table 1 for each category. The ratio of datasets that could be sourced from the SS-MIX2 standardized storage to the total number of datasets evaluated in each category was calculated. We examined the average ratio and standard deviation for each category.

Next, we used SEAMAT data to evaluate the increase in the coverage of clinical data in cardiology. The statistical t-test was performed using R version 3.6.0. Significance was set at $p < 0.05$.

Results

We categorized 95 target articles into the eight areas. Of these, 26 corresponded to cancers, eight to heart diseases, seven to strokes, 23 to infectious diseases, two to dementia, five to diabetes, 11 to respiratory illnesses, and 13 to blood disorders. Tables 1 and 2 show examples of data coverage for diabetes and cancer categories, respectively [10, 11]. Basic patient information, including gender, age, and nationality (instead of race), as well as laboratory test data and prescriptions could be easily accessed from the SS-MIX2 standard storage. Diagnostic information such as disease classification and metastatic state was also available. However, blood pressure, pulse, smoking status, and performance status were not available in the SS-MIX2 standardized storage. Genomic information, including

mutation type and histological data, is also beyond the scope of this storage. In this case, the coverage rates for the datasets corresponding to diabetes and cancer were 78.6% (11/14) and 45.4% (5/11), respectively.

Figure 1 shows the coverage rates (ratios) for each category, as calculated above. Diabetes (61.8 ± 13.7%) had the highest proportion of datasets that could be sourced from the SS-MIX2 standard storage mainly in the form of laboratory data. The next category involved blood disorders (55.9 ± 17.1%), followed by cardiology (48.0 ± 13.0%), respiratory illnesses (45.0 ± 16.3%), and infectious diseases (43.6 ± 24.9%). The lowest category was dementia (13.3% ± 13.3%), which involved cognitive intelligence and mental status.

An example involving cardiology [12] is shown in Table 3, and was examined using SEAMAT. Data from ECGs, UCGs, and CATHs could be accessed in SEAMAT, resulting in an increase in coverage from 54.5% (6/11) to 90.9% (10/11). Figure 2 shows that the addition of data from SEAMAT significantly increased from 82.7 ± 0.1% to 48.0 ± 13.0%.

Table 1– Example of data coverage from SS-MIX2 standard storage in clinical research (Category: Diabetes)

Category: Diabetes **Article: Risk Factors, Mortality, and Cardiovascular Outcomes in Patients with Type 2 Diabetes (N Engl J Med 2018; 379:633-644)**		
Items	**Description**	**Availablity in SS-MIX2 standardized storage**
Female sex — no. (%)	Female	OK
Age — yr	75 years old	OK
Duration of diabetes — yr	3 year	OK
Age at diagnosis of diabetes — yr	72 years old	OK
Glycated hemoglobin	7.3%	OK
LDL cholesterol	140mg/dL	OK
Total cholesterol — mmol/liter	241mg/dL	OK
Current smoker — no. (%)		
Body-mass index		
Blood pressure — mm Hg		
Macroalbuminuria — no. (%)		OK
Estimated GFR — ml/min/1.73 m^2	40	OK
Treatment — statin	Rosuvastatin	OK
Treatment — Antihypertensive agent	Bisoprolol	OK
Total		11/14

Table 2– Example of data coverage from SS-MIX2 standard storage in clinical research (Category: Cancer)

Category: Cancer **Article: Osimertinib in Untreated EGFR-Mutated Advanced Non–Small-Cell Lung Cancer (N Engl J Med 2018; 378:113-125)**		
Items	**example**	**Availablity in SS-MIX2 standardized storage**
Age — yr	65 years old	OK
Male sex — no. (%)	Male	OK
Race — no. (%)	White	OK
Smoking status — no. (%)	Never	
WHO performance status — no. (%)	0, 1, Missing data	
Histologic type — no. (%)	Adeno-carcinoma	
Overall disease classification — no. (%)	Metastatic	OK
Metastases — no. (%)	Visceral metastases	OK
EGFR mutation type at randomization — no. (%)	Exon 19 deletion	
EGFR mutation type by central test — no. (%)	Exon 19 deletion	
EGFR-TKI comparator — no. (%)	Gefitinib	
Total		5/11

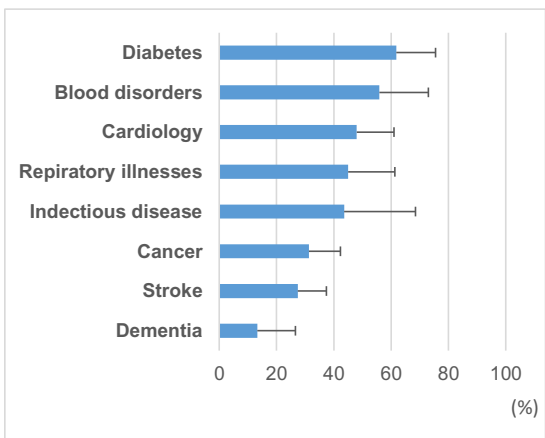

Figure 1. Ratio of data items that can be extracted from SS-MIX2 standard storage

Table 3– Example of data coverage from SS-MIX2 standard storage plus SEAMAT in clinical research (Category: Cardiology)

Category: Cardiology Article: Radial-Artery or Saphenous-Vein Grafts in Coronary-Artery Bypass Surgery (N Engl J Med 2018; 378:2069-2077)			
Items	**example**	**Availablity in SS-MIX2 standardized storage**	**Availablity in both SS-MIX2 standardized storage and SEAMAT data**
Age — yr	75 years old	OK	OK
Female sex — no. (%)	Female	OK	OK
Diabetes — no. (%)	Diabetes	OK	OK
Previous myocardial infarction — no. (%)	Old myocardial Infaction	OK	OK
Elective admission — no. (%)	not emergent admission	OK	OK
Renal insufficiency — no. (%)	Renal failure	OK	OK
Left ventricular ejection fraction <35% — no. (%)	30%		OK
Target vessel — no. (%)	Left circumflex coronary artery		OK
No. of grafts	3		OK
Proximal anastomosis site — no. (%)	Ascending aorta		OK
Outcome	Death, myocardial infarction, or repeat revascularization		
Total		6/11	10/11

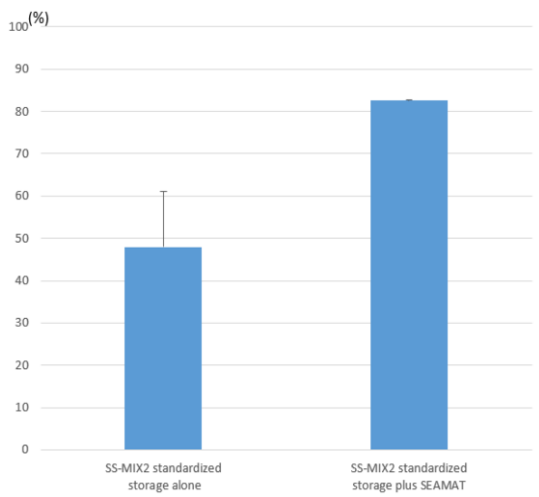

Figure 2. Ratio of data items that can be extracted from SS-MIX2 standard storage for the heart disease category and resulting ratio when SEAMAT data items are added (average value ± standard deviation).

Discussion

Comprehensive extraction of data from HISs and consequent efficient data-driven clinical research is a goal clinicians, epidemiologists, and medical information professionals aspire to achieve. However, there are various hurdles to overcome. First, interoperability is critical when collecting clinical data from different vendor-derived HISs [12]. Second, the limited coverage of data from real-world sources directly impedes research [13]. In this study, to clarify this problem, we illustrated the availability of health data from a standard Japanese medical repository, SS-MIX2. To the best of our knowledge, the present study is the first to quantitatively investigate the extent to which the data archived in SS-MIX2 cover the items of general clinical research.

Given that laboratory data is captured in SS-MIX2, a study targeting laboratory data is a good example of using data from SS-MIX2 standard storage. In fact, archived studies on diabetes and blood diseases were shown to have higher data availability. However, other research fields that require disease-specific data, patient activity, patient status, score, daily lifelog, and doctor's judgment showed deficits of indispensable information in SS-MIX2 standardized storage. Such information includes TNM classification and histology reports in cancer research, neurological function in research on stroke and dementia, and diagnostic-imaging findings in respiratory-organ research.

SEAMAT, a standard output format authorized by the largest Japanese association of cardiologists, the Japan Circulation Society (JCS), enables the transfer of numerical information generated by ECG, UCG, and CATH into the SS-MIX2 extension storage [9]. In fact, cardiovascular information can be collected from multiple facilities for research [14]. With SEAMAT, the coverage—especially the extraction of cardiology-specific data—was dramatically improved, suggesting that it is desirable to archive the results of highly specialized inspections. It is presumed that similar advances will be made in the other seven disease categories in future.

This study had several limitations. First, the format used was based on SS-MIX2, which is available only in Japan. However, the problem of data unavailability is universal. The current study aims to emphasize that specialists in medical informatics should aggressively address this issue. Second, our study was based on articles published in the New England Journal of Medicine, which may have caused selection bias. Different publications from other high-impact journals are required. Third, SEAMAT was the focus of this study. Nonetheless, other standard formats may be available for some categories. Since this study analyzed the availability of SS-MIX2 data, it would be better to investigate other formats such as fast healthcare interoperability resources (FHIR) [15] or observational health data sciences and informatics (OHDSI) [16] in subsequent studies.

Conclusions

SS-MIX2 standardized storage may automatically supply the majority of data items in studies that mainly incorporate blood tests, such as studies on diabetes and blood diseases. However, several categories require specific data types. Improved coverage of datasets is required to expedite data-driven clinical research.

Acknowledgments

We thank Ms. Chiaki Otomo, Mikiko Sato, Yukie Kobayashi, Mina Oikawa, and Misaki Arakawa for their technical support.

References

[1] Cassel C, Bindman A. Risk, Benefit, and Fairness in a Big Data World. JAMA 2019; 322:105–106.

[2] Vigilante K, Escaravage S, McConnell M. Big Data and the Intelligence Community - Lessons for Health Care. N Engl J Med 2019; 380:1888–1890.

[3] Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016; 375:2293–2297.

[4] Kimura M, Nakayasu K, et al. SS-MIX: A Ministry Project to Promote Standardized Healthcare Information Exchange. Methods Inf Med 2011; 50:131–139.

[5] Yamaguchi M, Inomata S, Harada S, Matsuzaki Y, Kawaguchi M, Ujibe M, et al. Establishment of the MID-NET® Medical Information Database Network as a Reliable and Valuable Database for Drug Safety Assessments in Japan. Pharmacoepidemiol Drug Saf 2019; 28:1395–1404.

[6] Sugiyama T, Miyo K, Tsujimoto T, Kominami R, Ohtsu H, Ohsugi M, et al. Design of and Rationale for the Japan Diabetes compREhensive Database Project Based on an Advanced Electronic Medical Record System (J-DREAMS). Diabetol Int 2017; 8:375–382.

[7] Ido K, Nakamura N, Nakayama M, Miyagi Medical and Welfare Information Network. Miyagi Medical and Welfare Information Network: A Backup System for Patient Clinical Information after the Great East Japan Earthquake and Tsunami. Tohoku J Exp Med 2019; 248:19–25.

[8] http://www.ss-mix.org/cons/ [Accessed: March 25, 2021].

[9] Nakayama M, Takehana K, Kohro T, Matoba T, et al. Standard Export Data Format for Extension Storage of Standardized Structured Medical Information Exchange. Rep.:587-616. Circ Rep 2020; 2:587–616.

[10] Rawshani A, Rawshani A, Franzén S, Sattar N, et al. Risk Factors, Mortality, and Cardiovascular Outcomes in Patients with Type 2 Diabetes. N Engl J Med 2018; 379:633–644.

[11] Soria JC, Ohe Y, Vansteenkiste J, Reungwetwattana T, Chewaskulyong B, et al. Osimertinib in Untreated EGFR-Mutated Advanced Non-Small-Cell Lung Cancer. N Engl J Med 2018; 378:113–125.

[12] Gaudino M, Benedetto U, Fremes S, Biondi-Zoccai G, Sedrakyan A, et al. Radial-Artery or Saphenous-Vein Grafts in Coronary-Artery Bypass Surgery. N Engl J Med 2018; 378:2069–2077.

[13] Sorace J, Wong HH, DeLeire T, Xu D, et al. Quantifying the Competitiveness of the Electronic Health Record Market and Its Implications for Interoperability. Int J Med Inform 2020; 136:104037.

[14] Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence. JAMA Netw Open 2019; 2:e1912869.

[15] Matoba T, Kohro T, Fujita H, Nakayama M, Kiyosue A, et al. Architecture of the Japan Ischemic Heart Disease Multimodal Prospective Data Acquisition for Precision Treatment (J-IMPACT) System. Int Heart J 2019; 60:264–270.

[16] Kasthurirathne SN, Mamlin B, Grieve G, Biondich P. Towards Standardized Patient Data Exchange: Integrating a FHIR Based API for the Open Medical Record System. Stud Health Technol Inform 2015; 216:932.

[17] Hripcsak G, Duke JD, Shah NH, Reich CG, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015; 216:574–578.

Address for correspondence

Masaharu Nakayama 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8574, Japan. Phone: +81-22-717-7572, FAX: +81-22-717-7505. Email: nakayama@cardio.med.tohoku.ac.jp