번역
Lab Med Online
Vol. 3, No. 4: 269-276, October 2013
http://dx.doi.org/10.3343/lmo.2013.3.4.269

진단면역학 LMO

진단 정확도를 넘어서: 진단의학검사의 임상적 유용성

Beyond Diagnostic Accuracy: The Clinical Utility of Diagnostic Tests

Patrick M.M. Bossuyt,[1]* Johannes B. Reitsma,[1,2] Kristian Linnet,[3] Karel G.M. Moons[2]

[1]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands; [2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands; [3]Section of Forensic Chemistry, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Like any other medical technology or intervention, diagnostic tests should be thoroughly evaluated before their introduction into daily practice. Increasingly, decision makers, physicians, and other users of diagnostic tests request more than simple measures of a test's analytical or technical performance and diagnostic accuracy; they would also like to see testing lead to health benefits. In this last article of our series, we introduce the notion of clinical utility, which expresses—preferably in a quantitative form—to what extent diagnostic testing improves health outcomes relative to the current best alternative, which could be some other form of testing or no testing at all. In most cases, diagnostic tests improve patient outcomes by providing information that can be used to identify patients who will benefit from helpful downstream management actions, such as effective treatment in individuals with positive test results and no treatment for those with negative results. We describe how comparative randomized clinical trials can be used to estimate clinical utility. We contrast the definition of clinical utility with that of the personal utility of tests and markers. We show how diagnostic accuracy can be linked to clinical utility through an appropriate definition of the target condition in diagnostic-accuracy studies.

서 론

진단의학검사와 표지자(marker)의 임상평가에서 진단정확도는 중요한 역할을 담당한다. 연재물[1-3]의 이전 문헌들에서 동일 환자에서 평가하고자 하는 검사나 표지자의 결과를 참고표준(reference standards)과 비교함으로서 정확도를 어떻게 정의하는지 설명하였다. 참고표준이란 질병의 유무, 또는 더 일반적으로는 해당 질환이나 상태를 확정하는 최선의 방법이다.

의사결정자, 임상의사 그리고 진단의학검사를 이용하는 다른 사용자들은 점점 검사자체의 기술적, 분석적인 성능과 진단정확도에 대한 평가 이상의 더 많은 정보를 요구하고 있다. 의료정책 입안자들은 제조사들이 좁은 의미의 기술적 또는 생물의학적 관점에서 벗어나 진단 기술들이 전형적인 환자군에서의 최종 치료경과를 향상시킬 수 있을지를 고려한 좀 더 확장된 관점으로 전향하기를 요구해왔다[4]. 의사결정자들과 사용자들은 진단의학검사와 표지자를 사용하도록 권고하기 전에, 그리고 급여여부를 결정하기 전에 그 검사가 관련 환자군에서 실제로 치료경과를 개선시킨다거나 의료의 질, 효율, 비용-효율을 향상시키는지에 대한 근거를 확인하고 싶어 한다.

우리들은 임상적 유용성의 개념을 소개하고 무작위 시험과 다른 연구 설계들이 진단의학검사들의 임상적 유용성을 평가하는데 어떻게 이용될 수 있는지 토의하고자 한다.

번역: 황상현
국립암센터 진단검사의학과
E-mail: mindcatch@ncc.re.kr

Received: May 31, 2013
Revision received: July 13, 2013
Accepted: July 19, 2013

This article is available from http://www.labmedonline.org

*본 원고는 양 잡지의 발행인 사이의 협약에 의하여 Clinical Chemistry에 실린 영문 논문을 번역하여 게재하는 것으로, 본 논문을 인용하고자 할 때는 다음과 같이 원 논문을 인용하여야 함. 원 논문의 저자 사사표기 및 기타 원고의 내용과 관련이 없는 부분은 번역 과정에서 생략하였음. 참고문헌 표기 방식은 원문 방식을 그대로 사용하였음.
원문 인용: Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem. 2012;58(12):1636-43.

임상적 유용성: health outcome에 대한 효과 (Clinical Utility: Effect on Health Outcomes)

지금까지 다수의 저자들이 진단의학검사의 사용을 지지하는 근거를 검토하기 위한 계통 체계와 평가에 대한 단계적 접근을 제안해왔다[5]. 진단정확도는 이러한 체계에서 언제나 중요한 부분을 차지하지만, 더 높은 수준의 요구들이 대두되고 있다. 진단의학검사의 유용성은 흔히 이러한 더 높은 수준으로 도입되거나 검사나 지표의 평가 단계에 있어 뒷편에 위치한다. 검사는 정보성이 있어야 할 뿐만 아니라, 실질적으로 유용한 정보를 제공해야 한다는 인식이 유전체학 및 대사체학 검사 시대에서 특히 증가하고 있다.

유용성이라는 개념은 보건의료평가에서 다소 모호하게 정의되어 왔다. 한편에서 보면, 임상의사들이 진단의학검사 결과에 대한 정보를 얼마나 진정으로 신뢰하는지, 또는 검사를 유용하게 느끼는지에 대한 척도이다. 흔히 인용되는 Fryback-Thornbury의 진단검사체계(hierarchy of diagnostic testing)는 영상의학분야에 처음 도입되었는데, 여기에서 이런 척도를 "진단적 사고 효능(diagnostic thinking efficacy)"(level 3)과 "치료적 효능(therapeutic efficacy)"(level 4)이라고 하였다[6]. 이러한 개념은 주관적이며 제한된 타당성을 가진다는 점에서 문제점으로 지적되어 왔다. 임상의사들이 검사 결과로부터 얻어지는 정보로 인해 진료행위 및 의사결정을 변경할 것이라고 선언할 때를 '의도하는' 행위(intended behavior)라고 하는데 이 '의도하는' 행위는 '실제' 행위(actual behavior)를 항상 반영하는 것은 아니다[5]. 더욱이, 판단이나 진료행위를 변경하는 것이 환자의 경과변화를 위한 필요조건은 될지 몰라도 충분조건은 아니다.

진단의학검사의 유용성은 보건의료에서 해당 검사를 실제로 이용함으로서 사망을 예방하거나 건강을 회복 또는 유지하는 등의 건강 결과(health outcome)가 바뀌는 것과 연관되어 있는 정도로 정의된다. 이러한 유용성은 Fryback-Thornbury의 체계에서 "환자경과 효능(patient outcome efficacy)" (level 5)으로 일컬어진다. 환자에게 중요한 경과의 실제적인 변화가 발생하는 정도를 분석함으로서 검사의 유용성을 평가해야 한다. 이는 무작위 대조시험이나다른 시험 등에서 특정 유형의 임상적 치료가 다른 유형보다 더 효과적임을 보여주는 환자들을 선택하는데 검사를 이용한다는 의미이다. 예를 들면, 폐색전증의 진단인데, 항응고제를 사용하는 것이사망률과 이환율을 감소시키는데 효과적임이 증명되었기 때문이다.

진단의학검사의 결정 기준을 질병경과의 변화로 정의하는 것은 검사를 다른 중재적 시술과 동일한 수준에 위치시키는 것이다. 근거중심의학의 정신에 비춰볼 때, 의약품은 시장에 도입되기 전에 안정성과 유효성이 보장되어야 한다; 가이드라인 개발자들이 약의 사용을 권고하기 전에 환자의 치료 결과에 대한 약의 효과를 주의깊게 분석하는 것은 당연하다; 우리들은 급여여부를 결정하기 위하여 비용과 비용효율을 고려해야 하는 것을 지지한다. 이와 유사하게 의사결정에 있어서 진단의학검사의 유용성을 강조하는 것은 진단의학검사들도 그런 수준에 위치시키는 것이다. 진단의학검사에 대한 결정은 보건의료의 다른 중재적 시술과 다르지 않다.

Fryback-Thornbury의 체계가 보여주듯이, "임상적 유용성" 과는 다른 용어들이 이러한 개념을 언급하는데 사용된다. 다른 저자들은 치료경과에 대한 효과를 기술할 때 "검사의 가치"로 언급하거나 "검사의 유용성" 또는 "검사의 이익"이나 "이득"으로 언급하기도 한다. 의사결정 분석에 있어서, 어떤 선택의 유용성은, 그 선택으로 인해 예상되는 결과(우연에 의한 것을 보정)의 정량적 표현이다. 보건의료의 경제적인 평가에서 예상되는 중재적 시술의 유용성은 종종 시술로 획득된 기대 수명으로 표현되거나 '삶의 질을 반영한 수명연장의 가치(quality-adjusted life year, QALY)로 표현된다. 치료결정과 그 결과에 대한 명확한 평가와의 관련성 때문에 우리는 "유용성"(utility)이라는 정량적인 용어를 더 선호한다. "임상적"(clinical)이라는 정성적인 부언은 아래에 설명할 것이다.

임상적 유용성: 특징 정의 (Clinical Utility: Defining Features)

우리는 진단의학검사의 유용성을 평가하는데 몇 가지 중요한 특징을 핵심적 요소들로부터 정의할 수 있다. 이러한 요소들을 Table 1에 요약하였다. 우선, 유용성의 정의는 보건의료의 일반적인 목적과 보조를 맞추어야 한다. 철학적 또는 정치적 쟁점으로 들어가지 않고도, 흔히 보건의료의 핵심 목적이라 함은 사람들이 기능적으로 건강을 유지하거나 다시 회복하도록 하는 것이다. 따라

Table 1. Key features of clinical utility

Elements of clinical utility	Explanation
Health outcomes	Health outcomes are outcomes that matter to patients and society: to prevent premature death, to restore or maintain functional health.
Strategy	Outcomes are generated not only by testing only but also by a management strategy that starts with testing but includes all downstream consequences of subsequent clinical management.
Probabilistic	Not all outcomes will be observed in everyone tested; evaluations will be made at the group level and expressed in terms of a distribution of outcomes.
Comparative	Utility is defined relative to a comparator strategy: current best standard practice.

http://dx.doi.org/10.3343/lmo.2013.3.4.269

서, 선별이나 모니터링을 위한 다른 유형의 검사도 포함하여 진단의학검사를 사용하는 기본적인 목적은 조기 사망 및 고통을 예방하고, 기능적 건강을 회복하는 것이어야 한다.

상당수의 검사들은 단순히 정보를 생성할 뿐, 그 정보 자체가 건강에 이득을 발생시키지는 않는다. 하지만 대부분의 경우에서, 진단의학검사를 통해 환자경과가 좋아지는 것은 검사결과가 후속조치를 안내함으로써 얻어지는데, 시행되는 검사결과에 따라 치료를 시작하거나, 변경하거나 중지하거나 보류하는 것이다. 적합한 환자를 대상으로 적합한 치료를 해야만 임상치료가 효과가 있는데, 검사는 가장 효과적인 치료를 선택하고 적합한 환자군을 찾는데 도움이 될 수 있다. 검사 후 경과가 개선되는 기전은 다양하며, 몇 가지 다른 기전을 포함한다[7].

임상적 유용성은 치료의 잠재적인 이득이라는 관점에서 정의될 뿐만 아니라, 검사나 표지자가 환자에 미치는 영향의 전 범위를 평가하는 것을 필요로 한다. 우리는 이러한 효과가 감정적, 사회적, 인지적, 그리고 행위적으로 어떤 영향을 미치는지를 다른 곳에서 기술하였다[7]. 진단의학검사로 인한 정서적인 영향이 가장 잘 연구되어 있는 부분은 다음과 같다. 즉, 검사를 시행하고 결과를 기다리면서 발생하는 걱정 및 스트레스에 관련된 연구가 여기에 해당한다. 사회적인 영향은 사회적 관계와 관련이 있다. 검사를 시행함으로써 낙인이 찍히거나 사회적으로 고립을 야기할 수 있다. 인지적 영향으로는, 검사결과 및 질병 상태에 대한 환자의 믿음, 자각, 이해 등이 있다. 행동학적 효과로는 건강 다이어트를 따라 하거나 운동이나 금연에 동참하는 등의 다른 건강행위에 관여되는 것 등이 해당한다.

따라서, 검사는 효과적인 임상치료가 없어도 임상적 유용성을 가질 수 있다. 예를 들어, 검사를 시행하여 결과가 양성(또는 음성)으로 판명되는 경우 불확실성이 조기에 해결되는 것이 해당한다. 또한, 진단-그리고 연관된 예후에 이르게 되는 검사는 환자가 질환에 대처하는 것을 도울 수 있고, 일상의 기능과 삶의 질을 향상시키는데 도움이 된다. 예를 들어, 알츠하이머질환을 진단함에 따라 그러한 이득이 환자의 배우자, 친구, 친인척으로 확대될 수 있다.

질환경과에 대한 이러한 영향은 긍정적이거나 부정적일 수 있고 의도되거나 의도되지 않을 수 있다. 또한 환자 자신과 직접적인 관련뿐만 아니라, 임상처치의 비용효율적 측면에서도 영향을 미칠 수 있다. DNA검사를 예를 들면, 건강행위에 맞춰진 프로그램의 효과를 증가시킬 수 있을 뿐만 아니라, 다른 한편으로, 행동변화가 효과적일 것이라는 환자의 기대를 또한 저하시킬 수 있다[8].

진단의학검사는 치료에 의한 주요 경과를 실제적으로 향상시키지 못하면서도 임상적으로 유용할 수 있다. 검사나 지표의 도입이 현재 표준적인 진단-치료 전략으로 얻어지는 건강의 성과와 비슷한 결과를 이루고 또한 더 쉽게 이루었다면, 임상적인 유용성이 있다고 볼 수 있다. 현실적으로는, 불필요한 추가적인 진단의학검사나 효과적이지 못한 치료와 관련된 비용부담으로부터 환자들이 보호될 것이다.

임상적 유용성의 추가적인 특징을 정의하자면 확률적인 특징을 빼놓을 수 없다. 진단의학검사 자체의 신뢰성과는 상관없이, 검사를 하는 것이 원하는 성과에 이르게 될 것이라고 확신할 수 없다. 검사방법 자체의 한계에서 기인할 수 있지만(모든 사람이 동일한 결과를 얻지 못하는), 검사결과 자체의 변이 때문에 발생할 수도 있다. 치료에서의 차이(동일한 결과를 보이는 모든 사람이 동일한 방식으로 치료를 받지 않는 것 같은)와 치료 반응에서의 차이(동일한 치료를 받는 모든 사람이 동일한 경과를 얻지 않는 것 같은)는 이러한 문제를 더욱 복잡하게 만든다. 기껏해야 검사의 이용을 고려하는 대상 집단에서 기대되는 결과로서 유용성을 기술할 수 밖에 없다. 다시 말해, 이러한 특성은 다른 유형의 근거중심의학과 크게 다르지 않다.

마지막으로 중요한 요소는 임상적 유용성의 정의는 상대적이라는 것이다. 진단을 위해 검사를 시행하는 전략의 유용성은 절대적인 개념으로 정의할 수 없다; 검사는 반드시 비교 대상의 유용성에 대해 상대적으로 평가되어야 하기 때문이다. 대부분의 경우, 비교 대상에 해당되는 전략은 환자에 적용되는 최상의 의료행위가 된다. 이 전략은 같은 목적으로 현재 시행중인 최상의 검사 전략이 될 수도 있고, 어떠한 유형의 검사도 전혀 사용하지 않는 것일 수 있다. 검사나 표지자의 유용성 평가에서, 우리는 검사가 원하는 결과를 도출하는지 여부를 평가할 뿐만 아니라, 검사를 했을 때의 결과와 최상의 대안을 사용했을 때의 결과를 비교한다. 대장암 집단선별에 근거한 대변잠혈검사의 유용성이란, 말하자면, 전혀 선별검사를 시행하지 않는 전략과 비교함으로서 평가된다[9]. 흉통을 주소로 입원한 환자에서 급성관상동맥증후군의 확진에 보조적으로 사용되는 CRP정밀검사의 유용성은 임상진단 및 기존의 심혈관 표지자에 의거한 표준진료지침과 비교하여 상대적으로만 평가될 수 있다.

진단의학검사의 임상적 유용성은 결과론적이면서 다분히 맥락적이다. 임상적 유용성이란 검사나 표지자 사용의 결과를 의미한다. 또한, 임상적 상황에 영향을 받는다. 치료옵션의 변경이나, 새롭고 다른 유형의 검사에 의해 진단의학검사의 유용성은 변하게 된다[10-13]. 동일한 고려사항이 지식의 진보에도 적용되는데, 이는 효과적인 치료에 대한 사고의 변화를 일으키고, 검사 기술 자체는 영향을 받지 않더라도 검사의 임상적 유용성에 영향을 미칠 수 있다. 새로운 중재적 시술들로 인해 이전의 유용한 진단의학검사가 사라질 수 있지만, 또 한편으로는 질병 경과를 향상시키는데 필요한 검사와 치료 전략으로 인해 새로운 검사가 흥미를 끌게 만들 수도 있다.

임상적 유용성은 흔히 진단의학검사의 검증이나 검정이라는 용어로 다루어지지만, 이러한 용어는 다소 일반적이고, 다소 다른 근거에 기초하고 있다. "검증"이라는 용어는 진단의학검사와 표지자 개발에서 상당히 다른 관점들을 두루 포괄하고 있다. 따라서, 우리들은 분석적 검증(analytical validation)과 임상적 검증(clinical validation)으로 구분한다. 임상적 검증은 검사결과가 임상적으로 의미 있다는 것을 보여주는 과정인데, 예를 들어 검사가 대상 환자군에서 질환이나 상태를 검출하거나 예측할 수 있는지를 확인하는 것이다[14].

검사의 "qualification"이란 검사결과가 기술된 사용범위 안에서 특정 해석을 하는데 신뢰할 수 있다는 결론을 내리는 과정을 의미한다[15]. 이러한 정의는 2004년 신약개발과 규정 검토에서 바이오마커 사용을 위한 미국 식품의약품안전청의 Critical Path Initiative에서 제시되었다. "Qualification"이라는 것은 의학적 검사나 표지자가 목적에 부합하는지를 평가하는 것을 의미한다. 예를 들어, 검사를 특별한 사용용도에 적용할 수 있는지를 확인하는 것 등이다.

개인적 유용성의 불필요함 (No Need for Personal Utility)

보건의료가 서양에서 계속 확대됨에 따라, 보건의료의 목적 또한 다양화되고, 저자들은 진단의학검사의 유용성이라는 정의도 확대되어야 한다고 항변해왔다. 더 새로운 유전 및 분자 표지자에 대한 열망에 대한 논의들이 특히 그러하다. 이러한 토론들로 인해 결국 "개인적 유용성(personal utility)"이라는 새로운 용어가 도입되기에 이르렀다. 이 용어는 다소 혼란스럽지만 사회적 유용성(social utility)으로도 가끔 언급된다[16, 17]. 하지만 우리들은 이런 새로운 용어가 불필요하고 잠재적으로 오해를 불러 일으킨다고 믿는다.

이러한 정의는 새로운 표지자나 검사의 다른 시각을 위해 도입되어 온 것으로 여겨지는데, 우선, 앞서 정의한 임상적 유용성이 없는 검사들을 정당화하는데 이용되고, 다음으로 이러한 검사를 처방하고 시행하는 것에 대한 결정이 임상적이라기 보다 개인적인 결정이라는 것이다.

초기의 일부 검사들은 임상적 유용성을 보이지 않지만 다른 이점 때문에 사용이 정당화되는 것 같다. 예를 들어 Huntington질환을 위한 검사는 효과적인 임상적 조치가 전혀 없지만 삶을 계획하고 임신출산에 대한 결정하는데 도움이 된다. 다른 검사들은 특정 질환 또는 상태의 기원이나 위험인자를 구분하는데 도움이 될 수 있다. 검사 결과들은 그로 인해 환자가 그들의 질병이나 질병의 위험을 보다 잘 이해할 수 있게 도울 수 있다는 것이다[18].

우리 관점에서는 검사를 시행하는 목적과 효과 모두 앞서 제시한 검사의 임상적 유용성이란 정의 안에 포함될 수 있다. 진단의학검사나 다른 의학적 검사를 사용하는 것에 대한 인지적 효과(의도 여부에 상관없이)도 임상적 유용성 평가에 반드시 포함되어야 한다. 따라서, 이러한 측면에서 다른 정의는 불필요하다. 하지만, 모든 검사들이 임상적 유용성이 있어야 하는 것은 아니다.

한편, 무엇이 유용할 지, 그리고 무엇이 유용하지 않을지를 결정하는 개인들의 권리라는 의견이 있다[16, 17]. 이들은 임상 의사들이 어느 한 사람의 최선의 이익이 무엇인지 알고 있다고 가정하는 것은 자체가 다소 가부장적이라고 주장한다[19]. 즉, 누군가가 표지자 검사 비용을 기꺼이 지불한다면, 환자는 그 검사를 받을 권리가 있다는 것이다. 그렇지만, 그런 검사를 시행하게 되면 사회적 비용 및 부정확하거나 왜곡된 이해, 오진, 잘못된 처치 등의 보건학적 결과를 초래할 수 있다.

유용성 평가 연구설계: 무작위시험 (Study Designs for Evaluating Utility: Randomized Trials)

Table 1에서 정의된 유용성의 요소들은 진단의학검사 및 다른 유형의 의학적 검사들의 유용성을 평가하기 위한 연구설계를 고려하고 선택하는 데에 이용될 수 있다. 치료 관리 전략을 개발할 때, 우리는 환자집단들을 아우르는 보다 광의의 건강결과를 살펴보고, 그러한 결과들과 지금까지 가장 최선의 대체 전략이 사용된 환자들의 결과들을 비교할 필요가 있다. 이러한 접근방식에서, 무작위시험은 진단의학검사의 유용성을 평가하는 탁월한 방식이다.

무작위임상시험은 항상 집단간 두 가지 이상의 전략을 비교한다. 전략들은 눈가림 및 객관적인 결과 측정과 위약이나 허위절차(sham procedures) 등을 사용하여 다른 많은 것들을 가능한 유사하게 유지하면서 동일한 결과에 초점을 맞춘다. 무작위임상시험은 다른 유형의 연구설계에 비해 삐뚤림으로부터 덜 영향을 받고 의사결정에 필요한 근거를 제공한다 Fig. 1은 검사에 대한 무작위 임상시험의 한 예를 보여준다.

Mueller 등[20]은 급성호흡곤란을 주소로 응급실을 방문한 환자들에 대하여 두 가지 진단적 전략을 비교하였다. 하나는 신속검사를 이용하여 B-type natriuretic peptide 농도를 측정하는 것이고, 다른 진단 전략은 통상적인 진단과정을 사용하는 것이다. 이 저널의 보고에 따르면, 720일째 모든 원인으로 인한 누적 사망률에서 차이가 없었다(37% 대 36%) [21].

정보력이 있으려면, 검사에 대한 무작위시험은 "검사 양성이면 치료하고, 음성이면 퇴원"하거나, "검사가 양성이면 추가 검사를 시행하고, 음성이면 퇴원한다"와 같은 검사결과와 특정 임상반응을 연결하는 잘 정의된 프로토콜이 있어야 한다. 이러한 프로토콜은 임상치료의 후속적 조치와 같은 유형의 효과에 대한 최상의 근

http://dx.doi.org/10.3343/lmo.2013.3.4.269

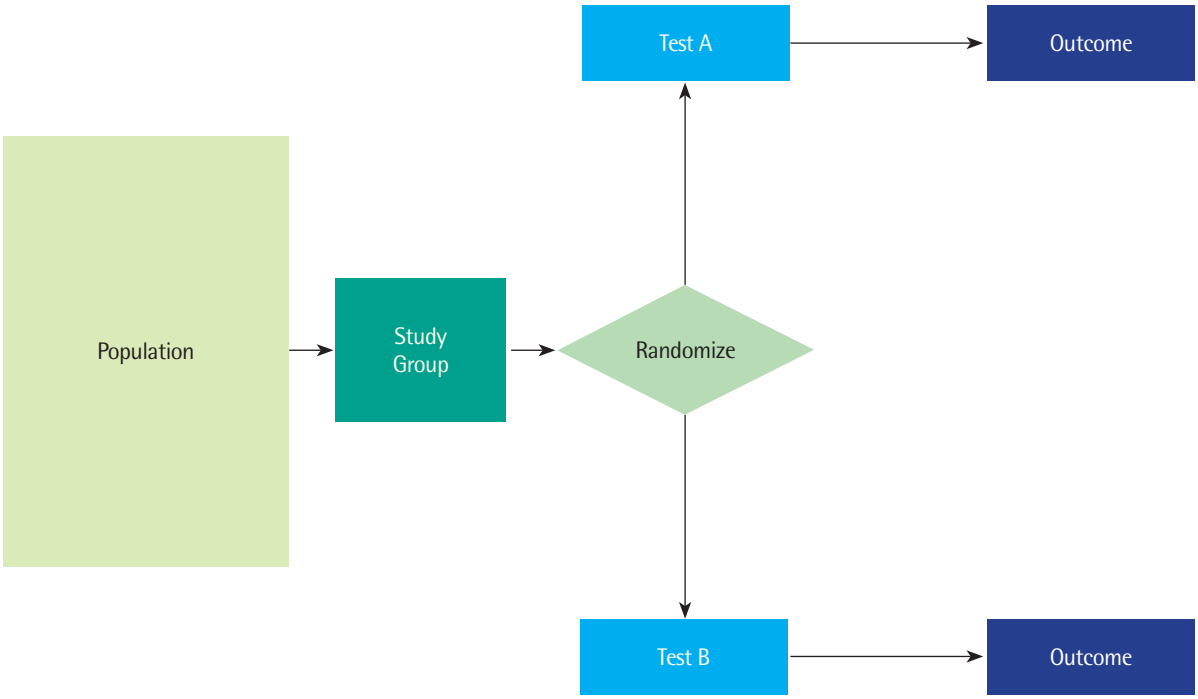

Fig. 1. A schematic representation of a randomized trial comparing 2 test strategies.

거-무작위시험이나 다른 강력한 연구설계에 기초해야 한다. 그러한 프로토콜이 없다면 검사에 대한 무작위시험은 정보력이 없거나 일반화하기 어렵다. 왜냐하면 결국 실제 치료행위에서 변수(variability)가 내재된 지역적인 의견에 좌우되기 때문이다[11].

진단의학검사 또는 다른 유형의 검사에 대한 임상시험을 설계할 때에는 몇 가지 추가적인 고려사항, 단점, 비효율을 고려해야 한다[10-13]. 검사에 대한 임상시험은 진단의학검사로 인한 결과를 평가해야 할 뿐만 아니라, 검사가 후속적 치료를 어떻게 안내하는지를 정의한 임상프로토콜에 통합한 총괄적인 검사전략도 평가할 수 있다. 그러므로, 검사의 효과는 검사 자체뿐만 아니라, 후속적 치료관리의 효과에 의해서도 달라지게 된다. 즉, 검사가 효과적인 관리와 결합되었을 때, 평범한 검사가 질환경과를 개선시키거나 비용-효과적이 되는 것도 가능한 결과가 될 수 있다. 이와 반대로, 우수한 검사가 효과적인 관리가 부재한 상황에서는 질환경과를 개선시키지 못할 수도 있다. 대변잠혈 선별검사는, 앞서 언급하였지만, 전자의 좋은 예이고, 알츠하이머질환에 대한 검사는 후자의 한 예이다.

의학적 검사 또는 표지자에 대한 임상시험은 또한 다소 비효율적일 수 있다[11]. 신약의 무작위 임상시험에서는 대조군의 모든 참가자는 위약을 받고, 실험군의 모든 참가자는 대상약물을 투여 받는다. 두 군 간 대비가 최대가 되지만 이는 실험군의 모든 참가자가 약물로부터 이익을 얻는 것을 의미하지 않는다. 예방 치료에 대한 임상시험에서는 단지 해당 사건의 위험에 있는 대상만이 중재적 시술로부터 이익을 얻을 가능성을 가진다.

검사나 표지자의 임상시험은 이와 반대로, 그러한 최대 대조가 이루어지지 않는다. 만약 환자의 경과가 임상관리와 의사결정을 내리는데 사용되는 검사결과에 따른다면, 불일치하는 결과를 보이는 환자만이 다르게 치료될 것이다. 예를 들어, 대장암 선별에서 대변잠혈검사와 같은 선별검사의 임상시험에서 선별검사가 추가적인 치료관리를 결정하는데 이용된다고 하자. 그 검사가 양성인 환자만-모든 연구참여자를 선별하는 것이 아닌- 대장내시경을 시행하게 된다. 이는 선별시험 및 진단의학검사를 포함한 의학적 검사에 대한 다른 임상시험은 일반적으로 많은 수의 연구 참여자가 필요하다는 것을 의미한다. 다수를 모집하는 것은 어려울 수 있다. 국립암연구소의 지원으로 수행된 Marker Validation for Erlotinib in Lung Cancer (MARVEL) 연구는 4년간 1,200명을 모집하도록 하였으나, 연구심의위원회에서 환자 등록이 승인된 300여 기관을 통하였으나, 잘 모집이 되지 않아 영구 중지되었다.

한편, 보다 효율적인 무작위임상시험도 가능하다. 그런 연구설계에는 두 검사를 비교하는데 있어서 불일치하는 결과를 보이는 대상만 무작위추출하거나, 단일검사를 평가하는데 있어서 서로 다른 두 치료전략을 적용해야 하는 검사에서 양성 결과를 보이는 대상만 무작위 추출하는 것과 같은 연구설계가 있다[11].

유용성 평가를 위한 연구설계: 모델들 (Study Designs for Evaluating Utility: Models)

진단의학검사의 무작위임상시험을 설계의 복잡성 때문에 일부 연구자들은 검사와 표지자의 유용성을 평가하는 다른 연구설계에 고개를 돌리게 된다. 한가지 방법은 모델링을 사용하는 것이다[13, 22, 23]. 의사 결정 모델은 다른 연구 설계로부터 얻어진 파라미터에 따른다. 가능한 관리 방법들로부터 모델을 수립하려는 연구자들은 검사와 다른 행위, 건강상태, 경과 간의 상호 관련성을 정의한다. 이 모델을 사용하여 연구자들은 평균적으로 예상되는 경과뿐만 아니라, 건강경과의 최종적인 가능성을 예측할 수 있다. Fig. 2는 검사 및 치료 전략, 모두 치료하는 전략, 모두 관망하는 전략 등 세 가지 전략에 대한 매우 단순한 결정 모델을 나타낸다.

이런 의사 결정 모델은 암 선별 프로그램을 평가하고 비교하는 데에 광범위하게 사용되었다[24]. 모델은 구축하는데 매우 복잡하고 시간이 소모될 수 있지만 다른 고비용의 임상시험을 대체할 수 있는 효과적인 방법을 제공할 수 있다. 하지만, 가정이라는 특성과 예상치 못한 부분을 대비하는 능력이 낮은 것 등이 단점들이다. Table 2는 무작위 시험과 모델 사이의 주요한 차이점을 요약하여 보여준다.

Table 2. Key differences between trials and models for evaluating the utility of medical tests.

Randomized trials	Models
Can compare only a few strategies	Can compare many strategies
Evaluates intended and unintended effects	Evaluates modeled effectsonly
No assumptions Expensive	Assumptions necessary Less costly
Time for follow-up needed	Model-building time only

대상 조건: 정확도를 유용성에 연결 (Target Condition: Linking Accuracy to Utility)

증가하는 의료비를 어떻게 관리할지, 그리고 이런 추세를 몰고 가는 새로운 기술들의 역할이 무엇인지에 대한 논의 한가운데, 새로운 진단의학검사를 임상진료에 도입하기 위해서는 표준화된 근거를 적용해야 한다는 논쟁이 있다. 무작위 대조시험이 비용이 지나치게 크고 상당한 시간적 지연이 발생한 후에야 답을 알 수 있는 상황이라면 지름길을 찾아야 한다. 만약 어떤 검사가 기존의 검사가 더 노동력이 많이 들어가는 검사와 동일한 결과를 보인다면 새

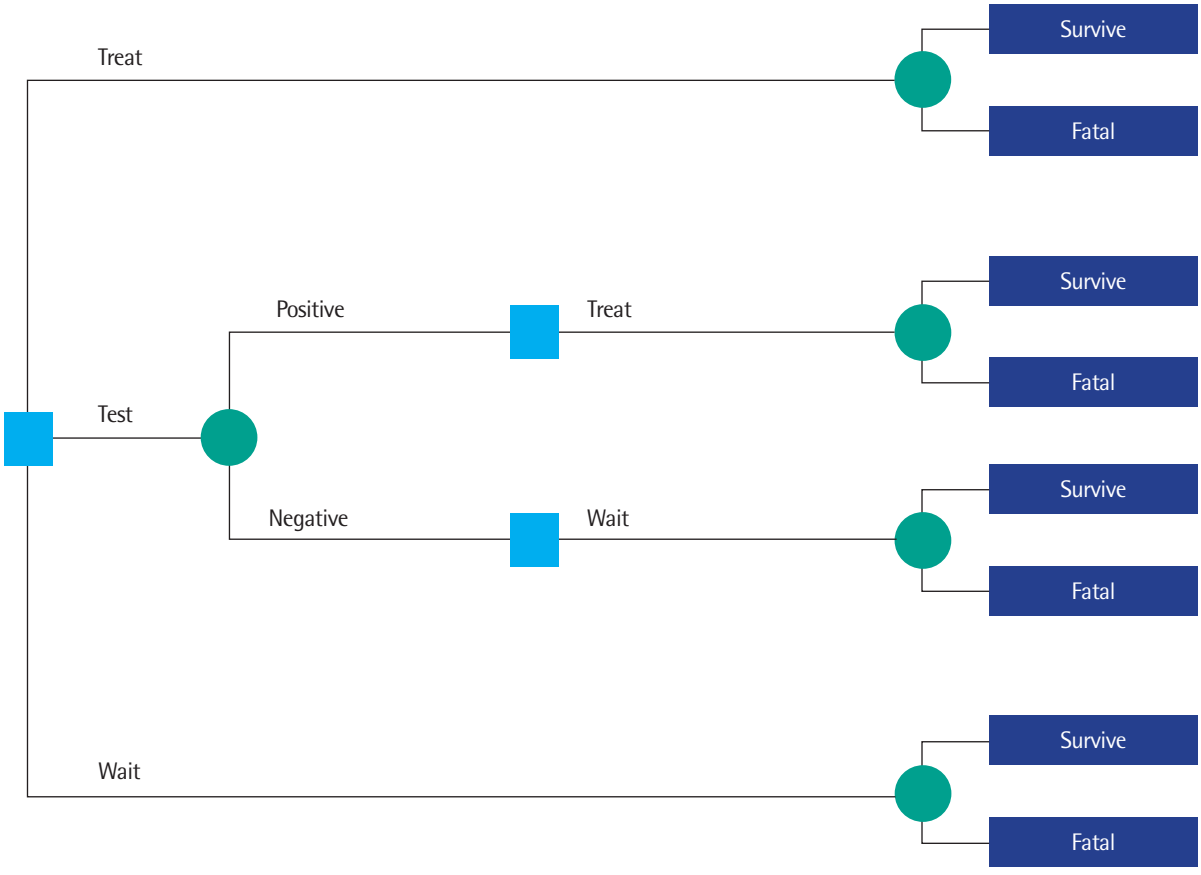

Fig. 2. A simple generic decision model to compare 3 strategies.

http://dx.doi.org/10.3343/lmo.2013.3.4.269

로운 검사가 환자의 예후에 영향을 주는지 아닌지에 대한 무작위 시험으로부터 근거를 굳이 만들 필요가 없을 것이다. 하지만, 진단의학검사의 임상적 유용성을 평가하는 무작위 임상시험의 필요성에 대한 완전한 토론은 이 논문의 범위를 벗어난다[11, 12, 23].

이번 연재물에서 이전 문헌에서 정의된 진단 정확도와 임상적 유용성이 어떻게 연관되어 있을까? 진단의학검사에서, 의심되는 질환을 가진 환자를 정확하게 확인하는 것은 임상적 유용성의 필수 조건이다[10]. 하지만, 질환을 확인하는 것이 건강에 도움이 되는 것과 동일하지는 않다. 중증도와 임상적 결과는 증례마다 그리고 치료 전략마다 다르고, 거의 동일하지 않다. 모든 유형의 질환이 순차적으로 진행되는 것도 아니고, 치료로 모두 호전되는 것은 아니다. 뿐만 아니라, 진단의학검사의 정확도는 이전 연재물에서 논의한 것처럼, 질환의 스펙트럼이나 검사를 사용하는 의학적 상황에 따라 차이가 난다[1].

따라서, 진단정확도를 평가하는 데에 단지 진단정확도를 표현하는 것보다 대상 조건의 정의를 포함하는 것이 권장된다[25]. 진단정확도는 최상의 기준이나 역치(threshold)에 근거하는 것을 권장한다. 단독검사에서는 검사 민감도는 적절한 치료를 받게 되는 환자의 비율을 나타내고, 검사 특이도는 추가적인 검사나 치료를 진행하지 않는 조건의 환자 비율을 나타낸다.

모든 검사의 성능평가에서 진단정확도를 고려하는 것이 도움이 되는 것은 아니다. 수명이 늘어나는 고령화 집단에서 검출률이 개선되면 만성화가 높아지고, 급성질환의 치료에 비해 만성질환의 관리가 늘어나게 된다. 검사와 표지자들은 흔히 진단 이외의 목적으로도 이용되는데, 치료선택, 치료모니터링, 용량조절, 부작용, 예후, 감시, 선별 등이 그 예이다. 이런 비진단적 목적에 대한 진단정확도를 어떻게 정의할 지에 대해서는 항상 명확하지 않을 뿐만 아니라, 임상민감도와 특이도에 대한 무조건적인 고정된 정의로 인해 새로운 표지자나 다른 의학적 검사의 임상적 유용성에 대한 의미 있는 평가가 제대로 이루어지지 못하게 될 수 있다.

맺음말

이 글에서 저자들은 비용을 명확히 고려하지 않은 임상적 유용성에 대해 토의하였다. 의사를 결정하거나 진단의학검사에 대한 권고를 고려할 때, 임상의사들과 다른 의사결정자들은 환자의 결과를 향상시키기 위해 어떤 자원이 필요할지도 함께 고려할 것이다. 따라서, 비용-효과 연구들은 임상적 유용성 연구에 반드시 후속적으로 이루어져야 하거나 동반연구로 진행되어야 할 것이다[22, 23].

저자들은 일상 진료용 검사들이 임상적 유용성에 대한 어떠한 근거 문헌도 없이 시행되고 있음을 알고 있다. 환자를 만족시키거나, 순수하게 재정적 이유나 일종의 방어진료와 같은 법적 책임 때문에 검사를 시행하는 경우도 있다. 진단의학검사들은 적절한 연구 데이터가 부족하여 해석이 되지 않을 수 있음을 알아도 시행될 수 있다.

하지만 요즘은 검사비용이 급여화되고 진료에 이용하기 전에 진단의학검사가 환자의 경과를 개선한다는 근거가 있어야 한다는 패러다임 변화에서 벗어날 수 없다. 향후, 그런 근거는 새로운 검사가 시장에 시판되기 전 승인을 받는데 반드시 필요하게 될 것이다. 새로운 기술에 관련된 혁신가, 개발자, 제조사는 새로운 표지자의 사용 용도를 확인하고 보건의료의 경과에 미치는 영향에 대한 근거를 창출해야만 할 것이다[26]. 흥미로운 시대임에 틀림없다.

요 약

다른 여타 의료 기술 또는 중재적 시술과 마찬가지로 진단의학검사들은 일상적인 진료에 도입되기 전에 충분히 평가되어야 한다. 검사 자체의 분석적 또는 기술적 성능과 진단정확도에 대한 평가 이상의 요구가 정책 결정자, 임상의사 그리고 진단의학검사를 이용하는 다른 사용자들로부터 점차 증가하고 있다. 뿐만 아니라, 검사를 시행하는 것이 결과적으로 건강에 이로움을 보고 싶어 한다. 연재물의 이번 마지막 원고에서, 저자들은 진단의학검사가 차선책 즉, 다른 유형의 검사를 시행하거나 검사를 아예 안 하는 것에 비해 건강 경과(health outcome)을 어느 정도 향상시키는지를 의미하는 임상적 유용성-흔히 정량적인 형태로 표현되는-을 소개하고자 한다. 대부분의 경우에서, 진단의학검사를 시행함으로써 양성 결과를 보이는 환자는 효과적인 치료를 하고, 음성 결과를 보이는 환자는 치료를 하지 않는 후속적 의료행위로부터 효과를 볼 환자를 확인하여 치료결과를 향상시킨다. 우리들은 무작위 비교 임상시험이 임상적 유용성을 평가하는데 어떻게 사용되는지 기술할 것이다. 우리들은 임상적 유용성의 정의와 진단의학검사와 표지자의 개인적 유용성(personal utility) 정의를 상호 대조하여 비교할 것이다. 우리들은 진단정확도 연구들에서 적합한 목표 조건을 정의함으로서 진단정확도가 어떻게 임상적 유용성과 연결되어 있는지를 보일 것이다.

REFERENCES

1. Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. Clin Chem 2012;58:1292-301.
2. Reitsma JB, Moons KG, Bossuyt PM, Linnet K. Systematic reviews of studies quantifying the accuracy of diagnostic tests and markers. Clin Chem 2012;58:1534-45.

3. Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PM. Quantifying the added value of a diagnostic test or marker. Clin Chem 2012;58:1408-17.
4. Neumann PJ and Tunis SR. Medicare and medical technology--the growing demand for relevant outcomes. N Engl J Med 2010;362:377-9.
5. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29:E13-21.
6. Gazelle GS, Kessler L, Lee DW, McGinn T, Menzin J, Neumann PJ, et al. A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research. Radiology 2011;261:692-8.
7. Bossuyt PM and McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009;29:E30-8.
8. Marteau TM and Weinman J. Self-regulation and the behavioural response to DNA risk information: a theoretical analysis and framework for future research. Soc Sci Med 2006;62:1360-8.
9. Heitman SJ, Hilsden RJ, Au F, Dowden S, Manns BJ. Colorectal cancer screening for average-risk North Americans: an economic evaluation. PLoS Med 2010;7:e1000370.
10. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144:850-5.
11. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844-7.
12. Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomization in diagnostic research. Ann Epidemiol 2006;16:540-4.
13. Hunink MG and Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. Radiology 2002;222:604-14.
14. Kelloff GJ and Sigman CC. Cancer biomarkers: selecting the right drug for the right patient. Nat Rev Drug Discov 2012;11:201-14.
15. Wagner JA, Williams SA, Webster CJ. Biomarkers and surrogate end points for fit-for-purpose development and regulatory evaluation of new drugs. Clin Pharmacol Ther 2007;81:104-7.
16. Foster MW, Mulvihill JJ, Sharp RR. Evaluating the utility of personal genomic information. Genet Med 2009;11:570-4.
17. Grosse SD and Khoury MJ. What is the clinical utility of genetic testing? Genet Med 2006;8:448-50.
18. Pletcher MJ and Pignone M. Evaluating the clinical utility of a biomarker: a review of methods for estimating health impact. Circulation 2011;123:1116-24.
19. Wizemann T, Berger AC, eds., Institute of Medicine (US) Roundtable on Translating Genomic-Based Research for Health. The value of genetic and genomic technologies: workshop summary. Washington, DC: National Academies of Press/Institute of Medicine; 2010.
20. Mueller C, Scholer A, Laule-Kilian K, Martina B, Schindler C, Buser P, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. N Engl J Med 2004;350:647-54.
21. Breidthardt T, Laule K, Strohmeyer AH, Schindler C, Meier S, Fischer M, et al. Medical and economic long-term effects of B-type natriuretic peptide testing in patients with acute dyspnea. Clin Chem 2007;53:1415-22.
22. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation 2009;119:2408-16.
23. Moons KG. Criteria for scientific evaluation of novel markers: a perspective. Clin Chem 2010;56:537-41.
24. Clarke LD, Plevritis SK, Boer R, Cronin KA, Feuer EJ. A comparative review of CISNET breast models used to analyze U.S. breast cancer incidence and mortality trends. J Natl Cancer Inst Monogr 2006;36:96-105.
25. Lord SJ, Staub LP, Bossuyt PM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. BMJ 2011;343:d4684.
26. Price CP and Christenson RH. Evaluating new diagnostic technologies: perspectives in the UK and US. Clin Chem 2008;54:1421-3.

http://dx.doi.org/10.3343/lmo.2013.3.4.269