의료 AI의 안착 과정

자동 심전도 판독(Automated ECG interpretation)의 발전 역사를 통해

글 | 정세영 syjung@snubh.org

의료 인공지능에 대한 대중들의 관심이 높아지고 있습니다. 의료 자원의 낭비를 막고 치료의 질을 높일 수 있는 방법에 대해 인공지능기술을 통해 찾는 연구자도 급속하게 늘어나고 있습니다. 그렇다면 그동안 의료 분야에서는 인공지능기술이 전혀 쓰이지 않았을까요? 역사적인 실례를 통해 인공지능 의료기기 같은 신의료기술이 태동하고 시장에 안착되기까지 어떤 과정을 거치는지 살펴봄으로써 의료 인공지능 연구자들이 놓치지 말아야 할 교훈에 대해 이야기 나누고 싶습니다.

응급실 스냅샷

2007년 11월 21일 오전 4시. 서울. 제법 매서워진 찬바람은 겨울이 오고 있음을 알린다. OO대학교 병원 응급실 인턴 OOO 선생은 초점이 나간 눈으로 컴퓨터 화면을 응시하고 있다.

'이제 이 환자만 보면 조금 쉴 수 있겠구나'

태풍이 휘몰아치듯 환자들이 몰려 들다가 새벽이 되자 잠잠해졌다. 대기 환자 명단에 남은 한 명의 환자는 평온해진 응급실을 대변하고 있다.

67세, 여성.

"저녁 식사를 한 후 계속 속이 답답하네요"

별다른 과거 병력이 없던 환자는 몸이 불편한 듯 표정이 좋지 않았다. 그리고 본인의 증상을 조금씩 이야기하기 시작하였다.

'CC> Abdominal discomfort, onset: 8 hours ago, 금일 저녁 식사를 한 후 prev. healthy…'

환자 보호자는 대수롭지 않은 증상으로 응급실을 따라온 것이 못마땅한 듯 보였다. OOO 선생은 쏟아지는 졸음을 참으며 자판을 두드렸다. 그리고 채근하던 보호자의 말을 애써 무시하면서 신체 진찰을 하였다. 급성 복증(acute abdomen)[*1]에 합당한 징후는 없었다.

'아.. 이제 검사 결과만 확인하면 끝이다! 좀 쉬자…'

혈액검사, 복부 엑스선 검사, 심전도 검사 오더를 내렸다. 얼마 후 검사 결과가 나왔다

'검사 결과 별 이상 없겠지?…. 잘 설명해서 보내야겠다…'

졸린 눈을 비비면서 검사 결과를 확인한 OOO 선생은 심전도 결과지를 보며 정신이 번쩍 들 수밖에 없었다.

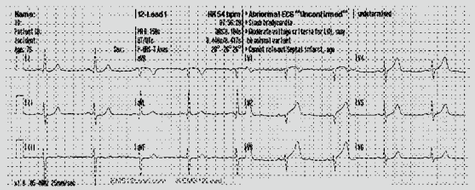
Alert! ACUTE MYOCARDIAL INFARCTION[*2]

막 나온 혈액 검사 결과를 자세히 보니 심근 관련 효소 수치도 올라가 있었다. 그는 재빠르게 담당 백업 레지던트를 호출했고, 심혈관조영실 당직의에게도 전화를 했다. 그렇게 그 환자는 응급실을 떠나 심혈관조영실로 올라가게 되었다.

심장은 우심방 꼭대기에 있는 동결절에서 생성된 전기 자극에 의해 분당 60~100회의 속도로 규칙적으로 박동한다. 심전도(心電圖, electrocardiogram, ECG)는 이런 심장의 전기 활동을 측정하는 방법으로 피부에 부착된 전극과 신체 외부의 장비로 기록된다. 심장질환이 있으면 전도 조직이 손상될 수 있으며 이는 심전도의 변화로 나타날 수 있다.[*3]

현재 병원에서 사용되는 심전도 기기는 판독 결과를 수초 내로 제공한다. 이를 '자동 심전도 판독(automated ECG interpretation)'이라고 한다.

[그림 1] 자동 심전도 판독 결과의 예시[*4]



Normal ECG "Unconfirmed"

"Unconfirmed"는 최종 진단을 내리기 전의 상태란 의미로, 의사가 환자의 임상 정보를 통합하여 판독하는 'overread'라는 과정을 거쳐야 최종 결과가 도출된다.

의대생들은 심장의 전기생리학부터 심전도의 원리와 판독방법까지 배운다. 병원 실습을 시작하면 심전도가 의료 현장에서 실제로 어떻게 활용되는지 경험하게 된다. 의사로서 처음 발을 내딛는 인턴 초반에는 심전도 기기의 각 유도(lead)를 몸의 어느 부위에 어떻게 붙여야 노이즈 없이 정확한 검사를 얻을 수 있는지 수없이 반복 학습하게 되지만, 나중에는 자동 심전도 판독 결과를 각각의 임상 상황에서 어떤 의미로 해석해야 하는지 체득하게 된다.

예컨대 자동 심전도 판독 결과 정상심방리듬(normal sinus rhythm)이 나오면 대부분 상황에서는 안심할 수 있다. 하지만 흉통을 지속적으로 호소하는 환자의 경우 ST분절 비상승 심근경색증(Non-ST elevation myocardial infarction, NSTEMI)을 배제할 수 없기 때문에 심전도 결과 이상이 없더라도 안심할 수 없다. 심전도는 대부분 상황에서 음성예측도[*5]가 높다.[*6] 즉, 심전도 결과가 정상으로 나오면 대부분 실제로도 정상이고 안심해도 된다는 의미이다. 양성예측도[*7]는 질환에 따라 달라진다. 예컨대 심전도에서 급성심근경색(acute myocardial infarction)이나 방실전도차단(atrio-ventricular conduction block)[*8]에 대해서는 양성예측도가 낮으므로 심전도 판독에 해당 질환이 있더라도 실제 그 질환이 있는지 임상의는 꼼꼼하게 점검해야 한다.

몇몇 상황에서 한계가 있기는 하지만, 대체로 자동 심전도 판독은 매우 고마운 기술이다. 특히 초보 의사에게는 더욱 그렇다. 그리고 임상 현장에서 널리 사용된다. 심전도를 찍으면 자동판독은 '당연히' 나오는 결과다. 또한, 의사로서 경험이 쌓일수록 그 결과가 어느 상황에서 어떤 의미를 갖는지 정확하게 알게 되고 적절히 활용하게 된다.

그렇다면 자동 심전도 판독은 어떤 기술일까? 위키피디아를 살펴 보도록 하자.

자동 심전도 판독은 인공지능과 패턴인식 소프트웨어, 지식 베이스를 활용한 기술로 환자의 심전도 결과를 분석하여 컴퓨터 기반 진단 및 해석을 자동으로 제공한다.[*9]

즉, 자동 심전도 판독은 인공지능 기술이다. 조금 더 살펴보자.

자동 심전도 판독 과정*10

1. 각 ECG 채널에서 얻은 아날로그 신호를 전용 소프트웨어와 DSP(digital signal processing) 칩을 이용해 디지털 신호로 변환한다
2. 변환된 디지털 신호를 일련의 알고리즘을 이용해 처리한다. 예컨대 노이즈 제거, 기저치 변동 제거 등이다.
3. 특징 추출(feature extraction): 수학적 방법을 사용하여 해석 및 진단에 중요한 분석을 시행한다. 분석결과는 AI 기반 알고리즘에 입력된다. 패턴분석 프로그램에 정보가 입력되기 전에 푸리에(Fourier) 분석이나 파형 분석같은 과정이 선행된다.
4. 규칙기반 전문가 시스템(rule-based expert system), 베이지안 확률 분석(probabilistic Bayesian analysis), 퍼지 논리 알고리즘(fuzzy logic algorithms), 클러스터 분석(cluster analysis), 인공 신경망(artificial neural networks), 유전 알고리즘(genetic algorithm) 등을 사용해서 분석을 한다.
5. 결과를 보고한다.

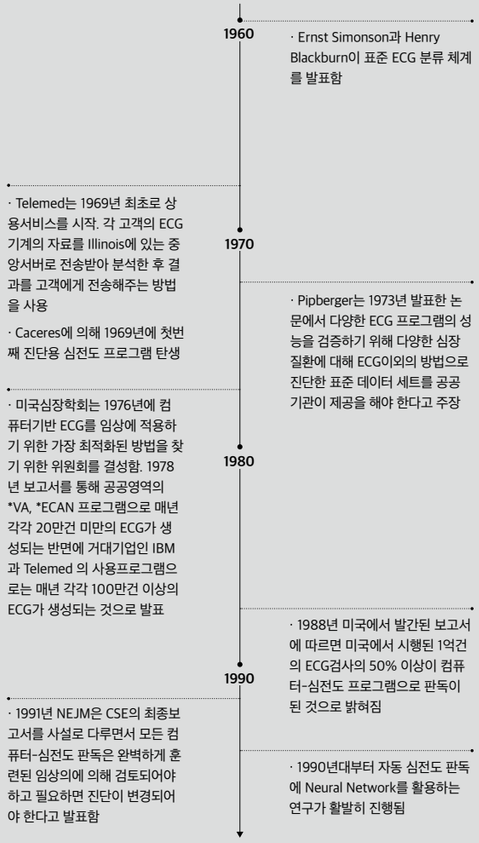
심전도를 찍으면 위와 같은 과정이 수초 이내에 진행되어 바로 결과가 출력된다.

자동 심전도 판독은 언제부터 개발되었고 사용되기 시작하였을까? 의사들조차도 대부분 심전도 기기가 인공지능 기술인지 알지 못한다. 의료 현장에 이처럼 완벽히 이식(implementation)된 인공지능 의료기기가 있을까? 그리고 인공지능 기술인 자동 심전도 판독 기기의 발전 과정을 보면 인공지능 기반 신의료기기가 의료 현장에 정착되기까지 어떤 과정이 필요한지 힌트를 얻을 수 있지 않을까?

본 글에서는 먼저 자동 심전도 판독 기술의 발전 역사를 찬찬히 짚어 보면서 새로운 의료기기가 의료 현장에 정착되기까지 어떤 과정을 거치게 되는지 알아보겠다. 다음으로 왜 그런 과정을 거칠 수 밖에 없었는지에 대해 의학 연구와 임상 의사 결정의 특징을 고려하여 살펴보도록 하겠다.

자동 심전도 판독의 발전 역사

[그림 2] 자동 심전도 판독(Automated ECG Interpretation)의 발전 역사*11



1) 1960년대 : 자동 심전도 판독 기술의 태동 시기

- 에른스트 시몬슨(Ernst Simonson)와 헨리 블랙번(Henry Blackburn)이 1960년 표준 심전도 분류 체계를 발표하였다.
- 허버트 핍버거(Hubert Pipberger)는 재향군인회 프로젝트를 통해 8개의 재향군인 병원에서 ECG-independent Frank data를 수집하여 VA 프로그램을 개발하였다. 미국 국립보건원(National Institutes of Health)에 있었던 세사르 카세레스(Cesar Caceres)는 표준 12-유도 심전도를 기반으로 하는 심전도 판독 프로그램인 ECAN을 개발하였다.
- 초기에는 ECAN 프로그램 도입률이 저조했는데, 단일 채널(single channel) 기록으로 나온 12-유도 심전도 결과를 해석하는 것이 효율적이지 못했기 때문이다.
- 핍버거(Pipberger)의 VA 프로그램도 사용률이 낮았다. 베이즈 방법 기반의 진단 분류 기준이 여러 진단을 조합하는 것을 허용하지 않고 상호 배타적인 진단만 허용했기 때문이다. 실제로 임상에서 심장전문의는 심전도 결과에 환자의 임상정보를 결합하여 진단의 정확도를 높이는 확률론적 방법을 사용한다.

2) 1970년대 이후 : 기술 발전 및 고도화 시기

- 메이오 클리닉(Mayo clinic)의 랄프 스미스(Ralph Smith)와 IBM은 합작 벤처회사를 설립하여 표준 12-유도 ECG와 변형 Frank 유도 방법을 통합하는 연구를 진행하였다.
- Mayo-IBM 합작사의 프로그램은 IBM 360 컴퓨터에서 구동되었고 후에 my ECG research center에 의해 IBM 1800 컴퓨터에 이식되었다. 향후에는 IBM 이외의 컴퓨터에도 사용 가능하게 되었다.
- 레이 보너(Ray Bonner)는 다기관협력프로젝트를 통해 IBM 5880을 사용하여 카트 장착용 ECG프로그램을 개발하였는데 이는 심전도 해석 결과를 저장하고 출력하는 것이 최초로 가능해진 획기적인 장치였다. 그 장치에 설치된 컴퓨터 기반의 ECG 프로그램은 텔레메드(Telemed), 마켓(Marquette), HP 필립스(HP-Philips), 모르타라(Mortara)가 개발한 제품이었다.
- 레이 보너는 자동 심전도 판독이 실용적이고 정확할 수 있다는 점을 최초로 증명하였다.
- 여러 연구자들의 노력으로 자동 심전도 판독 프로그램은 매우 빠른 속도로 발전했지만, 임상에서의 활용은 더딘 속도로 진행되었다.
- 미국심장학회는 1976년에 자동 심전도 판독 기술을 임상에 가장 효율적으로 활용하기 위한 위원회를 결성하였다. 또한 1978년 보고서에서 공공영역에서는 VA(Veternas Affairs), ECAN(Esophageal Cancer Action Network) 프로그램으로 매년 각각 20만 건 미만의 ECG가 생성되는 반면에 거대 기업인 IBM과 텔레메드의 사용 프로그램으로는 매년 각각 100만 건 이상의 ECG가 생성된다는 조사 결과를 발표하였다.
- 텔레메드는 1969년 최초로 상용 서비스를 시작하였는데 각 고객의 심전도 기계의 자료를 일리노이에 있는 중앙 서버로 전송받아 분석한 후 결과를 고객에게 전송해주는 방법을 사용하였다.
- 1978년까지 텔레메드는 1600명 이상의 고객을 확보하였고 매달 12만 5천여 건 이상의 심전도를 분석하게 되었다
- 1960~70년대에는 수많은 심전도 판독 프로그램이 개발되었다. Pipberger는 1973년 발표한 논문에서 공공기관이 표준 데이터 세트를 제공해야 한다고 주장하였다. 이는 심전도 프로그램의 성능을 평가할 때 쓰일 수 있는데, 심전도 이외의 방법으로 진단된 심장질환에 대한 정보가 포함되어야 한다.
- 데이비드 모르타라(David Mortara)는 1970~80년대에 컴퓨터 기반 자동 심전도 판독 기술의 발전에 지대한 영향을 미쳤다.
- 심전도 판독 기술을 검증하기 위한 표준 데이터셋이 만들어졌다. 유럽 위원회(European commission, EC)는 '정량적 심전도 분석을 위한 공통 표준(common standards for quantitative electrocardiography, CSE)' 데이터셋을 만들었다. 조스 윌렘스(Jos Willems)는 CSE 프로젝트를 통해 9개의 ECG 판독 프로그램과 8명의 심장 전문의의 판독을 비교 분석하였고 심장 전문의 대표 8인의 판독은 심전도 프로그램의 평균보다 5.8% 더 정확한 것으로 드러났다. (75.5% vs 69.7%, p<0.001)
- 최종 CSE 보고서에는 컴퓨터 기반 심전도 프로그램이 임상 의사가 지속적이고 일관된 진단을 내리는데 도움이 된다는 내용이 담겼다.

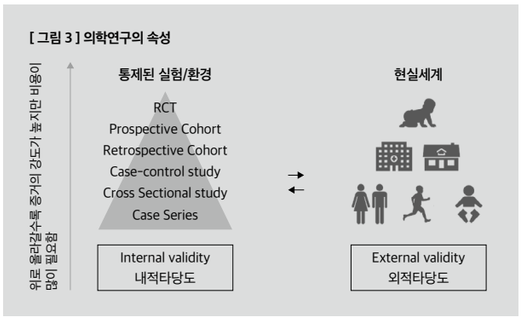
3) 1980년대 이후 : 인공지능 기술 도입 시기

- 1988년 미국에서 발간된 보고서에 따르면 미국에서 시행된 1억 건의 심전도 검사 결과의 50% 이상이 자동 심전도 판독 프로그램에 의해 보고되었다.
- 1991년 뉴잉글랜드 의학저널(The New England Journal of Medicine)은 CSE의 최종 보고서를 종설로 다루면서 모든 컴퓨터 기반 심전도 판독은 완벽하게 훈련된 임상 의사에 의해 검토되어야 하고 필요하면 진단이 변경(overread)되어야 한다고 발표하였다.
- 컴퓨터 기반 심전도는 serial ECG comparison과 같은 분야의 발전을 가져왔다.
- 1990년대부터 자동 심전도 판독에 신경망(neural net)과 같은 인공지능 기술을 도입하는 연구가 활발히 진행되기 시작하였다.

위에서 살펴본 바와 같이 자동 심전도 판독은 지난한 발전의 역사를 갖고 있다. 왜 그래야만 했을까? 의료 신기술이 임상에 널리 활용되기 위해서는 어떤 과정들을 거쳐야만 할까? 이를 정확하게 이해하기 위해서는 의학 연구와 임상 의사 결정 과정의 속성을 올바로 이해할 필요가 있다.

의학연구의 속성

[그림 3] 의학연구의 속성



의학은 기본적으로 사람을 치료하기 위한 학문이다. 따라서 증거를 창출하여 임상에 적용하기까지의 과정이 신중할 수밖에 없다. 아스피린이라는 약물을 개발하여 효과가 있는지 검증하는 연구를 생각해보자. 개발된 약을 아무런 검증 없이 환자에게 바로 적용한다고 가정해보자.

열나는 환자에게 아스피린을 사용했다. 환자는 열이 떨어졌다. 그렇다면 그게 과연 아스피린의 효과라고 할 수 있을까? 우리는 경험적으로 열이 날 때 물을 충분히 마시고 옷을 벗어도 체온이 떨어질 수 있다는 점을 알고 있다. 이렇듯 아스피린을 먹은 환자가 열이 떨어진 것이 아스피린의 순수한 효과인지 알기 위해 아스피린을 사용한 경험적인 사실을 수집하는 것은 강력한 증거가 될 수 없다. 교란 요인[*12]으로 인해 열이 떨어졌을 수도 있기 때문이다. 아스피린의 효과에 대한 정확한 증거를 얻기 위해서는 통제된 환경에서 임상시험을 해야 한다. 아스피린과 똑같이 생긴 가짜 약을 만들고 무작위로 나눈 실험군(아스피린을 먹는 그룹)과 대조군(가짜 약을 먹는 그룹)에 각각 아스피린과 가짜 약을 준 후 그 효과를 비교하는 것이다.

실험군과 대조군을 무작위로 나누면 각 그룹에 특정 성별, 나이를 갖는 환자가 몰리거나 특정한 임상적 특징을 공유하는 환자가 몰릴 가능성이 떨어지게 된다. 즉, 아스피린 복용 여부를 제외한 다른 특징은 두 군에서 동일하게 되기 때문에 위 실험의 결과는 순수하게 아스피린 복용 유무에 따른 차이라고 볼 수 있다. 또한 약을 먹는 사람이 실험군인지 대조군인지, 주는 약이 아스피린인지 가짜 약인지 연구자도 모르게 하면 실험이 더욱 완벽해진다. 이를 이중 맹검 무작위 대조군 연구(double blind randomized controlled study, double-blind RCT)라고 하며 의학에서 가장 강력한 연구 방법이다.

연구 결과로 아스피린의 해열 효과가 입증되었다고 하자. 연구자들이나 의사들은 그 결과를 어떻게 신뢰할 수 있을까? 이를 위해서는 연구결과로 발표한 논문의 방법론을 검토해야 한다. '교란 요인을 제대로 통제했는가?', '강력한 증거 능력을 갖는 연구 설계를 했는가?', '피험자는 충분한가?'와 같은 질문을 하게 된다. 연구 방법론은 증거의 능력이 가장 약한 증례 보고부터 가장 강력한 이중 맹검 무작위 대조군 연구까지 다양하다. 이렇게 통제된 환경에서 약물이나 치료의 효과가 입증되면 '내적 타당도(internal validity)'가 입증이 되었다고 말할 수 있다.

이제 아스피린이라는 신약이 출시되었다. 제약회사는 '이중 맹검 무작위 대조군 연구'에서 효능을 입증을 받은 획기적인 약물이라고 선전한다. 임상 의사들은 환자의 병마와 싸우는데 사용할 새로운 무기가 생겼다고 기뻐한다. 아스피린은 불티나는 속도로 팔리기 시작한다. 대부분의 환자에게서는 효과가 있었다. 하지만 사람마다 효과의 차이는 조금씩 달랐으며, 임상 시험에서는 보이지 않았던 부작용들이 조금씩 보고되기도 한다. 하지만 아스피린이라는 약을 지속적으로 사용하는 데는 문제 없는 것들이다. 아스피린을 생산하는 제약회사의 주가는 계속 오르게 된다.

이와 같이 현실의 약물 사용 경험은 반드시 필요하다. '외적 타당도(external validity)'를 입증하기 위해서다. 내적 타당도는 흡사 실험실 환경처럼 철저하게 통제된 조건에서 나온 결과라서, 새로 개발된 약물은 다양한 특징을 갖는 환자들이 산재해 있는 현실에서도 효과가 있는지 확인해야 한다. 실제로 내적 타당도는 입증했지만, 출시 후 예기치 않는 부작용이 나타나서 퇴출당하는 약물들도 있다.

다른 예를 살펴보자. 이번에는 현실 세계 사용 경험으로 의심을 품고 실험 연구로 증명하는 예다. 폐경 후 호르몬 대체 요법(postmenopausal hormonal replacement therapy, HRT)은 부족한 여성 호르몬을 보충해서 삶의 질을 떨어뜨릴 수 있는 폐경 증후군을 효과적으로 치료하는 방법이다. 다양한 HRT 약물이 출시된 후 연구자들은 약물 사용 경험을 수집하였다(후향적 연구 방법). 그리고 HRT가 심혈관 질환 발병을 낮춘다는 잠정적 결론을 내렸다. 얼마나 획기적인가. 폐경 증후군을 치료하면 일석이조로 심혈관 질환 발병도 줄어든다! 그런데 실제로도 그럴까? 제약회사들은 많은 기대를 갖고 이중 맹검 무작위 대조군 연구를 시행했다. 실험 결과는 에스트로겐과 프로게스테론 복합 제제를 사용하면 심혈관 질환 발병이 오히려 증가할 수도 있는 것으로 드러났다. 하지만 그 정도는 미약했다. 따라서 HRT가 심혈관 질환 발병을 낮추지는 않았으나 실제로 약물을 처방하는 데는 문제가 없었다.

이중 맹검 무작위 대조군 연구는 임상시험 환경을 완벽히 통제하는 '실험 연구(experimental study)'라고 한다. 이에 반해 이미 수집된 빅데이터를 후향적으로(retrospective) 살펴봐서 증거를 찾는 연구를 '관찰 연구(observational study)'라고 한다. 아래 표는 관찰 연구에서 발견된 사실이 이후에 실험 연구를 했을 때는 입증이 되지 않은 예시다.

[표 1] 관찰 연구와 실험 연구의 결과가 일치하지 않은 예[*13]

	Outcome	관찰 연구 결과	실험 연구 결과
폐경 후 호르몬 대체요법	심혈관질환 발생	감소	증가
비타민 E 고용량요법	심혈관질환 발생	감소	변화 없음
저지방 식이	심혈관질환, 암 발생	감소	변화 없음
칼슘보충요법	골절, 암 발생	감소	변화 없음
호모시스테인 감량을 위한 비타민 요법	심혈관질환 발생	감소	변화 없음

인공지능 의학연구의 속성

인공지능 기반 의학 연구에는 빅데이터가 필수적이다. 빅데이터 연구는 기본적으로 관찰 연구이다. 빅데이터 자체가 '이미 수집된 데이터'를 의미하기 때문이다. 바꿔 얘기하면 인공지능 기반 의학 연구는 위와 같이 관찰 연구가 갖는 한계를 고스란히 공유할 수밖에 없다. 따라서 빅데이터 연구에서 교란 요인을 적절히 통제하지 않으면 잘못된 결론을 내릴 수도 있다.

이제 인공지능 의학연구 얘기를 해보자. 다음은 가장 기본적인 질문이다.

(1) 인공지능 알고리즘은 내적 타당도를 어떻게 검증할까?
(2) 내적 타당도가 제대로 검증되지 않은 모델을 현실 세계에 바로 적용하는 것이 가능할까? 그렇게 하는 게 윤리적일 수 있을까?
(3) 내적 타당도가 완벽히 검증되었다고 주장하는 인공지능 모델이 있다고 하자. 그렇다면 외적 타당도는 어떻게 검증해야 할까?

위 질문에 대한 합의된 기준은 아직 없다. 그래서 내적 타당도가 입증되지도 않은 인공지능 모델이 버젓이 '인공지능 모델, 의료의 혁명'이라는 과장 광고를 달고 출시되기도 한다. 예를 들어 [그림 4]를 보면 위는 1990년대 초의 광고, 아래는 2017년의 광고이다. 30여 년이 흘렀지만 달라진 점이 없다. 그렇기 때문에 의사가 문지기(gate-keeper)로서의 역할을 제대로 수행하지 못하는 순간 잘못된 인공지능 모델로 피해를 보는 환자도 나오게 될 것이다.

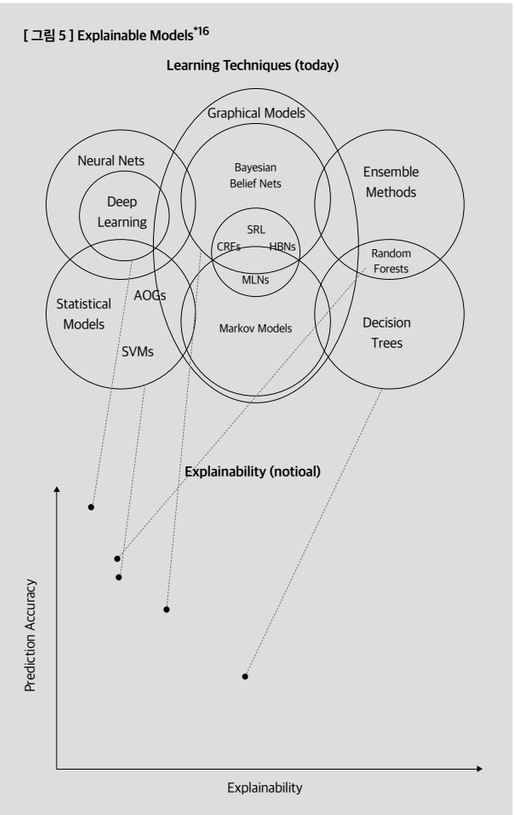
[그림 4] 인공지능 과장 광고 예시[*14 *15]

위 그림은 1990년대 초의 광고, 아래 그림은 2017년 현재의 광고이다. 30여년이 흘렀지만 달라진 점이 없다.

인공지능 모델의 '내적 타당도'에 대한 이야기를 해보자. 알파고가 출현한 후 '딥러닝'이 뜨고 있다. 딥러닝을 사용하면 내적 타당도를 높일 수 있는가? 다르게 질문을 해보자. 인공지능 알고리즘은 의학연구의 증거 강도 피라미드(pyramid of evidence-based medicine)처럼 '증거의 강도'를 기준으로 일렬로 세울 수 있는가? 만약 그렇다면 '딥러닝'은 증거 강도의 피라미드에서 가장 위에 위치할 수 있는가? 인공지능 연구에서 내적 타당도의 기준은 무엇일까? 보통 최종 결과로 제시되는 알고리즘의 성능 지표만으로는 내적 타당도를 평가하는 것에 한계가 있지 않을까?

아래 그림은 인공지능 알고리즘별로 정확성(accuracy)과 설명력(explainability)이 어떤 관계를 갖는지 보여 준다. 두 지표는 교환 관계(trade-off)를 갖는다. 설명력이 높은 모델은 정확성이 떨어지는 경향이 있고, 반대로 설명력이 떨어지면 정확성은 올라가는 경향이 있다.

[그림 5] Explainable Models[*16]



인공지능 알고리즘은 증거의 강도에 따라 일률적으로 평가할 수 없다. 연구 목적에 따라 가장 적합한 학습 방법을 선택한다. 하지만, 인공지능 연구 논문이 보통 '강력한 근거'로 보여주는 연구 결과는 알고리즘의 성능지표다. 성능지표가 높게 나오면 내적 타당도가

확보된 것일까? 학습 데이터가 잘못되었다면? 검증을 제대로 못했다면? 학습 데이터에 과적합(overfitting)* 된 모델이라면?

*** 과적합(overfitting)**
너무 정교해서 융통성 없는 모델을 도출한 경우를 말한다. 예컨대, 신발을 만드는 상황을 가정해보자. 신발공은 10명의 발에 대한 자료를 갖고 있다. 그 10명은 공교롭게 모두 발가락이 6개다. 신발공이 만약 그 10명의 발에 꼭 들어맞게(perfectly fitted) 발가락 모양까지 본뜬 무좀 양말 모양의 신발을 만든다면 발가락이 5개인 일반적인 사람들은 그 신발을 신지 못할 것이다. 따라서 보통 기성화를 만들 때는 발 볼의 너비와 발의 길이만을 반영한 융통성 있는 모델을 만들게 된다. 인공지능 모델을 개발할 때도 과적합을 하게 되면 성능지표는 월등하게 높아질 수 있지만 학습데이터 이외의 자료에는 적용할 수 없는 쓸모 없는 모형이 도출될 수 있다.

앞서 의학 연구의 속성에서 언급했듯이 의학 연구에서는 연구 결과의 내적 타당도를 검증할 수 있는 기준들이 잘 알려져 있다. 그 기준들의 본질은 객관적 증거로서 '설명'이 가능하냐는 것이다. 인공지능 의학 연구는 이런 관점에서 내적 타당도를 검증할 수 있는 장치가 부족하며, 더욱이 설명력이 떨어지는 인공지능 모델을 사용하면 내적 타당도 검증은 더욱 요원한 일이 될 수 있다.

따라서 이 상황에서는 외적 타당도를 검증하는 작업이 더욱 중요해지게 된다. 설명력이 떨어지는 인공지능 모델이더라도 실제 의료 현장에서 어떤 효과를 갖고 어떤 프로세스로 활용될 수 있는지 충분히 검증된다면 유용하게 활용될 수 있을 것이다. (전통적인 의학연구에 내적 타당도가 매우 높은 무작위 대조군 연구(randomized controlled trial, RCT)으로 효과가 입증된 약물은 현실세계에서 대부분 효과가 있다. 그러나 인공지능 알고리즘은 그렇지 않을 수 있다는 뜻이다. 성능이 높게 나온 알고리즘 일지라도 내적 타당도를 전통적인 의학연구법만큼 담보할 수 없기 때문에 현실 세계에서의 검증을 더 치열하게 해야만 한다.)

다시 자동 심전도 판독 예제로 돌아가자

자동 심전도 판독은 인공지능 기술이다. 그 발전 역사를 찬찬히 살펴보면 '내적 타당도'를 높이는 과정과 '외적 타당도'를 높이는 과정이 유기적으로 병행해서 이루어진 사실을 알 수 있다. 구체적으로 어떤 교훈을 얻을 수 있는지 정리해보자.

(1) 심전도 발전에는 전기 심장학이라는 기초 학문의 태동과 발전이 중요했다. - 인공지능 관련 기초 학문이 중요하다.

(2) 핍버그(Pipberger)나 카세레스(Caceres) 같은 연구자들의 자생적인 노력이 중요했다. - 연구자, 개발자의 창의적인 노력이 중요하다.

(3) ECAN과 VA프로그램은 초기사용률이 저조했다. - 두 프로그램 모두 초반에는 임상 현장에서 효율성을 입증하지 못했다. 외적 타당도를 입증하지 못한 모델은 도태될 수밖에 없다.

(4) Mayo-IBM 합작사의 역할이 중요했다. - 적절한 시점에는 시장을 주도하는 기업과 의료기관의 협력이 중요하다.

(5) 시장의 필요에 의해 텔레메드(Telemed)에서 1969년 ECG 상용 서비스를 최초로 시작한 것과는 별개로 임상 의사가 그 서비스의 유용성을 인지하고 받아들이는 과정은 매우 더뎠다. - 기술 개발 초기부터 신기술의 임상적 유용성과 활용가능성을 타진하는 협력 작업이 필요하다.

(6) Telehealth 기반의 원격 판독 서비스는 이미 있었다. - 인공지능기반 서비스 모델은 의료서비스 수준이 열악한 원격지에서 수행한 검사를 일차적으로 선별하는 데 유용하게 활용될 수 있다.

(7) 정량적 심전도 분석을 위한 공통 표준처럼 개발된 신기술의 내적 타당도를 검증하기 위한 도구가 필요하다. - 인공지능 연구자들이 마음 놓고 쓸 수 있는 표준 데이터셋이 있어야 한다.

(8) 의료인의 역할이 여전히 중요하다. - 도메인 지식[17]을 제공하는 것 뿐만 아니라 문지기로서 신기술의 유용성을 평가하는 중요한 역할을 담당해야 한다. 골목 대장으로 이득을 지키기 위한 제스처를 취하기보다는 철저한 전문가주의를 기반으로 선의의 피해자가 생기는 것을 방지하면서 기술과 의료의 공생적 발전을 도모하는 문지기가 되어야 한다. 미국의 정밀의료 프로젝트(All of US)의 라이프 로그 수집을 담당하는 스크립스(Scripps) 연구소의 수장인 에릭 토폴(Eric Topol)은 미국의사협회지(journal of the american medical association, JAMA)에 실린 종설에서 인공지능 기술이 영상의학, 병리학 분야에 적극적으로 도입이 되면, 단순 반복적인 작업은 인공지능 기술에 이양되고 영상의학과, 병리과 의사는 정보 전문가(information specialist)로서 복잡한 임상 케이스를 관리하게 될 것이라 전망하였다. 바로 그 정보 전문가가 인공지능 시대에 의사가 수행해야 하는 문지기 역할의 일례다.

(9) 사용 증례의 지속적 축적이 중요하다. - 인공지능기술 사용 증례(use cases)를 지속적으로 축적하여 인공지능이 활용되는 구체적인 프로세스에 대한 정립이 필요하다.

결국 제일 중요한 것은 인공지능 연구자가 마음껏 사용할 수 있는 표준 빅데이터셋을 확보하는 것과 개발된 인공지능 기술을 지속적으로 검증하고 발전시킬 수 있는 플랫폼을 만드는 것이다.

다음 글에서는 의학 연구에서 '내적 타당도'를 평가할 때 왜 설명력이 중요한 지를 임상 의사 결정 과정의 특징을 바탕으로 살펴보도록 하겠다. 그리고 상기 심전도의 예제처럼 신의료기술이 내적 타당도를 지속적으로 높이면서 외적 타당도를 확보하려면 어떤 연구 환경이 필요한지에 대해서도 알아보도록 하겠다.

***1** 참고 | 급성 복증: 복강 내 장기의 염증, 천공, 폐색, 경색, 파열에 의한 복통을 수반하는, 갑자기 발병하는 복부의 상태로 보통 긴급한 외과적 처치를 필요로 하는 질환을 말함 ***2** 참고 | 심근경색증 ***3** 참고 | 위키피디아 'Electrocardiogram', https://ko.wikipedia.org/wiki/%EC%8B%AC%EC%A0%84%EB%8F%84 ***4** 참고 | http://www.ems12lead.com/2009/02/20/76-yom-cc-shortness-of-breath-while-walking/ ***5** 참고 | 음성예측도: 검사가 음성(정상)으로 나왔을 때 실제 병이 없을 확률 ***6** 참고 | 논문 | Guglin, M. E., & Thatai, D. (2006). Common errors in computer electrocardiogram interpretation. Int J Cardiol, 106(2), 232-237. http://doi.org/10.1016/j.ijcard.2005.02.007 ***7** 참고 | 양성예측도: 검사가 양성(비정상)으로 나왔을 때 실제 병이 있을 확률 ***8** 참고 | 방실전도차단 : 심장은 2심방 2심실로 구성된다. 동결절에서 생성된 전기자극은 심방에서 심실로 전도되어야 하는데 병적인 불응 상태나 절단에 의해 심방 자극이 심실로 전도되는 것이 지연되거나 차단되는 것을 방실전도차단이라고 한다. ***9** 참고 | 위키피디아 'Automated ECG interpretation' ***10** 참고 | https://en.wikipedia.org/wiki/Automated_ECG_interpretation ***11** 논문 | Pentti el al.(2016). Eyewitness to history: landmarks in the development of computerized electrocardiography, Journal of Electrocardiology 49 (2016) 1 - 6 ***12** 참고 | 교란 요인: 보고자 하는 실험결과에 영향을 미치는 요인이다. 아스피린 예시에서는 '물 많이 먹기' '옷 벗어 냉각하기' 등이 해당된다. ***13** 참고 | 정세영 역. (2017). 직관으로 이해하는 의학 통계학. 대한의학서적 ***14** 참고 | https://www.kyuh.ac.kr/main.asp?ver=pc ***15** 참고 | http://v.media.daum.net/v/20171225090028486 ***16** 참고 | DARPA. (2016). Explainable Artificial Intelligence.(pp.9) ***17** 참고 | 도메인 지식이란 의학 교과서에 나온 객관화할 수 있는 내용뿐만 아니라 임상현장에서 체득한 다양한 경험을 포함한다.