Development of an automated appendix generation system (ARGUS) for clinical study reports

Doo yeon Jang[1,2], Seunghoon Han[1,2†], Dong-SeokYim[1,2,*]

[1]Department of Clinical Pharmacology and Therapeutics, Seoul St. Mary's Hospital

[2]PIPET (Pharmacometrics Institute for Practical Education and Training), College of Medicine, The Catholic University of Korea, Seoul, Korea 06591

Department of Clinical Pharmacology and Therapeutics, Seoul St. Mary's Hospital, 222 Banpo-daero, Seocho-gu, Seoul, Korea 06591; Telephone: +82-2-2258-7327

E-mail: jangdooyeon@catholic.ac.kr yimds@catholic.ac.kr

초록 데이터 처리 및 도표화는 임상 연구 보고서에 부록을 작성할 때 시간 소모적인 작업이다. 저자는 SAS (버전 9.3) 및 R (버전 3.3.1: PK 플롯 생성 용)을 사용하여 CDISC/SDTM 표준에 부합하는 자동 부록 생성 시스템 (ARGUS)을 개발했다. 이 시스템은 하나의 주 프로그램과 세 개의 서브 프로그램으로 구성되어 있다. 일반적인 데스크탑 환경에서 제출 버튼을 누른 후 약 1 분 만에 데이터베이스 파일을 MS Word 형식의 부록 문서로 변환한다. ARGUS 시스템을 사용하여 약 8 일간 팀을 구성한 부록을 작성하던 작업이 6 ~7 시간 내에 완료되었다.

서론

임상 연구 보고서 (clinical study report 이하 CSR)는 임상 시험 결과에 대한 공식적인 문서이다.[1] CSR의 신뢰성을 보장하기 위해서는 임상 시험 데이터를 적절하게 관리해야 한다.[2] CSR의 작성시에는 임상 시험 데이터로부터 생성되는 많은 종류의 표가 필요하다. 다양한 데이터를 테이블, 특히 CSR 부록 용 테이블로 오류 없이 변환하는 것은 scientific writer에게 커다란 부담이다.[3]

일반적으로 CSR 및 부록에 대한 테이블은 데이터베이스 파일 (Excel 파일 유형으로 제공/전달)의 복사 및 붙여 넣기를 사용하여 작성되고 있다. 이러한 방식은 오류에 취약하며 작업량이 증가할수록 오류가 생길 가능성이 높아진다.[4]

보고서의 신뢰성과 정확성을 향상시키기 위해서는 자동화되고 재생 가능한 프로세스가 필요하다.[5] 그렇기 때문에 우리는 CSR과 그 부록에서 복잡한 테이블을 자동으로 만들 수 있는 "자동 보고서 생성 및 업데이트 코드 스크립트 (automated report generation and update code script 이하 ARGUS)"라는 시스템을 개발하였다.

이론 및 방법

이론

ARGUS는 data 처리 및 tabulation system으로서 data manipulation 및 documentation을 한다.

ARGUS system에서 사용된 SAS program (이하 SAS)은 DATA step이라고 하는 data 처리형식구문(statement)이 있으며, 이를 두 가지 단계로 처리한다.

컴파일 단계(compile phase)와 실행단계 (execution phase)가 이다. 컴파일 단계에서는 작성된 코드를 읽고 구문의 정확성을 검사한다. 그리고 환경을 설정하고 machine code를 생성한다.

이후 실행 단계에서는, DATA step의 경계 내에서 단계의 반복마다 한 번의 명령문이 순서대로 실행된다. 이 Step 에는 암시 루프가 포함되어 있는데, DATA Step의 경계는 step의 시작에 위치한 DATA statement와 단계의 마지막에 있는 암묵적 혹은 명시 적 RUN statement 사이에 있다. DATA step의 마지막은 이 DATA step의 마지막 관측치가 처리 될 때까지 암시된 RETURN으로 사용된다.

컴파일 할 때에는 프로그램 데이터 벡터(이하 PDV) 가 초기화 된다. PDV는 SAS가 데이터 세트를 작성하는 메모리의 논리적 영역으로서, 하나의 관측치마다 PDV 영역에 입력 된다. 프로그램이 실행되면 SAS는 입력 버퍼 또는 기존 데이터 세트에서 데이터 값을 읽거나 SAS 언어 문을 실행하여 데이터 값을 만들게 된다. 데이터 값은 프로그램 데이터 벡터의 적절한 변수에 할당되고 거기에서 SAS는 값을 출력 데이터 세트에 관측치로 사용한다.

방법

시스템 구성 요소: ARGUS 시스템은 PK 플롯 생성 용 코드 (R 버전 3.3.1이 사용됨)를 제외하고는 SAS (버전 9.3, 동적 데이터 교환 (dynamic data exchange: DDE) 프로그래밍은 Windows 용 SAS, 버전 9.3 이상에서만 호환 가능)로 작성되었다.

Linux audit daemon 은 CRF 데이터베이스 파일 및 코드 스크립트 파일의 모니터링 (파일 추적)에 사용되었다. 코드 스크립트 시스템은 2x2 크로스 오버 디자인 (생물학적 동등성 테스트 또는 약물 - 약물 상호 작용 연구)의 CSR 부록에 맞게 설계되었다. ARGUS 시스템 구성 요소는 Figure. 1과 같다.

코드 스크립트의 구조: 시스템의 표 형태는 앞에서 만든 2x2 크로스 오버 디자인의 CSR 부록을 기반으로 개발되었다. 스크립트는 주 프로그램, 서브 프로그램 및 모듈로 구성된다.

메인 프로그램은 각 서브 프로그램을 실행하고, 서브 프로그램은 입력, 조작 절차 및 출력에 따라 분류된다. 각 서브 프로그램은 2 개 또는 3 개의 모듈을 포함한다. (Figure. 2) 모듈 스크립트는 사용자의 필요에 따라 수정 될 수 있다.

시스템 작동: ARGUS가 프로세스를 자동화하려면 메인 프로그램을 실행하기 전에 다음 전제 조건이 있어야 한다.

데이터베이스 파일 경로, 스크립트 파일 경로 및 매크로 변수를 입력을 필요로 한다. 메인 프로그램을 실행하기 전에 개별 PK 플롯의 R 코드를 실행해야 한다.

메인 프로그램을 실행 한 후 CSR 부록 보고서가 완성된다. 주 프로그램에서 사용자가 submit 단추를 누르면 시스템이 활성화되고 프로세스가 시작된다.

프로세스가 완료되면 출력 파일이 저장되고 MS Word가 자동으로 시작되어 보고서 파일이 생성된다. (Windows 버전의 SAS). 임상 시험에 따라 프로그램을 수정하는 경우 수정된 프로그램의 내용을 메인 프로그램에 기록하는 것을 권장한다. 작업 흐름은 다음과 같다.

- 입력 및 수정 (입력 경로 및 매크로 변수, 형식, 모듈 추가 / 수정 등) -> 실행 (제출) -> 출력

시스템 제약 조건:

- 시스템은 CDISC / SDTM 표준에 따라 생성되었으므로 비 표준 CRF 데이터베이스

파일은 호환되지 않는다.

- 한글(UTF-8 encoding)은 처리 할 수 없다.
- 코드의 계층 구조가 변경되면 시스템이 제대로 실행되지 않는다.
- 현재 버전은 2 × 2 교차 설계 (생물학적 동등성 시험 또는 약물 상호 작용 연구)를 위해 설계되었으므로 다른 연구 설계에는 코드 수정이 필요하다.

결과 및 논의

일반적인 데스크톱 환경에서 제출 버튼을 누른 후 실행까지 약 50-60 초가 소요된다. 컴퓨팅 성능 및 데이터 크기에 따라 시간 차이가 있을 수 있다. 시스템의 동작 순서는 다음과 같다.

1. ARGUS는 ARGUS 및 프로젝트 이름 디렉토리 아래의 Linux 서버에서 데이터 파일 및 코드 스크립트 파일을 위한 새 디렉토리를 만든다.

2. NCA 데이터 세트, individual PK parameter 테이블, 원본 CRF 데이터베이스 파일 및 코드 스크립트 파일을 권한이 부여 된 FTP 계정을 통해 프로젝트 디렉토리 아래의 각 하위 디렉토리에 업로드한다.

3. 메인 프로그램을 실행하기 전에 R 플롯 코드를 실행해 한다. 사용자는 R 스튜디오 서버에 연결하여 개별 PK 플롯을 그려 r 플롯 코드를 사용하여 저장해야 한다.

4. 사용자는 SAS Studio에 연결하거나 SAS, 버전 9.3을 실행 한 다음 메인 프로그램을 불러 와야 한다. 코드 스크립트 파일 및 매크로 변수는 새 보고서에 대해 수정 될 수 있다.

5. 모든 파일이 준비되면 제출 단추를 누른다. 프로세스가 끝나면 시스템이 MS Word를 열고 Word 파일이 서버에 저장된다.

임상 실험실적 검사 결과의 원본 CRF 데이터베이스 파일은 매우 방대하며, 임상 실험실적 검사항목의 테이블 작성은 많은 시간을 소비하는 작업이다. ARGUS 시스템을 사용하면 SAS 형식 프로시저가 있는 두 개의 변수 (LBTESTCD 및 LBORRES)가 매우 간단한 코드를 사용하여 교차 테이블을 생성할 수 있다. ARGUS에 의해 생산 된 임상 실험실의 항목 중 hematology부분의 표의 일부가 Fig. 3에 있다.

마찬가지로, vital sign 데이터는 manipulation 및 report 모듈에서 처리된다. 그림 4는 수축기 혈압 (SBP) 표의 일부이다. 테이블 머리글의 맨 아래 줄. 4 및 5 (이상 반응 표) 또한 template 모듈을 사용하여 이중선으로 표시 되었다.

그 동안 우리 팀에서 완료하는 데 약 8 일이 걸린 부록 작업을 ARGUS 시스템을 사용하여 6 ~ 7 시간 내에 완료할 수 있었다. 현재 버전은 각 CRF 데이터베이스 파일에 대해 마이너 코드 수정이 필요하다. 사용자는 30 분 이내에 변수 및 실행 입력을 완료 할 수 있지만 실행 후 코드를 수정하는 데는 시간이 소요된다.

따라서 대부분의 테이블은 제대로 작성되지만 일부는 코드를 수정해야 할 수 있다. 혈장 농도 데이터는 기관을 분석하여 다양한 형식으로 전달되기 때문에 plasma concentration 모듈을 각 임상 시험에 맞게 수정한다. 전달된 농도 자료가 표준화 될 경우, 이 변경 단계는 pharmacokinetic 분석 및 그 테이블화에서 생략 가능할 것이다.[6]

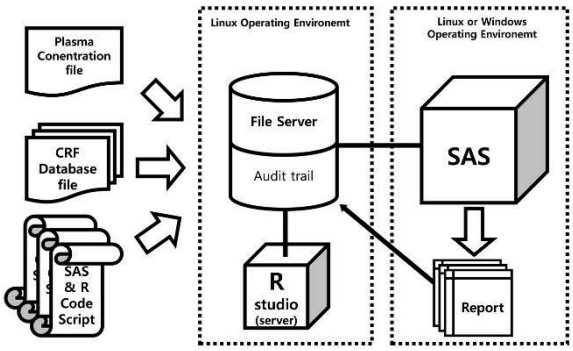

Figure 14. ARGUS 시스템 구성 요소

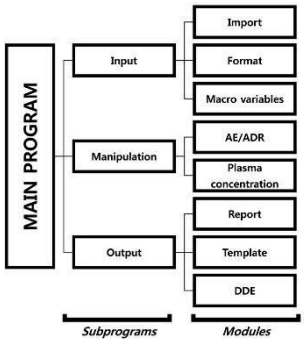

Figure 2. ARGUS 의 코드 구조

Enrollment No.	WBC			RBC			Hemoglobin			Hematocrit			Platelet			seg.Neutrophils		
	Scr	Day 1	DAY 8	Scr	Day 1	DAY 8	Scr	Day 1	DAY 8	Scr	Day 1	DAY 8	Scr	Day 1	DAY 8	Scr	Day 1	DAY 8
A010	4.67	3.69	3.64	5.18	4.75	4.78	16.5	15.5	15.1	47.7	43.8	45	264	210	211	61.3	57.2	65.4
A020	5.24	.	.	5.16	.	.	15.4	.	.	44.6	.	.	182	.	.	58.9	.	.
A021	5.07	4.46	6.07	4.62	4.5	4.52	13.8	13.5	13.6	42.2	40.3	41.1	296	301	293	47.9	43.3	66.4
A030	8.16	5.17	5.38	4.52	4.14	4	14.8	13.8	13.3	45.9	42.9	40.7	294	258	217	68.6	56.8	59.5
A040	5.64	5.6	6.16	5.41	5.3	5.53	15.7	15.2	15.5	45.8	45.6	47.9	199	176	194	56.5	61.4	56.2
A050	3.98	5.13	4.98	5.32	4.68	4.96	16.3	14.5	15.1	48	41.9	44.7	249	211	218	46	40.1	44.6
A060	3.98	4.2	5.38	5.3	5.38	4.95	15.2	15.6	13.9	45.6	45.8	42.5	218	232	231	51.2	57.2	58.5
A070	5.93	4.86	5.5	4.9	4.6	4.63	14.9	13.8	13.8	42.5	41.3	41.3	324	308	332	56.2	61.7	58.8
A080	6.21	5.38	5.13	5.55	5.39	5.26	16.1	15.7	15	47	46.4	45.4	314	300	298	47.6	52.8	52.7
B130	12.89	8.35	6.4	4.97	4.46	4.42	15.2	14	13.5	45.4	40.5	41.1	269	223	302	68.9	50	51.8
B140	8.08	5.13	4.88	5.45	5.14	4.96	16.6	16	15.1	49.2	46.7	46	198	227	202	69.5	53.3	58.3
N	29	28	27	29	28	27	29	28	27	29	28	27	29	28	27	29	28	27
MEDIAN	5.77	4.83	4.98	5.16	4.76	4.77	15.4	14.6	14.1	45.8	43.35	42.8	249	240.5	217	56.2	54.65	54.5
MIN	3.92	3.69	3.64	4.52	4.14	4	13.8	13.1	12.6	41.7	40.1	38.2	164	115	121	31.9	24.2	28.3
MAX	12.89	8.35	6.62	5.56	5.39	5.53	16.9	16.1	15.5	50	46.8	47.9	329	323	332	83	66.5	67
ARITHMETICMEAN	6.23	5.02	5.04	5.11	4.78	4.73	15.45	14.58	14.19	45.79	43.15	42.91	246.28	236.32	223.67	56.77	52.88	54.79
SD	1.99	0.96	0.81	0.28	0.35	0.35	0.8	0.83	0.78	2.09	2.15	2.28	48.51	48.67	48.56	10.87	9.21	8.62
CV	31.89	19.18	16.11	5.51	7.37	7.35	5.16	5.72	5.48	4.57	4.98	5.32	19.7	20.59	21.71	19.15	17.41	15.73

Figure 3. ARGUS 로 작성된 실험실적 검사 (hematology) 테이블

	SBP(mmhg)																	
	SCR	Period 1							Period 2									
	SCR	D-1	Day 0			Day 1	Day 2	Day 3	Day 6	Day 7					Day 8	Day 9	Day 10	PSV
A010	130	136	125	119	126	135	121	136	135	130	124	122	122	124	134	131	121	128
A020	115	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
A021	108	119	116	107	94.0	93.0	107	114	124	106	109	110	93.0	100	125	115	126	.
A030	123	124	113	112	101	104	110	110	112	111	107	115	113	109	107	108	108	115
A040	125	137	122	118	110	105	117	128	121	131	112	120	107	116	127	127	119	124
A050	108	110	114	116	97.0	125	114	124	125	122	120	127	122	126	126	122	127	107
A060	159	152	107	109	120	99.0	99.0	107	117	130	114	108	107	120	117	114		
B130	122	124	111	109	95.0	111	112	110	120	118	126	128	108	108	116	110	113	144
B140	113	123	110	110	100	100	103	113	118	105	102	104	107	106	102	115	121	120
N	29.0	28.0	28.0	28.0	28.0	28.0	28.0	28.0	28.0	27.0	27.0	27.0	27.0	27.0	27.0	27.0	26.0	25.0
MEDIAN	124	123	115	110	109	108	116	124	121	122	113	115	111	109	114	120	121	124
MIN	97.0	99.0	92.0	90.0	90.0	93.0	93.0	101	103	99.0	102	99.0	91.0	92.0	93.0	101	98.0	105
MAX	137	139	128	139	141	139	140	140	141	143	140	139	130	139	134	139	139	144
ARITHMETICMEAN	122	123	116	112	109	111	115	121	122	122	115	116	111	112	115	119	120	123
SD	10.0	9.3	10.8	11.0	13.0	11.5	11.2	11.6	10.9	11.6	9.0	10.0	9.9	10.5	10.1	9.5	9.5	10.7
CV	8.2	7.6	9.3	9.8	11.9	10.4	9.7	9.5	8.9	9.5	7.9	8.6	8.9	9.4	8.7	7.9	7.9	8.7

Figure 4. ARGUS 로 작성된 수축기 혈압 (SBP) 테이블

	Severity			Event		Causal Relationship	
	Mild	Moderate	Severe	n*	%	Possible	Unlikely
GASTROINTESTINAL DISORDERS							
ABDOMINAL PAIN LOWER	2	0	0	2(2)	7.14	2(2)	
DIARRHOEA	3	0	0	3(3)	10.7	3(3)	
INVESTIGATIONS							
ANION GAP INCREASED	1	0	0	1(1)	3.57		1(1)
ASPARTATE AMINOTRANSFERASE INCREASED	1	0	0	1(1)	3.57		1(1)
BLOOD CREATINE PHOSPHOKINASE INCREASED	2	0	0	2(2)	7.14		2(2)
BLOOD PH DECREASED	1	0	0	1(1)	3.57		1(1)
BLOOD URINE PRESENT	1	0	0	1(1)	3.57		1(1)
WHITE BLOOD CELLS URINE POSITIVE	1	0	0	1(1)	3.57		1(1)
NERVOUS SYSTEM DISORDERS							
DIZZINESS	1	0	0	1(1)	3.57		1(1)
HEADACHE	0	1	0	1(1)	3.57	1(1)	
RESPIRATORY, THORACIC AND MEDIASTINAL DISORDERS							
EPISTAXIS	2	0	0	2(2)	7.14		2(2)
NASAL CONGESTION	1	0	0	1(1)	3.57		1(1)
PRODUCTIVE COUGH	2	0	0	2(2)	7.14		2(2)
INFECTIONS AND INFESTATIONS							
NASOPHARYNGITIS	2	0	0	2(2)	7.14		2(2)
PHARYNGITIS	1	0	0	1(1)	3.57		1(1)

Figure 5. ARGUS 로 작성된 이상반응(AE)테이블

결론

이 보고서에 설명 된 시스템은 프로토 타입 단계에 있지만 지식 기반 시스템을 구축하는 데있어 의미 있는 첫 번째 단계로 볼 수 있다.

이 시스템은 실제 임상 시험에서 베타 테스트를 통한 검증이 필요하다. 프로토 타입 버전의 튜닝을 통해 교육 자료와 매뉴얼이 발행될 예정이다.

감사의 글

본 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단 첨단 사이언스·교육 허브 개발 사업의 지원을 받아 수행된 연구임(NRF-2016-936606).

참고문헌

1. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Ich Harmonised Tripartite Guideline - Structure and Content of Clinical Study Reports (E3). 1995:1-49.

2. Krishnankutty B, Naveen Kumar B, Moodahadu L, Bellary S. Data management

in clinical research: An overview. *Indian J Pharmacol*. 2012;44:168. doi:10.4103/0253-7613.93842.

3. Wieseler B, Kerekes MF, Vervoelgyi V, McGauran N, Kaiser T. Impact of document type on reporting quality of clinical drug trials: a comparison of registry reports, clinical study reports, and journal publications. *Bmj*. 2012;344:d8141. doi:10.1136/bmj.d8141.

4. Hong MKH, Yao HHI, Pedersen JS, et al. Error rates in a clinical data repository: lessons from the transition to electronic data transfer--a descriptive study. *BMJ Open*. 2013;3:1-7. doi:10.1136/bmjopen-2012-002406.

5. Peng RD. Reproducible research and Biostatistics. *Biostatistics*. 2009;10:405-408. doi:10.1093/biostatistics/kxp014.

6. Wood F, Schaefer P, Carolina N, Lewis R. Considerations in the Submission of Pharmacokinetics (PK) Data in an SDTM-Compliant Format. 2012:1-8.