의학석사 학위논문

HealthPro: Designing Search Strategies for Retrieving Data from Clinical Data Warehouse

HealthPro: 임상 데이터 웨어하우스에서 데이터 탐색을 위한 검색 전략 설계

2013 년 2 월

서울대학교 대학원

의학과 의료정보학 협동과정

조 용 래

ABSTRACT

Reusing the data collected in electronic medical record system (EMRs) is essential to improve clinical research efficiency. But, it is difficult to identify patients who meet research suitable criteria and collect the necessary information from EMRs because the data collection process must include various type of information stored in EMRs and integrate various techniques, including the development of a data warehouse. We designed a data model optimized for patient identification and for the collection of necessary information from EMRs, and provide the way to search a criteria in the data model. This research aimed to demonstrate an retrieval system(HealthPro) and an example of a hospital-based data that used the HealthPro to identify suitable patients.

The searching system uses data from 550,852 patients, which were imported into clinical data warehouse. The search condition was structured in tree for clinical research, and the tree structure consist of patient demography, diagnosis, laboratory test, clinical document. It enables to search detailed extraction in conditions of various clinical terms. Semantic search with questionaries of clinical documents using MDRS(MetaData Repository System) and drug treatment pattern for patients in particular period are well performed. The algorithm is developed for converting EAV format to table format when extracting data in data warehouse. Query performance averagely recorded 0.9945 seconds of response time with one to ten queries.

In this work, HealthPro performed to design the data model and

develop it for retrieving optimized clinical data. We provide the way to search a criteria using hierarchical tree and standardization terminology in the system. Also meta search is available by using controlled vocabulary for identifying patient who meet expanded criteria. HealthPro may be useful by providing a system package supported from data collection to analysis. Additionally, HealthPro plan to include analysis module with R statistics tool and can provide simple calculation about the result of retrieving clinical data.

keywords : Clinical Data warehouse, Data Search System, Clinical Research
Student Number : 2011-21959

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1. Introduction

As information technology has been applied in clinical field, hospitals have been developed electronic medical records systems (EMRs) all over the world. Many clinical data are already stored in EMR. Reusing the data collected in EMRs is essential to improve clinical research efficiency. [1, 2] Patient information of EMRs used in observational research to conduct surveillance for abnormal drug reactions, to recruit patients for clinical trials and to expand knowledge about cancer diagnoses and treatments. [3, 4, 5]

However, it is not easy to identify patients with eligibility criteria and collect necessary information from EMRs. [1] To identify patients with eligibility criteria, it is necessary to obtain various types of information stored in EMRs by subject, for example, diagnosis, laboratory test and medications. [1, 6]

However, the EMR database is optimized to show data on individual patient and the current EMRs does not provide retrieval function to search multiple patients. So, Data warehouses are essential components for clinical research. EMR data must first be warehoused to facilitate clinical research analysis efficiently. [6, 7, 8] But, Clinical data warehouse modelling is difficult and time consuming because of the complexity of the medical knowledge involved. [9] Additionally, Developing data warehouse cost amount to much in hospital. For that

reason, we describe three major issues concerning clinical research.: Universal data warehouse development for clinical research, retrieval interface for identifying patient and data extraction for analysis.

Over the past few years a number of data warehouses have emerged for clinical research. Such systems consist of a framework supporting multiple terminologies, for example, ICD9, RxNorm, SNOMED and provide data extraction function for analysis. [10, 11, 12, 13, 14] However, the suitable criteria in clinical documents(i.e., case report forms in the case of a clinical trial, or on survey forms, questionnaires) are not yet completely standardized. Current suitable criteria are written in a clinical document that cannot be computationally processed. [15]

To comprehensively and efficiently collect information from EMRs about patients participating in clinical research, we designed a data model optimized for patient identification and for the collection of necessary information from EMRs for clinical research, and provide the way to search a criteria in clinical documents systematically. This research aimed to demonstrate an retrieval system(HealthPro) and an example of a hospital-based data that used the HealthPro to identify suitable patients.

2. METHODS

2.1 Search Strategy

HealthPro developed for clinical studies to optimize the search in the data warehouse. HealthPro architecture was designed by considering the extensibility of the system, and developed data model for retrieving clinical data. EMRs store various types of information, integrating billing, pharmacy, radiology, laboratory information and others. We identified 4 categories that are useful for clinical research: form, medication, diagnosis, laboratory test.

2.2 System Architecture

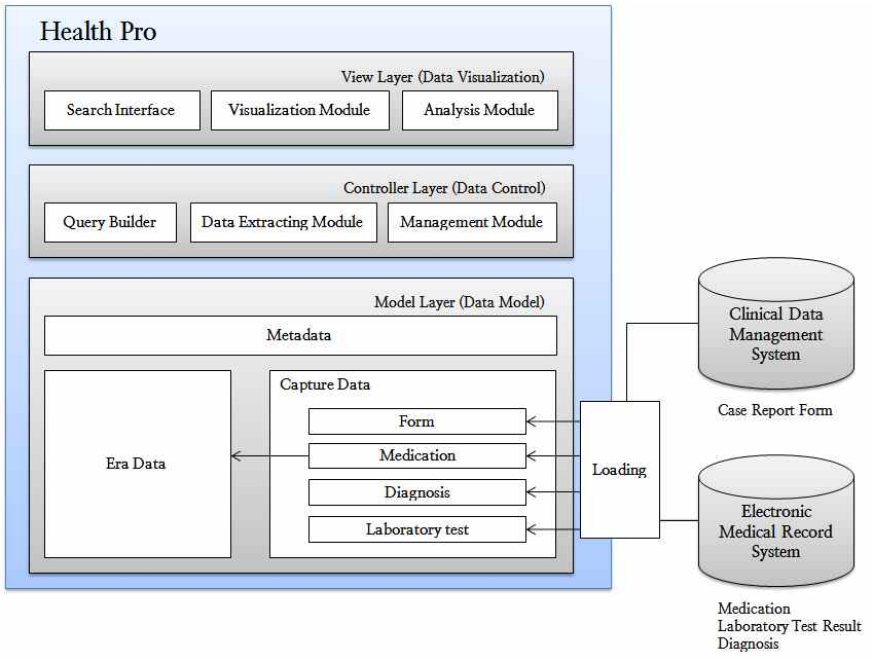

Figure 1. HealthPro System Architecture

The basis structure of the system was designed using MVC pattern. MVC pattern divided into three areas: Model, View, Controller. this pattern has the advantage that the layer of data control and visualization were divided.

Clinical data was structured and stored into HealthPro in model layer. In case of form data, Data concept is managed in conduction with the metadata repository system. Other data concept were used controlled vocabulary.

Controller layer consist of management module, query builder and extract module. The management module controls operating system in HealthPro and query builder can be retrieved clinical data. the extract module stored retrieved data into a text file for download.

View layer has a user interface that allows researcher to select the conditions in HealthPro. Visualization module enables them to view the result of search as graphs, tables.

When researcher select conditions and run through the web interface, this system shows the basic information such as age and sex of the patient. Also various data of patient retrieved by conditions can be extracted.

2.3 Data Model

HealthPro data model for clinical research was composed of three major regions : Observation region for storing the capture data, metadata region for semantic terms of data concept and Era regions for data patterns such as drug patterns. (Figure 2)

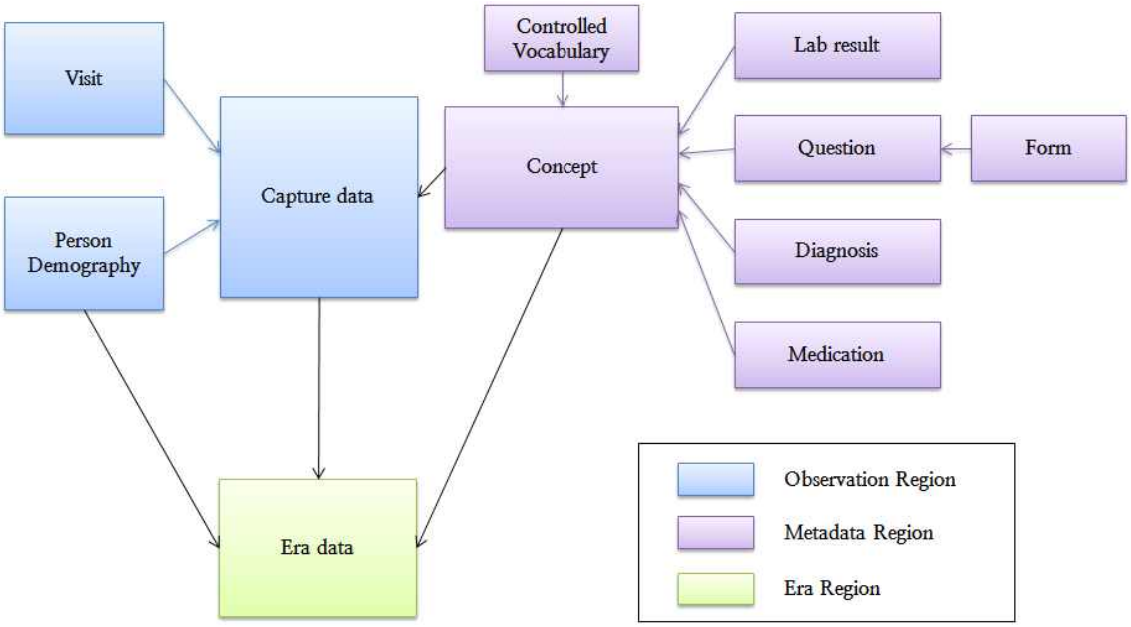

Figure 2. HealthPro Concept Data Model

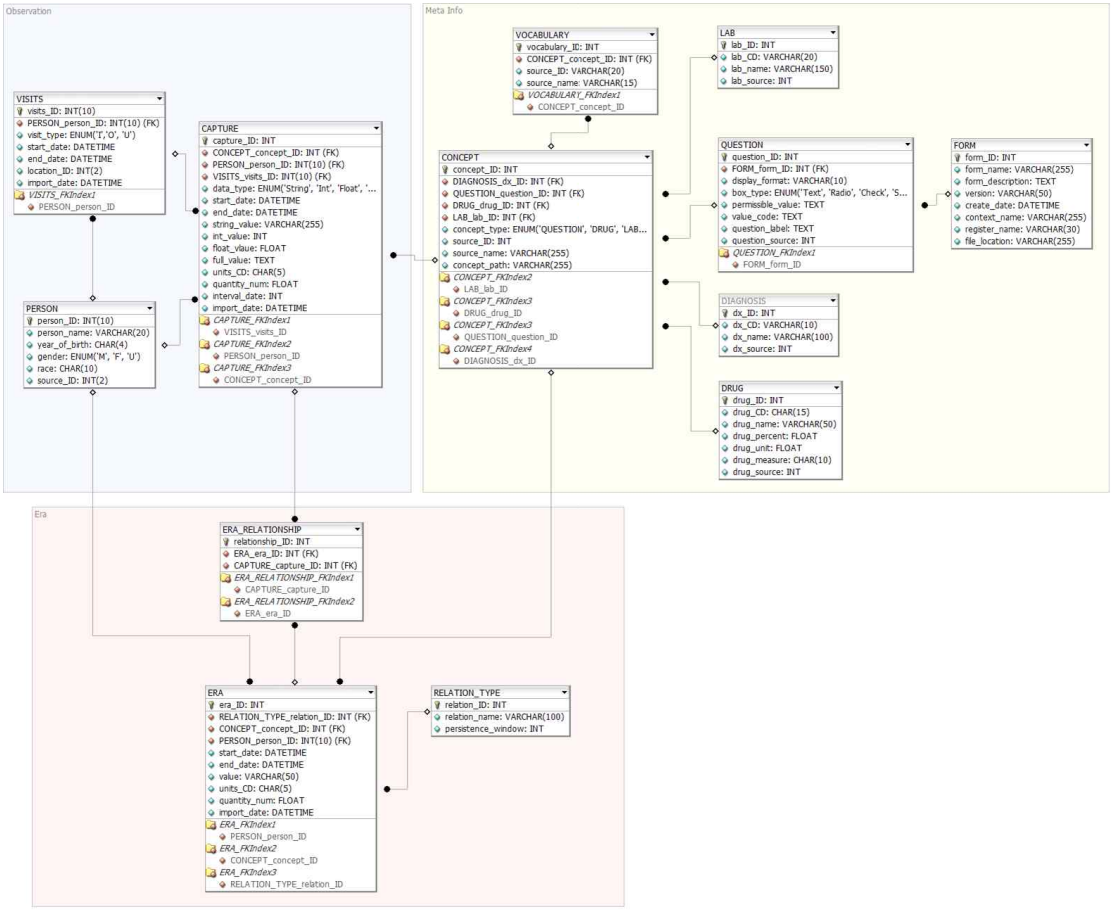

Figure 3. HealthPro Logical Data Model

1) Observation Region

Region in the underlying HealthPro model is to follow the principle of "Star schema". This schema revolved around observation data of patient from EMR and stored demography, hospital visit and captured data of patient.

Observation data was stored as EAV(Entity-Attribute-Value) model. Star schema in the structure of the EAV model is often implemented

in the EMR system. That was to improve the efficiency of the query by applying the EAV format. [16]

2) Metadata Info Region

Metadata region was designed for managing systematically and retrieving concepts in clinical studies. this regions managed diagnosis, medication and laboratory test results terms from EMR and data element of form from CDMS(clinical document management system).

In case of clinical document, CDMS was used for managing electronic documents such as eCRF. CDMS was managed clinical documents by hierarchy structure of Form-Section-Question. this hierarchy structure enable search to find easier. DB schema regarding clinical document was designed using this hierarchy in HealthPro. This system can be retrieved semantic search using concept of meta table. For example, Of sex and gender, the two words mean the same thing, but when used in different sense in clinical document. In this case, Source_ID of meta table was able to search by including terms with the same meaning.

3) Era Region

In era region, OMOP(Observational Medical Outcomes Partnership) [17] data model was employed to view medication spectrum such as drug treatment patterns about patients. OMOP evaluates the performance of various analytical methods identifying drug-outcome

associations across multiple disparate observational data sources. Era data was calculated in accordance with the OMOP policy on period window size. Additionally, the work for identifying the strength of drug prescribed to patients was performed. The detail description was explain by referring figure 6.

	Attribute(OMOP)	OMOP Data Model	Medication of HealthPro
Drug_exposure	Person ID	O	O
	Drug exposure start date	O	O
	Drug exposure end date	O	O
	Drug concept ID	O	O
	Drug exposure type	O	O
	Source drug code	O	O
	Stop reason	O	X
	Refills	O	X
	Drug Quantity	O	O
	Days supply	O	X
Drug Era	Person ID	O	O
	Drug era start date	O	O
	Drug era end date	O	O
	Drug exposure type	O	O
	Drug concept ID	O	O
	Drug exposure count	O	O

Figure 4. Comparison era region of HealthPro and OMOP model

2.4 Data Collection and Normalization

Data concept was collected by classifying diagnosis, medication, form and laboratory test. It is easier to control the concepts and loading EMR to HealthPro model.

Form data was brought from the eCRF of CDMS. eCRF consists of the Form-Section-Question structure that contains several questions in a document and it has in healthPro. So, Form data was loaded directly from CDMS to HealthPro. (Figure 5)

The data concept of form was kept by using MDRS(Meta Data Repository System) connected CDMS. MDRS followed ISO/IEC-11179 and this international standard consist of data element, data element concept, value domain, conceptual domains.

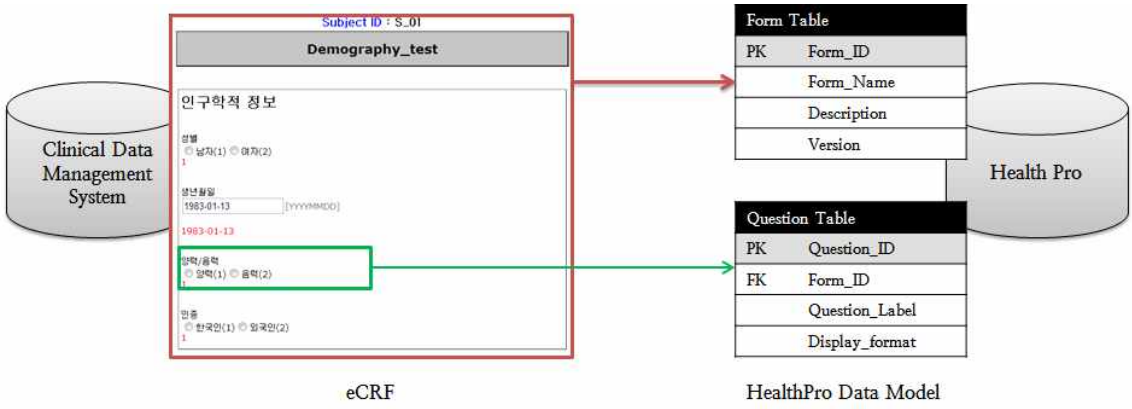

Figure 5. Process of Extracting Clinical Data of CDMS into HealthPro Data Model

Diagnosis data was included the name, code and date in HealthPro. this code was used local code and ICD 10. Laboratory test terms was standardized with LOINC(Logical Obervation Identifiers Names and Codes) [18] code. The attribute of LOINC code about terminology includes short term, long term, related term.

Table 1. Matching LOINC term with SNUH Laboratory test term

LOINC Term	Match Type	Matching Result with SNUH(#)	Matching Result with SNUH(%)
LOINC Short Term	Exact	0	0
	Soft("Term")	195	12.4
LOINC Long Term	Exact	3	0.1
	Soft	196	12.4
LOINC Related Term	Exact	131	8
	Soft	218	13.8

* LOINC Term(70,689), SNUH Laboratory test term(1,572)

Table 1 showed that local code of SNUH matched 3 attributes of LOINC. As a result, there is no one in matching Short term and only 3 terms matched in long terms . It showed low rate overall. To solve this problem, It need manual curation for matching local terms to LOINC.

Table 2. Example of Laboratory Test Concepts in Lab DB Table

Lab_ID	Lab_CD	Lab_name
1	HCSL5105	Syphilis ELISA(IgG,IgM)
2	HCSL5128	Antinuclear Ab(FANA)
3	L0002	Hb [POCT]
4	L0003	Hct [POCT]
5	L0004	D-dimer assay (정량) [POCT]
6	L00051	pH [POCT]
7	L000510	ctCO2(B) [POCT]
8	L000511	PAO2 [POCT]
9	L00052	pCO_2 [POCT]
10	L00053	pO_2 [POCT]

Patient's medication were reconstructed using persistence window in

era data. Drug era below illustrates the scenario (Figure 6). To define the drug era A, the timing, duration, overlap, and persistence of the person's prescriptions for drug A must be considered. A2 was filled before the expected completion of A1. A3 was filled before the expected completion of A2. A4 was filled after A3 was completed. but within the persistence window for Drug A. Therefore, the four prescriptions for Drug A will be consolidated into a single drug era(Era 1).

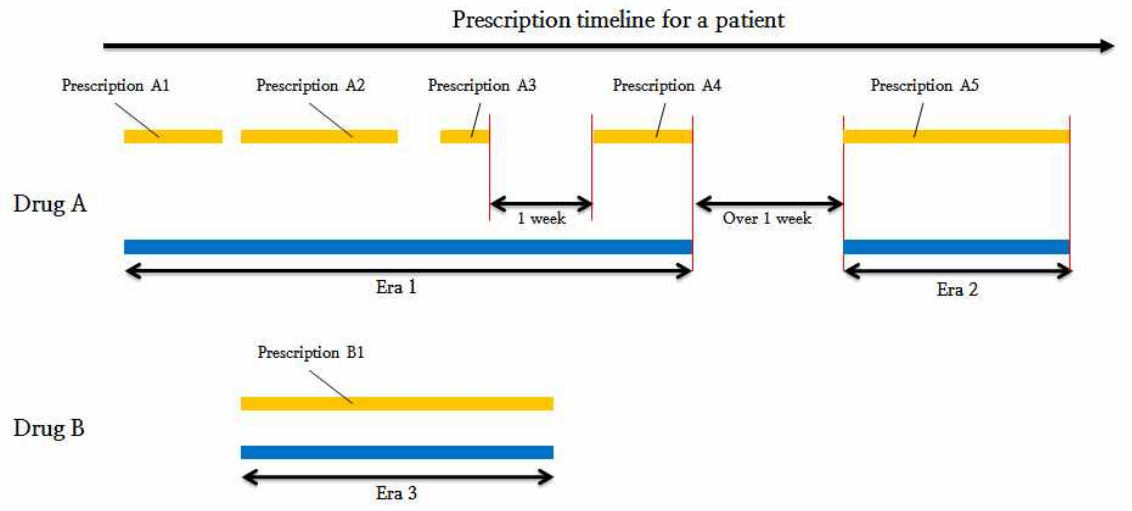

Figure 6. Define Era Data in Prescription Timeline for a Patient

During a period of time, Drug strength of a patient is important. But, It is difficult to know drug strength for including the unit such as ample, bottle and capsule in medication data. Therefore, these units required to convert to drug strength and we was performed.

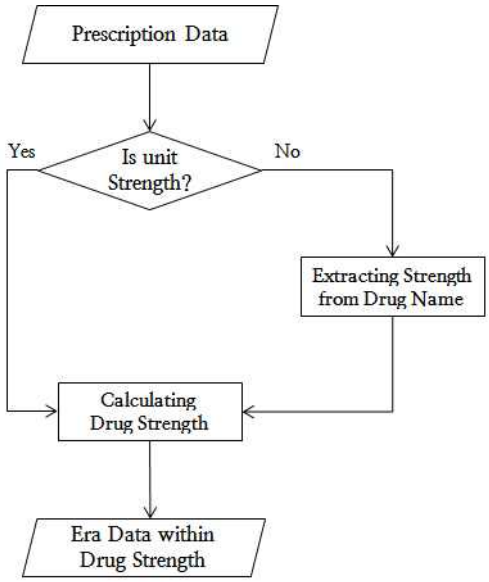

Figure 6. Sequence Diagram for Drug Strength Calculation

By referring to Table 3 giving the unit of measure, drug strength was calculated. If the unit of measure is uncountable, Total quantity of prescribed drug was calculated during a period of time for a patient. If the unit of measure is countable, We extracted strength information from prescribed drug name and then calculated with it.

For example, If patient A prescribed 10 tab of Tamoxifen 20mg for 1 week, drug strength was gotten 200mg(20mg * 10 tab). But, prescribed 100mg of Tamoxifen 20mg, Total quantity used as 100mg for 1 week.

Table 3. Example for DB Table of Era Data

Person	DrugName	unit_CD	Quantity	Start_date	Interval	Strength
A	Tamoxifen 20mg	tab	10	2012-05-01	7	200
A	Albumin 20% 50ml	tab	10	2012-07-14	10	100
A	Tamoxifen 10mg	mg	200	2012-08-30	7	200

From 2000 through 2009, 5,321 drug names from Seoul National University Hospital were identified. The unit of measure in SNUH was used 20 countable units and 9 uncountable.

Table 4. Table of Measure Unit in S Medical Center

	Quantity Unit	Uncountable Unit
1	btl	mg
2	ea	ml
3	cap	g
4	box	iu
5	tab	l
6	via	mcg
7	tub	u
8	set	meq
9	amp	%
10	bag	
11	dos	
12	sup	
13	pil	
14	srg	
15	pkg	
16	ke	
17	kit	
18	bst	
19	vt	
20	can	
21	sgr	
22	wlt	

In case of countable unit, we used regular expressions method for extracting drug strength, and examples of the results is shown in table 4.

Table 5. Extracting Drug Strength from Drug Name

Drug Name	%	Strength
Tacrolimus 0.03%10g	0.03	10
Tacrolimus 0.1% 10g	0.1	10
Tacrolimus 0.1% 30g	0.1	30
Tacrolimus 5mg inj		5
Tadalafil 10mg		10
Tadalafil 20mg		20
Tadalafil 5mg		5
Tadalafil 5mg(STUDY)		5
Taflu/Latanoprost		
Tafluprost 0.0015%	0.0015	
Tagen* cap		
Talc		
Talniflumate		
Talniflumate 370mg		370
Talniflumate370mg		370
Tamipool* inj		
Tamoxifen 10mg		10
Tamoxifen 20mg		20
Tafluprost 0.0015%	0.0015	
Tagen* cap		

3. RESULTS

3.1 Data Scope

For testing the HealthPro data model, 550,752 patients were randomly generated clinical information as well as included about 2 billion observational data, and clinical information of 100 patients stored in HealthPro at SNUH hospital were used. Then, we loaded these data into HealthPro. Data scope was shown in Table 6.

Table 6. HealhPro Data

	Random data		SNUH data	
Patients	550,752		100	
Entity	**Concepts**	**Captured Data**	**Concepts**	**Captured Data**
Diagnoses	12,462	4,000,000	387	985
Forms	16,951	2,117,589	0	0
Medications	5,321	100,000,000	59	2222
Laboratory Test	2,013	120,000,000	172	34058
Total	36,747	226,117,589	618	37265

3.2 System Interface

HealthPro system interface has been implemented on the web and that consist of Navigation trees, conditions, result visualization. Categories in navigation area had tree structure, which included forms, diagnosis, laboratory test and medications.

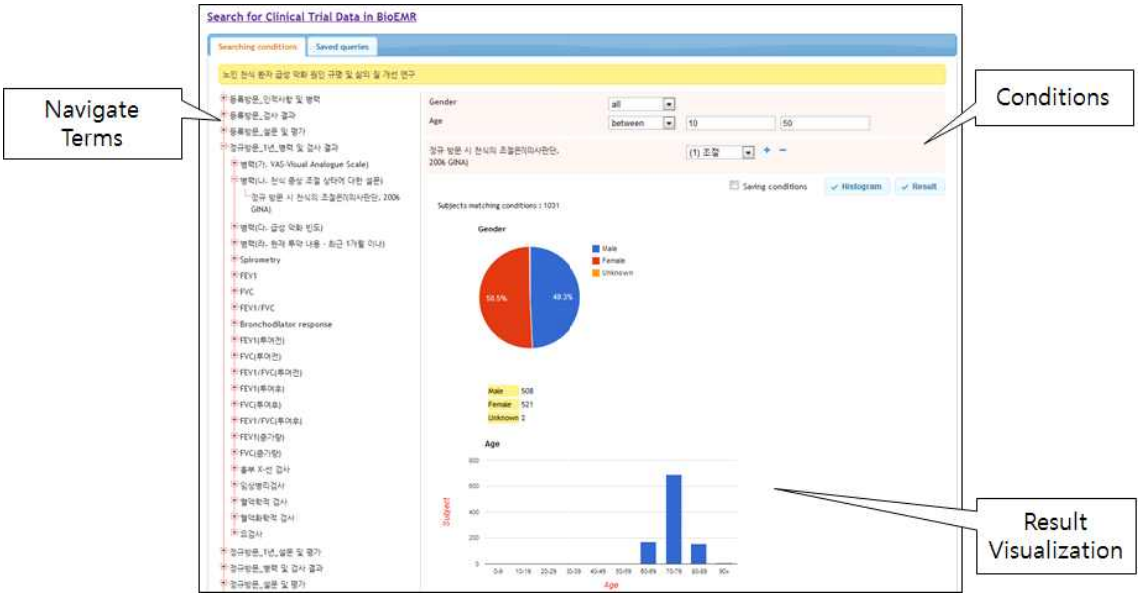

Figure 8. HealthPro Web Interface

Researcher can retrieve the basic demography such as age, gender and concepts in diagnosis and laboratory test. In form category, Documents included data elements were listed in top of form tree position.

Researcher select conditions using navigation tree and input a value in them. The result of selected conditions can be shown as a graph included sex, gender distribution through visual mode. This role of graph about sex, gender is to facilitate overview for clinical research. Result mode provided various data additionally and download text file of patients selected conditions.

Figure 9. Download Result Screenshot for Selected Conditions

The type of download data is text file included meta data of selected conditions. "@Entity" have a unique number of job schedule and total count of columns. "!Entity" enumerated data concept included column names and metadata ID. "#table" represented start and end positions, and data between them is observational data about selected conditions.

```
@Extract JOB ID=20120524_1043
@Column_count=7
!DRUG_INFO_ID=DE_ID:10310
!PT_NO=DE_ID:10311
!DRUG_NM=DE_ID:10312
!START_DATE=DE_ID:10313
!END_DATE=DE_ID:10314
!INTERVAL_DATE=DE_ID:10315
!DOSING_FOR_DAY=DE_ID:10316
!UNIT_OF_MEASURE==DE_ID:10318
#table_start
52  0   2008-03-05  2008-03-06  1   1.0 bag
153 0   2008-03-05  2008-03-06  1   1.0 amp
45  0   2008-03-05  2008-03-06  1   1.0 btl
135 0   2008-03-05  2008-03-06  1   160.0   mg
```

Figure 10. Download Result Text Data

3.3 Query Design

Searching terms were selected by using navigation tree and these was appeared in condition region. Selected term varies conditions depending on the character of it. for example, age term with numeric values have equal or unequal sing to search conditions. Terms with text values have exact or soft match and terms with select values provided these's value domains.

Figure 11. Various Condition Types for Searching Clinical Data

In case of data elements of form, checkbox of meta was shown in conditions. If it was checked, HealthPro search semantic means related to checked term. Medication search consist of drug name, drug code, drug exposure type, drug quantity, drug unit and exposure date. Drug exposure type was selected as persistence window size. for example, If era data was performed for 1 week, drug exposure type appear 1

week.

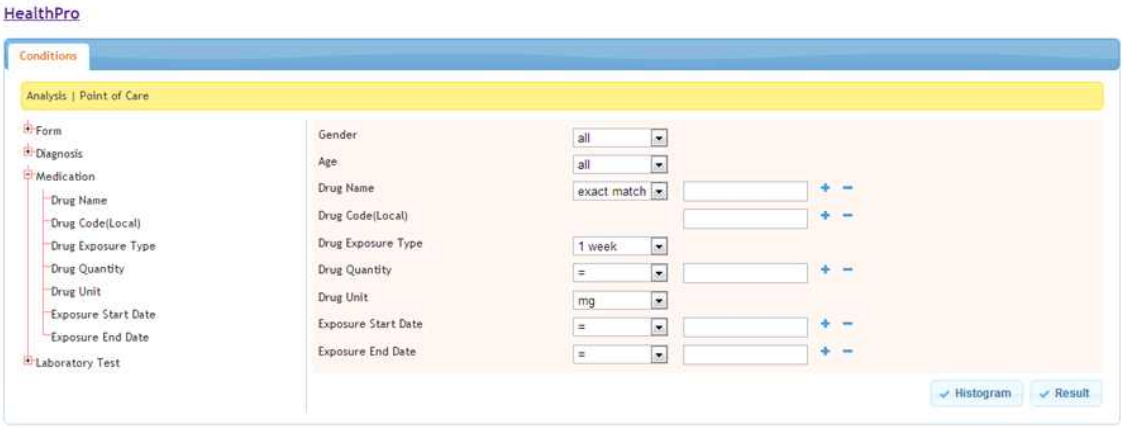

Figure 12. Medication Condition Types for Searching Clinical Data

3.4 Query Performance

HealthPro used star schema for optimizing response time and query from a variety of search criteria in clinical data warehouse and captured data were stored in accordance with the EAV model. We are constructed with MySQL 5.0 version and OS X Server 3.2GHz(Quad-core Intel Xeon Processor).

Query has a consistent pattern to search various conditions and was performed for response time test. Performance test was checked response time while increasing from 1 to 10 conditions and tested in different condition types. The result showed a response time of 1.759 average. But, increasing conditions for search, Increasing the response

time. The maximum response time is 3.6 seconds. The reason to increasing response time is to take times for obtain the intersection of the patients each conditions.

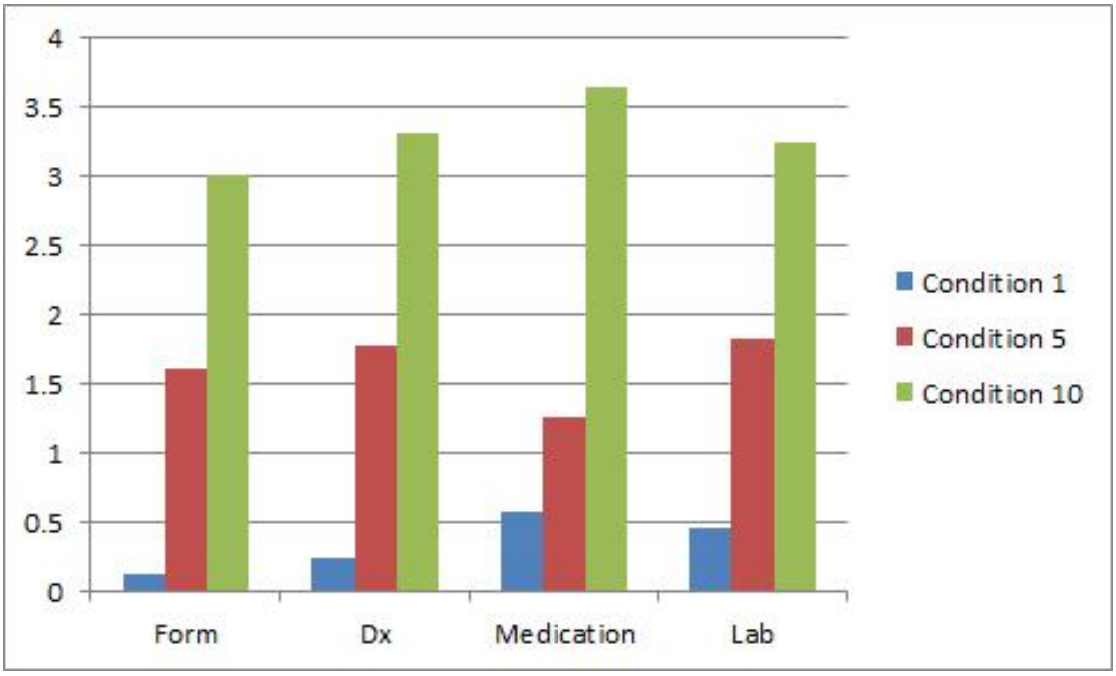

Figure 13. Response Time according to the Condition Query Number

Selected captured data in accordance with the conditions was difficult to extract it, because it is not easy to extract on EAV format. so there is need to convert EAV format to Table format. Table format enable analysis to use in statistical tool such as SPSS. HealthPro included the algorithms for extracting data quickly.

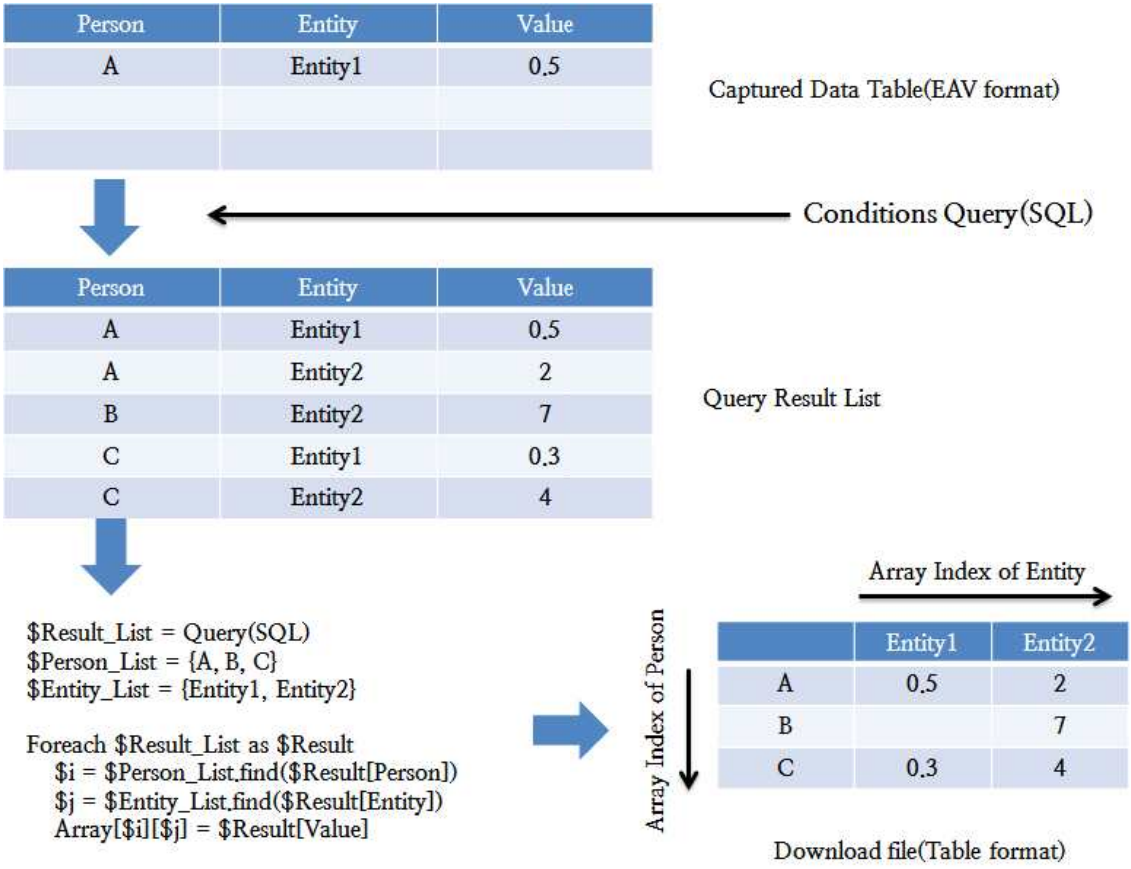

Figure 14. Algorithms for Changing EAV Format to Table Format

The selected group of patients that meet the criteria in the capture data table, then the data you want was added using navigation tree and extracted the results. Figure 14 illustrate the algorithms to change EAV format to Table format. First, The parameter Person_List was allocated the group of patient met the criteria and Entity_List was allocated the concepts you want to extract. Second, While result list loop, Array was allocated values of position located a person in result list. And then this array printed as table format on the web.

4. DISCUSSION

We identified suitable patients for this research and extracted the data necessary for statistical analyses. First af all, we designed a new data model optimized for patient identification. The main features of data model were as follows: (1) To retrieving various types of information in EMRs, we provide the way to search a criteria using hierarchical tree and standardization terminology in clinical documents, as well as other types such as diagnosis, laboratory test. (2) Meta search is available by using controlled vocabulary for identifying patient who meet expanded criteria. (3) To facilitate clinical research analysis such as drug patterns, surveillance for abnormal drug reactions, we designed the HealthPro system based-on OMOP data model.

In this research, we considered whether data were extracted directly from EMRs. But, the EMRs database structure was complicated and required tremendous effort. It was difficult to make precise logical queries for patient identification. Thus, Data is stored using an EAV model. The EAV model has advantages for query, including flexibility of search and ease of storage. However, it requires transforming EAV data into another analytical format before analysis.

The SQL generated by the HealthPro reduce the time required for data retrieval. But, as increasing criteria in HealthPro, The response time is increased. If distributed and parallel processing technology such as NoSQL, Hadoop were applied. Queries performed faster response time. However, This parallel system management was costed too much

to maintain.

Our system can retrieve information that is not in the EMRs. Current EMRs do not store all necessary data for clinical research, including information related to cohort studies. It is necessary to accumulate as much of this information as possible. In the hospital, much of this information does not integrate well with EMRs, including test reports stored only in the departmental system. Thus, We considered the data model for integrating various type of clinical information. It was also easy to interpret the available information due to use the terminology standardization, including ICD10, LOINC in meta layer of HealthPro model. This terminology standardization supports standards-based data entry, data integration, hierarchical concept-based retrieval and data interoperability.

Observations about a patient are recorded within a specific time range(defined by start and end dates), regarding a specific concepts, such as a laboratory test or medication. But, it is difficult for researcher to know specific patterns concerning drug admission period about a patient for clinical research. Thus, we designed “drug era” table based on OMOP model. It can be used to assess the feasibility and utility of using observational data to identify and evaluate associations between drugs and health-related conditions such as laboratory test.

Currently, most clinical research studies that use data from EMRs are planned according to the concept that the primary use of EMRs is for clinical practice and a secondary use is for clinical research.

Therefore, most researcher attempt to convert EMR data to the suitable format at the data collecting stage for analysis. To reduce the burden of clinical analysis, HealthPro may be useful by providing a system package supported from data collecting to data analysis for clincal research. HealthPro has been planned to include analysis module with R statistics tool and can provide simple calculation about the result of retrieving clinical data. We believe an efficient HealthPro and standardized data model are essential to facilitate clinical research.

5. REFERENCE

1. Embi PJ, Payne PR. Clinical research informatics: challegnes, opportunities and definition for an emerging domain. *Journal of the American Medical Informatics Association : JAMIA.* 2009; 16:316-27.

2. Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics.* 2008;77(5):291-304.

3. Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. *Medical care research and review : MCRR.* 2009;66(6):611-38.

4. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association : JAMIA.* 2007;14(1):1-9.

5. Savova GK, Olson JE, Murphy SP, Cafourek VL, Couch FJ, Goetz MP, et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. Journal of the American Medical Informatics Association : JAMIA. 2012;19(e1):e83-e9.

6. Ledbetter CS, Morgan MW. Toward best practice: leveraging the electronic patient record as a clinical data warehouse. *Journal of healthcare information management : JHIM.* 2001;15(2):119-31.

7. Myers DL, Burke KC, Burke JD, Jr., Culp KS. An integrated data warehouse system: development, implementation, and early outcomes. *Managed care interface.* 2000;13(3):68-72.

8. Breen C, Rodrigues LM. Implementing a data warehouse at Inglis Innovative Services. *Journal of healthcare information management : JHIM.* 2001;15(2):87-97.

9. Wade TD, Hum RC, Murphy JR. A Dimensional Bus model for integrating clinical and research data. *Journal of the American Medical Informatics Association : JAMIA.* 2011;18 Suppl 1:i96-102.

10. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium.* 2007:548-52.

11. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium.* 2009;2009:391-5.

12. McConnell P, Dash RC, Chilukuri R, Pietrobon R, Johnson K, Annechiarico R, et al. The cancer translational research informatics platform. *BMC medical informatics and decision making.* 2008;8:60.

13. Patrick JD, Nguyen DH, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association : JAMIA.* 2011;18(5):574-9.

14. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *Journal of the American Medical Informatics Association : JAMIA.* 2010;17(6):652-62.

15. Yamamoto K, Sumi E, Yamazaki T, Asai K, Yamori M, Teramukai S, et al. A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research. *BMJ open.* 2012;2(6).

16. Brandt CA, Morse R, Matthews K, Sun K, Deshpande AM, Gadagkar R, et al. Metadata-driven creation of data marts from an EAV-modeled clinical research database. *International journal of medical informatics.* 2002;65(3):225-41.

17. OMOP. http://omop.fnih.org

18. LOINC. http://loinc.org

국문 초록

임상 연구를 하기 위해서 전자의무기록(EMR)을 이용하는 것은 필수적이다. 그러나 EMR에서 임상 연구를 위해 조건에 맞는 환자를 찾고 데이터를 모으는 것은 쉽지 않다. 왜냐하면 임상 데이터 웨어하우스 구축은 물론 EMR의 다양한 데이터를 관리하기 위해서는 다양한 기술이 필요하기 때문이다. 그래서 우리는 EMR로부터 데이터를 모으고 조건에 적합한 환자를 찾기 위해 최적화된 데이터 모델을 설계하였다. 그리고 이 모델에서 데이터 조건 검색을 위한 방법을 제공하고자 한다. 이 연구의 목적은 검색 시스템(HealthPro)를 구성하고, 조건에 적합한 환자를 찾기 위해 병원 기반 데이터를 이용하여 예제를 적용해 보는 것이다.

임상 데이터 웨어하우스 구축을 위해 MVC Pattern을 이용하여 시스템을 설계하였다. 임상 데이터를 관리하고 검색을 용이하게 하기 위해 HealthPro 데이터 모델을 설계하였다. 데이터 모델은 EMR Data를 저장하는 Observation 영역과 임상 용어를 관리하는 Metadata 영역, 약물처방패턴 등을 검색하여 알 수 있게 해주는 Era 영역으로 구성하였다. 또한 의료기관의 EMR데이터를 임상 데이터 웨어하우스로 변환하기 쉽게 하기 위해 EMR 데이터구조를 설계하였다.

환자 550,852명에 해당되는 데이터를 생성하여 임상 데이터 웨어하우스에 저장하였고, 검색 시스템을 구축하였다. 임상 연구를 위한 검색 전략으로 환자 기본정보, 진단, 검사, 임상문서의 정보를 트리구조로 구성하여 검색 항목을 선택할 수 있도록 하였다. 검색 문항의 특징에 따라 검색 조건을 다르게 구성하여 정교한 검색이 가능하도록 하였으며, 임상 문서의 문항에 대해서는 MDRS(MetaData Repository System)을 이용하여 의미론적인 검색과 관리가 가능하도록 하였다. 또한 일정기간 내 환자의 약물 패턴 등을 검색할 수 있도록 하였다. 검색된 환자의 데이터 추출에 관해

서는 EAV 형식의 데이터 모델에서 Table 형식의 구조로 변환하는 알고리즘을 적용하였다. 검색 성능으로는 검색 조건 1개에서 10개까지의 평균 응답시간이 1.759초의 결과를 얻었다.

본 연구는 데이터 웨어하우스에서 임상 연구에 최적화된 검색을 위해 데이터 모델 설계하고 개발하였다. 우리는 이 시스템에서 계층 구조와 용어 표준화를 통해 검색을 위한 다양한 방법을 제시하였다. 또한 통제어휘를 통해 메타 검색이 가능하도록 했다. HealthPro는 데이터 수집부터 분석까지 하나의 패키지 형태로 시스템을 구성하여 제공함으로써 임상 연구를 함에 있어 유용하게 사용될 수 있다. 추후 R과 같은 통계 프로그램과 연동하여 검색된 데이터를 바로 분석까지 가능하도록 개발할 예정이다.

주요어 : 임상 데이터웨어하우스, 데이터 검색 시스템, 임상 연구
학 번 : 2011-21959

감사의 글

논문을 마치고 “감사의 글”을 보니, 지난 5년간 SNUBI 연구실에서 있었던 많은 사람들과 프로젝트들이 생각납니다. 그 동안 연구실에서 웃고 울고 힘들어하며 공부한 시간들이 쌓이고 쌓여 너무나도 값진 추억과 힘이 되었네요. 아직도 Medical Informatics 분야에서 알아야할 것들이 너무나도 많은데, 이렇게 석사과정을 마치게 되니 무언가 많이 아쉽기도 합니다. 김주한 교수님께서 부족한 저에게 공부할 수 있는 기회를 주시고, 많이 가르쳐 주신 덕분에 이렇게 결실을 맺게 되었습니다.

5년 동안 SNUBI 연구실의 많은 선배님, 후배들 덕분에 많이 배울 수 있었습니다. BioEMR 팀장이셨던 박유랑 박사님, 처음 연구실에 들어왔을 때부터 많은 도움을 주시고, 조언해 주셔서 감사합니다. 정말 많은 것들을 배웠습니다. Decipher 팀장이신 정제균 박사님, 김지훈 박사님, 고생 많으셨습니다. 덕분에 Genome decipher 분야에 대해서 시야를 가지게 되었고, 흥미로운 분야라는 것을 알게 되었습니다. 그리고 정말 다재다능한 찬희형, 존경합니다. 기술에 대한 도전과 열정은 정말 본받고 싶습니다. Sequence alignment와 Pathway 분야에 많은 훈련을 시켜주신 희준형, 저의 수많은 고민 상담을 진지하게 들어주신 수연누나, Co-work 많이 하고 싶었지만, 많이 도와드리지 못했던 영지누나, Ontology 분야에서 도움을 주셨던 정상원 박사님, 의학적 질문에 도움을 많이 주셨던 이계화 선생님, 연구자로서의 자질, 갖추어야 할 마음가짐 등 많은 부분을 알려주신 도균형, 연예상담 전문 정용형 정말 감사드립니다.

그리고 나의 동기, 혜현, 프로젝트에 대해 같이 고민하고, 토론하고 했던 많은 시간들, s수연 누나와 같이 관악캠퍼스에서 수업 들었던 추억들, 고맙다. 친구야. 톰과 제리 같은 사이였다고는 하지만, 덕분에 연구실에서

많이 웃고 힘낼 수 있었던 것 같아. 고마워. s수연누나, 누난 정말 능력자라는 거 아시죠? 그 수많은 프로젝트 다 감당해 내시고, 성과도 내시는 것 보면 대단하다는 생각이 들어요. 한 때 막내였던 수현이, 샌프란시스코에서 약물 분석한다고 방 안에서 랩탑 하나 들고 머리 싸매던 때가 생각나는구나. 박사과정도 잘 마치길 기도할게. 그리고 또 하나의 능력자 희원씨, 갈수록 늘어나는 지식에 감탄해요. 그리고 축하드려요! 말은 많지만 능력 또한 많은 임재현, 옆자리 파트너였던 병희, MD-PhD 과정 무사히 마쳐서 많은 사람들을 도울 수 있는 날이 올 테니 그때까지 힘내! Genomancer 프로젝트를 이끄느라 정말 수고 많은 준희, 장애라는 편견을 다시 생각하게 만들고, 주어진 과제에 언제나 최선을 다하는 모습이 너무 예쁜 백수연, 인턴으로 만나 짧은 인연 임영균, 나중에 멋진 모습으로 다시 봤으면 좋겠다.

연구실 생활하면서 나에게 소중한 영조, 대학교 시절부터 함께 해 왔던 많은 시간들, 원주에서 함께 여행했던 시간들 모두 소중히 간직할게. 그리고 지금은 함께 하지 못하고 있지만, 가장 힘든 시기에 큰 힘이 되어주었던 김근, 도영이형, 조만간 찾아뵐게요. IT 개발에 능숙하신 장태훈 선생님, 떠오르는 샛별 김승호, 묵묵히 많은 공부를 하고 계신 고인석 선생님, R 그래프의 끝판 왕 한현욱 선생님, 통계 분석의 마스터 손경아 선생님, 많은 대화는 못 나눴지만, 뜻하는바 모든 일들이 잘되길 기도할게요. 졸업한 선민이도 회사에서 잘 적응하고 있겠지?

그리고 무엇보다 사랑하는 미형이, 바쁜 일정에서도 투정보다 먼저 나를 이해해 주고, 챙겨주고, 생각해 주어 너무 고마워. 앞으로도 우리 계속 함께 하자. 나에게 물질적, 정신적 지주가 되어주신 부모님, 많이 속 썩였던 제가 이렇게 석사 졸업을 하게 되었습니다. 저를 끝까지 믿고 응원해 주셔서 너무 감사드립니다. 마지막으로 지금의 나를 있게 해주시고, 이끌어 주신 하나님께 이 모든 영광을 돌립니다.