알고리즘, 기계 및 의학

제공 : **김수근** 성균관의대 강북삼성병원 직업환경의학과 교수

원문

On algorithms, machines, and medicine

Coiera E. On algorithms, machines, and medicine. Lancet Oncol. 2019 Feb;20(2) : 166-167. doi: 10.1016/S1470-2045(18)30835-0. Epub 2018 Dec 21.

※ Lancet Oncol. 2019 2월호에 실린 On algorithms, machines, and medicine을 소개합니다.

지도가 영토가 아닌 것처럼, 알고리즘이 결코 기존의 주어진 질병이나 문제를 치료 하는 것은 아니다. 알고리즘, 신경망, 가이드라인 및 프로토콜 – 이 모든 것이 항상 더 복잡하고 변화가 심한 현실의 측면을 모델링만 할 수 있다.[1] 우리가 알고리즘과 기계로 학습된 임상적 접근법이 지배하는 세계로 이동함에 따라, 우리는 기계가 말하는 것과 우리가 해야 하는 것 사이의 차이를 깊이 이해해야 한다.

점점 더 많은 수의 연구 논문들이, 기계 학습(Machine learning)을 이용하여 만들어진 컴퓨터 시스템의 인상적인 진단 성능을 보고하고 있다. 특히 딥러닝 기술(Deep learning techniques)은 이미징 데이터를 해석하는 능력을 변화시키고 있다.[2] 란셋 종양학 학술지(Lancet Oncology)에서 Xiangchun Li와 동료들[3]은, 초음파 영상을 이용해 '갑상선 암을 진단하기 위한 딥러닝과 통계적 방법을 적용한 후향적 임상전 연구(Retrospective preclinical study)'를 보고했다. 그 연구결과는 인상적이다. 처음으로 접하는 영상 데이터에 대한 영상의학 전문의 6명과 비교했을 때, 시스템은 내부 검증 데이터 집합(Internal validation dataset)에서 알고리즘으로 민감도 93.4%(95% CI 89.6~96.1) 대 영상의학 전문의들은 96.9%(95% CI 93.9~98.6)로 동일한 수의 암을 정확하게 감지했지만(p〈0.0001), 위양성은 알고리즘을 이용한 컴퓨터 시스템[특이도 86.1%(95% CI 81.1~90.2)]이 영상의학 전문의[특이도 59.4%(53.0~65.6)] 보다 훨씬 적었다(p〈0.0001).

이러한 결과는 얼마나 일반적일까? 한 보건 서비스 또는 지역[이 연구에서

중국 천진(Tianjin) 암 병원의 경우]의 환자만을 대상으로 하는 훈련은 훈련 데이터에 지나치게 적합할 위험이 있어 다른 환경에서는 취약한 성능 저하를 초래한다.[4]

이 연구에서는 다른 병원의 모집단에서 유사한 기계 특이성(Machine specificity)을 달성했지만, 지린(Jilin)의 외부 검증 데이터(External validation dataset)에서는 민감도가 84.3%(95% CI 73.6~91.9), 웨이하이(Weihai)의 외부 검증 데이터(External validation dataset)에 대해서는 84.7%(77.0~90.7)로 떨어졌다. 중국인이 아닌 대상에 대해서는 이 시스템의 성능이 저하될 것으로 예상할 수 있다. 한 가지 해결책은 새로운 표적 모집단의 환자들에게 시스템을 재교육하는 것이다. 그러나 훈련 데이터의 편견(Biases) 문제는 기본적인 것[5]이며, 임상의사는 항상 기계 권장 사항이 환자와 다른 모집단의 데이터에 기초하는지 고려해야 한다.

자궁경부검사(Cervical smear tests)의 자동화된 분석은 우리에게 컴퓨터화된 이미지 스크리닝이 가능하다는 것을 알려주었지만, 그것은 많은 기술적 그리고 비기술적인 문제들을 해결해야만 한다는 것을 알려주기도 하였다.[6] 예를 들어, 기계 학습은 의도한 작업을 배우는 것이 아닐 수도 있고, 임상작업흐름(Clinical workflow)이나 데이터 품질을 부주의하게 모델링할 수도 있다. 이러한 상황별 요인(Context-specific factors)을 다른 곳에서 사용할 때 복제되지 않고 성능이 저하될 수 있다. 예를 들어, Li와 동료의 연구에서 갑상선 암에 걸린 환자의 암이 없는 영상(Cancer-free images)은 훈련대상에서 제외되었다. 실제 환경에서는 그러한 영상이 포함되며, 그러한 이미지는 알고리즘 성능을 왜곡할 수 있다.

비록 이 연구가 전임상적(Preclinical)이기는 하지만, 저자들은 가능한 한 임상적으로 결과를 의미 있게 하기 위해 훌륭한 노력을 하고 있다. 이미지 증강(Image augmentation)은 훈련 데이터를 인위적으로 변경하는 데 사용되었다. 즉, 임의적으로 이미지를 자르고, 크기를 조정하고, 실제 이미지 품질의 변화를 모방하기 위해 이미지를 변경하는 것이다. 딥러닝 시스템은 종종 비판받는데, 그 이유는 그들의 권고는 설명이 없고, 논리(Logic)가 숨겨진 진단을 뒷받침하기 때문이다.[7] 이 연구에서는 진단에 가장 많이 기여한 영상의 픽셀(Pixels)이 강조되었다. 임상의사는 컴퓨터 해석을 확인하는 데 도움이 되는 이미지의 중요한 부분을 강조할 수 있다.

그러나 알고리즘의 진단 성과나 인간과의 비교에만 초점을 맞추면 환자의 진단 결과에 대해 거의 알 수 없고, 중요한 것보다 자동화하기 쉬운 것만을 지나치게 최적화할 수 있다.[8] 의사결정 지원(Decision support)은 임상작업 흐름에 포함되어야 하며 환자의 관리(Care)를 유도하는 활동과 결정과정이 포함된 웹의 한 부분에 불과하다. 갑상선암의 경우, 초음파검사는 생체검사(Biopsy)와 치료로 이어질 수 있는 순서의 한 단계다. 갑상선암이 과다 진단되고 과다하게 치료된다[9]는 우려에 비추어 볼 때, 개선된 초음파 탐지는 결과면에서 거의 이익(Benefit)을 제공하지 못할 수 있다. 예를 들어, 한국은 갑상선 암이 과다 진단으로 인해 15배나 증가한 것으로 밝혀졌다.[10] 그리고 결과 질환(Consequential disease)보다 나태한 암(Indolent Cancer)[1]을 감지하는 진단 방법이 이 상황을 더욱 악화시킬 수 있다. 확실하게, 초음파검사결과를 정확히 음성이라고 자동식별을 한다면, 임상의사로 하여금 더 이상 진료를 할 필요가 없다는 확신을 줄 수 있을 것이다. 이러한 이유로, 인간과 기계를 비교하는 것보다 기계로 도움을 받는 인간의 성과를 측정하는 것이 임상적으로 더 의미가 있다. 이러한 측정은 궁극적으로 임상 시험(Clinical trials)에서 이루어져야 하며, 위 음성 식별 및 과소 치료 및 과다 치료를 기록해야 한다. 갑상선 암에서 가장 시급한 의사결정 지원은 진단이 아니라 치료 결정을 내리는 경우이다.

따라서 알고리즘 성능의 우수성은 자동화를 추구하는 데 필수적이지만, 궁극적으로 우리는 자질구레하고 혼잡한 진료환경의 현실에서 자동화를 사용할 때 인간이 결정하는 것에 관심이 있다. 우리의 기계가 그 현실에 완전히 포위되어 우리보다 잘 보일 때까지 임상의로서의 우리의 역할은 기계와 결정 사이의 다리 역할을 해야 한다.

적어도 현재 알고리즘은 환자를 치료하지 않으며, 의료시스템은 환자를 치료한다.

주석

❶ 천천히 자라는 암의 유형. 너무 느리게 진행해서 수십 년 동안 다른 세포조직에 전혀 악영향을 미치지 않고 잠복하는 암을 말한다. 최근 갑상선 · 전립선 · 폐 · 유방암 등 암의 조기진단에서 많이 발견되는 종양들 가운데 이같이 게으른 종양으로 드러나는 것들이 많이 발견되고 있다. 결국, 인체에 해를 끼치지 않는 종양이 갈수록 더 많이 발견되면서 갈수록 첨단화하는 암 조기진단의 가치에 대한 의구심이 고개를 들고 있다.

참고문헌

Enrico Coiera : Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia enrico.coiera@mq.edu.au

1. Coiera E. Basic concepts in informatics: models. In: Coiera E, ed. Guide to health informatics. Boca Raton: CRC Press, 2015: 3–12.
2. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. Radiographics 2017; 37: 2113–31.
3. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2018; published online Dec 21. http://dx.doi.org/10.1016/S1470–2045(18)30762–9.
4. Chen JH, Asch SM. Machine learning and prediction in medicine: beyond the peak of inflated expectations. N Engl J Med 2017; 376: 2507–09.
5. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017; 318: 517–18.
6. Bengtsson E, Malm P. Screening for cervical cancer using automated analysis of PAP–smears. Comput Math Methods Med 2014; 2014: 842037.
7. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? Dec 28, 2017. https://arxiv.org/pdf/1712.09923.pdf (accessed Nov 27, 2018).
8. Coiera E. The fate of medicine in the time of AI. Lancet 2018; 392: 2331–32.
9. Ahn HS, Kim HJ, Welch HG. Korea's thyroid–cancer "epidemic": screening and overdiagnosis. N Engl J Med 2014; 371: 1765–67.
10. Furuya–Kanamori L, Bell KJL, Clark J, Glasziou P, Doi SAR. Prevalence of differentiated thyroid cancer in autopsy studies over six decades: a meta–analysis. J Clin Oncol 2016; 34: 3672–79.