

다국적 임상시험 지원을 위한 CDISC 표준의 확장

(CDISC Extension for Supporting Multinational Clinical Trials)

염지현[†] 최인영^{**} 김석일^{***} 김혁만^{****}
 (Jihyeon Yeom) (Inyoung Choi) (Sukil Kim) (Hyeokman Kim)

요약 CDISC 컨소시엄에서는 임상시험에서의 비효율적인 데이터 처리 과정을 개선하기 위해, 플랫폼에 독립적인 임상시험 데이터 표준을 정의하였다. 그러나, CDISC 표준은 여러 나라의 여러 기관이 함께 참여하는 다국가 임상시험에서 발생하는 임상시험 데이터를 다국어로 표현하는 방법에 많은 제약을 갖고 있다. 특히, CDISC가 제정한 표준 중 임상시험 데이터의 콘텐츠 및 포맷에 해당하는 SDTM(Study Data Tabulation Model)과 ODM(Operational Data Model)에서의 다국어 지원이 매우 미비하다. 본 논문은 CDISC의 SDTM과 ODM에서의 언어 설정에 대한 문제점을 해결하기 위해, SDTM과 ODM 표준의 확장을 제안한다. 이를 위해 SDTM에서는 다국어 지원을 위한 새로운 도메인을 설계하였고, ODM에서는 ODM의 확장 스키마를 서브타이핑 방법으로 구현하였다. 확장 SDTM과 ODM을 기반으로 임상시험 데이터를 처리하면, 다국가 임상시험이 수행되는 경우 다국어로 표현된 임상시험 데이터도 효율적으로 처리할 수 있다.

키워드 : 임상시험, 다국어 지원, CDISC, 서브타이핑

Abstract Clinical Data Interchange Standards Consortium (CDISC) developed global and platform-independent data standards to improve ineffective processes of clinical trial studies. Regardless of its objective toward global cooperation, the current version of the CDISC standard cannot describe clinical trial data in various languages for multi-national investigators or reviewers. This problem applies not only to tabulated datasets in Study Data Tabulation Model (SDTM) but also to extensible markup language representation of the datasets in Operational Data Model (ODM) instances. In order to address this issue, we propose to extend the current version of SDTM and ODM to collect clinical data for multi-national clinical trials. SDTM needs to have new special-purpose domain for multi-language representation purpose. Additionally, ODM is recommended to extend its XML schema using subtyping or type inheritance mechanism respectively. Our extension of SDTM and ODM enable to represent any granule of study data tabulation model or XML data entities to describe in efficient languages. This result will contribute to collect multi-language data easily for multi-national clinical trials.

Key words : Clinical trial, Multinational language support, CDISC, Subtyping

[†] 학생회원 : 국민대학교 컴퓨터공학부

jhyum@kookmin.ac.kr

^{**} 비회원 : 가톨릭중앙의료원 임상연구지원센터 교수

iychoi@catholic.ac.kr

^{***} 비회원 : 가톨릭대학교 의과대학 예방의학교실 교수

sikimMD@catholic.ac.kr

^{****} 정회원 : 국민대학교 컴퓨터공학부 교수

hmkim@kookmin.ac.kr

논문접수 : 2009년 3월 31일

심사완료 : 2009년 5월 26일

Copyright©2009 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지 : 컴퓨팅의 실제 및 레터 제15권 제8호(2009.8)

1. 서론

임상시험의 처리 과정은 과다하고 불필요한 문서로 인하여 매우 비효율적으로 진행되고 있다[1]. 특히, 데이터를 유지하기 위한 데이터베이스 설계 및 구축, 데이터 추출, 제출할 기관이 요구하는 형식으로서의 변환, 변환된 자료의 전송 등에 관련된 일련의 과정에 매우 많은 시간 및 노력이 소요된다[2,3].

CDISC(Clinical Data Interchange Standards Consortium)에서는 임상시험의 비효율적 수행 과정을 개선하기 위해, 모든 임상시험 데이터를 전자 문서화하고, 이 전자 문서의 생성/교환, 제출, 및 저장 등을 지원하

기 위한 표준을 정의하였다[4,5]. 그 중 SDTM(Study Data Tabulation Model)은 FDA(Food and Drug Administration)에 제출할 임상시험 데이터의 형식을 정의하고[6], ODM(Operational Data Model)은 XML 전자 문서로 제출할 임상시험 데이터의 형식 및 증례기록지(CRF, Case Report Form)를 표현하기 위한 메타데이터의 형식을 정의한다[7,8]. CDSIC 표준은 미국 식품의약품안전청(FDA)가 2004년부터 CDISC의 SDTM 표준에 따라 임상시험 데이터를 제출할 수 있도록 허용한 이후 더욱 국제적으로 주목 받고 있다[9].

한편 최근의 임상시험은 민족간의 효능 차이를 파악하기 위해 여러 언어를 사용하는 피험자들을 대상으로 하거나, 특정 국가의 범위를 넘어 여러 국가에서 실시되는 경우가 많이 발생하고 있다[10]. 이와 같이 다민족/다국가 기반의 임상시험이 수행됨에 따라, 여러 나라의 언어로 표현된 각각의 임상시험 데이터를 입력하고, 그 데이터들을 올바르게 처리하기 위한 특별한 주의가 필요하다[11].

첫째, 임상시험 데이터를 수집하기 위하여 피험자, 임상시험 담당자 및 연구자가 자신에게 편한 언어로 작성된 증례기록지를 사용할 수 있어야 한다. 둘째, 요구한 특정 언어를 사용한 증례기록지 내에 텍스트 형식의 데이터를 입력하거나 주석을 삽입할 경우, 해당 데이터 혹은 주석을 해당 언어로 표현할 수 있어야 한다. 셋째, 특정 언어로 표현된 증례기록지에서도 부분적으로 다른 언어를 사용하는 것이 가능해야 한다. 예를 들어, ECG(Electrocardiogram)와 같이 특정 나라에서 자국어로 번역하지 않고 영어를 그대로 사용하는 의학 용어가 증례기록지 혹은 증례기록지를 통해 수집된 데이터에 포함될 수 있어야 한다. 그러나, 현재의 CDISC 표준은 다민족/다국가 임상시험에 적용할 수 있을 정도의 다국어 지원을 심각하게 고려하지 않고 있다. 따라서, 이 논문에서는 CDISC 표준이 다국어 표현을 지원할 수 있도록 구조를 보완하고 확장하는 방안을 강구하였다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 SDTM에서 다국어 설정의 문제점에 대해서 설명하고 다국어 지원을 위한 확장 사항을 기술한다. 3장에서는 ODM에서의 다국어 설정의 제한 점과 다국어 지원 기능을 추가하기 위한 ODM 확장 스키마에 대해 설명한다. 마지막으로 4장에서는 결론 및 향후 연구 방향을 서술한다.

2. SDTM에서의 다국어 지원

2.1 SDTM 개요

SDTM은 신약 허가기관에 임상시험 결과 데이터를 테이블 형식으로 제출하기 위해 정의한 CDISC의 표준

으로, 도메인(domain)에 기반한 모델이다[6]. 도메인은 피험자들에 대한 시험 값 혹은 관측 값 중 논리적으로 관련이 깊은 것들을 모아 놓은 것으로 데이터베이스의 테이블과 동일한 개념으로 볼 수 있다. SDTM에서는 임상시험 측정 데이터를 데이터의 성격에 따라 도메인 단위로 구분하고, 각 도메인에 저장될 변수의 이름과 데이터 타입을 정의한다. 도메인은 임상시험 동안 수집되는 데이터를 생체계측 데이터, 검사 데이터 등 측정값을 표현하는 일반 목적 도메인과 임상시험, 피험자 등 일반 사항을 표현하는 특수 목적 도메인으로 구성되는데 현재 24개의 도메인이 정의되어 있다.

각 도메인은 표 1과 같이 하나의 테이블에 대응되며, 도메인의 변수들은 이 테이블의 열에 대응된다. 또한 증례기록지에서 수집된 각 측정값은 해당 테이블에서 하나의 행으로 표현된다. 예를 들어, 표 1은 피험자 GSR-005와 GSR-006에 대한 증례기록지의 생체계측(Vital Signs) 데이터를 나타내는 테이블로서 생체계측을 의미하는 VS 도메인으로 표현된다. 여기서 피험자 GSR-005는 세 개의 측정값을 갖고 있으며, 피험자 GSR-006은 하나의 측정값만 갖고 있다. 그리고 VS 도메인은 CDISC에서 정의한 VSTESTCD, VSTEST, VSCAT, VSORRES과 같은 변수들을 가지는데 각 변수는 다음과 같은 일정한 의미를 표현한다.

- STUDYID: 임상시험 아이디(Study Identifier)
- DOMAIN: 도메인 약어명(Domain Abbreviation)
- USUBJID: 피험자 아이디(Unique Subject Identifier)
- VSSEQ: 순차 번호(Sequence Number)
- VSTESTCD: 생체계측 측정 항목코드(Vital Signs Test Short Name)
- VSTEST: 생체계측 측정 항목(Vital Signs Test Short Name)
- VSCAT: 생체계측 카테고리(Category for Vital Signs)
- VSORRES: 생체계측 측정 결과값(Result or Finding in Original Units)

2.2 SDTM에서의 언어 설정 및 문제점

SDTM에서는 여러 언어로 임상 데이터를 표현할 경우 다른 언어로 표현된 측정값은 SUPPQUAL(Supplemental Qualifiers)이라는 특수 목적 도메인을 사용할 것을 권장한다. SUPPQUAL 도메인은 일반 도메인에 정의되어 있지 않는 변수를 사용자가 추가로 정의하기 위한 테이블이다[6]. 예를 들어, 생체계측 항목에서 한 환자(GSR-005)의 체중에 대한 타이틀을 영어로, 다른 환자(GSR-006)는 한글로 데이터를 수집하였을 경우 표 1과 같이 표현되어야 한다. 이 때, 두 환자의 체중

표 1 VS 도메인의 예

VS							
STUDYID	DOMAIN	USUBJID	VSSEQ	VSTESTCD	VSTEST	VSCAT	VSORRES
GSR	VS	GSR-005	1	WGHT	Weight	Somatometry	60
GSR	VS	GSR-005	2	SYSBP	Systolic Blood Pressure	Cardinal Sign	120
GSR	VS	GSR-005	3	DIABP	Diastolic Blood Pressure	Cardinal Sign	80
GSR	VS	GSR-006	1	WEIGHT	체중	Somatometry	65

표 2 SUPQUAL 도메인을 이용한 언어 설정

Supplemental Qualifiers							
STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL
GSR	VS	GSR-005	VSSEQ	1	VSLANG	Vital Sign Language	en
GSR	VS	GSR-005	VSCAT	Cardinal Sign	VSLANG	Vital Sign Language	en
GSR	VS	GSR-006	VSSEQ	1	VSLANG	Vital Sign Language	ko

타이틀이 서로 다른 언어로 수집되었다는 것은 표 2와 같이 SUPQUAL이라는 특수 목적 도메인에서 정의한다.

표 2는 SUPQUAL 도메인으로 표 1의 VS 도메인에서 사용한 언어를 설정한 예이다. SUPQUAL 도메인에서는 QNAM과 QLABEL이라는 변수를 이용하여 생체계측 데이터에서 사용한 언어를 정의한다. 표 2의 첫 번째 레코드는 표 1의 VS 도메인의 피험자 GSR-005의 첫 번째 레코드(VSSEQ="1"인 레코드)의 모든 변수 값의 언어가 영어(VSLANG="en")임을 낸다. 또한 표 2의 두 번째 레코드는 표 1의 피험자 GSR-005의 두 번째 및 세 번째 레코드(VSCAT="Cardinal Sign"인 레코드들)의 모든 변수 값의 언어가 영어(VSLANG="en")임을 나타낸다. 한편 표 2의 세 번째 레코드는 표 1의 피험자 GSR-006의 레코드(VSSEQ="1"인 레코드)의 모든 변수가 한국어(VSLANG="ko")임을 나타낸다. 그러나, 이와 같이 SUPQUAL을 사용하여 언어 설정을 하는 것은 다음과 같은 문제점이 있다.

첫째, SUPQUAL을 사용할 경우 각 도메인의 특정 레코드(혹은 변수 값의 집합) 전체에 대해서만 언어를 지정할 수 있으며, 변수 단위로는 언어를 설정할 수 없다. 예를 들어 표 1의 네 번째 레코드에는 한글과 영문이 함께 사용되고 있으나, 표 2의 세 번째 레코드는 표 1의 네 번째 레코드 전체가 한글로 기록된다고 설정하고 있다. 즉 표 2의 방법으로는 표 1의 네 번째 레코드와 같이 변수 값이 두 개 이상의 언어로 작성되는 경우를 설정할 수 없다.

둘째, 현재 SDTM에서는 특정 임상시험 전체 혹은 그 시험에 참여하는 특정 피험자의 디폴트 언어를 설정할 수 있는 기능을 제공하지 않고 있다. 디폴트 언어를 설정하면 임상시험의 주 언어 혹은 특정 피험자의 선호 언어를 명백히 선언할 수 있으므로, 부분적으로 언어를

설정하는 것을 최소화할 수 있다. 따라서, SUPQUAL 도메인의 레코드 수를 최소한으로 유지할 수 있다.

2.3 다국가 임상시험을 위한 SDTM의 확장

본 논문에서는 SDTM이 다국가 임상시험에서 다국어로 임상시험 데이터를 표현할 수 있는 기능을 지원하기 위해 다음과 같은 세가지 확장을 제안한다. 첫째, 임상 시험에 사용된 언어를 지정하는 특수 목적 도메인인 LS (Language Support) 도메인을 추가할 것을 제안한다. 현재의 SDTM에서는 언어 설정을 위한 새로운 변수를 타겟 도메인 변수의 선택 조건과 함께 SUPQUAL 도메인에 레코드의 형태로 추가한다. 그러나 제안하는 새로운 LS 도메인에서는 선택 조건과 설정한 언어의 코드만 기록한다. 이를 위해 LS 도메인은 아래의 변수를 포함한다.

- IDVAR(Identifying Variable): 특정 언어를 사용하고자 하는 변수
- IDVARVAL(Identifying Variable Value): 특정 언어를 사용하고자 하는 변수의 값
- SEQVAL(Value of Sequence Number): 특정 언어를 사용하고자 하는 변수의 --SEQ 값
- LANGCD(Language Code): 사용한 언어의 코드 값

표 3은 LS 도메인을 기반으로 표 1의 VS 도메인에서 사용한 언어를 서술한 예이다. 표 3에서 첫째 및 둘째 레코드는 변수 IDVAR과 IDVARVAL의 쌍으로 각각 선택조건 VSSEQ="1", VSCAT="Cardinal Sign"을 나타낸다. 이들 조건으로 선택한 레코드 혹은 레코드 집합은 표 2의 조건으로 선택한 것들과 일치한다. 그러나 마지막 세 번째 레코드는 변수 SEQVAL과 IDVAR의 쌍으로 선택 조건을 표현한다. 이 레코드는 표 1의 VS 도메인에서 피험자 GSR-006의 레코드(VSSEQ="1"인 레코드)의 특정 변수 VSTEST의 값이 한국어(VSLANG=

표 3 LS 도메인을 이용한 언어 설정

LS						
STUDYID	RDOMAIN	USUBJID	SEQVAL	IDVAR	IDVARVAL	LANGCD
GSR	VS	GSR-005		VSSEQ	1	en
GSR	VS	GSR-005		VSCAT	Cardinal Sign	en
GSR	VS	GSR-006	1	VSTEST		ko

“ko”)임을 나타낸다. 이 방법이 표 2의 방법과 다른 점은 표 2의 SUPPQUAL 도메인에서는 표 1의 특정 레코드의 전체 변수 값의 언어만 설정하지만, 표 3의 LS 도메인으로는 표 1의 특정 레코드의 전체 변수 값뿐만 아니라 특정 변수 값만 다른 언어로도 설정이 가능하다는 점이다.

표 3에서 살펴본 것처럼, 제안한 방법은 레코드뿐만 아니라, 레코드 내의 변수 단위로도 언어 서술이 가능하다. 또한, 사용자가 언어 서술을 위해 표 2의 VSLANG과 같은 새로운 변수를 정의할 필요가 없다.

둘째, 임상시험에 관한 일반 사항을 정의하는 도메인인 TS(Trial Summary)에 임상시험의 디폴트 언어를 설정할 수 있는 기능을 추가한다. 예를 들어, TS 도메인의 변수인 TSPARMCD(Trial Summary Parameter Short Name)에 임상시험 전체의 디폴트 언어 설정을 위한 코드인 DLANG(Default Language)을 추가한다. 표 4는 이 임상시험의 디폴트 언어가 영어로 설정되었음을 나타낸다.

셋째, 피험자의 특성을 정의하는 도메인인 SC(Subject Characteristics)에 피험자의 디폴트 언어를 설정할 수 있는 기능을 추가한다. 예를 들어, SC 도메인의 변수인 SCTESTCD(Subject Characteristic Short Name)에 피험자들의 디폴트 언어 설정을 위한 코드인 DLANG

을 이용하여 피험자 GSR-005의 디폴트 언어는 영어로, 피험자 GSR-006의 디폴트 언어는 한국어로 설정한다(표 5).

이와 같은 방법으로 본 논문은 표 6과 같이, CDISC의 SDTM에서의 언어 설정에 대한 3가지 문제점을 지적하고, 각 문제점을 해결하기 위해 SDTM의 확장을 제안하였다.

3. ODM에서의 다국어 지원

3.1 ODM 개요

ODM은 임상시험 데이터를 전자 문서로 제출, 교환 및 축적하기 위한 CDISC의 표준으로서 특정 업체나 플랫폼에 관계없이 사용할 수 있도록 개발되었다. ODM은 임상시험에 관한 메타데이터(study data), 임상시험 데이터(clinical data), 관리 데이터(administrative data), 참조 데이터(reference data) 등의 표준 형식을 XML Schema[13]로 정의한다[7]. 특히, 임상시험에 관한 메타데이터에서는 제출하는 임상시험 데이터 값들의 관계를 Protocol, StudyEvent, Form, ItemGroup, Item 등의 단계로 그룹화한 논리적 구조를 엘리먼트 <MetaDataVersion>에서 정의한다. 이 논리적 구조는 임상시험에서 사용한 증례기록지의 논리적 구조와 대응되며, 증례 기록지로 수집된 데이터는 이 논리적 구조에 따라 XML

표 4 TS 도메인을 이용한 임상시험의 디폴트 언어 설정

TS					
STUDYID	DOMAIN	TSSEQ	TSPARMCD	TSPARM	TSVAL
GSR	TS	1	DLANG	Default Language	en

표 5 SC 도메인을 이용한 피험자의 디폴트 언어 설정

SC							
STUDYID	DOMAIN	USUBJID	SCSEQ	SCTESTCD	SCTEST	SCORRES	SCSTRESC
GSR	SC	GSR-005	1	DLANG	Default Language	ENGLISH	en
GSR	SC	GSR-006	1	DLANG	Default Language	KOREAN	ko

표 6 SDTM에서의 다국어 설정의 문제점 및 다국어 지원 방법

SDTM에서의 언어 설정의 문제점	다국어 임상시험을 위한 SDTM의 확장
각 도메인의 변수 단위로 언어 설정 불가	LS 도메인 추가
특정 임상시험 전체의 디폴트 언어 설정 기능 미제공	TS 도메인의 TSPARMCD에 DLANG 코드 추가
특정 피험자의 디폴트 언어 설정 기능 미제공	SC 도메인의 SCTESTCD에 DLANG 코드 추가

인스턴스 문서로 변환된다. 따라서 CDISC 표준을 사용하면 증례기록지를 통해 수집된 임상시험 자료는 SDTM에서 정의한 도메인 및 변수에 따라 테이블 형태로 정리되고, 이 테이블의 자료는 ODM에서 정의한 논리적 구조에 맞게 XML 문서로 변환된다.

그림 1은 ODM 표준에 따라 임상시험에 관한 메타데이터를 XML 인스턴스 문서로 표현한 예이다. 이 문서는 임상시험에 관한 기본 사항과 제출하고자 하는 임상시험 데이터의 메타데이터를 정의한다. 그림에서 엘리먼트 <GlobalVariables>는 임상시험의 제목(<StudyName>)과 간단한 설명(<StudyDescription>)을 나타낸다. 또한, 엘리먼트<MetaDataVersion>은 임상시험 메타데이터 중 하나의 Item인 SYSBP에 대한 설명을 나타낸다. 그림은 홍삼의 효능 연구를 위한 임상시험이며, 수축기 혈압에 대한 메타데이터를 기술하고 있다.

3.2 ODM에서의 언어 설정 및 문제점

ODM은 엘리먼트 <TranslatedText>를 사용하여 언어를 설정한다[7]. 엘리먼트 <TranslatedText>는 xml:lang라는 애트리뷰트를 가지는데, xml:lang은 XML 인스턴스 문서에서 특정 엘리먼트의 콘텐츠와 애트리뷰트 값에 사용된 언어를 정의한다[8]. 이 애트리뷰트의 값으로는 IETF RFC 3066에서 정의한 언어 코드를 사용한다[14]. xml:lang으로 설정한 언어는 이 애트리뷰트가 부가된 엘리먼트, 그 엘리먼트의 모든 애트리뷰트 값, 그리고 그 엘리먼트의 콘텐츠에 포함된 모든 자식 엘리먼트에 적용된다[8].

ODM에서 엘리먼트 <TranslatedText>를 자식으로 가지는 엘리먼트로는 <Decode>, <Description>, <ErrorMessage>, <Question>, <Symbol>이 있다. 이들 엘리먼트들은 모두 애트리뷰트가 없고, 자식 엘리먼트

```
<CodeList OID="C66714" DataType="string" Name="Vital Signs Test Code">
  <CodeListItem CodedValue="SYSBP">
    <Decode>
      <TranslatedText xml:lang="en">Systolic Blood Pressure</TranslatedText>
      <TranslatedText xml:lang="ko">수축기 혈압</TranslatedText>
    </Decode>
  </CodeListItem>
  <CodeListItem CodedValue="WEIGHT">
    <Decode>
      <TranslatedText xml:lang="en">Weight</TranslatedText>
      <TranslatedText xml:lang="ko">체중</TranslatedText>
    </Decode>
  </CodeListItem>
</CodeList>
```

그림 2 엘리먼트 TranslatedText를 사용하여 언어 설정한 예

<TranslatedText>만 가진다. 그림 2는 엘리먼트 <Decode>를 사용하여 영어와 한국어 지원이 가능한 코드 리스트 Vital Signs Test Code를 ODM 스키마의 정의에 따라 표현한 예이다. 그림 2는 엘리먼트 <Decode>의 자식 엘리먼트인 <TranslatedText>에 있는 텍스트들이 모두 다른 언어로 표현된 동일한 용어임을 나타낸다.

이와 같은 ODM에서의 언어 설정 방법은 다음과 같은 문제를 갖고 있다. 첫째, 그림 2와 같이 엘리먼트 <TranslatedText>을 이용하여 언어를 정의할 경우, 다국어를 표현할 수 있지만 문서의 '깊이가 증가한다. 그러나, 해당 엘리먼트에 애트리뷰트 xml:lang을 부가 사용하여 언어를 지정할 경우, 문서의 깊이도 증가하지 않고 같은 의미를 전달할 수 있다. 또한 일반적으로 XML Schema로 서술하는 대부분 문서 타입에서, 이와 같은 방법으로 다국어를 표현하고 있다.

둘째, 스트링 타입에 바인딩된 모든 엘리먼트는 다국어로 서술될 가능성이 있다. 예를 들어, 그림 1의 엘리먼트 <StudyName>은 여러 나라 언어로 표현될 수 있

```
<?xml version="1.0" encoding="UTF-8"?>
<ODM xmlns="http://www.cdisc.org/ns/odm/v1.3">
  <Study OID="GSR">
    <GlobalVariables>
      <StudyName>A study on the efficacy of the ginseng steamed red</StudyName>
      <StudyDescription>Examine the efficacy of GSR on controlling blood sugar</StudyDescription>
      <ProtocolName>CDISC-Protocol-007</ProtocolName>
    </GlobalVariables>
    <MetaDataVersion OID="GSR-MDV-001" Name="Version 0.1">
      <ItemDef OID="VS.SYSBP" Name="Systolic Blood Pressure" DataType="integer">
        <Question>
          <TranslatedText xml:lang="en">Systolic Blood Pressure (mmHg)</TranslatedText>
          <TranslatedText xml:lang="ko">수축기 혈압 (mmHg)</TranslatedText>
        </Question>
      </ItemDef>
    </MetaDataVersion>
  </Study>
</ODM>
```

그림 1 ODM 인스턴스 문서의 예

다. 그러나 ODM에서는 스트링 타입에 바인딩된 엘리먼트의 출현 횟수(occurrence)의 최소(minOccurs) 및 최대값(maxOccurs)을 모두 1로 정의하고 있다. 따라서 스트링 값을 어떠한 언어로든지, 단 한번만 기술해야 한다. 즉, 엘리먼트의 값이 여러 개의 언어로 표현될 수 없다.

셋째, 그림 1의 엘리먼트 <ItemDef>의 애트리뷰트 Name 같이, ODM에서 애트리뷰트로 정의된 것들 중 일부는 다국어로 표현될 가능성이 있다. 그러나 애트리뷰트에는 언어 설정을 서술하는 애트리뷰트인 xml:lang을 부가할 수 없다.

넷째, ODM에서는 ODM 문서 전체의 디폴트 언어를 설정하는 방법이 없다. 뿐만 아니라, 특정 피험자의 디폴트 언어를 설정하는 방법도 제공하지 않고 있다.

그림 3은 XML Schema의 정의대로, 엘리먼트 <Decode>의 언어 설정을 위해, 이 엘리먼트에 직접 애트리뷰트 xml:lang을 사용한 예이다. 이와 같이 ODM 인스턴스 문서를 그림 3의 이상적인 문서와 같은 형식으로 작성하려면, ODM 스키마를 직접 수정해야 한다. 그러나 CDISC에서는 개별적인 목적으로 ODM 스키마를 수정하는 것을 허용하지 않고 있다. 다음 절에서는 다국어 지원을 위해, 위와 같은 문제점을 해결하기 위한 ODM의 확장 스키마를 제안하고자 한다.

```

<CodeList OID="C66714" DataType="string" Name="Vital Signs Test Code">
  <CodeListItem CodedValue="SYSBP">
    <Decode xml:lang="en">Systolic Blood Pressure</Decode>
    <Decode xml:lang="ko">수축기 혈압</Decode>
  </CodeListItem>
  <CodeListItem CodedValue="WEIGHT">
    <Decode xml:lang="en">Weight</Decode>
    <Decode xml:lang="ko">체중</Decode>
  </CodeListItem>
</CodeList>
    
```

그림 3 엘리먼트 TranslatedText를 사용하지 않고 언어 설정한 예

3.3 다국가 임상시험을 위한 ODM의 확장

ODM 스키마를 확장하기 위해서는 CDISC에서 정의한 ODM 스키마의 확장 요구사항을 준수해야 한다[7]. ODM 확장 요구사항의 핵심은 다음 두 가지이다. 첫째, 새로운 엘리먼트와 애트리뷰트를 추가하더라도 기존의 ODM 엘리먼트와 애트리뷰트를 변경 혹은 제거하지 않아야 한다. 둘째, 새로 정의한 엘리먼트와 애트리뷰트는 기존의 ODM 엘리먼트 혹은 애트리뷰트와 구별되도록 새로운 XML 네임스페이스를 가져야 한다.

이러한 요구사항을 강제하기 위해, ODM의 모든 complex datatype은 엘리먼트와 애트리뷰트 각각의 확장을 위한 dummy group을 갖는다. 이 dummy group은 기

존 데이터 타입에 새로운 콘텐츠를 추가할 때 “do-it-here”, 즉 이곳에서만 확장하라는 제약을 강제하는 장치이다. 이와 같이 스키마 확장 시 특정 위치에서만 확장하라는 dummy group의 개념은 XML Schema의 re-define 매커니즘을 사용해 확장 스키마를 작성하라는 것을 의미한다[15].

다국어 지원을 목적으로 ODM에서의 언어 설정 문제를 해결하기 위해, re-define을 사용한 ODM 스키마의 확장에는 몇 가지 문제가 존재한다. 첫째, 그림 3과 같이 불필요한 엘리먼트 <TranslatedText>를 사용하지 않고 언어를 설정하려면, ODM에서 엘리먼트 <TranslatedText>를 자식으로 갖고 있는 타입에서 먼저 엘리먼트 <TranslatedText>를 삭제하고, 해당 타입이 스트링 값을 직접 갖도록 하며, 이 타입에 애트리뷰트 xml:lang을 추가해야 한다. 그러나 원래 타입을 re-define으로 확장하여, 그 타입이 스트링 값을 직접 갖게 할 수는 없다. 왜냐하면 위의 타입(complexType 임)이 스트링을 갖게 하려면 속성값을 mixed="true"로 변경해야 하나, re-define을 사용하여 complexType의 속성을 변경할 수는 없기 때문이다.

둘째, 다국어를 지원해야 할 스트링 타입에 바인딩된 엘리먼트들의 출현 횟수 범위의 최대값을 1에서 무한대(unbounded)로 변경해야 한다. 그러나, re-define은 기존 엘리먼트의 출현 횟수 범위를 제한할 수는 있어도, 그 범위를 확장할 수는 없다.

셋째, 스트링 값을 갖는 애트리뷰트에 다국어를 설정하기 위해서는 해당 애트리뷰트를 비활성화하고, 그것을 엘리먼트로 변환하고, 이 새로운 엘리먼트의 출현 횟수를 복수 개로 정의해야 한다. 그러나 re-define으로는 애트리뷰트를 비활성화하는 동시에, 엘리먼트로 변환할 수는 없다. XML Schema의 정의에서 애트리뷰트의 비활성화는 restriction, 엘리먼트의 추가는 extension에 따르나, re-define을 사용하여 이 두 가지 상속 방법을 동시에 정의할 수 없기 때문이다.

넷째, ODM 문서 전체의 디폴트 언어를 설정하려면, 문서 전체의 루트 엘리먼트인 ODM의 바인딩 타입(binding type)에 애트리뷰트 xml:lang을 추가해야 한다. 그러나 현재의 ODM 스키마에서는 이 타입이 익명(anonymous)으로 정의되어 있다. 따라서 이 타입을 re-define할 방법이 없다.

다섯째, XML Schema의 정의에 따라, re-define을 사용한 확장 스키마는 원래 ODM 스키마와 동일한 네임스페이스를 사용해야 한다[16]. 또한 ODM 확장 요구사항에 따르면, 새로 정의된 엘리먼트와 애트리뷰트는 ODM 스키마의 네임스페이스와는 다른 네임스페이스를 가져야 한다. 따라서 ODM에 새로운 엘리먼트 혹은 애

트리뷰트를 추가하기 위해서는, ODM 스키마 및 확장 스키마와는 다른 네임스페이스를 가지는 제 3의 확장 스키마가 필요하다[7]. 따라서 redefine을 사용하여 ODM 스키마를 확장하면, 다음과 같은 2개의 확장 스키마가 생성된다.

- redefine 엘리먼트를 갖는 확장 스키마
- 새로운 엘리먼트 혹은 애트리뷰트를 정의하는 네임스페이스 확장 스키마

이와 같은 문제들을 모두 해결하려면, redefine이 아닌 다른 방법을 사용하여 ODM 스키마를 확장해야 한다. 본 연구에서는 다국어 지원을 위한 ODM 스키마의 확장 방법으로 서브타이핑(subtyping)을 사용한 확장 스키마의 작성을 제안한다. 서브타이핑은 기존 ODM 스키마의 타입을 상속받아 그 정의를 확장하거나 제한하여 필요한 새로운 타입을 생성하는 방법이다[16].

서브타이핑은 redefine 사용의 문제점을 해결할 수 있으나, 서브타이핑으로 ODM 스키마를 확장하기 전에 먼저 다음과 같이 기존 ODM 스키마의 수정이 필요하다. 엘리먼트 <TranslatedText>를 확장 스키마에서 restriction으로 제거하려면, 먼저 그것의 최소 출현 횟수가 "0"으로(minOccurs="0") 정의되어 있어야 한다. 따라서, ODM 스키마에서 엘리먼트 <TranslatedText>의 최소 출현 횟수를 "0"으로 변경하고, 그것의 부모 엘리먼트의 최대 출현 횟수를 "unbounded"로(maxOccurs="unbounded") 수정한다. 또한 다국어를 지원해야 할, 스트링

타입에 바인딩된 엘리먼트들의 출현 횟수 최대값을 "1"에서 "unbounded"로 수정한다. 표 7은 이와 같은 방법으로 출현 횟수의 수정이 필요한 ODM 스키마의 엘리먼트들을 나타낸다.

본 절에서는 3.2절에서 설명한 ODM에서의 언어 설정 방법이 갖는 4가지 문제점을 해결하기 위해 서브타이핑 방법을 사용하여 ODM 스키마를 확장하는 경우를 4가지로 분류하여 제시한다. 첫 번째 경우는 자식 엘리먼트 <TranslatedText>를 사용하지 않고 언어를 설정하는 방법이다. 이를 위해서는 먼저 ODM 스키마에서 자식 엘리먼트 <TranslatedText>를 갖는 모든 타입에 대해, 이들의 서브 타입(subtype)을 ODM의 확장 스키마에서 정의한다. 이 서브 타입들에서는 엘리먼트 <TranslatedText>의 최대 출현 횟수를 "0"으로 (maxOccurs="0") restriction하여, 실제로는 이 엘리먼트가 사용되지 못하게 한다. 그리고 이 서브 타입에 mixed="true" 속성을 부여하여, 엘리먼트 <TranslatedText>에서 표현하던 스트링을 직접 표현할 수 있게 한다. 또한 이 서브 타입에 애트리뷰트 xml:lang을 추가하여, 위의 스트링에 사용된 언어를 설정 할 수 있도록 한다. 예를 들어, 그림 4에서는 엘리먼트 <Question>이 자식 엘리먼트 <TranslatedText>를 사용하지 않고 애트리뷰트 xml:lang으로 다국어 지원이 가능하도록, 자식 엘리먼트 <TranslatedText>를 애트리뷰트 xml:lang으로 변환하는 스키마를 XMLSpy의 스키마 다이어그램으로 나타낸 것이다[17].

표 7 출현 횟수의 수정이 필요한 ODM 스키마의 엘리먼트

Occurrence		Sample element
Before revision	After revision	
0..1	0..*	FlagType, Comment, DisplayName, FirstName, FullName, LastName, LoginName, Organization, ReasonForChange ErrorMessage, Question
1..1	1..*	FlagValue, LegalReason, Meaning, ProtocolName, StudyDescription, StudyName Decode, Description, Symbol
1..*	0..*	TranslatedText

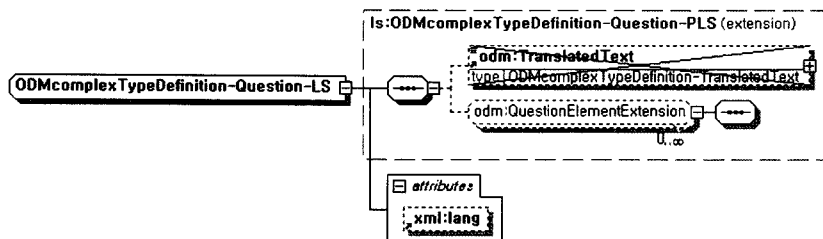


그림 4 자식 엘리먼트 TranslatedText를 사용하지 않고 언어를 설정

두 번째 경우는 다국어를 지원해야 할, 스트링를 갖는 엘리먼트에 언어를 설정하기 위한 방법이다. 이를 위해 해당 엘리먼트의 바인딩 타입에서 새로운 서브 타입을 extension으로 정의하고, 그것에 언어 설정을 위한 애트리뷰트 xml:lang을 추가한다. 예를 들어 그림 5와 같이, 엘리먼트 <StudyName>의 바인딩 타입 ODMcomplexTypeDefinition-StudyName을 확장하여 애트리뷰트 xml:lang을 갖는 새로운 서브 타입 ODMcomplexTypeDefinition-StudyName-LS를 생성한다.

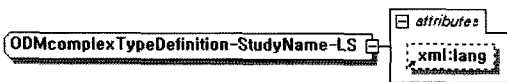


그림 5 스트링 타입에 바인딩된 타입의 언어 설정

세 번째 경우는 스트링 값을 갖는 애트리뷰트에 언어를 설정하기 위한 방법이다. 먼저 이런 애트리뷰트를 갖는 타입에서 속성 use의 값을 "prohibited"으로 변경하여 해당 애트리뷰트를 비활성화한 서브 타입을 restriction을 통해 생성한다. 그리고 생성된 서브 타입에 비활성화한 애트리뷰트와 동일한 이름의 엘리먼트와 애트리뷰트 xml:lang을 추가한 또 다른 새로운 서브 타입을 extension을 통해 생성한다. 예를 들어, 그림 6에서는 엘리먼트 <Alias>의 애트리뷰트 Context와 Name에서 다국어 지원이 가능하도록 이들 속성을 엘리먼트로 변환하는 스키마를 스키마 다이어그램으로 나타낸 것이다. 애트리뷰트의 본래 use 속성 값이 "optional"이면 해당 엘리먼트의 최소 출현 횟수를 "0"으로(minOccurs="0") 정의하고, "required"이면 "1"로(minOccurs="1") 정의한다. 그리고 다국어를 지원하기 위해서 해당 엘리먼트의 최대 출현 횟수는 "unbounded"로(maxOccurs="unbounded") 한다.

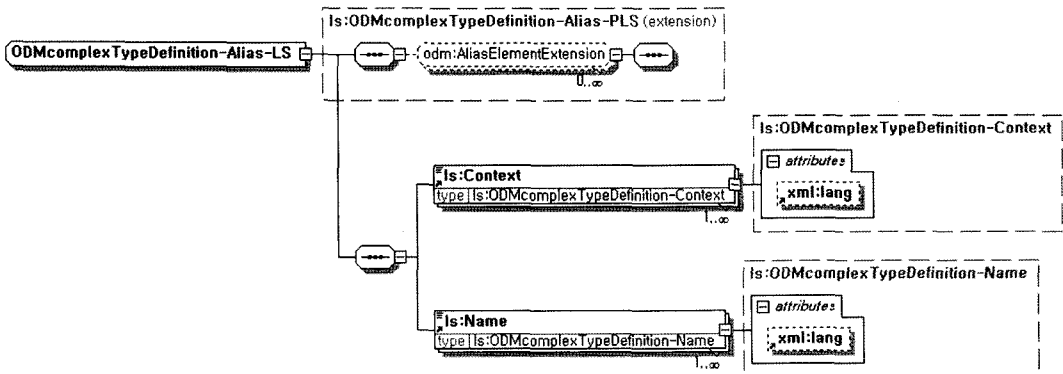


그림 6 스트링 값을 갖는 애트리뷰트를 위한 언어 설정

네 번째 경우는 엘리먼트의 콘텐츠 전체에 디폴트 언어를 설정하기 위하여, 해당 타입을 extension하여 애트리뷰트 xml:lang을 추가한 서브 타입을 생성하는 것이다. 예를 들어, 그림 7의 ODMcomplexTypeDefinition-ODM-LS 서브 타입은 ODMcomplexTypeDefinition-ODM 타입을 extension으로 확장하여 애트리뷰트 xml:lang을 추가하였다. 그러나 이를 위해서는 먼저 ODM 스키마에서 의미로 정의된 엘리먼트 ODM의 바인딩 타입이 ODMcomplexTypeDefinition-ODM과 같은 이름을 가지도록 수정해야 한다.

이와 같은 방법으로 본 논문은 표 8과 같이, CDISC

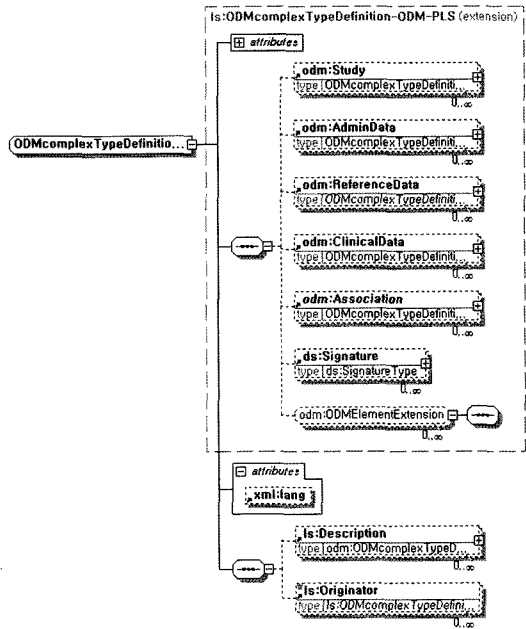


그림 7 엘리먼트 콘텐츠 전체의 디폴트 언어 설정

표 8 ODM에서의 다국어 설정의 문제점 및 다국어 지원 방법

ODM에서의 언어 설정의 문제점	다국어 임상시험을 위한 ODM의 확장
엘리먼트<TranslatedText>를 이용한 언어 정의는 다국어의 표현이 가능하지만, 문서의 길이 증가 스트링 타입에 바인딩된 엘리먼트의 출현 횟수의 제한 다국어로 표현될 가능성 있는 애프트뷰트의 다국어 설정 불가 ODM 문서 전체의 디플트 언어 설정 불가	서브 타이핑 방법을 사용하여 ODM 스키마의 확장

의 ODM에서의 언어 설정에 대한 4가지 문제점에 대한 해결 방법으로 서브 타이핑 방법을 사용한 ODM의 확장을 제안하였다.

4. 결론

임상시험이 다민족을 대상으로 여러 국가에서 시행되는 현재의 환경을 고려하였을 때, CDISC 표준은 많은 제약을 가지고 있다. 특히, CDISC 표준 중 임상시험 데이터의 콘텐츠 및 형식에 해당하는 SDTM과 ODM에서의 다국어 지원 기능은 매우 미비하다.

본 논문은 CDISC의 SDTM과 ODM에서의 언어 설정에 대한 문제점을 지적하고, 이를 해결하기 위해 SDTM과 ODM의 확장을 제안하였다. SDTM에서는 다국어 지원을 위한 새로운 도메인 LS를 설계하였고, ODM에서는 설계한 LS를 지원하는 ODM의 확장 스키마를 서브타이핑 방법으로 구현하였다. 설계한 LS 도메인은 기존의 도메인과 유사한 구조 및 접근 방법을 갖고 있으므로, CDISC 사용자가 쉽게 이해하고 사용할 수 있다. 또한, 확장된 ODM 스키마는 기존 ODM 스키마의 기본 정의를 배제하지 않고, 그 정의에 기반하여 다국어 지원을 위해 서브타이핑 방법을 사용하여 확장되었다. 따라서 기존 ODM 스키마에 따르는 인스턴스 문서뿐만 아니라 확장된 ODM 스키마에 따르는 인스턴스 문서도 각각 제공할 수 있다. 이와 같은 SDTM과 ODM을 기반으로 임상시험 데이터 처리시 다국어를 지원하면, 다국적으로 임상시험이 수행되는 경우, 각 나라의 언어로 작성된 CRF를 사용할 수 있고, 다민족의 피험자들 각각이 원하는 언어로 작성된 CRF를 제공할 수 있다. 그리고 심사 기관이 원하는 언어로 표현된 제출용 임상시험 데이터를 생성할 수 있다.

한편, CDISC 사용자가 본 논문에서 제안한 서브타이핑 방법에 의한 확장 스키마를 사용할 때, XML Schema에 대한 깊은 지식을 가지지 않고서는 확장 스키마를 이해하기가 쉽지 않을 것이다. 그러나 CDISC를 이용해 다국적 임상시험을 하거나 CDISC를 채용하는 국가를 확대하기 위해서는, 본 논문에서 서술한 다국어 설정 기능이 필수적이다.

참고 문헌

- [1] C. Willoughby, D. Fridsma, L. Chatterjee, J. Speakman, J. Evans, R. Kush, "A Standard Computable Clinical Trial Protocol: The Role of the BRIDG Model," *Drug Information Journal*, vol.41, pp.383-392, 2007.
- [2] R. Kush, "Data Standards for Clinical Development-Progress Update," *APPLIED CLINICAL TRIALS*, vol.11, no.4, pp.35-44, 2002.
- [3] M. Wheeldon, CDISC: Eliminating Standard Deviation. *International Clinical Trials - Spring edition*, pp.32-5, 2006.
- [4] W.R. Kubick, S. Ruberg, E. Helton, "Toward a Comprehensive CDISC Submission Data Standard," *Drug Information Journal*, vol.41, no.3, pp.373-382, 2007.
- [5] T. Souza, R. Kush, J.P. Evans, Global clinical data interchange standards are here! *Drug Discov Today*, vol.12, pp.174-81, 2007.
- [6] CDISC Study Data Tabulation Model Version 1.1 Final, <http://www.cdisc.org/models/sdtm/v1.1/index.html>.
- [7] Operational Data Model Final Version 1.3, <http://www.cdisc.org/models/odm/v1.3/index.html>.
- [8] Extensible Markup Language (XML) 1.0 (Third Edition), <http://www.w3.org/TR/2004/REC-xml-20040204/>.
- [9] M.W. Haber, B.W. Kisler, M. Lenzen, L.W. Wright, "Controlled Terminology for Clinical Research: A Collaboration Between CDISC and NCI Enterprise Vocabulary Services," *Drug Information Journal*, vol. 41, no.3, pp.405-412, 2007.
- [10] A.G. Oliveira, N.C. Salgado. Design aspects of a distributed clinical trials information system. *Clinical Trials*, pp.385-396, 2006.
- [11] H.O. Ht, S.C. Chow. "Design and analysis of multinational clinical trials," *Drug Information Journal*, pp.1309-1316, 1998.
- [12] Formedix Technology Primer: Formedix and CDISC Standards.
- [13] XML Schema Part 0: Primer Second Edition, <http://www.w3.org/TR/xmlschema-0/>.
- [14] Tags for the Identification of Languages, <http://www.ietf.org/rfc/rfc3066.txt>.
- [15] Specification for the Operation Data Model (ODM). Version 1.2.1 ReadMe.

[16] D. Jon, Professional XML Schemas, Wrox Press Ltd, 2001.

[17] ALTOVA, <http://www.altova.com>.

엄 지 현

정보과학회논문지 : 컴퓨팅의 실제 및 레터
제 15 권 제 7 호 참조



최 인 영

1989년 이화여자대학교 영어영문학과 (문학사). 1993년 서강대학교 경영학(경영학석사). 2004년 이화여자대학교 경영학과(경영학박사). 1990년~1993년 Ernst & Young. 2004년~2006년 Georgetown University Imaging Science & Information Systems Research Center 박사후 연구원. 2006년~현재 가톨릭대학교 임상연구지원센터 연구교수. 관심분야는 정보화 전략, SIS/SUIT, 지식경영, 기술 수용 모델 (Technology Acceptance Model), 임상연구지원시스템, CDISC 표준



김 석 일

1989년 가톨릭의대(의학사). 1995년 연세대학교 대학원 보건학과(보건학석사) 1998년 연세대학교 대학원 보건학과(보건학박사). 2004년 University of Utah 이학석사(의료정보학). 1998년~현재 가톨릭대학교 의과대학 부교수. 2008년~현재 세계보건기구(WHO-FIC) 용어분과 공동위원장. 관심 분야는 Clinical Decision Support, Clinical Terminology

김 혁 만

정보과학회논문지 : 컴퓨팅의 실제 및 레터
제 15 권 제 7 호 참조