博士學位論文

효율적 질병분류를 위한 지식기반 모델

濟州大學校 大學院

컴퓨터工學科

金 美 貞

2017 年 2 月

효율적 질병분류를 위한 지식기반 모델

指導敎授 李 尙 俊

金 美 貞

이 論文을 컴퓨터工學 博士學位 論文으로 提出함

2016 年 12 月

金美貞의 컴퓨터工學 博士學位 論文을 認准함

審査委員長 鄭 錫 燁

委 員 이 상 준

委 員 정 도 현

委 員 김 한 일

委 員 박 경 희

濟州大學校 大學院

2016 年 12 月

A Knowledge-based Model for Efficient Classification of Diseases

Mi-Jung Kim
(Supervised by professor Sang-Joon Lee)

A thesis submitted in partial fulfillment of the requirement for the degree of Doctor of Computer Engineering

2016. 12.

This thesis has been examined and approved.

Thesis director, Hoyoung Kwak
Thesis director, Sang Joon Lee
Thesis director, DoHyeun Kim
Thesis director, Hanil Kim
Thesis director, Chung Hee Park

December 2016

Department of Computer Engineering
GRADUATE SCHOOL
JEJU NATIONAL UNIVERSITY

Contents

List of Figures

List of Tables

국문 초록

효율적 질병분류를 위한 지식기반 모델

제주대학교 대학원

컴퓨터공학과

김 미 정

이 연구는 효율적 질병분류를 위한 지식 모델을 제안하기 위한 목적으로 시도되었다. 순환기 계통의 다빈도 질병을 대상으로 질병명의 상세한 수준에 따라 KCD-7 코드가 부여되는 질병분류모델과 부여된 코드들의 오류를 검토하는 질병분류 규칙을 개발하였다. 개발된 지식모델의 활용가능성과 타당성을 검증하기 위하여 지식기반 시스템을 구현하고 모델의 커버리지를 평가하였다.

질병분류 지식과 질병명의 구조 분석 결과 KCD-7 코드는 해당 질병명이 가진 고유한 속성에 따라 달라지거나 질병명 이외의 다른 요인에 의해 달라졌다. 질병명의 속성에 따른 코드변화는 질병명을 최소 단위의 기본개념과 속성, 속성값으로 구조화하여 질병분류코드를 할당하는 BAVC(Base concept, Attribute, Value and Code) 모델을 통해 공식화하였다. 질병명 이외의 요인에 의한 코드변화는 IF-THEN 규칙으로 정의하였다.

개발된 BAVC 모델은 총 27개이며, 기본개념 27개, 속성 14개, 유효한 속성값 138개 그리고 KCD-7 코드 186개가 사용되었다. 모델을 구성하는 모든 요소에는 SNOMED CT 코드를 매핑하여 요소마다 의미를 부여하였다. BAVC 모델에 따라 최종 290개의 인스턴스를 작성하였다. 질병명 이외의 코드변경 요인에는 동반코드 유무, 산과환자 여부 등이 해당되어 규칙으로 정의하였다.

이러한 지식 모델을 시스템으로 구현한 결과 질병명별 선택적 속성 제시에 따라 상세한 수준의 진단명 입력이 유도되었으며, 이로 인해 질병분류의 정확성과 일관성이 보장되었다. 모델의 타당성 검증에 사용된 임상현장의 진단명은 지식 모델에 의해 88.9% 커버되었고, 커버되지 않은 나머지 11.1%의 진단명은 시스템 독립적인 지식관리 도구에서 해당 모델에 속성을 추가하는 간단한 방식을 통해 사용자 용어를 100% 수용할 수 있었다. BAVC 모델에 의해 저장된 KCD-7 코드들은 사전에 정의한 IF-THEN 규칙에 의해 자동 검토되어 오류가 있을 경우 자동 재분류되었다.

기존의 연구들이 주로 사용자 진단명의 코드화나 질병분류코드의 누락 방지에 의미를 두었다면, 본 연구는 후조합 개념에 의한 질병분류모델과 규칙을 제시함으로써 질병분류의 일관성과 정확성을 보장함은 물론 속성 단위의 의미적 검색과 지식의 재활용 등 정보의 활용성을 강조하였다는데 의의가 있다.

Keyword: Classification of Diseases, Knowledge Model, KCD-7 Coding rules

ABSTRACT

A Knowledge-based Model for Efficient Classification of Diseases

The Graduate School of Jeju National University
The Department of Computer Engineering
Mi-Jung Kim

The purpose of this study is to develop and validate the knowledge-based model for efficient classification of diseases. The disease classification model, which the KCD-7 code is given according to the detailed level of name of disease and the disease classification rules, which review the errors in the code given, were developed with the circulatory diseases having high morbidity rate. To test the applicability and the validity, the knowledge-based system was constructed and the model coverage was evaluated.

The results are as follows.

Based on the structural analysis results of the disease classification knowledge and the disease name, 27 BAVC (Base concept, Attribute, Value and Code) models, which are the knowledge-based model of this study and the rule were developed. BAVC model was summarized in 27 basic concepts, 14 attributes, 138 valid attribute values and 186 KCD-7 codes. Meaning was given to all the terms used by mapping with SNOMED CT code. According to BAVC model, 290 instances were developed finally.

In the results of constructing the knowledge-based model as a system, the

accuracy and consistency of the disease classification code were secured by inducing the input of the diagnosis according to the suggestion of the attribute by step for the disease name. The diagnosis in the clinical site used in the test were covered 88.9% by the knowledge-based model and the whole coverage could be improved to 100% in a simply manner that for the rest 11.1% of the diagnosis, the attributes are added to the relevant model in the independent knowledge management tool of the system. The KCD-7 codes stored were reviewed by the rules and classified again automatically.

While the existing researches focused mainly to assign the interface terminology to code of standard terminology or to prevent the omission of disease classification code, this study has a meaning that it ensures the consistency and accuracy of the KCD-7 code by the independent knowledge -based model by suggesting the disease classification model and rule by the concept of post-coordination and the applicability of the information was enhanced by allowing the semantic search of attribute unit.

Keyword: Classification of Diseases, Knowledge Model, KCD-7 Coding rules

Abbreviations

AUI	Atom Unique Identifier
BAV	Base concept, Attribute and Value
BAVC	Base concept, Attribute, Value and Code
CCM	Clinical Content Model
CEM	Clinical Element Model
CUI	Concept Unique Identifier
DCM	Detailed Clinical Model
EAV	Entity-Attribute-Value
EHR	Electronic Health Record
EMR	Electronic Medical Record
FSN	Fully Specified Name
ICD-10	International Classification of Disease 10th revision
ICD-9-CM	International Classification of Disease 9th revision-Clinical Modification
ICPC	International Classification of Primary Care
KCD-7	Korea Standard Classification of Diseases 7th Revision
KOSTOM	Korea Standard Terminology of Medicine
LOINC	Logical Observation Identifiers Names and Codes
LUI	Lexical Unique Identifier
MeSH	Medical Subject Headings
SNOMED	Systematized Nomenclature of Medicine
SNOMED CT	SNOMED-Clinical Terms
SNOP	Systematized Nomenclature of Pathology
SUI	String Unique Identifier
UMLS	Unified Medical Language System

I. 서론

1. 연구 배경 및 목적

1) 연구 배경

병원정보시스템의 도입은 병원의 업무 환경은 물론 의료정보의 양과 질을 획기적으로 변화시켰다. 처방전달시스템을 시작으로 의료영상관리시스템, 임상병리시스템, 전자의무기록 등이 도입되면서 다양한 형태의 방대한 임상정보가 데이터베이스화되어 정보공유와 활용 측면에서 많은 이점을 제공하고 있다. 전자의무기록의 도입은 의무기록의 물리적 저장 공간을 해결하고, 시간과 공간의 제약 없이 동시에 여러 명이 접속할 수 있는 등 병원 내 정보공유 측면에서는 이점이 두드러지지만 비정형적인 서술방식의 기록이 많고, 표준용어체계의 도입이 낮아 활용가치에 비해 자료의 활용도는 매우 낮은 편이다. 그에 대한 해결 방법으로 임상용어 및 자료의 구조, 임상서식 항목 등을 표준화하기 위한 노력이 활발히 진행되고 있다[1-5]. 이는 환자의 주증상이나 신체검사, 계통검진, 처방 등의 진료기록을 서술방식이 아닌 임상정보모델과 표준용어체계를 사용하여 자료를 표준화 및 구조화된 형태로 일관성 있게 수집하고 정보의 활용성을 높이자는 것이다[1].

의료정보의 표준화는 진료정보의 공유는 물론 국가보건정책 자료로 활용되어 엄청난 부가가치를 창출할 수 있다. 국제적 차원의 표준화 노력은 병원 내 진료정보의 공유 차원을 넘어 의료기관 간, 국가 간, 요람에서 무덤까지 전 생애의 건강기록을 공유할 수 있는 전자건강기록(Electronic Health Record: EHR)이라는 개념을 가져 왔다[2].

보건의료분야에 사용되는 용어의 표준화를 위하여 UMLS (Unified Medical Language System), SNOMED CT(Systematized Nomenclature of Medicine-Clinical Terms), LOINC(Logical Observation Identifiers Names and Codes) 등과 같은 임상 용어체계가 수십 년 전부터 개발되어 왔으며, 임상자료의 구조와 표현을 위해 임상요소모델(Clinical Element Model: CEM), 임상콘텐츠모형

(Clinical Content Model: CCM), 아키타입(Archetype) 등의 임상정보모델이 개발되었다[3,4]. 임상내용을 상세한 수준으로 일관성 있게 수집하기 위해서는 임상정보의 표현방식과 속성, 속성을 표현하는 유효한 값들에 대한 합의가 필요하다[5]. 임상정보모델은 전자의무기록을 구성하는 구조화된 입력도구로 활용되어 제한된 시간에 필요한 정보를 신속하고 정확하게 입력할 수 있도록 도와주지만 서술방식의 입력이나 단일 정보를 선택하는 방식보다 인터페이스가 복잡하여 그 동안 제한적으로 사용되었다. 이러한 모델들은 의무기록에 작성되는 다양한 영역에 대해 개발되었지만, 환자의 진단명에 대해서는 단순 진술문의 형태를 취하고 있다.

환자의 진단명은 임상자료 중 가장 중요한 자료로써 진료활동을 위한 기준이 되며, 보험청구 업무 및 보건통계를 위한 기초자료로 사용된다. 따라서 의무기록에는 정확한 진단명이 기록되어야 하며 동시에 데이터 처리를 위한 질병분류코드 또한 정확해야 한다. 병원 업무가 전산화되면서 의사가 진단명을 화면에 입력하면 질병명과 질병분류코드가 함께 저장되는 방식이기 때문에 의사의 진단명 입력은 질병분류코드의 정확성과 관련이 깊다. 동일한 질병을 가진 환자라도 의사의 성향이나 상황에 따라 질병명을 표현하는 수준이 다를 수 있다. 동일한 위암 환자라도 위암 혹은 진행성 위암, 유문 부위의 진행성 위암 등 세분화 정도가 다른 진단명으로 입력된다. 사실 위암 환자의 진단명을 위암이라고 작성한 것을 틀렸다고 할 수는 없지만, 유문 부위의 진행성 위암이라고 작성하면 더 구체적이고 명확하다. 이처럼 동일한 환자에 대한 진단명이라도 질병분류코드를 부여하는 사람이나 부서마다 차이가 있다.

국내에서는 진단서 및 사망진단서 작성, 보험청구, 보건의료 통계 등에 한국표준질병사인분류 7차 개정판((Korea Standard Classification of Diseases 7th Revision: KCD-7)의 분류코드를 사용한다[6]. 그러나 의사, 보험청구 부서, 병원 통계 부서 등에서 분류한 KCD-7 코드가 부서 간 서로 일치하지 않아 그 동안 문제가 꾸준히 제기되어 왔다[7-11]. 안진하의 연구[7]에 의하면 의사가 작성한 입·퇴원기록지의 코드와 의무기록사가 재분류한 코드 간의 주 진단의 일치율은 81%지만, 기타 진단의 일치율은 47.5%였다. 다른 연구에서는 의사 중 53.5%만이 구체적인 코드로 정확하게 분류하였고 하나의 진단명을 두 개의 코드로 이원분류 해야 하는 경우에는 8.9%만이 정확하게 분류하였다[10]. 의사가 선정한 코드

의 정확성이 낮은 이유는 의사는 주로 행정적인 목적으로 사용되는 KCD-7의 질병분류 원칙과 지침에 대한 지식이 낮고, 질병분류코드를 고려하기 보다는 의학적으로 의미가 있는 수준의 진단명을 주관적으로 판단하여 선정하기 때문이다. 구체적이고 상세한 진단명의 입력은 진료와 연관된 행정 처리는 물론 다양한 의료전문가가 협업해서 환자를 치료하는 과정에 있어서도 매우 중요하다.

부정확한 질병분류코드의 사용은 보험청구 시 보험금 삭감의 원인이 되거나 과다 청구로 인한 병원의 경제적인 불이익을 초래하며, 진료통계 및 임상연구 검색의 결과에 대한 정확성을 보장할 수 없다[11].

외래환자는 대부분이 경증 질환이기 때문에 의사가 입력하는 진단명과 코드는 대부분 정확하지만, 입원환자의 경우에는 중증 질환이 많고, 동반 질환, 합병증 등 질병분류에 영향을 주는 다양한 요소가 존재하기 때문에 코드가 정확하지 않은 경우가 발생한다. 의사가 다양한 요소를 고려하여 질병명을 입력하기 어렵다 보니 퇴원환자의 정확한 질병분류코드를 얻기 위해 특별히 훈련된 질병분류 전문가가 코드를 재분류하도록 하고 있다. 의사가 선정한 질병명을 토대로 보험청구와 분석을 위해 KCD-7 분류기준에 맞는 코드로 재분류를 하는 일은 전문 인력의 고용으로 인한 비용을 초래하거나, 기존의 인력이 수행하더라도 매우 시간 소모적이고 노동집약적인 일로써 부서 간 중복업무를 발생시킨다[11].

정보의 활용에 있어서도 부정확한 질병분류코드는 부정확한 검색결과를 가져온다. 또한 분류체계인 KCD-7은 유사한 질병명들을 하나의 동일한 코드로 처리하기 때문에 질병분류코드만 사용해서는 구체적인 질병명을 구분할 수 없을 뿐 아니라, 질병명이 내포하고 있는 특징이나 속성에 대해서는 의미적 검색이 불가능하다.

이러한 이유로 의료정보관리 차원에서 정확한 질병분류의 체계화 및 관리를 위한 필요성이 꾸준히 제기되어 왔으나[9,12], 이를 해결하기 위한 방법을 제안한 연구는 많지 않다. 일부 연구에서 서술문으로 작성된 질병명을 대상으로 자연어처리 기술 및 기계학습에 의한 질병분류를 제안하였으나[13,14], 국내에서는 서술방식이 아닌 코드화된 질병명을 선택하는 방식이기 때문에 자연어처리 기술은 적합하지 않다. 일부 병원에서는 코드화된 진단명을 선택할 때 해부학적 부위를 명시해야 하는 진단명에 한하여 발병 부위를 별도로 선택하도록 하고 있지만,

임상문서에 작성되는 다양한 자료 항목에 대해 개발된 임상정보모델처럼 질병명의 구조와 의미적 호환성을 보장하기 위한 모델은 개발된 바 없다. 규칙을 적용한 질병분류 연구에서는 입력된 질병분류코드가 이미 정확하다는 가정 하에 코드 간의 문제를 규칙으로 해결하였기 때문에 코드의 부정확성에 대한 근본적인 문제를 해결하진 못하였다[15].

2) 연구 목적

본 연구의 목적은 환자의 진단명과 질병분류코드가 정확하고 일관성 있게 수집될 수 있도록 질병분류와 관련된 지식 모델을 제안하는 것이다. 이에 진단명을 상세한 수준으로 표현할 수 있는 질병분류모델과 분류코드의 오류를 검토하는 질병분류 규칙을 개발한다. 이러한 지식을 기반으로 질병분류 시스템을 구현하여 활용 가능성과 모델의 타당성을 검증한다.

본 연구는 인간의 중재 없이 질병분류코드를 구체적이고 정확하게 부여하면서 질병명이 내포하고 있는 속성 단위의 의미적 검색이 가능한 지식기반 모델을 제시한다는데 의의가 있다. 지식기반의 질병분류 시스템을 사용하면 질병분류에 대한 지식이 낮은 사용자라도 사용자 인터페이스를 통해 질병분류에 대한 지식을 얻을 수 있으며, 상세수준의 질병명을 일관성 있게 수집할 수 있다. 기존의 시스템에서 일반적으로 사용되고 있는 텍스트 검색 기반의 진단명 입력 도구를 개선하여 구체적인 진단명을 단계적으로 선택할 수 있도록 구조화된 입력 방식을 보장하고, 동시에 내부적으로는 질병분류 가이드라인에 맞는 KCD-7 코드를 부여하여 일관성 있는 질병분류를 가능하게 한다. 기존 시스템의 문제점인 부정확한 진단명 입력과 질병분류코드의 사용으로 발생하는 의료정보의 질 저하 문제와 행정적인 목적의 코드 재분류 업무의 효율성을 개선할 수 있다. 상세한 진단명의 수집과 정확한 질병분류는 임상정보의 질과 함께 궁극적으로 의료의 질을 향상시키며, 병원통계 및 국가단위의 통계자료의 신뢰성과 의료정보의 활용가치를 높이는데 기여할 것이다.

2. 연구 내용 및 논문의 구성

1) 연구 내용

본 연구에서는 도메인 지식으로 사용될 질병분류 지식을 문헌과 전문가를 통해 도출하고 도출된 지식을 규칙과 구조로 체계화하여 질병분류모델과 질병분류 규칙을 개발하였다. 개발된 지식모델을 검증하기 위하여 지식기반의 질병분류 시스템을 구현하고 평가하였다. 연구는 다음과 같이 3단계로 진행되었다.

1단계는 질병분류지식을 체계화하는 단계이다. 이를 위해 KCD-7 지침과 질병분류 원칙에 대한 지식을 체계화 하고 해당 지식을 질병분류 규칙으로 정의하였다. 또한 순환기 계통의 질병명을 핵심용어와 핵심용어를 수식하는 수식어로 구분하여 질병명의 구조를 분석하였다.

2단계는 질병분류모델을 개발하는 단계이다. 질병명의 구조분석 결과를 근거로 진단명으로써 의미가 있는 최소 단위의 질병명을 기본개념(base concept)으로 정하고 이러한 질병을 설명할 수 있는 요소들을 속성(attribute)으로 정의하였다. 질병명은 여러 가지 수식어에 의하여 즉 속성 값(value)에 의하여 더 상세해지고 명확해진다. 속성에 대한 속성 값은 기본개념에 따라 달라질 수 있으므로 속성 값 세트(value set)로 제한하였다. 또한 모델에 사용된 모든 용어에는 의미적 해석이 가능하도록 SNOMED CT 코드를 매핑하여 의미를 부여하였다. 질병명은 수식어에 의해 질병분류 코드가 달라지거나 동일해지므로 기본개념, 속성, 속성 값의 용어를 조합하여 KCD-7 코드를 부여하였다. 기본개념별 속성 값의 변동에 따라 질병분류코드도 변동되므로 이렇게 작성된 질병분류모델을 BAVC(Base concept, Attribute and Value) 모델이라고 명명하였다.

3단계는 개발한 지식 모델을 시스템으로 구축하여 평가하는 단계이다. BAVC 모델과 질병분류 규칙의 타당성을 검증하기 위하여 지식기반의 질병분류 시스템을 개발하여 활용가능성을 확인하였다. 시스템에서 BAVC 모델이 의사가 구체적인 질병명을 입력하여 정확한 질병분류코드가 부여되도록 도와준다면, 질병분류 규칙은 성별, 나이, 동반 질환 여부에 따라 달라질 수 있는 분류코드들을 확인하여 재분류해준다. 연구 내용 및 절차는 Fig.1과 같다.

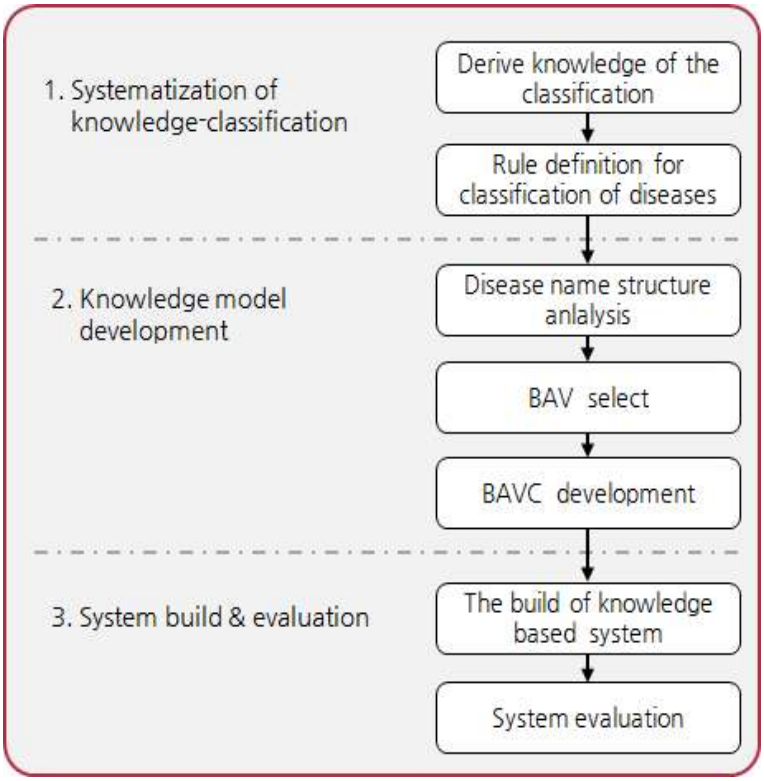


Fig. 3. Contents and procedure of study

2) 논문의 구성

논문의 구성은 다음과 같다.

1장에서는 효율적인 질병분류를 위한 지식기반 모델의 필요성을 언급하기 위해 전자의무기록 시스템 하의 질병명 입력과 질병분류에 대한 현황 및 문제점을 기술하고, 연구 배경과 연구 목적에 대하여 설명하였다. 그리고 연구 내용 및 논문의 구성에 대하여 설명하였다.

2장에서는 정확한 임상정보의 획득과 질병분류의 자동화와 관련된 선행 연구에 대하여 살펴보았다. 먼저 의료분야의 임상 용어체계와 기존에 개발된 임상정보모델, 구조화된 입력도구에 대하여 고찰하고, 정확한 질병분류코드를 얻기 위한 기존 연구와 기법들에 대하여 살펴보았다.

3장에서는 질병분류를 위한 전문지식을 체계화하였다. 기존의 질병분류의 문

제점을 해결하기 위해 필요한 KCD-7의 일반적인 지식을 도출하였다. 순환기 계통의 질병명을 어휘 분석하여 질병명의 구조를 확인하고 질병명을 수식하는 수식어의 속성을 파악하였다. 도출된 지식을 근거로 질병분류에 영향을 주는 요인에 대하여 분류규칙을 작성하였다.

4장에서는 질병명의 어휘 분석 결과를 토대로 해당 질병명이 가질 수 있는 속성과 속성 값을 정의하고 순환기 계통의 다빈도 질병의 KCD-7 코드를 중심으로 질병의 속성에 따라 KCD-7 코드가 할당되는 규칙을 모델로 공식화하였다.

5장에서는 본 연구에서 개발한 BAVC 모델과 질병분류 규칙의 활용가능성을 확인하기 위해 지식기반의 질병분류 시스템을 구축하고 기존 시스템과 비교하였다. 지식 모델의 타당성을 검증하기 위해 모델의 커버리지를 평가하였다.

마지막으로 6장에서는 논문의 결과를 요약하고 연구의 한계점과 향후 연구방향에 대해 제언하였다.

II. 관련 연구

2장에서는 보건의료분야의 대표적인 임상 용어체계를 성격에 따라 구분하고, 미국과 유럽에서 표준으로 인정받고 있는 SNOMED CT와 UMLS 그리고 한국 보건의료표준용어체계인 KOSTOM(Korea Standard Terminology Of Medicine)에 대하여 정리하였다. 그리고 임상정보모델과 정확한 질병분류코드를 얻기 위한 기존의 연구와 기법에 대하여 살펴보았다.

1. 임상 용어체계

1) 임상 용어체계의 종류

보건의료 분야의 임상 용어체계는 성격에 따라 인터페이스 용어체계(interface terminology), 참조 용어체계(reference terminology), 분류체계(classification)로 구분된다[16].

의료정보시스템을 도입한 의료기관에서는 각 의료기관마다 사용하는 고유의 용어체계를 가지고 있다. 병원 고유의 용어체계는 일반적으로 해당 의료기관의 사용자가 필요로 하는 임상용어를 사전에 수집하여 모아놓은 것이다. 이러한 용어들은 전자의무기록과 같은 전산화면에서 일관성 있는 임상정보를 입출력 할 수 있도록 해주기 때문에 인터페이스 용어체계, 입력 용어체계(entry terminology), 혹은 응용프로그램 용어체계(application terminology)라고도 한다[16]. 의료진이 사용하는 친숙한 용어들을 수집하여 만들기 때문에 자연어에 가까우며, 사용자가 선호하는 대표용어, 약어 등이 포함된다. 자연어처럼 보이지만 내부적으로 코드화된 요소와 연결되어 정해진 시스템에서 사용된다. 특정 컴퓨터 프로그램의 그래픽이나 텍스트 인터페이스에 표시되어 유연하고 사용하기 쉬운 사용자 중심의 다양한 표현을 지원한다. 전자의무기록의 문제목록, 임상문서 작성, 텍스트 생성, 임상 의사결정지원 및 전문가 기능이 있는 처방 입력 등에 사용된다[16]. 병원정

보시스템의 자원을 통합하고 의료정보의 정확한 의미와 표현이 가능하도록 인터페이스 용어는 개념 기반의 통제의학용어로 발전되었다[17]. 사용자 입장에서는 다양한 어휘를 이용하여 자유롭게 임상문서를 작성하기 원하지만, 임상자료의 상호운용성을 보장하기 위해서는 개념 기반의 표준용어체계의 도입은 필수적이다[2].

참조 용어체계는 표현은 다르더라도 같은 의미를 갖는 용어들에 대해 동일한 개념코드를 부여하여 이를 통해 상호 의미전달이 가능하도록 만들어진 용어체계이다. 다양한 표준용어체계에서 정의한 개념과 매핑이 가능한 특징이 있다. 의료분야의 대표적인 참조 용어체계에는 UMLS, SNOMED CT 등이 있다. 미국 및 유럽에서는 오래 전부터 표준용어체계의 중요성을 인식하고 의료분야의 표준용어체계를 국가차원으로 지정하여 활발하게 사용하고 있는 반면 국내에서는 의료법 시행규칙 제14조에 의해 2014년 처음으로 의료인이 진료기록부등에 기록하는 질병명, 검사명, 약제명 등 의학용어와 진료기록부등의 서식 및 세부내용에 관한 보건의료 용어표준으로 KOSTOM이 채택되었다.

분류체계란 특정한 영역에서 사용되는 용어를 통계나 보고와 같은 특정 목적에 따라 유사한 것끼리 분류한 용어체계이다. 특정 목적에 따라 미리 구분한 수준에서 자료를 정리하기 위한 개념의 집합이기 때문에 선택 가능한 분류코드가 미리 정해져 있으며 그 외의 분류는 할 수 없다. 분류체계의 모든 용어는 미리 정해진 규칙과 범주에 따라 상호 배타적으로 분류된다[18]. 의료분야의 대표적인 분류체계에는 ICD-10(International Classification of Disease 10th revision), ICPC(International Classification of Primary Care), KCD-7 등이 있다. 국내에서는 보건의료 영역의 질병 및 사망원인 통계산출과 행정업무를 위하여 ICD-10을 한국어로 번역한 KCD-7을 사용하고 있다. 의료분야의 분류체계는 대부분 진료나 진료기록 목적이 아닌 통계나 보험청구와 같은 행정적인 목적의 업무를 달성하기 위하여 사용하기 때문에 행정용어체계(administrative terminology)라고도 한다.

이러한 임상 용어체계들은 각각의 고유의 용어모델을 가지고 있다. 용어모델과 정보모델은 종종 중복되어 사용되는데, 일반적으로 정보모델은 잘 정의된 의미 관계를 가진 필드들의 시리즈로 간주되며, 용어는 그 필드에 들어가는 가능한

값들이다. 이러한 간단한 접근으로 보면, 용어체계는 조직화된 개념들의 세트라고 볼 수 있다[19].

SNOMED CT와 같이 구조화된 용어체계는 동의어, 속성, 다양한 관계를 갖는 개념들로 구성된다. 관계에는 계층 관계(taxonomy, is-a), 부분전체 관계(partonomy, part of), 원인 관계(etiology, caused by), 위치관계(position, located in) 등이 있다. 분류체계의 경우는 계층 관계만 존재한다.

Fig. 2는 심한 흉부의 통증을 표현하기 위한 두 가지 방법을 나타낸다. 인간은 'severe chest pain'의 의미를 '심한 흉통'이라고 이해할 수 있지만 기계는 그 의미를 해석하지 못하기 때문에 'severe chest pain'의 병리(has_pathology)는 통증(pain)이고 발생부위(has_locations)는 흉부(chest)이며, 중증도(has-severity)는 심한 정도(severe)라고 정의하면 기계도 해석이 가능하다. 인터페이스 용어체계는 사용자 위주의 용어로 구성되기 때문에 흉통이라는 단일 개념과 흉통을 수식하는 여러 가지 용어로 구성될 수 있다. 흉통을 참조 용어체계와 매핑하면, 흉통은 병리가 통증이며, 발생부위는 흉통이라는 의미적 해석이 가능하다. 흉통을 수식하는 용어를 참조 용어체계와 매핑하면 '심한 압박성 급성 흉통(acute severe crushing chest pain)'은 임상경과를 뜻하는 수식어인 급성(acute)과 중증도를 뜻하는 심한정도(severe), 흉통의 특징을 설명하는 압박성(crushing)이라는 의미로 해석될 수 있다[16].

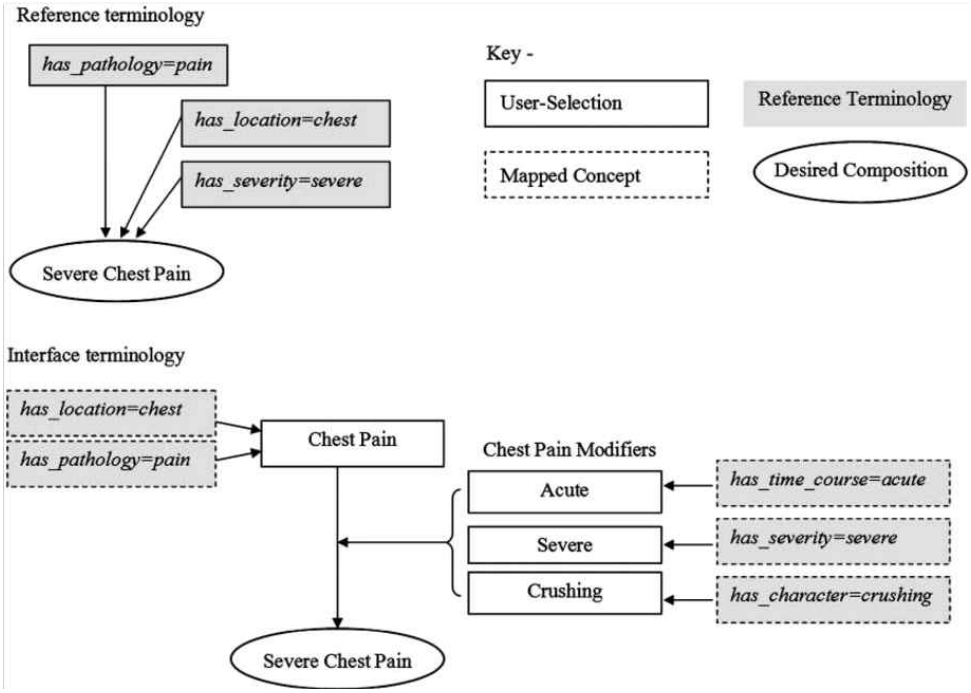


Fig. 4. Two approaches to composing the concept 'severe chest pain' (Source: S. Trent Rosenbloom et al,, 2007)

2) SNOMED CT

SNOMED CT는 의료 분야에 있어서 세계에서 가장 포괄적이고 정교한 참조 용어체계로 알려져 있다. 임상자료의 기록을 위해 필요한 전반적인 영역의 용어들을 지원하기 위하여 만들어졌으며, 임상 문서와 각종 보고 자료에 사용되는 코드, 용어, 유의어, 정의를 포함한다. 컴퓨터로 처리될 수 있도록 체계적으로 구조화된 의학용어의 집합체라고 할 수 있다. EHR을 위한 전반적인 핵심용어를 제공하기 위해 만들어졌기 때문에 인터페이스 용어로도 사용이 가능하다[18].

SNOMED CT는 원래 1965년 미국병리학회에서 병리분야에 사용할 형태학과 해부학을 기술하기 위해 만든 명명법인 SNOP(Systematized Nomenclature of Pathology)에서 시작되었다. 의료분야의 모든 용어를 기술할 목적으로 영역을 확대하면서 SNOMED(systematized nomenclature of medicine)로 발전하였고, 영국의 표준 임상 용어체계인 CTV3(Clinical Terms Version3)와 통합되면서 현재의 SNOMED CT가 되었다. 2007년에 미국, 영국 등 9개 국가가 모여 국제보건의료용어체계 표준개발기구인 IHTSDO를 새로 설립하고, SNOMED CT의 모든

버전에 대한 지석 재산권을 인수하여 SNOMED CT의 국제적 채택과 사용을 촉진하고 있다. 현재 29개의 나라가 정식 회원국으로 참여고 있으며 개별 라이선스를 얻어 사용하고 있는 곳까지 합치면 50개국이 넘는 국가에서 활발하게 사용되고 있다[20].

SNOMED CT는 의료분야의 대표적인 온톨로지다. 임상 현장에서 사용되는 다양한 용어들을 개념 기반으로 정리하고, 관계를 통하여 개념의 정확한 의미를 이해할 수 있다. 각 각의 임상 개념은 19개의 상위 수준의 계층(Top level hierarchy) 중 하나에 속하며, 개념 간 1개 이상의 하위형식 관계인 계층 관계(is-a)와 몇 가지 속성(attributes) 관계를 가짐으로써 개념의 정확한 의미를 구별할 수 있도록 설계되었다. Fig. 3은 SNOMED CT의 전체 설계도이다[21].

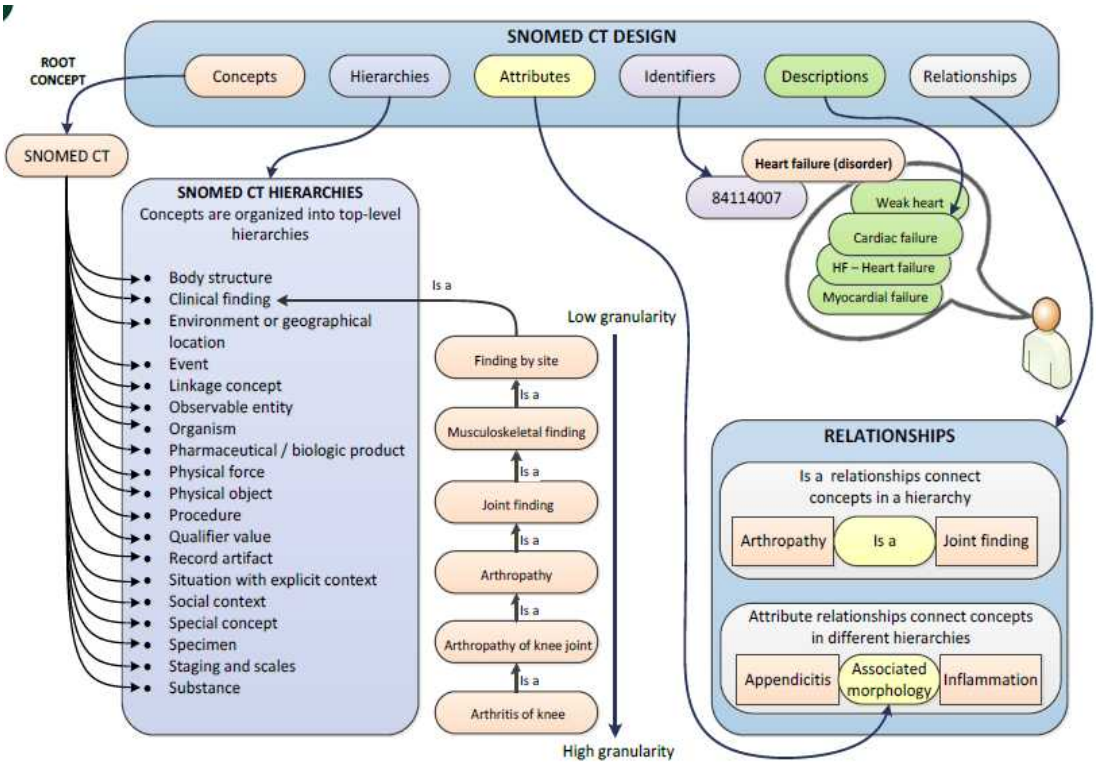


Fig. 5. SNOMED CT design (Source: SNOMED CT starter guide, 2016)

SNOMED CT 콘텐츠의 핵심 구성요소는 개념(concept), 용어(description), 관계(relationship)이다. 개념은 개념을 가장 잘 설명할 수 있는 완전히 명확한 명칭(Fully Specified Name: FSN)과 개념 식별자인 개념코드로 이루어진다. 용어(description)는 하나의 개념을 표현하는 다양한 용어로 동의어, 유사어 등이 포함된다. 용어는 용어, 용어식별자인 용어코드 그리고 개념코드로 구성된다. 'cardiac failure'와 'myocardial failure'는 서로 다른 용어코드를 갖고 있지만 'heart failure'라는 동일한 개념코드를 부여받아 동일한 개념으로 처리된다. Fig. 4는 핵심 구성요소인 개념과 용어, 개념과 관계에 대한 내용을 도식화한 SNOMED CT의 논리적 모델이다.

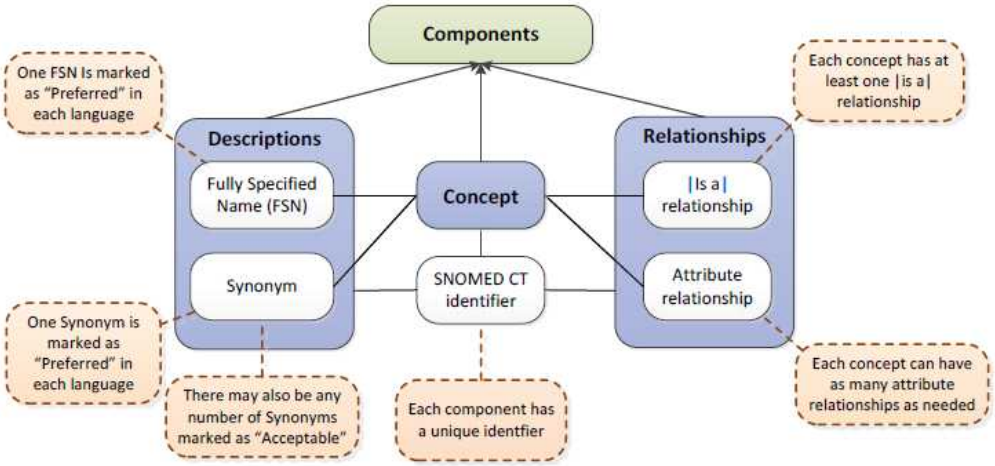


Fig. 6. The SNOMED CT logical model
(Source: SNOMED CT starter guide, 2016)

관계는 개념의 일종이며 개념과 개념 간의 의미적 연결 관계를 나타낸다. 두 개념의 계층 관계와 속성 관계를 나타낸다. Fig. 5는 제2형 당뇨(diabetes mellitus type 2)와 당뇨(diabetes mellitus)의 관계를 보여준다. 제2형 당뇨는 당뇨와 'Is-a' 관계로 제2형 당뇨는 당뇨 중 하나이며, 제2형 당뇨가 발생하는 해부학적인 부위(finding site)는 내분비 계통이라는 것을 알 수 있다. 관계는 개념 간의 관계로 그 개념에 속하는 용어들은 개념의 관계를 상속받는다. 즉 'diabetes mellitus type 2', 'DM 2', 'adult onset diabetes' 라는 용어들의 모두 제2형 당뇨

라는 개념이기 때문에 개념의 관계를 그대로 상속받는다.

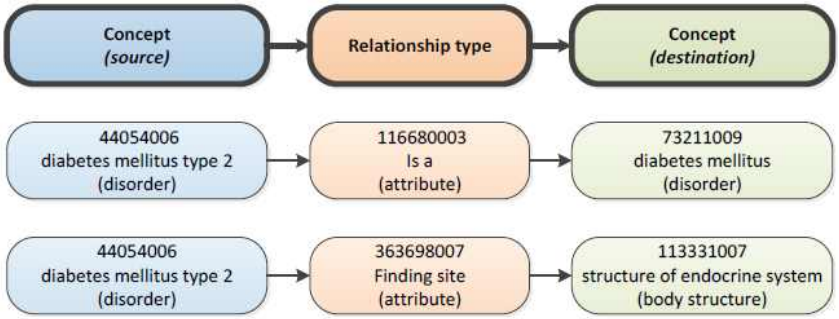


Fig. 7. Different Types of relationships available within SNOMED CT (Source: SNOMED CT starter guide, 2016)

개념 간의 관계를 이용하여 하나의 임상 개념을 정의할 수 있다. Fig. 6은 심장염의 정의 관계를 나타낸다. 심장염과 의미적 연결 관계에 있는 개념을 통해 심장염의 명확한 의미를 해석할 수 있다.

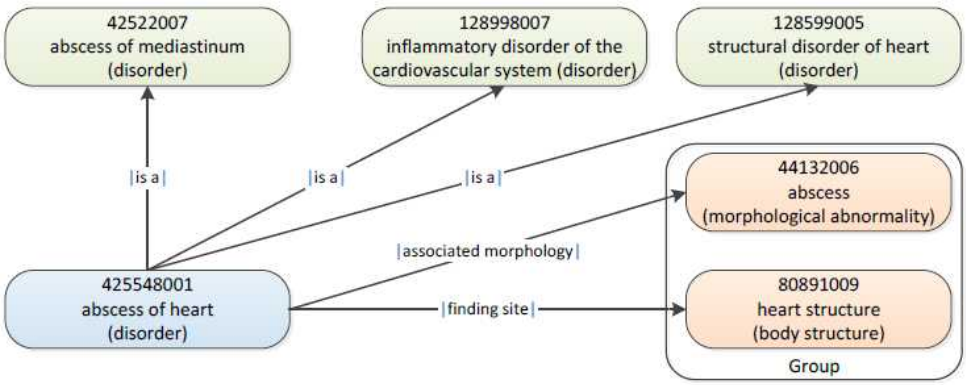


Fig. 8. Example of defining relationships
(Source: SNOMED CT starter guide, 2016)

SNOMED CT는 임상문서에 사용되는 모든 용어를 코드화하고, 용어에 대한 기계적 해석이 가능하도록 모델화 되었다. SNOMED CT의 콘텐츠는 독창적인 조합 문법이 있어 문법을 사용할 수 있기 때문에 기존의 개념을 조합하여 새로

운 개념을 만들 수 있다. SNOMED CT 문법은 코드와 부호로 표현된다. 충수절제술과 같이 단일 개념을 표현할 때는 개념코드 '80146002' 라고 쓰고, 선택적으로 용어를 함께 사용하고 싶으면 코드 뒤에 파이프라인'|' 두 개를 이용하여 그 안에 용어를 작성하여 '80146002|충수절제술|'라고 사용할 수 있다. 충수절제술을 상세화하기 위해서는 뒤에 콜론 ':'을 사용하여 '80146002|충수절제술|:'라고 작성하면 그 뒤에 나오는 모든 코드들은 앞의 개념을 수식하는 개념이 된다. 수식어가 여러 개일 경우에는 쉼표 ',' 로 구분한다. 복강경을 이용한 응급 충수절제술의 경우 '80146002|충수절제술|:260870009|우선순위|= 25876001|응급|, 42539100 5|사용 도구|= 86174004|복강경|'처럼 작성할 수 있다. Fig. 7은 SNOMED CT의 조합 문법의 기본원리를 정리한 것이다.

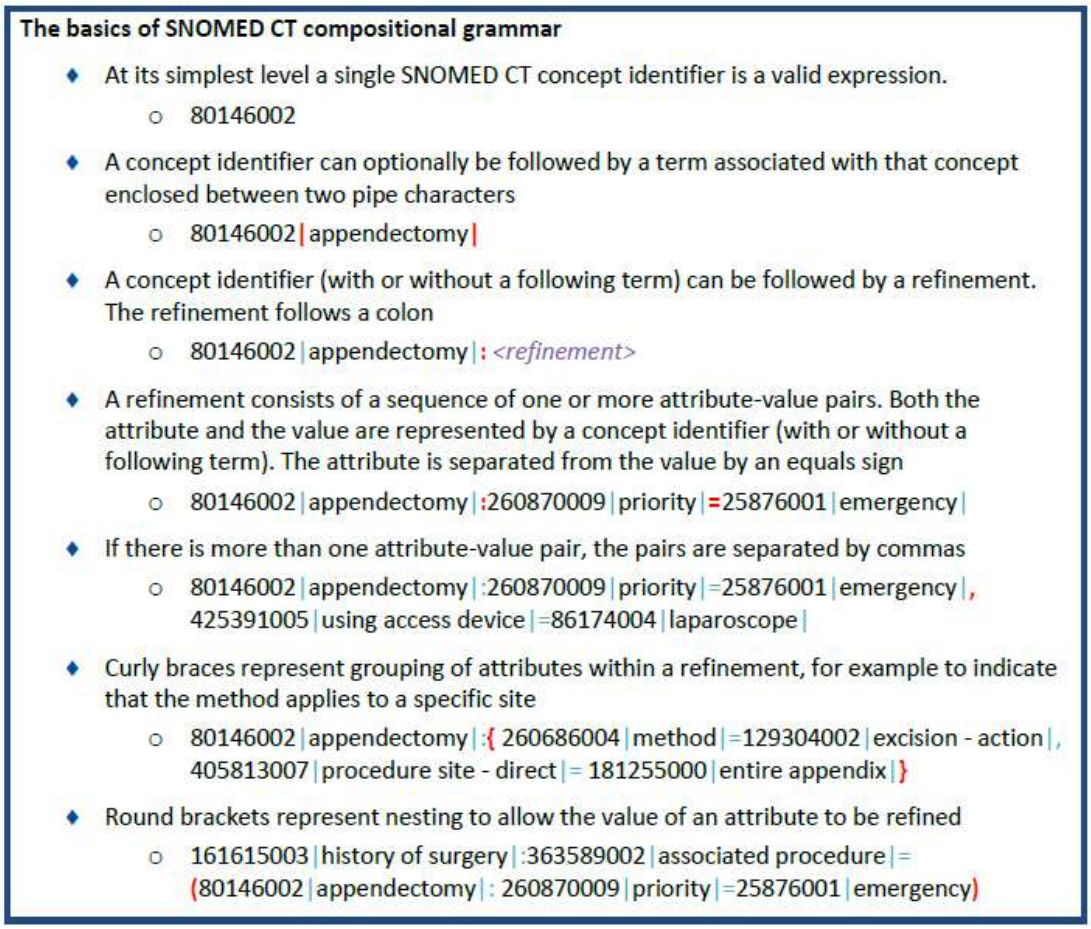


Fig. 9. The basis of SNOMED CT compositional grammar
(Source: SNOMED CT starter guide, 2016)

3) UMLS

통합의학언어시스템인 UMLS는 1986년부터 미국국립의학도서관에 의해 개발되고 있는 광범위한 용어체계로 기존의 의료분야의 용어체계들을 통합하여 관리하기 위해 개발되었다. 동일한 개념의 용어들이 서로 다른 용어체계 내에서 서로 다른 코드를 부여받아 상이한 표현으로 사용되고 있기 때문에 이러한 용어체계들을 모두 통합하여 동일한 개념으로 묶은 후 UMLS 고유의 방법으로 체계화한 것이다. 현재 140개 이상의 용어체계의 용어가 통합되어 있다[22].

UMLS는 메타시소러스(metathesaurus), 의미망(semantic network), 어휘사전(specialist lexicon)이라는 세 가지 지식원으로 구성된다(Table 1).

Table 1. Three knowledge sources of the UMLS

Components	Contents	Sources
Metathesaurus	Terms and codes from over 100 vocabularies	CPT®, LOINC®, ICD-10-CM, SNOMED CT®, etc.
Semantic network	Broad categories and their relationships	133 semantic types and 54 semantic networks
Specialist lexicon	Many biomedical terms and English lexicon for natural language processing	English dictionary, Medical dictionary, Medline abstract, etc.

메타시소러스는 다양한 출처의 용어체계에서 가져온 용어들을 동의어, 유사어, 어휘의 변형, 번역어 등 의미는 같지만 다르게 표기되는 경우들을 묶어 동일한 개념으로 조직화한 메타 용어사전이다.

메타시소러스는 용어(Terms), 원자코드(Atom Unique Identifier: AUI), 문자코드(String Unique Identifier: SUI), 어휘코드(Lexical Unique Identifier: LUI), 개념코드(Concept Unique Identifier: CUI)로 구성된다. 즉 하나의 용어는 4개의 식별코드를 갖게 된다. UMLS는 다양한 출처에서 온 용어들이 통합된 용어체계

이기 때문에 MeSH(Medical Subject Headings)의 'headache'와 SNOMED의 'headache'가 모두 포함되어 있다. 만약, 동일한 'headache'라도 용어의 출처가 다르면 문자코드는 같지만 원자코드가 다르다. 'headache'와 'headaches'처럼 어휘는 같지만 표현이 다른 경우에는 어휘코드는 같지만 문자코드가 다르며, 'headache', 'headaches', 'cranial pain'처럼 표현은 다르지만 두통의 의미를 갖는 모든 용어들은 동일한 개념코드를 갖게 된다(Table 2). 이러한 식별자를 통하여 원자별, 문자별, 어휘별, 개념별로 사용자가 원하는 다양한 수준의 용어들을 식별하여 처리할 수 있다.

Table 2. Identifiers in the metathesaurus

Concept (CUI)	Terms (LUI)	Strings (SUI)	Atoms (AUI)
C0018681 Headache	L0018681 headache	S1459113 headaches	A1412439 headaches(BI)
		S0046854 Headache	A2882187 Headache(SNOMED)
			A0066000 Headache(MeSH)
	L1406212 cranial pain	S1680378 Cranial Pain	A1641293 Cranial Pain(MeSH)

의미망은 개념의 유형이나 범주, 유형 간의 관계를 제공한다. UMLS의 의미망은 의미 유형(semantic type)과 의미 관계(semantic network relationships)로 구분된다. 의미 유형은 개념들을 속성에 따라 범주화 한 것이다. 질병 및 증후(disease or syndrome), 임상 약제(clinical drug) 등과 같이 성격이 다른 개념들이 속하는 넓은 범주이다. SNOMED CT가 19개의 상위수준의 계층으로 영역이 나뉘듯이 UMLS는 133개의 의미 유형이 곧 해당 개념이 속하는 영역이나 범주이다. 모든 개념은 최소 하나 이상의 의미 유형에 포함된다. 상위 범주인 개체(entity)와 사건(event)이라는 주요 계층에서 시작되며 하나의 개념은 하나 이상

의 의미 유형에 포함된다. 의미망 관계는 계층 관계(is a)와 비계층 관계(associated with)로 구분된다. 관계라 함은 의미 유형 간의 관계이거나 관계 간의 관계이다. Fig. 8은 UMLS의 의미망의 일부를 보여준다. 계층관계는 'is a'라는 1개의 관계만 가능하지만, 비계층 관계는 'associated with'라는 광범위한 관계 1개를 시작으로 물리적, 공간적, 시간적, 기능적, 개념적으로 관련된 다섯 가지 속성 관계와 그 하위의 더 상세하고 구체적인 다양한 속성 관계를 나타낼 수 있다. 이러한 의미망은 주로 의미를 해석하기 위한 응용 프로그램에서 사용된다. Fig. 8은 UMLS의 의미망의 일부를 보여준다.

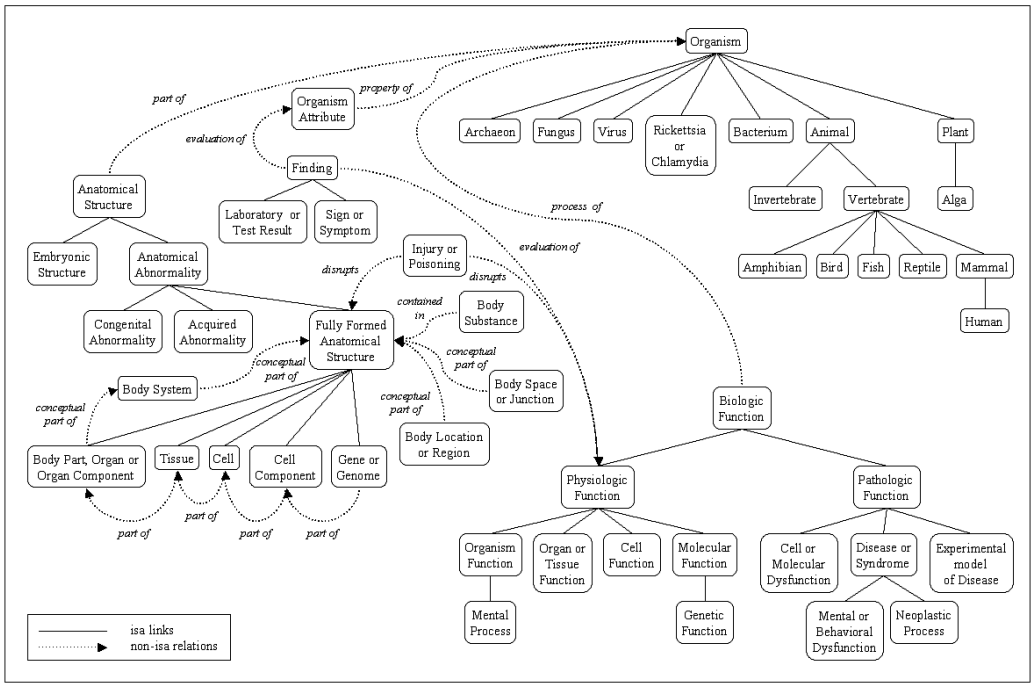


Fig. 10. A Portion of the UMLS semantic network
(Source: UMLS® Reference Manual, 2009)

전문 어휘사전은 기타 생명의학 분야의 전문용어와 영어단어 사전이 포함된 어휘목록이다. 메타시소러스에 없는 명사, 동사, 형용사, 부사, 전치사, 대명사 등의 품사와 동사의 과거형, 진행형 등을 포함하여 자연어 처리 시스템에 사용될 목적으로 만들어졌다. 전문 어휘사전과 함께 어휘도구를 제공한다.

이처럼 UMLS는 매우 방대한 용어를 포함하고 있기 때문에 다양한 사건이 작성되는 의무기록에 적합한 것 같지만, 실제 규모와 관계의 복잡성 때문에 중복, 애매모호함, 계층적 관계의 순환 등 내부적으로 정교하지 못하여 임상적용에 대한 문제점이 제기되기도 한다.

4) KOSTOM

국내 임상용어 표준으로 채택된 한국보건의료표준용어인 KOSTOM은 보건복지부가 국내 보건의료정보 국가표준을 마련하기 위하여 2004년에 보건의료정보표준화위원회를 발족하면서 개발되었다. 현재는 모든 업무가 사회보장정보원으로 이관되어 용어에 대한 개발 및 관리, 배포업무를 담당하고 있다. KOSTOM은 진료용 그림 표준과 진단, 의료행위, 임상검사, 방사선의학, 치과, 보건, 간호, 기타라는 8개의 용어표준 영역으로 구분되어 개발되었다. 관계가 없는 개념기반의 용어체계지만 개념에 UMLS의 개념코드를 매핑하여 국제적 표준과의 연계를 고려하였다. Fig. 9는 KOSTOM의 영역별 서브셋 테이블과 구성요소를 보여준다[23]. 통합용어테이블의 전체 개념과 용어들은 서로 다른 영역의 서브셋에 중복되어 나타날 수 있다.

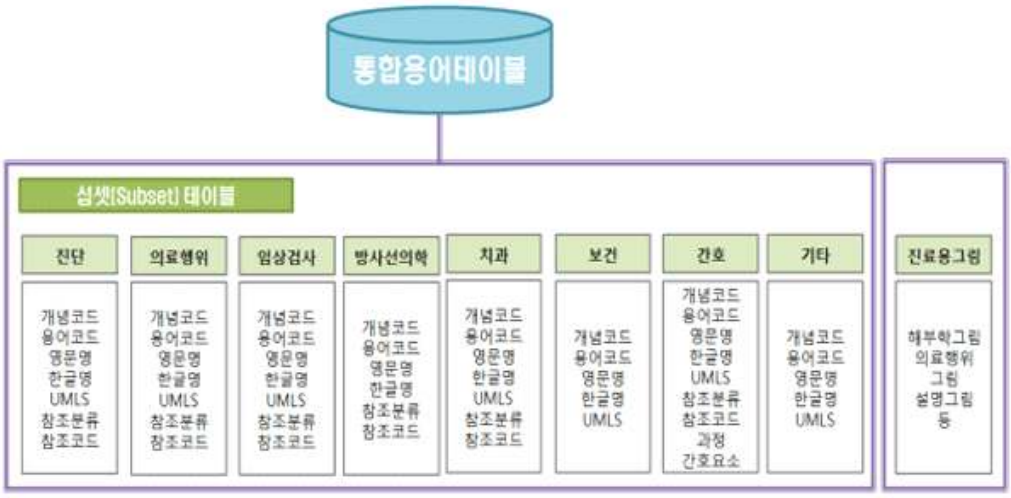


Fig. 11. Subset tables and components of the KOSTOM
(Source: http://www.hins.or.kr)

KOSTOM의 내용은 개념코드, 용어코드, 한글명, 영문명, UMLS코드 그리고 해당영역의 표준분류코드와의 매핑코드로 구성된다[24]. 특이한 점은 영문-한글 쌍을 하나의 용어로 식별한다는 것이다. KOSTOM에서는 'cerebral infarction-뇌경색'과 'cerebral infarction-대뇌경색증'이라는 영문-한글 용어 쌍을 서로 다른 용어로 구분하며, 이러한 한글-용어 쌍 중 하나의 용어코드를 개념코드로 사용한다. 만약 4개의 영문명이 4가지 한글명으로 번역이 가능하면, 사용된 용어는 8개지만, 영문-한글 쌍 조합은 16개가 되며, 16개의 용어코드가 발생한다. 동의어까지 포함하게 되면 용어의 개수는 훨씬 더 늘어나게 된다. 개념코드는 대표용어의 용어코드와 동일한 코드이다. 고혈압의 경우 대표용어는 'Essential(primary) hypertension-본태성(원발성) 고혈압'이며 이 용어의 용어코드인 'H02460066'가 개념코드가 된다. Table 3을 보면 개념코드가 'H02460066'인 용어들은 모두 본태성(원발성) 고혈압과 동의어라는 의미이다. 시대의 흐름에 따라 변하는 용어의 특성상 일반적으로 사용되는 대표용어가 변하더라도 개념코드를 변경할 수 없다.

Table 3. Term codes and concept codes based on English-Korean pairs in the KOSTOM

TermID	ConceptID	English	Korean
H00570148	H02460066	Essential Hypertension	본태 고혈압
H00570456	H02460066	Essential(primary) hypertension	본태성(일차성)고혈압
H02460066	H02460066	Essential(primary) hypertension	본태성(원발성) 고혈압
H02460111	H02460111	Unspecified primary hypertension	상세불명의 원발성 고혈압

2014년 의료법 시행규칙에 의해 보건복지부가 국가표준으로 고시하였지만 다른 임상 용어체계에 비하여 역사가 짧고, 도입에 대한 강제성이 없기 때문에 일부 공공병원을 제외하고는 사용하는 병원이 거의 없다. 민간병원이 주도하고 있는 국내 의료 현실에서 대부분의 병원들은 자체 개발한 용어체계를 사용하며, 진단명과 수술명에 대해서만 분류를 위해 각각 KCD-7과 ICD-9-CM(International

Classification of Disease 9th revision Clinical Modification) 코드를 사용하고 있다. 일부 대학병원에서는 SNOMED CT, UMLS, LOINC, ICNP(International Classification for Nursing Practice) 등을 병원 고유의 용어체계와 일부 매핑하여 사용하고 있다[25].

Fig. 10은 2016년 사회보장정보원에서 배포한 KOSTOM 버전 2.0의 콘텐츠의 일부이다. 엑셀 파일로 제공되며 관계가 없는 용어사전 형태로 하나의 단독 테이블로 제공된다. 각 용어들은 영역별로 참조분류 코드가 매핑되어 있는데 진단 영역은 KCD-7, 수술 및 처치와 관련된 의료행위는 ICD-9-CM, 임상검사 영역은 LOINC, 간호영역은 ICNP 코드가 매핑되어 있다.

용어코드	개념코드	영문명	한글명	KCD	ICD9CM	LOINC	EDI	CCC	ICNP
H00010847	H02650788	Abscess of fascia, pelvic region and thigh (buttock, femur, pelvis, h	근막의 농양, 골반 부분 및 대퇴	M72.85					
H00010851	H02650427	Abscess of fascia, shoulder region (clavicle, scapula, acromioclavicu	근막의 농양, 어깨 부분	M72.81					
H00010866	H00010866	Abscess of fascia, site unspecified	근막의 농양, 상세불명 부분	M72.89					
H00010871	H02650512	Abscess of fascia, upper arm (humerus, elbow joints)	근막의 농양, 위팔	M72.82					
H00010885	H00010885	Abscess of finger	손가락 고름집						
H00010890	H00010890	Abscess of foot, except toes	발가락을 제외한 발 고름집						
H00010905	H00010905	Abscess of forearm	아래팔 고름집						
H00010910	H00010910	Abscess of gallbladder without calculus	결석이 없는 담낭의 농양	K81.0					
H00010924	H00010924	Abscess of hand	손 고름집						
H00010939	H00010939	Abscess of hip	엉덩이 고름집						
H00010943	H00010943	Abscess of intestine	장의 농양	K63.0					
H00010958	H00010958	Abscess of ischiorectal fossa	좌골직장와의 농양	K61.3					
H00010962	H00010981	Abscess of larynx	후두 고름집	J38.7					
H00010977	H00010981	Abscess of larynx	후두의 고름집(농양)	J38.7					
H00010981	H00010981	Abscess of larynx	후두 농양	J38.7					
H00010996	H00010996	Abscess of leg, except foot	발을 제외한 다리 고름집						
H00011007	H00011011	Abscess of liver	간의 고름집(농양)	K75.0					
H00011011	H00011011	Abscess of liver	간의 농양	K75.0					
H00011026	H00011011	Abscess of liver NOS	간농양NOS	K75.0					
H00011031	H00011050	Abscess of lung	폐의 고름집	J85.2					
H00011045	H00011050	Abscess of lung NOS	폐 고름집	J85.2					
H00011050	H00011050	Abscess of lung NOS	폐의 농양 NOS	J85.2					
H00011064	H00011050	Abscess of lung NOS	폐의 고름집(농양) NOS	J85.2					
H00011079	H00011079	Abscess of lung and mediastinum	폐 및 종격의 농양	J85					
H00011083	H00011083	Abscess of lung with pneumonia	폐렴을 동반한 폐의 농양	J85.1					
H00011098	H00011098	Abscess of lung without pneumonia	폐렴을 동반하지 않은 폐의 농양	J85.2					
H00011103	H00011103	Abscess of mastoid	유돌의 농양	H70.0					
H00011118	H00011118	Abscess of mediastinum	종격의 농양	J85.3					
H00011122	H00011122	Abscess of oesophagus	식도의 농양	K20					
H00011137	H02445245	Abscess of orbit	안와 농양	H05.0					
H00011141	H00011141	Abscess of pancreas	췌장의 농양	K85					
H00011156	H00011156	Abscess of parametrium specified as acute	급성이라고 명시된 자궁주위조직의 농양	N73.0					
H00011161	H00011161	Abscess of parametrium specified as chronic	만성이라고 명시된 자궁주위조직의 농양						
H00011175	H00011175	Abscess of parametrium unspecified whether acute or chronic	급성인지 만성인지 상세불명인 자궁주위조직의 농양						

종합 | 고시용어 | KCD7 신규용어

Fig. 12. Part of the contents of the KOSTOM (Source: Social security information service)

5) ICD-10과 KCD-7

세계보건기구(WHO)는 사망과 질병, 상해에 대하여 국제적으로 비교 가능한 질병통계를 산출하기 위해 국제질병분류체계인 ICD-10의 사용을 권고하고 있다. ICD는 1900년 국제통계학회의 국제사인분류를 시작으로 발전하여 1948년부터 WHO에서 개정업무를 이관 받아 여러 번 개정을 거쳐 현재는 1990년에 발표된 ICD의 10차 개정판을 보편적으로 사용하고 있다[26].

ICD-10의 구조는 대분류, 중분류, 소분류, 세분류, 세세분류로 구분된다. 모든 코드는 첫 자리에 알파벳을 하나씩 붙이고 그 뒤에 기본 숫자 두 자리, 소수점, 소수점 이하의 수가 붙는다. 소수점 이전까지의 세 자리를 소분류라고 하며 소수점 이하의 수는 상황에 따라 세분류, 세세분류 혹은 세세세분류까지 구분이 가능하다. ICD-10은 대표적인 분류체계로서 상호 배타적인 계층구조를 가지며, 유사한 질병명을 하나의 코드로 분류하기 때문에 유사한 질병군을 모아서 처리하는 용도로는 매우 적합하나 동일한 질병을 표현하는 유의어나 동의어에 대한 구분이 불가능하여 상세한 용어를 표현하거나 처리하는 데는 적합하지 않다. ICD-10은 온라인 브라우저(http://apps.who.int/classifications/icd10/browse/2016/en)를 통해 검색이 가능하다.

Table 4는 ICD-10의 대분류 범위이다.

Table 4. Chapters of ICD-10

Chapters	Code categories
Ⅰ Certain infectious and parasitic diseases	A00-B99
Ⅱ Neoplasms	C00-D48
Ⅲ Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50-D89
Ⅳ Endocrine, nutritional and metabolic diseases	E00-E90
Ⅴ Mental and behavioural disorders	F00-F99
Ⅵ Diseases of the nervous system	G00-G99
Ⅶ Diseases of the eye and adnexa	H00-H59
Ⅷ Diseases of the ear and mastoid process	H60-H95
Ⅸ Diseases of the circulatory system	I00-I99
Ⅹ Diseases of the respiratory system	J00-J99
Ⅺ Diseases of the digestive system	K00-K93
Ⅻ Diseases of the skin and subcutaneous tissue	L00-L99
ⅩⅢ Diseases of the musculoskeletal system and connective tissue	M00-M99
ⅩⅣ Diseases of the genitourinary system	N00-N99
ⅩⅤ Pregnancy, childbirth and the puerperium	O00-O99
ⅩⅥ Certain conditions originating in the perinatal period	P00-P96
ⅩⅦ Congenital malformations, deformations and chromosomal abnormalities	Q00-Q99
ⅩⅧ Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	R00-R99
ⅩⅨ Injury, poisoning and certain other consequences of external causes	S00-T98
ⅩⅩ External causes of morbidity and mortality	V01-Y98
ⅩⅩⅠ Factors influencing health status and contact with health services	Z00-Z99
ⅩⅩⅡ Codes for special purposes	U00-U99

ICD-10은 현재 43개의 언어로 번역되어 100개국 이상의 나라에서 사용되고 있다. 호주, 미국 등 많은 국가에서는 이를 일부 수정하여 사용하며, 국내에서도 ICD-10의 한글 번역본인 KCD-7을 사용하고 있다. KCD-7에는 ICD-10코드 외에 국내 실정에 적합한 질병과 한의학 분류코드가 추가되었다. ICD-10이 우리나라의 임상현장 상황과 국가 차원의 질병통계 작성 등에 필요한 충분히 세분화된 코드를 제공하지 못하고 있기 때문에 질병분류코드의 활용가치를 높이고자 일부 질환에 대하여 더욱 구체적인 코드를 사용하도록 개정되었다[6]. 국내에서만 사용되는 질병분류코드에는 코드 옆에 태극마크 '◒'를 붙여 구분한다. 울혈성 심부전의 경우 ICD-10에서는 'I50.0'이지만, 국내에서는 우심부전일 경우에는 'I50.03'을, 수축기능부전을 동반한 울혈성 심부전을 경우는 'I50.04'를 부여해야 한다(Table 5). 그 외의 유형이나 상세불명의 울혈성 심부전은 'I50.08'을 부여하기 때문에 만약 국내에서 울혈성 심부전을 'I50.0'코드로 분류했다면 잘못 분류한 사례가 되며, 수축 기능부전을 동반한 울혈성 심부전인데 그냥 심부전 코드인 'I50.08'로 분류해도 분류오류에 해당된다.

Table 5. Difference in granularity between ICD-10 and KCD-7

Diagnosis	ICD-10	KCD-7
Congestive heart failure	I50.0	I50.08◒
Right heart failure	I50.0	I50.03◒
Congestive heart failure with systolic dysfunction	I50.0	I50.04◒

2. 임상정보모델

임상정보모델은 임상정보의 활용을 목적으로 의무기록의 구성 및 내용을 구조화 한 것으로 임상정보모델은 표준용어시스템을 사용하여 임상개념을 표준화 및 공식화하여 자료의 상호호환성을 보장한다[27]. 임상내용을 동일한 상세수준

에서 일관성 있게 수집하기 위해서는 임상내용의 표현방식, 속성, 속성의 유효한 값, 기타 임상내용의 표현에서 지켜야 할 규칙에 대해 합의된 임상내용모델이 필요하다[5]. 임상정보를 공식화된 형태로 저장하고 활용하기 위해 개발된 대표적인 임상정보모델에는 국제표준화기구(ISO)의 승인을 받은 openEHR 재단의 아키타입과 인터마운틴 헬스케어의 CEM, 국내에서 개발한 CCM 등이 있다. 이들은 전자의무기록에 사용되는 임상소견, 진단검사, 약물 처방 등 여러 분야의 임상영역에서 자주 사용되는 개념을 표준화된 형식으로 제공할 수 있도록 개발되었다. 임상정보모델 중 환자 자료를 상세하게 표현한다는 의미에서 상세임상모델(Detailed Clinical Model: DCM)이라고도 한다.

1) openEHR의 아키타입

openEHR은 EHR을 위한 개방형 명세서이다. 비영리 단체인 openEHR 재단이 개발한 임상모델이며 이에 대한 개방형 연구와 구현을 지원하고 있다. openEHR의 아키타입 명세서들은 15년 이상 진행된 유럽과 호주의 연구 성과를 바탕으로 유럽의 표준으로 발전하였다. openEHR의 아키타입은 하나의 임상개념 단위인 개념과 개념을 기반으로 구조화한 자료인 아키타입, 아키타입을 그룹화한 템플릿으로 구성된다. 아키타입은 문서 작성과 메시지 교환을 위해 만들어졌으며 이러한 활용을 보장하기 위하여 표준 EHR 참조모델, 서비스 인터페이스 모델, 도메인 콘셉트 모델을 제공한다. CEN/ISO EN13606은 국가 간 혹은 기관 간 EHR 데이터의 상호운용성을 보장하기 위하여 의료정보 교환에 대한 표준을 정의한 문서인데, 아키타입에 대한 내용이 포함되었다[28-29].

아키타입은 혈압측정이나 임상병리 검사결과와 같은 특정 임상영역을 표현하는 참조 모델의 개념들을 구조화되고 제한한 조합이다. 특정 임상개념을 더 적합하거나 더 상세하게 표현하기 위해 하나의 아키타입은 부모 아키타입이라고 불리는 다른 아키타입에 구속되어 정의될 수 있다[30].

아키타입은 헤더(header), 정의(definition), 온톨로지(ontology)로 구성된다. 헤더는 저자 정보나 식별자와 같은 아키타입에 대한 메타 데이터를 담고 있다. 정의 섹션은 참조 모델 엔터티들 측면에서 아키타입이 표현하는 임상개념이 설명되는 곳이다. 온톨로지 섹션은 정의섹션에 정의된 엔터티들이 용어체계와 연결되

어 설명되는 곳이다. 아키타입의 특징은 아키타입의 재사용을 허용한다는 것이다. 다른 아키타입에 의해 이미 표현된 정보를 추가로 제약하여 제공할 수 있다.

Fig. 11은 혈압에 대한 아키타입이다. 혈압을 표현하기 위해 필요한 다양한 요소로 구성되어 있다. 혈압 모델에서 보듯이 혈압의 데이터 값으로 수축기 혈압, 이완기 혈압, 평균 동맥 혈압, 맥박 압이 올 수 있으며 그 값은 수치 데이터이다. 혈압에 대한 부가 정보는 커프 사이즈, 위치, 측정 방법 등이며 텍스트 데이터로 입력된다.

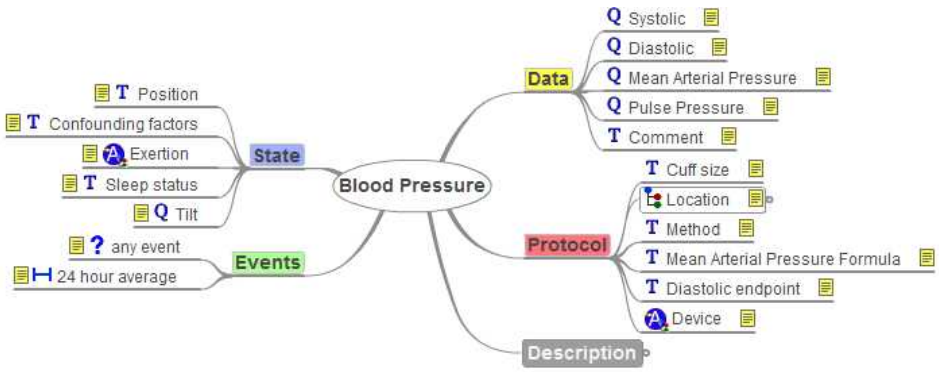


Fig. 13. Blood pressure archetype (Source: www.openEHR.org/ckm)

Fig. 12는 openEHR의 무료 공개용 아키타입을 이용하여 개발된 상용 전자의무기록 화면이다. 혈압 아키타입을 포함하여 여러 개의 아키타입으로 구성되어 있다. 혈압은 수축기혈압과 이완기 혈압 두 가지 요소로 입력되어야 하며 혈압의 단위는 mmHg이며, 수축기 혈압과 이완기 혈압의 값은 각각 단일 값으로 수치형 데이터로 입력되어야 함을 알 수 있다. 아키타입과 같은 상세임상모델을 사용하면 표준화된 입력서식 제공이 가능하며 데이터의 일관성 있는 저장과 정보활용 및 교환 측면에서 유용하다. 의무기록 중 입원기록을 구성하는 항목인 주 증상, 현 병력 상태, 과거 병력 등의 구성요소를 섹션 클래스로 정의할 수 있다. 화면에서 보다시피 진단명 입력(assessment)모델의 경우 진단명과 그렇게 진단한 근거(rationale)에 대한 값을 입력하도록 구성되어 있다.

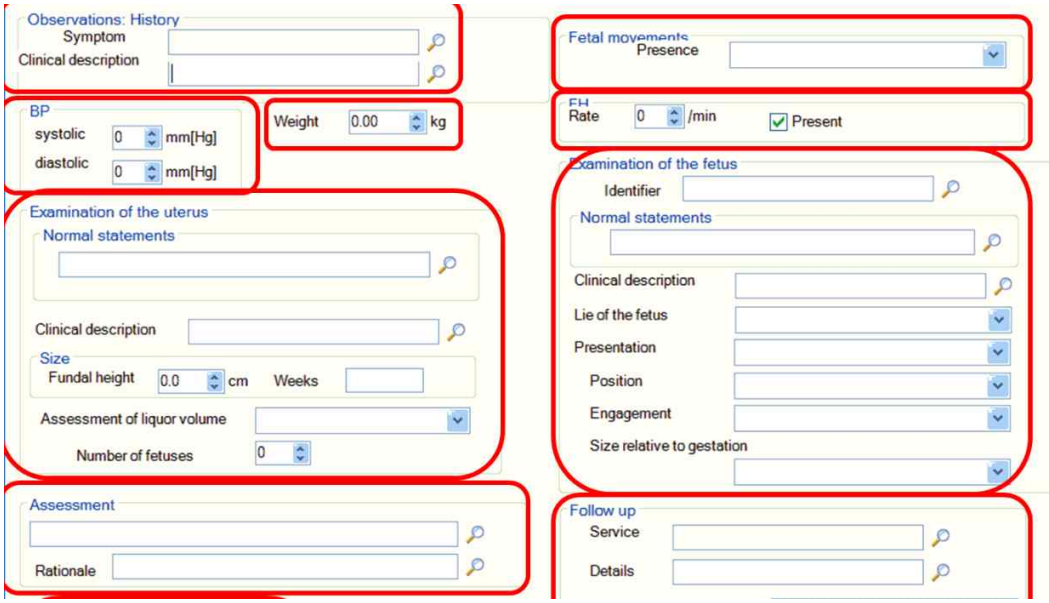


Fig. 14. EMR screen constructed with archetypes
(Source: Ocean Informatics, 2012)

2) 인터마운틴 헬스케어의 CEM

CEM은 미국 최대 규모의 인터마운틴 헬스케어에서 약 20년 동안 개발한 임상정보모델이다. 상세한 임상자료를 표현하기 위하여 개발되어 자료의 논리적인 구조를 표현할 수 있다. 구조화된 형태의 데이터 입력과 자동화된 임상자료의 처리, 임상 의사결정지원 시스템에 활용되고 있다. CEM의 구조는 모델링 단위인 키(key)와 키가 가질 수 있는 값(data)을 갖는다. 만약 하나의 키가 여러 개의 값으로 구성된다면, 키 하위의 아이템과 아이템의 값들로 자료를 표현할 수 있다. 이러한 데이터 값을 부가적으로 서술하기 위해 수식자(qualifier)와 변경자(modifier)를 사용할 수 있다[31].

환자의 임상자료를 보다 정확하게 수집하기 위해 서식에 작성되는 많은 항목들에 대한 CEM 모델이 개발되어 실제로 인터마운틴 헬스케어와 관련된 의료기관에서 사용되고 있지만, 진단명에 대해서는 여전히 서술방식으로 입력하거나, 진단명 리스트 중 하나를 선택하는 방식으로 수식자와 변경자가 없는 단언(assertion)모델의 형태이다(Fig. 13). CEM은 임상문서에 기술되는 다양한 진료

정보들을 구조화된 형식으로 입력하고 저장하기 위한 목적으로 만들어졌지만 진단명의 경우에는 단일 입력되는 형식이다.

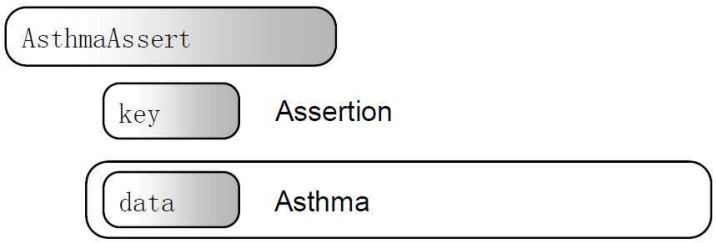
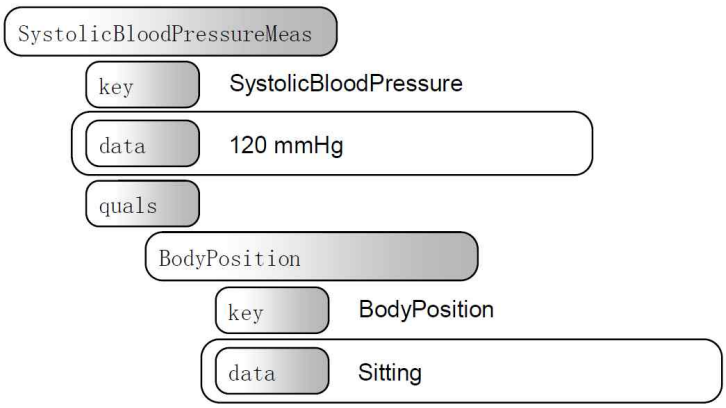


Fig. 15. The CEM Instance about Asthma
(Source: J.F. Coyle, 2013)

Fig. 14는 수축기 혈압에 대한 측정 모델이다. 120으로 측정된 혈압 값은 앉은 상태(sitting)로 쟀다는 내용을 부연 설명한 것으로 측정자세(body position)라는 수식자에 의해 환자의 측정 당시의 상황이 명확해 진다. 키는 현실 세계의 임상개념이며, 데이터는 해당 값에 해당된다. 이러한 CEM들은 서로 조합되어 사용될 수 있다. 예를 들어 혈압은 일반적으로 수축기혈압과 이완기혈압을 함께 사용하기 때문에 Fig. 14의 (a)인 수축기 혈압 측정 모델과 이완기 혈압 측정모델을 묶어서 Fig. 14의 (b)와 같이 표현할 수 있다. 혈압 측정 모델처럼 동일한 수식어를 갖는 동일한 구조의 모델을 조합한 것을 패널(panel)이라고 한다. 이질적인 구조를 갖는 모델의 조합도 가능한데 혈압, 체온, 심박수, 호흡수라는 4개의 이질적인 모델을 조합하여 생체활력을 측정하는 모델을 만들 수 있다.

(a) Systolic blood pressure measurement model



(b) Blood pressure measurement model

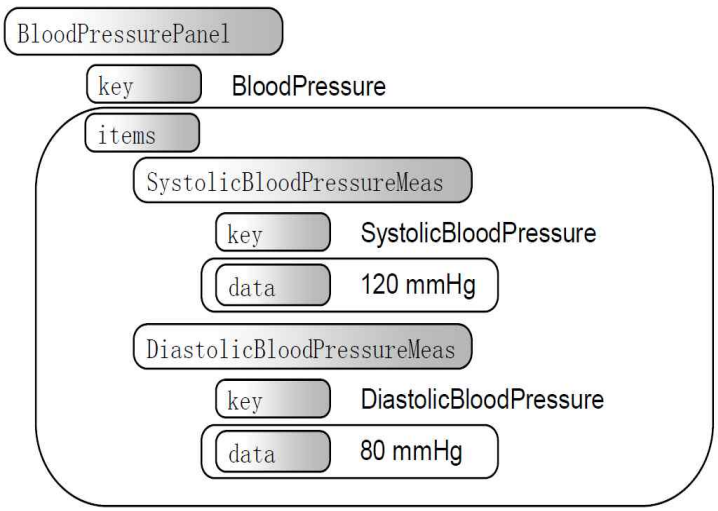


Fig. 16. The CEM instance representing a blood pressure panel measured in sitting position (Source: J.F. Coyle, 2013)

3) 국내의 CCM

국내에서도 용어와 자료구조 및 임상서식 항목의 표준화가 부족하여 의료기관 간 정보공유는 물론 의료기관 내 의료정보의 활용도 제한적이다. 이에 국내에서도 EHR 시스템에 사용되는 임상문서의 저작, 기관 간 임상정보의 교환, 임상

의사결정지원시스템에 활용할 목적으로 2005년부터 2010년까지 보건복지부의 산하의 EHR핵심공통기술연구개발 사업단에서 CCM을 개발하였다. 개발된 영역은 임상소견(Clinical Finding), 처치(Procedure), 측정(Measurement), 약제(Medication) 영역이다. 만약, 복통으로 병원을 방문한 환자의 주 증상을 표현한다면 복부 통증의 심한정도, 발생기간, 발생시점 등이 필요하며 이러한 항목을 수식자로 정의하였다. Fig. 15은 CCM 모델의 구조도이다. 하나의 임상개념을 수식자(qualifier)와 변경자(modifier)를 이용하여 부가적으로 서술할 수 있다[32].

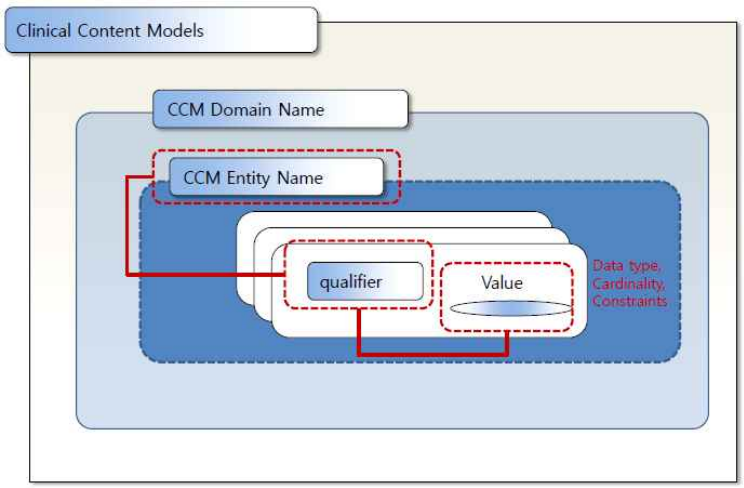


Fig. 17. Structure of CCM
(Source: Center for EHR, 2010)

3. 기존의 질병분류 연구

1) 인코더 방식

질병분류 업무를 자동화하기 위한 연구는 1990년대부터 본격화되었다[34]. 전자의무기록에 진단명을 입력하는 방법은 프리텍스트 형식으로 입력하는 방법과 코드화된 데이터 형식으로 입력하는 방법이 있다. 질병분류 방법 중 인코더 방식은 일반 자연어로 표기된 진단명에 가장 적절한 질병코드를 부여하도록 도와주

는 소프트웨어이다. 논리기반 인코더는 주요 용어를 입력하면 질병분류지침서에 맞게 여러 가지 관련 질문이 제시되며 가장 적절한 코드를 부여해주도록 도와주는 방식이다. 자동 코드북은 질병분류지침서를 그대로 전산화하여 테이블 리스트 형식의 화면으로 옮긴 형태로 질병명을 선택하면 질병분류 코드가 부여된다. 자동 코드부여 방식은 컴퓨터에 입력된 진단용어에 의하여 자동적으로 질병분류코드가 부여되는 방식이다[33].

2) 통계적 분류기

과거에는 주로 자유 서술형태의 진단명이 입력되었기 때문에 입력된 텍스트 진단명을 통계적 분류기를 이용하여 훈련한 후 질병코드를 할당하는 방법들이 연구되었다[13]. 프리텍스트를 분류하는 연구는 현재까지 진행되고 있는데 방사선 기록지나 퇴원요약지 등에 서술방식으로 작성된 질병명을 자연어 처리 기술을 이용하여 질병분류코드를 할당하는 방법들이 연구되었다. ICD-9 코드를 자동으로 분류하기 위해 규칙 기반의 자동분류 시스템을 제안한 연구에서는 질병과 증상 코드가 동시에 부여되는 문제를 해결하기 위해 질병분류 지침을 전문가 규칙으로 표현하였다[35]. 질병코드와 증상코드의 관계는 C4.5 의사결정나무 분류기로 찾아내어 서술방식의 임상문서에서 최종 ICD-9 코드가 모두 찾아지면 관련된 증상코드들은 삭제하는 방식을 도입하였다.

3) 국내 연구

서진숙 등[36]은 텍스트 검색 기반의 진단명 선택의 어려움을 해결하고자 안과의 다빈도 질환을 대상으로 트리 구조의 질병용어체계를 개발하였다. 안과 질환을 6개의 전문분야로 구분하고 6단계를 거쳐 병명을 선택할 수 있도록 시스템에 적용하였다. 사용자 인터페이스에서 5단계 까지는 전문분야별 세부 항목을 선택하고, 마지막 6단계에서는 부위를 선택하는 방식이다. 예를 들어 망막 전문영역을 선택하면 관련 질환이 뜨고 그 중 해당하는 병명을 선택한 후 계층이 나오지 않을 때 까지 단계별로 용어를 선택하는 방식이다. 질병마다 6단계까지 나열되는 경우도 있고 2단계에서 끝나는 경우도 있기 때문에 선택 가능한 발병 부위는 계층구조에서 분리하여 어느 단계에서건 선택할 수 있도록 하였다. 이러한 계

층 구조를 적용한 결과 기재가 미흡하여 세부분류가 불가능했던 진단명이 26.4%에서 0.7%로 감소하고, 세부 분류된 진단명이 68.8%에서 97.2%로 향상되었다. 구조화된 사용자 인터페이스가 질병분류의 정확도를 높일 수 있음을 시사한다.

강영희[37]는 산부인과, 외과, 신경과를 대상으로 퇴원요약지에 프리텍스트로 입력한 질병명을 수집하여 해당 질병코드를 일대일로 매핑하여 질병용어 테이블을 구성한 후 의사가 질병명을 텍스트로 입력하면 입력된 텍스트에 질병분류 코드가 부여되도록 한 후 질병분류 알고리즘에 의해 질병분류 생성을 지원하였다. 암일 경우 암의 형태 코드를 입력할 것을 요구하고, 외상코드일 경우에 외인코드를 입력할 것을 요구하는 방식이다. 구체적인 질병코드를 제시하기 보다는 동반코드를 요청하는 개념에서 접근하였다. 이러한 동반코드에 대한 요청이 끝나면 예정된 치료 여부시행 여부, 치료 중 또는 치료 후 합병증 유무 등을 팝업창으로 묻고 이에 대한 응답 방식으로 질병명을 추가할 수 있다.

전은정 등[38]은 자동분류와 관련하여 약물부작용 용어체계를 기반으로 자동코딩 시스템을 제안하였는데 해당 용어체계가 계층적 구조인 특징을 이용하여 수집된 이상반응을 입력하면 일치, 부분일치, 근접이라는 세 가지 형태의 초벌코딩을 자동으로 하는 방식이다. 텍스트가 일치하는 경우 코드가 매핑되며, 일치하지 않을 경우 부분일치를 선택하면 문자 단순 일치법에 의한 맵퍼(mapper)를 개발하였다. 일반적으로 국내 병원들은 질병코드 분류를 위한 방법으로 질병명과 질병분류코드가 일대일로 매핑된 테이블을 사용한다. 질병명을 텍스트 검색하여 검색된 질병명 중 하나를 선택하는 방식이다. 김미정 등[15]은 질병분류의 정확도를 높이기 위해 산과 질환을 대상으로 질병분류 전문지식을 규칙으로 정의하여 지식베이스를 구축한 후 질병분류 부서에서 사용하는 퇴원분석 시스템에 적용하였다.

III. 질병분류 지식 체계화

3장에서는 문헌과 전문가 지식을 통해 질병분류를 위한 지식을 체계화하였다. 기존의 질병분류의 문제점을 해결하기 위해 필요한 KCD-7의 일반적인 지식을 도출하였다. 순환기 계통의 질병명을 어휘 분석하여 질병명의 구조를 확인하고 질병명을 수식하는 수식어의 속성을 파악하였다. 도출된 지식을 근거로 질병분류에 영향을 주는 요인에 대하여 분류규칙을 작성하였다.

1. 질병분류 지식 도출

1) 질병분류의 오류 유형

질병분류코드의 정확성과 불일치와 관련된 기존의 연구결과를 토대로 살펴본 질병분류의 오류 상황은 다음과 같다. 첫 번째는 세분화 오류이다. 세분화된 코드가 존재하지만, 세분화되지 않은 코드를 부여하는 경우이다. 덜 구체적인 정보보다 구체적인 정보로 코딩해야 한다. 두 번째는 주 진단 선정 오류이다. 기저질환보다 치료가 이루어진 증상을 주 진단으로 선정해야 하며, 의심되는 병태를 주 진단으로 코딩해야 한다. 재활치료, 화학요법 등을 목적으로 입원한 경우에는 병태가 아닌 입원목적을 주 진단으로 분류해야 한다. 세 번째는 코드 자체를 누락시키는 경우이다. 환자가 만성질환을 앓고 있다면 기타진단으로 코딩해야 하며, 후유증으로 나타난 병태일 경우 병태와 후유증 코드를 함께 분류해야 한다. 환자의 과거력에 대한 정보도 코딩해야 한다. 네 번째는 세트 코드의 경우 하나를 누락시키는 경우이다. 원인과 발현 증세를 함께 이원 분류해야 하는 질병의 경우에는 두 개 모두 부여해야 한다.

주 진단의 선정은 진단 코드들이 일관성 있고 정확하게 분류되었다는 가정하에 이루어져야 하므로 이번 연구범위에서 제외하였다. Table 6은 질병분류의 오류 유형이다.

Table 6. Error types of classification of diseases

Type	Description
Subdivision error	- Giving comprehensive code instead of concrete one
Principle diagnosis selection error	- Complications being selected as principle diagnosis - Purpose of hospitalizations being selected as main diagnosis
Code missing	- Missing codes of chronic disease, sequelae, past history, injury and external accident
Paring code missing	- Missing an odd code of a pair in the process of two-way classification

2) 지식 도출 방법

질병분류 지식을 도출하기 위해 총 3권으로 구성된 한국질병사인분류체계인 KCD-7의 한글 색인(1권), 코드 색인(3권), 질병분류 지침서(2권)와 통계청에서 발행한 질병코딩지침서 2016 버전을 사용하였으며[6,39], 문헌의 내용을 해석하고 규칙으로 도출하는 과정에서 전문가의 지식이 사용되었다. 질병분류의 일반적인 상위 지식을 먼저 도출하고 순환기 계통 질환의 지식을 도출하였다. 세분화와 관련된 질병분류는 세분화를 위한 요소를 속성으로 정의하여 공식화하고, 누락 코드 및 그 외에 코드분류에 영향을 주는 요소는 분류 규칙으로 정의하였다. 질병명을 공식화한 질병분류모델은 4장에서 다룬다.

지식 모델링은 시설이나 제품, 과정에 대한 지식이나 표준안을 컴퓨터가 이해할 수 있는 모델을 만드는 과정이다. 지식 모델은 지식이 소프트웨어에 의해 해석될 수 있는 어떤 지식 표현 언어나 데이터 구조로 표현될 때 비로소 컴퓨터 해석이 가능하며, 데이터베이스나 데이터 교환파일에 저장될 수 있다[40].

3) 질병분류 지식

질병분류코드가 변경되는 일반적인 조건은 다음과 같다.

(1) 한정어에 의해 코드를 변경해야 하는 경우

질병의 특징을 한정하는 수식어에 따라 분류코드가 달라진다. 심장질환의 경우 류마티스성 여부에 따라 전혀 다른 코드로 분류된다. 해부학적인 부위 혹은 동반병태 등에 따라 더 상세한 코드로 분류되기도 한다.

(2) 대상에 따라 달라지는 코드

동일한 진단명이라도 속성에 따라 코드가 달라질 수 있다. 환자의 성별이나 나이, 신생아 및 산모 유무에 따라서 코드가 달라진다. 남성에게만 부여할 수 있는 코드, 여성에게만 부여할 수 있는 코드가 있다.

(3) 두 개의 질환 코드를 하나의 코드로 분류하는 경우

일반적으로 증상코드와 관련된 진단코드가 입력될 경우 증상코드는 삭제하여야 하며, 동반된 질환에 따라 두 개의 질환을 하나로 코드로 변경해야 한다. 증상코드의 경우 증상을 유발한 질병이 있을 경우 증상코드는 부여하지 않은 것이 원칙이나 의사가 임상적으로 중요하다고 판단할 경우 부가적으로 부여하기도 한다. 이러한 원칙은 병원이나 부서의 판단에 의해 선택적으로 사용된다. 식도염(K20)과 위식도역류병(K21.9)이 함께 있는 경우, 식도염을 동반한 위식도 역류병(K21.0)으로 분류되어야 한다.

(4) 이원분류 해야 하는 경우

특정 질병명에 대해서는 원인과 병태라는 두 가지 측면에서 코드가 2개 부여되어야 한다. 결핵성 심내막염의 경우 A18.83+과 I39.8* 코드를 함께 분류하여야 한다. 결핵균에 감염되었나는 의미인 A18.83과 그로 인하여 심내막염 I39.8이 발병하였다는 의미로 사용된다. 별표(*)가 붙은 코드는 단독으로 쓰일 수 없다.

(5) 부가코드를 부여해야 하는 경우

합병증에 의한 병태일 경우 합병증 코드와 병태코드를 각각 분류하여야 한다. 외상환자의 경우는 손상의 외인코드를 부가코드로 분류해야 한다. 후유증에 의한 병태일 경우 현재 나타난 병태와 어떤 질환의 후유증인지를 함께 분류해야 한다. 감염질환의 경우 감염체를 안다면 감염체에 대한 코드를 부여하고, 신생물의 경우 암의 형태가 밝혀졌다면 암의 형태학에 대한 코드도 함께 분류한다.

2. 질병분류 규칙 정의

1) 규칙 정의 과정

문헌 고찰에서 도출된 지식 중 순환기 계통의 지식을 대상으로 질병분류 규칙으로 정하고, 그에 대한 전문가의 해석을 기술하였다. 해석된 내용을 토대로 규칙 기반의 지식표현 방식인 IF-THEN 형식으로 지식을 정의하였다. 분류규칙 정의 과정은 Table 7과 같다.

Table 7. Process of generating classification rules

KCD-7 Guide	Expert Knowledge	Generated Rules
기관지염 J40 - 15세 미만 [급성 참조] - 급성 J20.9	Explanation 1. 기관지염은 15세 미만일 경우 급성 기관지염으로 간주한다.	Rule 1. IF Code=J40 AND Age < 15 THEN Code=J20.9
알츠하이머 G30.9 - 조기성(초로성) G30.0 - 후기성(노년) G30.1	Explanation 2. 65세 이전에 발병된 알쯔하이머병은 G30.0으로 분류한다.	Rule2. IF Code=G30.9 AND Age<65 THEN Code=G30.0
폐색전증 I26.9 - 산과적 O88.2	Explanation 2. 폐색전증 환자 중 산과환자는 O88.2를 부여한다.	Rule3. IF Code=I26.9 AND Dept=OB THEN Code=O88.2

2) 순환기 계통의 질병분류 규칙

순환기 계통의 질병분류 규칙은 두 가지 코드가 하나의 코드로 분류되는 동반코드 규칙, 특정 질환에 대해서는 원인과 병태를 이원분류 해야 하는 규칙, 대상자에 따라 대분류가 달라지는 규칙, 외인에 의한 질환일 때 외인코드를 부여해야 하는 규칙이 있다.

동반코드가 하나의 코드로 분류 되어야 하는 경우의 예는 경동맥 협착에 의한 뇌경색 환자에게 경동맥 협착(I65.2)과 뇌경색(I63.9)이라는 두 가지 병명이 입력되었다면 경동맥 협착에 의한 뇌경색 코드(I63.22)코드로 분류되어야 한다. 또한, 고혈압성 심장질환(I11.9)과 신장질환(I12.9)이 동반될 경우 두 가지 질환을

포함하는 하나의 코드(I13.9)로 분류된다. 동반 코드가 동시에 입력되는 경우에 질병분류지침에 맞는 하나의 코드로 변경되도록 동반코드를 체크하여 이러한 문제를 해결하거나 병명 선택 시 정보를 제공해야 한다.

환자가 임산부 일 경우에는 알파벳 'O'로 시작하는 코드를 부여하는 것이 원칙이다. 임신 중인 환자가 심근병증이 있다면 I42.9가 아닌 O99.4코드를 부여하고 I42.9는 부가적으로 사용할 수 있다.

순환기 계통의 질환분류 시 이원분류 해야 하며 단독코드로 쓰일 수 없는 별표 코드는 I32, I39, I41, I43, I52, I68, I79, I98로 총 8개이다. 일부 질환에 대해서는 병태인 검표 코드와 원인인 별표 코드가 함께 분류되는 경우가 있는데, 이러한 이원분류에서 검표 코드가 우선되는 코드이다. 별표 코드는 세트로 함께 분류되는 코드로 주 진단 코드나 단독 코드로 사용될 수 없다. 이러한 이원분류는 코드 간의 비교보다는 해당 질병명의 원인 병태를 찾는 규칙이기 때문에 IF-THEN 규칙으로 해결할 수가 없어 콘텐츠를 다루는 질병분류모델에서 해결하였다. 예를 들어 디프테리아에서의 심근병증은 감염성 및 기생충성 질환에서의 심근병증에 대한 I43.0과 디프테리아 감염인 A36.8 코드를 함께 분류해야 하는데, 이러한 정보는 질병명을 입력할 때 얻을 수 있는 정보이다.

외상과 관련된 질병분류코드가 입력될 경우에는 외인코드를 부가적으로 입력하도록 질병분류모델에서 설명으로 제시하여 의사의 입력을 유도하도록 하였다. 따라서 동반코드와 산과코드에 대한 전문 지식만을 대상으로 전문 지식을 기술하고 IF-THEN 형식으로 규칙을 정의하였다(Table 8).

규칙에 사용된 용어 중 'List'는 의사가 입력한 진단코드들을 의미하며, 'CONTAINS'는 뒤에 나온 코드를 포함하고 있다는 의미이다. 여러 개의 코드는 'AND'로 연결하며, 'MERGE INTO'는 앞에서 확인 한 코드들을 합친 후 뒤에 나오는 코드로 변경해 준다는 의미이다. 'ADD'는 뒤에 나오는 코드를 기존의 코드와 상관없이 'List'에 추가한다는 뜻이다.

Table 8. Classification rules for disease of circulatory diseases

구분		순환기 계통 질병의 코드분류 규칙
동반코드	지식	K1. 뇌동맥의 협착, 폐쇄(I65._, I66._)와 뇌경색(I63._)이 함께 올 경우에는 뇌경색에 대한 상세 코드(I63._)로 분류한다. K2. 고혈압성 심장질환(I11._)과 고혈압성 신장질환(I12._)이 함께 있을 경우 고혈압성 심장질환과 신장질환 코드(I13._)로 분류한다.
	규칙	Rule 1.1. IF List CONTAINS I65.0 AND I63* THEN MERGE INTO I63.20 Rule 1.2. IF List CONTAINS I65.2 AND I63* THEN MERGE INTO I63.22 Rule 1.3. IF List CONTAINS I65.1 AND I63* THEN MERGE INTO I63.21 Rule 2.1 IF List CONTAINS I11.9 AND I12.9 THEN MERGE INTO I13.9 Rule 2.2 IF List CONTAINS I11.0 AND I12.0 THEN MERGE INTO I13.2
산과코드	지식	K1. 고혈압성 질환(I10._)은이 임신, 출산 및 산후기에 합병되면 산과코드(O10._)로 분류한다. K2. 심근병증(I42._)이 산후기에 합병되면 O90.3을 분류한다. K3. 심근병증(I42._)이 임신기간에 합병되면 O99.4을 분류한다.
	규칙	Rule 1. IF List CONTAINS I10* AND Dept = OB THEN CHANGE O10.9 Rule 2. IF List CONTAINS I42* AND OB Type = puerperium AND List NOT CONTAINS O90.3 THEN ADD O90.3 Rule 3. IF List CONTAINS I42* AND OB Type = before delivery AND List NOT CONTAINS O99.4 THEN ADD O99.4

3. 질병명의 구조 분석

1) 자료 수집

질병분류의 세분화와 관련하여 질병명을 수식하는 용어와 속성을 파악하기 위해 순환기 계통의 질병명의 구조를 분석하였다. 질병분류는 질병 자체가 가지고 있는 해부학적인 상세 부위, 임상경과, 질환별 특성 등이 반영되어야 정확한 코드부여가 가능하다. 분석 자료는 통계청 홈페이지(http://kssc.kostat.go.kr:8443)에서 엑셀 파일로 제공하는 KCD-7 마스터 파일을 다운로드 한 후, 순환기 계통의 질병코드에 해당하는 I10-I99코드를 추출하였다. Fig. 16은 통계청에서 제공하는 KCD-7 엑셀 파일의 일부이다.

표제어	분류기준	질병분류코드	검별	주석	한글명칭	영문명칭	변동	최하위코드	국내세분화코드	희귀질환	한의병명	한글용어추가	비고	수정일
1	**소**	**I00**			**심장침범에 대한 언급이 없는 류마티스열**	**Rheumatic fever without mention of heart involvem**		1	0	0	0	0		
1	**소**	**I01**			**심장 침범이 있는 류마티스열**	**Rheumatic fever with heart involvement**		0	0	0	0	0		
1	세	I01.0			급성 류마티스심장막염	Acute rheumatic pericarditis		1	0	0	0	0		
1	세	I01.1			급성 류마티스심내막염	Acute rheumatic endocarditis		1	0	0	0	0		
1	세	I01.2			급성 류마티스심근염	Acute rheumatic myocarditis		1	0	0	0	0		
1	세	I01.8			기타 급성 류마티스심장병	Other acute rheumatic heart disease		1	0	0	0	0		
1	세	I01.9			상세불명의 급성 류마티스심장병	Acute rheumatic heart disease, unspecified		1	0	0	0	0		
1	**소**	**I02**			**류마티스무도병**	**Rheumatic chorea**		0	0	0	0	0		
1	세	I02.0			심장 침범이 있는 류마티스무도병	Rheumatic chorea with heart involvement		1	0	0	0	0		
1	세	I02.9			심장 침범이 없는 류마티스무도병	Rheumatic chorea without heart involvement		1	0	0	0	0		
1	**소**	**I05**			**류마티스성 승모판질환**	**Rheumatic mitral valve diseases**		0	0	0	0	0		
1	세	I05.0			승모판협착	Mitral stenosis		1	0	0	0	0		
1	세	I05.1			류마티스성 승모판폐쇄부전	Rheumatic mitral insufficiency		1	0	0	0	0		
1	세	I05.2			폐쇄부전이 있는 승모판협착	Mitral stenosis with insufficiency		1	0	0	0	0		
1	세	I05.8			기타 승모판질환	Other mitral valve diseases		1	0	0	0	0		
1	세	I05.9			상세불명의 승모판질환	Mitral valve disease, unspecified		1	0	0	0	0		
1	**소**	**I06**			**류마티스성 대동맥판질환**	**Rheumatic aortic valve diseases**		0	0	0	0	0		
1	세	I06.0			류마티스성 대동맥협착	Rheumatic aortic stenosis		1	0	0	0	0		
1	세	I06.1			류마티스성 대동맥판폐쇄부전	Rheumatic aortic insufficiency		1	0	0	0	0		
1	세	I06.2			폐쇄부전이 있는 류마티스성 대동맥협착	Rheumatic aortic stenosis with insufficiency		1	0	0	0	0		
1	세	I06.8			기타 류마티스성 대동맥판질환	Other rheumatic aortic valve diseases		1	0	0	0	0		
1	세	I06.9			상세불명의 류마티스성 대동맥판질환	Rheumatic aortic valve disease, unspecified		1	0	0	0	0		
1	**소**	**I07**			**류마티스성 삼첨판질환**	**Rheumatic tricuspid valve diseases**		0	0	0	0	0		
1	세	I07.0			삼첨판협착	Tricuspid stenosis		1	0	0	0	0		
1	세	I07.1			삼첨판폐쇄부전	Tricuspid insufficiency		1	0	0	0	0		
1	세	I07.2			폐쇄부전을 동반한 삼첨판협착	Tricuspid stenosis with insufficiency		1	0	0	0	0		
1	세	I07.8			기타 삼첨판질환	Other tricuspid valve diseases		1	0	0	0	0		
1	세	I07.9			상세불명의 삼첨판질환	Tricuspid valve disease, unspecified		1	0	0	0	0		
1	**소**	**I08**			**다발판막질환**	**Multiple valve diseases**		0	0	0	0	0		
1	세	I08.0			승모판 및 대동맥판의 장애	Disorders of both mitral and aortic valves		1	0	0	0	0		
1	세	I08.1			승모판 및 삼첨판의 장애	Disorders of both mitral and tricuspid valves		1	0	0	0	0		
1	세	I08.2			대동맥판 및 삼첨판의 장애	Disorders of both aortic and tricuspid valves		1	0	0	0	0		
1	세	I08.3			승모판, 대동맥판 및 삼첨판의 복합장애	Combined disorders of mitral, aortic and tricuspid valves		1	0	0	0	0		
1	세	I08.8			기타 다발판막질환	Other multiple valve diseases		1	0	0	0	0		
1	세	I08.9			상세불명의 다발판막질환	Multiple valve disease, unspecified		1	0	0	0	0		
1	**소**	**I09**			**기타 류마티스심장질환**	**Other rheumatic heart diseases**		0	0	0	0	0		
1	세	I09.0			류마티스심근염	Rheumatic myocarditis		1	0	0	0	0		
1	세	I09.1			상세불명 판막의 류마티스심내막질환	Rheumatic diseases of endocardium, valve unspecified		1	0	0	0	0		
1	세	I09.2			만성 류마티스심장막염	Chronic rheumatic pericarditis		1	0	0	0	0		
1	세	I09.8			기타 명시된 류마티스심장질환	Other specified rheumatic heart diseases		1	0	0	0	0		
1	세	I09.9			상세불명의 류마티스심장병	Rheumatic heart disease, unspecified		1	0	0	0	0		
1	**소**	**I10**			**본태성(원발성) 고혈압**	**Essential(primary) hypertension**		0	0	0	0	0		
2	세	I10.1			악성 고혈압	Malignant hypertension		1	1	0	0	0		
2	세	I10.9			상세불명의 원발성 고혈압	Unspecified primary hypertension	국내세분화	1	1	0	0	0		

Fig. 18. Part of KCD-7 data in Excel format provided by Statistics Korea (Source: Statistics Korea, 2016)

순환기 계통의 질환은 10개의 중분류로 분류된다(Table 9). 질병분류는 가장 상세한 코드를 부여해 주는 것이 원칙이기 때문에 소분류 밑에 세분류가 있다면 세분류 코드로 분류하고, 만약 세세분류, 세세세분류가 가능하다면 최하위 수준의 코드로 분류해야 한다. 이러한 최하위 수준의 코드를 완전한 코드라고 한다.

Table 9. Blocks of categories of disease of circulatory system in KCD-7

Blocks categories	Korean	English
I00-I02	급성 류마티스열	Acute rheumatic fever
I05-I09	만성 류마티스 심장질환	Chronic rheumatic heart diseases
I10-I15	고혈압성 질환	Hypertensive diseases
I20-I25	허혈심장질환	Ischemic heart diseases
I26-I28	폐성 심장병 및 폐순환의 질환	Pulmonary heart disease and diseases of pulmonary circulation
I30-I52	기타 형태의 심장병	Other forms of heart disease
I60-I69	뇌혈관질환	Cerebrovascular diseases
I70-I79	동맥, 세동맥 및 모세혈관의 질환	Diseases of arteries, arterioles and capillaries
I80-I89	달리 분류되지 않은 정맥, 림프관 및 림프절의 질환	Diseases of veins, lymphatic vessels and lymph nodes, NEC
I95-I99	순환계통의 기타 및 상세불명의 장애	Other and unspecified disorders of the circulatory system

Table 10은 복부 대동맥류 박리의 경우 대분류 코드부터 최하위 코드까지를 보여주는 표이다. 국가 단위의 통계 비교를 위해서는 뒤에 자리수를 때고 소분류인 'I71' 수준에서 자료를 분석하지만, 진단서 작성, 보험청구 등을 위해서는 최하위 코드를 사용해야 한다. 국내 임상현장에서 실제 코드분류에 사용해야 하는 코드는 'I71.01'이다. 이렇듯 최하위 수준의 진단코드를 부여한다는 KCD의 대원칙하에 최하위 코드에 해당하는 진단명만을 분석에 사용하였다.

Table 10. Final code of 'dissection of abdominal aorta with gangrene'

Category	Code	Description
Chapter	I00-I99	Cerebrovascular diseases
Blocks category	I70-I79	Disease of arteries, arterioles and capillaries
Three-character category	I71	Aortic aneurysm and dissection
Four-character subcategory	I71.0	Dissection of aorta
Five-character subcategory	I71.01	Dissection of abdominal aorta

순환계통의 질환에서는 소분류를 최종 코드로 사용할 수 있는 경우는 I00과 I99가 있다. I00과 I99는 하위 수준의 세분류가 없기 때문에 소분류가 최하위 코드이자 완전한 코드로 사용된다.

2) 어휘 분석

순환기 계통의 질병명 중 실제 진단명으로 사용할 수 있는 최하위 진단명은 454개며, 이러한 최하위 진단명에 대하여 수작업으로 어휘 분석을 하였다. 먼저 질병명을 원자단위로 분해하기 위해 문서편집기를 이용하여 띄어쓰기 기준으로 질병명을 분해하였다. 한글용어는 1,582의 단어로 분해되었으며, 중복을 제외하면 398개였다. 영문용어는 2,359의 단어로 분해되었고, 중복을 제외하면 338개였다.

띄어쓰기를 기준으로 작업한 자료 중 질병분류의 기준이 되는 병태와 두 단어로 분리하면 의미가 달라지는 용어에 대해서는 다시 하나의 용어로 처리하였다. 예를 들어, 'due to'는 due와 to로 분해되지만, 분해될 경우 뜻이 달라지기 때문에 복합 단어로써 의미를 가지는 용어는 하나로 처리하였다.

병태의 경우 임상적으로 의미가 부여될 수 있는 최소한의 용어를 하나의 용어로 처리하였다. 죽상경화증은 단독으로 사용되어 질병분류코드가 부여될 수 없으며, 해부학적인 부위에 따라 코드가 상이하게 부여될 수 있다. 그러나 혈전증의 경우에는 해부학적인 부위를 몰라도 대표코드를 부여할 수 있다. 뇌경색증은 영문으로 'cerebral'과 'infarction'이 조합되어야 하고, 한글인 경우 '뇌'와 '경색증'이 하나로 조합되어야만 '뇌경색증'이라는 병명으로 의미가 있다. 이렇게 처리한

최소 단위의 진단명은 '뇌경색증'이 36건으로 가장 많았고, 그 다음은 죽상경화증으로 21건이었다. 이러한 원칙에 입각하여 우선 최소 단위의 질병명을 기본개념으로 분리하였다.

기본개념을 분리한 후 남은 용어는 기본개념을 주로 수식하기 위한 형용사, 부사, 전치사, 동명사 등이었다. 이러한 나머지 수식어들에 대하여 분석하였다. 질병명을 수식하는 용어는 급성인지 만성인지 여부, 해당 질병의 발병인자, 해부학적인 부위 등 다양하였다. 이러한 수식어들을 유사한 성격으로 분류하여 최소 단위의 진단명과 어떤 의미의 관계로 맺어지는지 유형을 분석하였다. 질병명은 대부분 간단한 한정어에 의해 구체화되었으며, 경우에 따라 어떤 질환에 의한 이차적인 병태인 경우 혹은 동반되는 병태에 따라 세분화 되었다.

IV. 질병분류모델 개발

4장에서는 순환기 계통의 질병명을 구조 분석한 결과를 토대로 병명 자체의 세분화 수준에 따른 분류 지식을 공식화하여 질병분류모델을 개발하였다. 질병명을 기본개념, 속성, 속성 값으로 정의하고 그에 따라 질병분류코드가 할당되도록 하였다.

1. 질병분류모델 정의

본 연구에서 정의한 질병분류모델은 의사가 임상문서에 기록하는 질병명에 정확한 KCD-7 코드가 부여될 수 있도록 질병명을 일정한 형식으로 구조화 및 코드화한 모델이다. 질병명의 표현은 EAV(Entity-Attribute-Value) 모델 방식을 채택하였다. EAV 모델은 임상환자 기록과 관련된 데이터를 저장하는데 가장 널리 사용되는 모델이다[44]. 의료분야의 임상 용어체계와 상세임상모델도 대부분 EAV 모델을 근간으로 한다. EAV 모델은 데이터 구조를 나타내는 데이터 모델로 비교적 간단한 물리적 데이터베이스 스키마를 사용하는 매우 이질적인 데이터를 정리하는 방법으로 테이블 구조의 변화 없이 속성 셋을 확장할 수 있어 데이터를 유연하게 표현하고 저장할 수 있다. EAV 접근법은 네임-밸류 쌍(name-value pair) 접근법이라고도 한다[45].

기존의 상세임상모델들은 임상자료를 표현하는 속성들이 주로 해당 개념의 일부가 아닌 의미적으로 연결된 외부의 속성 값으로 구성된다면, 질병분류모델의 속성 값은 해당 개체를 구체적으로 명시해주는 내부 속성이라는 점과 EAV 조합에 따른 모델 식별자와 KCD-7 코드가 존재한다는 점에서 명확한 차이가 있다. 즉, 질병분류모델은 EAV 조합에 따라 질병분류코드를 제시하는 지식 모델이다. 또한 의미 해석과 정보의 호환을 위해 표준용어체계의 개념코드를 사용하였다. SNOMED CT는 후조합을 지원하는 용어체계이기 임상정보를 표현하는데 필요

한 다양한 개념을 포함하며, 국제적인 표준코드라는 장점 때문에 질병분류모델에 사용된 용어들을 SNOMED CT의 개념코드와 매핑하여 지식을 표현하였다.

본 모델에서는 질병분류가 가능하고 진단명으로 의미가 있는 최소단위의 질병명을 기본개념(Base concept)으로 하여 이를 부연 설명하는 수식어를 속성 명(Attribute)과 속성 값(Value)로 구분한 후 사용자가 선택한 용어의 조합에 의하여 질병분류코드(Code)가 할당되도록 구성하였다. 질병분류모델은 포괄적인 용어이기 때문에 본 연구에서 개발한 질병분류모델을 BAVC(Base concept, Attribute, Value and Code) 모델이라고 명명하였다. 다음 Fig. 17은 KCD-7 분류책자의 지식 구조를 BAVC 지식 모델로 표현한 것이다.

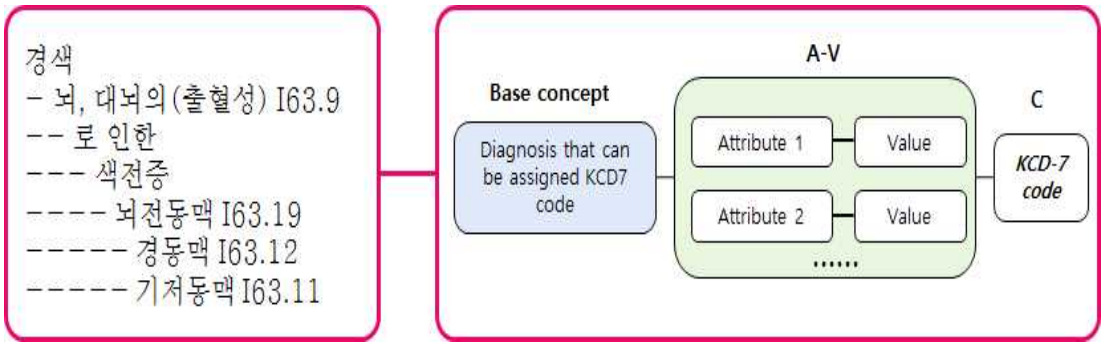


Fig. 19. Structure of KCD-7 character index and BAVC abstract model

2. BAVC 모델의 기본 요소

1) 기본개념

질병분류모델에서의 기본개념은 의사가 진단명을 입력하기 위해 독립적으로 사용할 수 있는 최소 단위의 질병명이다. 폐쇄(occlusion)는 하나의 임상개념이지만 진단명으로 단독 사용되기에는 의미가 매우 광범위하며, 해부학적 부위가 명시되지 않고서는 질병분류 코드를 부여할 수 없기 때문에 기본개념에 해당되지 않는다. 반면, 동맥류(aneurysm)는 단독 사용하여도 질병분류 코드(I72.9)를 부여할 수 있기 때문에 기본개념으로 선정하였다. 이처럼 기본개념은 단독으로 사용되어 질병분류 코드가 부여될 수 있으며, 속성 값에 의하여 더 구체적인 코드로

분류될 수 있다는 것을 전제조건으로 한다. 질병명은 동의어, 약어, 영문명, 한글명 등 표현의 다양성 때문에 개념단위로 모델을 개발하였다.

2) 속성

최소 단위의 질병명은 다양한 수식어(modifier)에 의하여 구체화될 수 있다. 속성은 최소 단위의 질병명을 상세하게 표현해 주기 위해 필요한 개념으로 속성값에 의하여 구체화된다. 속성의 유형을 보면 첫째, 광범위한 기본개념을 하위유형으로 구분하는 하위유형 한정어(subtype qualification)가 있다. 하위유형 수식자에는 한정어(qualification), 속성 정의의 세분화(refinement), 수용할 수 없는 수식어의 추가, 중첩된 수식어의 추가 등이 해당된다. 예를 들어, 고혈압은 원발성 혹은 속발성이라는 하위 유형 수식자에 의해 구체화 될 수 있다. 두 번째 유형은 개체를 수정하거나 부정하는 변경자(axis modification)에 의해 의미가 변경시킬 수 있다. 변경자는 질병에 대한 부정, 현재는 없음, 불확실성, 혹은 질병의 대상을 변경하기 위하여 사용된다. 예를 들어, 신생아나 산모의 경우 동일한 질병이라도 코드가 달라지기 때문에 질병분류 대상이 신생아일 때는 변경자에 의하여 코드가 신생아 코드로 변경될 수 있다. 질병명 분석을 통해 도출된 속성 값들은 ISO/TC 215의 환자 임상소견에 대한 개념적 틀, openEHR의 diagnosis 아키타입과 SNOMED CT의 clinical finding 중 disease의 속성명(attribute name)을 참고로 하여 최종 속성명을 확정하였다[19,30,43].

3) 속성 값과 속성 셋

질병명에 따른 선택 가능한 속성 값들을 구분하기 위하여 KCD-7의 질병명을 분석하여 속성을 나열하고, 해당 질병명이 가질 수 있는 속성 값을 정의하였다 속성 값은 용어사전에 있는 모든 용어들이 될 수 있지만, 동질의 속성 값을 묶어 속성 셋을 구성할 수 있다. 발병 부위는 해부학적 부위 중 병이 발생한 부위로 전신의 모든 부위가 대상이 될 수 있지만, 뇌경색의 발병 부위는 뇌혈관과 관련된 해부학적인 부위로 제한된다. 마찬가지로 심근경색증의 발병 부위는 심근과 관련된 부위로 제한되기 때문에 이러한 값들을 묶어 속성 셋을 구성하였다. 속성 셋은 해당 모델이 취할 수 있는 값을 제한하는 목적으로 사용된다.

4) 코드

코드는 질병명의 기본개념과 속성 값을 모두 선택한 후 최종 부여되는 KCD-7 코드이다. 충수염 환자는 임상경과에 의해 코드가 변경될 수 있다. 임상경과가 급성이면 코드는 J35.8, 만성이면 K36으로 분류된다. 만약 발병기간의 언급이 없어서 만성인지 급성인지 알 수 없을 경우에는 K37로 분류된다(Table 11). 일반적으로 하나의 질병명에는 하나의 코드가 부여되지만, 일부 질병명에 대해서는 이원분류를 해야 한다는 원칙을 반영하여 KCD-7 코드는 한 개 이상일 수도 있다. 예를 들어 영양성 심근병증(nutritional cardiomyopathy)의 경우 이원분류 대상 질병이기 때문에 E63.9와 I43.2코드가 함께 부여된다.

Table 11. Core components of BAVC model

Component	Description		Case
Base concept	The minimum unit of disease name available for diagnosis		Appendicitis
Attribute	Name	Element that explain a basic concept	Clinical course
	Value	Specific attribute value of the concept	Acute
Code	KCD-7 code assigned to diagnosis		J35.8

3. BAVC 모델링

1) 코드별 질병명 분해

KCD-7 코드별 해당하는 질병명을 기본개념, 속성, 속성 값으로 해체한 후 BAVC 모델을 개발하였다. 모델링은 질병의 소분류에 해당하는 코드와 그에 속하는 하위 질병명을 모두 나열한 후 BAVC 요소 수준으로 용어를 분해한 후 BAVC로 조합하는 방식이다.

예를 들어, 심장 침범이 있는 류마티스 열인 I01에 해당하는 하위 질병명을 모

두 나열한 후 I01의 하위 코드에 해당하는 코드와 질병명을 나열한다. ‘I01.0’의 질병명은 ‘급성 류마티스성 심내막염’, ‘급성 류마티스성 심막염’ 등이 해당되기 때문에 코드를 중심으로 기본개념을 추출하고 모델을 작성하였다(Fig. 18). ‘I01.0’의 기본개념은 심내막염과 심막염 두 가지 이다.

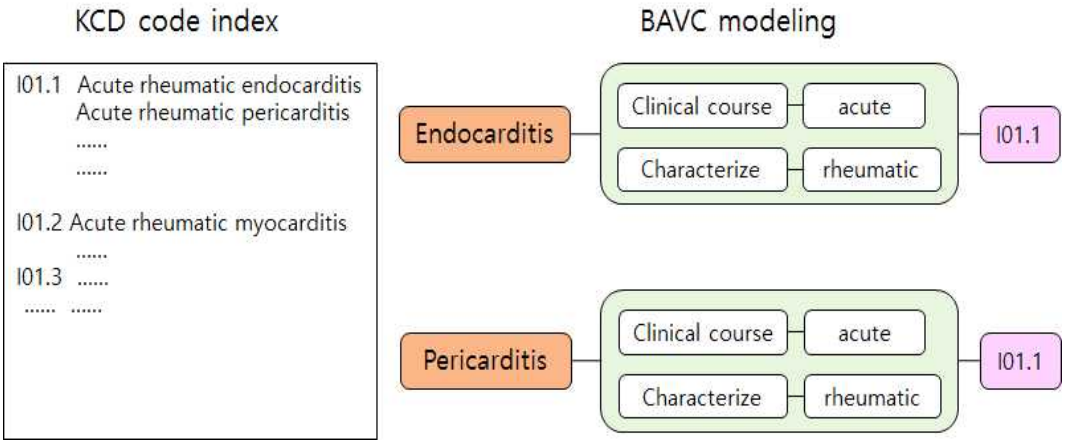


Fig. 20. Decomposition disease name to BAV component

2) 기본개념별 범주화

질병분류 코드별로 BAVC 모델링이 끝나면 질병분류 코드가 아닌 질병의 기본개념별로 모델을 범주화 하였다 . ‘I01.1’처럼 동일한 코드라도 기본개념은 심내막염과 심막염 두 가지로 정리될 수 있기 때문에 KCD-7 코드별 질병명을 모델링한 후 동일한 기본개념을 모아 기본개념별로 모델을 범주화 하였다. 기본개념별로 범주화를 끝내면, 기본개념을 수식하기 위해 필요한 모든 속성을 도출해 낼 수 있다. 다음 Fig. 19는 기본개념이 심내막염인 질병명들을 하나로 모아 범주화 한 것이다. 모델링 작업은 인스턴스를 개발하면서 동시에 수행된다.

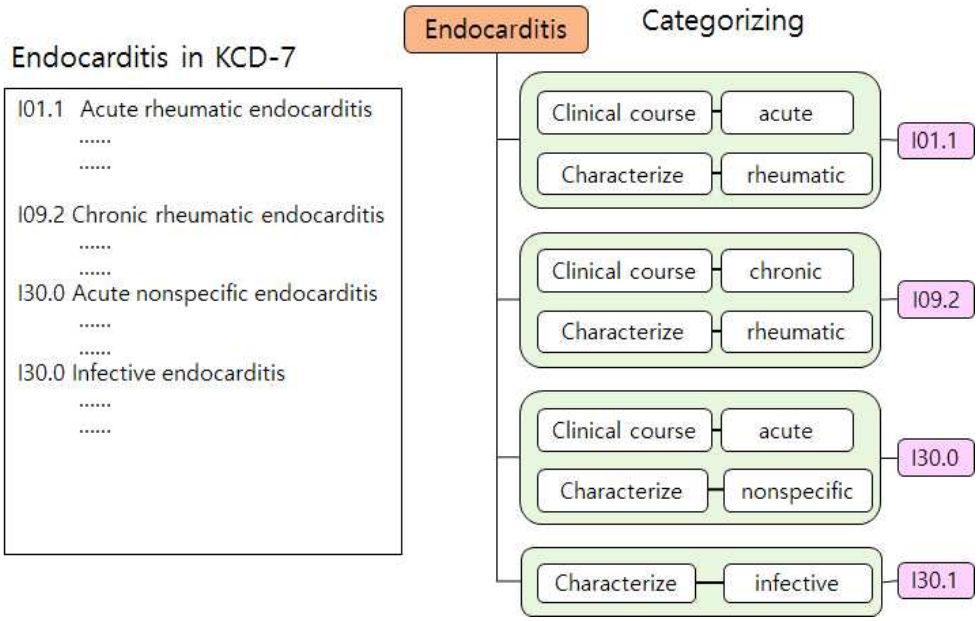


Fig. 21. Categorization of BAV models by basic concept

4. BAVC 인스턴스 개발

1) 모델 개발 범위

모델 개발의 우선순위를 순환기 계통의 다빈도 질병명으로 정하고, 국가통계포탈 사이트(http://kosis.kr)를 통해 국민건강보험공단에서 제공하는 '질병소분류별 입원 다빈도 질병 급여현황(2014년)'을 확인하였다. 이 중 2017년도 개정된 KCD-7을 기준으로 순환기 계통의 질환에서 제외된 치핵(I84)과 심뇌혈관 질환과 관련이 적은 림프절염(I88)을 제외하고 상위 20개의 질병명을 선정하였다(Table 12). 국가단위의 통계는 KCD-7의 소분류 단위에서 집계하기 때문에 기타 뇌혈관 질환 코드인 I67은 파열되지 않은 대뇌동맥류, 대뇌 죽상경화증, 모야모야 병, 고혈압성 뇌병증 등 다양한 질환을 포함한다. 이러한 경우에는 세분류의 질병명을 육안으로 확인하여 임상 전문가의 경험에 의해 다빈도 질환이라고 판단되는 질병명을 선정하였다. I67의 경우 파열되지 않은 대뇌동맥류와 모아모아병이 다빈도 질환으로 판단되었다.

Table 12. Inpatient top diagnosis of the circulatory system in Korea (Source: National health insurance service, 2014)

Ranking	Code	Diagnosis	Number of Inpatients
1	I63	Cerebral infarction	93,838
2	I20	Angina pectoris	84,805
3	I83	Varicose veins of lower extremities	43,927
4	I10	Essential(primary) hypertension	43,387
5	I69	Sequelae of cerebrovascular disease	29,805
6	I67	Other cerebrovascular diseases	27,215
7	I25	Chronic ischemic heart disease	24,917
8	I21	Acute myocardial infarction	24,635
9	I50	Heart failure	20,288
10	I61	Intracerebral hemorrhage	20,084
11	I48	Atrial fibrillation and flutter	15,834
12	I60	Subarachnoid hemorrhage	9,560
13	I47	Paroxysmal tachycardia	6,176
14	I42	Cardiomyopathy	5,846
15	I11	Hypertensive heart disease	5,780
16	I65	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction	5,762
17	I49	Other cardiac arrhythmias	5,597
18	I70	Atherosclerosis	5,573
19	I62	Other nontraumatic intracranial hemorrhage	4,368
20	I71	Aortic aneurysm and dissection	4,316

2) BAVC 인스턴스

(1) 기본개념

순환기 계통의 20대 다빈도 질병분류 코드를 중심으로 총 27개 모델에 대한 BAVC 인스턴스를 개발하였다. 27개의 모델 중에는 다빈도 질병의 코드가 아니더라도 질병분류 규칙과 검증에 사용할 코드가 소수 포함되었다. 기본개념별로

가능한 질병명을 BAV 용어로 조합한 후 질병분류코드를 부여하였다. 구체적인 속성이 많을수록 기본개념별 인스턴스의 수도 많아졌다. 이렇게 개발된 질병분류 모델의 인스턴스는 총 290개이다(Fig. 20). 27개의 기본개념은 뇌경색, 협심증, 하지의 정맥류, 고혈압, 뇌혈관 질환의 후유증, 심장병, 심근경색, 심부전, 뇌출혈, 심방세동, 심방조동, 빈맥, 심근병증, 뇌동맥 폐쇄, 뇌동맥 협착, 부정맥, 동맥경화증, 죽상경화증, 대뇌동맥류, 대뇌동맥 박리, 판막염, 신장병, 모아모아병, 심장막염, 심내막염, 심근염, 뇌혈관 사고이다.

Base concept	A1	V1	A2	V2	A3	V3	C
뇌경색	원인	혈전증	부위	뇌전동맥	상세부위	척추동맥	I63.00
뇌경색	원인	혈전증	부위	뇌전동맥	상세부위	기저동맥	I63.01
뇌경색	원인	혈전증	부위	뇌전동맥	상세부위	경동맥	I63.02
뇌경색	원인	혈전증	부위	뇌전동맥	상세부위	기타 뇌전동맥	I63.08
뇌경색	원인	혈전증	부위	뇌전동맥	상세부위	상세불명의 뇌전동맥	I63.09
뇌경색	원인	혈전증	부위	대뇌동맥	상세부위	중대뇌동맥	I63.30
뇌경색	원인	혈전증	부위	대뇌동맥	상세부위	전대뇌동맥	I63.31
뇌경색	원인	혈전증	부위	대뇌동맥	상세부위	후대뇌동맥	I63.32
뇌경색	원인	혈전증	부위	대뇌동맥	상세부위	소뇌동맥	I63.33
뇌경색	원인	혈전증	부위	대뇌동맥	상세부위	기타 대뇌동맥	I63.38
뇌경색	원인	혈전증	부위	대뇌동맥	상세부위	상세불명의 대뇌동맥	I63.39
뇌경색	원인	색전증	부위	뇌전동맥	상세부위	척추동맥	I63.10
뇌경색	원인	색전증	부위	뇌전동맥	상세부위	기저동맥	I63.11
뇌경색	원인	색전증	부위	뇌전동맥	상세부위	경동맥	I63.12
뇌경색	원인	색전증	부위	뇌전동맥	상세부위	기타 뇌전동맥	I63.18
뇌경색	원인	색전증	부위	뇌전동맥	상세부위	상세불명의 뇌전동맥	I63.19
뇌경색	원인	색전증	부위	대뇌동맥	상세부위	중대뇌동맥	I63.40
뇌경색	원인	색전증	부위	대뇌동맥	상세부위	전대뇌동맥	I63.41
뇌경색	원인	색전증	부위	대뇌동맥	상세부위	후대뇌동맥	I63.42
뇌경색	원인	색전증	부위	대뇌동맥	상세부위	소뇌동맥	I63.43
뇌경색	원인	색전증	부위	대뇌동맥	상세부위	기타 대뇌동맥	I63.48
뇌경색	원인	색전증	부위	대뇌동맥	상세부위	상세불명의 대뇌동맥	I63.49
뇌경색	원인	폐쇄	부위	뇌전동맥	상세부위	척추동맥	I63.20
뇌경색	원인	폐쇄	부위	뇌전동맥	상세부위	기저동맥	I63.21
뇌경색	원인	폐쇄	부위	뇌전동맥	상세부위	경동맥	I63.22
뇌경색	원인	폐쇄	부위	뇌전동맥	상세부위	기타 뇌전동맥	I63.28
뇌경색	원인	폐쇄	부위	뇌전동맥	상세부위	상세불명의 뇌전동맥	I63.29
뇌경색	원인	폐쇄	부위	대뇌동맥	상세부위	중대뇌동맥	I63.50
뇌경색	원인	폐쇄	부위	대뇌동맥	상세부위	전대뇌동맥	I63.51
뇌경색	원인	폐쇄	부위	대뇌동맥	상세부위	후대뇌동맥	I63.52
뇌경색	원인	폐쇄	부위	대뇌동맥	상세부위	소뇌동맥	I63.53
뇌경색	원인	폐쇄	부위	대뇌동맥	상세부위	기타 대뇌동맥	I63.58
뇌경색	원인	폐쇄	부위	대뇌동맥	상세부위	상세불명의 대뇌동맥	I63.58
뇌경색	원인	협착	부위	뇌전동맥	상세부위	척추동맥	I63.20
뇌경색	원인	협착	부위	뇌전동맥	상세부위	기저동맥	I63.21
뇌경색	원인	협착	부위	뇌전동맥	상세부위	경동맥	I63.22
뇌경색	원인	협착	부위	뇌전동맥	상세부위	기타 뇌전동맥	I63.28
뇌경색	원인	협착	부위	뇌전동맥	상세부위	상세불명의 뇌전동맥	I63.29
뇌경색	원인	협착	부위	대뇌동맥	상세부위	중대뇌동맥	I63.50

Fig. 22. Part of BAVC model instances

290개의 인스턴스 중 뇌경색에 대한 인스턴스가 47개로 가장 많았다. 그 다음으로 뇌출혈이 39개, 뇌동맥 폐쇄 및 협착이 30개, 동맥경화증과 죽상경화증이 23개였다(Table 13). 이처럼 뇌혈관 질환의 경우 뇌혈관의 해부학적인 부위에 따라서 코드가 달라지기 때문에 조합에 따른 질병분류 코드의 수가 많아 졌다. 반면, 모아모아병 등은 1개의 규칙으로 질병분류가 가능하였다. 1개의 인스턴트만 개발된 판막염이나 심근염 등은 실제 여러 가지 속성에 의해 더 많은 조합 규칙으로 나타낼 수 있지만, 개발 우선순위에서 제외되어 하나의 조합만이 정의되었다. BAV의 조합을 따르는 BAVC 모델에 의해 다양한 조합 규칙을 만들어 낼 수 있다.

Table 13. 290 Instances created by BAVC model

Model	Number of instances	Model	Number of instances
Cerebral infarction	47	Heart disease	6
Cerebral hemorrhage	39	Varicose veins	4
Cerebral arterial occlusion	30	Atrial fibrillation	4
Cerebral arterial stenosis	30	Tachycardia	4
Arteriosclerosis	23	Arrhythmia	4
Atherosclerosis	23	Renal disease	2
Myocardial infarction	11	Atrial flutter	2
Hypertension	9	Cerebral aneurysm	2
Cardiomyopathy	9	Cerebral artery dissection	2
Heart failure	8	Moyamoya disease	1
Pericarditis	8	Cerebrovascular accident	1
Angina pectoris	7	Valvulitis	1
Sequelae of cerebrovascular disease	6	Myocarditis	1
Endocarditis	6		

(2) 속성

어휘 분석을 통해 최종 14개의 속성이 도출되었지만, 실제 인스턴스 개발에 사용된 속성은 6개였다. 290개의 인스턴스 모델에서 첫 번째 수식어로 온 속성에는 동반병태, 발병부위, 순서, 원인, 임상경과, 특징이 있었고, 두 번째 수식어로는 동반병태, 발병부위, 상세부위, 원인, 임상경과, 특징이 있었다. 세 번째 수식어는 동반병태, 발병 부위, 상세부위, 열린 상처 여부이다. 열린 상처를 구분하는 속성은 순환기 계통의 질환명(I00-I99)을 분석하면서 도출되지 않은 속성이었으나 외상에 의한 뇌출혈일 경우(S06)에 필요한 속성이기 때문에 모델 검토과정에서 동반병태('associated with') 속성으로 처리하였다. 또한 상세부위는 발병부위의 서브셋에 포함시켰다. 하나의 질병명이 갖는 속성의 최대 개수는 3개였다. 인스턴스를 만들면서 사용된 속성은 6개지만, 차후 모델의 확장성을 고려하여 어휘 분석에서 도출된 14개의 속성들을 모두 유지하였다. Table 14는 어휘 분석에서 도출된 14개의 속성들이다.

Table 14. Attributes drawn through the structural analysis of disease name

Attribute	Description
Finding site	발병 부위
Episodicity	에피소드/새로운 것인지 재발인지
Clinical course	임상경과 (만성, 급성 등)
Severity	중증도 (경증, 중증 등)
Associated morphology	관련된 형태학
Causative agent	원인인자 (바이러스, 세균 등)
Pathological process	병리과정
Occurrence	발생시점
After	후에
Due to	원인(에 의한), 원인병태
Associated with	와 관련된, 와 동반된, 동반병태
Finding method	발견방법
Characterizes	구체적인 특성
Order	질병의 순서 (원발성, 이차성 등)

(3) 속성 값과 속성 셋

기본개념별로 수식 가능한 속성과 속성 값이 있다. 속성 값의 경우 속성명이 같아도 질병명에 따라 속성의 값은 달라질 수 있으므로 해당질병에서 사용가능한 유효한 값을 속성 셋으로 구성하였다. 뇌경색의 원인은 혈관의 색전증, 혈전증, 폐색, 협착에 의해 발병이 되며 이에 따라 KCD-7 코드가 달라진다. 발병 부위는 뇌동맥의 상세 부위 중에 선택해야 하므로 뇌경색의 발병부위는 뇌동맥의 해부학적 상세 부위로 제한시켰다.

구성된 속성 셋은 유사한 수식이 가능한 질병명에 재활용될 수 있다. 예를 들어 죽상경화증과 동맥경화증은 발병 부위에 해당하는 동맥경화부위 속성 셋을 미리 정의해 두면 정의된 경로를 통해 공통으로 사용할 수 있다. 동맥 중 사지동맥과 말초동맥의 경우에는 괴저 여부에 따라서만 질병분류코드가 달라지지만, 일부 동맥의 경우에는 괴저, 궤양, 휴식통, 간헐적 파행과 같은 여러 동반병태의 유무에 따라 코드가 달라진다. 또 다른 동맥의 경우에는 동반 병태를 구분하지 않는다. 이럴 경우에는 동일한 속성 셋을 사용할 수 없기 때문에 동맥경화 부위 즉 속성 값에 따라서 다음 선택될 속성의 서브셋이 결정되도록 하였다. 서브셋도 재활용이 가능하다. Fig. 21은 이러한 방법으로 작성된 BAV 콘텐츠이다. 기본개념별 선택 가능한 속성 값에 의하여 질병분류 코드를 할당하여 최종 BAVC 인스턴스 모델을 개발하였다.

Base Concept (SNOMED Concept ID	Attribute	ValueSet	ValueItems	Subset
atherosclerosis	Finding Site	동맥경화 부위	내장동맥	subset(동반병태, 동맥 동반병태)
죽상경화증	363698007		쇄골하동맥	subset(동반병태, 동맥 동반병태)
38716007			겨드랑동맥	subset(동반병태, 동맥 동반병태)
			기타동맥	subset(동반병태, 동맥 동반병태)
	Restriction of attribute value (Finding Site: arteries)		사지동맥	subset(동반병태, 사지동맥 동반병태)
			말초동맥(사지)	subset(동반병태, 사지동맥 동반병태)
			신장동맥	subset(동반병태, 동맥 동반병태)
			대동맥	subset(동반병태, 동맥 동반병태)
			관상동맥	
			장간막	
			폐동맥	
			대뇌동맥	*
	Associated with	사지동맥 동반병태	간헐적 파행	
	(20401003)		괴저	
			궤양	
			후식통	
		동맥 동반병태	괴저	

Fig. 23. Value sets and subsets of atherosclerosis

속성 셋은 경우에 따라 서브셋을 가질 때도 있다. 발병 부위는 해부학적 부위에 따라 더 상세한 부위로 세분화 될 수 있고, 때에 따라서는 양측편측에 대한 선택도 가능해야 한다. Table 15는 속성과 속성 셋에 대한 일부이다.

Table 15. Part of attributes and values by basic concept

Base concept	Attribute	Value set	Value items
Cerebral infarction	Due to	CauseofCI	Embolism Thrombosis Occlusion Stenosis
	Finding Site	SiteofCI	Precerebral artery - Vertebral artery - Basilar artery - Carotid artery - Other artery - unspecified artery
			Cerebral artery - Middle Cerebral artery - Anterior Cerebral artery - Posterior Cerebral artery - Cerebellar artery - Other artery - Unspecified artery
Angina pectoris	Characterizes	Featureof -Angina	Unstable Spasm-induced Stable Effort Atypical Other forms
Varicose vein of lower extremities	Associated with	Conditionwith Varicose	Ulcer Inflammation Ulcer and inflammation
Endocarditis	Clinical course	Courseof -Endocarditis	Acute Subacute Chronic
	Characterizes	FeatureofEndo -carditis	Rheumatic Bacterial Gonococcal Typhoid Infective
	Finding Site	SiteofHeart -Valve	Tricuspid valve Mitral valve Aortic valve Pulmonary valve Multiple valve

3) BAVC 모델의 유형

BAVC 인스턴스를 개발한 결과 질병명을 수식하는 유형에 따라 몇 가지 형태로 구분되었다. 기본개념 자체가 최종 진단명이 되어 수식어가 필요 없거나 혹은 하나 이상의 수식어에 따라 코드가 부여되었다. 선택된 수식 값에 따라 해당 수식어를 부연 설명될 수 있는 수식어 또한 하나 이상 존재할 수 있다. 개발된 모델들은 수식어의 유형에 따라 다음과 같이 분류된다.

(1) 기본개념 형태

기본개념 형태는 최소 단위의 질병명을 그대로 사용하는 경우이다(Fig. 22). 모아모아 병은 수식어가 필요 없기 때문에 속성의 선택 없이 기본개념 단독으로 사용된다. 뇌경색의 경우에 질병명을 작성할 당시 뇌경색이 생긴 원인과 부위를 알 수 없다면 뇌경색이라는 기본개념을 진단명으로 사용할 수는 있지만, 뇌경색 모델은 원인과 부위라는 수식어의 선택이 가능하므로 기본개념 형태로 보지 않는다. 모아모아병(I67.5), 뇌혈관사고(I64) 모델이 기본개념 형태이다.

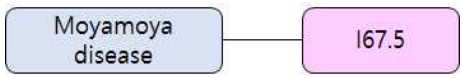


Fig. 24. Base concept form

(2) 단일 수식 형태

단일 수식 형태는 최소 단위의 질병명을 수식하는 수식어가 하나인 경우이다(Fig. 23). 예를 들어, 심근경색은 특성이라는 하나의 속성에 따라 허혈성 심근경색, 확장성 심근경색으로 한정될 수 있다. 협심증, 하지 정맥류, 심장막염, 심근염, 뇌혈관 사고의 후유증, 대뇌동맥류, 대뇌동맥 박리, 심방세동, 심방조동, 빈맥, 심장부정맥, 심근병증, 심내막염, 판막염 모델이 단일 수식형태였다. 개발된 모델 중 단일 수식형태가 가장 많았다.

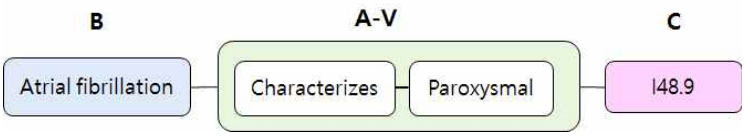


Fig. 25. Single modifier form

(3) 다수 수식 형태

다수 수식 형태는 최소 단위의 질병명을 수식하는 수식어가 여러 개이면서, 각 수식어가 독립적으로 질병명을 수식하는 형태이다(Fig. 24). 따라서 속성의 나열 순서나 선택 순서에 제한이 없다. 급성 감염성 심내막염은 급성과 감염성이라는 수식어가 심내막염을 직접 한정하므로 '감염성'과 '급성'의 선택 순서를 바꾸어도 의미상으로는 문제가 되지 않는다. 뇌동맥 폐쇄, 뇌동맥 협착, 심근경색증 모델이 다수 수식 형태에 해당되었다.

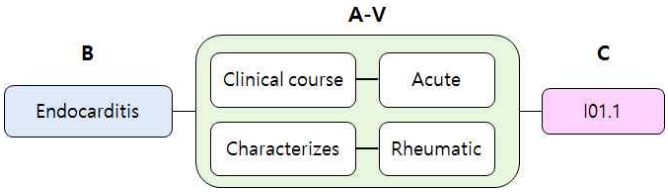


Fig. 26. Multiple modifier form

(4) 단계별 수식 형태

단계별 수식형태는 질병명을 수식하는 수식자가 또 다른 수식어에 의하여 더 구체화 될 수 있는 경우이다. 속성과 속성 값이 중첩되는 형태이다(Fig. 25). 다수 수식 형태와의 차이는 반드시 단계적으로 속성을 선택해야 한다는 점이다.

뇌출혈의 경우 첫 번째 속성 값이 외상성이면 뇌막을 중심으로 출혈 부위를 선택해야 하지만 외상성이 아닐 경우에는 출혈된 뇌동맥의 상세부위를 선택하여야 한다. 죽상경화증의 경우 해부학적 부위에 따라 괴저 유무를 구별하지 않는 경우도 있기 때문에 발병 부위를 선택한 후에야 괴저 유무에 대한 선택 여부를 판단할 수 있다. 이러한 형태는 속성의 순서가 중요하며, 순서를 지켜야 한다. 고

혈압의 경우에도 이차성 고혈압의 경우에만 고혈압의 원인 질환을 선택할 수 있다. 심부전, 뇌출혈, 동맥경화증, 죽상경화증, 고혈압, 심장병, 신장병 모델이 단계별 수식 형태에는 해당되었다.

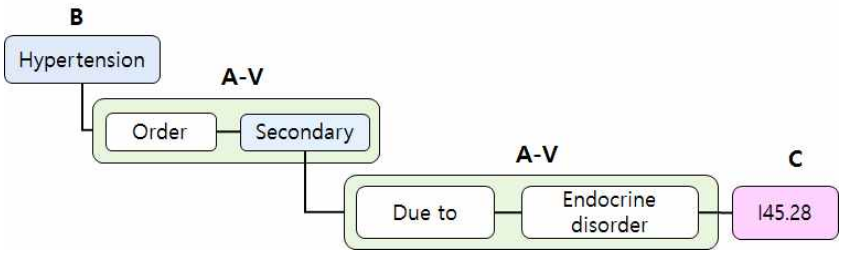
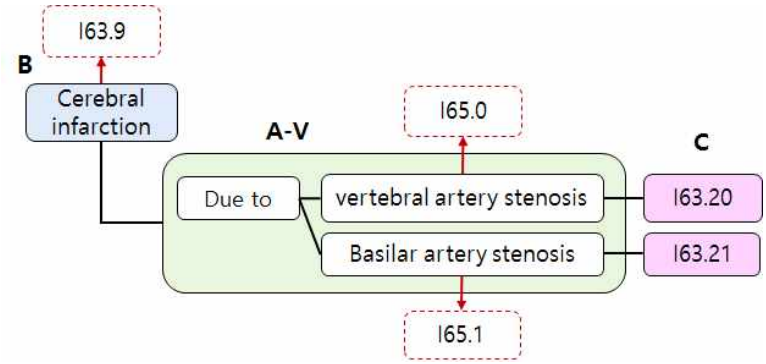


Fig. 27. Step-by-step modifier form

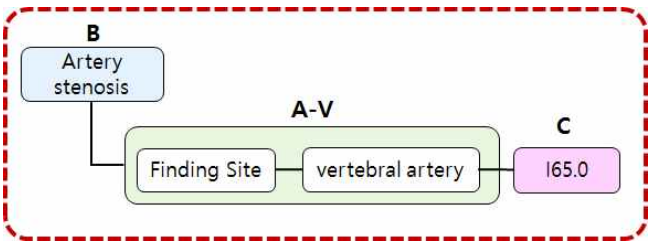
(5) 중복 병명 형태

중복 병명 형태는 두 가지의 질병명이 중복된 형태이다. 뇌경색이라는 포괄적인 질병명이 '뇌전동맥의 혈전증'이라는 원인 병태에 의하여 더욱 구체화될 수 있다(Fig. 26 (a)). '뇌전동맥의 혈전증'은 이미 개발된 뇌동맥 협착 모델에 의해 표현될 수 있기 때문에(Fig. 26 (b)) 뇌전동맥의 혈전증에 의한 뇌경색은 이미 개발된 모델을 재사용하여 단계별 수식 형태로 변경이 가능하다(Fig. 26 (c)). 주로 동반 병태, 원인 병태 등이 분류코드에 영향을 미칠 경우에 중복 병명 형태에 해당된다. 본 연구에서는 사용자 인터페이스를 고려하여 중복 병명 형태는 단계별 수식 형태로 수정하였다.

(a) Cerebral infarction model: overlapped condition form



(b) Artery stenosis model: single modifier form



(c) Modified cerebral infarction model: step by step modifier form

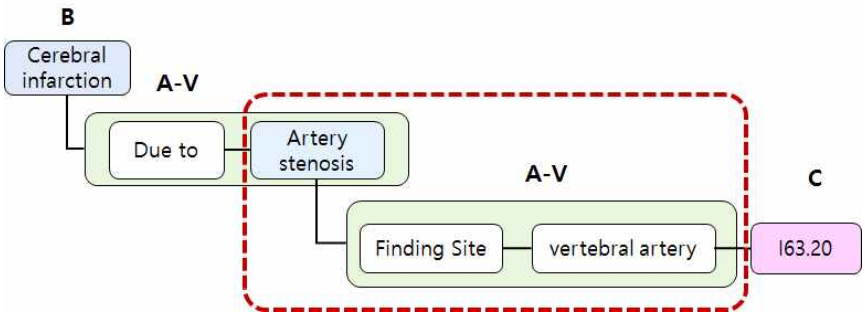


Fig. 28. Process to correct the model overlapped condition form in other model

Table 16은 개발된 질병분류 모델명과 모델의 수를 수식어의 형태별로 분류한 표이다.

Table 16. BAVC models by model type

Type	Model	Number of models
Basic concept form	Moyamoya disease, Cerebrovascular accident	2
Single modifier form	Angina pectoris, Varicose veins, Pericarditis, Sequelae of cerebrovascular disease, Cerebral aneurysm, Myocarditis Cerebral artery dissection, Valvulitis	14
Multiple modifier form	Atrial flutter, Atrial fibrillation, Tachycardia, Arrhythmia, Cardiomyopathy, Endocarditis, Cerebral arterial occlusion, Cerebral arterial stenosis, Myocardial infarction	3
Step by step modifier form	Heart failure, Cerebral hemorrhage, Cerebral infarction, Heart disease, Renal disease, Arteriosclerosis, Atherosclerosis, Hypertension	8
Total		27

4) 표준용어체계와의 매핑

(1) 표준용어체계의 선정

질병분류모델은 질병명과 질병분류코드를 상세한 수준으로 정확하게 부여할 목적으로 고안된 모델이면서 정보의 상호운용성과 활용성을 염두에 두고 만든 모델이다. 하나의 질병명을 표현하기 위해 사용된 각 용어에 표준용어체계를 매핑하면 컴퓨터의 의미적 해석이 가능하며, 이에 따라 질병분류 및 의미 있는 정보처리가 가능하다. 이에 BAVC 모델에서 사용할 표준용어체계를 선정하였다.

먼저 국내에서 표준으로 채택된 한국보건의료용어표준인 KOSTOM 코드와 매핑을 시도하였다. 사회보장정보원에서 운영하는 KOSTOM 온라인 브라우저(http://www.hins.or.kr)를 통해 용어를 검색하여 개념코드를 검토하였다. 그러나 질병명의 수식어를 따로 표현할 수 있는 개념이 많지 않고, 있다고 하더라도 공식적인 정의나 의미 관계의 부재로 해당 용어의 의미를 정확하게 표현할 수 없

었다. Fig. 27에서 보듯이 급성(acute)을 검색하면 두 개의 급성이 검색되는데 하나는 간호 범주에 속하는 용어이고 하나는 기타 범주에 속하는 용어이다.

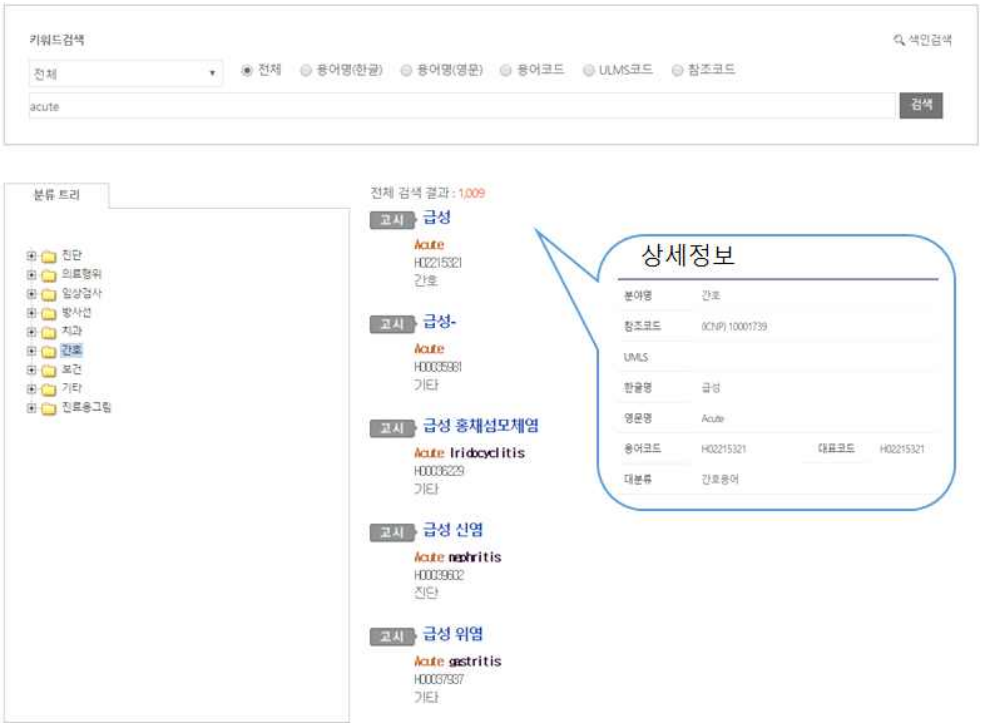


Fig. 29. Results of searching 'acute' in KOSTOM
(Source: https://www.hins.or.kr)

기본개념 단위인 질병명은 KCD-7 용어와 대, 중, 소분류인 계층구조를 그대로 차용하고 있기 때문에 매핑의 의미가 없다고 판단되었다. 이에 미국과 유럽 등 많은 나라에서 표준임상용어로 채택된 SNOMED CT를 본 모델의 표준용어 체계로 선정하였다.

(2) SNOMED CT 매핑 방법

기본개념, 속성명, 속성 값으로 올 수 있는 모든 용어에 SNOMED CT의 개념코드를 매핑하였다. 개발된 BAVC 모델에 사용된 용어가 많지 않고, 정확성을 기하기 위하여 SNOMED CT 온라인 검색 브라우저(http://browser.ihtsdotools.org)를 이용하여 용어를 일일이 검색하였다. 검색한 용어의 공식적인 정의와 개

념 관계를 확인한 후 가장 적합한 코드를 찾아 수동매핑 하였다. SNOMED CT는 2016년 6월 국제판 버전(international edition, July, 2016 version)을 사용하였으며 한글과 한글검색을 지원하지 않기 때문에 KCD-7에서 제시한 영문명을 사용하였다. Fig. 28은 SNOME CT 온라인 브라우저의 초기 화면이다.

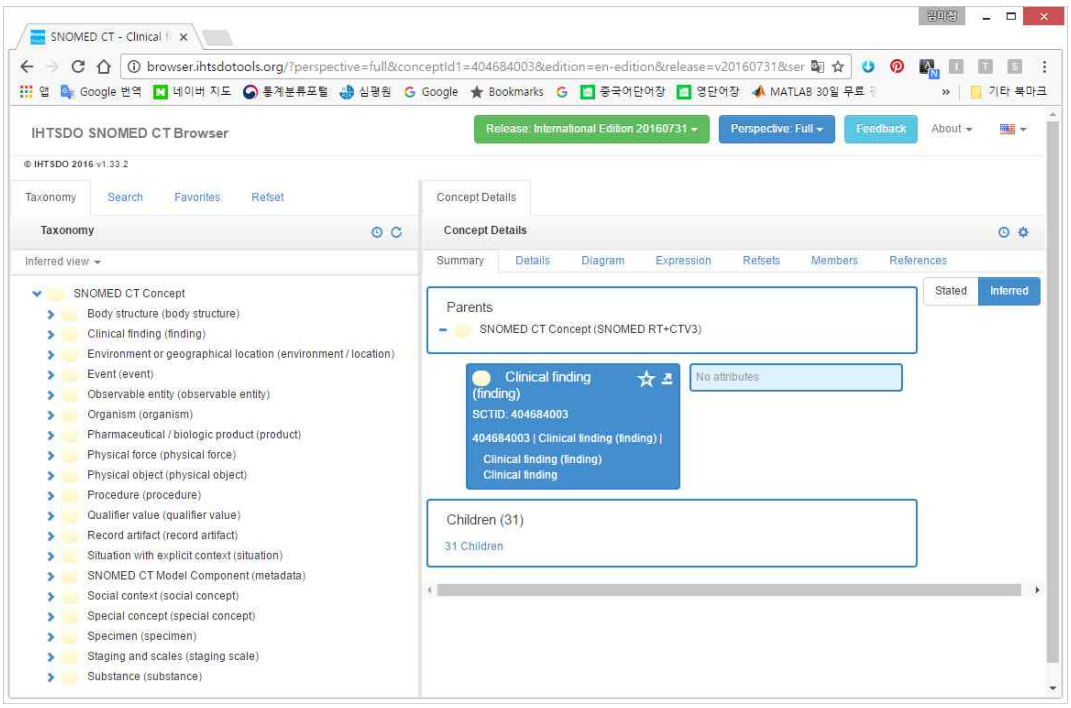


Fig. 30. The main screen of SNOMED CT online browser
(Source: http://browser.ihtsdotools.org)

BAVC 모델에 사용된 용어와 SNOMED CT의 개념코드를 매핑하는 일반적인 방법을 심막염의 사례로 보면 다음과 같다.

① 왼쪽 상단의 검색 탭에서 'pericarditis'를 검색한다(Fig. 29).

② 검색된 용어 중 하나를 선택한 후 오른쪽에 나타난 개념의 세부사항에서 부모자식 관계와 속성 관계의 의미를 확인한다. 선택한 'pericarditis'의 개념적 의미를 오른쪽의 세부사항에서 파악해 보면 관계를 통해 심낭 질환 중 하나이자, 순환기 계통의 염증 질환 중 하나이고, 염증 질환이라는 형태학적 특징이 있으며, 발병 부위는 심낭구조인 질환임을 확인할 수 있다.

③ 기본개념인 심낭염에 SNOMED CT의 개념코드인 '3238004'를 부여한다.

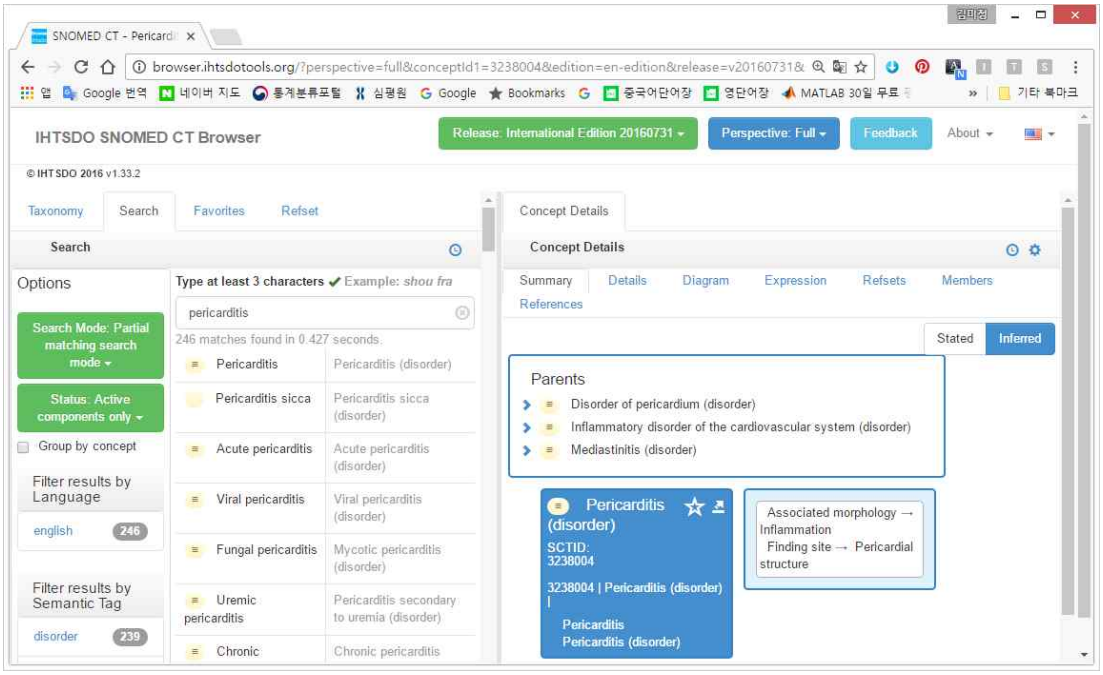


Fig. 31. Concept details of 'pericarditis' in search result (Source: http://browser.ihtsdotools.org)

BAVC 모델에 사용된 일부 용어에 대해서는 다른 방식으로 매핑하였다. 예를 들어, 급성(acute)은 맥락에 따라 다소 다른 의미로 사용되기 때문에 일반적인 방식으로 개념코드를 찾기 어려웠다. SNOMED CT 브라우저에서 'acute'를 검색하면 동일한 문자열은 검색되지 않고, 부분 매치로 검색된 용어들이 나열된다 (Fig. 30). 이런 경우에는 각 용어의 개념 세부사항을 읽어보아도 영문을 한글로 해석하는데서 오는 차이와 국내에서 사용하지 않는 생소한 표현들로 인해 개념코드를 판단하기가 어렵다. 이러한 경우에는 'acute'이 사용될 전체 용어인 'acute pericarditis'를 검색한 후 전체 용어를 기술하기 위한 다이어그램을 통해 해당 속성을 확인하였다. 'acute pericarditis'의 임상 경과가 'Sudden onset AND/OR short duration'임을 확인한 후 급성('acute')의 개념코드를 '42412400'로 최종 결정하였다 (Fig. 31).

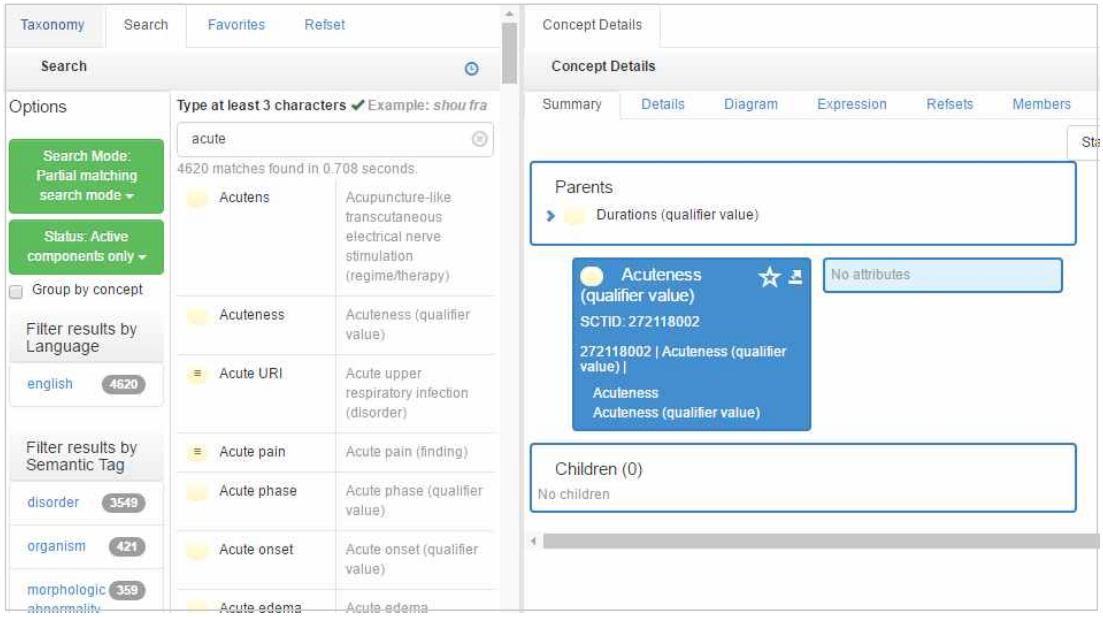


Fig. 32. Concept details of 'acute' in search result
(Source: http://browser.ihtsdotools.org)

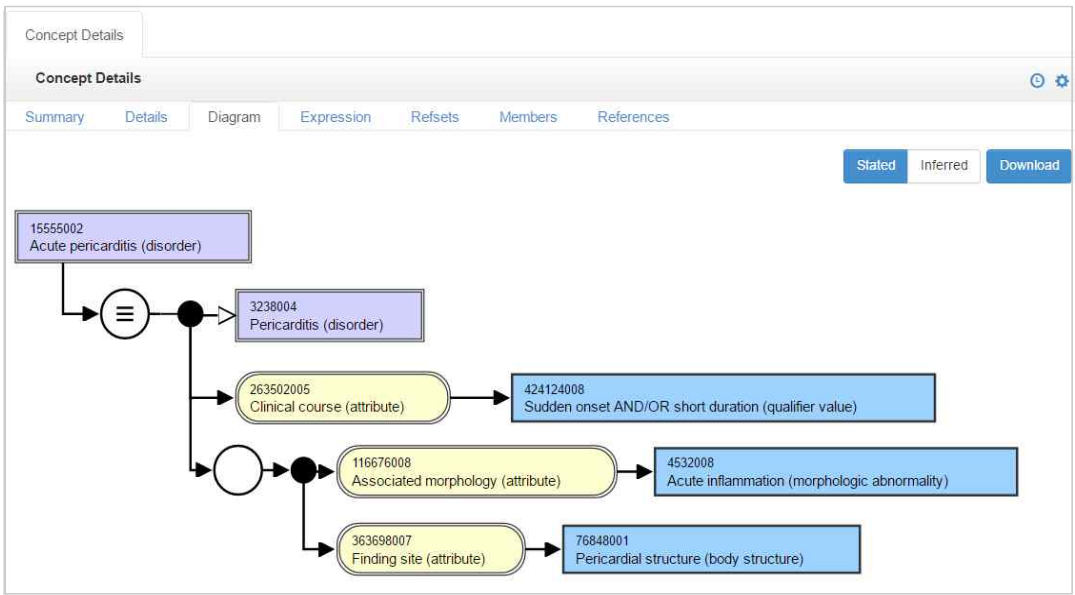


Fig. 33. Diagram stating the meaning of 'acute pericarditis'
(Source: http://browser.ihtsdotools.org)

(3) SNOMED CT 매핑 결과

BAVC 인스턴스 모델에 사용된 용어는 중복을 제외하고 총 179개이다. 그 중 27개의 기본개념은 SNOMED CT 개념코드와 100% 완전 매핑되었다. 대부분 SNOMED CT의 질병 영역(disorder)의 코드와 매핑되었지만, 죽상경화증과 동맥경화증은 형태학적 이상 영역(morphologic abnormality)의 코드로 매핑되었다. 동맥의 죽상경화증은 질병 영역에 해당되는 개념코드가 있었지만, 본 모델에서는 동맥과 대동맥 모두를 포함한 모델이기 때문에 형태학적 이상(morphologic abnormality)의 개념코드를 수용하였다. 빈맥의 경우는 수식어구가 없으면 증상으로 구분되기 때문에 마찬가지로 관찰 영역(finding)의 코드와 매핑하였다(Table 17). SNOMED CT에는 하나의 개념을 표현하는 다양한 용어들이 존재하기 때문에 해당 개념을 명확하게 설명할 명칭이 필요하다. 일반적으로 해당 개념에 대한 대표용어가 개념을 명확하게 설명하는 명칭이 되는데 대표용어 옆에 해당 개념이 속하는 영역을 괄호 안에 표시하여 뜻을 더 명확하게 전달한다. Table 17의 SNOMED CT 개념에 작성된 내용은 개념코드와 개념을 명확하게 설명하는 명칭인 FSN이다. 24개의 기본개념은 모두 'disorder'영역에 해당되지만, 앞서 설명한 것처럼 동맥경화증과 죽상경화증은 'Arteriosclerosis (morphologic abnormality)'와 'Atherosclerosis (morphologic abnormality)'와 매핑했으며, 빈맥은 'Tachycardia (finding)'와 매핑되었다.

Table 17. Mapping base concepts onto SNOMED CT concepts

No	Base concept	SNOMED CT concept
1	고혈압	38341003 |Hypertensive disorder, systemic arterial (disorder)|
2	뇌경색	432504007 |Cerebral infarction (disorder)|
3	뇌동맥 색전증	75543006 |Cerebral embolism (disorder)|
4	뇌동맥 폐색	20059004 |Cerebral artery occlusion (disorder)|
5	뇌동맥 협착	71444005 |Cerebral arterial thrombosis (disorder)|
6	뇌출혈	230690007 |Cerebrovascular accident (disorder)|
7	뇌혈관 사고	230690007 |Cerebrovascular accident (disorder)|
8	뇌혈관질환의 후유증	195239002 |Late effects of cerebrovascular disease (disorder)|
9	대뇌동맥류	128608001 |Cerebral arterial aneurysm (disorder)|
10	대뇌동맥 박리	713081000 |Dissection of cerebral artery (disorder)|
11	동맥경화증	28960008 |Arteriosclerosis (morphologic abnormality)|
12	모아모아병	69116000 |Moyamoya disease (disorder)|
13	빈맥	3424008 |Tachycardia (finding)|
14	심근경색증	22298006 |Myocardial infarction (disorder)|
15	심근병증	85898001 |Cardiomyopathy (disorder)|
16	심근염	50920009 |Myocarditis (disorder)|
17	심내막염	56819008 |Endocarditis (disorder)|
18	심막염	3238004 |Pericarditis (disorder)|
19	심방세동	49436004 |Atrial fibrillation (disorder)|
20	심방조동	5370000 |Atrial flutter (disorder)|
21	심부전	22298006 |Myocardial infarction (disorder)|
22	심장병	56265001 |Heart disease (disorder)|
23	심장부정맥	698247007 |Cardiac arrhythmia (disorder)|
24	죽상경화증	38716007 |Atherosclerosis (morphologic abnormality)|
25	하지정맥류	72866009 |Varicose veins of lower extremity (disorder)|
26	협심증	194828000 |Angina (disorder)|
27	신장병	58718002 |Rheumatic fever (disorder)|

순환기 계통의 질병명 구조분석에서 총 14개의 속성이 도출되었지만, 실제 모델에 사용된 속성은 발병 부위(Finding site), 임상 경과(Clinical course), 원인(Due to), 동반병태(Associated with), 특징(Characterizes), 순서(Order)로 총 6개였다. 실제로 더 상세하게 구별될 수 있는 속성이지만 모델이 복잡해지는 것을 막기 위해 특징이라는 범주의 속성으로 구분하였다. Table 18은 BAVC 모델에 사용가능한 전체 속성을 SNOMED CT 코드와 매핑한 결과이다.

Table 18. Mapping attributes onto SNOMED CT concepts

No	Attribute	SNOMED CT concept
1	Finding site	363698007 | Finding site (attribute)|
2	Episodicity	246456000 | Episodicity (attribute)|
3	Clinical course	263502005 | Clinical course (attribute)|
4	Severity	246112005 | Severity (attribute)|
5	Associated morphology	116676008 | Associated morphology (attribute)|
6	Causative agent	246075003 | Causative agent (attribute)|
7	Pathological process	370135005 | Pathological process (attribute)|
8	Occurrence	246454002 | Occurrence (attribute)|
9	After	255234002 | After (attribute)|
10	Due to	42752001 | Due to (attribute)|
11	Associated with	47429007 | Associated with (attribute)|
12	Finding method	418775008 | Finding method (attribute)|
13	Characterizes	704321009 | Characterizes (attribute)|
14	Order	272126005 | Order values (qualifier value)|

속성 값에 해당하는 용어는 중복을 제외하고 총 138개의 용어가 사용되었다. 속성 값의 경우 대부분의 용어코드가 매핑되었으나, 양쪽 뇌전동맥, 다발성 뇌전동맥과 같은 복합명사 중 일부는 개념코드를 찾을 수 없었다. 이런 경우 SNOMED CT의 후조합 원칙에 따라 개념코드를 부여하였다. 예를 들어 기타 뇌

전동맥은 기타에 해당하는 '74964007'과 뇌전동맥에 해당하는 '11281008'을 조합하여야 표현이 가능하였다. 기타 뇌전동맥은 분류에 적합하도록 만든 용어이지 실제 임상적으로는 기타에 해당하는 뇌전동맥 부위들이 존재하기 때문에 차후 해당되는 용어를 추가할 필요가 있다. 분류체계에서는 주요 뇌전동맥 부위에 대해서는 분류코드가 존재하지만 그 외의 부위에 대해서는 기타로 처리하는 경우가 많다. 질병분류코드만 중요한 것이 아니라, 명확하고 상세한 질병명이 입력되기 위해서는 모델에 구체적인 부위를 추가할지에 대해서는 차후 사용자 합의가 필요하다. 상세불명의 뇌전동맥의 경우에는 상세불명 코드만 부여하여도 되지만, 상세불명의 뇌전동맥과 상세불명의 대뇌동맥을 구분하기 위하여 상세불명 코드인 '10003008'과 대뇌동맥 코드인 '11281008'을 함께 사용하였다. Fig. 32는 SNOMED CT 코드로 표현된 BAVC 인스턴스 중 일부이다.

BAVC 모델의 핵심 구성요소는 기본개념, 속성, 속성 값, KCD-7 코드 4가지지만, 그 외에 모델번호, 모델명, SNOMED CT의 버전정보, 생성일, 종료일이 포함된다. 인스턴스별로는 4가지 요소 외에 모델번호, 인스턴스 번호, 인스턴스 표현 명칭, 생성일, 종료일이 포함된다.

Baseconcep	A1	V1	A2	V2	A3	V3	C
432504007	42752001	10003008	363698007	10003008	363698007	10003008	I63.9
432504007	42752001	74964007	363698007	10003008	363698007	10003008	I63.8
432504007	42752001	439127006	363698007	244392000	363698007	10003008	I63.6
432504007	42752001	439127006	363698007	11281008	363698007	85234005	I63.00
432504007	42752001	439127006	363698007	11281008	363698007	59011009	I63.01
432504007	42752001	439127006	363698007	11281008	363698007	69105007	I63.02
432504007	42752001	439127006	363698007	11281008	363698007	74964007+11281008	I63.08
432504007	42752001	439127006	363698007	11281008	363698007	10003008+11281008	I63.09
432504007	42752001	439127006	363698007	88556005	363698007	17232002	I63.30
432504007	42752001	439127006	363698007	88556005	363698007	60176003	I63.31
432504007	42752001	439127006	363698007	88556005	363698007	181313007	I63.32
432504007	42752001	439127006	363698007	88556005	363698007	38608005	I63.33
432504007	42752001	439127006	363698007	88556005	363698007	74964007+88556005	I63.38
432504007	42752001	439127006	363698007	88556005	363698007	10003008+88556005	I63.39
432504007	42752001	414086009	363698007	11281008	363698007	85234005	I63.10
432504007	42752001	414086009	363698007	11281008	363698007	59011009	I63.11
432504007	42752001	414086009	363698007	11281008	363698007	69105007	I63.12
432504007	42752001	414086009	363698007	11281008	363698007	74964007+11281008	I63.18
432504007	42752001	414086009	363698007	11281008	363698007	10003008+11281008	I63.19
432504007	42752001	414086009	363698007	88556005	363698007	17232002	I63.40
432504007	42752001	414086009	363698007	88556005	363698007	60176003	I63.41
432504007	42752001	414086009	363698007	88556005	363698007	181313007	I63.42
432504007	42752001	414086009	363698007	88556005	363698007	38608005	I63.43
432504007	42752001	414086009	363698007	88556005	363698007	74964007+88556005	I63.48
432504007	42752001	414086009	363698007	88556005	363698007	10003008+88556005	I63.49
432504007	42752001	26036001	363698007	11281008	363698007	85234005	I63.20
432504007	42752001	26036001	363698007	11281008	363698007	59011009	I63.21
432504007	42752001	26036001	363698007	11281008	363698007	69105007	I63.22
432504007	42752001	26036001	363698007	11281008	363698007	74964007+11281008	I63.28
432504007	42752001	26036001	363698007	11281008	363698007	10003008+11281008	I63.29
432504007	42752001	26036001	363698007	88556005	363698007	17232002	I63.50
432504007	42752001	26036001	363698007	88556005	363698007	60176003	I63.51
432504007	42752001	26036001	363698007	88556005	363698007	181313007	I63.52
432504007	42752001	26036001	363698007	88556005	363698007	38608005	I63.53
432504007	42752001	26036001	363698007	88556005	363698007	74964007+88556005	I63.58
432504007	42752001	26036001	363698007	88556005	363698007	10003008+88556005	I63.58
432504007	42752001	415582006	363698007	11281008	363698007	85234005	I63.20
432504007	42752001	415582006	363698007	11281008	363698007	59011009	I63.21
432504007	42752001	415582006	363698007	11281008	363698007	69105007	I63.22

Fig. 34. BAVC instances expressed by SNOMED CT concept code

V. 시스템 구축 및 평가

5장에서는 본 연구에서 개발한 BAVC 모델과 질병분류 규칙의 활용가능성을 확인하고 검증하기 위해 지식기반의 질병분류 시스템을 구축하고 기존 시스템과 비교 평가하였다. 또한 모델의 타당성을 검증하기 위하여 모델의 커버리지를 평가하였다.

1. 구축 도구 및 개발 환경

시스템 개발 환경은 개발 범용성과 데스크톱, 태블릿, 스마트폰 등의 사용자 접근성을 고려하여 웹 기반으로 구현하였다. 웹 서버는 아파치 서버이고, 프로그램 언어는 PHP를 사용하였고, 데이터베이스는 MySQL을 사용하였다. 아파치 서버 버전은 2.4.10, PHP서버 버전은 5.6.0, MySQL 서버의 버전은 5.6.20이다.

2. 질병분류 시스템 구현

1) 시스템 구성도

본 시스템의 구성도는 Fig. 33과 같다. 질병분류는 의사의 진단명 입력 화면에서 시작된다. 사용자가 진단명을 입력할 때 BAV 저장소(BAV DB)에 있는 용어모델에 의하여 속성이 제시되고, 제시된 속성을 선택하면 선택된 BAV 조합에 의하여 BAVC 저장소(BAVC DB)의 질병분류모델에 의해 질병분류코드가 부여된다. 의사가 선택한 질병분류코드는 전자의무기록 데이터베이스(EMR DB)에 저장된다. 의사가 BAVC 모델에 의하여 상세하고 구체적인 질병명을 입력했다고 하여도 의사의 관점과 분석의 관점에서 질병분류코드가 상이할 수 있다. 의사가

상세한 수준으로 입력한 두 개의 진단명이 질병분류원칙에 의해 하나의 진단명으로 입력되어야 하는 경우가 있지만, 임상적인 관점에서 두 개의 진단명을 사용하는 것이 더 명확할 때도 많다. 의사가 입력한 진단명은 EMR DB에 그대로 보관되면서 분석용 데이터베이스로 데이터가 넘어올 때 질병분류 규칙 지식 저장소에 있는 규칙에 따라 새로운 코드로 재분류 되어 저장된다. 분석용 데이터베이스의 자료는 병원통계 및 각종 진료정보 분석업무에 사용된다.

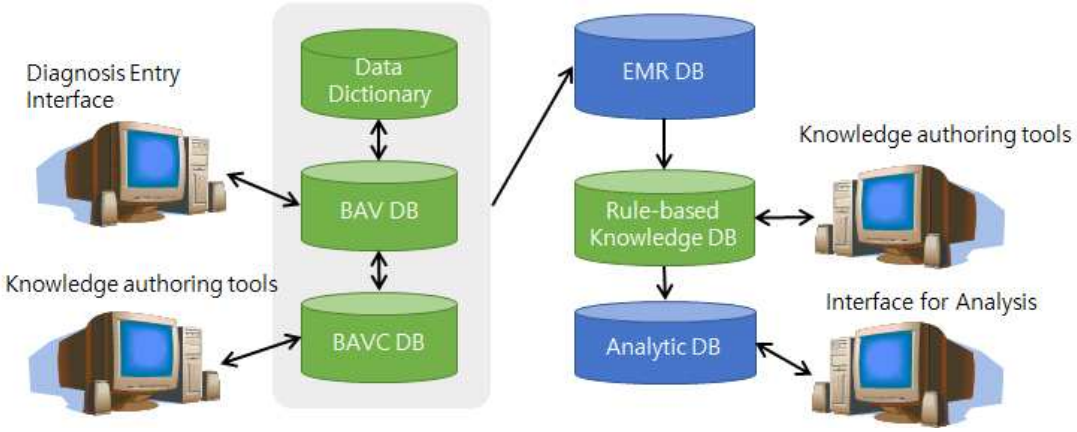


Fig. 35. Overview of the knowledge-based system for classification of disease

2) 질병분류 시스템 흐름도

기존의 질병분류는 의사가 진단명을 선택하면 코드화된 질병분류코드가 함께 저장되는 방식이기 때문에 질병분류 업무는 의사가 진단명을 입력하는 행위에서 시작된다(Fig. 34 (a)). 기존 시스템에서 의사가 저장한 질병분류코드는 행정 및 분석 업무에 그대로 사용되거나 인간의 개입에 의해 KCD-7 지침에 맞는 질병분류코드로 재분류된 후 별도의 분석용 데이터베이스에 저장된다.

본 연구에서 제안한 지식 모델 기반의 질병분류 시스템은 두 단계의 지식모델을 사용한다. 첫 번째 지식 모델은 전자의무기록에 상세한 질병분류코드가 저장될 수 있도록 돕는 속성 중심의 질병분류모델인 BAVC 모델이고, 두 번째 지식 모델은 자료 간의 관계를 규칙으로 정의한 질병분류 규칙이다. 의사가 EMR에 진단명을 입력하는 시점에서 상세 수준의 질병명이 EMR 데이터베이스에 저

장되도록 BAVC 모델이 적용되고, 이렇게 저장된 질병분류코드들은 두 번째 지식인 질병분류 규칙에 의해 자동 재분류되어 분석 데이터베이스에 저장된다(Fig. 34 (b)).

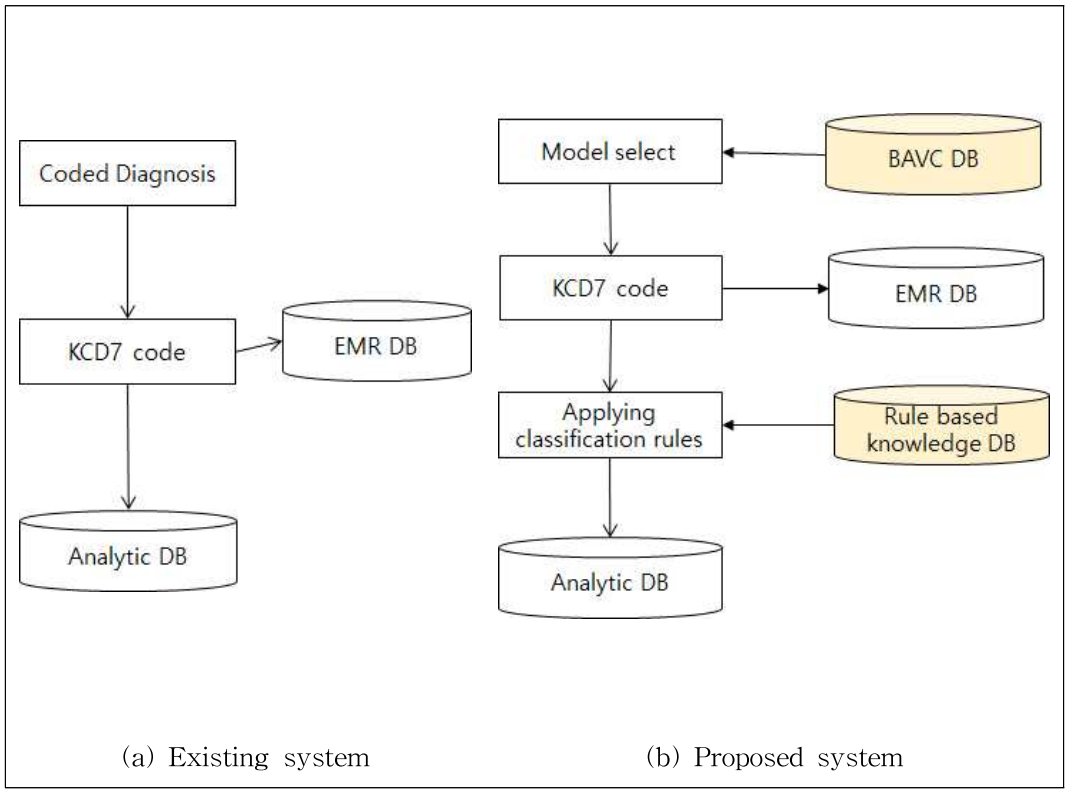


Fig. 36. Flow diagram of disease classification in the existing system and the proposed system

3. 지식기반의 사용자 인터페이스

1) 사용자 인터페이스

Fig. 35는 BAVC 지식 모델을 적용한 사용자 인터페이스 화면이다. 진단명을 입력하는 화면으로 의사가 상세 수준의 진단명을 입력하도록 돕고 그에 맞는 질병분류코드가 부여되도록 화면을 구성하였다. BAVC 모델을 적용하여 기본개념을 선택하면 그에 맞는 속성이 지능적으로 나열된다.

화면의 왼쪽 첫 번째 진단명 리스트에는 모델의 기본개념들이 나열된다. 의사의 입력 빈도를 체크하여 자주 사용하는 기본개념 순으로 나열되도록 하였다. 두 번째 리스트부터는 의사가 선택한 진단명에 따라 진단명을 구체화하는 속성들이 나타난다. 뇌출혈의 경우 외상 여부에 따라 질병분류 범주가 달라지기 때문에 첫 번째 속성으로 뇌출혈의 원인이 나타나고 그에 맞는 유효한 속성 값 나열된다. 외상성 뇌출혈의 경우 출혈 부위와 열린 상처 여부를 에 따라 'I'가 아닌 'S'로 시작하는 질병분류코드가 부여된다. 그 다음 속성은 발병부위이고 최종적으로 외상성 지주막하 출혈이 두 개내 열린 상처인지를 확인해야 진단명 입력이 가능하다. 필요한 모든 속성을 선택해야만 진단명 입력이 가능하기 때문에 의사는 매번 같은 방법으로 일관성 있게 동일한 상세수준의 진단명을 입력하게 된다.

최종 진단명을 선택하면 화면 하단의 참고 항목에 KCD-7 코드와 함께 해당 진단명을 입력할 때 주의해야 할 점등 질병분류와 관련된 정보를 제공해준다. 아래 화면에서는 외상 환자이기 때문에 손상의 원인인 외인코드를 입력하라는 참고 메시지를 보여주고 있다. 의사의 진단명 입력 화면에 나열되는 진단명과 속성들은 BAVC 모델에 의해 구성되며 사용되는 진단명과 속성들은 BAV 지식 저작도구를 이용하여 생성 및 관리할 수 있다.

본 연구에서는 기본개념 별 단계별로 입력 가능한 구조화된 입력도구로 구현하였지만, BAVC 모델은 다양한 형태의 사용자 화면으로 구현되거나 활용될 수 있다. 기존의 방식인 진단명을 텍스트 검색하여 선택하는 방식에도 사용될 수 있다. 의사는 기존의 방식대로 진단명을 선택하더라도 내부적으로는 선택한 진단명을 구성하는 용어별 의미가 부여되어 있기 때문에 정보의 활용성이 높아진다. 또한 서술방식의 기록에서 질병명과 관련된 자연어를 추출하고 처리할 때도 유용하게 활용될 수 있다.

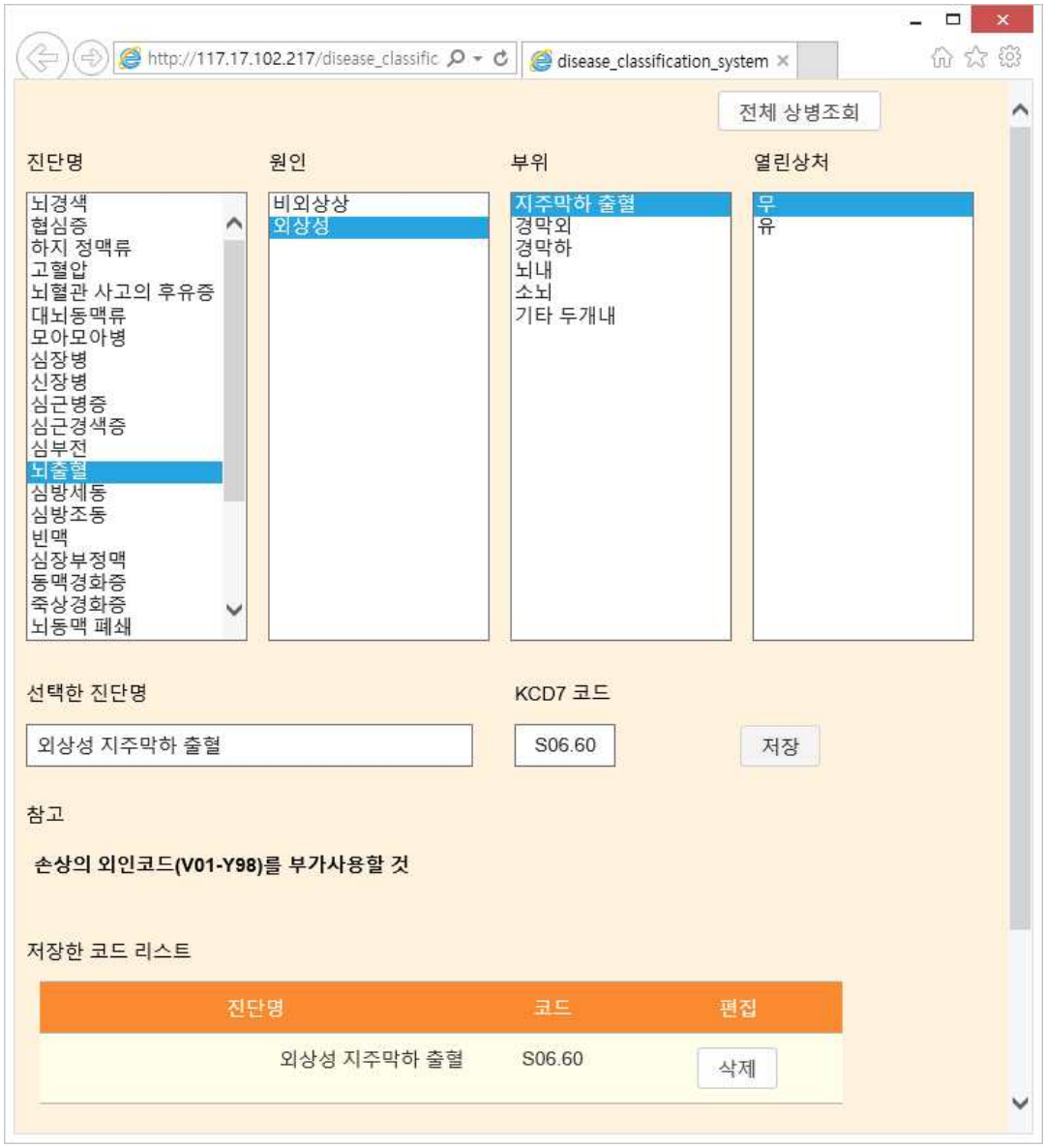


Fig. 37. Knowledge-based diagnosis input screen

2) 모델 유형별 구현 사례

BAVC 모델은 형태별 기본개념 형태, 단일 수식 형태, 다수 수식 형태, 단계별 수식 형태, 중복 병명 형태로 구분된다. 중복 병명 유형은 다수 수식 혹은 단계별 수식 유형으로 수정 개발되었기 때문에 개념적으로만 존재한다.

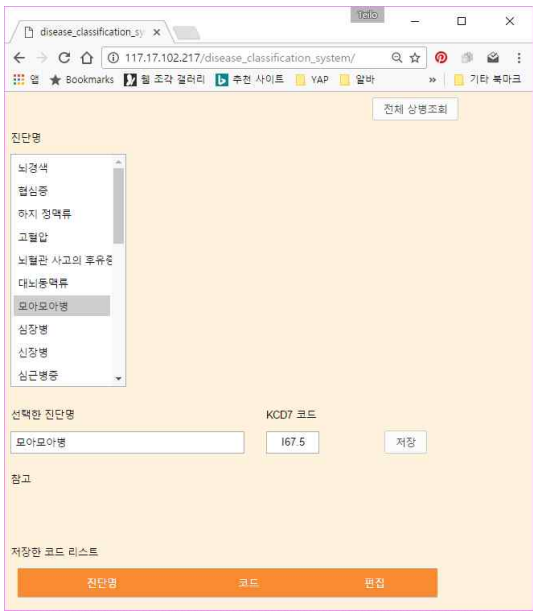
기본개념 유형인 모아모아병은 그 질병을 더 세분화할 속성이 없기 때문에 기본 진단명을 선택하면 속성 리스트가 나열되지 않고 바로 KCD-7 코드가 부여

된다(Fig. 36 (a)).

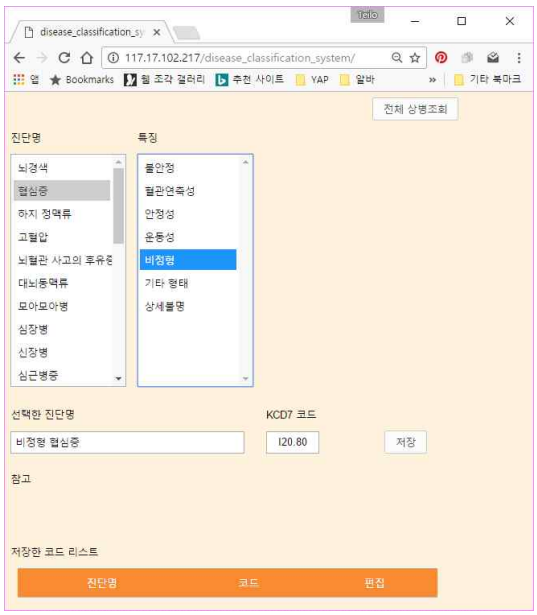
단일 수식 모델 구현 사례인 Fig. 36의 (b)는 협심증 모델로 기본개념인 협심증을 선택하면 협심증의 특징에 대한 속성 값이 나열된다. 만약 협심증의 특징이 나열된 속성에 있다면 해당 특징을 선택하고, 만약 없다면 기타를 선택하면 된다. 기타 특징을 가진 협심증의 경우 그 특징을 구별하고 싶다면 별도의 BAVC 지식도구를 이용하여 추가로 인스턴스를 생성하면 된다. 해당 특징이 특징 리스트에 추가되며, 그때 부여되는 코드는 도메인 전문가에 의해 검토 후 기타 협심증에 준하는 코드가 할당된다. 협심증이지만 어떤 특징을 가진 협심증인지 알 수 없을 경우에는 상세불명을 선택하도록 하였다.

다수 수식 모델 구현 사례인 Fig. 36의 (c)는 심근경색증 모델로 임상경과와 발생 부위에 따라 코드가 구체화 되는 것을 볼 수 있다. 임상경과와 부위의 순서가 바뀌어도 코드 분류에 문제가 되지 않는다. 심근경색증은 심장 근육이라는 부위에 발생하는 질환이므로 심근경색증의 발생부위의 속성 값으로 올 수 있는 항목은 심근부위로 한정되어야 하므로 심근경색증 모델을 개발할 때 속성 셋으로 제한한 심근의 상세부위만 화면에 나열된다.

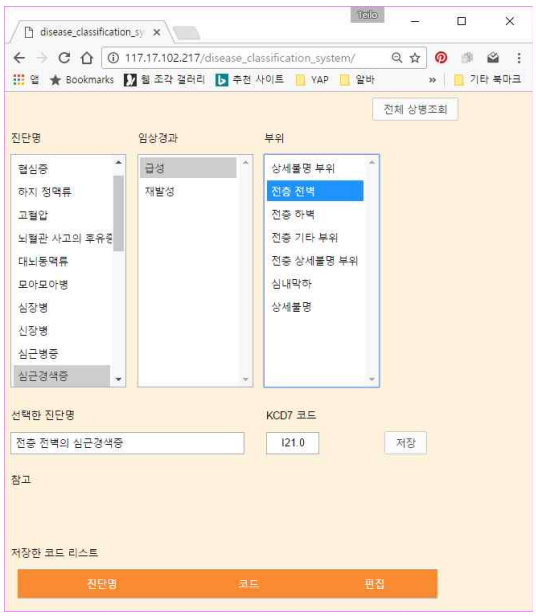
Fig. 36의 (d)는 단계별 수식 형태로 고혈압 모델을 보여준다. 고혈압의 경우 순수하게 발병한 원발성 고혈압과 다른 질환에 의해 발병한 이차성 고혈압으로 구분된다. 원발성 고혈압은 원인 병태가 없지만, 이차성 고혈압은 원인 병태가 분명히 존재하기 때문에 임상적으로 구별되며, 질병분류코드도 전혀 다르게 부여된다. 원발성을 선택하면 악성 여부만을 묻지만, 이차성을 선택하면 원인을 묻는 리스트가 나열된다. 중복 병변 형태의 모델은 모델 개발 단계에서 질병명의 용어 구조에 따라 다수 수식 형태나 단계별 수식 형태로 변경 개발하였기 때문에 위와 같은 식으로 구현되었다.



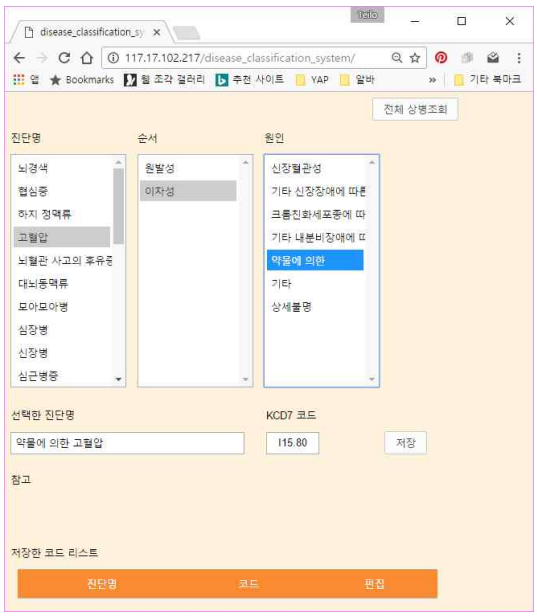
(a) Base concept form



(b) Single modifier form



(c) Multiple modifier form



(d) Step-by-Step modifier form

Fig. 38. Implementation example by model type

4. 두 단계 분류 프로세스

1) 진료 단계의 질병분류

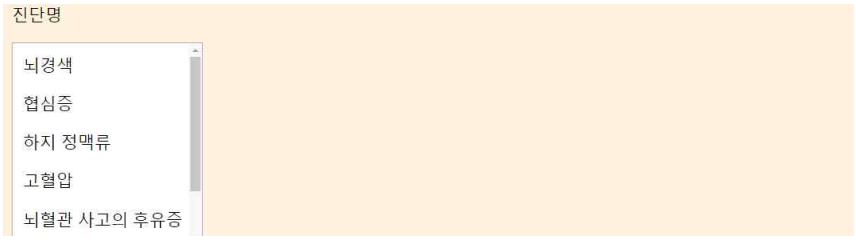
의사가 진단명을 전자의무기록 데이터베이스에 저장하는 단계로 질병분류 프로세스 상 최초의 분류단계이다. 두 번째 단계인 분석 단계에서는 첫 번째 단계에서 입력된 질병분류코드를 기준으로 코드 검코 및 재분류 과정을 거치므로 첫 번째 단계가 더욱 중요하다고 할 수 있다. 진료 단계에서는 BAVC 지식 모델이 사용된다(Fig. 37). 사용 단계는 아래와 같다.

① 의사는 진단명 리스트 중 해당되는 질병명을 선택한다. 진단명 리스트에 나타나는 질병명은 사용자의 저장 빈도를 체크하여 저장 빈도가 높은 질병명부터 나열된다.
② 선택한 질병명의 관련 속성과 속성 값 세트가 오른쪽 리스트에 나타난다.
③ 속성 값를 선택하면 더 상세한 속성 리스트가 다시 나타난다.
④ 최종 속성까지 선택하면 질병명 저장이 가능하다.
⑤ 저장버튼을 누르면 EMR 데이터베이스에 저장된다.

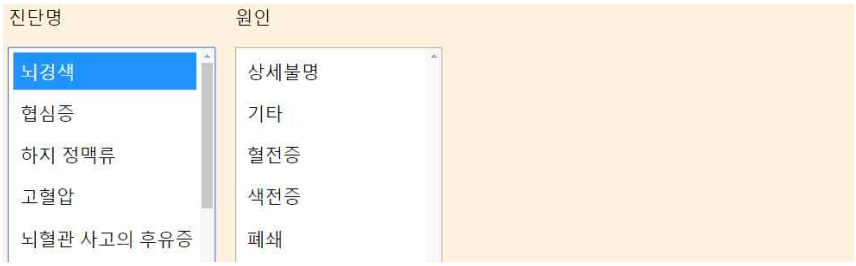
2) 분석 단계의 질병분류

분석 단계의 질병분류는 전자의무기록 데이터베이스에 저장된 질병분류코드를 검토하여 오류를 찾아내어 재분류하는 과정으로 최종 질병분류코드를 분석용 데이터베이스에 반입하는 단계이다. 이 단계에서는 지식 모델로 질병분류 규칙이 사용된다. 질병분류 규칙은 다음과 같은 단계로 적용된다.

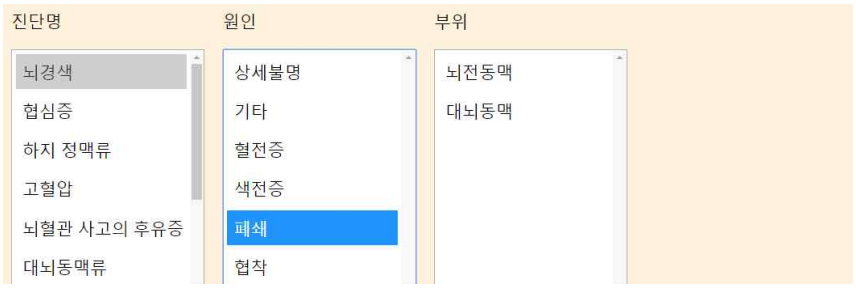
① 분석에 필요한 환자의 정보를 EMR 데이터베이스에서 가져온다.
② EMR 데이터베이스에서 가져온 질병코드들을 질병분류 알고리즘에 따라 검토하여 오류가 있을 경우 새로운 코드로 변경한다. 예를 들어 고혈압성 신부전과 고혈압성 신부전이 함께 있을 경우 KCD 지침에 따라 'I13.2'로 코드 변경한다(Fig. 38).
③ 재분류된 질병분류코드를 분석용 데이터베이스에 저장한다.



(a) Select base concept



(b) Select the first attribute value



(c) Select the second attribute value

(d) Select the final attribute value

Fig. 39. Application of knowledge-based model in the diagnosis phase

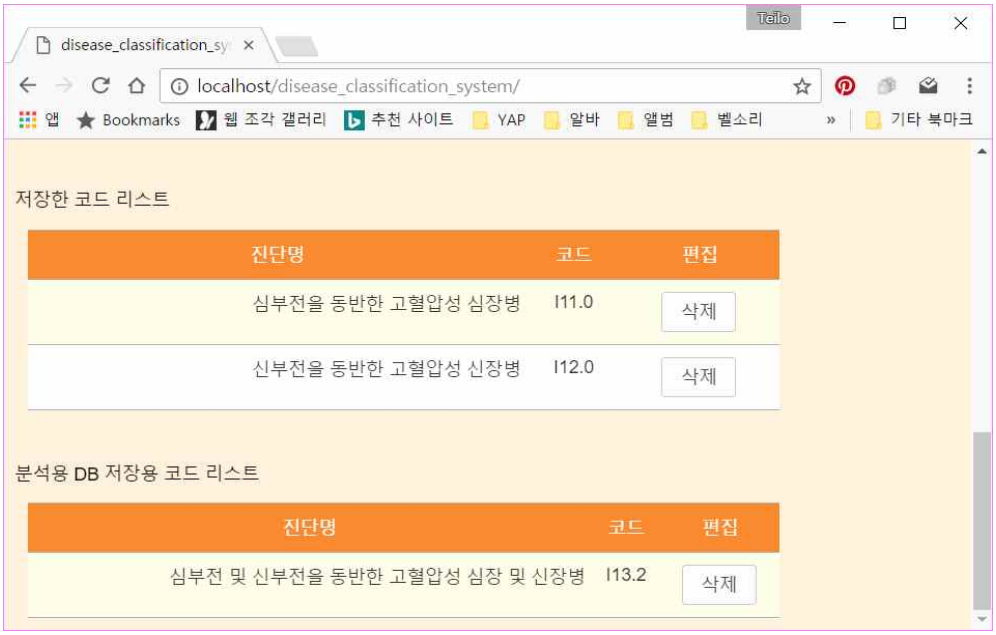


Fig. 40. Example of changing the code according to the disease classification rule

5. 지식 모델 및 시스템 평가

1) 사용자 인터페이스 비교

질병분류 지식모델을 적용한 시스템을 평가하기 위하여 기존 시스템과의 인터페이스를 비교하였다.

제안한 시스템의 인터페이스는 의사가 진단명을 선택하면 관련된 속성들이 단계적으로 제시되고 이를 선택하는 방식이다. 질병명 선택 후 제시되는 속성이 없거나 한 단계에서 끝나는 경우도 있고, 여러 단계의 속성 선택을 거쳐 최종 질병명을 선택해야 하는 경우도 있다. 예를 들어, 심근경색증을 선택하면 임상경과에 대한 정보들이 제시되며, 그 중 급성을 선택하면 해부학적인 부위가 제시되고 이 중 하나를 선택해야만 하단에 해당 진단명에 대한 정보와 KCD-7 코드가 제시된 후 진단명 입력이 가능하다(Fig. 39 (a)). 하지정맥류는 염증이나 궤양 동반 유무에 따라 코드가 달라지므로 동반병태의 값만 선택하면 진단명 입력이 가능하다(Fig. 39 (b)).

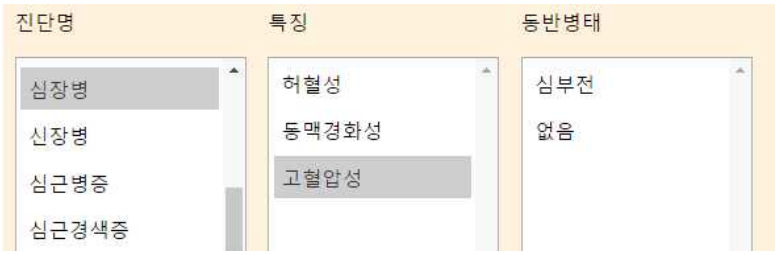
(a) Myocardial infarction

(b) Varicose vein of lower extremities

Fig. 41. Example of changing the attributes by disease

기본개념이 같더라도 첫 번째 선택된 속성 값에 따라 두 번째 속성이 달라질 수 있다. 예를 들어 심장병은 첫 번째 속성 값을 허혈성으로 선택하면 임상경과를 선택하라는 화면이 제시되고, 고혈압성을 선택하면 동반병태에 따라 상세코드로 분류된다(Fig. 40). 진단명 선택에 따라 사용자 인터페이스가 지능적으로 변경되는 방식은 사용자에게 특정 질병명이 어떻게 더 세분화되고 구체화되는지에 대한 정보를 암묵적으로 전달한다.

(a) Ischemic heart disease



(b) Hypertensive heat disease

Fig. 42. Example of attribute changing by phase of the same disease

기존의 일반적인 사용자 인터페이스는 질병명을 텍스트로 검색하여 검색된 질병명 중 하나를 선택하여 입력하는 방식이다. 텍스트 검색의 첫 번째 문제점은 오타의 우려가 있다는 점이다. 검색할 단어를 정확하게 타이핑하지 않으면 질병명이 검색되지 않는다. 두 번째는 검색된 질병명이 여러 개이고 질병명의 길이가 긴 경우에 가장 적합한 질병명을 선택하여야 하는데, 먼저 검색되는 가장 간결한 기본개념의 질병명을 선택하는 경향이 있다. 의사는 진단명이 다소 포괄적으로 작성되어도 퇴원요약지에 있는 요약된 진료정보를 보고 진료 업무를 수행하기

때문에 구체적인 질병명 입력에 대한 중요성을 간과할 수 있다. Fig. 41의 사례처럼 만성골수염(chronic osteomyelitis)를 검색할 경우 해부학적인 부위와 동반병태에 따라 수많은 질병명이 검색되지만 검색된 질병명을 모두 읽어보고 그 중 하나를 선택하는 것이 쉽지 않다.

Fig. 43. User interface of existing system

2) 시스템의 특징 비교

기존 시스템과 제안 시스템의 가장 큰 차이는 질병용어의 구조, 질병분류 방식, 표현 및 입력 방식에 있다. 기존 시스템은 식별자, 질병명, 질병코드로 이루어진 하나의 단순 용어 테이블을 사용하는 반면 제안 시스템에서는 최소단위의 질병명, 속성, 속성 값이 포함된 개념기반의 용어 테이블과 질병분류모델, 질병분류 규칙 테이블을 함께 사용한다. 이러한 특징 때문에 기존 시스템의 질병분류

방식은 텍스트 기반의 검색을 통해 선조합 방식의 진단명을 선택하면 질병코드가 부여되지만, 제안 시스템에서는 질병명을 구성하는 용어들을 조합하여 질병명을 완성하는 후조합 개념으로 질병코드를 부여하는 방식이다. 또한 규칙 기반의 분류 알고리즘을 적용하여 질병분류의 정확도를 높일 수 있다. Table 19는 기존 시스템과 제안 시스템의 특징을 비교한 표이다.

Table 19. Comparison of features between the existing system and the proposed system

Feature	Existing System	Proposed System
Concept based terms	×	○
Search by disease name	○	○
Search by KCD-7 code	○	○
Search by attribute of disease	×	○
Post-coordination	×	○
Providing knowledge	×	○
Knowledge management	×	○
Auto-classification	×	○

기존 시스템의 질병분류는 의사가 선택한 질병분류코드에 전적으로 의존하는 형태로 미리 조합된 단일 질병명의 선택에 의해 결정되었다. 제안 시스템에서도 의사의 선택에 의존하는 사실은 변함없지만, 의사가 속성을 조합하여 진단명을 선택하도록 후조합 방식을 제공하여 의사가 자칫 누락시킬 수 있는 정보도 빠짐없이 입력될 수 있도록 유도하였다. 그 후 최종 분석에 사용될 질병분류코드는 KCD-7의 질벼분류 알고리즘에 따라 자동 검토되고 재분류된다. EMR에 저장된 질병명과 코드를 수정할 수는 없지만, 분석용 데이터베이스에는 검토를 거친 최종 코드들이 별도로 저장된다.

이러한 지식 기반 질병분류 시스템의 이점을 평가하기 위하여 기존의 시스템의 문제점을 단계별로 정리한 후 지능형 시스템과 비교하였다. 기존 시스템의 입력 단계의 문제점은 텍스트 검색을 통해 질병명 선택 시 가독성이 낮아 구체적인 질병명을 선택하기 어렵다는 것이다. 제안한 시스템에서는 질병명별 지능형 속성 제시 기능으로 상세한 질병명 선택을 유도할 수 있도록 하였다. 또한 선택한 진단명이 가질 수 있는 속성만을 선택적으로 보여주기 때문에 화면에 보이는 콘텐츠는 간결하고, 선택된 질병분류코드에 대한 정보도 제공해 준다. 따라서 기존의 포괄적인 질병명과 세분화되지 못한 질병분류코드가 저장되는 것을 방지할 수 있다. 기존 시스템에서는 부정확한 보험청구나 질병통계 산출의 우려로 담당 부서에서 다시 질병분류코드를 재검토하였지만, 제안한 시스템은 지식 모델에 의해 그러한 과정을 생략할 수 있다. 정보 활용 측면에서는 속성 기준의 풍부한 검색과 활용이 가능하다. 기존의 시스템은 질병명 혹은 질병코드 단위로만 정보검색이 가능하였으나, 속성 기반의 질병분류모델에서는 질병명을 구성하는 기본개념, 속성명, 속성 값이 모두 SNOMED CT의 개념코드에 의해 식별되기 때문에 기본개념 혹은 속성명, 속성 값 각 각에 대한 검색이 가능하다. 예를 들어, 지주막하에 문제가 있는 모든 환자를 검색하려면 기존 시스템에서는 지주막하라는 표현이 들어간 질병명을 모두 검색한 후 해당 KCD-7 코드들을 모두 찾아 일일이 검색해야 가능하지만, 실제로 그러한 코드들을 찾아 모두 검색하는 것은 매우 어려운 일이다. 그러나 BAVC 모델을 사용한 시스템에서는 속성 중 발병 부위(363698007)가 지주막하(33930006)인 사례를 검색하면 매우 간단하게 지주막하에 문제가 있는 모든 환자를 검색할 수 있다. 또한 제안한 시스템에 적용된 BAVC 모델은 프로그램과 독립적이기 때문에 지식관리가 가능하다. BAVC 저작 도구를 이용하여 새로운 진단명의 입력과 수정이 가능하며 속성명과 속성 값의 추가가 자유롭다. 버전 관리를 통하여 표준화된 형태로 사용될 수 있다. 구체적이고 상세한 질병명의 입력은 의무기록의 질을 향상시키고 병원진료통계의 정확성을 향상시켜 궁극적으로 의료의 질을 향상시키는 결과를 가져온다. Table 20은 제안 시스템의 이점을 요약한 표이다

Table 20. Advantages of the proposed system compared to the existing system

Division	Existing System	Proposed System
Input	·Low readability ·Difficulty in selecting a detailed diagnosis	·High readability ·Ease of selection of a detailed diagnosis ·Provide information about coding
Save	·Save comprehensive disease names and codes ·Decrease in the quality of data	· Save specific disease names and precise codes ·Improve in the quality of data
Accuracy	·Need manual reclassification by humans ·Different code assignments by occupation ·Inaccurate statistics	·Automatic reclassification by rules ·Consistent code assignment ·Accurate statistics
Search criteria	·Search by KCD-7 code and name ·Use a local code to identify disease names	·Search by attribute, value, base concept, KCD-7 code and name ·Use a standard code to identify terms composing disease name

3) BAVC 모델의 커버리지 평가

제주시에 위치한 A 종합병원의 신경외과 진단명 중 자주 사용하는 진단명으로 등록된 용어를 임상 현장의 다빈도 질병명으로 간주하고 이를 대상으로 BAVC 모델의 커버리지를 평가하였다. 사용자 용어로 등록된 진단명 377개 중 순환기 계통의 질병명은 총 66개였다. 그 중 척추 동맥 동맥루(I72.88), 동정맥루(I77.0), 정맥의 색전증 및 혈전증(I80.2), 소혈관병(I99)인 4개의 질병명은 20대 다빈도 질병에 포함되지 않아 모델 개발에서 제외되었기 때문에 평가에서도 제외하였다.

62개의 사용자 용어를 제안 시스템에 입력해본 결과 62개 중 30개의 진단명이 BAVC 모델에 의해 구체적인 속성까지 입력되었다. 그러나 32개의 진단명은 포괄적 입력은 가능하였지만, 질병명을 수식하는 표현에 대해서는 입력이 가능하지 않았다. 상세한 입력이 불가능한 32개의 진단명은 대뇌동맥류, 대뇌동맥 박리, 뇌출혈 3개의 모델에 해당된다.

대뇌동맥류(I67.1)와 대뇌동맥 박리(I67.0)는 BAVC 모델에서는 파열 여부만을

묻고 질병분류코드가 부여되지만, 평가에 사용된 사용자 용어에서는 두 질병명이 발생한 구체적인 대뇌동맥 부위를 구분하였다. 즉, KCD-7의 코드부여지침은 대뇌동맥의 어떠한 부위에 발생하더라도 동일한 질병분류코드를 부여하기 때문에 BAVC 인스턴스 개발 시 발병부위에 대한 속성은 고려되지 않았다. 따라서 대뇌동맥류와 대뇌동맥 박리 두 개의 모델에서 개발된 인스턴스는 파열 여부만을 구분하여 4개에 불과하였지만 임상에서는 부위별로 구분하여 총 29개의 질병명을 사용하였다. 마지막으로 BAVC 모델을 적용할 수 없었던 사용자 용어는 뇌출혈 중 지주막하 출혈과 관련된 진단명이었다. 지주막하 출혈은 KCD-7 분류체계에서는 임상경과를 구분하지 않았지만 사용자 용어에서는 급성, 아급성, 만성으로 구분하여 사용하였다. 지주막하 출혈은 급성, 아급성, 만성의 경우 모두 I62.0로 분류한다. 결과적으로 KCD-7 코드는 정확하게 입력 가능하지만, 지주막하 출혈의 경우 급성, 아급성, 만성을 분류할 수 없었고, 대뇌동맥류와 대뇌동맥 박리 2개의 모델에서 발병 부위를 선택할 수 없었다.

모델의 커버리지는 27개 중 24개가 모델에 의해 구조화된 입력이 가능하여 전체 사용자 용어의 88.9%를 포괄하였고, 27개 중 3개의 모델은 입력은 가능하였지만 구체적인 속성을 선택할 수 없어 기본개념 수준에서 입력이 가능하여 11.1%가 포괄적 혹은 부분적인 입력이 가능하였다. 순수하게 입력할 수 없는 경우는 0%였다. (Table 21). 수정 전 모델의 커버리지는 88.9%였지만, 지주막하 출혈 모델에는 임상경과 속성을 추가하고, 대뇌동맥류와 대뇌동맥 박리 모델에는 발병 부위 속성을 추가하여 사용자 용어를 모두 수용하였다.

Table 21. The results of BAVC model coverage

Result	BAVC model	BAVC instances
Fully match	24 (88.9%)	30 (48.4%)
Partial match	3 (11.1%)	32 (51.6%)
Not match	0 (0.0%)	0 (0.0%)
Total	27(100.0%)	62(100.0%)

BAVC 모델은 속성을 자유롭게 추가할 수 있는 유연한 구조이기 때문에 기존의 속성과 속성 값을 이용하여 모델의 포괄성을 100%로 향상시킬 수 있었다. 이와 반대로 뇌경색의 경우에는 KCD-7 분류지침에서는 I63.0-I63.9까지 원인별 부위별로 매우 상세한 분류코드를 요구하고 있지만, 사용자 용어에서는 기타 뇌경색에 해당하는 I63.8과 상세불명의 뇌경색인 I63.9코드만이 사용자 용어로 등록되어 있었다(Table 22). 통계청에서 KCD-7을 개정하면서 뇌경색의 경우 ICD-10의 분류보다 더 구체적인 해부학적 부위를 세분화하도록 요구하고 있지만, 임상현장에서는 오히려 기타 뇌경색과 상세불명의 뇌경색에 해당하는 분류코드를 세분화하여 사용하고, 통계청에서 요구한 상세코드는 정작 사용하지 않았다. 구체적인 분류가 필요한 기타 뇌경색에 해당하는 질병명에 대해서 세분화할 필요가 있음을 단적으로 보여준다.

Table 22. User's terms related to the cerebral infarction in A hospital

Diagnosis	KCD-7 code
basal ganglia infarction	I638
cerebral infarction	I638
infarction due to vasospasm	I638
internal capsular infarction	I638
lateral medullary infarction	I638
low flow infarction	I638
multiple cerebral infarction	I638
striatocapsular infarction	I638
cerebellar infarction	I639
infarction of total middle cerebral artery territory	I639

4) 기대 효과

효율적인 질병분류를 위한 지식 모델의 시스템 구현과 검증 결과 다음과 같은 기대 효과를 예상할 수 있다.

첫째, 정확하고 상세한 질병명 선택을 유도하기 때문에 양질의 의료정보를 수집할 수 있다. 사용자가 놓치기 쉬운 구체적인 정보 대해서 자동으로 선택여부를

묻기 때문에 중요한 정보를 누락시킬 우려가 줄어들며, 상세한 질병명의 입력이 가능하다. 그 동안 원하는 질병명을 찾기 어려워 포괄적인 진단명이 입력되는 문제를 해결하여 관련된 원내부서나 보험회사와 같은 외부 유관 기관에서 진단명과 질병분류코드를 재확인해야 하는 번거로움을 완화할 수 있을 것으로 기대된다.

둘째, 진단명의 입력이 수월하다. 텍스트로 검색하여 진단명을 선택하는 것보다 지식 모델이 적용된 구조화된 화면은 제한된 시간에 필요한 정보를 신속하고 정확하게 입력할 수 있도록 도와준다. 텍스트 검색 시 길이기 긴 진단명이 너무 많이 검색되어 나열될 경우 가독성이 매우 떨어진다. 구조화된 입력방식은 인터페이스가 복잡하고 의사가 원하는 임상개념을 찾기 어렵다는 단점 때문에 제한적으로 사용되었지만, 지식기반의 진단명의 입력 화면은 단지 선택 단계만을 거치므로 입력이 수월하다.

셋째, 질병분류에 대한 정보를 제공할 수 있다. 어떤 진단명이 어떠한 속성에 의해 구체화되는지, 관련 질환에 대하여 부가적으로 입력해야 하는 진단명은 무엇인지 등의 정보를 제공하여 질병분류에 대한 사용자의 인식과 지식을 높일 수 있다. 제안한 모델에서는 KCD-7의 마스터 테이블을 통해 제외, 포함, 주(notes)와 관련된 정보만을 제공하지만, 차후 보험청구와 관련된 정보도 함께 제공한다면 청구와 관련된 지식도 제공할 수 있다.

넷째, 정보의 활용성이 향상된다. 질병분류코드 뿐만 아니라 해당 질병명을 구성하고 있는 조합 용어들에 대해서도 검색이 가능하기 때문에 단편적인 검색이 아닌 풍부한 의미 검색이 가능하다. BAVC 각 요소별로 SNOMED CT의 개념코드를 부여하여 컴퓨터가 질병명의 의미를 해석할 수 있게 한다. 동맥경화증과 죽상경화증은 KCD-7에서는 동일한 질환으로 간주되어 동일한 질병코드로 분류되지만, 임상적으로는 차이가 있다. 물론 이러한 차이를 진료 시 구분하지 않기도 하지만, 다른 사례나 다른 유사 질병의 경우 분명히 구분할 필요가 있다. 대부분 연구용 자료검색을 할 때 KCD-7 코드로 자료를 검색하지만, 본 모델을 적용하면 KCD-7 코드는 같더라도 다른 질병일 경우 이들을 구별하여 검색하고 분석할 수 있다. KCD-7 코드는 다르지만 유사한 속성을 가진 질병들을 분석하면 새로운 지식을 발견할 수 있다. 이러한 지식들은 임상의사결정지원시스템에도

유용하게 사용될 수 있다. 질병명별 치료내역이나 처방내역, 환자의 예후 등은 해부학적 부위나 동반병태에 따라서 달라지거나, 차별화 될 수 있기 때문이다.

향후 서술방식의 임상문서에서 다양하게 표현되는 진단명을 추출하여 파싱한 후 개념단위로 용어를 조합하면 코드 조합인 BAVC 모델을 기준으로 질병분류 코드를 부여할 수 있다. 즉, 프리텍스트 형태의 진단명도 BAVC 모델을 활용하면 속성 단위로 저장하여 활용할 수 있다. 예를 들어, 조직검사결과지에 'tubular adenocarcinoma'라고 작성되면 이를 파싱하여 'tubular' & 'adenocarcinoma'를 개념코드로 조합하여 'M8211/3'이라는 코드를 찾아내 저장할 수 있다.

다섯째, 지식 모델과 질병분류 규칙을 사용하면 객관적이고 일관성 있는 질병분류가 가능하다. 병원내규에 의해서 부여해 주거나 부여해 주지 않는 경우도 있기 때문에 이러한 내용을 반영하면 병원 내 일관성 있는 질병분류가 가능하다. 질병분류에는 정답도 있지만, 사용자별 재량껏 부여해주는 코드들도 있기 때문이다. 통계관련 부서에서 KCD-7 코드를 재분류할 때도 정확한 통계산출을 위하여 의사가 부여한 질병코드와 별개로 인간 질병 분류자가 의무기록을 검토하여 KCD-7 코드를 재분류하고 있다. 질병분류 전문가의 지식수준이 차이가 나기 때문에 해당 부서에서는 하급 질병분류자가 질병분류를 하면 상급 분류자가 재검토하고 있는 실정이다. 지식 모델은 인간의 수준차이를 극복하고, 인간이 할 수 있는 실수를 막아 객관적인 질병분류코드를 부여할 수 있게 한다.

여섯째, 지식 관리의 유연성과 확장성이다. 만약, 의사가 더 상세하게 분류하고 싶은 진단명 있다면 BAVC 인스턴스를 개발할 수 있어 지식의 확장성 측면에서도 유용하다. BAVC 모델과 규칙은 별도의 저장소에 관리되기 때문에 지식과 응용 프로그램이 분리되어 있어 새로운 질병명이나 속성 그리고 규칙의 추가가 비교적 자유롭다.

VI. 결론 및 향후연구

본 논문에서는 환자의 진단명과 질병분류코드를 정확하고 일관성 있게 수집하고, 부서마다 다르게 부여되는 질병분류코드의 불일치 문제를 해결하기 위해 효율적 질병분류를 위한 지식기반 모델을 제안하였다. 전자의무기록의 최초 질병분류인 진단명 입력 시 질병분류 지침에 맞는 상세 수준의 코드가 저장될 수 있도록 질병분류모델을 개발하고, 저장된 코드의 오류 여부를 검토하는 규칙을 개발하였다. 개발된 지식기반 모델의 활용 가능성을 검증하기 위하여 질병분류 시스템을 구축하고 평가하였다.

본 연구를 통해 얻은 결과는 다음과 같다.

문헌과 전문가 지식을 분석한 결과 KCD-7 코드는 해당 질병명이 가진 고유한 속성에 따라 달라지거나, 질병명 이외의 다른 요인에 의해 달라졌다. KCD-7의 순환기 계통의 질병명을 대상으로 어휘 분석을 실시한 결과 14가지 속성의 수식어에 의하여 질병명이 구체화되었다. 질병명의 속성에 따른 코드변화는 질병명을 기본개념, 속성, 유효한 속성 값으로 구조화하여 질병분류코드를 할당하는 BAVC(Base concept, Attribute, Value and Code) 모델로 공식화하였고, 질병명 이외의 다른 요인에 의한 코드변화는 IF-THEN 규칙으로 정의하였다.

순환기 계통의 국내 다빈도 질병명 대상으로 BAVC 모델을 개발한 결과 기본개념에 따라 총 27개의 모델이 개발되었으며, 모델의 구성은 기본개념 27개, 속성 14개, 유효한 속성 값 138개 그리고 KCD-7 코드 186개로 요약되었다. 모델은 기본개념을 수식하는 형태에 따라 기본개념 형태, 단일 수식 형태, 다수 수식 형태, 단계별 수식 형태, 중복 병명 형태로 구분되었다. 모델에 사용된 모든 용어에는 SNOMED CT 개념코드를 매핑하여 용어마다 의미를 부여하였다. BAVC 모델에 의해 작성된 인스턴스는 최종 290개이다. 질병명 이외의 다른 요인에 의한 코드변경 규칙은 동반코드 간 의 관계, 산과환자 여부 등이 해당되었다.

이러한 지식 모델을 시스템으로 구현한 결과 질병명별 선택적 속성 제시에 따라 예상대로 상세한 수준의 진단명 입력이 유도되었다. 질병분류지침을 따르는

BAVC 모델에 의해 동일한 질병명이라도 선택된 속성에 따라 그 다음 선택 옵션이 달라지며, 그러한 선택 단계는 일관성 있게 제시되었다. 지식 모델의 타당성을 검증하기 위해 A 종합병원의 신경외과 다빈도 사용자 진단용어 중 순환기계통의 진단용어 62개를 시스템에 입력한 결과 모델의 커버리지는 88.9%였다. 11.1%에 해당하는 3개의 모델은 대뇌동맥류, 대뇌동맥 박리, 뇌출혈 이었다. 대뇌동맥류와 대뇌동맥 박리는 KCD-7 지침에서 발병 부위를 구분하지 않고 동일한 코드를 부여하기 때문에 모델 개발 시 발병부위에 대한 속성을 고려하지 않았지만, 현장에서는 발병부위에 따라 세분화된 진단명을 사용하였다. 뇌출혈 모델을 이용하는 지주막하 출혈의 경우에도 KCD-7 지침에서는 급성, 만성, 아급성을 구분하지 않았지만, 실제 임상현장에서는 이러한 속성 들을 구분하고 있었다. 커버되지 않은 11.1%의 모델은 시스템 독립적인 지식관리 도구를 이용하여 해당 모델에 속성을 추가하는 간단한 방식으로 사용자 진단명은 100% 수용할 수 있었다.

BAVC 모델을 통해 상세하고 정확하게 입력된 코드들은 사전에 정의한 IF-THEN 규칙에 의해 자동 검토되어 오류가 있을 경우 자동 재분류되는 것을 확인하였다. 의사가 최선의 진단명을 입력하였더라도, 질병분류지침에 의해 두 개의 코드가 하나의 코드로 변경되어야 할 때도 있고, 환자의 상황에 따라 동일한 진단명이라도 다른 코드가 부여될 수 있기 때문에 행정적인 목적으로 사용될 때는 이러한 규칙에 의한 코드의 재분류 과정이 꼭 필요하다.

본 연구에서 개발한 지식기반 모델은 기존의 질병분류코드의 불일치를 해소하고, 정확성을 높여 의료정보의 질을 높이고, 데이터 활용 측면에서도 유용할 것으로 기대된다. 개발된 지식 모델은 프로그램에 비종속적인 형태로 개발되어 이식과 재활용이 가능하며, 표준임상 용어체계인 SNOMED CT 개념코드를 매핑하여 의미적 호환성과 활용성을 보장한다.

대부분의 업무가 자동화되고 빅데이터에 의한 인공지능의 영역이 확대되고 있는 시점에서 인간의 개입 없이 정확한 질병분류를 수행하기 위해서는 다양한 연구가 필요할 것으로 사료되며, 본 연구의 한계와 향후 연구에 대하여 다음과 같이 제언한다.

첫째, 질병분류모델을 순환기 계통의 다빈도 질병만을 대상으로 하였기 때문

에 질병 전체를 포괄하기 위한 모델이 되기 위해서는 다른 영역의 질환에 확장 적용하여 정교화 할 필요가 있다.

둘째, 질병분류 규칙에 대해 주로 문헌 중심의 지식을 활용하였기 때문에 문헌과 임상현장에서 나타나는 용어의 차이를 반영할 수 없었다. 통계청에서 제시한 표제어를 사용하였기 때문에 의사 고유의 표현인 현장 용어를 반영하지 못하였다. 특정 의사만 사용하는 표현이나 약어, 축약어 등의 사용을 자제하고 표준적인 용어를 권장한다는 유용한 측면도 있지만, KCD-7의 표제어와 표현과 의미가 다른 질병명도 있기 때문에 빈번하게 사용되는 현장 용어들은 반영할 필요가 있다.

셋째, 다양한 상세수준을 가진 진단명을 편리하게 입력할 수 있도록 사용자 입장을 고려한 추가적인 인터페이스 연구가 필요하다.

넷째, 상세하고 정확한 진단명의 수집을 우선순위로 두어 주 진단의 선정 오류에 대한 문제는 다루지 않았다는 점이다. 주 진단 선정의 오류가 많은 만큼 정확한 진단명의 수집을 전제로 주 진단 선정을 지원할 수 있는 방안에 대한 연구가 필요하다.

References

[1] van Ginneken Astrid M, Derksen-Lubsen Gerarda, Bleeker Sacha E, van der Lei Johan, Moll Henriëtte A, 2006, Structured data entry for narrative data in a broad specialty: patient history and physical examination in pediatrics, BMC Med Inform Decis Mak, Vol. 6, No. 1, pp. 1472-6947.

[2] Scott P, Macisaac P, Saad P., 2002, An introduction to health terminologies, Brisbane: National Centre for Classification in Health.

[3] Oniki TA, Zhuo N, Beebe CE, Liu H, Coyle JF, Parker CG, Solbrig HR, Marchant K, Kaggal VC, Chute CG, Huff SM, 2016, Clinical element models in the SHARPn consortium, J Am Med Inform Assoc. Vol. 23, NO. 2, pp.248-256.

[4] Goossen W.T., 2014, Detailed clinical models: representing knowledge, data and semantics in healthcare information technology, Healthc Inform Res. Vol. 20, No. 3, pp.163-72.

[5] Goossen W, Goossen-Baremans A, van der Zel M., 2010, Detailed clinical models: a review, Healthcare Informatics Research, Vol. 16, No. 4, pp.201-215

[6] KCD 7th revision, 2015, Statistics Korea

[7] J.H. Ahn, 2002, Analysis of agreement status between the diagnostic code of health insurance claim and medical record coding, M.S. Thesis, Inje Univ.

[8] Y.S. Seo, Y.M. Kim, M.H. Nam, S.H. Kang, J.H. Lim, 2009, A Study on the agreement of Principal Diagnosis, Quality Improvement in Health Care Vol. 15, No. 1, pp.123-133.

[9] H.Y. Shin, K.H. Kim, 2015, A study on the assessment for the principal diagnosis confirm: electronic medical records and medical certificates, J Kor Health Information Management assoc., Vol. 29, pp.42-51.

[10] S.O. Bae, K.W. Kang, Y.K. Boo, Y. Lee, H.S. Cheo, H.Y. Choi, 2015, A Study on the Difference in Disease Coding of Doctors, Medical Insurance

Review Nurses and Medical Record Administrators based on Coding Simulation, J Korean Biometric Assoc., Vol. 40, No. 3, pp.161-174.

[11] Pakhomov SV, Buntrock JD, Chute CG., 2006, Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques, J Am Med Inform Assoc., Vol. 13, No. 5, pp.516-525.

[12] Mehrdad Farzandipour a, Abbas Sheikhtaheri b, F. Sadoughi, 2010, Effective factors on accuracy of principal diagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10), International Journal of Information Management, Vol. 30, No. 1, pp.78-84.

[13] Larkey LS, Croft WB, 1995, Technical Report - Automatic assignment of icd9 codes to discharge summaries, Ph.D. thesis, University of Massachusetts at Amherst, Amherst, MA.

[14] Serguei V.S. Pakhomov, JAMES D. Buntrock, Christopher G. Chute, 2006, Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques, J Am Med Inform Assoc. Vol. 13, No. 5, pp.516-525.

[15] M.J. Kim, H.C. Kim, 2014, Building Rule-based support system for disease code classification, The e-Business Studies, Vol. 15, No. 4, pp.61-81.

[16] S. Trent Rosenbloom, Randolph A. Miller, Kevin B. Johnson, Peter L. Elkin, and Steven H. Brown, 2006, Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems, J Am Med Inform Assoc., Vol. 13, No. 3, pp.277-288.

[17] Rocha RA, Huff SM, Haug PJ, Warner HR, 1994, Desiging a controlled medical vocabulary server: the voser project, Comput Biomed Res., Vol. 27, No. 6, pp.472-507.

[18] Tim Benson, 2012, Principles of Health Interoperability HL7 and SNOMED, Springer, UK.

[19] Robert AG. Greenes, 2007, Clinical Decision Support: The Road Ahead. Elsevier Seiten.

[20] IHTSDO. Http://www.ihtsdo.org/members/

[21] SNOMED CT Starter Guide international release, 2016, IHTSDO

[22] UMLS Reference Manual, 2009, National Library of Medicine(US)

[23] KOSTOM guideline, 2015, social security information service, ministry health & welfare

[24] KOSTOM. Http://www.hins.or.kr/

[25] J.H. Yoon, M.J. Kim, S.J. Ahn, M.S. Kwak, Y. Kim, H.G. Kim, 2009, The Development of Clinical Terminology Dictionary for Integration and Management of Clinical Terminologies in EMR Systems, Healthcare Informatics Research, Vo. 15, No. 4, pp.411-421.

[26] WHO, 2015, International Classification of Disease 10th Revision(ICD-10).

[27] J.H. Yoon, 2008, Generation and Processing of CDA Documents Based on Entry Level Supporting Templates, Ph.D. Thesis, The National University of Kyungpook.

[28] ISO 13606-2: Health Informatics-Electronic health record communication-Part2: Archetype interchange specification, 2008, ISO/TC 215.

[29] Catalina Martinez-Costa, Marcos Menarguez-Tortosa, Jesualdo Tomas Fernandez-Breis, 2009, Towards ISO 13606 and openEHR archetype-based semantic interoperability, Studies in Health Technology and Informatics Vol. 150, pp.260-264.

[30] openEHR Architecture Overview, 2015, openEHR Foundation

[31] Joseph F. Coyle, 2013, The clinical element model detailed model, Ph.D. thesis, The University of Utah.

[32] A development of clinical contents model and structured data entry, Center for interoperable EHR report, 2010, Center for interoperable EHR

[33] J.H. Hong, 2015, Medical record information management 9th revision, Komoonsa, Seoul.

[34] O.Y. Moon, 1992, Individual Variations in the Code of the International Classification of Disease for Similar Outpatient Conditions among General

Practitioners, Health Policy and Management, Vol. 2, No. 1, pp.66-79.

[35] Richard Farkas, Gyorgy Szarvas, 2008, Automatic construction of rule-based ICD-9-CM coding systems, BMC Bioinformatics, Vol. 9, No. 3, pp.1471-2105.

[36] J.S. Seo, H.Y, Shin, C.W. Ki, 2003, Development of Construction Model of Disease Classification on Clinical Diagnosis in Ophthalmology, Quality Improvement in Health Care, Vol. 10, No. 2, pp.204-214.

[37] Y.H. Kang, 2006, A Development of Semi-automated Diagnosis Code Generation for Discharge Summary, M.D. Thesis, The National University of Kyoungpook.

[38] E.J. Jun, H.W. Lim, K.Y. Song, E.Y. Choi, Y.J. Lee, K.S. Lee, 2009, Development of auto coding system on the basis of MedDRA to analyze adverse events for clinical trial, Translational and Clinical Pharmacology, Vol. 17, No. 2, pp.164-173.

[39] Statistics Korea, 2016, KCD-7 coding guideline, Ver.2016

[40] J.M. Spector, 2016, Handbook of Research on Educational Communications and Technology: Communication, Communication, 3rd edition, Content Technologies.

[41]Prakash M. Nadkarni, Luis Marenco, Roland Chen, Emmanouil Skoufos, Gordon Shepherd, Perry Miller, 1999, Organization of Heterogeneous Scientific Data Using the EAV/CR Representation, J Am Med Inform Assoc., Vol. 6, No. 6, pp.478-493.

[42] NadKarni, P.M., 1997, QAV: Querying entity-attribute-value metadata in a biomedical database, Comput Methods Programs Biomed, Vol. 53 No. 2, pp.93-103.

[43] ISO/TS 22789:2010(en), Health informatics-Conceptual framework for patient findings and problems in terminologies