

신장암 표준임상빅데이터 구축 및 머신러닝 기반 치료결정지원시스템 개발

Constructing a Standard Clinical Big Database for Kidney Cancer and Development of Machine Learning Based Treatment Decision Support Systems

송원훈¹, 박미영^{2*}

Won Hoon Song¹, Meeyoung Park^{2*}

〈Abstract〉

Since renal cell carcinoma(RCC) has various examination and treatment methods according to clinical stage and histopathological characteristics, it is required to determine accurate and efficient treatment methods in the clinical field. However, the process of collecting and processing RCC medical data is difficult and complex, so there is currently no AI-based clinical decision support system for RCC treatments worldwide. In this study, we propose a clinical decision support system that helps clinicians decide on a precision treatment to each patient. RCC standard big database is built by collecting structured and unstructured data from the standard common data model and electronic medical information system. Based on this, various machine learning classification algorithms are applied to support a better clinical decision making.

Keywords : Renal Cell Carcinoma, Standard Medical Big Data, Clinical Decision Support System, Machine Learning

1 정회원, 제1저자, 양산부산대학교병원 비뇨의학과, 조교수 1 Dept. of Urology, Pusan National University Yangsan Hospital
E-mail: whsong1015@gmail.com

2* 정회원, 교신저자, 경남대학교 컴퓨터공학부, 조교수 2* Dept. of Computer Engineering, Kyungnam University
E-mail: mpark@kyungnam.ac.kr

1. 서론

신장암(kidney cancer)은 대부분 신장의 실질에서 발생하는 신장세포 암으로 주로 60~70대의 노년층에서 주로 발생한다. 2020년도에 전 세계적으로 새롭게 신장암으로 진단된 인구가 약 431,288명으로 나타났다. 아시아 지역에서만 진단된 인구가 전체의 36.6%로 수위를 차지하고 있고 [1], 신장암의 국내 발생률도 매년 지속적으로 증가 추세이며, 임파선이나 타장기에 전이된 경우에는 5년 생존율이 20% 미만으로 예후가 매우 불량하다 [2].

신장암 중에서 조직학적으로 90% 이상을 차지하는 것이 신세포암(Renal Cell Carcinoma, RCC) 이고, 그 중에서 70%는 clear cell(투명세포) RCC, 10-15%는 papillary(유두세포) RCC, 그리고 5%는 chromophobe(염색소성) RCC 이다 [3]. 신장암의 치료는 Fig. 1과 같이 임상 병기별과 조직학적 특성에 따라서 매우 다양하고, 또한 환자별 임상적 상황에 따라서도 서로 다른 치료 방법들이 선택될 수 있다. 크게는 수술을 하지 않는 추적관찰(active surveillance), 수술적 치료, 그리고 항암치료로 나눌 수 있지만, 수술적 치료는 다시 부분절제술과 전절제술로 나눌 수 있고, 그에 따른 수술 방법도 다양하다. 항암치료도 조직학적 특성에 따라서 치료 약제가 다르고, 약제의 종류들도 매우 다양하다. 수술적 치료와 항암

치료를 병행하기도 한다. 가이드라인에서 제시하는 표준적인 치료 방법들이 있으나, 실제 임상 현장에서는 다양한 조건과 환자별 특성에 따라서 서로 다른 치료 방법들이 선택되기도 한다 [4].

국내 대학병원의 진료 환경은 의사 수에 비해 환자 수는 너무 많아서, 환자 1명당 진료 시간이 수분을 넘기기 어려운 열악한 실정이다. 이런 진료 환경 속에서는 신장암 환자들에게 다양한 치료법을 설명하며 개개인별 적합한 치료법을 소개하고, 환자를 이해시키기란 어렵다. 따라서 국내 진료 환경에서는 검사 및 치료의 기준을 다각화하여, 환자별 개개인의 건강 상태에 맞는 인공지능을 기본으로 한 환자 맞춤형 검사/진단/치료 자동화 프로토콜을 적용하여, 임상 의사와 환자들을 위한 임상결정지원시스템이 절실하다고 볼 수 있다.

최근 의료인공지능 분야에서는 의료빅데이터를 기반으로 정밀의료를 구현하기 위하여 다양한 인공지능 알고리즘을 사용하여 특정 질환에 대한 맞춤형 임상결정지원시스템(CDSS, Clinical Decision Support System) 개발 연구가 활발히 진행되고 있다 [5-7]. 그러나 신장암의 경우 이러한 인공지능 알고리즘을 적용한 연구가 극히 드물다. 더욱이, 각 의료기관의 임상 의는 서로 다른 데이터의 형태로 환자를 진료하고 있으므로 다기관 연구 또한 매우 어렵다. 따라서, 최근 의료용어와 데이터베이스 구조를 표준화하여 다기관 공동연구가 가능

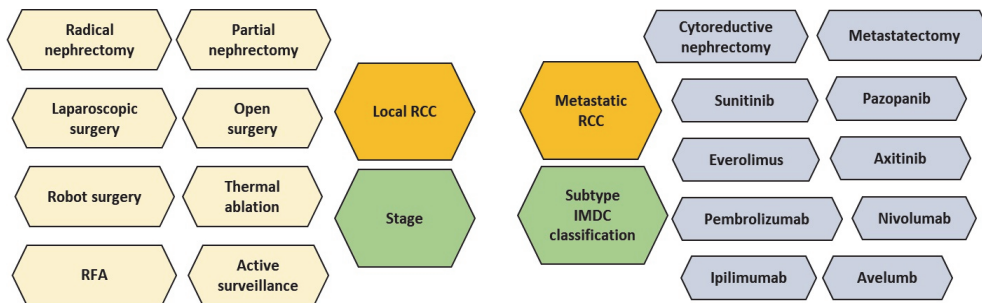


Fig. 1 Various RCC treatment methods

하도록 하는 공통 데이터 모델 (CDM, common data model)을 사용한 다기관 공동연구 및 임상 빅데이터 연구가 진행되고 있다 [8-9].

본 연구에서는 의료기관에 구축된 CDM을 활용하여 정형데이터를 수집하고, 암 질환 치료에 가장 중요한 항목인 병리 기록지, 영상판독결과 기록지 등의 비정형데이터를 확보하여 분석 가능한 데이터로 정제하여 신장암 빅데이터를 구축하였다. 그 후 신장암 치료에 영향을 미치는 특징 (features)을 추출하고 머신러닝 알고리즘을 적용하여 임상현장에서 임상이가 최적의 치료 방법을 찾을 수 있도록 지원하는 맞춤형 치료결정지원시스템을 개발한다.

본 연구는 다음과 같이 구성되었다. 1장에서는 신장암 치료결정시스템에 필요한 배경을 설명하고, 2장에서는 정형 및 비정형 데이터를 수집하여 신장암 표준데이터베이스를 구축하는 과정을 설명한다. 3장에서는 구축된 데이터베이스를 활용하여 머신러닝 알고리즘을 적용한 결과를 보여주며, 4장에서는 알고리즘의 결과를 분석한다. 마지막으로 5장에서는 본 논문을 마무리한다.

2. 신장암 표준데이터베이스 구축

2.1 데이터 준비 및 전처리

신장암 표준빅데이터를 구축하기 위하여 양산부 산대학교병원의 CDM 데이터베이스 및 병원정보 시스템(Electronic Medical Records)에서 2008년 11월 25일~ 2021년 6월 29일의 12년 동안 신장암(renal cell carcinoma, RCC)으로 진단받은 환자의 데이터를 수집하였다. 국제 표준 의료데이터베이스인 CDM에서 환자의 나이, 성별 등의 정보와 진단검사를 수행한 정보, 수술 또는 처치의 정

보, 처방의약품 정보 등의 정형 데이터를 추출하였다. 암 환자의 치료에 가장 중요한 병리기록지와 CT 영상판독결과와 같은 비정형데이터는 병원 정보전산팀의 협력으로 EMR에서 직접 추출하였다. 수집한 전체 데이터는 2,153명이고, 그 중 남성 1,358명(63%), 여성 795명(37%)으로 나타났다.

Table 1과 같이 신장암 코호트 중 CDM으로 구축된 정형 데이터베이스에서 RCC 진단을 받은 환자 2,153명 중 혈액검사의 정보를 가지고 있는 진단검사(Measurement) 1,905명, 처방의약품(Drug) 1,737명, 수술(Procedure) 1,013명의 정형데이터를 추출하였다. CDM에 존재하지 않는 비정형데이터는 EMR에서 병리기록지(1,289명), 간호정보기록지(1,349명), 방사선치료기록지(1,979명), 수술기록지(996명)를 추가로 수집하였다(Table 2).

CDM과 EMR에 공통적으로 정보가 있는 수술 정보와 같은 경우, 더 많은 환자의 수술 정보가 있는 CDM 데이터로 분석하였다. CDM 내 진단검사의 경우 1,030개의 항목이 존재하였고, 그 중 80%의 환자가 검사를 시행한 항목을 수집하였다. 처방의약품의 경우 1,484 항목을 수집하였다. 그

Table 1. Structured data from CDM

Data	Number of patients (%)
Condition	2,153 (100.00)
Measurement	1905 (88.48)
Drug	1,737 (80.68)
Procedure	1,013 (47.05)

Table 2. Unstructured data from EMR

Data	Number of patients (%)
Pathology	1,289 (59.87)
Patient history	1,349 (62.66)
Radiology history	1,979 (91.92)
Surgery history	996 (46.26)

중 항암제는 26 항목으로 나타났다. 비정형데이터인 병리기록지의 경우 2,153명의 환자 중 1,289명의 환자의 병리기록지를 수집하였다.

전체 신장암 표준데이터베이스를 구축하기 위하여 환자의 진단 당시의 데이터를 기준으로 필요한 의료정보를 추출하였다. RCC로 진단받은 모든 환자가 처음 진단받은 날짜를 ‘기준날짜’(index date)

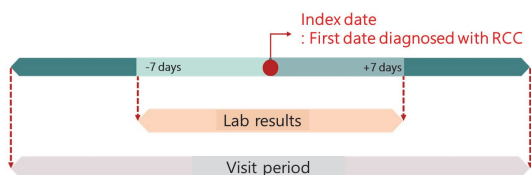


Fig. 2 Index date and lab data extraction

Urine OB	Urine Urd/Urine	Clr/Urine	Cuk/Urine	WBl/Urine	Epi
-	01 Clear	STRAW	0-2	0-2	0
-	01 Clear	LT/YELLOW-2	0-2	0	0
-	NORMAL Clear	Light yellow-0-2	0-2	0	0
-	Clear	Yellow	0-2	0-2	0
2+	01 Clear	LT/Yellow	0-10	0-2	0
-	01 Clear	yellow	0-5	0-2	0
-	NORMAL Clear	STRAW	0-2	0-2	0
-	01 Clear	STRAW	0-2	0-2	0
-	01 Clear	LT/Yellow	0-2	0-2	0
-	01 Clear	STRAW	0-2	0-2	0
-	NORMAL Clear	STRAW	0-2	0-2	0
-	01 Clear	LT/Yellow	0-2	0-2	0
-	01 Clear	STRAW	0-2	0-2	0
+	01 Clear	STRAW	0-2	0-2	0
++	01 Clear	LT/Yellow	0-2	0-2	0
3+	01 Clear	STRAW	0-10	0-2	0
+/2	NORMAL Clear	STRAW	0-2	0-2	0

(a) Unstructured lab results

Urine OB	Urine Urd/Urine	Clr/Urine	Cuk/Urine	WBl/Urine	Epi
5	1	0	17	1	0
5	1	0	13	1	0
5	5	0	12	1	0
5	2	0	18	1	0
7	1	0	13	5	0
5	1	0	16	4	0
5	5	0	17	1	0
5	1	0	17	1	0
5	1	0	13	0	0
5	1	0	13	0	0
4	5	0	17	1	0
5	1	0	13	1	0
5	1	0	17	1	0
3	1	0	13	1	0
8	3	1	16	3	0
4	5	0	17	1	0

(b) Unstructured CT report

Fig. 3 Extracting information from unstructured data

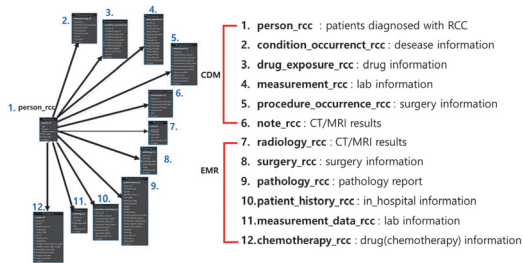


Fig. 4 Overall RCC standard database architecture

로 정의하고, 혈액검사 등의 다양한 진단검사 정보는 기준날짜에서 7일 전후의 검사정보 중 기준날짜와 가장 가까운 검사값을 추출하였다 (Fig. 2). 또한, 데이터 내의 결측값은 Multi variate Imputations by Chain ed Equations (MICE) 패키지를 사용하여 대체하였다 [10]. 영상 판독 결과의 경우 환자가 내원하여 촬영한 모든 CT 결과 중 기준날짜와 가장 가까운 날의 판독 결과를 사용하였다.

여러 진단검사 정보 중 정형화된 값으로 표현되지 않는 비정형 데이터 값은 Fig. 3(a)와 같이 정보를 손실하지 않도록 전처리를 수행하여 정형 데이터로 변환한 후 저장하였다. 특히, 텍스트 형태로 저장된 비정형데이터인 CT 판독 결과의 내용은 신장암 환자의 치료 방법을 결정하는데 매우 중요한 내용임에도 불구하고, 과거와 현재의 판독 결과의 양식이 다르고, 판독을 시행한 임상전문의가 각각 다른 형태의 판독 결과를 나타내고 있다. 즉, 한 의료기관에서 생성되었다고 하더라도 시간과 담당 전문의에 따라 이질적인 데이터가 생성되고 있다. 이러한 데이터를 표준데이터베이스에 저장하기 위하여 본 연구에 참여한 신장암 전문의가 모든 판독 결과를 재검증하고, 신장암 치료 방법 결정을 위해 공통적이고 필수로 나타나는 단어를 추천하여 그에 해당하는 단어를 추출한 후 표준데이터베이스에 저장하였다(Fig. 3(b)).

구축된 전체 표준데이터베이스는 Fig. 4와 같이 “person_rcc” 테이블을 기준으로 11개의 테이블로 구성된 구조를 보여준다.

데이터베이스 내에 들어갈 특정 환자의 데이터 중 전체 50%가 넘는 결측치를 가진 경우와 데이터의 표준분포에서 벗어난 이상치를 제외하였고, 범주형 데이터는 특정값으로 변환하는 전처리를 수행하였다. 최종적으로, 환자의 수는 1,867명으로 신장암 표준데이터베이스를 구축하였다.

3. 머신러닝 수행

3.1 데이터 레이블링

본 연구는 시스템의 적용 가능성의 타진과 성능 평가를 위하여 다양한 임상적 상황 중 치료 방법 결정에 가장 중요한 CT 판독결과지를 기준으로 치료 방법을 크게 총 3개의 그룹으로 나누어 구분하였다. 즉, 수술이나 항암치료를 받지 않은 추적관찰 그룹(Group1-Surveillance), 수술을 받은 그룹(Group2-Operation), 항암치료 또는 항암치료와 항암치료 전후 수술을 받은 그룹(Group3-Operation+Chemotherapy)으로 나누었다. 분류된 데이터의 최종 그룹별 환자의 수는 전체 1,867명의 환자 중 Group1 748명, Group2 893명, Group3 226명으로 분류되었다. 환자의 치료 방법은 모든 데이터에 대해 신장암 전문의가 직접 레이블링 작업을 수행하였다. 본 연구에서는 한 의료기관의 비뇨의학 임상전문이가 CT 판독 결과를 기준으로 레이블링을 수행하였으므로, 실제 환자가 받은 치료와는 차이가 있을 수 있으나, 후속 연구에서는 다기관 연구를 통하여 여러 가지 임상적 상황들을 고려한 실제 임상 현장을 고려한 시스템을 개발하고자 한다.

3.2 머신러닝 수행

구축한 신장암 표준빅데이터를 활용하여 치료결정지원시스템을 개발하기 위한 머신러닝 알고리즘을 적용하였다. 이를 위해 사용한 특징집합(feature set)으로 CDM에서 추출한 환자의 기본 정보, 진단검사 정보, 처방의약품 정보와, EMR에서 추출한 수술기록지, 영상판독결과 기록지 내의 신장암 치료법 분류에 영향을 미치는 중요 정보를

추출하여 총 243개의 특징을 사용하였다. 본 연구에서는 의료현장의 임상전문가와 환자가 직관적으로 이해할 수 있는 결과를 도출하기 위하여 Decision Tree 기반의 모델을 선택하였고, 또한 추가적으로 최근 머신러닝 알고리즘에서 우수한 성능평가를 보이는 앙상블 모델인 Random Forest, Gradient Boosting Machine(GBM), Extreme Gradient Boosting(XGBoost)을 함께 적용하여 치료결정지원시스템에 적용할 최적의 모델을 선정하고자 하였다 [11, 12]. 추출한 특징을 사용한 머신러닝 알고리즘의 성능 검증을 위해 5가지의 성능평가 지표와 ROC curve를 사용하여 알고리즘 간의 성능을 비교하였다.

Decision Tree [13]는 데이터에 있는 규칙을 학습을 통해 특정 조건을 자동으로 찾아내어 예측을 위한 트리를 생성하여 전통적인 분류 규칙을 만드는 알고리즘으로 머신러닝 비전문가가 결과를 해석하는데 매우 유용한 알고리즘이다. Random Forest [14] 알고리즘은 Decision Tree 알고리즘을 기반으로 데이터를 복원추출하는 배깅(Bagging) 방법을 적용하여 여러 개의 Decision Tree 모델을 생성한 후 그 결과를 합하여 예측하는 앙상블 알고리즘이다. Random Forest는 제한된 데이터에서 과적합의 위험을 줄이고, 특징의 중요도를 지표로 나타낼 수 있다는 장점으로 가장 많이 사용되는 알고리즘이다. GBM [15]과 XGBoost [16]는 잘못 예측된 데이터에 대해 가중치를 부여하여 오류를 점진적으로 개선해 나가는 부스팅(Boosting) 방법을 적용한 알고리즘으로 최근 많은 머신러닝 문제에서 활용되고 있다. 각 모델은 파이썬의 사이킷런 모듈 내 GridSearchCV() 함수를 사용해 최적의 하이퍼파라미터를 결정했다. Decision Tree와 Random Forest의 경우 과적합을 방지하기 위해 모델의 깊이(max_depth)와 트리 내 각 노드의 최소 샘플 수(min_samples_split)를 튜닝

하였다. Random Forest는 추가로 몇 개의 분류기를 만들것인지(n_estimators) 고려하였고, GBM과 XGBoost의 경우 또한 모델의 깊이(max_depth), 분류기의 개수(n_estimators)와 학습을 진행할 때마다 적용되는 학습률을 튜닝하였다.

4. 머신러닝 결과 분석

본 연구에서는 전체 데이터 중 70% (1,306명)를 훈련 데이터, 30% (561명)를 테스트 데이터로 사용하였다. 정확도를 향상시키기 위하여 교차검증(cross validation)을 수행하였고, 데이터의 크기를 고려하여 5-folds로 진행하였다.

Table 3. Hyperparameters used to train the machine learning models

Models	Selected features	All features
Decision Tree	max_depth=5 min_samples_split= 23	max_depth=4 min_samples_split= 12
Random Forest	n_estimators= 200 max_depth=12 min_samples_split=13	n_estimators=400 max_depth=7 min_samples_split=18
GBM	n_estimators=1000 max_depth=8 learning_rate=0.1	n_estimators=300 max_depth=9 learning_rate=0.1
XGBoost	n_estimators= 400 max_depth=4 learning_rate= 0.01	n_estimators=400 max_depth=4 learning_rate= 0.01

Table 4. Performance of applied machine learning algorithms

	Decision Tree	Random Forest	GBM	XGBoost
Precision	0.90	0.85	0.79	0.80
Specificity	0.92	0.93	0.92	0.92
Recall	0.71	0.74	0.73	0.73
F1-Score	0.70	0.75	0.73	0.73
Accuracy	0.86	0.87	0.86	0.86

를 고려하여 5-folds로 진행하였다.

Table 3과 같이 본 연구에 적용한 머신러닝 알고리즘 중 Decision Tree 알고리즘은 트리의 최대 깊이를 나타내는 max_depth= 4, 노드 분할을 위한 최소한의 샘플 데이터 수인 min_samples_splits=12인 경우 가장 높은 예측 정확도를 나타내었다. max_depth가 클수록, min_samples_splits가 작을수록 과적합일 경우가 있으므로, 파라미터의 범위를 설정하여 최적의 모델을 선정하였다. Random Forest 알고리즘으로 선정된 최적의 모델의 경우 기본 알고리즘인 Decision Tree의 개수(n_estimators) 500, 트리의 최대 깊이(max_depth) 7인 경우 가장 높은 예측 정확도를 보였다. GBM 알고리즘으로 선정된 최적의 모델의 경우 기본 분류 알고리즘인 Decision Tree의 개수(n_estimators) 300, 트리의 최대 깊이(max_depth) 9, 학습률 0.1일 경우 가장 높은 예측 정확도를 보였다. XGBoost 알고리즘의 최적의 모델의 경우 기본 알고리즘의 개수(n_estimators) 400, 트리의 최대 깊이(max_depth) 4, 학습률 0.01일 경우 가장 성능이 우수했다.

Table 4는 적용한 각 머신러닝 알고리즘의 성능지표이다. 각 지표의 특성에 따라 알고리즘의 성능이 각각 다르게 나타나지만, 정밀도(precision)를 제외한 4가지 지표(specificity, recall, F1-score, accuracy)는 Random Forest의 결과가 가장 높게 나타났다.

5. 결론

본 연구에서는 단일 병원의 EMR과 CDM에서 신장암 환자들의 정형 및 비정형데이터를 추출하여 신장암 빅데이터를 구축하였다. 또한, 환자별 임상적 특징들에 따라서 다양한 진단 및 치료법이

존재하는 신장암에 대해, 열악한 국내 진료 환경에서 환자들의 진단 및 치료에 도움을 줄 수 있고, 수준 높은 진료 환경조성을 위하여, 머신러닝 알고리즘들을 적용하여 환자별 맞춤형 치료 결정을 위한 초기 임상결정지원 시스템을 개발하였다. 머신러닝 알고리즘 중 직관적으로 이해하기 쉽고 해석이 용이한 Decision Tree 알고리즘과 분류 성능이 뛰어난 앙상블 계열의 Random Forest 알고리즘, 그리고 최근 효율적으로 오류를 재분류하여 과적합을 줄이고 분류 작업의 성능을 높이는 GBM과 XGBoost 모델을 적용하였다.

향후 연구에서는 더 많은 임상적인 상황과 변수를 고려한 시스템 고도화가 필요하고, 이를 바탕으로 다기관 및 국제적 공동 연구를 통한 국제 표준 CDM 기반 신장암 표준임상빅데이터 구축을 통하여, 신장암 인공지능 알고리즘을 차별화 및 고도화할 전략을 가지고 있다. 본 연구에서 개발한 시스템을 통해 진료 현장에서 임상 의사 및 환자의 시간적·경제적 비용을 절감하고, 더하여 국가적 의료비용의 절감을 기대한다.

후 기

본 논문은 2021년도 양산부산대학교병원 임상연구비 지원을 받아 연구되었음.

참고문헌

- [1] Hyuna Sung, Jacques Ferlay and Rebecca L. Siegel, et al., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries”, *CA Cancer J Clin*, vol 71, no. 3, 209-249, 2021.
- [2] 국립암센터. <https://ncc.re.kr/cancerStatsView.ncc?bbsnum=598&searchKey=total&searchValue=&pageNum=1>
- [3] Holger Mocha, Antonio L. Cubillab, and Peter A. Humphrey et al., “The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs—Part A: Renal, Penile, and Testicular Tumours”, *Eur Urol*, vol 70, no. 1, 93-105, 2016.
- [4] Robert J Motzer, Eric Jonasch and Neeraj Agarwal et al., “Kidney Cancer, Version 3.2022, NCCN Clinical Practice Guidelines in Oncology”, *J Natl Compr Canc Netw*, vol 20, no. 1, 71-90, 2022.
- [5] Jiang, Xia, Alan Wells, Adam Brufsky, and Richard Neapolitan. “A clinical decision support system learned from data to personalize treatment recommendations towards preventing breast cancer metastasis.” *PloS one*, vol. 14, no. 3, 2019.
- [6] Sesen, M. Berkan, Ann E. Nicholson, Rene Banares-Alcantara, Timor Kadir, and Michael Brady. “Bayesian networks for clinical decision support in lung cancer care.” *PloS one*, vol. 8, no. 12, 2013.
- [7] Waghlikar, Kavishwar B., Kathy L. MacLaughlin, Michael R. Henry, Robert A. Greenes, Ronald A. Hankey, Hongfang Liu, and Rajeev Chaudhry. “Clinical decision support with automated text processing for cervical cancer screening.” *Journal of the American Medical Informatics Association* vol. 19, no. 5, 833-839, 2012.
- [8] Borim Ryu, Eunsil Yoon, Seok Kim et al., “Transformation of Pathology Reports Into the Common Data Model With Oncology Module: Use Case for Colon Cancer”, *J Med Internet Res*, vol. 22, 12, 1-13, 2020.
- [9] Seung In Seo, Chan Hyuk Park, Seng Chan

- You et al., "Association between proton pump inhibitor use and gastric cancer: a population-based cohort study using two different types of nationwide databases in Korea", *Gut*, vol. 0, 1-10, 2021.
- [10] Van Buuren, S. *Flexible Imputation of Missing Data*. Second Edition. Chapman & Hall/CRC. Boca Raton, FL., 2018.
- [11] Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinohara RT, Berger C, Ha SM, Rozycki M, Prastawa M. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*, 2018.
- [12] Song X, Waitman LR, Hu Y, Yu AS, Robins D, Liu M. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *Journal of the American Medical Informatics Association*, vol. 26, no. 3, 242-253, 2019.
- [13] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." *Shanghai archives of psychiatry*, vol. 27, no. 2, 130, 2015.
- [14] Breiman, Leo. "Random forests." *Machine learning*, vol. 45, no. 1, 5-32, 2001.
- [15] Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." *Frontiers in neurorobotics*, vol. 7, no. 21. 2013.
- [16] Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, and Kailong Chen. "Xgboost: extreme gradient boosting" *R package version 0.4-2* 1, no. 4, 1-4, 2015.

(접수: 2022.10.19. 수정: 2022.11.22. 게재확장: 2022.11.25.)