



A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research

Journal:	<i>BMJ Open</i>
Manuscript ID:	bmjopen-2012-001622
Article Type:	Research
Date Submitted by the Author:	28-Jun-2012
Complete List of Authors:	Yamamoto, Keiichi; Kyoto University Hospital, Department of Clinical Trial Design and Management, Translational Research Center Sumi, Eriko; Kyoto University Hospital, Department of Clinical Innovative Medicine, Translational Research Center Yamazaki, Toru; Kyoto University, Department of Oral and Maxillofacial Surgery, Graduate School of Medicine Asai, Keita; Kyoto University, Department of Oral and Maxillofacial Surgery, Graduate School of Medicine Yamori, Masashi; Kyoto University, Department of Oral and Maxillofacial Surgery, Graduate School of Medicine Teramukai, Satoshi; Kyoto University Hospital, Department of Clinical Trial Design and Management, Translational Research Center Bessho, Kazuhisa; Kyoto University, Department of Oral and Maxillofacial Surgery, Graduate School of Medicine Yokode, Masayuki; Kyoto University Hospital, Department of Clinical Innovative Medicine, Translational Research Center Fukushima, Masanori; Foundation for Biomedical Research and Innovation, Translational Research Informatics Center
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Public health, Research methods, Epidemiology
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, ORAL & MAXILLOFACIAL SURGERY, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS, Clinical trials < THERAPEUTICS

SCHOLARONE™
Manuscripts

1
2
3
4 **A pragmatic method for electronic medical record-based observational studies:**
5
6
7 **developing an electronic medical records retrieval system for clinical research**
8
9

10 Keiichi Yamamoto¹, Eriko Sumi², Toru Yamazaki³, Keita Asai³, Masashi Yamori³, Satoshi
11 Teramukai¹, Kazuhisa Bessho³, Masayuki Yokode², Masanori Fukushima⁴
12
13
14
15

16
17 ¹Department of Clinical Trial Design and Management, Translational Research Centre, Kyoto
18 University Hospital, Kyoto, Japan
19
20
21
22

23
24 ²Department of Clinical Innovative Medicine, Translational Research Centre, Kyoto
25 University Hospital, Kyoto, Japan
26
27
28

29
30
31 ³Department of Oral and Maxillofacial Surgery, Graduate School of Medicine, Kyoto
32 University, Kyoto, Japan
33
34
35

36
37 ⁴Translational Research Informatics Centre, Foundation for Biomedical Research and
38 Innovation, Kobe, Japan
39
40
41

42
43
44 Corresponding author: Keiichi Yamamoto, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto,
45 606-8507 Japan. E-mail: kyamamo@kuhp.kyoto-u.ac.jp, Tel: +81-75-751-4717, Fax
46 +81-75-751-3399
47
48
49
50

51
52
53
54 Total number of words: 3,542
55

56
57
58 Short title: A pragmatic method for EMR-based observational studies
59
60

1
2
3
4 Keywords: clinical research informatics, data warehouse, OLAP, computable eligibility
5
6
7 criteria, pharmacoepidemiology, hospital-based cohort study
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

ABSTRACT

Objective: Utilising data collected via electronic medical records (EMRs) is necessary to improve clinical research efficiency. However, it is not easy to identify patients who meet research eligibility criteria and collect the necessary information from EMRs, as the data collection process must integrate various techniques, including developing a data warehouse and translating eligibility criteria into computable criteria. This research aims to establish a pragmatic method optimised for patient identification and collect necessary clinical research information from EMRs. **Design:** Qualitative analyses. **Participants:** At our hospital, 800,000 cases of clinical information have been stored in EMRs. **Primary and secondary outcome measures:** To evaluate the feasibility and usefulness of the ERS, the method to convert text form eligible criteria to computable criteria, and a reconfirmation method to increase research data accuracy. **Results:** To comprehensively and efficiently collect information from patients participating in clinical research, we developed an EMR retrieval system (ERS). To design the ERS database, we modified the star schema and designed a new multi-dimensional data model optimised for patient identification. We also devised practical methods to translate narrative eligibility criteria into computable parameters. We applied the system to an actual hospital-based cohort study performed at our hospital. We converted the test results into computable criteria. Based on this information, we identified eligible patients and extracted data necessary for confirmation by our investigators and statistical analyses with our ERS.

1
2
3
4 **Conclusion:** We propose a pragmatic methodology to identify patients from EMRs who met
5
6
7 clinical research eligibility criteria. Our ERS allowed for the efficient collection of
8
9
10 information on the eligibility of a given patient. The method proposed here reduces the labour
11
12
13 required from the investigators. We believe an efficient ERS is essential to facilitate clinical
14
15
16 research that utilises EMRs.

17 18 19 **ARTICLE SUMMARY**

20 21 22 **Article focus**

23
24
25
26
27 To establish a pragmatic methodology to efficiently collect information from patients who
28
29
30 meet clinical research eligibility criteria from EMRs.
31

32 33 34 **Key messages**

35
36
37 Utilising data collected as electronic medical records (EMRs) is necessary to improve clinical
38
39
40 research efficiency. However, it is not easy to identify patients who meet research eligibility
41
42
43 criteria and collect necessary data from EMRs, as the data collection process must integrate
44
45
46 various techniques, including developing a data warehouse and translating eligibility criteria
47
48
49 into computable criteria. An efficient ERS that integrates these techniques is essential to
50
51
52 facilitate clinical research that utilises EMRs.
53

54 55 56 **Strengths and limitations of this study**

57
58
59
60

- The strengths of our method include using a specialised data model for patient identification in clinical research and efficient data conversion without being conscious of the EMR database structure when converting narrative criteria to computable criteria.
- Our ERS cannot retrieve information that is not in the data model. It is thus necessary to create eligibility criteria assuming the use of our ERS in the protocol development stage.
- Enabling ERS use in multiple institutions is an important future task.

BACKGROUND

Medical information technology has recently advanced in many countries, and enormous amounts of clinical data are already stored as electronic medical records (EMRs). Utilising the data collected in EMRs is necessary to improve clinical research efficiency. An EMR is a large database of patient data and is used in observational research investigating the relationships among diseases, treatments, and outcomes; conducting surveillance for rare drug reactions; and recruiting patients for clinical trials [1-10]. However, it is not easy to identify patients who meet research eligibility criteria and collect necessary information from EMRs. Herein, we describe three major issues concerning EMR-based observational studies: EMR patient data retrieval function, eligibility criteria protocol representation, and EMR data accuracy. [11-14]

To identify patients who meet research eligibility criteria, it is necessary to obtain various types of information stored in EMRs by subject, e.g., diagnosis and prescribed medications. However, because the EMR database structure is designed to facilitate online transaction processing to enable rapid and detail-oriented clinical information searches and updates on individual patients [12-13], current EMR system do not facilitate this retrieval function [11-13, 15]. Data warehouses are essential components of data-driven decision support. To allow efficient research analyses, EMR data must first be warehoused to allow data analyses across patient populations [16-23]. However, health care data modelling is difficult and

1
2
3
4 time-consuming because of the complexity of the medical knowledge involved. Thus, the
5
6
7 most common approaches to clinical data warehouse modelling are variations on the
8
9
10 entity-attribute-value (EAV) model [24-30], where data are stored in a single table with three
11
12
13 columns: entity identification, attribute, and attribute value. The EAV design has advantages,
14
15 including flexibility and ease of storage; however, it requires transforming EAV data into
16
17
18 another analytical format before analysis [25-30]. Online analytical processing (OLAP),
19
20
21 which facilitates flexible data navigation and analyses, is most frequently used for searching
22
23
24 data stored in the data warehouse [16-19, 31-34]. OLAP systems in relational databases are
25
26
27 typically designed based on Kimball's star or snowflake schemas [16-19, 31, 33]. However,
28
29
30 the star schema was devised to facilitate online measurement analyses [16-19, 31, 33]. In
31
32
33 health care, this method can be used to dynamically gather online analyses of numeric data
34
35
36 (e.g., a specific dose of a drug for a specific disease or the time required for a specific
37
38
39 operation) in clinical practice. It is thus not suitable for identifying patients who meet the
40
41
42 complicated eligibility criteria for a given clinical research.

43
44
45 Current eligibility criteria are written in a text format that cannot be computationally
46
47
48 processed. Several investigations have sought to establish computable eligibility criteria
49
50
51 [35-42]. However, eligibility criteria are not yet completely standardised. Using natural
52
53
54 language processing (NLP) technologies, converting the text format of eligibility criteria to a
55
56
57 computer or extracting patient identifications from EMRs is far from perfect without human
58
59
60

1
2
3
4 intervention [13, 43-44].
5
6

7
8 Current EMRs have been used to support claims for medical service fees and the treatments
9
10 administered to each patient; therefore, data gathered specifically for research purposes may
11
12 be incomplete and unreliable [12-13].
13
14

15
16
17 Although various investigations on each technique are executed individually, standardised
18
19 methods must still be established that integrate these techniques, facilitate identifying patients
20
21 who are eligible for clinical research, and collect necessary information from EMRs.
22
23
24

25 26 27 **Objective** 28

29
30
31 To utilise EMRs efficiently in clinical research, we considered it necessary to develop an
32
33 EMR retrieval system (ERS) to collect data from patients who meet the eligibility criteria for
34
35 a study and establish practical methods to utilise the system.
36
37

38
39
40 This research aims to establish a pragmatic method optimised for patient identification and
41
42 collect necessary information from EMRs for clinical research. These tools are implemented
43
44 as an ERS. We apply the system to an actual hospital-based cohort study and conduct
45
46 qualitative analyses to evaluate the feasibility and usefulness of the ERS, the method to
47
48 convert text form eligible criteria to computable criteria, and a reconfirmation method to
49
50 increase research data accuracy.
51
52
53
54
55
56
57
58
59
60

MATERIALS AND METHODS

EMR retrieval system

In our hospital, EMR use was introduced in 2005; approximately 800,000 cases of clinical information have already been stored. To collect information for patients participating in clinical research comprehensively and efficiently, we developed an ERS [45].

We identified nine data categories (i.e., entities) from EMRs that are useful for clinical research: demographic characteristics, physical findings, diagnostic studies, laboratory tests, diagnoses, progress reports on an EMR template [46-50], medications and injections, operation records, and other treatments.

In designing the ERS database, we designed a new data model based on the star schema, optimised for patient identification in clinical research. Figure 1 presents our data model. In our data model, all entities in a given schema are independent and complete; this allows for logical operations and creating eligible patient lists for each respective parameter in a study [51]. The target patient list is made by combining these patient lists. The data model also supports the inference of medical concepts expressed in eligibility criteria in reference to corresponding patient data accumulated in EMRs [35].

To ensure that the data retrieval process is practical and independent of the EMR system structure, a data warehouse (i.e., data mart [52]) was created on a relational database

1
2
3
4 management system by extracting, transforming, and loading information from the EMR
5
6
7 system [16-23].
8
9

10 An OLAP tool was installed to efficiently search through data from multiple patients [16-19,
11
12 31-34]. The OLAP tool runs in an Internet browser and can generate structured query
13
14 language (SQL) based on predefined metadata (i.e., a data model) by defining logical queries
15
16 (i.e., programs) using a graphical user interface (GUI) [16-19, 31-34]. Moreover, it allows
17
18 reports on information retrieved from the browser to be transcribed using hypertext markup
19
20 language (HTML). The reports are created in various formats, including portable document
21
22 format (PDF), comma separated values (CSV), and extensible markup language (XML).
23
24
25
26
27
28
29
30

31 To protect personal information in medical records at our hospital, the EMR network is
32
33 separated physically from other networks. Our data warehouse and OLAP servers are
34
35 deployed in the same EMR network and managed using the same EMR security policies.
36
37
38
39 Additionally, using our ERS is limited only to clinical research approved by the ethics
40
41 committee at our hospital. Only designated staff member at our centre are allowed to retrieve
42
43 data. Our centre creates and manages ERS user identification separate from the EMRs. For
44
45 the external output of CSV and other data, permission must be obtained from our department
46
47 of medical informatics and data extraction must be executed in the presence of supervisors
48
49 who are responsible for protecting personal information at our hospital.
50
51
52
53
54
55
56
57
58
59
60

Application to clinical research

We applied the system to a hospital-based cohort study performed at our hospital, titled 'Risk of osteomyelitis of the jaw induced by oral bisphosphonates (BP) in patients taking medications for osteoporosis: a hospital-based cohort study in Japan', in which we identified eligible patients, extracted research data, and evaluated the feasibility of our system. The ethics committee at Kyoto University Hospital approved this research. A different paper details the purpose, methods, results, and discussion of this research.

This research aims to estimate the risks for osteomyelitis of the jaw in osteoporosis patients at our hospital who had been exposed to oral BP compared to those who had not [53-54].

The eligibility criteria were as follows.

Inclusion criteria

- Patients diagnosed with osteoporosis and treated with osteoporosis medications at Kyoto University Hospital between November 2000 and October 2010.
- Patients aged 20 years or older.

Exclusion criteria

- Patients with a history of treatment with radiation therapy to the maxillofacial region.

- Patients with primary or metastatic tumours in the maxillofacial region.
- Patients treated with intravenous BP.

The data collected were diagnosis, date of diagnosis, sex, birthday, and the doses and dates when osteoporosis medications, steroids, biopharmaceuticals in rheumatic diseases, disease modifying antirheumatic drugs, anticancer drugs, diabetes drugs and HbA1c test were administered.

Patient identification and data collection using our ERS

To identify patients who meet the eligibility criteria for the clinical research in question, data were collected using the ERS in the following ways:

- 1) Convert the text form of the narrative criteria into computable criteria.
- 2) Create a targeted patient list.
- 3) Add a flag for investigators to confirm the targeted patient list.
- 4) Create reports for the investigators to confirm.

We show the details below.

Convert the text form of the narrative criteria to computable criteria

To identify eligible patients and collect the necessary data from EMRs, narrative criteria and

1
2
3
4 data must be converted to computable criteria. Such computable criteria include entities,
5
6
7 attributes, logical operators (i.e., 'and' and 'or'), codes, and parameters [35-42]. The
8
9
10 clinical research purpose and clinical practice demands made it necessary to perform this
11
12
13 task.

14
15
16 As an example of conversion from narrative criteria to computable criteria, we present the
17
18
19 following two-step conversion procedure.
20

21
22
23 Step 1: Convert narrative criteria into entity-level criteria.
24

25
26
27 Medical concepts expressed as narrative criteria are mapped onto entities in the data model
28
29
30 and converted into entity-level criteria. For each entity, a criterion is created to extract
31
32
33 patients who meet each condition. If exclusive conditions for the same entity must be defined,
34
35
36 a different criterion is created. In this study, we mapped 'osteoporotic patients' onto two
37
38
39 entities (i.e., 'diagnosis' and 'medications and injections') and converted it to a combination
40
41
42 of two criteria (i.e., 'diagnosis of osteoporosis' and 'osteoporosis drug administration'). This
43
44
45 process reflects that the test research aims to estimate some risks of osteomyelitis of the jaw
46
47
48 with BP administration instead of diagnosing osteoporosis patients accurately. The recorded
49
50
51 diagnosis in the EMR was typically designed to ensure payment for medical claims. We thus
52
53
54 sought to reduce the number of false-positives by extracting patients with a given treatment
55
56
57 type. This task was performed at the protocol development stage of the study.
58
59
60

Step 2: Convert entity-level criteria into attribute-level criteria (i.e., computable criteria).

Medical concepts expressed in the entity-level criteria are mapped onto attributes in the data model; these become computable criteria by specifying the corresponding date and codes [55].

For this study, Table 1 presents a “diagnosis of osteoporosis and osteoporosis drug administration and no intravenous BP administration (i.e., exclude ‘intravenous BP administration’)”.

Table 1. Example of computable criteria to create the targeted patient list

Criterion	Entity	Operator symbol	Attribute	Operator symbol	Parameter	SQL
Inclusion: Osteoporosis diagnosis	Diagnosis	-	ICD10Code	in	(ICD10 code list)	Select PatientId From
		and	DiagnosisDate	>=	'10/01/2000'	Diagnosis
		and	DiagnosisDate	<=	'09/30/2010'	Where ICD10Code in
		and	SuspectedFlag	=	Fixed	(ICD10 code list) and DiagnosisDate >= '10/01/2000' and DiagnosisDate <= '09/30/2010' and SuspectedFlag = 'Fixed'
Inclusion: Osteoporosis drug administrations	Medications and Injections	-	DrugCode	in	(drug code list)	Select PatientId From
		and	ExecuteDate	>=	'10/01/2000'	MedicationsAndInjections
		and	ExecuteDate	<=	'09/30/2010'	Where DrugCode in (drug code list) and ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010'
Exclusion: Intravenous BP	Medications and	-	DrugCode	in	(drug code list)	Select PatientId From
		and	ExecuteDate	>=	'10/01/2000'	MedicationsAndInjections

administrations	Injections	and	ExecuteDate	<=	'09/30/2010'	Where DrugCode in (drug code list) and ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010'
-----------------	------------	-----	-------------	----	--------------	---

BP: bisphosphonates; ICD: International Classification of Diseases; ID: identifications; SQL: structured query language.

Creating a targeted patient list

A targeted patient list is created from the entire set of patients for whom EMRs have been obtained by defining logical queries (i.e., programs defined by the GUI) based on the computable criteria included in the ERS.

Logical queries are first defined in the ERS to identify patients who meet the conditions for each criterion. The ERS automatically generates the SQL necessary for data extraction according to the logical queries. Logical queries are then defined to include or exclude eligible patients who meet each criterion for the demographic entity. The targeted patient list is created by executing the logical query. Figure 2 presents an example of SQL automatically generated by the ERS.

We thus designed our data model to enable creating the targeted patient list by defining the patients extracted from each criterion (i.e., 'in' or 'not in') as conditions for the demographic entity that was the unique patient list for the entire hospital. If logical queries are defined

using our method, even if the eligibility criteria are complicated, it is not necessary to dramatically change the SQL structure generated in the ERS.

Flagging entries for investigators to confirm

To improve research data accuracy, confirmation by the investigators was necessary. When confirmation is required, additional information is linked.

For the targeted patient list, logical queries are defined to flag certain items according to the investigators' interest. Necessary logical queries are first defined for each criterion. Logical queries are then defined for addition to the patient list as '1' if the data correspond or '0' if they do not. Data sets created by these operations are joined by 'union' and pivoted on a cross-tabulation list using statistical analysis software. We show an example of computable criteria that contain two criteria (i.e., 'oral BP administration' and 'diagnosis of osteomyelitis of the jaw') in Table 2 and SQL generated by the ERS in Figure 3. We also present the dataset image created by these operations in Figure 4.

Table 2. Example of computable criteria to be flagged for confirmation by investigators

Criterion	Entity	Operator symbol	Attribute	Operator symbol	Parameter	SQL
Oral BP administrations	Medications	-	DrugCode	in	(drug code list)	Select PatientId From
	and	and	ExecuteDate	>=	'10/01/2000'	MedicationsAndInjections
	Injections	and	ExecuteDate	<=	'09/30/2010'	Where DrugCode in (drug code list)and

						ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010'
Diagnosis of inflammatory conditions of jaws	Diagnosis	-	ICD10Code	in	(ICD10 code list)	Select PatientId From Diagnosis
		and	DiagnosisDate	>=	'10/01/2000'	Where ICD10Code in
		and	DiagnosisDate	<=	'09/30/2010'	(ICD10 code list) and
		and	SuspectedFlag	=	Fixed	DiagnosisDate >= '10/01/2000' and DiagnosisDate <= '09/30/2010' and SuspectedFlag = 'Fixed'

BP: bisphosphonates; ICD: International Classification of Diseases; ID: identifications; SQL:
structured query language.

Create reports for investigators to confirm

To help investigators confirm the targeted patient list, reports are created by linking the findings for diagnostic imaging, pathological diagnosis, operations, and others. Investigators confirm these entries using the reports and EMR information, including progress notes and images. When the diagnosis history, medication, laboratory results, progress notes, and other information are necessary, the same operation is executed for each instance. We present the dataset image generated by this operation in Figure 5. The reports may improve the investigators' confirmation efficiency, because it prevents referring to the medical records for each patient who needs confirmation.

Systemic evaluation

To evaluate our system, we collected information about the research period using the recall method. For the accuracy of the data collected by the ERS, we evaluated the results after

1
2
3
4 investigator confirmation. We also asked the investigators to evaluate the system in a
5
6
7 questionnaire.
8

9 10 **RESULTS**

11 12 13 **Computable criteria, datasets, and system evaluation**

14
15
16
17
18 We present the computable criteria in Table 3. To increase data accuracy, we collected all
19
20
21 exclusion criteria for the investigators to confirm. As Table 3 shows, we extracted
22
23
24 information from EMRs. For investigator confirmation, we also reported all targeted patients
25
26
27 using the following lists: osteoporosis drugs administered, oral BP administered, intravenous
28
29
30 BP administered, diabetes drugs administered, anticancer drugs administered,
31
32
33 biopharmaceuticals in rheumatic diseases administered, disease-modifying antirheumatic
34
35
36 drugs, steroid drugs administered, osteoporosis diagnoses, oral cancer diagnoses, patients
37
38
39 diagnosed with inflammation of the jaw, patients diagnosed with other suspicious diseases,
40
41
42 patients diagnosed with diabetes, patients diagnosed with rheumatoid arthritis, patients
43
44
45 diagnosed with Sjogren's syndrome, HbA1c values, radiological findings, pathological
46
47
48 findings, and radioisotope findings. These data were extracted from the ERS for statistical
49
50
51 analyses, presented in CSV format, and analysed using statistics software.
52
53

54 **Table 3.** Computable criteria for our test research
55
56
57
58
59
60

Criterion	Entity	Operator symbol	Attribute	Operator symbol	Parameter
Create a targeted patient list					
Inclusion criteria: Osteoporosis diagnosis	Diagnosis	-	ICD10Code	in	(osteoporosis ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Inclusion criteria: Osteoporosis drug administrations	Medications and Injections	-	DrugCode	in	(osteoporosis drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
	and	ExecuteDate	<=	'09/30/2010'	
Add a flag for investigators to confirm the targeted patient list					
Exclusion criteria: Oral cancer diagnosis	Diagnosis	-	ICD10Code	in	(oral cancer ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Exclusion criteria: Intravenous BP administrations	Medications and Injections	-	DrugCode	in	(intravenous BP drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
Oral BP administrations	Medications and Injections	-	DrugCode	in	(oral BP drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
Inflammatory jaw condition diagnosis	Diagnosis	-	ICD10Code	in	(inflammatory conditions of jaws ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Other suspicious disease diagnosis	Diagnosis	-	ICD10Code	in	(other suspicious disease ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Rheumatoid arthritis diagnosis	Diagnosis	-	ICD10Code	in	(rheumatoid arthritis ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'

		and	SuspectedFlag	=	Fixed
Sjogren's syndrome diagnosis	Diagnosis	-	ICD10Code	in	(sjogren's syndrome ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Diabetes diagnosis	Diagnosis	-	ICD10Code	in	(diabetes ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Steroid drug administrations	Medications and Injections	-	DrugCode	in	(steroid drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
Biopharmaceuticals in rheumatic diseases administrations	Medications and Injections	-	DrugCode	in	(biopharmaceuticals in rheumatic diseases code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
Disease-modifying antirheumatic drug administrations	Medications and Injections	-	DrugCode	in	(disease modifying antirheumatic drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
Anticancer drug administrations	Medications and Injections	-	DrugCode	in	(anticancer drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
Diabetes drug administrations	Medications and Injections	-	DrugCode	in	(diabetes drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
HbA1c test execution	Laboratory Test	-	LaboratoryTestCode	in	(HbA1c test code)
		and	TestDate	>=	'10/01/2000'
		and	TestDate	<=	'09/30/2010'
Create reports for confirmation by the investigators					
Radiological finding reports	Diagnostic Studies	-	ReportName	in	(report name list of oral region)
Pathologic finding reports	Diagnostic Studies	-	SampleName	contains	'bone'
		or	SampleName	contains	'jaw'

Radio isotope finding reports	Diagnostic Studies	-	-	-	-
-------------------------------	--------------------	---	---	---	---

BP: bisphosphonates; ICD: International Classification of Diseases; ID: identifications

Among the approximately 800,000 cases at our hospital, 8,772 were categorised using the terms 'Inclusion criteria: Osteoporosis diagnosis'; among this group, 7,195 were further categorised using 'Inclusion criteria: Osteoporosis drug administration'. We then calculated the time that had elapsed since the osteoporosis diagnosis, determined that 7,062 patients were aged 20 years or older and created a targeted patient list. Among those on the targeted patient list, 23 were placed under the heading 'Exclusion criteria: Oral cancer diagnosis', 110 under 'Exclusion criteria: Intravenous BP administration', 4,200 under 'Oral BP administration', 84 under 'Inflammatory jaw condition diagnosis', 2,064 as 'Other suspicious disease diagnosis', 1,176 as 'Rheumatoid arthritis diagnosis', 394 as 'Sjogren's syndrome diagnosis', 1,700 as 'Diabetes diagnosis', 4,551 as 'Steroid drug administration', 186 as 'Biopharmaceuticals in rheumatic diseases administrations', 1,279 as 'Disease-modifying antirheumatic drug administrations', 904 as 'Anticancer drug administrations', 1,055 as 'Diabetes drug administrations', and 3,641 as 'HbA1c test execution'. Because of the end point considered, patients who were classified under 'Inflammatory jaw condition diagnosis' or 'Other suspicious disease diagnosis' were confirmed by investigators, who performed the statistical analyses and arranged the research results.

1
2
3
4 The accuracy of the data extracted by the ERS was characterised as follows. Reviewing the
5
6
7 medical records revealed that 2,817 patients were not labelled as 'Oral BP administration',
8
9
10 including seven (one who received intravenous BP) treated at other hospitals. Six patients had
11
12
13 been treated with radiation therapy to the oral and maxillofacial regions. Among 72 classified
14
15
16 under 'Inflammatory jaw condition diagnosis', 35 cases and 37 non-cases were identified.

17
18
19 We present the timeline required for the test research. The data extraction period lasted
20
21
22 approximately three months. Ten meetings were held to create and validate the computable
23
24
25 criteria and the list of codes for various drugs and diagnoses (i.e., International Statistical
26
27
28 Classification of Diseases (ICD)-10 [56]). The time required for logical query definition
29
30
31 when using the ERS was approximately 20 hours. The investigator confirmations and
32
33
34 statistical analyses took approximately four months.

35
36
37 The investigators evaluating the system mentioned that 1) it enabled them to extract the
38
39
40 necessary data for diagnosis and drug administration without exception; 2) by screening the
41
42
43 entire patient population at the hospital using the ERS, they could identify not only eligible
44
45
46 patients in the department of oral and maxillofacial surgery but also all eligible patients,
47
48
49 which reduced the study bias; and 3) by creating reports for confirmation, it enabled
50
51
52 investigators to devote their time to reading images, thus effectively reducing the time
53
54
55 required for reviewing medical records.
56
57
58
59
60

DISCUSSION

We identified eligible patients for this research and extracted the data necessary for confirmation by investigators and statistical analyses. Using the ERS allowed the collection of information on patient eligibility by efficiently combining clinical information. Our proposed method also reduced the labour required from investigators, indicating that it was useful.

To design the ERS database, we modified the star schema and designed a new data model optimised for patient identification. To compare our data model with the star schema, we present an example of a clinical data model designed based on the star schema [34] in Figure 6. The main differences between our data model and the star schema were 1) demographic data, which were presented in list form in our EMR system, were presented as a fact-less fact table, and 2) date and time, measurements (i.e., facts) and text information were presented in dimension tables [43]. The most significant characteristic of our method for patient identification is using a specialised data model for patient identification in clinical research. Data can be converted efficiently without being conscious of the EMR database structure when converting narrative criteria to computable criteria. In this research, we considered whether data were extracted directly from EMRs at the protocol development stage. However, EMR data were recorded in a sequential format for every medical practice, and the database structure was complicated. Comprehending the location and meaning of the necessary data

1
2
3
4 thus required tremendous effort. It was difficult to make precise logical queries for patient
5
6
7 identification. However, because our ERS data model was arranged by subjects (e.g., tests,
8
9
10 diagnosis, or medications), it was easy to interpret the available information. Due to the
11
12
13 standardisation of computable criteria and SQL possible with the ERS, it was also possible to
14
15
16 create computable criteria in little time. Additionally, verifying the patient identification
17
18
19 accuracy was easy because it was possible to test each individual criterion.

20
21
22 The SQL generated by our ERS does not reduce the time required for data retrieval. Our ERS
23
24
25 also cannot retrieve information that is not in the data model. It is thus necessary to create
26
27
28 eligibility criteria assuming the use of our ERS from the protocol development stage. Current
29
30
31 EMRs do not store all necessary data for clinical research, including information related to
32
33
34 pregnancy, performance status, cancer stage, availability of transportation to the hospital,
35
36
37 specific tests that are not typically performed, drug regimen, outcomes including death, and
38
39
40 adverse events. Additionally, all tests are not administered to all patients, and necessary
41
42
43 information may have been recorded in medical records at another hospital [12-13]. To
44
45
46 facilitate EMR use in clinical research, it is necessary to accumulate as much of this
47
48
49 information as possible. In the hospital, much information does not integrate well with EMRs,
50
51
52 including test reports stored only in the departmental system and departmental research
53
54
55 databases [57]. It is important to utilise this information. Additionally, enabling ERS use in
56
57
58 multiple institutions is also an important future task.
59
60

1
2
3
4 Our conversion method depends on high-level medical decisions by investigators. Some
5
6
7 medical concepts may be interpreted differently depending on the research and the
8
9
10 investigator caring for the patients. It is necessary both to change the tacit knowledge of the
11
12
13 investigators about converting computable criteria to explicit knowledge and to standardise
14
15
16 this knowledge [58-63]. To reduce criteria conversion and investigator confirmation, it may
17
18
19 be useful to apply NLP or decision support technology in combination with our system.
20
21
22 Moreover, it is important to further discuss computable eligibility criteria standardisation.
23
24
25 The attribute-level criteria that describe the search conditions in detail may be useful in
26
27
28 global studies that address diseases that vary according to the diagnostic criteria used in each
29
30
31 country.

32
33
34 Concerning EMR data accuracy, the ICD10 code (osteomyelitis of the jaw) sensitivity was
35
36
37 48.6% (35/72). The investigators reported six simple diagnosis errors, seven oral BP
38
39
40 administrations at other hospitals, and six patients who were treated with radiation therapy in
41
42
43 the oral and maxillofacial region. For the accuracy of current EMRs, the investigators had to
44
45
46 confirm the information. However, the EMRs provided rich confirmation data and were
47
48
49 useful in improving research data accuracy. In this study, we checked the data from actual
50
51
52 EMRs manually and identified patients precisely and extensively using coded information,
53
54
55 narrative information, and images. However, only information from existing EMRs was
56
57
58 available. Current EMRs have a high degree of flexibility in data entry and are not currently
59
60

1
2
3
4 managed for research purposes, which decreases their reliability [12-13]. It is necessary to
5
6
7 improve data quality through quality control without placing too much of a burden on clinical
8
9
10 practice. Alternatively, it may be possible to organise data sufficiently before research use
11
12 [64-66]. Standardising terminology and exchange formats characteristic of healthcare has
13
14 facilitated international discourse [56, 67-73]. It is necessary to further discuss not only
15
16 clinical practice but also research purposes, particularly how to utilise such various standards
17
18 when using EMRs beyond the hospital setting.
19
20
21
22
23

24 25 **CONCLUSION**

26
27
28 We proposed a pragmatic method for EMR-based observational studies. Our ERS is already
29
30 used to support hospital-based cohort studies, clinical trial recruitment, and the eClinical trial
31
32 infrastructure [45] at our centre. We believe an efficient ERS is essential to facilitate clinical
33
34 research that utilises EMRs.
35
36
37
38
39

40 41 **Acknowledgements**

42
43
44
45 The authors would like to acknowledge the staff of the Department of Medical Informatics of
46
47
48 Kyoto University Hospital for their generous support.
49
50

51 52 **Funding**

53
54
55
56 This work was supported by Coordination, Support and Training Program for Translational
57
58
59
60

1
2
3
4 Research of Ministry of Education, Culture, Sports, Science and Technology of Japan and
5
6
7 Grants-in-Aid for Scientific Research of Japan (23790566).
8
9

10 **Competing interests**

11
12
13
14 None.
15
16

17 **Contributors**

18
19
20
21
22 KY designed the study, developed the ERS system, identified the computable eligibility
23
24 criteria, wrote logical queries, collected data, and wrote the manuscript. ES is grant holder,
25
26
27 designed the study, developed the ERS system, and wrote and edited the manuscript. TY
28
29
30 designed and conducted the 'Risk of osteomyelitis of the jaw induced by oral
31
32 bisphosphonates in patients taking medications for osteoporosis: a hospital-based cohort
33
34 study in Japan' (BRONJ study) study and this study, identified the computable eligibility
35
36
37 criteria, and wrote and edited the manuscript. KA and MY designed and conducted the
38
39
40 BRONJ study. ST is designed the study and provided comments and feedback on the study.
41
42
43
44 KB is the principal investigator of the BRONJ study. MY is the owner of the ERS system and
45
46
47 supervised the study. MF supervised the study and provided comments and feedback on the
48
49
50 study. All of the authors read and approved the final manuscript.
51
52
53
54

55 **Provenance and peer review**

1
2
3
4 Not commissioned; externally peer reviewed.
5
6

7
8 **Data sharing statement**
9

10
11 No other data are available to share.
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

REFERENCES

1. Dean BB, Lam J, Natoli JL, et al. Review: Use of electronic medical records for health outcomes research: A literature review. *Med Care Res Rev* 2009;**66**(6):611-38.
2. Tannen RL, Weiner MG, Marcus SM. Simulation of the Syst-Eur randomized control trial using a primary care electronic medical record was feasible. *J Clin Epidemiol* 2006;**59**(3):254-64.
3. Williams JG, Cheung WY, Cohen DR. Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment. *Health Technol Assess* 2003;**7**(26): iii, v-x, 1-117.
4. Yamamoto K, Matsumoto S, Tada H, et al. A data capture system for outcomes studies that integrates with electronic health records: development and potential uses. *J Med Syst* 2008;**32**(5):423-7.
5. Yamamoto K, Matsumoto S, Yanagihara K, et al. A data-capture system for post-marketing surveillance of drugs that integrates with hospital electronic health records. *Open Access J Clin Trials* 2011;**3**:21-6.
6. Embi PJ, Jain A, Clark J, et al. Effect of a clinical trial alert system on physician participation in trial recruitment. *Arch Intern Med* 2005;**165**(19):2272-7.

- 1
2
3
4 7. Campbell MK, Snowdon C, Francis D, et al. Recruitment to randomised trials: strategies
5
6 for trial enrollment and participation study. The STEPS study. *Health Technol Assess*
7
8 2007;**11**:iii, ix-105.
9
- 10
11
12
13 8. Dugas M, Lange M, Müller-Tidow C, et al. Routine data from hospital information
14
15 systems can support patient recruitment for clinical studies. *Clin Trials* 2010;**7**(2):183-9.
16
17
- 18
19
20 9. Thadani SR, Weng C, Bigger JT, et al. Electronic screening improves efficiency in clinical
21
22 trial recruitment. *J Am Med Inform Assoc* 2009;**16**(6):869-73.
23
24
- 25
26
27 10. Torgerson JS, Arlinger K, Käppi M, et al. Principles for enhanced recruitment of subjects
28
29 in a large clinical trial: The XENDOS study experience. *Control Clin Trials*
30
31 2001;**22**(5):515-25.
32
33
- 34
35
36 11. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and
37
38 definition for an emerging domain. *J Am Med Inform Assoc.* 2009;**16**(3):316-27
39
40
- 41
42
43 12. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic
44
45 medical record for clinical research. *Methods Inf Med.* 2009;**48**(1):38-44.
46
47
- 48
49
50 13. Wasserman RC. Electronic medical records (EMRs), epidemiology, and epistemology:
51
52 reflections on EMRs and future pediatric clinical research. *Acad Pediatr.* 2011;**11**(4):280-7.
53
54
- 55
56
57 14. Daniels K, R. *eClinical Trials: Planning and Implementation.* Boston, MA:
58
59
60

1
2
3
4 CenterWatch, 2003.
5
6

7
8 15. Kristianson KJ, Ljunggren H, Gustafsson LL. Data extraction from a semi-structured
9
10 electronic medical record system for outpatients: a model to facilitate the access and use of
11
12 data for quality control and research. *Health Inform J* 2009;**15**(4):305–319.
13
14

15
16
17 16. Inmon WH. *Building the Data Warehouse*, Wiley 2005.
18

19
20
21 17. Shim JP. Past, present, and future of decision support technology. *Decis Support Syst*
22
23 2002;**33**(2):111-26.
24
25

26
27
28 18. Prat N. A UML-based data warehouse design method. *Decis Support Syst*
29
30 2006;**42**(3):1449-73.
31
32

33
34
35 19. Park YT. An empirical investigation of the effects of data warehousing on decision
36
37 performance. *Inform Manag* 2006;**43**(1):51.
38
39

40
41 20. Schlaps D, Schmid T. Data warehousing in clinical research and development - From
42
43 clinical data to knowledge portals. *Pharmind* 2004;**66**(5a):637-46.
44
45

46
47
48 21. Grant A, Moshyk A, Diab H, et al. Integrating feedback from a clinical data warehouse
49
50 into practice organisation. *Int J Med Inform* 2006;**75**(3-4):232-9.
51
52

53
54
55 22. Junttila K, Meretoja R, Seppälä A, et al. Data warehouse approach to nursing
56
57 management. *J Nurs Manag* 2007;**15**(2):155-61.
58
59
60

- 1
2
3
4 23. Rubin DL, Desser TS. A data warehouse for integrating radiologic and pathologic data. J
5
6
7 Am Coll Radiol 2008;**5**(3):210-7.
8
9
10
11 24. Johnson SB. Generic data modeling for clinical repositories. J Am Med Inform Assoc
12
13 1996;**3**(5):328-39.
14
15
16
17 25. Nadkarni PM, Brandt C. Data extraction and ad hoc query of an entity—attribute—value
18
19 database. J Am Med Inform Assoc 1998;**5**:511-27.
20
21
22
23
24 26. Anhøj J. Generic design of Web-based clinical databases. J Med Internet Res
25
26 2003;**5**(4):e27.
27
28
29
30
31 27. Chen RS, Nadkarni P, Marenco L, et al. Exploring performance issues for a clinical
32
33 database organized using an entity-attribute-value representation. J Am Med Inform Assoc
34
35 2000;**7**:475-87.
36
37
38
39
40 28. Dinu V, Nadkarni P. Guidelines for the effective use of entity-attribute-value modeling for
41
42 biomedical databases. Int J Med Inform 2007;**76**:769-79.
43
44
45
46
47 29. Corwin J, Silberschatz A, Miller PL, et al. Dynamic tables: an architecture for managing
48
49 evolving, heterogeneous biomedical data in relational database management systems. J Am
50
51 Med Inform Assoc 2007;**14**:86-93.
52
53
54
55
56
57 30. Wade TD, Hum RC, Murphy JR. A dimensional bus model for integrating clinical and
58
59
60

1
2
3
4 research data. *J Am Med Inform Assoc* 2011;**1**:96-102.

5
6
7
8 31. Pardillo J, Mazón JN. Model-driven development of OLAP metadata for relational data
9
10 warehouses. *Comput Stand Interfac* 2012;**34**(1):189-202.

11
12
13
14 32. Hettler M. Data mining goes multidimensional. *Healthc Inform.* 1997;**14**(3):43-6, 48, 51-6.

15
16
17
18 33. Gordon BD, Asplin BR. Using online analytical processing to manage emergency
19
20 department operations. *Acad Emerg Med* 2004;**11**(11):1206-12.

21
22
23
24 34. Kimball R, Reeves L, Ross M, et al. *The Data Warehouse Lifecycle Toolkit*. New
25
26
27
28
29 York: John Wiley, 1998.

30
31
32
33 35. Weng C, Tu SW, Sim I, et al. Formal representation of eligibility criteria: A literature
34
35
36 review. *J Biomed Inform* 2010;**43**(3):451-67.

37
38
39
40 36. Lonsdale DW, Tustison C, Parker CG, et al. Assessing clinical trial eligibility with logic
41
42
43 expression queries. *Data Knowl Eng* 2008;**66**(1):3-17.

44
45
46
47 37. Tu SW, Peleg M, Carini S, et al. A practical method for transforming free-text eligibility
48
49
50 criteria into computable criteria. *J Biomed Inform* 2011;**44**(2):239-50.

51
52
53
54 38. Sordo M, Boxwala AA, Ogunyemi O, et al. Description and status update on GELLO: a
55
56
57
58
59
60 proposed standardized object-oriented expression language for clinical decision support. *Stud*

1
2
3
4 Health Technol Inform. 2004;**107**:164-8.
5
6

7
8 39. Séroussi B, Bouaud J. Using OncoDoc as a computer-based eligibility screening system
9
10 to improve accrual onto breast cancer clinical trials. Artif Intell Med 2003;**29**(1-2):153-67.
11
12

13
14 40. CDISC ASPIRE: Integration of clinical research and EHR: Eligibility coding standards.
15
16

17 http://crisummit2010.amia.org/files/symposium2008/S14_Niland.pdf (accessed 31 Mar
18
19 2012).
20
21

22
23
24 41. CDISC Study Design Model: SDM-XML Version 1.0.
25
26

27 [http://www.cdisc.org/stuff/contentmgr/files/0/8c85b168e80d6834ded59339b55fdb7/misc/cd
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60](http://www.cdisc.org/stuff/contentmgr/files/0/8c85b168e80d6834ded59339b55fdb7/misc/cdisc_sdm_xml_1.0.pdf)
[isc_sdm_xml_1.0.pdf](http://www.cdisc.org/stuff/contentmgr/files/0/8c85b168e80d6834ded59339b55fdb7/misc/cdisc_sdm_xml_1.0.pdf) (accessed 31 Mar 2012).

42. US. National Cancer Institute (NCI). caMATCH.

<https://cabig.nci.nih.gov/community/tools/caMATCH> (accessed 31 Mar 2012).

43. Jagannathan V, Mullett CJ, Arbogast JG, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. Int J Med Inform 2009;**78**(4):284-91.

44. Pakhomov S, Weston SA, Jacobsen SJ, et al. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care 2007;**13**(6):281-8.

- 1
2
3
4 45. Yamamoto K, Yamanaka K, Hatano E, et al. An eClinical trial system for cancer that
5
6 integrates with clinical pathways and electronic medical records. Clin trials (In press accepted
7
8 27-Mar-2012) .
9
10
11
12
13 46. Matsumura Y, Kuwata S, Yamamoto Y, et al. Template-based data entry for general
14
15 description in medical records and data transfer to data warehouse for analysis. Stud Health
16
17 Technol Inform 2007;**129**(Pt 1):412–416.
18
19
20
21
22
23 47. Henry SB, Douglas K, Galzagorry G, et al. A template-based approach to support
24
25 utilization of clinical practice guidelines within an electronic health record. J Am Med Inform
26
27 Assoc. 1998;**5**(3):237-44.
28
29
30
31
32
33 48. Los RK, van Ginneken AM, van der Lei J. OpenSDE: a strategy for expressive and
34
35 flexible structured data entry. Int J Med Inform. 2005;**74**(6):481-90.
36
37
38
39
40 49. Rose EA, Deshikachar AM, Schwartz KL, et al. Use of a template to improve
41
42 documentation and coding. Fam Med. 2001;**33**(7):516-21.
43
44
45
46 50. Chen R, Enberg G, Klein GO. Julius--a template based supplementary electronic health
47
48 record system. BMC Med Inform Decis Mak. 2007 **2**;7:10.
49
50
51
52
53 51. C.J.Date. An Introduction to Database Systems 8th ed. Boston; Addison Wesley. 2003
54
55
56
57 52. W. H. Inmon, Claudia Imhoff, Ryan Sousa. Corporate Information Factory. Hoboken,
58
59
60

1
2
3
4 New Jersey, John Wiley & Sons 2ed; 2000
5
6

7
8 53. Fellows JL, Rindal DB, Barasch A, et al. ONJ in two dental practice-based research
9
10 network regions. J Dent Res 2011;**90**(4):433-8.
11

12
13
14 54. Vestergaard P, Schwartz K, Rejnmark L, et al. Oral bisphosphonate use increases the risk
15
16 for inflammatory jaw disease: a cohort study. J Oral Maxillofac Surg. [Published Online First
17
18 2011 Jul 15].
19

20
21
22 55. Pathak J, Wang J, Kashyap S, et al. Mapping clinical phenotype data elements to
23
24 standardized metadata repositories and controlled terminologies: the eMERGE Network
25
26 experience. J Am Med Inform Assoc. 2011;**18**(4):376-86.
27
28

29
30
31 56. World Health Organization (WHO). International Classification of Diseases (ICD).
32
33

34
35
36 <http://www.who.int/classifications/icd/en/> (accessed 31 Mar 2012).
37
38

39
40 57. Shortliffe EH, Cimino JJ. Biomedical Informatics: Computer Applications in Health Care
41
42 and Biomedicine. Health Informatics series. Springer, 2006.
43
44

45
46
47 58. Carrier A, Levasseur M, Bédard D, et al. Community occupational therapists' clinical
48
49 reasoning: identifying tacit knowledge. Aust Occup Ther J 2010;**57**(6):356-65.
50
51

52
53
54 59. Thornton T. Clinical judgement, expertise and skilled coping. J Eval Clin Pract
55
56 2010;**16**(2):284-91.
57
58
59
60

- 1
2
3
4 60. Loughlin M. Epistemology, biology and mysticism: comments on 'Polanyi's tacit
5
6
7 knowledge and the relevance of epistemology to clinical medicine'. J Eval Clin Pract
8
9
10 2010;**16**(2):298-300.
11
12
13 61. Kienle GS, Kiene H. Clinical judgement and the medical profession. J Eval Clin Pract
14
15
16 2011;**17**(4):621-7.
17
18
19
20 62. Ting SL. Knowledge elicitation approach in enhancing tacit knowledge sharing. Ind
21
22
23 Manag Data Syst 2011;**111**(7):1039-64.
24
25
26
27 63. Fugill M. Tacit knowledge in dental clinical teaching. Eur J Dent Educ 2012;**16**(1):2-5.
28
29
30
31 64. McFadden E. Management of data in clinical trials. 2nd ed. Hoboken, NJ:
32
33
34 Wiley-Interscience, 2007.
35
36
37
38 65. Zhengwu L, Jing S. Clinical data management: current status, challenges, and future
39
40
41 directions from industry perspectives. Open Access J Clin Trials 2010;**2**: 93–105.
42
43
44
45 66. Data Management Association (DAMA) International. Data management body of
46
47
48 knowledge. <http://www.dama.org/i4a/pages/index.cfm?pageid=3364> (accessed 31 Mar 2012).
49
50
51
52 67. MedDRA MSSO. MedDRA MSSO. <http://www.meddramsso.com/> (accessed 31 Mar
53
54
55 2012).
56
57
58 68. US National Library of Medicine. SNOMED clinical terms,
59
60

1
2
3
4 http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html (accessed 31 Mar 2012).
5
6

7
8 69. Clinical Data Interchange Standards Consortium (CDISC). Clinical data acquisition
9
10 standards harmonization (CDASH). <http://www.cdisc.org/cdash> (accessed 31 Mar 2012).
11
12

13
14 70. Clinical Data Interchange Standards Consortium (CDISC). Study data tabulation model
15
16 (SDTM). <http://www.cdisc.org/sdtm> (accessed 31 Mar 2012).
17
18

19
20
21 71. Huff S. Development of the logical observation identifier names and codes (LOINC)
22
23 vocabulary. J Am Med Inform Assoc 1998;5(3):276-92.
24
25

26
27
28 72. Health level seven (HL7). <http://www.hl7.com/> (accessed 31 Mar 2012).
29
30

31
32 73. Integrating the healthcare enterprise (IHE). <http://www.ihe.net/> (accessed 31 Mar 2012).
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

FIGURE CAPTIONS

Figure 1. Data model for our EMR retrieval system.

Figure 2. Example SQL to create the target patient list.

Figure 3. Example SQL to flag the target patient report for investigator confirmation.

Figure 4. Sample data were flagged on the target patient report for investigator confirmation.

The left table is extracted from logical queries, which were defined to add patients as '1' on the list if a correspondence was observed or as '0' if no correspondence was observed. The table was pivoted on a cross-tabulation list using statistical software.

Figure 5. Dataset image set aside for investigator confirmation. This is a sample of the list of radiological findings.

Figure 6. Example of a star schema for clinical information.

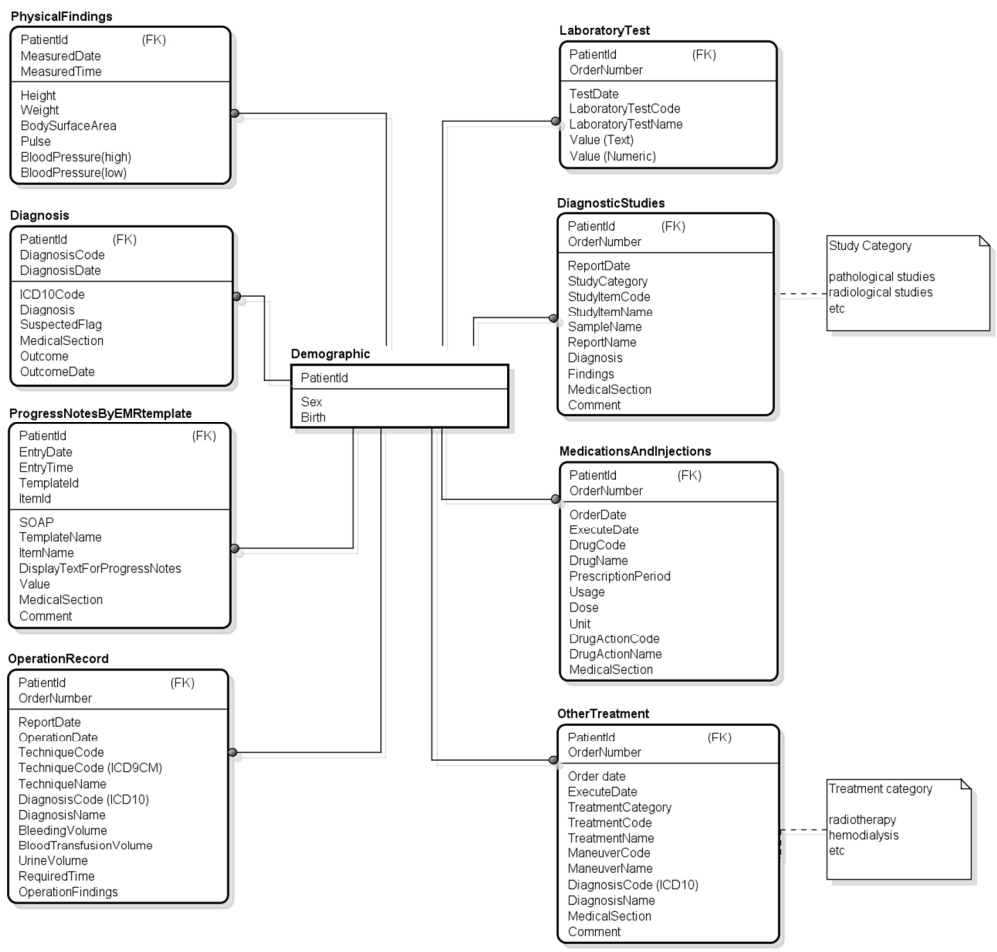


Figure 1. Data model for our EMR retrieval system.
109x104mm (300 x 300 DPI)

only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8 Create View_PatientsList as
9 Select PatientId From Demographic

10 Where

11 a. PatientId in (

12 Select PatientId From Diagnosis
13 Where ICD10Code in (osteoporosis ICD10 code list) and
14 DiagnosisDate >= '10/01/2000' and DiagnosisDate <= '09/30/2010' and
15 SuspectedFlag = 'Fixed')
16

17 and

18 a. PatientId in (

19 Select PatientId From MedicationsAndInjections
20 Where DrugCode in (osteoporosis drugs code list) and
21 ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010')
22

23 and

24 a. PatientId not in (

25 Select PatientId From MedicationsAndInjections
26 Where DrugCode in (intravenous BP drug code list) and
27 ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010')
28
29
30

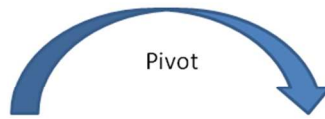
31
32 Figure 2. Example SQL to create the target patient list.
33 81x60mm (300 x 300 DPI)
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

```

Select PatientId, Oral BP administrations, 1 From View_PatientsList a
Where a. PatientId in (
    Select PatientId From MedicationsAndInjections
    Where DrugCode in (oral BP drugs code list) and
    ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010 ' )
Union all
Select PatientId, Oral BP administrations, 0 From View_PatientsList a
Where a. PatientId not in (
    Select PatientId From MedicationsAndInjections
    Where DrugCode in (oral BP drugs code list) and
    ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010')
Union all
Select PatientId, Inflammatory jaw condition diagnosis, 1 From View_PatientsList a
Where a. PatientId in (
    Select PatientId From Diagnosis
    Where ICD10Code in (inflammatory conditions of jaws ICD10 code list) and
    DiagnosisDate >= '10/01/2000' and DiagnosisDate <= '09/30/2010' and SuspectedFlag = 'Fixed')
Union all
Select PatientId, Inflammatory jaw condition diagnosis, 0 From View_PatientsList a
Where a. PatientId not in (
    Select PatientId From Diagnosis
    Where ICD10Code in (inflammatory conditions of jaws ICD10 code list) and
    DiagnosisDate >= '10/01/2000' and DiagnosisDate <= '09/30/2010' and SuspectedFlag = 'Fixed')
    
```

Figure 3. Example SQL to flag the target patient report for investigator confirmation.
81x60mm (300 x 300 DPI)

Review only



Patient Id	Criteria	Value
1	Oral BP administrations	0
1	Inflammatory jaw condition diagnosis	0
2	Oral BP administrations	1
2	Inflammatory jaw condition diagnosis	1
3	Oral BP administrations	0
3	Inflammatory jaw condition diagnosis	0
4	Oral BP administrations	1
4	Inflammatory jaw condition diagnosis	0
5	Oral BP administrations	0
5	Inflammatory jaw condition diagnosis	1
6	Oral BP administrations	0
6	Inflammatory jaw condition diagnosis	1
7	Oral BP administrations	0
7	Inflammatory jaw condition diagnosis	0
8	Oral BP administrations	1
8	Inflammatory jaw condition diagnosis	1

Patient Id	Oral BP administrations	Inflammatory jaw condition diagnosis
1	0	0
2	1	1
3	0	0
4	1	0
5	0	1
6	0	1
7	0	0
8	1	1

Figure 4. Sample data were flagged on the target patient report for investigator confirmation. The left table is extracted from logical queries, which were defined to add patients as '1' on the list if a correspondence was observed or as '0' if no correspondence was observed. The table was pivoted on a cross-tabulation list using statistical software.
79x41mm (300 x 300 DPI)

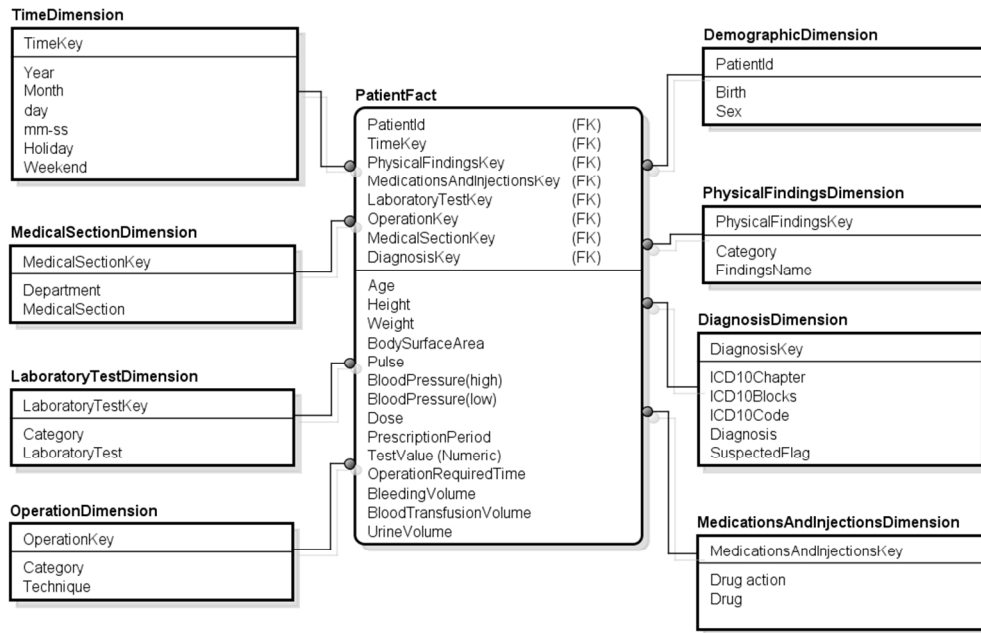
患者ID	オーダー種類(名称)	レポート名称	診断	所見	報告日	依頼コメント
00000010	放射線	頭部検査(MR)	no interval of change since 2005/01/05.	このレポートは、取り消されています。 同側基底核、放線冠に多発性陈旧性ラクナ梗塞を認めます。大脳白質内にびまん性にT2強調画像で高信号領域が認められ、正常加齢性変化と考えます。他に特記すべき異常所見を認めませんでした。副ビクワ 同側基底核、放線冠に多発性陈旧性ラクナ梗塞を認めます。大脳白質内にびまん性にT2強調画像で高信号領域が認められ、正常加齢性変化と考えます。他に特記すべき異常所見を認めませんでした。副ビクワ 同側基底核、放線冠に多発性陈旧性ラクナ梗塞を認めます。大脳白質内にびまん性にT2強調画像で高信号領域が認められ、正常加齢性変化と考えます。他に特記すべき異常所見を認めませんでした。副ビクワ	2005/05/27	5月27日：放射線レポート
00000010	放射線	骨(全身)	s/oNo evidence of bone metastasis/ oTrauma/c/wNo interval change since previous study	このレポートは、取り消されています。 造影CT検査時には患者様の同意書を取得して頂くことになっておりますが、今回の検査では同意書取得が確認出来ませんでしたので、やむを得ず単純CTのみを撮影させて頂きました。あしからずご了承下さい。	2005/05/27	5月27日：放射線レポート本番環境テスト
00000010	放射線	頭部検査(MR)	no interval of change since 2005/01/05.	このレポートは、取り消されています。 同側基底核、放線冠に多発性陈旧性ラクナ梗塞を認めます。大脳白質内にびまん性にT2強調画像で高信号領域が認められ、正常加齢性変化と考えます。他に特記すべき異常所見を認めませんでした。副ビクワ 同側基底核、放線冠に多発性陈旧性ラクナ梗塞を認めます。大脳白質内にびまん性にT2強調画像で高信号領域が認められ、正常加齢性変化と考えます。他に特記すべき異常所見を認めませんでした。副ビクワ 同側基底核、放線冠に多発性陈旧性ラクナ梗塞を認めます。大脳白質内にびまん性にT2強調画像で高信号領域が認められ、正常加齢性変化と考えます。他に特記すべき異常所見を認めませんでした。副ビクワ	2005/05/27	5月27日：放射線レポート本番環境テスト

Patient id
 Study category
 Report name
 Diagnosis
 Findings
 Comment

Figure 5. Dataset image set aside for investigator confirmation. This is a sample of the list of radiological findings.

81x60mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



Example of a star schema for clinical information.
90x58mm (300 x 300 DPI)

Review only



A pragmatic method for electronic medical-record-based observational studies: developing an electronic medical records retrieval system for clinical research

Journal:	<i>BMJ Open</i>
Manuscript ID:	bmjopen-2012-001622.R1
Article Type:	Research
Date Submitted by the Author:	14-Sep-2012
Complete List of Authors:	Yamamoto, Keiichi; Kyoto University Hospital, Department of Clinical Trial Design and Management, Translational Research Center Sumi, Eriko; Kyoto University Hospital, Department of Clinical Innovative Medicine, Translational Research Center Yamazaki, Toru; Kyoto University, Department of Oral and Maxillofacial Surgery, Graduate School of Medicine Asai, Keita; Kyoto University, Department of Oral and Maxillofacial Surgery, Graduate School of Medicine Yamori, Masashi; Kyoto University, Department of Oral and Maxillofacial Surgery, Graduate School of Medicine Teramukai, Satoshi; Kyoto University Hospital, Department of Clinical Trial Design and Management, Translational Research Center Bessho, Kazuhisa; Kyoto University, Department of Oral and Maxillofacial Surgery, Graduate School of Medicine Yokode, Masayuki; Kyoto University Hospital, Department of Clinical Innovative Medicine, Translational Research Center Fukushima, Masanori; Foundation for Biomedical Research and Innovation, Translational Research Informatics Center
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Public health, Research methods, Epidemiology
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, ORAL & MAXILLOFACIAL SURGERY, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS, Clinical trials < THERAPEUTICS

SCHOLARONE™
Manuscripts

1
2
3
4 **A pragmatic method for electronic medical-record-based observational studies:**
5
6
7 **developing an electronic medical records retrieval system for clinical research**
8
9

10 Keiichi Yamamoto¹, Eriko Sumi², Toru Yamazaki³, Keita Asai³, Masashi Yamori³, Satoshi
11 Teramukai¹, Kazuhisa Bessho³, Masayuki Yokode², Masanori Fukushima⁴
12
13
14
15

16
17 ¹Department of Clinical Trial Design and Management, Translational Research Centre, Kyoto
18 University Hospital, Kyoto, Japan
19
20
21
22

23
24 ²Department of Clinical Innovative Medicine, Translational Research Centre, Kyoto
25 University Hospital, Kyoto, Japan
26
27
28

29
30
31 ³Department of Oral and Maxillofacial Surgery, Graduate School of Medicine, Kyoto
32 University, Kyoto, Japan
33
34
35

36
37 ⁴Translational Research Informatics Centre, Foundation for Biomedical Research and
38 Innovation, Kobe, Japan
39
40
41
42

43
44 Corresponding author: Keiichi Yamamoto, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto,
45 606-8507 Japan. E-mail: kyamamo@kuhp.kyoto-u.ac.jp, Tel: +81-75-751-4717, Fax
46 +81-75-751-3399
47
48
49
50
51

52
53
54 Total number of words: 3,997
55

56
57
58 Short title: A pragmatic method for EMR-based observational studies
59
60

1
2
3
4 Keywords: clinical research informatics, data warehouse, OLAP, computable eligibility
5
6
7 criteria, pharmacoepidemiology, hospital-based cohort study
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

ABSTRACT

Objective: The use of electronic medical record (EMR) data is necessary to improve clinical research efficiency. However, it is not easy to identify patients who meet research eligibility criteria and collect the necessary information from EMRs because the data collection process must integrate various techniques, including the development of a data warehouse and translation of eligibility criteria into computable criteria. This research aimed to demonstrate an electronic medical records retrieval system (ERS) and an example of a hospital-based cohort study that identified both patients and exposure with an ERS. We also evaluated the feasibility and usefulness of the method. **Design:** The system was developed and evaluated. **Participants:** In total, 800,000 cases of clinical information stored in EMRs at our hospital were used. **Primary and secondary outcome measures:** The feasibility and usefulness of the ERS, the method to convert text from eligible criteria to computable criteria, and a confirmation method to increase research data accuracy. **Results:** To comprehensively and efficiently collect information from patients participating in clinical research, we developed an ERS. To create the ERS database, we designed a multi-dimensional data model optimised for patient identification. We also devised practical methods to translate narrative eligibility criteria into computable parameters. We applied the system to an actual hospital-based cohort study performed at our hospital and converted the test results into computable criteria. Based on this information, we identified eligible patients and extracted data necessary for

1
2
3
4 confirmation by our investigators and for statistical analyses with our ERS. **Conclusion:** We
5
6
7 propose a pragmatic methodology to identify patients from EMRs who meet clinical research
8
9
10 eligibility criteria. Our ERS allowed for the efficient collection of information on the
11
12
13 eligibility of a given patient, reduced the labour required from the investigators, and
14
15
16 improved the reliability of the results.

17 18 19 **ARTICLE SUMMARY**

20 21 22 **Article focus**

23
24
25
26
27 The focus of this work was to establish a pragmatic methodology to efficiently collect
28
29
30 information from EMRs about patients who meet clinical research eligibility criteria.
31
32

33 34 **Key messages**

35
36
37 The use of electronic medical record (EMR) data is necessary to improve clinical research
38
39
40 efficiency. However, it is not easy to identify patients who meet research eligibility criteria
41
42
43 and collect necessary data from EMRs because the data collection process must integrate
44
45
46 various techniques, including the development of a data warehouse and the translation of
47
48
49 eligibility criteria into computable criteria. An efficient ERS and a standardised data
50
51
52 processing model that integrates these techniques are essential to facilitate clinical research
53
54
55 that utilises EMRs.
56
57
58
59
60

Strengths and limitations of this study

- Our method uses a specialised data model for patient identification in clinical research and efficient data conversion that does not depend on the EMR database structure when converting narrative criteria to computable criteria.
- We propose that computable criteria should not be a result of the automated conversion of narrative criteria but rather a result of research preparation involving medical concepts that are not expressed logically or explicitly in the narrative criteria. Therefore a large amount of the conversion of the eligibility criteria to computable criteria should be executed at the protocol development stage.
- It is important to further discuss protocol standardisation, including eligibility criteria representation for computable use.
- Enabling ERS use in and across multiple institutions is an important future task.

BACKGROUND

Medical information technology has recently advanced in many countries, and enormous amounts of clinical data are already stored as electronic medical records (EMRs). Utilising the data collected in EMRs is necessary to improve clinical research efficiency [1-3]. An EMR is a large database of patient data and is used in observational research to investigate the relationships among diseases, treatments, and outcomes [4-7], to conduct surveillance for rare drug reactions [4, 8], and to recruit patients for clinical trials [9-13]. However, it is not easy to identify patients who meet research eligibility criteria and collect necessary information from EMRs [2-3]. Herein, we describe three major issues concerning EMR-based observational studies: EMR patient data retrieval function, eligibility criteria protocol representation, and EMR data accuracy.

To identify patients who meet research eligibility criteria, it is necessary to obtain various types of information stored in EMRs by subject, e.g., diagnosis and prescribed medications. However, the EMR database is designed to facilitate online transaction processing for rapid and detail-oriented clinical information searches on individual patients, and the current EMR system does not facilitate this retrieval function [2-3, 14]. Data warehouses are essential components of data-driven decision support. To allow for efficient research analyses, EMR data must first be warehoused to enable data analyses across patient populations [15-21]. However, health care data modelling is difficult and time-consuming because of the

1
2
3
4 complexity of the medical knowledge involved. Thus, the most common approaches to
5
6
7 clinical data warehouse modelling are variations on the entity-attribute-value (EAV) model
8
9
10 [22-28], where data are stored in a single table with three columns: entity identification,
11
12
13 attribute, and attribute value. The EAV design has advantages, including flexibility and ease
14
15
16 of storage; however, it requires transforming EAV data into another analytical format before
17
18
19 analysis [25, 28]. Online analytical processing (OLAP) is most frequently used for searching
20
21
22 data stored in the data warehouse [29-31]. OLAP systems in relational databases are typically
23
24
25 designed based on Kimball's star schema [32]. However, the star schema was devised to
26
27
28 facilitate online measurement analyses. In health care, this method can be used to
29
30
31 dynamically gather online analyses of numeric data (e.g., a specific dose of a drug for a
32
33
34 specific disease) in clinical practice. Therefore, this method is not suitable for identifying
35
36
37 patients who meet the complicated eligibility criteria for a given clinical research study.
38
39
40 Data-modelling methods that facilitate the identification of patients and enable the collection
41
42
43 of necessary information from EMRs remain to be established [28].

44
45
46 Current eligibility criteria are written in a text format that cannot be computationally
47
48
49 processed. Additionally, to be applied in actual EMR, eligible criteria need to be integrated
50
51
52 with the data model of EMRs [33]. Several investigations have sought to establish
53
54
55 computable eligibility criteria [34-41]. However, there is no consensus regarding a standard
56
57
58 patient information model [33], and the eligibility criteria are not yet completely standardised.
59
60

1
2
3
4 Using natural language processing (NLP) technologies to convert the text format of eligibility
5
6
7 criteria to a computer or to extract patient identifications from EMRs is far from perfect
8
9
10 without human intervention [3, 42-43].

11
12
13 Current EMRs have been used to support claims for medical service fees and the treatments
14
15 administered to each patient; therefore, data gathered specifically for research purposes may
16
17
18 be incomplete and unreliable [2-3, 44].
19
20
21

22
23 Although various investigations on each technique are executed individually, standardised
24
25 methods must still be established that integrate these techniques, facilitate the identification
26
27
28 of patients who are eligible for clinical research, and collect necessary information from
29
30
31 EMRs.
32
33

34 35 **OBJECTIVE**

36
37
38 We designed a pragmatic data processing model optimised for patient identification and for
39
40 the collection of necessary information from EMRs for clinical research. These tools are
41
42
43 implemented as an electronic medical records retrieval system (ERS) [44].
44
45
46
47

48
49 This research aimed to demonstrate an ERS and an example of a hospital-based cohort study
50
51 that used the ERS to identify both patients and exposure. Another aim was to evaluate the
52
53
54 feasibility and usefulness of the ERS, the method to convert text form eligible criteria to
55
56
57 computable criteria, and a confirmation method to increase research data accuracy.
58
59
60

MATERIALS AND METHODS

Outline of our procedure for patient identification and data collection from the EMR

To identify patients who met the eligibility criteria for the clinical research in question, data were collected in the following ways:

- 1) The text form of the narrative criteria was converted into computable criteria.
- 2) A targeted patient list was created.
- 3) A flag was added for investigators to confirm the targeted patient list.
- 4) Reports were created for the investigators to confirm.
- 5) After confirmation by the investigator, the statistical analyses were executed.

EMR retrieval system

In our hospital, EMR use was introduced in 2005; approximately 800,000 cases of clinical information have already been stored. To comprehensively and efficiently collect information about patients participating in clinical research, we developed an ERS [44].

EMRs store various types of information, integrating billing, pharmacy, radiology, laboratory information, and others [4]. In creating the ERS database, we designed a new data model based on the star schema that was optimised for patient identification in clinical research. We

1
2
3
4 identified nine data categories from EMRs that are useful for clinical research: demographic
5
6 characteristics, physical findings, diagnostic studies, laboratory tests, diagnoses, progress
7
8 reports on an EMR template [44-45], medications and injections, operation records, and other
9
10 treatments. We then designated these categories to 'entities'. In our hospital, the diagnosis is
11
12 managed by codes that were originally defined by our hospital and mapped with International
13
14 Statistical Classification of Diseases (ICD) 10 codes [46] for medical insurance purposes.
15
16 Operations codes were also managed by codes that originally were defined by our hospital
17
18 and mapped with ICD-9 Clinical Modification codes. We identified available columns (e.g.,
19
20 ICD code, diagnosis date) from the EMR data model and designated these columns as
21
22 'attributes' of the entities.
23
24
25
26
27
28
29
30
31
32

33
34 Figure 1 presents our data model. In our model, all entities in a given schema are independent
35
36 and complete; this allows for logical operations and for the creation of eligible patient lists
37
38 for each respective parameter in a study. The target patient list is generated by combining
39
40 these patient lists. The data model also supports the inference of medical concepts expressed
41
42 in the eligibility criteria in reference to corresponding patient data accumulated in EMRs
43
44
45
46
47
48 [33-34].
49
50

51
52 In our hospital, a replicate of the EMR database known as 'Open DB' was established for the
53
54 secondary use of accumulated EMR data [7]. A data mart for our ERS was created to ensure
55
56 that the data retrieval process was practical and independent of the EMR system structure; the
57
58
59
60

1
2
3
4 data mart was created on the relational database management system by extracting,
5
6 transforming, and loading (ETL) information from the Open DB [7, 44]. The ETL process is
7
8 performed automatically once nightly except for the 'Progress notes by EMR template' entity,
9
10 which is referred directly from the Open DB to ensure real-time visibility for the eClinical
11
12 trial [44].
13
14
15
16

17
18
19 An OLAP tool was installed to efficiently search through data from multiple patients [44].
20

21
22 The OLAP tool runs in an Internet browser and can generate structured query language
23
24 (SQL) based on predefined metadata (i.e., a data model) by defining logical queries (i.e.,
25
26 programs) using a graphical user interface (GUI). Moreover, this tool allows reports on
27
28 information retrieved from the browser to be transcribed using hypertext markup language
29
30 (HTML). The reports are created in various formats, including portable document format
31
32 (PDF), comma separated values (CSV), and extensible markup language (XML) [44].
33
34
35
36
37
38

39
40 To protect personal information in medical records at our hospital, the EMR network is
41
42 separated physically from other networks. Our data mart and OLAP servers are deployed in
43
44 the same EMR network and managed using the same EMR security policies. Additionally, the
45
46 use of our ERS is limited to clinical research approved by the ethics committee at our hospital,
47
48 and only designated staff members at our centre are allowed to retrieve data. Our centre
49
50 creates and manages ERS user identification separate from the EMRs. For the external output
51
52 of CSV and other data, permission must be obtained from our department of medical
53
54
55
56
57
58
59
60

informatics, and data extraction must be executed in the presence of supervisors who are responsible for protecting personal information at our hospital.

Application to clinical research

We applied the system to a hospital-based cohort study performed at our hospital titled 'Risk of osteomyelitis of the jaw induced by oral bisphosphonates (BP) in patients taking medications for osteoporosis: A hospital-based cohort study in Japan' [47], in which we identified eligible patients, extracted research data, and evaluated the feasibility of our system. The ethics committee at Kyoto University Hospital approved this research. A different paper details the purpose, methods, results, and discussion of this research [47].

This research aimed to estimate the risks for osteomyelitis of the jaw in osteoporosis patients at our hospital who had been exposed to oral BP compared to those who had not [48-49].

The eligibility criteria were as follows:

Inclusion criteria

- Patients diagnosed with osteoporosis and treated with osteoporosis medications at Kyoto University Hospital between November 2000 and October 2010.
- Patients aged 20 years or older.

Exclusion criteria

- Patients with a history of treatment with radiation therapy to the maxillofacial region.
- Patients with primary or metastatic tumours in the maxillofacial region.
- Patients treated with intravenous BP.

The data collected were diagnosis, date of diagnosis, sex, birthdate, and the doses and dates when osteoporosis medications, steroids, anticancer drugs, diabetes drugs and HbA1c tests were administered.

Conversion of the text form of the narrative criteria to computable criteria

To identify eligible patients and collect the necessary data from the EMRs, narrative criteria and data must be converted to computable criteria. Such computable criteria include entities, attributes, logical operators (i.e., 'and' and 'or'), codes, and parameters [33-37]. The clinical research purpose and clinical practice demands made it necessary to perform this task.

We manually executed the conversion from text eligibility criteria to computable criteria. As an example of the conversion from narrative criteria to computable criteria, we present the following two-step conversion procedure:

Step 1: Convert the narrative criteria into entity-level criteria.

Medical concepts expressed as narrative criteria are mapped onto entities in the data model and converted into entity-level criteria. This task is manually performed at the protocol

1
2
3
4 development stage of the study by the investigators. For each entity, a criterion is created to
5
6
7 extract patients who meet each condition. If exclusive conditions for the same entity must be
8
9
10 defined, a different criterion is created. Additionally, the list of codes for drugs and diagnoses
11
12 (i.e., ICD-10) is created, and the period of treatments and others are defined by investigators.
13
14
15 In this study, we mapped ‘osteoporotic patients’ onto two entities (i.e., ‘diagnosis’ and
16
17 ‘medications and injections’) and converted it to a combination of two criteria (i.e.,
18
19 ‘diagnosis of osteoporosis’ and ‘osteoporosis drug administration’). In the test research, we
20
21 defined the entity-level criteria according to the entered diagnosis and ordered treatments
22
23 rather than the diagnostic criteria of the disease. This process reflects that the test research
24
25 aimed to estimate some risks of osteomyelitis of the jaw with BP administration instead of
26
27 diagnosing osteoporosis patients accurately. The recorded diagnosis in the EMR was typically
28
29 designed to ensure payment for medical claims. We thus sought to reduce the number of false
30
31 positives by extracting patients with a given treatment type.
32
33
34
35
36
37
38
39
40
41

42 Step 2: Convert entity-level criteria into attribute-level criteria (i.e., computable criteria).
43
44

45
46 The abovementioned corresponding codes, date and parameters are mapped onto attributes of
47
48 the entity-level criteria, and these factors become computable criteria.
49
50
51

52 53 **Creating a targeted patient list** 54

55
56
57 A targeted patient list is created from the entire set of patients for whom EMRs have been
58
59
60

1
2
3
4 obtained by defining logical queries (i.e., programs defined by the GUI) based on the
5
6
7 computable criteria included in the ERS.
8
9

10 Logical queries are first defined in the ERS to identify patients who meet the conditions for
11
12
13 each criterion. The ERS automatically generates the SQL necessary for data extraction
14
15
16 according to the logical queries. Logical queries are then defined to include or exclude
17
18
19 eligible patients who meet each criterion for the demographic entity. The targeted patient list
20
21
22 is created by executing the logical query. Figure 2 presents an example of an SQL
23
24
25 automatically generated by the ERS.
26
27
28

29 We thus designed our data model to enable the creation of a targeted patient list by defining
30
31
32 the patients extracted from each criterion (i.e., 'in' or 'not in') as conditions for the
33
34
35 demographic entity that was the unique patient list for the entire hospital. If logical queries
36
37
38 are defined using our method, even if the eligibility criteria are complicated, it is not
39
40
41 necessary to dramatically change the SQL structure generated in the ERS.
42
43
44

45 **Flagging entries for investigators to confirm**

46
47

48 To improve research data accuracy, confirmation by the investigators is necessary. When
49
50
51 confirmation is required, additional information is linked.
52
53
54

55 For the targeted patient list, logical queries are defined to flag certain items according to the
56
57
58 investigators' interest. Necessary logical queries are first defined for each criterion. Logical
59
60

1
2
3
4 queries are then defined for addition to the patient list as '1' if the data correspond or '0' if
5
6 they do not. Data sets created by these operations are joined by 'union' and pivoted on a
7
8 cross-tabulation list using statistical analysis software. We show an example of an SQL
9
10 generated by the ERS in Figure 3.
11
12
13

14 15 16 **Create reports for investigators to confirm** 17

18
19
20 To help investigators confirm the targeted patient list, reports are created by linking the
21
22 findings for diagnostic imaging, pathological diagnosis, operations, and other findings.
23
24 Investigators confirm these entries using the reports and EMR information, including
25
26 progress notes and images. When the diagnosis history, medication, laboratory results,
27
28 progress notes, and other information are necessary, the same operation is executed for each
29
30 instance. For example, the list of radiological findings involves 'patient id', 'study category',
31
32 'report name', 'diagnosis', 'findings', and 'comment'. The reports may improve the
33
34 investigators' confirmation efficiency because they prevent the need to refer to the medical
35
36 records for each patient who needs confirmation.
37
38
39
40
41
42
43
44
45
46

47 **Confirmation by the investigator and execution of the statistical analyses.** 48

49
50
51 The investigators confirm the accumulated data and execute the statistical analysis. In this
52
53 test research, two oral and maxillofacial surgeons diagnosed cases by a chart review with an
54
55 observation of imaging findings [47].
56
57
58
59
60

Systemic evaluation

To evaluate our system, we collected information about the research period using the recall method. For the accuracy of the data collected by the ERS, we evaluated the results after they were confirmed by the investigator.

RESULTS

Computable criteria, datasets, and system evaluation

We present the computable criteria in Table 1. To increase data accuracy, we collected all of the exclusion criteria for the investigators to confirm. As Table 1 shows, we extracted information from EMRs. For investigator confirmation, we also reported all targeted patients using the following lists: osteoporosis drugs administered, oral BP administered, intravenous BP administered, diabetes drugs administered, anticancer drugs administered, steroid drugs administered, osteoporosis diagnoses, oral cancer diagnoses, patients diagnosed with inflammation of the jaw, patients diagnosed with other suspicious diseases, patients diagnosed with diabetes, HbA1c values, radiological findings, pathological findings, and radioisotope findings. These data were extracted from the ERS for statistical analyses, presented in CSV format, and analysed using statistics software.

Table 1. Computable criteria for our test research

Criterion	Entity	Operator symbol	Attribute	Operator symbol	Parameter
Created a targeted patient list					
Inclusion criteria: Osteoporosis diagnosis	Diagnosis	-	ICD10Code	In	(osteoporosis ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Inclusion criteria: Osteoporosis drug administrations	Medications and	-	DrugCode	in	(osteoporosis drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
	Injections	and	ExecuteDate	<=	'09/30/2010'
Added a flag for investigators to confirm the targeted patient list					
Exclusion criteria: Oral cancer diagnosis	Diagnosis	-	ICD10Code	in	(oral cancer ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Exclusion criteria: Intravenous BP administrations	Medications and	-	DrugCode	in	(intravenous BP drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
	Injections	and	ExecuteDate	<=	'09/30/2010'
Oral BP administrations	Medications and	-	DrugCode	in	(oral BP drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
	Injections	and	ExecuteDate	<=	'09/30/2010'
Inflammatory jaw condition diagnosis	Diagnosis	-	ICD10Code	in	(inflammatory conditions of jaws ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Other suspicious disease diagnosis	Diagnosis	-	ICD10Code	in	(other suspicious disease ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Diabetes diagnosis	Diagnosis	-	ICD10Code	in	(diabetes ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed

Steroid drug administrations	Medications	-	DrugCode	in	(steroid drugs code list)
	and	and	ExecuteDate	>=	'10/01/2000'
	Injections	and	ExecuteDate	<=	'09/30/2010'
Anticancer drug administrations	Medications	-	DrugCode	in	(anticancer drugs code list)
	and	and	ExecuteDate	>=	'10/01/2000'
	Injections	and	ExecuteDate	<=	'09/30/2010'
Diabetes drug administrations	Medications	-	DrugCode	in	(diabetes drugs code list)
	and	and	ExecuteDate	>=	'10/01/2000'
	Injections	and	ExecuteDate	<=	'09/30/2010'
HbA1c test execution	Laboratory Test	-	LaboratoryTestCode	in	(HbA1c test code)
		and	TestDate	>=	'10/01/2000'
		and	TestDate	<=	'09/30/2010'
Created reports for confirmation by the investigators					
Radiological finding reports	Diagnostic Studies	-	ReportName	in	(report name list of oral region)
Pathologic finding reports	Diagnostic Studies	-	SampleName	contains	'bone'
		or	SampleName	contains	'jaw'
Radio isotope finding reports	Diagnostic Studies	-	-	-	-

BP, bisphosphonates; ICD, International Classification of Diseases; ID, identifications

Among the approximately 800,000 cases at our hospital, 8,772 were categorised using the terms 'Inclusion criteria: Osteoporosis diagnosis'; among this group, 7,195 were further categorised using 'Inclusion criteria: Osteoporosis drug administration'. We then calculated the time that had elapsed since the osteoporosis diagnosis, determined that 7,062 patients were aged 20 years or older, and created a targeted patient list. Among those on the targeted patient list, 23 patients were placed under the heading 'Exclusion criteria: Oral cancer diagnosis', 110 under 'Exclusion criteria: Intravenous BP administration', 4,200 under 'Oral

1
2
3
4 BP administration', 84 under 'Inflammatory jaw condition diagnosis', 2,064 as 'Other
5
6 suspicious disease diagnosis', 1,700 as 'Diabetes diagnosis', 4,551 as 'Steroid drug
7
8 administration', 904 as 'Anticancer drug administrations', 1,055 as 'Diabetes drug
9
10 administrations', and 3,641 as 'HbA1c test execution'. Because of the end point considered,
11
12 patients who were classified under 'Inflammatory jaw condition diagnosis' or 'Other
13
14 suspicious disease diagnosis' were confirmed using predefined hierarchical diagnostic criteria
15
16 by investigators who performed the statistical analyses and arranged the research results. We
17
18 show the schema of data collection and confirmation as Figure 4 [47].
19
20
21
22
23
24
25
26
27

28 The accuracy of the data extracted by the ERS was then characterised. Reviewing the medical
29
30 records revealed that 2,817 patients were not labelled as 'Oral BP administration', including 7
31
32 (1 who received intravenous BP) treated at other hospitals. 6 patients had been treated with
33
34 radiation therapy to the oral and maxillofacial regions. Among the 72 patients classified under
35
36 'Inflammatory jaw condition diagnosis', 35 cases and 37 non-cases were identified.
37
38
39
40
41
42

43 The data extraction period lasted approximately three months. Ten meetings were held during
44
45 the protocol development stage to create and validate the computable criteria and the list of
46
47 codes for various drugs and diagnoses (i.e., ICD-10). The time required for logical query
48
49 definition when using the ERS was approximately 20 hours. The investigator confirmations
50
51 and statistical analyses took approximately four months.
52
53
54
55
56
57
58
59
60

DISCUSSION

We identified eligible patients for this research and extracted the data necessary for confirmation by investigators and for statistical analyses.

We asked the chart reviewers to evaluate the system in a questionnaire about ‘the effect of computer programming support for data retrieval from the EMR’, ‘the result of the data retrieval’, ‘the positive and negative aspects of our ERS use’, and ‘the aspects of our method that should be improved’. The investigators evaluating the system mentioned that the following points: 1) the method enabled them to extract the necessary data for diagnosis and drug administration without exception; 2) by screening the entire patient population at the hospital using the ERS, they could identify not just eligible patients in the department of oral and maxillofacial surgery but all eligible patients, which reduced the study bias; and 3) by creating reports for confirmation, it enabled investigators to devote their time to reading images, thus effectively reducing the time required for reviewing medical records. The aspects of our method that should be improved are the ‘lack of claim data’ and the ‘administrative complexity of EMR data use’. No negative aspects of our ERS use were noted.

The ERS allowed for the collection of information on patient eligibility by efficiently combining clinical information. Although we did not compare our method with other

1
2
3
4 methods, our proposed method reduced the labour normally required from investigators and
5
6 improved the reliability of test research results, which indicated that it was useful.
7
8
9

10 To design the ERS database, we designed a new data model optimised for patient
11
12 identification. The main differences between our data model and the star schema were as
13
14 follows: 1) demographic data, which were presented in list form in our EMR system, were
15
16 presented as a fact-less fact table, and 2) date, time, measurements and text information were
17
18 presented in dimension tables [32]. The most significant characteristic of our method for
19
20 patient identification is the use of a specialised data model in clinical research and the ability
21
22 to execute a large number of conversion tasks at the protocol development stage. Data can be
23
24 converted efficiently in a way that does not depend on the EMR database structure when
25
26 converting narrative criteria to computable criteria. In this research, we considered whether
27
28 data were extracted directly from EMRs at the protocol development stage. However, EMR
29
30 data were recorded in a sequential format for every medical practice, and the database
31
32 structure was complicated. Comprehending the location and meaning of the necessary data
33
34 thus required tremendous effort. It was difficult to make precise logical queries for patient
35
36 identification. However, because our ERS data model was arranged by subjects (e.g., tests,
37
38 diagnosis), it was easy to interpret the available information. Due to the standardisation of
39
40 computable criteria and SQL possible with the ERS, it was also possible to create computable
41
42 criteria in little time. Additionally, verifying the patient identification accuracy was easy
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 because it was possible to test each individual criterion.
5
6

7
8 The SQL generated by our ERS does not reduce the time required for data retrieval. Our ERS
9
10 also cannot retrieve information that is not in the data model. Current EMRs do not store all
11
12 necessary data for clinical research, including information related to pregnancy, performance
13
14 status, cancer stage, availability of transportation to the hospital, specific tests that are not
15
16 typically performed, drug regimen, outcomes (including death), and adverse events.
17
18 Additionally, all tests are not administered to all patients, and necessary information may
19
20 have been recorded in medical records at another hospital [44]. To facilitate EMR use in
21
22 clinical research, it is necessary to accumulate as much of this information as possible. In the
23
24 hospital, much of this information does not integrate well with EMRs, including test reports
25
26 stored only in the departmental system [50]. However, it is important to utilise this
27
28 information. Additionally, enabling ERS use in and across multiple institutions is also an
29
30 important future task.
31
32
33
34
35
36
37
38
39
40
41
42

43
44 Currently, most clinical research studies that use data from EMRs are planned according to
45
46 the concept that the primary use of EMRs is for clinical practice and a secondary use is for
47
48 clinical research [44]. Therefore, most investigators attempt to convert the text form
49
50 eligibility criteria that already have been defined on a protocol to computable criteria at the
51
52 data collecting stage [35. 36]. However, we propose that computable criteria should not be a
53
54 result of the automated conversion of narrative criteria but rather a result of research
55
56
57
58
59
60

1
2
3
4 preparation involving medical concepts that are not expressed logically or explicitly in the
5
6
7 narrative criteria. Some medical concepts may be interpreted differently depending on the
8
9
10 research and the investigator caring for the patients. Additionally, current eligibility criteria
11
12
13 are vague or complex, and they do not consider the use of the actual EMR. To convert
14
15
16 computable criteria appropriately, high-level medical decisions to answer the research
17
18
19 question are required. Therefore, we thought that a large amount of the conversion of the
20
21
22 eligibility criteria to computable criteria should be executed at the protocol development
23
24
25 stage. In addition, the conversion process should be divided into entity-level conversions that
26
27
28 require higher medical decisions and attribute-level conversions. To reduce the burden of
29
30
31 conversion, it may be useful to apply NLP technology for the conversion from entity-level
32
33
34 criteria to attribute-level criteria. Moreover, it is important to further discuss protocol
35
36
37 standardisation, including eligibility criteria representation for computable use. For instance,
38
39
40 the attribute-level criteria that describe the search conditions in detail may be useful in global
41
42
43 studies to address diseases that vary according to the diagnostic criteria used in each country.

44
45 Concerning EMR data accuracy, the ICD10 code (osteomyelitis of the jaw) sensitivity was
46
47
48 48.6% (35/72). The investigators reported 6 simple diagnosis errors, 7 oral BP
49
50
51 administrations at other hospitals, and 6 patients who were treated with radiation therapy in
52
53
54 the oral and maxillofacial region [47]. For the accuracy of current EMRs, the investigators
55
56
57 had to confirm the information. However, the EMRs provided rich confirmation data and
58
59
60

1
2
3
4 were useful in improving research data accuracy. In this study, we checked the data from
5
6
7 actual EMRs manually and identified patients precisely and extensively using coded
8
9
10 information, narrative information, and images. However, only information from existing
11
12 EMRs was available. Current EMRs have a high degree of flexibility in data entry and are not
13
14 currently managed for research purposes, which decreases their reliability. It is necessary to
15
16 improve data quality through quality control without placing too much of a burden on clinical
17
18 practice. Alternatively, it may be possible to organise data sufficiently before research use
19
20 [51-53]. Standardising the terminology and exchange formats used in the healthcare setting
21
22 has facilitated international discourse [46, 54-58]. It is necessary to further discuss not only
23
24 clinical practice but also research purposes, particularly how to utilise various standards when
25
26 using EMRs beyond the hospital setting.
27
28
29
30
31
32
33
34
35

36 **CONCLUSION**

37
38
39
40 We propose a pragmatic method for EMR-based observational studies. Our ERS is already
41
42 used to support hospital-based cohort studies, clinical trial recruitment, and the eClinical trial
43
44 infrastructure [44] at our centre. We believe an efficient ERS and standardised data
45
46 processing model are essential to facilitate clinical research that utilises EMRs.
47
48
49
50
51

52 **Acknowledgements**

53
54
55
56
57 The authors would like to acknowledge the staff of the department of medical informatics of
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Kyoto University Hospital for their generous support.

Funding

This work was supported by the Coordination, Support and Training Program for Translational Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan and by Grants-in-Aid for Scientific Research of Japan (23790566).

Competing interests

None.

Contributors

KY designed the study, developed the ERS system, identified the computable eligibility criteria, wrote logical queries, collected data, and wrote the manuscript. ES is grant holder who designed the study, developed the ERS system, and wrote and edited the manuscript. TY designed and conducted the 'Risk of osteomyelitis of the jaw induced by oral bisphosphonates in patients taking medications for osteoporosis: A hospital-based cohort study in Japan' (BRONJ study) study and the current study, identified the computable eligibility criteria, and wrote and edited the manuscript. KA and MY designed and conducted the BRONJ study. ST designed the study and provided comments and feedback. KB is the principal investigator of the BRONJ study. MY owns the ERS system and supervised the

1
2
3
4 study. MF supervised the study and provided comments and feedback. All of the authors read
5
6
7 and approved the final manuscript.
8
9

10 **Provenance and peer review**

11
12
13
14 Not commissioned; externally peer reviewed.
15
16

17 **Data sharing statement**

18
19
20
21
22 No other data are available to share.
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

REFERENCES

1. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc.* 2009;**16**(3):316-27
2. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods Inf Med.* 2009;**48**(1):38-44.
3. Wasserman RC. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. *Acad Pediatr.* 2011;**11**(4):280-7.
4. Dean BB, Lam J, Natoli JL, et al. Review: Use of electronic medical records for health outcomes research: A literature review. *Med Care Res Rev* 2009;**66**(6):611-38.
5. Tannen RL, Weiner MG, Marcus SM. Simulation of the Syst-Eur randomized control trial using a primary care electronic medical record was feasible. *J Clin Epidemiol* 2006;**59**(3):254-64.
6. Williams JG, Cheung WY, Cohen DR. Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment. *Health Technol Assess* 2003;**7**(26): iii, v-x, 1-117.
7. Yamamoto K, Matsumoto S, Tada H, et al. A data capture system for outcomes studies that integrates with electronic health records: development and potential uses. *J Med Syst*

1
2
3
4 2008;**32**(5):423–7.
5
6

7
8 8. Yamamoto K, Matsumoto S, Yanagihara K, et al. A data-capture system for post-marketing
9
10 surveillance of drugs that integrates with hospital electronic health records. Open Access J
11
12 Clin Trials 2011;**3**:21–6.
13
14

15
16
17 9. Embi PJ, Jain A, Clark J, et al. Effect of a clinical trial alert system on physician
18
19 participation in trial recruitment. Arch Intern Med 2005;**165**(19):2272-7.
20
21
22

23
24 10. Campbell MK, Snowdon C, Francis D, et al. Recruitment to randomised trials: strategies
25
26 for trial enrollment and participation study. The STEPS study. Health Technol Assess
27
28 2007;**11**:iii, ix-105.
29
30
31

32
33
34 11. Dugas M, Lange M, Müller-Tidow C, et al. Routine data from hospital information
35
36 systems can support patient recruitment for clinical studies. Clin Trials 2010;**7**(2):183-9.
37
38
39

40
41 12. Thadani SR, Weng C, Bigger JT, et al. Electronic screening improves efficiency in
42
43 clinical trial recruitment. J Am Med Inform Assoc 2009;**16**(6):869-73.
44
45
46

47
48 13. Torgerson JS, Arlinger K, Käppi M, et al. Principles for enhanced recruitment of subjects
49
50 in a large clinical trial: The XENDOS study experience. Control Clin Trials
51
52 2001;**22**(5):515-25.
53
54
55

56
57 14. Kristianson KJ, Ljunggren H, Gustafsson LL. Data extraction from a semi-structured
58
59
60

1
2
3
4 electronic medical record system for outpatients: a model to facilitate the access and use of
5
6 data for quality control and research. *Health Inform J* 2009;**15**(4):305–319.
7
8

9
10
11 15. Shim JP. Past, present, and future of decision support technology. *Decis Support Syst*
12
13 2002;**33**(2):111-26.
14

15
16
17 16. Prat N. A UML-based data warehouse design method. *Decis Support Syst*
18
19 2006;**42**(3):1449-73.
20
21

22
23
24 17. Park YT. An empirical investigation of the effects of data warehousing on decision
25
26 performance. *Inform Manag* 2006;**43**(1):51.
27
28

29
30
31 18. Schlaps D, Schmid T. Data warehousing in clinical research and development - From
32
33 clinical data to knowledge portals. *Pharmind* 2004;**66**(5a):637-46.
34
35

36
37
38 19. Grant A, Moshyk A, Diab H, et al. Integrating feedback from a clinical data warehouse
39
40 into practice organisation. *Int J Med Inform* 2006;**75**(3-4):232-9.
41
42

43
44 20. Junttila K, Meretoja R, Seppälä A, et al. Data warehouse approach to nursing
45
46 management. *J Nurs Manag* 2007;**15**(2):155-61.
47
48

49
50
51 21. Rubin DL, Desser TS. A data warehouse for integrating radiologic and pathologic data. *J*
52
53 *Am Coll Radiol* 2008;**5**(3):210-7.
54
55

56
57
58 22. Johnson SB. Generic data modeling for clinical repositories. *J Am Med Inform Assoc*
59
60

1
2
3
4 1996;**3**(5):328-39.
5
6

7
8 23. Nadkarni PM, Brandt C. Data extraction and ad hoc query of an entity—attribute—value
9
10 database. *J Am Med Inform Assoc* 1998;**5**:511-27.
11
12

13
14 24. Anhøj J. Generic design of Web-based clinical databases. *J Med Internet Res*
15
16 2003;**5**(4):e27.
17
18

19
20
21 25. Chen RS, Nadkarni P, Marenco L, et al. Exploring performance issues for a clinical
22
23 database organized using an entity-attribute-value representation. *J Am Med Inform Assoc*
24
25 2000;**7**:475-87.
26
27

28
29
30 26. Dinu V, Nadkarni P. Guidelines for the effective use of entity-attribute-value modeling for
31
32 biomedical databases. *Int J Med Inform* 2007;**76**:769-79.
33
34

35
36
37 27. Corwin J, Silberschatz A, Miller PL, et al. Dynamic tables: an architecture for managing
38
39 evolving, heterogeneous biomedical data in relational database management systems. *J Am*
40
41 *Med Inform Assoc* 2007;**14**:86-93.
42
43

44
45
46 28. Wade TD, Hum RC, Murphy JR. A dimensional bus model for integrating clinical and
47
48 research data. *J Am Med Inform Assoc* 2011;**1**:96-102.
49
50

51
52
53 29. Pardillo J, Mazón JN. Model-driven development of OLAP metadata for relational data
54
55 warehouses. *Comput Stand Interfac* 2012;**34**(1):189-202.
56
57
58
59
60

- 1
2
3
4 30. Hettler M. Data mining goes multidimensional. *Health Inform.* 1997;**14**(3):43-6, 48, 51-6.
5
6
7
8 31. Gordon BD, Asplin BR. Using online analytical processing to manage emergency
9
10 department operations. *Acad Emerg Med* 2004;**11**(11):1206-12.
11
12
13
14 32. Kimball R, Reeves L, Ross M, et al. *The Data Warehouse Lifecycle Toolkit*. New York:
15
16 John Wiley, 1998.
17
18
19
20
21 33. Wang D, Peleg M, Tu SW, et al. Representation primitives, process models and patient
22
23 data in computer-interpretable clinical practice guidelines: a literature review of guideline
24
25 representation models. *Int J Med Inform.* 2002 **18**;68(1-3)
26
27
28
29
30
31 34. Weng C, Tu SW, Sim I, et al. Formal representation of eligibility criteria: A literature
32
33 review. *J Biomed Inform* 2010;**43**(3):451-67.
34
35
36
37
38 35. Lonsdale DW, Tustison C, Parker CG, et al. Assessing clinical trial eligibility with logic
39
40 expression queries. *Data Knowl Eng* 2008;**66**(1):3-17.
41
42
43
44 36. Tu SW, Peleg M, Carini S, et al. A practical method for transforming free-text eligibility
45
46 criteria into computable criteria. *J Biomed Inform* 2011;**44**(2):239-50.
47
48
49
50
51 37. Sordo M, Boxwala AA, Ogunyemi O, et al. Description and status update on GELLO: a
52
53 proposed standardized object-oriented expression language for clinical decision support. *Stud*
54
55
56
57 *Health Technol Inform.* 2004;**107**:164-8.
58
59
60

1
2
3
4 38. Séroussi B, Bouaud J. Using OncoDoc as a computer-based eligibility screening system
5
6
7 to improve accrual onto breast cancer clinical trials. *Artif Intell Med* 2003;**29**(1-2):153-67.
8
9

10 39. CDISC ASPIRE: Integration of clinical research and EHR: Eligibility coding standards.
11
12
13 http://crisummit2010.amia.org/files/symposium2008/S14_Niland.pdf (accessed 31 Mar
14
15
16 2012).
17
18
19

20 40. CDISC Study Design Model: SDM-XML Version 1.0.
21
22
23 <http://www.cdisc.org/stuff/contentmgr/files/0/8c85b168e80d6834ded59339b55fdb7/misc/cd>
24
25
26 isc_sdm_xml_1.0.pdf (accessed 31 Mar 2012).
27
28
29

30 41. US. National Cancer Institute (NCI). caMATCH.
31
32
33 <https://cabig.nci.nih.gov/community/tools/caMATCH> (accessed 31 Mar 2012).
34
35
36

37 42. Jagannathan V, Mullett CJ, Arbogast JG, et al. Assessment of commercial NLP engines for
38
39 medication information extraction from dictated clinical notes. *Int J Med Inform*
40
41
42 2009;**78**(4):284-91.
43
44
45

46 43. Pakhomov S, Weston SA, Jacobsen SJ, et al. Electronic medical records for clinical
47
48
49 research: application to the identification of heart failure. *Am J Manag Care*
50
51
52 2007;**13**(6):281-8.
53
54
55

56 44. Yamamoto K, Yamanaka K, Hatano E, et al. An eClinical trial system for cancer that
57
58
59
60

1
2
3
4 integrates with clinical pathways and electronic medical records. Clin trials 2012 9: 408-417.
5
6

7
8 45. Matsumura Y, Kuwata S, Yamamoto Y, et al. Template-based data entry for general
9
10 description in medical records and data transfer to data warehouse for analysis. Stud Health
11
12 Technol Inform 2007;**129**(Pt 1):412–416.
13
14

15
16
17 46. World Health Organization (WHO). International Classification of Diseases (ICD).
18
19
20 <http://www.who.int/classifications/icd/en/> (accessed 31 Mar 2012).
21
22

23
24 47. Yamazaki T, Yamori M, Yamamoto K, et al, Risk of osteomyelitis of the jaw induced by
25
26 oral bisphosphonates in patients taking medications for osteoporosis: A hospital-based cohort
27
28 study in Japan, Bone 2012: 51(5) 882–887.
29
30
31

32
33
34 48. Fellows JL, Rindal DB, Barasch A, et al. ONJ in two dental practice-based research
35
36 network regions. J Dent Res 2011;**90**(4):433-8.
37
38

39
40 49. Vestergaard P, Schwartz K, Rejnmark L, et al. Oral bisphosphonate use increases the risk
41
42 for inflammatory jaw disease: a cohort study. J Oral Maxillofac Surg. 2012 70(4):821-829.
43
44

45
46
47 50. Shortliffe EH, Cimino JJ. Biomedical Informatics: Computer Applications in Health Care
48
49 and Biomedicine. Health Informatics series. Springer, 2006.
50
51

52
53
54 51. McFadden E. Management of data in clinical trials. 2nd ed. Hoboken, NJ:
55
56 Wiley-Interscience, 2007.
57
58
59
60

- 1
2
3
4 52. Zhengwu L, Jing S. Clinical data management: current status, challenges, and future
5
6
7 directions from industry perspectives. *Open Access J Clin Trials* 2010;**2**: 93–105.
8
9
- 10 53. Data Management Association (DAMA) International. Data management body of
11
12 knowledge. <http://www.dama.org/i4a/pages/index.cfm?pageid=3364> (accessed 31 Mar 2012).
13
14
15
- 16 54. MedDRA MSSO. MedDRA MSSO. <http://www.meddramsso.com/> (accessed 31 Mar
17
18 2012).
19
20
21
22
- 23 55. US National Library of Medicine. SNOMED clinical terms,
24
25 http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html (accessed 31 Mar 2012).
26
27
28
29
- 30 56. Clinical Data Interchange Standards Consortium (CDISC). Study data tabulation model
31
32 (SDTM). <http://www.cdisc.org/sdtm> (accessed 31 Mar 2012).
33
34
35
36
- 37 57. Huff S. Development of the logical observation identifier names and codes (LOINC)
38
39 vocabulary. *J Am Med Inform Assoc* 1998;**5**(3):276-92.
40
41
42
43
- 44 58. Health level seven (HL7). <http://www.hl7.com/> (accessed 31 Mar 2012).
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 **FIGURE CAPTIONS**
5
6
7

8 **Figure 1.** Data model for our EMR retrieval system.
9

10
11 **Figure 2.** Example SQL to create the target patient list.
12

13
14
15 **Figure 3.** Example SQL to flag the target patient report for investigator confirmation.
16
17

18
19 **Figure 4.** Schema of data collection and confirmation.
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

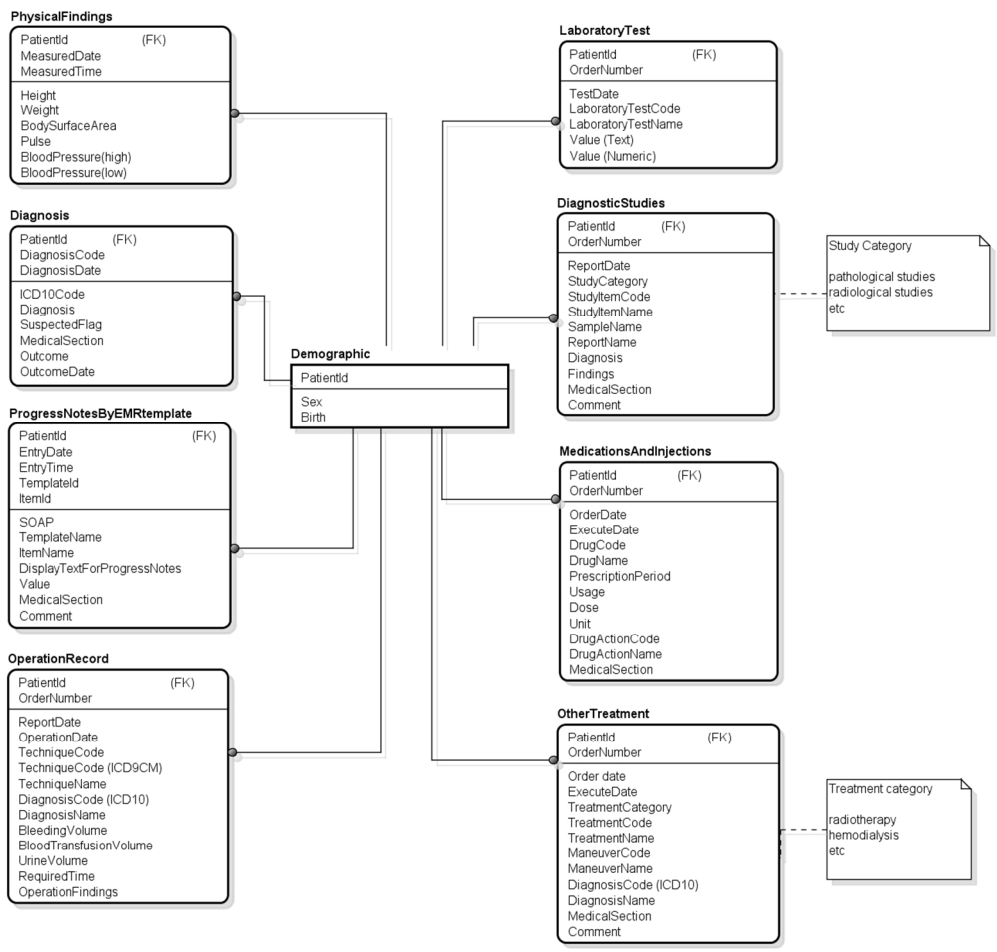


Figure 1. Data model for our EMR retrieval system.
109x104mm (300 x 300 DPI)

only

1
2
3
4
5
6
7
8 Create View_PatientsList as
9 Select PatientId From Demographic

10 Where

11 a. PatientId in (

```
12 Select PatientId From Diagnosis  
13 Where ICD10Code in (osteoporosis ICD10 code list) and  
14     DiagnosisDate >= '10/01/2000' and DiagnosisDate <= '09/30/2010' and  
15     SuspectedFlag = 'Fixed' )
```

16
17 and

18 a. PatientId in (

```
19 Select PatientId From MedicationsAndInjections  
20 Where DrugCode in (osteoporosis drugs code list) and  
21     ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010' )
```

22
23 and

24 a. PatientId not in (

```
25 Select PatientId From MedicationsAndInjections  
26 Where DrugCode in (intravenous BP drug code list) and  
27     ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010' )
```

28
29
30
31
32 Figure 2. Example SQL to create the target patient list.
33 81x60mm (300 x 300 DPI)

```

1
2
3
4
5
6
7 Select PatientId, Oral BP administrations, 1 From View_PatientsList a
8 Where a. PatientId in (
9   Select PatientId From MedicationsAndInjections
10  Where DrugCode in (oral BP drugs code list) and
11    ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010 ' )
12 Union all
13 Select PatientId, Oral BP administrations, 0 From View_PatientsList a
14 Where a. PatientId not in (
15   Select PatientId From MedicationsAndInjections
16   Where DrugCode in (oral BP drugs code list) and
17     ExecuteDate >= '10/01/2000' and ExecuteDate <= '09/30/2010')
18 Union all
19 Select PatientId, Inflammatory jaw condition diagnosis, 1 From View_PatientsList a
20 Where a. PatientId in (
21   Select PatientId From Diagnosis
22   Where ICD10Code in (inflammatory conditions of jaws ICD10 code list) and
23     DiagnosisDate >= '10/01/2000' and DiagnosisDate <= '09/30/2010' and SuspectedFlag = 'Fixed')
24 Union all
25 Select PatientId, Inflammatory jaw condition diagnosis, 0 From View_PatientsList a
26 Where a. PatientId not in (
27   Select PatientId From Diagnosis
28   Where ICD10Code in (inflammatory conditions of jaws ICD10 code list ) and
29     DiagnosisDate >= '10/01/2000' and DiagnosisDate <= '09/30/2010' and SuspectedFlag = 'Fixed')
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

```

Figure 3. Example SQL to flag the target patient report for investigator confirmation.
81x60mm (300 x 300 DPI)

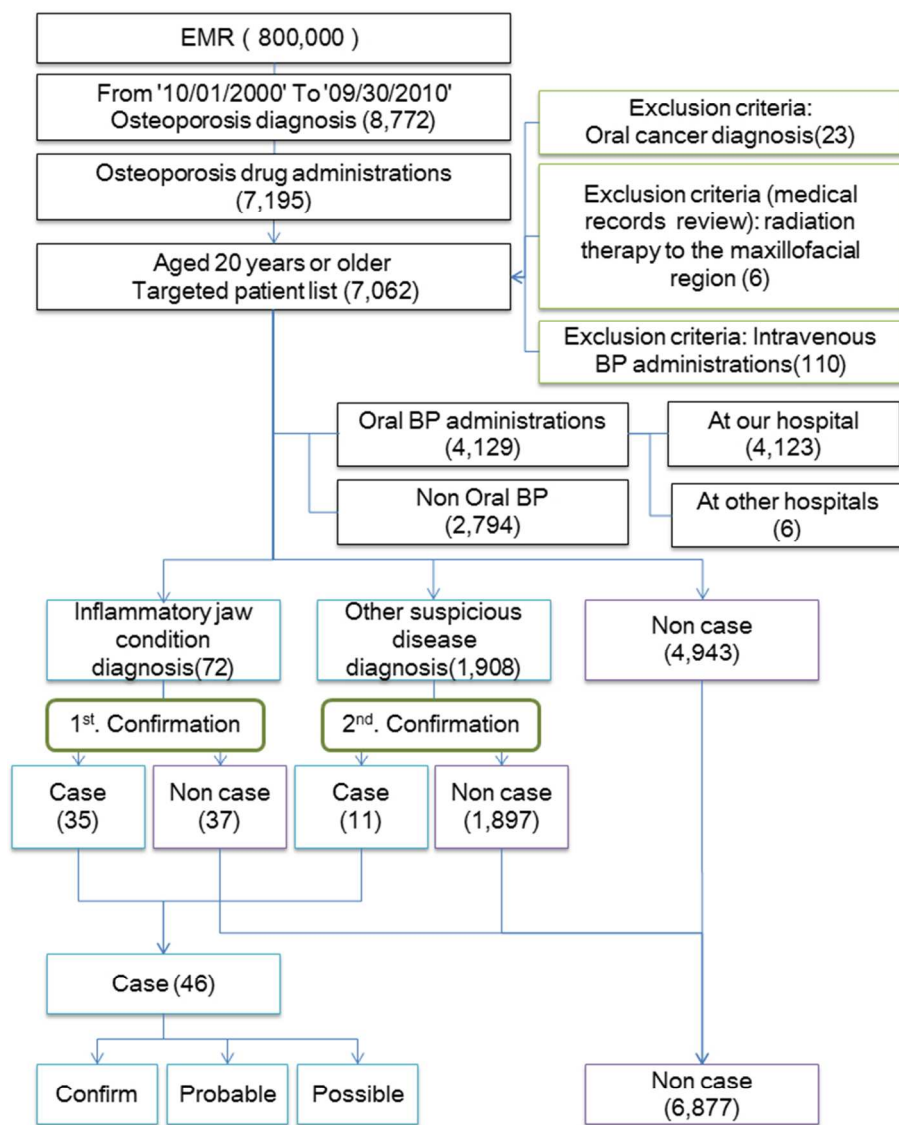


Figure 4. Schema of data collection and confirmation.
60x81mm (300 x 300 DPI)

1
2
3
4 **A pragmatic method for electronic medical-record-based observational studies:**
5
6
7 **developing an electronic medical records retrieval system for clinical research**
8
9

10 Keiichi Yamamoto¹, Eriko Sumi², Toru Yamazaki³, Keita Asai³, Masashi Yamori³, Satoshi
11 Teramukai¹, Kazuhisa Bessho³, Masayuki Yokode², Masanori Fukushima⁴
12
13
14
15

16
17 ¹Department of Clinical Trial Design and Management, Translational Research Centre, Kyoto
18 University Hospital, Kyoto, Japan
19
20
21
22

23
24 ²Department of Clinical Innovative Medicine, Translational Research Centre, Kyoto
25 University Hospital, Kyoto, Japan
26
27
28

29
30
31 ³Department of Oral and Maxillofacial Surgery, Graduate School of Medicine, Kyoto
32 University, Kyoto, Japan
33
34
35

36
37 ⁴Translational Research Informatics Centre, Foundation for Biomedical Research and
38 Innovation, Kobe, Japan
39
40
41
42

43
44 Corresponding author: Keiichi Yamamoto, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto,
45 606-8507 Japan. E-mail: kyamamo@kuhp.kyoto-u.ac.jp, Tel: +81-75-751-4717, Fax
46 +81-75-751-3399
47
48
49
50

51
52
53
54 Total number of words: 3,997
55

56
57
58 Short title: A pragmatic method for EMR-based observational studies
59
60

1
2
3
4 Keywords: clinical research informatics, data warehouse, OLAP, computable eligibility
5
6
7 criteria, pharmacoepidemiology, hospital-based cohort study
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

ABSTRACT

Objective: The use of electronic medical record (EMR) data is necessary to improve clinical research efficiency. However, it is not easy to identify patients who meet research eligibility criteria and collect the necessary information from EMRs because the data collection process must integrate various techniques, including the development of a data warehouse and translation of eligibility criteria into computable criteria. This research aimed to demonstrate an electronic medical records retrieval system (ERS) and an example of a hospital-based cohort study that identified both patients and exposure with an ERS. We also evaluated the feasibility and usefulness of the method. **Design:** The system was developed and evaluated.

Participants: In total, 800,000 cases of clinical information stored in EMRs at our hospital were used. **Primary and secondary outcome measures:** The feasibility and usefulness of the ERS, the method to convert text from eligible criteria to computable criteria, and a confirmation method to increase research data accuracy. **Results:** To comprehensively and efficiently collect information from patients participating in clinical research, we developed an ERS. To create the ERS database, we designed a multi-dimensional data model optimised for patient identification. We also devised practical methods to translate narrative eligibility criteria into computable parameters. We applied the system to an actual hospital-based cohort study performed at our hospital and converted the test results into computable criteria. Based on this information, we identified eligible patients and extracted data necessary for

1
2
3
4 confirmation by our investigators and for statistical analyses with our ERS. **Conclusion:** We
5
6
7 propose a pragmatic methodology to identify patients from EMRs who meet clinical research
8
9
10 eligibility criteria. Our ERS allowed for the efficient collection of information on the
11
12
13 eligibility of a given patient, reduced the labour required from the investigators, and
14
15
16 improved the reliability of the results.

17 18 19 **ARTICLE SUMMARY**

20 21 22 **Article focus**

23
24
25
26
27 The focus of this work was to establish a pragmatic methodology to efficiently collect
28
29
30 information from EMRs about patients who meet clinical research eligibility criteria.
31
32

33 34 **Key messages**

35
36
37 The use of electronic medical record (EMR) data is necessary to improve clinical research
38
39
40 efficiency. However, it is not easy to identify patients who meet research eligibility criteria
41
42
43 and collect necessary data from EMRs because the data collection process must integrate
44
45
46 various techniques, including the development of a data warehouse and the translation of
47
48
49 eligibility criteria into computable criteria. An efficient ERS and a standardised data
50
51
52 processing model that integrates these techniques are essential to facilitate clinical research
53
54
55 that utilises EMRs.
56
57
58
59
60

Strengths and limitations of this study

- Our method uses a specialised data model for patient identification in clinical research and efficient data conversion that does not depend on the EMR database structure when converting narrative criteria to computable criteria.
- We propose that computable criteria should not be a result of the automated conversion of narrative criteria but rather a result of research preparation involving medical concepts that are not expressed logically or explicitly in the narrative criteria. Therefore a large amount of the conversion of the eligibility criteria to computable criteria should be executed at the protocol development stage.
- It is important to further discuss protocol standardisation, including eligibility criteria representation for computable use.
- Enabling ERS use in and across multiple institutions is an important future task.

BACKGROUND

Medical information technology has recently advanced in many countries, and enormous amounts of clinical data are already stored as electronic medical records (EMRs). Utilising the data collected in EMRs is necessary to improve clinical research efficiency [1-3]. An EMR is a large database of patient data and is used in observational research to investigate the relationships among diseases, treatments, and outcomes [4-7], to conduct surveillance for rare drug reactions [4, 8], and to recruit patients for clinical trials [9-13]. However, it is not easy to identify patients who meet research eligibility criteria and collect necessary information from EMRs [2-3]. Herein, we describe three major issues concerning EMR-based observational studies: EMR patient data retrieval function, eligibility criteria protocol representation, and EMR data accuracy.

To identify patients who meet research eligibility criteria, it is necessary to obtain various types of information stored in EMRs by subject, e.g., diagnosis and prescribed medications. However, the EMR database is designed to facilitate online transaction processing for rapid and detail-oriented clinical information searches on individual patients, and the current EMR system does not facilitate this retrieval function [2-3, 14]. Data warehouses are essential components of data-driven decision support. To allow for efficient research analyses, EMR data must first be warehoused to enable data analyses across patient populations [15-21]. However, health care data modelling is difficult and time-consuming because of the

1
2
3
4 complexity of the medical knowledge involved. Thus, the most common approaches to
5
6
7 clinical data warehouse modelling are variations on the entity-attribute-value (EAV) model
8
9
10 [22-28], where data are stored in a single table with three columns: entity identification,
11
12
13 attribute, and attribute value. The EAV design has advantages, including flexibility and ease
14
15
16 of storage; however, it requires transforming EAV data into another analytical format before
17
18
19 analysis [25, 28]. Online analytical processing (OLAP) is most frequently used for searching
20
21
22 data stored in the data warehouse [29-31]. OLAP systems in relational databases are typically
23
24
25 designed based on Kimball's star schema [32]. However, the star schema was devised to
26
27
28 facilitate online measurement analyses. In health care, this method can be used to
29
30
31 dynamically gather online analyses of numeric data (e.g., a specific dose of a drug for a
32
33
34 specific disease) in clinical practice. Therefore, this method is not suitable for identifying
35
36
37 patients who meet the complicated eligibility criteria for a given clinical research study.
38
39
40 Data-modelling methods that facilitate the identification of patients and enable the collection
41
42
43 of necessary information from EMRs remain to be established [28].
44

45
46
47 Current eligibility criteria are written in a text format that cannot be computationally
48
49
50 processed. Additionally, to be applied in actual EMR, eligible criteria need to be integrated
51
52
53 with the data model of EMRs [33]. Several investigations have sought to establish
54
55
56 computable eligibility criteria [34-41]. However, there is no consensus regarding a standard
57
58
59 patient information model [33], and the eligibility criteria are not yet completely standardised.
60

1
2
3
4 Using natural language processing (NLP) technologies to convert the text format of eligibility
5
6
7 criteria to a computer or to extract patient identifications from EMRs is far from perfect
8
9
10 without human intervention [3, 42-43].

11
12
13 Current EMRs have been used to support claims for medical service fees and the treatments
14
15 administered to each patient; therefore, data gathered specifically for research purposes may
16
17
18 be incomplete and unreliable [2-3, 44].
19
20
21

22
23 Although various investigations on each technique are executed individually, standardised
24
25 methods must still be established that integrate these techniques, facilitate the identification
26
27
28 of patients who are eligible for clinical research, and collect necessary information from
29
30
31 EMRs.
32
33

34 35 **OBJECTIVE**

36
37
38 We designed a pragmatic data processing model optimised for patient identification and for
39
40
41 the collection of necessary information from EMRs for clinical research. These tools are
42
43
44 implemented as an electronic medical records retrieval system (ERS) [44].
45
46
47

48
49 This research aimed to demonstrate an ERS and an example of a hospital-based cohort study
50
51
52 that used the ERS to identify both patients and exposure. Another aim was to evaluate the
53
54
55 feasibility and usefulness of the ERS, the method to convert text form eligible criteria to
56
57
58 computable criteria, and a confirmation method to increase research data accuracy.
59
60

MATERIALS AND METHODS

Outline of our procedure for patient identification and data collection from the EMR

To identify patients who met the eligibility criteria for the clinical research in question, data were collected in the following ways:

- 1) The text form of the narrative criteria was converted into computable criteria.
- 2) A targeted patient list was created.
- 3) A flag was added for investigators to confirm the targeted patient list.
- 4) Reports were created for the investigators to confirm.
- 5) After confirmation by the investigator, the statistical analyses were executed.

EMR retrieval system

In our hospital, EMR use was introduced in 2005; approximately 800,000 cases of clinical information have already been stored. To comprehensively and efficiently collect information about patients participating in clinical research, we developed an ERS [44].

EMRs store various types of information, integrating billing, pharmacy, radiology, laboratory information, and others [4]. In creating the ERS database, we designed a new data model based on the star schema that was optimised for patient identification in clinical research. We

1
2
3
4 identified nine data categories from EMRs that are useful for clinical research: demographic
5
6 characteristics, physical findings, diagnostic studies, laboratory tests, diagnoses, progress
7
8 reports on an EMR template [44-45], medications and injections, operation records, and other
9
10 treatments. We then designated these categories to 'entities'. In our hospital, the diagnosis is
11
12 managed by codes that were originally defined by our hospital and mapped with International
13
14 Statistical Classification of Diseases (ICD) 10 codes [46] for medical insurance purposes.
15
16 Operations codes were also managed by codes that originally were defined by our hospital
17
18 and mapped with ICD-9 Clinical Modification codes. We identified available columns (e.g.,
19
20 ICD code, diagnosis date) from the EMR data model and designated these columns as
21
22 'attributes' of the entities.
23
24
25
26
27
28
29
30
31
32

33
34 Figure 1 presents our data model. In our model, all entities in a given schema are independent
35
36 and complete; this allows for logical operations and for the creation of eligible patient lists
37
38 for each respective parameter in a study. The target patient list is generated by combining
39
40 these patient lists. The data model also supports the inference of medical concepts expressed
41
42 in the eligibility criteria in reference to corresponding patient data accumulated in EMRs
43
44
45
46
47
48 [33-34].
49
50

51
52 In our hospital, a replicate of the EMR database known as 'Open DB' was established for the
53
54 secondary use of accumulated EMR data [7]. A data mart for our ERS was created to ensure
55
56
57
58 that the data retrieval process was practical and independent of the EMR system structure; the
59
60

1
2
3
4 data mart was created on the relational database management system by extracting,
5
6
7 transforming, and loading (ETL) information from the Open DB [7, 44]. The ETL process is
8
9
10 performed automatically once nightly except for the 'Progress notes by EMR template' entity,
11
12
13 which is referred directly from the Open DB to ensure real-time visibility for the eClinical
14
15
16 trial [44].

17
18
19 An OLAP tool was installed to efficiently search through data from multiple patients [44].

20
21
22 The OLAP tool runs in an Internet browser and can generate structured query language
23
24
25 (SQL) based on predefined metadata (i.e., a data model) by defining logical queries (i.e.,
26
27
28 programs) using a graphical user interface (GUI). Moreover, this tool allows reports on
29
30
31 information retrieved from the browser to be transcribed using hypertext markup language
32
33
34 (HTML). The reports are created in various formats, including portable document format
35
36
37 (PDF), comma separated values (CSV), and extensible markup language (XML) [44].

38
39
40 To protect personal information in medical records at our hospital, the EMR network is
41
42
43 separated physically from other networks. Our data mart and OLAP servers are deployed in
44
45
46 the same EMR network and managed using the same EMR security policies. Additionally, the
47
48
49 use of our ERS is limited to clinical research approved by the ethics committee at our hospital,
50
51
52 and only designated staff members at our centre are allowed to retrieve data. Our centre
53
54
55 creates and manages ERS user identification separate from the EMRs. For the external output
56
57
58 of CSV and other data, permission must be obtained from our department of medical
59
60

1
2
3
4 informatics, and data extraction must be executed in the presence of supervisors who are
5
6
7 responsible for protecting personal information at our hospital.
8
9

10 **Application to clinical research**

11
12
13
14 We applied the system to a hospital-based cohort study performed at our hospital titled 'Risk
15
16
17 of osteomyelitis of the jaw induced by oral bisphosphonates (BP) in patients taking
18
19
20 medications for osteoporosis: A hospital-based cohort study in Japan' [47], in which we
21
22
23 identified eligible patients, extracted research data, and evaluated the feasibility of our system.
24
25
26 The ethics committee at Kyoto University Hospital approved this research. A different paper
27
28
29 details the purpose, methods, results, and discussion of this research [47].
30
31

32
33 This research aimed to estimate the risks for osteomyelitis of the jaw in osteoporosis patients
34
35
36 at our hospital who had been exposed to oral BP compared to those who had not [48-49].
37
38

39 The eligibility criteria were as follows:
40
41

42 **Inclusion criteria**

- 43
44
45
46
47 • Patients diagnosed with osteoporosis and treated with osteoporosis medications at
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
- Patients aged 20 years or older.

Exclusion criteria

- Patients with a history of treatment with radiation therapy to the maxillofacial region.
- Patients with primary or metastatic tumours in the maxillofacial region.
- Patients treated with intravenous BP.

The data collected were diagnosis, date of diagnosis, sex, birthdate, and the doses and dates when osteoporosis medications, steroids, anticancer drugs, diabetes drugs and HbA1c tests were administered.

Conversion of the text form of the narrative criteria to computable criteria

To identify eligible patients and collect the necessary data from the EMRs, narrative criteria and data must be converted to computable criteria. Such computable criteria include entities, attributes, logical operators (i.e., 'and' and 'or'), codes, and parameters [33-37]. The clinical research purpose and clinical practice demands made it necessary to perform this task.

We manually executed the conversion from text eligibility criteria to computable criteria. As an example of the conversion from narrative criteria to computable criteria, we present the following two-step conversion procedure:

Step 1: Convert the narrative criteria into entity-level criteria.

Medical concepts expressed as narrative criteria are mapped onto entities in the data model and converted into entity-level criteria. This task is manually performed at the protocol

1
2
3
4 development stage of the study by the investigators. For each entity, a criterion is created to
5
6
7 extract patients who meet each condition. If exclusive conditions for the same entity must be
8
9
10 defined, a different criterion is created. Additionally, the list of codes for drugs and diagnoses
11
12
13 (i.e., ICD-10) is created, and the period of treatments and others are defined by investigators.

14
15 In this study, we mapped ‘osteoporotic patients’ onto two entities (i.e., ‘diagnosis’ and
16
17
18 ‘medications and injections’) and converted it to a combination of two criteria (i.e.,
19
20
21 ‘diagnosis of osteoporosis’ and ‘osteoporosis drug administration’). In the test research, we
22
23
24 defined the entity-level criteria according to the entered diagnosis and ordered treatments
25
26
27 rather than the diagnostic criteria of the disease. This process reflects that the test research
28
29
30 aimed to estimate some risks of osteomyelitis of the jaw with BP administration instead of
31
32
33 diagnosing osteoporosis patients accurately. The recorded diagnosis in the EMR was typically
34
35
36 designed to ensure payment for medical claims. We thus sought to reduce the number of false
37
38
39 positives by extracting patients with a given treatment type.

40
41
42 Step 2: Convert entity-level criteria into attribute-level criteria (i.e., computable criteria).

43
44
45
46 The abovementioned corresponding codes, date and parameters are mapped onto attributes of
47
48
49 the entity-level criteria, and these factors become computable criteria.
50

51 52 53 **Creating a targeted patient list**

54
55
56
57 A targeted patient list is created from the entire set of patients for whom EMRs have been
58
59
60

1
2
3
4 obtained by defining logical queries (i.e., programs defined by the GUI) based on the
5
6
7 computable criteria included in the ERS.
8
9

10 Logical queries are first defined in the ERS to identify patients who meet the conditions for
11
12
13 each criterion. The ERS automatically generates the SQL necessary for data extraction
14
15
16 according to the logical queries. Logical queries are then defined to include or exclude
17
18
19 eligible patients who meet each criterion for the demographic entity. The targeted patient list
20
21
22 is created by executing the logical query. Figure 2 presents an example of an SQL
23
24
25 automatically generated by the ERS.
26
27
28

29 We thus designed our data model to enable the creation of a targeted patient list by defining
30
31
32 the patients extracted from each criterion (i.e., 'in' or 'not in') as conditions for the
33
34
35 demographic entity that was the unique patient list for the entire hospital. If logical queries
36
37
38 are defined using our method, even if the eligibility criteria are complicated, it is not
39
40
41 necessary to dramatically change the SQL structure generated in the ERS.
42
43
44

45 **Flagging entries for investigators to confirm**

46
47

48 To improve research data accuracy, confirmation by the investigators is necessary. When
49
50
51 confirmation is required, additional information is linked.
52
53
54

55 For the targeted patient list, logical queries are defined to flag certain items according to the
56
57
58 investigators' interest. Necessary logical queries are first defined for each criterion. Logical
59
60

1
2
3
4 queries are then defined for addition to the patient list as '1' if the data correspond or '0' if
5
6 they do not. Data sets created by these operations are joined by 'union' and pivoted on a
7
8 cross-tabulation list using statistical analysis software. We show an example of an SQL
9
10 generated by the ERS in Figure 3.
11
12
13

14 15 16 **Create reports for investigators to confirm** 17

18
19
20 To help investigators confirm the targeted patient list, reports are created by linking the
21
22 findings for diagnostic imaging, pathological diagnosis, operations, and other findings.
23
24 Investigators confirm these entries using the reports and EMR information, including
25
26 progress notes and images. When the diagnosis history, medication, laboratory results,
27
28 progress notes, and other information are necessary, the same operation is executed for each
29
30 instance. For example, the list of radiological findings involves 'patient id', 'study category',
31
32 'report name', 'diagnosis', 'findings', and 'comment'. The reports may improve the
33
34 investigators' confirmation efficiency because they prevent the need to refer to the medical
35
36 records for each patient who needs confirmation.
37
38
39
40
41
42
43
44
45
46

47 **Confirmation by the investigator and execution of the statistical analyses.** 48

49
50
51 The investigators confirm the accumulated data and execute the statistical analysis. In this
52
53 test research, two oral and maxillofacial surgeons diagnosed cases by a chart review with an
54
55 observation of imaging findings [47].
56
57
58
59
60

Systemic evaluation

To evaluate our system, we collected information about the research period using the recall method. For the accuracy of the data collected by the ERS, we evaluated the results after they were confirmed by the investigator.

RESULTS

Computable criteria, datasets, and system evaluation

We present the computable criteria in Table 1. To increase data accuracy, we collected all of the exclusion criteria for the investigators to confirm. As Table 1 shows, we extracted information from EMRs. For investigator confirmation, we also reported all targeted patients using the following lists: osteoporosis drugs administered, oral BP administered, intravenous BP administered, diabetes drugs administered, anticancer drugs administered, steroid drugs administered, osteoporosis diagnoses, oral cancer diagnoses, patients diagnosed with inflammation of the jaw, patients diagnosed with other suspicious diseases, patients diagnosed with diabetes, HbA1c values, radiological findings, pathological findings, and radioisotope findings. These data were extracted from the ERS for statistical analyses, presented in CSV format, and analysed using statistics software.

Table 1. Computable criteria for our test research

Criterion	Entity	Operator symbol	Attribute	Operator symbol	Parameter
Created a targeted patient list					
Inclusion criteria: Osteoporosis diagnosis	Diagnosis	-	ICD10Code	In	(osteoporosis ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Inclusion criteria: Osteoporosis drug administrations	Medications and Injections	-	DrugCode	in	(osteoporosis drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
Added a flag for investigators to confirm the targeted patient list					
Exclusion criteria: Oral cancer diagnosis	Diagnosis	-	ICD10Code	in	(oral cancer ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Exclusion criteria: Intravenous BP administrations	Medications and Injections	-	DrugCode	in	(intravenous BP drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
Oral BP administrations	Medications and Injections	-	DrugCode	in	(oral BP drugs code list)
		and	ExecuteDate	>=	'10/01/2000'
		and	ExecuteDate	<=	'09/30/2010'
Inflammatory jaw condition diagnosis	Diagnosis	-	ICD10Code	in	(inflammatory conditions of jaws ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Other suspicious disease diagnosis	Diagnosis	-	ICD10Code	in	(other suspicious disease ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed
Diabetes diagnosis	Diagnosis	-	ICD10Code	in	(diabetes ICD10 code list)
		and	DiagnosisDate	>=	'10/01/2000'
		and	DiagnosisDate	<=	'09/30/2010'
		and	SuspectedFlag	=	Fixed

Steroid drug administrations	Medications	-	DrugCode	in	(steroid drugs code list)
	and	and	ExecuteDate	>=	'10/01/2000'
	Injections	and	ExecuteDate	<=	'09/30/2010'
Anticancer drug administrations	Medications	-	DrugCode	in	(anticancer drugs code list)
	and	and	ExecuteDate	>=	'10/01/2000'
	Injections	and	ExecuteDate	<=	'09/30/2010'
Diabetes drug administrations	Medications	-	DrugCode	in	(diabetes drugs code list)
	and	and	ExecuteDate	>=	'10/01/2000'
	Injections	and	ExecuteDate	<=	'09/30/2010'
HbA1c test execution	Laboratory Test	-	LaboratoryTestCode	in	(HbA1c test code)
		and	TestDate	>=	'10/01/2000'
		and	TestDate	<=	'09/30/2010'
Created reports for confirmation by the investigators					
Radiological finding reports	Diagnostic Studies	-	ReportName	in	(report name list of oral region)
Pathologic finding reports	Diagnostic Studies	-	SampleName	contains	'bone'
		or	SampleName	contains	'jaw'
Radio isotope finding reports	Diagnostic Studies	-	-	-	-

BP, bisphosphonates; ICD, International Classification of Diseases; ID, identifications

Among the approximately 800,000 cases at our hospital, 8,772 were categorised using the terms 'Inclusion criteria: Osteoporosis diagnosis'; among this group, 7,195 were further categorised using 'Inclusion criteria: Osteoporosis drug administration'. We then calculated the time that had elapsed since the osteoporosis diagnosis, determined that 7,062 patients were aged 20 years or older, and created a targeted patient list. Among those on the targeted patient list, 23 patients were placed under the heading 'Exclusion criteria: Oral cancer diagnosis', 110 under 'Exclusion criteria: Intravenous BP administration', 4,200 under 'Oral

1
2
3
4 BP administration', 84 under 'Inflammatory jaw condition diagnosis', 2,064 as 'Other
5
6 suspicious disease diagnosis', 1,700 as 'Diabetes diagnosis', 4,551 as 'Steroid drug
7
8 administration', 904 as 'Anticancer drug administrations', 1,055 as 'Diabetes drug
9
10 administrations', and 3,641 as 'HbA1c test execution'. Because of the end point considered,
11
12 patients who were classified under 'Inflammatory jaw condition diagnosis' or 'Other
13
14 suspicious disease diagnosis' were confirmed using predefined hierarchical diagnostic criteria
15
16 by investigators who performed the statistical analyses and arranged the research results. We
17
18 show the schema of data collection and confirmation as Figure 4 [47].
19
20
21
22
23
24
25
26
27

28 The accuracy of the data extracted by the ERS was then characterised. Reviewing the medical
29
30 records revealed that 2,817 patients were not labelled as 'Oral BP administration', including 7
31
32 (1 who received intravenous BP) treated at other hospitals. 6 patients had been treated with
33
34 radiation therapy to the oral and maxillofacial regions. Among the 72 patients classified under
35
36 'Inflammatory jaw condition diagnosis', 35 cases and 37 non-cases were identified.
37
38
39
40
41
42

43 The data extraction period lasted approximately three months. Ten meetings were held during
44
45 the protocol development stage to create and validate the computable criteria and the list of
46
47 codes for various drugs and diagnoses (i.e., ICD-10). The time required for logical query
48
49 definition when using the ERS was approximately 20 hours. The investigator confirmations
50
51 and statistical analyses took approximately four months.
52
53
54
55
56
57
58
59
60

DISCUSSION

We identified eligible patients for this research and extracted the data necessary for confirmation by investigators and for statistical analyses.

We asked the chart reviewers to evaluate the system in a questionnaire about ‘the effect of computer programming support for data retrieval from the EMR’, ‘the result of the data retrieval’, ‘the positive and negative aspects of our ERS use’, and ‘the aspects of our method that should be improved’. The investigators evaluating the system mentioned that the following points: 1) the method enabled them to extract the necessary data for diagnosis and drug administration without exception; 2) by screening the entire patient population at the hospital using the ERS, they could identify not just eligible patients in the department of oral and maxillofacial surgery but all eligible patients, which reduced the study bias; and 3) by creating reports for confirmation, it enabled investigators to devote their time to reading images, thus effectively reducing the time required for reviewing medical records. The aspects of our method that should be improved are the ‘lack of claim data’ and the ‘administrative complexity of EMR data use’. No negative aspects of our ERS use were noted.

The ERS allowed for the collection of information on patient eligibility by efficiently combining clinical information. Although we did not compare our method with other

1
2
3
4 methods, our proposed method reduced the labour normally required from investigators and
5
6
7 improved the reliability of test research results, which indicated that it was useful.
8
9

10 To design the ERS database, we designed a new data model optimised for patient
11
12 identification. The main differences between our data model and the star schema were as
13
14 follows: 1) demographic data, which were presented in list form in our EMR system, were
15
16 presented as a fact-less fact table, and 2) date, time, measurements and text information were
17
18 presented as a fact-less fact table, and 2) date, time, measurements and text information were
19
20 presented in dimension tables [32]. The most significant characteristic of our method for
21
22 patient identification is the use of a specialised data model in clinical research and the ability
23
24 to execute a large number of conversion tasks at the protocol development stage. Data can be
25
26 converted efficiently in a way that does not depend on the EMR database structure when
27
28 converting narrative criteria to computable criteria. In this research, we considered whether
29
30 data were extracted directly from EMRs at the protocol development stage. However, EMR
31
32 data were recorded in a sequential format for every medical practice, and the database
33
34 structure was complicated. Comprehending the location and meaning of the necessary data
35
36 thus required tremendous effort. It was difficult to make precise logical queries for patient
37
38 identification. However, because our ERS data model was arranged by subjects (e.g., tests,
39
40 diagnosis), it was easy to interpret the available information. Due to the standardisation of
41
42 computable criteria and SQL possible with the ERS, it was also possible to create computable
43
44 criteria in little time. Additionally, verifying the patient identification accuracy was easy
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 because it was possible to test each individual criterion.
5
6

7
8 The SQL generated by our ERS does not reduce the time required for data retrieval. Our ERS
9
10 also cannot retrieve information that is not in the data model. Current EMRs do not store all
11
12 necessary data for clinical research, including information related to pregnancy, performance
13
14 status, cancer stage, availability of transportation to the hospital, specific tests that are not
15
16 typically performed, drug regimen, outcomes (including death), and adverse events.
17
18 Additionally, all tests are not administered to all patients, and necessary information may
19
20 have been recorded in medical records at another hospital [44]. To facilitate EMR use in
21
22 clinical research, it is necessary to accumulate as much of this information as possible. In the
23
24 hospital, much of this information does not integrate well with EMRs, including test reports
25
26 stored only in the departmental system [50]. However, it is important to utilise this
27
28 information. Additionally, enabling ERS use in and across multiple institutions is also an
29
30 important future task.
31
32
33
34
35
36
37
38
39
40
41
42

43
44 Currently, most clinical research studies that use data from EMRs are planned according to
45
46 the concept that the primary use of EMRs is for clinical practice and a secondary use is for
47
48 clinical research [44]. Therefore, most investigators attempt to convert the text form
49
50 eligibility criteria that already have been defined on a protocol to computable criteria at the
51
52 data collecting stage [35. 36]. However, we propose that computable criteria should not be a
53
54 result of the automated conversion of narrative criteria but rather a result of research
55
56
57
58
59
60

1
2
3
4 preparation involving medical concepts that are not expressed logically or explicitly in the
5
6
7 narrative criteria. Some medical concepts may be interpreted differently depending on the
8
9
10 research and the investigator caring for the patients. Additionally, current eligibility criteria
11
12
13 are vague or complex, and they do not consider the use of the actual EMR. To convert
14
15
16 computable criteria appropriately, high-level medical decisions to answer the research
17
18
19 question are required. Therefore, we thought that a large amount of the conversion of the
20
21
22 eligibility criteria to computable criteria should be executed at the protocol development
23
24
25 stage. In addition, the conversion process should be divided into entity-level conversions that
26
27
28 require higher medical decisions and attribute-level conversions. To reduce the burden of
29
30
31 conversion, it may be useful to apply NLP technology for the conversion from entity-level
32
33
34 criteria to attribute-level criteria. Moreover, it is important to further discuss protocol
35
36
37 standardisation, including eligibility criteria representation for computable use. For instance,
38
39
40 the attribute-level criteria that describe the search conditions in detail may be useful in global
41
42
43 studies to address diseases that vary according to the diagnostic criteria used in each country.

44
45 Concerning EMR data accuracy, the ICD10 code (osteomyelitis of the jaw) sensitivity was
46
47
48 48.6% (35/72). The investigators reported 6 simple diagnosis errors, 7 oral BP
49
50
51 administrations at other hospitals, and 6 patients who were treated with radiation therapy in
52
53
54 the oral and maxillofacial region [47]. For the accuracy of current EMRs, the investigators
55
56
57 had to confirm the information. However, the EMRs provided rich confirmation data and
58
59
60

1
2
3
4 were useful in improving research data accuracy. In this study, we checked the data from
5
6
7 actual EMRs manually and identified patients precisely and extensively using coded
8
9
10 information, narrative information, and images. However, only information from existing
11
12 EMRs was available. Current EMRs have a high degree of flexibility in data entry and are not
13
14 currently managed for research purposes, which decreases their reliability. It is necessary to
15
16 improve data quality through quality control without placing too much of a burden on clinical
17
18 practice. Alternatively, it may be possible to organise data sufficiently before research use
19
20 [51-53]. Standardising the terminology and exchange formats used in the healthcare setting
21
22 has facilitated international discourse [46, 54-58]. It is necessary to further discuss not only
23
24 clinical practice but also research purposes, particularly how to utilise various standards when
25
26 using EMRs beyond the hospital setting.
27
28
29
30
31
32
33
34
35

36 **CONCLUSION**

37
38
39
40 We propose a pragmatic method for EMR-based observational studies. Our ERS is already
41
42 used to support hospital-based cohort studies, clinical trial recruitment, and the eClinical trial
43
44 infrastructure [44] at our centre. We believe an efficient ERS and standardised data
45
46 processing model are essential to facilitate clinical research that utilises EMRs.
47
48
49
50

51 **Acknowledgements**

52
53
54
55
56
57 The authors would like to acknowledge the staff of the department of medical informatics of
58
59
60

1
2
3
4 Kyoto University Hospital for their generous support.
5
6

7 8 **Funding** 9

10
11 This work was supported by the Coordination, Support and Training Program for
12
13 Translational Research of the Ministry of Education, Culture, Sports, Science and Technology
14
15 of Japan and by Grants-in-Aid for Scientific Research of Japan (23790566).
16
17
18
19

20 21 **Competing interests** 22

23
24
25 None.
26
27

28 29 **Contributors** 30

31
32 KY designed the study, developed the ERS system, identified the computable eligibility
33
34 criteria, wrote logical queries, collected data, and wrote the manuscript. ES is grant holder
35
36 who designed the study, developed the ERS system, and wrote and edited the manuscript. TY
37
38 designed and conducted the 'Risk of osteomyelitis of the jaw induced by oral
39
40 bisphosphonates in patients taking medications for osteoporosis: A hospital-based cohort
41
42 study in Japan' (BRONJ study) study and the current study, identified the computable
43
44 eligibility criteria, and wrote and edited the manuscript. KA and MY designed and conducted
45
46 the BRONJ study. ST designed the study and provided comments and feedback. KB is the
47
48 principal investigator of the BRONJ study. MY owns the ERS system and supervised the
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 study. MF supervised the study and provided comments and feedback. All of the authors read
5
6
7 and approved the final manuscript.
8
9

10 **Provenance and peer review**

11
12
13
14 Not commissioned; externally peer reviewed.
15
16

17 **Data sharing statement**

18
19
20
21
22 No other data are available to share.
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

REFERENCES

1. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *J Am Med Inform Assoc.* 2009;**16**(3):316-27
2. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods Inf Med.* 2009;**48**(1):38-44.
3. Wasserman RC. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. *Acad Pediatr.* 2011;**11**(4):280-7.
4. Dean BB, Lam J, Natoli JL, et al. Review: Use of electronic medical records for health outcomes research: A literature review. *Med Care Res Rev* 2009;**66**(6):611-38.
5. Tannen RL, Weiner MG, Marcus SM. Simulation of the Syst-Eur randomized control trial using a primary care electronic medical record was feasible. *J Clin Epidemiol* 2006;**59**(3):254-64.
6. Williams JG, Cheung WY, Cohen DR. Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment. *Health Technol Assess* 2003;**7**(26): iii, v-x, 1-117.
7. Yamamoto K, Matsumoto S, Tada H, et al. A data capture system for outcomes studies that integrates with electronic health records: development and potential uses. *J Med Syst*

1
2
3
4 2008;**32**(5):423–7.
5
6

7
8 8. Yamamoto K, Matsumoto S, Yanagihara K, et al. A data-capture system for post-marketing
9
10 surveillance of drugs that integrates with hospital electronic health records. Open Access J
11
12 Clin Trials 2011;**3**:21–6.
13
14

15
16
17 9. Embi PJ, Jain A, Clark J, et al. Effect of a clinical trial alert system on physician
18
19 participation in trial recruitment. Arch Intern Med 2005;**165**(19):2272-7.
20
21
22

23
24 10. Campbell MK, Snowdon C, Francis D, et al. Recruitment to randomised trials: strategies
25
26 for trial enrollment and participation study. The STEPS study. Health Technol Assess
27
28 2007;**11**:iii, ix-105.
29
30
31

32
33
34 11. Dugas M, Lange M, Müller-Tidow C, et al. Routine data from hospital information
35
36 systems can support patient recruitment for clinical studies. Clin Trials 2010;**7**(2):183-9.
37
38
39

40
41 12. Thadani SR, Weng C, Bigger JT, et al. Electronic screening improves efficiency in
42
43 clinical trial recruitment. J Am Med Inform Assoc 2009;**16**(6):869-73.
44
45
46

47
48 13. Torgerson JS, Arlinger K, Käppi M, et al. Principles for enhanced recruitment of subjects
49
50 in a large clinical trial: The XENDOS study experience. Control Clin Trials
51
52 2001;**22**(5):515-25.
53
54
55

56
57 14. Kristianson KJ, Ljunggren H, Gustafsson LL. Data extraction from a semi-structured
58
59
60

1
2
3
4 electronic medical record system for outpatients: a model to facilitate the access and use of
5
6 data for quality control and research. *Health Inform J* 2009;**15**(4):305–319.
7
8

9
10
11 15. Shim JP. Past, present, and future of decision support technology. *Decis Support Syst*
12
13 2002;**33**(2):111-26.
14

15
16
17 16. Prat N. A UML-based data warehouse design method. *Decis Support Syst*
18
19 2006;**42**(3):1449-73.
20
21

22
23
24 17. Park YT. An empirical investigation of the effects of data warehousing on decision
25
26 performance. *Inform Manag* 2006;**43**(1):51.
27
28

29
30
31 18. Schlaps D, Schmid T. Data warehousing in clinical research and development - From
32
33 clinical data to knowledge portals. *Pharmind* 2004;**66**(5a):637-46.
34
35

36
37
38 19. Grant A, Moshyk A, Diab H, et al. Integrating feedback from a clinical data warehouse
39
40 into practice organisation. *Int J Med Inform* 2006;**75**(3-4):232-9.
41
42

43
44 20. Junttila K, Meretoja R, Seppälä A, et al. Data warehouse approach to nursing
45
46 management. *J Nurs Manag* 2007;**15**(2):155-61.
47
48

49
50
51 21. Rubin DL, Desser TS. A data warehouse for integrating radiologic and pathologic data. *J*
52
53 *Am Coll Radiol* 2008;**5**(3):210-7.
54
55

56
57
58 22. Johnson SB. Generic data modeling for clinical repositories. *J Am Med Inform Assoc*
59
60

1
2
3
4 1996;**3**(5):328-39.
5
6

7
8 23. Nadkarni PM, Brandt C. Data extraction and ad hoc query of an entity—attribute—value
9
10 database. *J Am Med Inform Assoc* 1998;**5**:511-27.
11
12

13
14 24. Anhøj J. Generic design of Web-based clinical databases. *J Med Internet Res*
15
16 2003;**5**(4):e27.
17
18

19
20
21 25. Chen RS, Nadkarni P, Marenco L, et al. Exploring performance issues for a clinical
22
23 database organized using an entity-attribute-value representation. *J Am Med Inform Assoc*
24
25 2000;**7**:475-87.
26
27

28
29
30 26. Dinu V, Nadkarni P. Guidelines for the effective use of entity-attribute-value modeling for
31
32 biomedical databases. *Int J Med Inform* 2007;**76**:769-79.
33
34

35
36
37 27. Corwin J, Silberschatz A, Miller PL, et al. Dynamic tables: an architecture for managing
38
39 evolving, heterogeneous biomedical data in relational database management systems. *J Am*
40
41 *Med Inform Assoc* 2007;**14**:86-93.
42
43

44
45
46 28. Wade TD, Hum RC, Murphy JR. A dimensional bus model for integrating clinical and
47
48 research data. *J Am Med Inform Assoc* 2011;**1**:96-102.
49
50

51
52
53 29. Pardillo J, Mazón JN. Model-driven development of OLAP metadata for relational data
54
55 warehouses. *Comput Stand Interfac* 2012;**34**(1):189-202.
56
57
58
59
60

- 1
2
3
4 30. Hettler M. Data mining goes multidimensional. *Health Inform.* 1997;**14**(3):43-6, 48, 51-6.
5
6
7
8 31. Gordon BD, Asplin BR. Using online analytical processing to manage emergency
9
10 department operations. *Acad Emerg Med* 2004;**11**(11):1206-12.
11
12
13
14 32. Kimball R, Reeves L, Ross M, et al. *The Data Warehouse Lifecycle Toolkit*. New York:
15
16 John Wiley, 1998.
17
18
19
20
21 33. Wang D, Peleg M, Tu SW, et al. Representation primitives, process models and patient
22
23 data in computer-interpretable clinical practice guidelines: a literature review of guideline
24
25 representation models. *Int J Med Inform.* 2002 **18**;68(1-3)
26
27
28
29
30
31 34. Weng C, Tu SW, Sim I, et al. Formal representation of eligibility criteria: A literature
32
33 review. *J Biomed Inform* 2010;**43**(3):451-67.
34
35
36
37
38 35. Lonsdale DW, Tustison C, Parker CG, et al. Assessing clinical trial eligibility with logic
39
40 expression queries. *Data Knowl Eng* 2008;**66**(1):3-17.
41
42
43
44 36. Tu SW, Peleg M, Carini S, et al. A practical method for transforming free-text eligibility
45
46 criteria into computable criteria. *J Biomed Inform* 2011;**44**(2):239-50.
47
48
49
50
51 37. Sordo M, Boxwala AA, Ogunyemi O, et al. Description and status update on GELLO: a
52
53 proposed standardized object-oriented expression language for clinical decision support. *Stud*
54
55
56
57 *Health Technol Inform.* 2004;**107**:164-8.
58
59
60

1
2
3
4 38. Séroussi B, Bouaud J. Using OncoDoc as a computer-based eligibility screening system
5
6
7 to improve accrual onto breast cancer clinical trials. *Artif Intell Med* 2003;**29**(1-2):153-67.
8
9

10 39. CDISC ASPIRE: Integration of clinical research and EHR: Eligibility coding standards.
11
12
13 http://crisummit2010.amia.org/files/symposium2008/S14_Niland.pdf (accessed 31 Mar
14
15
16 2012).
17
18
19

20 40. CDISC Study Design Model: SDM-XML Version 1.0.
21
22
23 <http://www.cdisc.org/stuff/contentmgr/files/0/8c85b168e80d6834ded59339b55fdb7/misc/cd>
24
25
26 isc_sdm_xml_1.0.pdf (accessed 31 Mar 2012).
27
28
29

30 41. US. National Cancer Institute (NCI). caMATCH.
31
32
33 <https://cabig.nci.nih.gov/community/tools/caMATCH> (accessed 31 Mar 2012).
34
35
36

37 42. Jagannathan V, Mullett CJ, Arbogast JG, et al. Assessment of commercial NLP engines for
38
39 medication information extraction from dictated clinical notes. *Int J Med Inform*
40
41 2009;**78**(4):284-91.
42
43
44

45
46 43. Pakhomov S, Weston SA, Jacobsen SJ, et al. Electronic medical records for clinical
47
48 research: application to the identification of heart failure. *Am J Manag Care*
49
50 2007;**13**(6):281-8.
51
52
53

54
55
56 44. Yamamoto K, Yamanaka K, Hatano E, et al. An eClinical trial system for cancer that
57
58
59
60

1
2
3
4 integrates with clinical pathways and electronic medical records. Clin trials 2012 9: 408-417.
5
6

7
8 45. Matsumura Y, Kuwata S, Yamamoto Y, et al. Template-based data entry for general
9
10 description in medical records and data transfer to data warehouse for analysis. Stud Health
11
12 Technol Inform 2007;**129**(Pt 1):412–416.
13
14

15
16
17 46. World Health Organization (WHO). International Classification of Diseases (ICD).
18
19
20 <http://www.who.int/classifications/icd/en/> (accessed 31 Mar 2012).
21
22

23
24 47. Yamazaki T, Yamori M, Yamamoto K, et al, Risk of osteomyelitis of the jaw induced by
25
26 oral bisphosphonates in patients taking medications for osteoporosis: A hospital-based cohort
27
28 study in Japan, Bone 2012: 51(5) 882–887.
29
30
31

32
33
34 48. Fellows JL, Rindal DB, Barasch A, et al. ONJ in two dental practice-based research
35
36 network regions. J Dent Res 2011;**90**(4):433-8.
37
38

39
40 49. Vestergaard P, Schwartz K, Rejnmark L, et al. Oral bisphosphonate use increases the risk
41
42 for inflammatory jaw disease: a cohort study. J Oral Maxillofac Surg. 2012 70(4):821-829.
43
44

45
46
47 50. Shortliffe EH, Cimino JJ. Biomedical Informatics: Computer Applications in Health Care
48
49 and Biomedicine. Health Informatics series. Springer, 2006.
50
51

52
53
54 51. McFadden E. Management of data in clinical trials. 2nd ed. Hoboken, NJ:
55
56 Wiley-Interscience, 2007.
57
58
59
60

- 1
2
3
4 52. Zhengwu L, Jing S. Clinical data management: current status, challenges, and future
5
6
7 directions from industry perspectives. *Open Access J Clin Trials* 2010;**2**: 93–105.
8
9
10
11 53. Data Management Association (DAMA) International. Data management body of
12
13 knowledge. <http://www.dama.org/i4a/pages/index.cfm?pageid=3364> (accessed 31 Mar 2012).
14
15
16
17 54. MedDRA MSSO. MedDRA MSSO. <http://www.meddramsso.com/> (accessed 31 Mar
18
19 2012).
20
21
22
23
24 55. US National Library of Medicine. SNOMED clinical terms,
25
26
27 http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html (accessed 31 Mar 2012).
28
29
30
31 56. Clinical Data Interchange Standards Consortium (CDISC). Study data tabulation model
32
33 (SDTM). <http://www.cdisc.org/sdtm> (accessed 31 Mar 2012).
34
35
36
37 57. Huff S. Development of the logical observation identifier names and codes (LOINC)
38
39
40 vocabulary. *J Am Med Inform Assoc* 1998;**5**(3):276-92.
41
42
43
44 58. Health level seven (HL7). <http://www.hl7.com/> (accessed 31 Mar 2012).
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4 **FIGURE CAPTIONS**
5
6
7

8 **Figure 1.** Data model for our EMR retrieval system.
9

10
11 **Figure 2.** Example SQL to create the target patient list.
12

13
14
15 **Figure 3.** Example SQL to flag the target patient report for investigator confirmation.
16
17

18
19 **Figure 4.** Schema of data collection and confirmation.
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60