Hanyang Medical Reviews

Hanyang Med Rev 2017;37:49-60
https://doi.org/10.7599/hmr.2017.37.2.49
pISSN 1738-429X eISSN 2234-4446

Review

IBM 왓슨 포 온콜로지의 의학적 검증에 관한 고찰

Yoon Sup Choi[1, 2, 3]

[1]Digital Healthcare Institute, Seoul, Korea
[2]Digital Healthcare Partners, Seoul, Korea
[3]Department of Digital Health, Sungkyunkwan University, Seoul, Korea

The development of artificial intelligence revolutionizes many fields, and one of the representative fields is medical artificial intelligence. The world-renowned IBM Watson for Oncology (WFO) is a leading medical artificial intelligence that recommends treatment options to doctors based on medical records of cancer patients. This review examines the concepts and characteristics of WFOs and the challenges to be addressed. In particular, the accuracy and medical efficacy of WFO have not yet been fully validated, and it will be necessary to establish evidence through further clinical studies. Artificial intelligence will make fundamental changes in many areas of medicine in the future. It is necessary for the medical community to establish principles of how to maximize clinical benefits and minimize side effects.

Key words: Artificial intelligence; Clinical trials; Evidence based medicine; Cancer; Machine learning

Corresponding Author: Yoon Sup Choi
Digital Healthcare Institute, Seocho-gu, Banpodero 30 gil, Room 1229, Seoul, Korea
E-mail: yoonsup.choi@gmail.com

Received 27 July 2017
Revised 18 Sep 2017
Accepted 2 Nov 2017

서 론

실리콘밸리의 선각자이자 유명 벤처투자가인 비노드 코슬라(Vinod Khosla)는 몇 년 전 '미래에는 80%의 의사가 첨단 기술로 대체될 것'이라고 공개석상에서 주장한 바 있다 [1, 2]. 그는 의료의 많은 부분이 여전히 근거에 기반을 둔 과학이라고 보기 어렵다며, 대규모의 데이터에 기반하고 막강한 연산 능력으로 무장한 기계가 평균적인 의사보다 더 저렴하면서도 정확하고 객관적일 수 있다고 언급했다.

그는 '닥터 알고리즘(Doctor Algorithm)'의 실력은 갈수록 좋아져서, 어려운 치료 사례에 대해서도 모든 가능성을 고려하여 2차 소견을 제공하면서 진료실에서의 영향력은 더 커질 것이라고 했다. 또한 많은 경우 의사들의 진료에 일관성이 부족하고, 편차가 크다는 점도 지적했다.

비노드 코슬라는 선마이크로시스템즈(Sun Microsystems)를 창업한 실리콘밸리 IT 분야의 입지전적인 인물로, 지금은 자신의 이름을 딴 코슬라 벤처스(Khosla Ventures)라는 벤처캐피털을 이끄는 전설적인 벤처투자가다. 이렇게 IT 업계에서 영향력 있는 인물의 도발적인 발언은 즉시 의료계 내 외부에서 격렬한 찬반양론을 불러일으켰다 [2, 3, 4].

이것이 2012년의 이야기였다. 만약 같은 주장을 오늘날에 듣게 된다면 이제는 어떨까. 당시에는 의학을 모르는 타 분야 전문가의 허무맹랑한 주장 정도로 치부되었을 이야기가, 지금에 이르러서는 그리 가볍게 느껴지지 않는다. 그만큼 지난 몇 년 동안의 짧은 기간 동안 인공지능 기술이 폭발적으로 발전했기 때문이다.

이후 지금까지 인간 의사 수준의 혹은 그 이상의 실력을 가지는 인공지능 연구들이 쏟아져 나오기 시작했고, 몇몇 병원에서는 인공지능을 실제로 도입하기에 이르렀다. 하지만 이제는 인공지능에 의해서 미래의 의료와 의사의 역할이 어떤 식으로든 상당히 달라지게 될 것이라는 주장 자체에는

의료계 내부에서도 큰 이견이 없어 보인다.[5, 6, 7]

현재 다양한 의료 분야에서 여러 종류의 인공지능이 발전해왔으며, 앞으로도 새로운 인공지능과 연구 결과들은 지속적으로 등장하게 될 것이다. 향후 예상되는 모든 종류의 의료 인공지능을 포괄하여 분류한다는 것은 쉬운 일이 아닐 것이다. 다만 필자는 적어도 현재까지 연구되고 있는 대부분의 의료 인공지능을 다음과 같이 세 가지 정도의 유형으로 분류할 수 있다고 본다.

- 복잡한 의료 데이터를 분석하여 의학적 통찰력을 도출하는 인공지능
- 이미지로 나타낼 수 있는 의료 데이터를 분석 및 판독하는 인공지능
- 연속적인 의료 데이터를 모니터링하여 질병을 예측 및 예방하는 인공지능

첫 번째 유형이 바로 복잡한 의료 데이터를 분석하여 의학적인 통찰력을 도출하는 인공지능이다. 여기에서 '복잡한 의료 데이터'라고 한다면, 전자의무기록(EMR)이나 차트에 저장되어 있는 환자의 진료 기록이나, 환자의 진료비를 청구한 데이터, 유전체 데이터, 임상 시험 데이터 등의 의료 빅데이터를 포괄한다.

이러한 인공지능은 의료 빅데이터를 분석하여 '의학적 통찰력'을 도출할 수 있을 것이다. 예를 들어, 진료기록 등을 바탕으로 환자의 질병을 진료하거나 진단하거나 예측한다 [8]. 또한 유전체 데이터를 바탕으로 질병을 유발한 원인이 되는 유전적 요인을 정밀하게 찾아주고 [9], 특정 환자에게 가장 적합한 임상 연구가 어떤 것인지 매칭해줄 수 있다 [10]. 더 나아가 사망률이나 재 입원율을 낮추고, 의료비를 절감하는 목적으로 사용할 수도 있다 [11].

이러한 유형의 의료 인공지능 중에 가장 잘 알려진 것은 바로 IBM의 왓슨(Watson)이다. 왓슨은 현재 의료 분야에서 암 환자 진료(Watson for Oncology), 유전체 분석(Watson Genomics), 임상시험 환자 매칭(Clinical Trial Matching) 등의 세 가지 서비스를 제공하고 있다. 이번 종설에서는 현재 국내에서도 많은 이슈를 낳고 있는 IBM 왓슨 포 온콜로지 (Watson for Oncology)를 둘러싼 여러 측면에 대해서 살펴보도록 하겠다.

본 론

1. 왓슨의 병원 도입

퀴즈쇼 제퍼디의 우승 이후 왓슨은 본격적으로 의료 분야에 진출하여 암 환자의 진료에 도전하겠다고 발표한다. 제퍼디 당시에는 알려지지 않았지만, 2011년 5월에 발표된 기사를 보면 이미 18개월 전부터 메릴랜드 대학의 엘리엇 시걸(Eliot Siegel) 박사 팀과 협력하여 각종 의학 논문, 교과서 등의 연구 결과들, MD앤더슨이나 존스홉킨스 등에서 나온 질병 데이터를 학습하고 있다고 언급되어 있다 [12].

이후 2012년 3월 왓슨은 세계에서 가장 오래되었고, 가장 큰 사립 병원인 뉴욕의 '메모리얼 슬론 캐터링 암 센터(Memorial Sloan Kettering Cancer Center, 이하 MSKCC)'와 협업을 통해 암의 치료에 도전하겠다고 밝힌다 [13]. 실제 의사들이 어떻게 암 환자를 진료하고, 진단을 내리며, 치료하는 지에 대한 의사 결정을 학습하기 위한 것이었다. IBM과 MSKCC는 공동연구를 통해 폐암을 시작으로 유방암 등 다른 암종으로 범위를 넓혀가겠다는 계획을 발표했다.

그렇게 개발이 시작된 것이 바로 '왓슨 포 온콜로지(Watson for Oncology)'이다. 2013년 2월에 IBM의 보도자료에 따르면, 당시까지 왓슨은 암과 관련된 60만 건의 의학적 근거, 42개의 의학 학술지와 임상 시험 데이터로부터 200만 페이지 분량의 자료를 학습했다고 나온다 [14]. 또한 MSKCC의 의사들이 1,500여 개의 실제 폐암 치료 사례와 25,000개의 치료 사례 시나리오, 의사들의 진료 기록, 검진 결과 등 '자연어'로 되어 있는 데이터를 학습시켰다고 한다. 그 이후로 14,700시간 동안 간호사들이 수작업으로 왓슨의 학습을 주의 깊게 수정했다고 한다.

이후로 3년 반 정도가 지난 2016년 9월 가천대 길병원에서 이 왓슨 포 온콜로지를 도입할 당시의 자료에 따르면, 300개 이상의 의학 학술지, 200개 이상의 의학 교과서, 1,500만 페이지의 의료 정보를 학습했다고 언급되어 있어서, 학습한 데이터의 크기는 더 증가한 것을 알 수 있다 [15].

길병원 도입 당시에도 왓슨 포 온콜로지는 폐암뿐만 아니라, 유방암, 대장암, 직장암, 위암에 적용 가능하도록 개발되었다. 더 나아가 2017년 6월 보도에 따르면, 왓슨 헬스 측은 연내에 왓슨 포 온콜로지를 총 12개 암종에 적용 가능하도록 발전시켜, 전 세계에서 발병하는 암의 80%를 커버할 것이라고 밝혔다 [16].

왓슨은 현재 세계적으로 여러 병원에 꾸준히 새롭게 도입되고 있다 [17]. 2017년 중반 기준으로 세계적으로 수십 개 정도의 병원에 채택된 것으로 보인다. 왓슨 포 온콜로지는 2014년 태국의 범룽랏 국제병원(Bumrungrad International Hospital)에 도입되었으며, 2015년 12월에는 인도의 마니팔 병원(Manipal Hospital)에, 2016년 8월에는 항저우 코그니티브케어(Hangzhou CognitiveCare)를 통해서 중국의 21개 병원에 도입되었다.

한국에는 2016년 9월에는 가천대학교 길병원의 도입을 시작으로, 2017년에는 부산대학교병원, 대전의 건양대학병원, 대구의 계명대 동산병원과 대구가톨릭병원에 연달아 도입되었다. 부산대학병원은 국내에서는 유일하게 왓슨 포 온

콜로지 뿐만 아니라, 유전체 분석 관련 서비스인 왓슨 지노믹스도 도입했다.

2. 왓슨 포 온콜로지란 무엇인가

그렇다면 왓슨 포 온콜로지는 과연 어떤 기능을 가지고 있을까. 환자의 진료 기록과 의료 데이터를 바탕으로 가능한 치료법(treatment plan option)을 권고해주는 것이 왓슨 포 온콜로지의 기능이다. 예를 들어, 해당 암 환자의 진료 기록, 검사 기록, 유전 정보, 수술 가능 여부 등을 입력하면, 이를 기반으로 치료법을 권고해주는 것이다. 특정 종류의 항암제 혹은 항암제의 조합, 방사선 치료, 호르몬 치료 등을 권고해준다.

중요한 것은 치료법을 초록색, 주황색, 빨간색의 3단계로 권고한다는 것이다. 초록색은 추천하는(recommended) 치료법, 주황색은 고려해볼 수 있는(for consideration) 치료법이며, 빨간색은 권고하지 않는(not recommended) 치료법이다. 또한 각각의 권고된 치료법 마다 근거 버튼이 달려 있다. 이것을 클릭하면, 왜 이러한 치료법을 권고하는지에 대해서 왓슨이 학습했던 관련 논문, 임상 연구 등의 결과, 가이드라인 등의 근거 자료들을 보여준다. 만약에 의사가 처음 보는 치료법이라고 할지라도, 이러한 근거 자료에 기반하여 해당 권고안이 과학적, 의학적으로 설득력이 있는지를 고민해볼 수 있다.

엄밀히 말해 왓슨 포 온콜로지는 진단(diagnosis)을 해주는 것은 아니며, 치료법을 권고하여 의사의 진료를 보조하는 역할을 한다. 적어도 아직까지는 왓슨 포 온콜로지에는 병리과에서 시행하는 병리 데이터의 분석을 통한 진단 기능은 포함되어 있지 않다.

또한 진료와 진료를 보조하는 것에도 큰 차이가 있다. IBM 측의 발표를 들으면 항상 빠지지 않는 것이 바로 '왓슨은 의사를 대체하지 않는다. 의사의 역할을 강화(augment)하는 것이 왓슨의 역할이다'는 것이다. 한국이든 미국이든 IBM 관계자가 하는 왓슨 관련 발표에는 항상 이 표현이 빠지지 않는다. 이는 의료계에서 IBM이 왓슨의 의사와의 관계 및 포지셔닝을 어떻게 하려는 지를 암시한다.

3. 왓슨 포 온콜로지의 특징

IBM이 주장하는 왓슨의 강점 중의 하나는 매일 같이 쏟아져 나오는 엄청난 분량의 암과 관련된 연구 논문들, 임상시험 결과들을 환자의 치료에 빠르게 반영할 수 있다는 것이다 [14, 15]. IBM에 따르면 2015년 한 해 동안 출판된 종양학 논문은 44,000개이다 [15]. 이는 매일 122개의 새로운 논문이 발표된다는 이야기다. 이는 10분에 한 편씩 논문을 읽는다고 가정해도, 주말 없이 매일 20시간 이상씩 읽어야만 따라갈 수 있는 양이다. 즉, 인간의 능력으로 따라가기에는 이미 불가능한 수준의 연구 결과들이 쏟아진다는 것이다.

이렇게 최신 연구 결과를 치료법 선택에 빠르게 반영할 수 있다는 것이 IBM이 주장하는 왓슨의 강점이라고 할 수 있다. 하지만 이는 양날의 검과 같다. 뒤에서 자세히 논의하겠지만, (정확한 업데이트 주기는 알 수 없으나) 지속적으로 연구 결과를 업데이트하면서 진화한다는 왓슨의 특징은 이를 규제적으로 정의하거나, 정확성, 효과성, 임상적 유용성 등을 증명하는 데 있어서 적지 않은 근본적 문제를 야기한다.

한 가지 유의할 점은 왓슨의 학습이 결코 '자동으로' 이뤄지지 않는다는 것이다. 왓슨이 제퍼디에서 활용한 자연어 처리 기술을 기반으로 하였기 때문에, 논문도 자동으로 읽고 스스로 판단하여 암 환자 진료의 근거로 삼을 것이라고 생각하기 쉽다. 하지만 실제로는 왓슨을 훈련시키는 과정에서 수작업 교정에 많은 시간을 사용한 것으로 알려져 있다. 또한 쏟아져 나오는 종양학 논문 중에 어떤 논문을 왓슨에 반영할 것인지를 결정하기 위해서 MSKCC의 종양학 전문의들이 관여한다고 알려져 있다.

기계 학습의 기본적인 원칙은 '가비지-인, 가비지-아웃(garbage-in, garbage-out)'이다. 즉, 좋지 않은 데이터로 학습시키면 좋지 않은 결과가 나온다. 훈련시킬 때의 데이터의 양과 질이, 결과적으로 인공지능의 성능을 좌우한다고 해도 과언이 아니다. 일례로 제퍼디를 준비할 당시에, 왓슨에게 구어체의 학습을 목표로 속어나 은어 등이 담긴 'The Urban Dicationary'를 학습시켰던 적이 있다. 그 결과 제퍼디 답지에 욕지거리가 포함되는 경우들이 있었기 때문에 결국 개발자들은 그 데이터를 삭제할 수밖에 없었다. 강한 인공지능이 구현되지 않는 이상, 아직까지 인공지능을 학습시킬 때 어떤 데이터가 좋은 데이터이며, 나쁜 데이터인지는 개발하는 사람이 판단할 수밖에 없다. 그런데 이 왓슨 포 온콜로지를 규제적으로, 의학적으로 어떻게 봐야 할까? 몇 가지 어려운 이슈들이 있다.

4. 왓슨 포 온콜로지는 의료기기일까

일단 왓슨 포 온콜로지가 과연 규제적으로 의료기기로 분류되어야 할지, 비의료기기로 분류되어야 할지의 이슈에 대해 살펴보자. 만약 왓슨 포 온콜로지와 같은 임상 의사결정 지원시스템(CDSS, Clinical Decision Support System)이 의료기기라면 정확성, 안전성 등을 검증 받고 FDA나 식약처의 의료기기 인허가 과정을 거쳐야만 할 것이다. 만약 비의료기기로 분류된다면 인허가 과정이 필요 없다.

이는 매우 애매한 문제다. 필자는 강의에서 왓슨 포 온콜로지에 대해서 설명하고, 청중에게 이 시스템이 의료기기로 분류되어야 할지, 아니면 비의료기기로 분류되어야 할지 의

견을 물어보곤 한다. 이 문제에 대해서 청중들의 답변은 항상 양쪽으로 갈리곤 한다. 심지어는 실제로 유럽과 미국, 한국의 규제 기관의 판단에도 결과적으로 차이가 있다.

일반적으로 의료기기의 여부는 목적과 위해도라는 두 가지를 기준으로 정해진다. 왓슨 포 온콜로지의 목적은 진료나 진단, 치료를 직접 하지 않고, 의사를 보조하는 목적이다. 또한 내어놓은 결과를 직접 환자에게 적용하지 않고, 의사가 그 결과를 검토하여 의사가 치료법을 결정하게 되므로 위해도 역시 높지 않다고 간주한다. 이러한 논리에 따라 현재의 식약처나 FDA의 결정은 왓슨 포 온콜로지가 의료기기가 아니라는 것이다. (이러한 결정은 대부분의 국가에서 동일한 것으로 보이나, 예외적으로 EU의 경우에는 왓슨 포 온콜로지를 의료기기로 규정하고 있다.)

기사에 따르면, 복지부 관계자는 "(의사들의 왓슨 활용은) 평소 의사들이 진단과 처방을 내림에 있어 관련 서적과 논문 등을 참고하는 것과 같은 성격으로 봐야 한다"며 "때문에 의료법상 왓슨을 사용하는 것은 문제가 없다는 생각"이라고 언급하기도 했다 [19]. 즉, 보다 발전된 의학 교과서의 개념이라는 것이다.

이러한 내용이 2016년 12월에 발표된 "빅데이터 및 인공지능(AI) 기술이 적용된 의료기기의 허가 · 심사 가이드라인"에 결국 반영되었다. 특히, 비의료기기의 예시에는 "전자의무기록, 의료영상, 생체신호를 이용하여 문헌을 검색하고 문헌의 내용(진단법, 치료법 등)을 요약하여 제시하는 소프트웨어"라는 것이 명시되어 있다. 왓슨 포 온콜로지도 여기에 해당한다.

5. 왓슨 규제의 근본적 어려움

그럼에도 불구하고, 필자는 몇 가지 면에서 왓슨이 의료기기인지 여부에 대해서 근본적인 어려움과, 현실적인 문제가 있다는 것을 지적하고자 한다. 비의료기기로 분류하는 경우에도 몇 가지 중요한 문제가 있으며, 그렇다고 의료기기로 분류하더라도 관리가 어려운 부분이 있다는 것이다. 이는 왓슨 포 온콜로지와 같은 시스템이 지금까지의 임상의사결정 지원시스템과도 다른 부분이 있기 때문이다.

일단 왓슨이 의료기기의 성격을 띤다고도 볼 수 있는 것이 바로 결과를 제시하는 양식 때문이다. 왓슨 포 온콜로지는 단순히 가능한 치료법을 권고하는 것에서 더 나아가, 치료법에 대한 우선순위까지도 매겨준다. 앞서 언급했던 바와 같이 치료 권고안을 추천(초록색) – 고려(주황색) – 비추천(빨간색)의 3단계로 점수를 매겨주기 때문이다. 다양한 치료법 중에서 이렇게 우선순위를 매겨준다는 것은 그 자체로 의료적 의사결정이나, 의료 행위로 봐야 하지 않을까? 교과서는 개별 환자에 맞게 치료법의 우선순위를 평가해주지는 않지만, 왓슨 포 온콜로지는 이 부분을 스스로 판단한다.

비록 최종적인 의사결정은 의사가 내린다고는 하지만, 왓슨 포 온콜로지의 권고안이 의사의 치료법 결정에 영향을 줄 가능성이 전혀 없다고 할 수 없을 것이다. 만약에 영향을 전혀 주지 않는다면, 애시당초 왓슨과 같은 CDSS를 사용할 이유가 없을 것이다. 그렇다면 왓슨이 제시하는 치료 권고안의 정확성이 중요하다고 할 수 있다. 그럴 가능성은 낮겠지만 극단적으로 가정하여, 왓슨 포 온콜로지가 모든 경우에 부정확한 치료 권고안을 준다면 환자에게 위해가 가지 않는다는 법은 없다. 이러한 것을 고려하면 왓슨에게는 분명히 의료기기적인 성격이 전혀 없다고는 하기 어려울 것이다.

하지만 설사 의료기기로 분류한다고 할지라도 이를 규제하고 관리하기가 매우 어렵고 까다롭다. 바로 왓슨 포 온콜로지가 끊임없이 변화하고 진화하기 때문이다. 앞서 언급했듯이 IBM이 주장하는 왓슨 포 온콜로지의 가장 큰 장점 중의 하나는 최신 연구 결과를 반영한다는 것이다. IBM의 보도자료에서 직접 밝혔듯이 하루에 100개 이상의 종양학 논문이 쏟아져 나오므로, 최신 논문들을 최대한 빠르게 반영하려고 할 것이다.

이런 논문이 반영됨에 따라서, 왓슨 포 온콜로지는 계속해서 변화한다. 문제는 이렇게 논문 등의 근거 자료가 업데이트되면, 판단 근거가 달라지므로 동일한 환자를 대상으로 내어놓는 치료 권고안에도 변화가 생길 수도 있다는 것이다. 예를 들어, 어제와 오늘, 그리고 내일의 왓슨 포 온콜로지는 완전히 동일한 시스템이라고 보기 어려울 수 있다는 것이다.

연구 결과의 최신 반영 이외에 왓슨 포 온콜로지가 진화하는 또 하나의 경로가 있다. 바로 실제 이 시스템을 활용하는 의료진의 의사 결정을 반영하는 것이다. 예를 들어, 길병원에서 의료진이 왓슨 포 온콜로지를 사용하면서 치료법에 대한 피드백을 IBM으로 보낼 수 있다. 만약 한국의 의사들이 보기에 위암에 대한 왓슨의 판단이 적절하지 않으면, 이에 대한 코멘트를 보낼 수 있는 것이다. IBM의 입장에서는 그 결정에 대해서 더 알아보고 필요한 경우 왓슨의 판단에 그러한 결정을 반영하는 것도 필요할 것이다.

그런데 이러한 최신 논문의 반영이나 진료 현장에서의 피드백을 누가 어떤 기준으로 판단하며, 얼마나 자주 왓슨에 반영하여 업데이트할까. 논문으로 출판되었다고 해서 모두 양질의 데이터라고는 할 수 없으며, 일선 의료진의 피드백을 모두 타당한 것으로 보기는 어렵다. 필자가 IBM 왓슨 헬스의 최고 의료 책임자(Chief Health Officer)인 Kyu Rhee 박사를 인터뷰한 결과 이러한 부분은 MSKCC의 종양학자들이 직접 결정한다고 한다. 논문의 경우 정기적으로 간행되므로, 이 스케쥴에 맞게 정기적으로 심사한다고 밝혔다.

중요한 것은 이러한 과정을 통해 왓슨이 끊임없이 변화한다는 것이다. 만약에 왓슨 포 온콜로지가 의료기기로 분류

되었을 경우, 기존의 의료기기 관리 기준에 따른다면 내부 알고리즘이나 작동 원리에 변화가 있다면 변경 인허가를 새롭게 받아야 한다.

그렇다면 왓슨 포 온콜로지의 경우에도 연구 결과가 반영되거나, 의료진의 의사 결정이 반영될 때마다 정확성을 새롭게 검증하고, 매번 의료기기 인허가를 다시 받아야 할까? 예를 들어, 매일 왓슨이 업데이트된다면, 매일 변경 인허가를 새롭게 받아야 하는가? 만약 이렇게 기존의 의료기기의 규제와 관리에 대한 잣대를 그대로 들이대면, 왓슨 포 온콜로지와 같은 시스템은 의료기기로서는 아예 사용이 불가능할 수도 있다. 그렇기 때문에 왓슨을 의료기기로 관리하는 것에도 근본적인 어려움이 있다는 것이다.

6. 왓슨 포 온콜로지는 얼마나 정확한가

가장 큰 문제는 왓슨 포 온콜로지가 내어놓은 치료 권고안이 얼마나 정확한지 아직 완전히 검증되지 않았다는 것이다. 그 유명세에 비해서 IBM 왓슨 포 온콜로지의 암 환자 진료의 정확성은 증명된 바 없다. MSKCC라는 세계 최고의 암 병원에서 함께 개발했으니, 어느 정도의 정확성은 갖추고 있으리라고 추측을 해볼 수는 있을 것이다.

하지만 왓슨 포 온콜로지의 추천, 고려, 비추천의 등급의 정확성과 이러한 시스템의 의학적 효용성을 검증하기 위한 무작위 대조군 임상시험(RCT)이 정식으로 진행된 바 없고, 이것이 의료계와 학계에서 인정하는 학술 저널에 논문으로 발표된 적도 전무하다. 대규모 임상 연구의 결과가 아니라고 하더라도, 왓슨 포 온콜로지를 통해서 실제로 치료한 환자의 증례 보고(case report)도 전무하다.현재 우리가 참고할 수 있는 것은 종양학과 관련된 학회에서 초록의 형태로 병원별로 왓슨 포 온콜로지의 사용 결과가 발표된 사례 정도다 [20]. 사용 결과 분석의 경우에도 수백 명 이상의 환자 대상의 결과가 발표되기 시작한 것은 왓슨 포 온콜로지가 여러 병원에 도입된 이후인 2016년 12월 정도부터였다. 이 연구의 결과들은 우리가 고민해야 할 여러 이슈를 던져준다.

비록 FDA와 식약처는 왓슨 포 온콜로지가 의료기기에 해당하지 않는다고 판단하였지만, 그럼에도 불구하고 필자는 어떤 식으로든 왓슨 포 온콜로지의 정확성, 효과성에 대한 의학적인 검증이 필요하다고 본다. 아무리 이 시스템에서 도출된 치료 권고안을 의사가 한 번 더 검토한다고 할지라도, 결과가 부정확할 경우 환자에게 미치는 위해도가 전혀 없다고 할 수 없기 때문이다.

뿐만 아니라, '인공지능이 암 환자를 진료한다'는 주장에 이끌려 많은 환자들이 왓슨을 도입한 병원을 찾고 있는 것이 현실이기도 하기 때문이다. 2017년 1월 조선일보가 보도한 바에 따르면, 길병원에서 2016년 11월부터 진료한 100여 명의 환자 중에서, 의사와 왓슨의 판단이 다를 경우에 환자들은 모두 의사보다 왓슨을 더 신뢰했다고 한다.

우리는 왓슨 포 온콜리지가 정확하다면 얼마나 정확한지, 정확하지 않다면 얼마나 정확하지 않은지를 알아야 한다. 또한 이러한 시스템이 의료적으로 효용이 있는지, 있으면 얼마나 있는지에 대해 파악하는 것이 필요하다. 이러한 근거가 있어야만 비로소 우리가 왓슨을 진료실에서 어떻게 활용하는 것이 좋을지에 대한 원칙을 세울 수 있기 때문이다. 지금은 병원에서 왓슨을 어떤 방식으로 진료 프로세스에 녹여낼지에 대한 원칙도 없고, 근거가 없기 때문에 그 원칙을 만들기도 어려운 상황이다.

7. 인도 마니팔 병원의 사례

2016년 12월, 인도의 마니팔 병원에서 왓슨 포 온콜로지의 정확성을 평가하기 위해서 우리가 참고할만한 최초의 결과를 발표하였다 [20]. 과거 3년간 치료받은 유방암, 대장암, 직장암, 폐암 등 4가지 암종의 환자 1,000명에 대해서 의사의 판단과 왓슨 포 온코로지의 판단이 얼마나 일치했는지를 본 것이다. 다양한 암종의 대규모 환자를 대상으로, MSKCC가 아닌 독립적인 병원이 왓슨 포 온콜로지의 정확도를 공개한 것은 실질적으로 이 연구가 처음이라고 해도 좋을 것이다.

참고로 ASCO 2015에서 MSKCC의 의료진이 발표한 초록이 몇 가지 있다. 다만 모두 특정 암에 편중되어 있고, 환자의 수가 충분하지 않거나, 가상의 환자를 대상으로 한 것이기 때문에 왓슨의 정확도를 파악하기에는 한계가 있다는 점을 알려 둔다 [22, 23, 24].

인도의 마니팔 병원은 지난 2015년 12월 인도에서는 최초로 왓슨 포 온콜로지를 도입했다. 아시아에서는 태국의 범룽랏 병원 이후로 두 번째다. 마니팔 병원은 뱅갈로르를 중심으로 16개 병원의 네트워크로 이루어진, 총 5,000병상을 갖추고 연간 20만 명 이상의 암 환자를 진료하는 대형 암 센터이다.

해당 연구의 대상 환자는 네 가지 암종으로, 각각 유방암(638명), 대장암(126명), 직장암(124명), 폐암(112명)으로 구성되어 있다. 왓슨 포 온콜로지를 도입한 지 당시 1년이 지난 시점인데, 연구에서는 과거 3년 동안 진료한 환자를 분석한 것이므로, 후향적(retrospective) 연구로 볼 수 있다.

이 연구에서 마니팔 병원의 암 환자 진료와 관련된 여러 진료과의 전문의들이 모인 다학제 진료팀(Manipal multidisciplinary tumour board)의 판단과 왓슨 포 온콜로지의 판단을 비교하였다. 1,000명의 환자에 대해서 마니팔 병원의 다학제 진료팀이 제시한 치료법을 기준으로, 왓슨 포 온콜로지가 제시한 치료법 중에 추천, 고려, 비 추천의 세 단계의 권고안과 일치하는 비율은 아래와 같았다.

- '추천'과 일치: 50%
- '고려'와 일치: 28%
- '비 추천'과 일치: 17%

다학제 진료팀의 판단을 정답으로 보고, 왓슨의 추천과 고려 항목이 일치하는 경우를 모두 합하면 약 80% 정도의 정확도를 가진다고 볼 수 있다. 또한 추천, 고려, 비추천 모두에 속하지 않는 나머지 5%의 경우에는 의사들이 결정한 치료법을 왓슨 포 온콜로지의 권고안 중에서는 찾을 수 없었다.

그런데, 문제는 암종별로 왓슨의 치료 권고안과 의사가 결정한 치료의 일치도에 현저한 차이가 드러났다는 것이다. 추천 항목과의 일치율을 기준으로 한다면, 일치율이 가장 높았던 것은 직장암으로 85%였으며, 가장 낮은 것은 폐암으로 17.8%에 불과했다.

더 나아가, 유방암의 경우에는 세부 암종별로도 일치율의 차이가 나타났다. 호르몬 수용체와 HER2 유전자가 모두 음성으로 나오는 삼중 음성(triple-negative) 유방암의 경우에는 67.9%가 일치했고, HER2 유전자만 음성인 경우에는 35%로 일치도가 낮았다. 또한 비전이성 유방암은 80% 일치하였으나, 전이성 유방암은 45% 밖에 일치하지 않았다.

이 연구의 결과만 놓고 볼 때, 왓슨이 내어놓는 치료 권고안의 의사 대비 정확성이 암종별로 상당히 차이가 크다는 것으로 이야기할 수 있다. 특정 암종에 대해서는 80% 이상의 비교적 높은 일치도를 보이지만, 특정 암종이나, 세부 암종에 따라서는 또 50% 에도 미치지 못하는 일치율을 보이는 것이다.

8. ASCO 2017에 보고된 왓슨의 실력

약 반년 뒤인 2017년 6월에도 비슷한 연구 결과들이 발표되었다. 2017년 6월 초에 시카고에서 열린 미국 임상 종양학회(ASCO) 2017에서 태국의 범룽랏 병원, 인도의 마니팔 병원, 한국의 가천대학교 길병원 등 세 병원이 각각 왓슨 포 온콜로지의 치료 권고안과 해당 병원 의료진의 결정이 얼마나 일치했는지에 대해서 초록의 형태로 발표한 것이다.

이 세 병원의 연구 결과들은 세부적으로는 차이가 있지만, 큰 그림에서 보면 우리가 얻을 수 있는 결론은 크게 다르지 않다고 하겠다. 이 결과들을 간략히 살펴보면 아래와 같다.

태국 범룽랏 병원의 경우 2015-2016년에 치료받은 폐암, 유방암, 위암 등 세 암종의 환자 211명을 대상으로 살펴보았다 [25]. 이 중에 92명은 과거에 치료했던 환자의 기록을 왓슨 포 온콜로지와 비교해본 후향적 연구였고, 나머지 119명은 새롭게 진료받은 환자의 기록을 왓슨 포 온콜로지가 분석하여 의사와 비교해본 전향적 연구였다. 추천과 고려와의 일치를 기준으로 전체 환자군에 대한 일치도는 83%였으며, 암종별로 보면 폐암 91%, 유방암 76%, 위암 78%였다. 후향적 분석과 전향적 분석 결과는 비슷했다.

인도 마니팔 병원에서도 상기에 언급한 2016년 12월 발표된 결과에서, 유방암을 제외하고 폐암 112명, 대장암 126명, 직장암 124명에 대한 결과를 다시 발표했다 [26]. 암종별로 일치율을 보았을 때(역시 추천과 고려를 모두 기준으로 하여) 폐암 96.4%, 대장암 81%, 직장암 92.7%였다.

또한 이번 발표에는 세 암종에 대해서 원발 조직에 국한(localized)되었을 때와 전이되었을 때의 병기별로 구분한 결과도 발표되었다. 병기에 따라서도 일치율에 다소 차이가 나는 것을 알 수 있는데, 폐암의 경우에는 전이암일 때, 대장암과 직장암은 원발 조직에 국한되었을 때 일치율이 더 높았다.

- 폐암: 국한 88.9%, 전이성 97.9%
- 대장암: 국한 85.5%, 전이성 76.6%
- 직장암: 국한 96.8%, 전이성 80.6%

마지막 세 번째 발표는 한국의 길병원에서 치료받은 2-4기 대장암 환자 340명과 항암치료를 받지 않은 진행성 위암 환자 185명을 대상으로 후향적 연구를 한 것이다 [27]. 대장암 환자 전체에서는 73%의 일치율을 보였다. 그중에서 보조 항암 치료를 받은 환자 250명의 경우에는 85%, 전이성 대장암 환자 90명의 경우 40%가 일치했다. 또한 위암 환자의 경우에는 49%에서 일치했다.

이러한 결과를 보면 아래와 같은 네 가지 정도의 결론을 잠정적으로 내려볼 수 있다.

- 왓슨 포 온콜로지와 의사의 일치율은 암종별로 다르다.
- 왓슨 포 온콜로지와 의사의 일치율은 같은 암종에서도 병기별로 다르다.
- 왓슨 포 온콜로지와 의사의 일치율은 같은 암종에 대해서도 병원별, 혹은 국가별로 다르다.
- 왓슨 포 온콜로지와 의사의 일치율은 시간에 따라서 달라질 가능성이 있다.

9. 왓슨과 의사의 일치율 차이의 원인

1) 가이드라인 및 인종적 차이

왓슨의 결과에 왜 이러한 차이가 있을까. 먼저 왓슨과 의사의 일치율이 암종별로, 병기별로, 병원별로, 국가별로 왜 다른지에 대해서 먼저 논해보자.

필자는 이러한 일치율의 차이에 대해서 왓슨 헬스의 최고의료책임자를 포함한 IBM 소속의 의사들과도 개인적으로 이야기를 나눠보았다. 그들이 공통적으로 꼽는 일치율의 차이는 왓슨 포 온콜로지가 미국이라는 특정한 나라의 환경에서 MSKCC라는 특정 병원을 기준으로 개발된 시스템이기 때문이다. 그렇기 때문에 아래와 같은 요소들에 국가별로 차이를 드러낼 수 있다.

- 해당 국가 진료 가이드라인 준수 여부
- 암 환자 인종별 차이의 고려 여부
- 권고한 약이나 치료법의 국가별 인허가 여부
- 보험 급여 기준 및 심사 기준 준수 여부

한국을 비롯한 각 나라에서는 암 환자에 대한 진료 가이드라인이 존재한다. 전 세계적으로는 NCCN(전미 종합 암 네트워크)에서 발간하고 수시로 업데이트하는 치료 가이드라인이 권위를 가지고 있지만, 한국을 포함한 개별 국가에서는 자국의 상황에 맞게 변경되어 있는 가이드라인을 따르기도 한다. 이에 따라 항암제의 종류와 사용, 수술의 필요 여부 등이 달라지기도 한다.

또한 인종 별 차이도 무시할 수 없다. 미국인에 맞게 개발된 왓슨 포 온콜로지는 다른 국가, 특히 아시아인 환자를 대상으로는 인종적 특수성을 고려하지 못하는 것으로 알려져 있다. 이러한 요인도 앞서 언급한 태국, 인도, 한국 병원의 일치율 차이의 요인이 될 수 있다. 암은 유전적 요인에 의해서 발병하는데, 인종별로 발병 원인 유전자의 구성이나 유전자 발현의 정도가 다를 수 있다. 이런 요인 때문에 항암제에 대한 반응이나 부작용이 달라지는 경우가 있다. 즉, 어떤 경우에는 같은 항암제를 사용하더라도 치료 효과가 달라질 수 있다는 것이다.

예를 들어, 아스트라제네카의 폐암 치료제 이레사는 아시아인과 비아시아인의 반응이 다른 대표적인 약제 중의 하나다. 2003년 발표된 이레사의 연구에서는 서양인에 비해 일본인에 더 큰 효과를 보인다는 것이 증명되었으며 [29], 성균관대 의대 연구진이 비소세포폐암(NSCLC)에 대해서 폐암 종양이 50% 이상 감소하는 환자의 비율이 서양인에 비해 한국인이 두 배 높다는 것을 증명한 바 있다. 또한 세브란스 연구팀은 항암제 파클리탁셀이 아시아인 위암 환자의 경우 감수성이 38%에 불과하지만, 비아시아인 환자에게는 75%나 된다는 연구 결과를 2009년 발표한 바 있다 [31].

최근 발표에 따르면 이러한 인종적 차이는 실제로 왓슨과 국내 의료진의 결정의 차이를 만들어낸다. ASCO 2017에서 길병원 발표에서 위암 환자의 낮은 일치율에 대해서 두 가지 가능성이 제시되고 있다[27]. 그 중 하나가 항암제 S-1(tegafur, gimeracil and oteracil)+cisplatin의 조합이 한국에서는 일상적으로 사용되지만, 미국에서는 그렇지 않기 때문이라는 것이다. 일본에서 개발된 S-1이라는 항암제는 임상 연구 결과 일본과 한국의 환자에게는 우수한 결과를 보여준 바 있다. 하지만 서양인에게는 설사와 같은 부작용이 흔해서 잘 쓰지 않는 것으로 알려져 있다.

또한 국가별로 인허가받은 약제의 차이가 있을 수 있다. 미국에서는 FDA의 승인을 받아서 환자들에게 판매되는 약이, 한국에서는 여러 이유로 식약처의 허가를 받지 못해 사용이 불가한 약제일 수 있기 때문이다. 반대로, 한국에서 승인 받은 약이, 미국에서는 아직 허가를 받지 않았을 수도 있다. 앞서 언급한 항암제 S-1은 일본과 한국, 유럽의 여러 국가에서도 인허가 받았으나, 아직 미국의 FDA에서는 허가를 받지 않았다.

2) 보험제도의 차이

더 크게 지적되는 문제는 건강 보험 제도의 차이 때문이다. 한국은 전 세계에서도 드물게 전 국민에게 국민건강보험을 보장해주는 시스템을 가지고 있다. 이는 전 세계가 부러워하는 보험 체계이기도 하지만, 의사의 진료와 처방이 심평원(건강보험심사평가원)의 급여 기준에 부합해야 한다는 특수한 환경을 야기했다.

문제는 왓슨 포 온콜로지가 당연히 심평원의 급여기준과는 상관 없는 치료법을 권고안으로 제시한다는 것이다. 왓슨이 정말로 하루에 수백 편씩 쏟아지는 최신 연구에 발맞춰서 최적의 치료법을 제시해준다고 하더라도, 이러한 치료법에 대한 보험 급여를 적용 받지 못하거나 심평원에서 삭감해버린다면 국내에서는 이 권고안을 채택하기가 어려워질 것이다.

이러한 차이는 왓슨 포 온콜로지의 판단과 국내 의료진의 판단의 차이를 만들어내는 실제 요인이 된다. ASCO 2017에서 길병원의 의료진이 발표한 연구 결과에는 위암에서의 일치도가 낮은 요인에 대해(앞서 설명한 S-1의 국가별 차이 이외에도) 왓슨이 권유하는 항암제 Trastzumab/FOLFOX 가 한국에서는 국민 건강 보험 수가를 받지 못하기 때문에 택하기가 어렵다는 점을 명시적으로 언급하고 있다 [27].

3) 치료 옵션 다양성의 차이

또한 암종별로 일치도가 다른 이유 중의 하나는 암의 종류에 따라서 얼마나 다양한 치료 옵션들이 존재하는 지에도 차이가 있기 때문으로 보인다. 다양한 치료 옵션이 존재할수록 아무래도 양측의 판단이 확률적으로 일치하기 어렵기 때문이다.

예를 들어, 마니팔 병원의 발표에서 삼중음성 유방암의 경우(일치율 67.9%)에는 HER2 음성 유방암(일치율 35%)에 비해서 가능한 치료 옵션 자체가 적기 때문에 결과적으로 일치도는 올라갈 수밖에 없다는 것이다. 왓슨의 추천과 일치율이 85%로 높게 나온 직장암의 경우에도 상대적으로 다른 암에 비해 치료 옵션의 다양성이 제한적인 편이다.

4) 가이드라인의 변화와 왓슨의 진화

왓슨과 의료진의 일치율이 차이가 나는 또 다른 이유는 가이드라인과 왓슨이 시간이 흐름에 따라서 계속 바뀌어간다는 것이다. 새로운 연구 결과가 발표되고, 새로운 치료법이 개발되면 그에 맞춰서 표준 진료 가이드라인과 왓슨 포

온콜로지의 결정도 계속 바뀌어갈 수 있다. 또한 앞서 언급했듯이, 왓슨 포 온콜로지를 활용하는 의료진의 결과가 피드백될 수 있으므로, 이를 반영해서도 왓슨은 진화한다.

'과거'에 진료했던 환자의 기록을 바탕으로, '오늘'의 왓슨 포 온콜로지를 실행하여 그 결과를 비교하는 후향적 연구의 경우 이러한 차이가 크게 드러날 수밖에 없다. 앞서 언급한 대부분의 연구가 수년 전에 진료한 환자의 치료 방법과 현 시점의 왓슨 포 온콜로지에서 나온 결과를 비교한 것이다.

이러한 요인은 마니팔 병원이 2016년 12월 샌안토니오 유방암 심포지움에서 발표한 자료에서도 확인할 수 있다 [32]. 이 발표에는 638명의 유방암 환자 사례를 두 가지 시점에서 분석하고 있다. 하나는 지난 3년간 환자를 진료했던 과거 시점(T1)에서 실제 치료법과 오늘날 왓슨 포 온콜로지의 판단을 비교한 것이다. 다른 하나는 연구가 진행된 2016년 시점(T2)에서 과거의 진료기록을 재검토하여 의사들이 판단한 것과 왓슨의 결과를 비교한 것이다.

이렇게 과거의 진료 가이드라인 등에 맞춘 의료진의 판단보다 현재의 의료진의 판단이 왓슨과 더 높은 비율로 일치한다. 왓슨 포 온콜로지의 '추천' 만을 기준으로 했을 때에는 46%에서 60%로 증가하였고, '추천'과 '고려'를 모두 기준으로 하면 과거의 73%에서 90%까지 일치율이 증가하는 것을 볼 수 있다. T1과 T2 시점의 일치율 차이는 그 기간 동안 바뀌었던 가이드라인 때문이라고 설명할 수 있다.

그런데 시간이 흐름에 따라서 가이드라인, 신규 논문, 새로운 치료법 등을 반영하여 계속 진화한다는 왓슨의 속성은 또 다른 문제를 만들어낸다. 아래에서 더 자세히 이야기하겠지만, 필자는 왓슨의 정확성과 의학적 효용 등을 증명하기 위해서 임상 시험이 필요하다고 본다. 하지만 이렇게 지속적으로 변화한다는 왓슨은 과거 특정 시점에서 진행한 임상연구의 결과가 현재 혹은 미래에 적용되지 않을 수 있다는 것을 의미한다.

10. 왓슨의 정확도와 의학적 효용의 증명

앞서 논의한 내용을 정리해보자면 왓슨 포 온콜로지의 정확성과 의학적 효용이 아직 완전히 증명되지 않았음을 알 수 있다. 기존에 발표된 연구들은 모두 왓슨 포 온콜로지와 특정 국가의 특정 병원 의료진의 판단과 얼마나 일치하는지를 분석한 정도이다.

이러한 연구가 가지는 결정적 한계점은 왓슨과 의사의 '일치율(concordance)'을 보는 것이 왓슨 포 온콜로지의 정확성과 효용성을 평가하기 위해서 적절한 지표가 되기 어렵다는 것이다. 의사의 판단과 일치한다고 해서 왓슨의 치료법이 정확하다고 할 수 없다. 반대로 의사의 판단과 불일치한다고 해서 왓슨의 권고안이 부정확하다고 할 수는 없는 일이다. 의사의 판단이 최선의 판단일 수도 있지만, 최선이 아닐 수도 있기 때문이다.

만약 '일치율'을 높이는 것이 왓슨의 실력에 대한 유일한 지표나 개발 목표가 된다면, 결국 인간 의사와 완전히 동일한 판단을 하는 시스템이(구현 가능성의 여부는 차치하고서라도) 왓슨의 최종적인 모습이 될 것이다. 즉, '일치율'만을 기준으로 한다면 의사와 동일한 수준의 인공지능을 구현할 수는 있겠지만, 의사보다 더 나은 인공 지능을 개발할 수는 없다.

우리가 인공지능에게 기대하는 것은 한 단계 더 높은 수준이다. 즉, 인간 의사의 부족한 점을 보완해줄 수 있고, 가능하면 때로는 더 나은 치료법을 찾아줄 수도 있는 인공지능이다. 만약 왓슨의 판단이(예를 들어 MSKCC의) 의료진의 결정과 100% 일치율을 달성했다면, 그리고 그것을 증명했다면, 과연 그 시스템은 유용할 것인가. 의사의 수가 부족하거나, 의료진의 종양내과적 전문성이 부족한 환경에서는 유용할 수도 있다. 하지만 평균적인, 혹은 평균 실력 이상의 종양내과 전문의를 충분히 갖춘 병원에서 이는 그리 큰 가치를 제공하지 못할 가능성이 높다.

이 부분에 대해서 앤드류 노든 박사는 ASCO 2017에서 발표된, 일치율이 80-90%라는 왓슨 포 온콜로지의 퍼포먼스에 대해서 만족감을 표시했다 [16]. "이러한 수치는 우리가 원하는 정도다. 만약 일치율이 100%라면 모든 경우에 의사와 완전히 동일한 권고안을 준다는 것이므로 아무런 가치가 없다고 주장할 수 있다. 만약 훨씬 낮거나 0%의 일치율을 보인다면, 그것 또한 문제가 될 것이다."

즉, 현재 의사와 너무 다르지도 않고, 너무 동일하지도 않은 권고안을 주기 때문에, 왓슨 포 온콜로지가 의료진에게 가치가 있을 가능성이 있다는 것이다. 그럼에도 불구하고, 앞서 강조했듯이 아직 그러한 가능성을 증명할 수 있는 근거는 충분하지 않다. 특히, 우리는 아직까지 왓슨 포 온콜로지로 인해 의학적으로 환자나 의료진이 어느 정도의 효용을 얻는지에 대해서 알지 못한다. 예를 들어, 아래와 같은 질문에 대해서 아직 답할 수 있는 근거가 마련되지 않았다.

- 왓슨 포 온콜로지의 권고안은 얼마나 정확한가?
- 왓슨 포 온콜로지가 환자의 생존기간의 연장에 유의미한 도움을 주는가?
- 왓슨 포 온콜로지가 환자의 치료 효과를 개선시키는가?
- 왓슨 포 온콜로지가 의료비 절감 혹은 증가에 어떤 영향을 주는가?
- 왓슨 포 온콜로지가 의료진의 진료 효율성을 높이는가?

이러한 질문에 답을 얻기 위해서는 결국 근거가 필요하다. 근거를 마련하기 위해서 가장 좋은 방법은 역시 임상 시험이다. 필자는 결국 왓슨이 정확성이나 의학적인 효용을 검증하기 위해서는 임상 시험을 거쳐야 할 것이라고 생각한다.

11. 임상시험의 필요성과 어려움

왓슨 포 온콜로지의 정확성과 의학적 효용을 증명하기 위해서 어떤 방식으로든 임상 연구가 필요할 것이라는 점은 많은 의료 전문가들이 동의한다. 필자가 왓슨 헬스 소속의 의사들과 이 부분을 논의했을 때도 대체로 이 점에 대해서 동감했다 [16, 28]. 하지만 왓슨에 대한 임상 시험을 진행한다고 하더라도 몇 가지 근본적인 문제가 있다는 점에 대해서도 역시 동의했다 [28].

무엇보다 왓슨 포 온콜로지의 의학적 효용을 증명하기 위해서 무엇을 기준으로 할 것인지가 애매하다. 임상 시험을 디자인하기 위해서는 정확성, 의학적 효용, 안전성 등을 판단할 명확한 기준이 필요하다. 임상시험을 진행하려면 일차 목적(primary outcome)과 이차 목적(secondary outcome)을 정의해야 한다. 예를 들어, 항암제에 대한 임상시험을 계획할 경우, 생존율(Survival Rate), 전체 생존 기간(OS, Overall Survival), 반응률(RR, Response Rate), 무진행 생존 기간(PFS, Progressive-free Survival), 독성(toxicity) 등의 지표를 해당 약의 효능을 평가하기 위한 기준으로 삼기도 한다.

그런데, 왓슨 포 온콜로지에 대해서 임상시험을 하려고 하면 무엇을 기준으로 해야 할까? 왓슨 헬스의 발표자료를 보면, 임상 시험 결과를 어떻게 평가할지에 대해서, 일치율(concordance), 의사 결정에 주는 영향(decision impact), 가이드라인 준수율(guideline adherence), 비용(cost), 시간 절감(time savings)과 함께 종양의 반응(tumor response), 생존율(survival) 등이 명시되어 있다. 이 중에서는 마지막에 언급되어 있는 종양의 반응과 생존율 정도가 의학적인 효용을 평가하기 위해서 그나마 유 의미해 보인다.

한 걸음 더 들어가서, 임상 시험을 어떤 식으로 디자인해야 할까? 엄격한 요건을 갖춘 임상시험이라면, 실험군과 대조군을 갖춰야 하며, 이중 맹검(double blind) 및 무작위(randomized)라는 조건 하에 전향적(prospective)으로 진행되어야 한다.

이러한 조건에 맞춰서 만약 다음과 같이 임상 시험을 진행한다고 가정해보자. 대조군은 종양내과 전문의 한 명이나, 혹은 특정 병원, 혹은 복수의 병원에 있는 종양내과 전문의들이 진료하는 그룹이다. 실험군은 의사가 전혀 관여하지 않고 왓슨 포 온콜리지의 '추천' 항목에 의해서만 치료하는 그룹이다. 암 환자를 각각 5,000명씩 전향적으로 모집하여 무작위로 양쪽 그룹에 배정한다. 이 경우에는 이중맹검은 어려울 것이므로, 환자만이라도 자신이 어느 그룹에 속했는지를 모르는 '단일 맹검(single blind)' 방식으로 해야 하겠다. 일차, 이차 목표는 5년 간의 생존율(OS)과 무진행 생존기간(PFS)으로 하도록 하자.

과연 이런 디자인의 임상 시험이 가능할까? 복잡하게 생각하지 않아도, 이러한 임상 시험은 몇 가지 심각한 문제가 있는 것이 명백하다.

첫 번째로 무엇보다 아직 정확성이나 효용이 검증되지 않은 왓슨 포 온콜로지만으로 의사의 개입 없이 실험군의 환자를 진료하는 것에는 의학적으로나 윤리적인 문제가 있다. 신약 임상 시험의 경우에는 전임상이나 임상 1상에서 후보물질의 독성 등 최소한의 안전성을 동물과 사람에서 검증한 후에, 2상에서 효능을 검증한다. 하지만 왓슨의 경우에는 그렇게 최소한의 안전성을 보장하기 위한 단계를 거치기 어려우므로, 일단 실험군의 환자를 전적으로 왓슨에게 맡기는 것은 적절하지 않다.

두 번째로 대조군에서 의사들의 실력이 매우 다양(heterogeneous)할 수 있다는 것이다. 개별 종양내과 의사를 비교하는 것은 당연히 의미가 없을 뿐만 아니라, 특정 병원의 종양내과 의사 전체, 혹은 여러 병원의 의사를 대상으로 한다고 해도 이 의사들 중에 실력, 경험, 치료 방침 등에 차이가 있을 수 있다. 이런 조건에서 나온 임상 결과를 다른 병원의 의사들이 참고하기는 어려울 가능성이 있다. 따라서 어떤 식으로든 '기존의 의료계 최선의 치료법'을 대표할 수 있는 보편적, 일반적인 기준을 마련해야 할 것이다.

세 번째로 왓슨이 계속 진화한다는 점이다. 이 문제 때문에 5년에 걸친 임상시험을 마무리하고, 그 결과를 몇 개월 혹은 몇 년 동안 정리하여 논문으로 발표하는 시점이 되면, 이미 논문 출판 시점의 왓슨은 임상 시험 당시의 왓슨이 아니게 된다. 즉, 과거에 수행한 임상 시험의 결과가 실제 환자에게 적용하는 현시점의 왓슨의 실력에 근거가 되기 어려운 것이다. 뿐만 아니라, 5년이라는 임상시험 기간 중에도 왓슨은 계속 바뀐다. 임상시험 시작 첫날의 왓슨과 마지막 날의 왓슨은 다를 것이며, 얼마나 다를 것인지의 정도 예측도 어렵다.

12. 왓슨의 검증을 위한 임상 시험

그러면 어떻게 해야 할까? 필자도 주위의 종양내과 전문의 들과도 논의해보았으나, 현실적인 어려움이 있다는 것에만 동의했을 뿐, 모두가 만족할만한 결론을 내리지는 못했다. 완벽하지는 않으나 다음과 같은 임상 시험 디자인이 필자가 구상할 수 있는 그나마 최선의 결과물인 것 같다.

일단 실험군과 대조군을 의사 vs 왓슨의 구도보다는 의사 vs 의사+왓슨으로 구성하는 것이 좋다고 본다. 양쪽 모두 의사가 기본적인 진료를 하기 때문에, 앞서 언급했던 실험군 환자에게 왓슨 포 온콜로지만으로 진료할 때의 윤리적인 문제나 안전성의 문제를 최소화할 수 있다. 또한 왓슨 포 온콜로지가 실제 진료 현장에서 사용될 때는 의사를 보조하는 방식으로 사용될 것이므로 이러한 디자인이 진료에 참고하기에 더 적절하다고 생각한다. 다만 이 경우에는 왓슨의 의

견을 어떤 경우에 어떻게 반영할 것인지에 대한 원칙도 정해져야 한다.

또한 실험군과 대조군에서 의사가 개입할 때 개별 의사나, 특정 병원의 의사가 각자 알아서 진료하는 것보다는 NCCN 가이드라인과 같은 표준화된 기준을 마련하는 것이 좋다고 본다. 사실 NCCN 가이드라인도 방대한 종류의 치료법을 담고 있고, 치료법을 뒷받침하는 근거 수준도 다양해서 '기존의 치료법'을 대표할 수 있는 일반적 기준이 될지에 대해서는 고민의 여지가 있다. 다만, 엄격한 임상시험을 위해서는 어떤 방식으로든 다수의 의사들이 공통된 기준을 바탕으로 진료하는 조건은 마련해야 할 것이다.

하지만 이러한 수정된 디자인의 임상 연구에도 여전히 해결되지 않는 문제가 있다. 바로 앞서 지적한 왓슨이 시간에 따라 계속 진화한다는 것이다. 실험군과 대조군을 조정하고, NCCN 가이드라인을 기반으로 하더라도, 역동적으로 변화하는 왓슨의 근본적인 속성은 임상 시험을 진행하고, 여기에서 나온 근거를 바탕으로 진료를 하기 위해 본질적인 한계를 부여한다.

13. 왓슨은 정말 마케팅용에 불과한가

현재 왓슨을 도입한 병원에서는 저마다의 방식으로 암 환자의 진료에 왓슨을 활용하는 것으로 알려져 있다. 예를 들어, 가천대 길병원의 경우에는 왓슨을 활용하는 다학제 진료실을 별도로 만들어 놓고 진료한다. 여러 진료과의 의료진이 좌우로 앉고, 가운데 화면에 환자의 검사 결과와 왓슨을 함께 띄워놓고 15-20분을 진료하는 방식이다.

혹자는 국내 병원이 왓슨을 도입한 이유를 단순히 '마케팅용'이라고 치부하기도 한다. 지방병원에서 더 많은 환자를 유치시키고, 수도권 병원으로 환자의 유출을 막기 위함이라는 것이다. 국내 한 대학병원에서 왓슨을 도입할 때 "지역 환자들은 수도권의 여러 병원을 찾아다닐 필요가 없어질 것"이라고 언급한 병원장의 코멘트는 이런 고민을 반영한다고 할 수 있다 [33].

IBM에 따르면 원래 왓슨이 만들어진 목적 중의 하나가 의료의 민주화(democratization)이다. 의료 전달 체계가 한국과 다른 미국에서는 1차 병원에서 시작하여, 2차 병원 등등을 차례대로 거쳐 마지막에 MD앤더슨이나 MSKCC와 같은 상급종합병원으로 가게 된다. 즉, 1, 2차 병원에 해당하는 지역 병원(community hospital)에서는 암 환자를 진료하는 인프라나 경험이 MSKCC에 비해 부족할 수밖에 없다. 이런 지역 병원에서 왓슨 포 온콜로지를 도입하면, 환자가 뉴욕의 MSKCC를 가지 않고서 지역 병원에서도 비슷한 수준의 진료를 받을 수 있을 것으로 기대하는 것이다. 하지만 1, 2차 병원의 추천을 받지 않아도 바로 상급종합병원으로 갈 수 있는 한국에서는 이런 왓슨이 조금 다른 목적으로 사용된다고 평가할 수 있다.

또한 의료계에서는 왓슨 포 온콜로지를 구글과 같이 단순히 치료법을 검색하는 역할에 그친다는 의견도 있다. 실제 왓슨을 현장에서 활용해본 의사들 중에도 이런 의견을 내기도 한다. 이런 의견에 대해서는 먼저 구글이 우리에게 의미가 없는 존재인지를 반문해볼 수도 있다. 세상의 모든 지식을 머리 속에 담고 있고, 새롭게 도출되는 지식도 모두 학습하고 있다면 구글이 의미 없는 존재일 수 있다. 하지만 인간은 그렇지 못하기 때문에 구글이 유용한 검색 엔진이 되는 것이다. 혹은 백과사전을 뒤져서 찾았을 지식을 구글은 몇 초만에 찾을 수 있게 해준다는 장점도 있다. 이는 의료 지식에도 적용될 수 있는 부분이다. 하지만 검색엔진의 검색 결과가 너무 부정확하거나, 사용자의 기대와 크게 차이가 난다면 가치가 없었을 것이다.

필자가 지적하고 싶은 것은 아직 왓슨 포 온콜로지에 대한 근거가 부족하다는 것이다. 현재의 왓슨이 얼마나 효과적인지, 정말로 단순히 마케팅 용도에 그치는지, 단순 검색에 불과해서 치료 개선에 영향이 미미한지를 판단할 수 있는 근거가 없다는 것이다. IBM은 발표자료에서 왓슨 포 온콜로지의 필요성 및 장점을 강조하기 위해서 현대 의학의 45% 이상의 의료 행위가 근거 없이 행해지고 있다는 연구를 인용한다 [34]. 하지만 그러한 왓슨 자체가 정확성과 효용성을 입증할 수 있는 근거가 부족하다는 것은 아이러니한 일이다.

때문에 병원이 왓슨을 단순히 마케팅용으로 도입했다고 폄훼하는 것도 필자는 무리가 있다고 생각한다. 반대로 왓슨이 실제로 환자의 치료에 도움이 된다고 주장할 수도 없다. 요는 양쪽의 주장의 타당성을 판단할 근거가 아직 충분하지 않다는 것이다. 의학은 과학이며, 과학적 주장은 논리와 데이터, 근거로 뒷받침되어야 한다. 하지만 아직까지 그 근거가 부족하다.

결 론

현재의 왓슨 포 온콜로지를 평가하기 위한 근거는 아직 부족하다. 근거가 부족하다 보니, 현재 진료 현장에서 환자에게 왓슨을 어떤 원칙에 기반하여 적용할지가 불명확하다. 예를 들어, 아래와 같은 부분에 대해서 의료계에서 합의된 원칙이 현재 전무하다.

- 어떤 환자의 경우에 왓슨 포 온콜로지의 의견을 물을 것인가?
- 왓슨 포 온콜로지를 (암종별로, 세부 암종별로) 얼마나 신뢰할 것인가?
- 왓슨 포 온콜로지의 결과를 환자에게도 공개해야 하는가? 혹은 의료진만 확인할 것인가?

- 왓슨 포 온콜로지의 판단과 의료진의 판단이 다른 경우에는 어떻게 할 것인가?
- 왓슨의 의견을 반영하여 치료한 결과가 좋지 않았다면, 그 책임의 일부라도 왓슨에게 물을 수 있나?

왓슨을 진료 현장에 적용할 때, 위와 같은 세부적인 사항을 어떻게 결정하는지에 따라, 의료의 질과 치료 효과, 진료 효율성 등등이 달라질 수 있다. 결국 문제는 원칙이 없다는 것이다. 원칙이 없기 때문에 현재 국내만 하더라도 왓슨을 도입한 병원이 저마다의 개별적인 기준에 맞춰 진료에 활용하고 있는 실정이다.

이는 결국 의료 인공지능이라는 새로운 종류의 솔루션이 진료에 영향을 줄 수 있음에도 불구하고 이를 어떻게 활용할지에 대한 의료계의 고민이 충분하지 못했다는 것을 의미한다. 결국 왓슨과 같은 인공지능 진료 보조 시스템을 활용하기 위한 표준 가이드라인이나 원칙을 합의하는 것이 필요하게 될 것이다. 그런 기준이 정해졌을 때에야 의료의 질 관리도 가능하다.

의료 인공지능이 의료 현장에 적용되는 것은 아직 초창기에 불과하지만, 앞으로 그 사례는 더욱 많아질 것이며, 파급 효과도 커질 것이다. 인공지능은 의료에 접목되는 완전히 새로운 방식의 수단이라고도 할 수 있지만, 그것이 목표로 하는 바는 기존의 의료가 추구해온 바와 다르지 않다. 질병을 더 효과적, 효율적으로 진단하고, 치료하고, 예방하고, 예측하며, 치료 과정의 부작용을 줄이는 것. 의료비를 낮추며, 더 나아가 의료진과 병원에도 도움이 될 수 있으면 좋을 것이다. 이를 위해서는 인공지능이라는 전례 없는 완전히 새롭고 혁신적인 방식이라고 하더라도, 의학에 활용되기 위해서는 근거와 원칙이 필요하다는 대전제는 동일하다.

REFERENCES

1. Do We Need Doctors Or Algorithms? [Internet] TechCrunch [cited 2017 July 20]. Available from: https://techcrunch.com/2012/01/10/doctors-or-algorithms/
2. Vinod Khosla says technology will replace 80 percent of doctors — sparks indignation [Internet] VentureBeat [cited 2017 July 20]. Available from: https://venturebeat.com/2012/09/02/vinod-khosla-says-technology-will-replace-80-percent-of-doctors-sparks-indignation/
3. What Silicon Valley Doesn't Understand About Medicine [Internet] Forbes [cited 2017 July 20]. Available from: https://www.forbes.com/sites/davidshaywitz/2011/06/17/what-silicon-valley-doesnt-understand-about-medicine/
4. Why I Disagree With Vinod Khosla About Digital Health -- And Hope He Succeeds Brilliantly [Internet] Forbes[cited 2017 July 20]. Available from: https://www.forbes.com/sites/davidshaywitz/2012/09/01/why-i-disagree-with-vinod-khosla-about-digital-health-and-hope-he-succeeds-brilliantly/
5. A.I. VERSUS M.D. [Internet] The New Yorker [cited 2017 July 20]. Available from: http://www.newyorker.com/magazine/2017/04/03/ai-versus-md
6. Jha S, Topol EJ. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. JAMA. 2016;316(22):2353-4.
7. Darcy AM, Louie AK, Roberts LW. Machine Learning and the Profession of Medicine. JAMA. 2016;315(6):551-2.
8. Miotto R, Li L, Kidd BA, Dudley JT, Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. 2016; 6: 26094.
9. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science 2014;347(6218):1254806.
10. How Watson for Clinical Trial Matching is Accelerating the Screening Process [Internet] IBM Blog [cited 2017 July 20]. Available from: https://www.ibm.com/blogs/think/2017/04/watson-health-screening/
11. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: A Convolutional Net for Medical Records. IEEE J Biomed Health Inform. 2017;21(1):22-30.
12. IBM's Watson Now A Second-Year Med Student [Internet] Forbes [cited 2017 July 20] https://www.forbes.com/sites/bruceupbin/2011/05/25/ibms-watson-now-a-second-year-med-student
13. IBM Watson Hits Daily Double Fighting Cancer With Memorial Sloan Kettering [Internet] Forbes [cited 2017 July 20]. Available from: https://www.forbes.com/sites/bruceupbin/2012/03/22/ibm-watson-hits-daily-double-fighting-cancer-with-memorial-sloan-kettering/
14. IBM Watson Hard At Work: New Breakthroughs Transform Quality Care for Patients [Internet] IBM News releases [cited 2017 July 20]. Available from: http://www-03.ibm.com/press/us/en/pressrelease/40335.wss?i=1360645029661
15. Gil Hospital adopts IBM WFO for the first time in South Korea. [Internet] IBM News Release (KR) [cited 2017 July 20]. Available from: http://www-03.ibm.com/press/kr/ko/pressrelease/50591.wss#release
16. IBM shares data on how Watson augments cancer treatment decision-making [Internet] Mobihealthnews [cited 2017 July 20]. Available from: http://www.mobihealthnews.com/content/ibm-shares-data-how-watson-augments-cancer-treatment-decision-making
17. How AI would innovate the future of medicine (2) [Internet] Healthcare Innovation Blog (KR) [cited 2017 July 20]. Available from: http://www.yoonsupchoi.com/2017/06/13/ai-medicine-2/
18. IBM's Watson Memorized the Entire 'Urban Dictionary,' Then His Overlords Had to Delete It [Internet] The Atlantic [cited 2017 July 20]. Available from: https://www.theatlantic.com/technology/archive/2013/01/ibms-watson-memorized-the-entire-urban-dictionary-then-his-overlords-had-to-delete-it/267047/

19. WFO is advanced medical textbook [Internet] Korean Doctors' Weekly (KR) [cited 2017 July 20]. Available from: http://www.docdocdoc.co.kr/news/articleView.html?idxno=220186
20. Validation study to assess performance of IBM cognitive computing system Watson for oncology with Manipal multidisciplinary tumour board [Internet] Oncology Pro [cited 2017 July 20]. Available from: http://oncologypro.esmo.org/Meeting-Resources/ESMO-Asia-2016-Congress/Validation-study-to-assess-performance-of-IBM-cognitive-computing-system-Watson-for-oncology-with-Manipal-multidisciplinary-tumour-board-for-1000-consecutive-cases-An-Indian-experience
21. Patients prefer opinion of WFO to doctors' opinion. [Internet] Chosun (KR) [cited 2017 July 20]. Available from: http://news.chosun.com/site/data/html_dir/2017/01/12/2017011200289.html
22. Assessing the performance of Watson for oncology, a decision support system, using actual contemporary clinical cases. [Internet] ASCO 2015 [cited 2017 July 20]. Available from: http://meetinglibrary.asco.org/record/113013/abstract
23. Steps in developing Watson for Oncology, a decision support system to assist physicians choosing first-line metastatic breast cancer (MBC) therapies: Improved performance with machine learning. [Internet] ASCO 2015 [cited 2017 July 20]. Available from: http://meetinglibrary.asco.org/record/113826/abstract
24. Integration of multi-modality treatment planning for early stage breast cancer (BC) into Watson for Oncology, a Decision Support System: Seeing the forest and the trees. [Internet] ASCO 2015 [cited 2017 July 20]. Available from: http://meetinglibrary.asco.org/record/112747/abstract
25. Concordance assessment of a cognitive computing system in Thailand. [Internet] ASCO 2017 [cited 2017 July 20]. Available from: http://abstracts.asco.org/199/AbstView_199_183143.html
26. Early experience with IBM Watson for Oncology (WFO) cognitive computing system for lung and colorectal cancer treatment. [Internet] ASCO 2017 [cited 2017 July 20]. Available from: http://abstracts.asco.org/199/AbstView_199_187558.html
27. Use of a cognitive computing system for treatment of colon and gastric cancer in South Korea. [Internet] ASCO 2017 [cited 2017 July 20]. Available from: http://abstracts.asco.org/199/AbstView_199_191277.html
28. Interview of Dr. Kyu Rhee, Chief Health Officer of IBM Watson Health [Internet] Healthcare Innovation Blog (KR) [cited 2017 July 20]. Available from: http://www.yoonsupchoi.com/2017/07/10/interview-with-kyu-rhee/
29. Fukuoka M, Yano S, Giaccone G, Tamura T, Nakagawa K, Douillard JY. Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small-cell lung cancer (The IDEAL 1 Trial) J Clin Oncol. 2003;21(12):2237-46.
30. Efficacy of Iressa was re-confiremd. [Internet] Medical Observer (KR) [cited 2017 July 20]. Available from: http://www.monews.co.kr/news/articleView.html?idxno=9534
31. Efficacy of cancer therapy is distinct according to ethnicity [Internet] Korean Doctors' Weekly (KR) [cited 2017 July 20]. Available from: http://www.docdocdoc.co.kr/news/articleView.html?idxno=77725
32. IBM Watson for Oncology Platform Shows High Degree of Concordance with Physician Recommendations [Internet] AACR [cited 2017 July 20]. Available from: http://www.aacr.org/Newsroom/Pages/News-Release-Detail.aspx?ItemID=983
33. Konyang University Hospital adopts WFO [Internet] Yonhap News (KR) [cited 2017 July 20]. Available from: http://www.yonhapnews.co.kr/society/2017/03/15/0706000000AKR20170315180700063.HTML
34. The Future of Health is Cognitive. [Internet] IBM Corporation [cited 2017 July 20]. Available from: http://sitesolutionssummit.com/wp-content/uploads/2016/10/Communication-2016-10-16-GlobalSiteSolutionsSummitfromKyuRhee.pdf