Editorial

ISSN 2465-8243(Print) / ISSN: 2465-8510(Online)
https://doi.org/10.14777/uti.2020.15.1.16
Urogenit Tract Infect 2020;15(1):16-25

Korean Translation of the GRADE Series Published in the *BMJ*, 'GRADE: Grading Quality of Evidence and Strength of Recommendations for Diagnostic Tests and Strategies' (A Secondary Publication)

Translated by: **Jae Hung Jung[1,2], Do Kyung Kim[3], Ho Won Kang[4], Ja Yoon Ku[5], Hyun Jin Jung[6], Hong Wook Kim[7], Eu Chang Hwang[8]; Guideline Development Committee in the Korean Association of Urogenital Tract Infection and Inflammation**

[1]Department of Urology, [2]Institute of Evidence-Based Medicine, Yonsei University Wonju College of Medicine, Wonju, [3]Department of Urology, Soonchunhyang University Seoul Hospital, Soonchunhyang University College of Medicine, Seoul, [4]Department of Urology, Chungbuk National University College of Medicine, Cheongju, [5]Department of Urology, Pusan National University Hospital, Busan, [6]Department of Urology, Daegu Catholic University School of Medicine, Daegu, [7]Department of Urology, College of Medicine, Konyang University, Daejeon, [8]Department of Urology, Chonnam National University Hwasun Hospital, Chonnam National University Medical School, Hwasun, Korea

This article is the fourth translation of a GRADE series published in the BMJ, which graded the quality of evidence and strength of recommendations for diagnostic tests or strategies, as a comprehensive and transparent approach for developing recommendations. Randomized trials for diagnostic approaches represent the ideal study design for intervention studies. On the other hand, cross-sectional or cohort studies with a direct comparison of the test results with an appropriate reference standard can provide high-quality evidence. The guideline panel must be reminded that the test accuracy is a surrogate for patient-important outcomes, so such studies often provide a low quality of evidence for recommendations regarding diagnostic tests, even when the studies do not have serious limitations. Diagnostic accuracy studies showing that a diagnostic test or strategy improves important patient outcomes will require the availability of effective treatment, reduction of test-related adverse effects or anxiety, or improvement of the patients' well-being from prognostic information. Therefore, it is important to assess the directness of the test results regarding the consequences of diagnostic recommendations that are important to patients.

Received: 16 April, 2020
Revised: 16 April, 2020
Accepted: 16 April, 2020

Correspondence to: Eu Chang Hwang
https://orcid.org/0000-0002-2031-124X
Department of Urology, Chonnam National University Hwasun Hospital, Chonnam National University Medical School, 322 Seoyang-ro, Hwasun-eup, Hwasun 58128, Korea
Tel: +82-62-379-7747, Fax: +82-62-379-7745
E-mail: urohwang@gmail.com

This article is the secondary publication (complete translation in Korean) of the article originally published in the *BMJ* in English (Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. 2008;336:1106-10). The Editor-in-Chief of *Urogenital Tract Infection* decided to publish this secondary publication for the reader's sake, and it was approved by the *BMJ*. The *BMJ* Publishing Group takes no responsibility for the accuracy of the translation from the published English language original and is not liable for any errors that may occur.

GRADE: 진단검사(Tests and Strategies)의 근거수준(Quality of Evidence) 및 권고강도 Strength of Recommendation) 평가

중재(intervention) 효과의 근거수준 및 권고강도를 평가하기 위해 GRADE system을 이용한 것과 마찬가지로 진단검사에서 역시 GRADE를 이용할 수 있으며 이러한 과정에서 어떻게 환자에게 중요한 의료결과(patient-important outcome)를 고려하여 평가하는지에 대하여 본 논문은 설명하고자 합니다.

비록 진단 영역에 있어 권고는 중재에 대한 것과 마찬가지로 GRADE의 방법론적 기반을 공유하고 있으나 특유의 어려움(challenges)이 있는 것이 사실입니다. 권고의 기반으로서 진단 정확도(test accuracy)의 근거를 이용할 때 진료지침 개발 패널은 왜 주의가 필요한지 또 진단 정확도의 근거는 보통 근거수준 낮음(low quality of evidence)으로 평가되는지 저자들은 설명하고자 합니다.

1. 진단과정(Testing)은 치료과정(Patient Care)에 다양한 방식으로 기여합니다.

임상 의사는 증상, 영상검사, 그리고 생화학적 검사를 포함한 진단법을 신체의 생리학적 이상상태(derangement)를 평가하고, 질환을 진단하며, 환자의 상태를 추적하며, 예후를 예측하기 위해 이용합니다[1]. 본 논문은 질환(예, 결핵), 환자상태(target condition: 예, 철 결핍), 또는 증후군(예, 쿠싱 증후군) 유무를 판단하기 위한 진단검사에 집중하고자 합니다.

임상 의사는 보통 진단검사를 하나의 묶음(package)으로 또는 일련의 과정(strategy)으로 이용합니다. 예를 들면 수술이 가능한 폐암 환자에서, 임상 의사는 바로 가슴절개술(개흉술: thoracotomy)을 시행할 수도 있으며 또는 뇌, 뼈, 간, 부신의 영상학적 평가에 따른 추가적인 진단과정(strategy)으로 이용할 수도 있습니다. 그래서 우리는 평가(evaluating) 또는 권고(recommending)를 하나의 진단검사라기 보다는 진단과정으로 생각할 수 있습니다. 진단검사(test or strategy) 진료지침 개발 패널은 환자(P: patients), 진단검사(diagnostic intervention), 비교검사(C: comparison), 건강결과(O: outcome)를 반드시 구분하여야 합니다(Box) [2-5].

2. 진단 정확도는 환자에게 중요한 궁극적 의료결과(Patient-Important Outcome)의 대리결과(Surrogate Outcome)이다.

본 논문은 환자에게 중요한 의료결과에 대하여 진단검사가 미치는 영향에 대한 근거의 질을 평가할 수 있는 분석적 틀(framework)을 제공하고자 합니다. 보통 임상 의사는 진단검사를 시행할 때 검사가 질환을 가지고 있는지 없는지를 얼마나 잘 분류하는지를 의미하는 진단검사의 정확도(민감도 및 특이도)를 고려합니다. 그러나 기본적인 가정은 질환의 유무에 대한 정확한 진단이 건강결과에 어떤 영향을 미치는가에 대한 이해입니다. 이러한 가정을 설명하는 예로서, 수술이 가능한 폐암 환자에서 추가적인 영상검사는 가슴절개술의 이환율(morbidity)과 초기사망률(early mortality)을 고려할 때 무의미한 수술을 피할 수 있다는 것입니다. 심혈관질환 유무 평가를 위한 새로운 검사법(예, 고식적 혈관조영술에 대한 전산화 단층촬영)은 더 침습적이고 비싼 검사와 연관된 합병증을 줄일 수 있을 것입니다[6].

Box. 대체검사의 임상 시나리오

심혈관질환이 의심되는 환자에서, 심혈관에 대한 다중 나선식 전산화 단층촬영은 고식적 혈관조영술을 대체하는 진단법으로서 위음성과 연관된 심혈관 부작용을 낮추고 위양성과 연관된 불필요한 치료 및 그로 인한 부작용을 낮출 수 있는가?

진단과정 평가를 위한 가장 좋은 방법은 (특히 검사 정확도가 높은 새로운 진단과정인 경우) 무작위 대조군 연구로서 연구자는 중재 검사군과 비교검사군을 무작위 할당하여 사망률, 이환율, 증상 및 삶의 질을 평가하는 것입니다(Fig. 1) [7-12].

환자에게 중요한 의료결과에 대한 대체 진단과정을 평가할 수 있는 진단검사연구(이상적으로 무작위 대조군 연구이나 관찰연구를 포함하여)가 존재할 때, 임상진료지침 개발 패널들은 이전 본 논문의 시리즈에서 언급된 GRADE 방법론을 이용할 수 있습니다[13,14]. 그러나 그러한 연구가 존재하지 않는다면 진료지침 개발 패널들은 진단 정확도 연구를 이용할 수 밖에 없으며 환자에게 중요한 의료결과에 진단 정확도가 미치는 영향을 추론하여야만 합니다[15]. 중요한 임상 질문(key questions)은 위음성(false negative) 또는 위양성(false positive)의 감소 또는 이와 대응되는 진양성(true positive) 그리고 진음성(true negative)의 증가가, 진단검사 과정에 의해 환자가 얼마나 정확하게 분류되었는지, 질병군으로 분류되거나 비질병군으로 분류된 환자에서 어떤 건강결과가 발생하였는지 확인하는 것입니다(Table 1 and Appendix 1) [12].

3. 환자에게 중요한 의료결과 추론을 위한 간접적 근거(Indirect Evidence) 이용하기

환자에게 중요한 의료결과를 향상시키는 진단검사 또는 과정에 대한 근거로부터 추론은 효과적 중재의 이용가능성(availability)에 기반한다[1]. 대안으로, 심지어 효과적인 중재법이 없다고 하더라도 정확한 진단검사는 진단과 연관된 부작용 또는 불안을 낮출 수 있거나 또는 예후 정보제공을 통하여 환자에게 만족감(well-being)을 줄 수 있는 이점이 있을 수 있다.

예를 들면, 치료가 어려운 헌팅턴 무도병(Huntington's chorea)의 유전자 검사는 진단의 확진 외에도 추후 증상의 발현할 때에 대한 계획을 세우기 위해 도움이 될 수 있다. 이러한 장점은 효과적 중재를 받는 것과 유사할 수 있으며 조기 검진을 통한 미래의 질환을 예측하여 준비하는 것은 조기 진단의 위해(downsides)와 함께 고려되어야 합니다[16-18]. 이번에는 바람직한 결과(desirable)와 바람직하지 않은 결과(undesirable consequences)의 균형에 영향을 줄 수 있는 요소에 대하여 근거의 질 관점에서 설명하고자 합니다. 결과를 분류하는 단순한 방법인 2×2 table로서 진양성, 진음성, 위양성, 위음성을 이용하여 설명하고자 합니다.

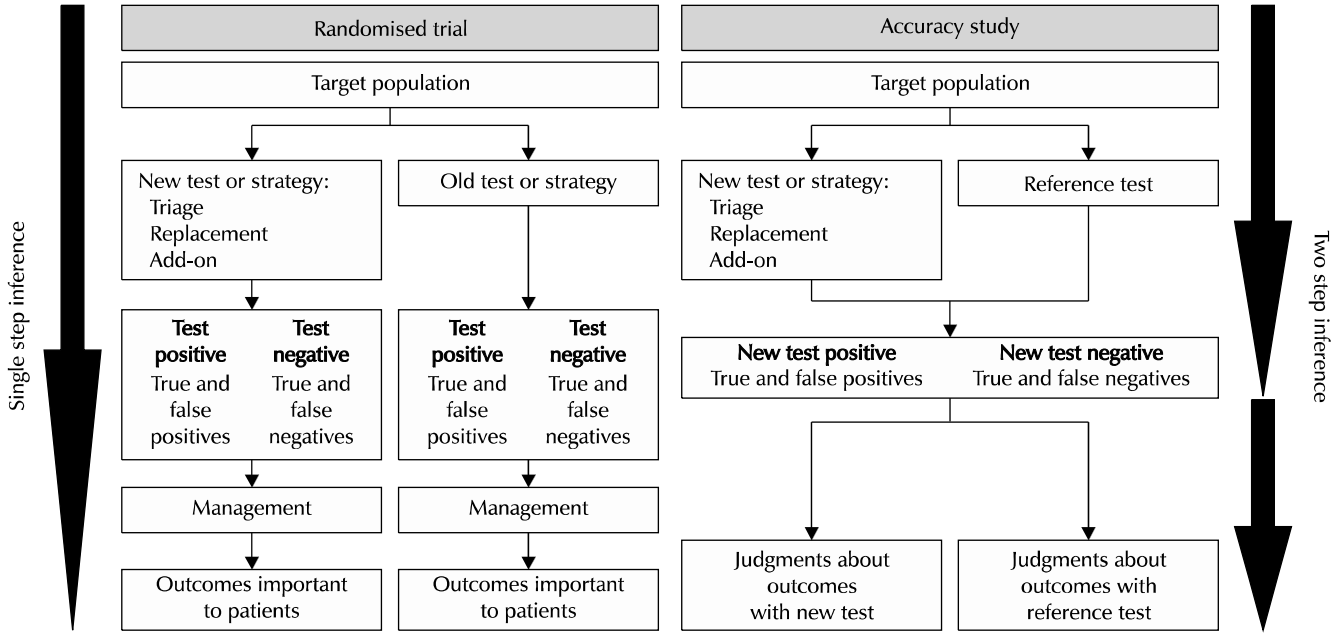

Fig. 1. Two generic ways in which a test or diagnostic strategy can be evaluated. On the left, the patients were randomized to a new test or strategy, or an old test or strategy. Those with a positive test result (cases detected) were randomized (or were previously randomized) to receive the best available management (second step of randomization for management not shown). The investigators evaluated and compared the patient-important outcomes in all patients in both groups. On the right, the patients received both a new test and a reference test (old or comparator test or strategy). The investigators could then calculate the accuracy of the test compared to the reference test (first step). To judge the importance of this information to patients, patients with a positive test (or strategy) in either group were (or have been in previous studies) submitted to treatment or no treatment; the investigators then evaluated and compared the important patient outcomes in all patients in both groups (second step). Adapted from the article of Schünemann et al. BMJ 2008;336:1106-10 [12].

4. 주어진 근거의 질에 대한 판정(Judgement)

1) 연구 비뚤림 위험(Study design and limitations [risk of bias])

GRADE 방법론은 환자에게 중요한 의료결과에 대한 진단검사과정의 신뢰도를 표현하는 근거수준을 4가지로 분류합니다[14]. Table 2는 GRADE 방법론이 근거수준을 어떻게 평가하는지에 대하여 보여줍니다. 무작위 대조군 연구는 진단검사 연구의 임상 권고(recommendation)를 만들기 위해 이상적 연구설계입니다. 그럼에도 불구하고, GRADE 방법론은 진단검사에 타당한 다른 연구설계 역시 근거수준 '높음'으로 평가합니다. 그러나 그런 연구설계는 비뚤림 위험에 취약하며 환자에게 중요한 의료결과에 대한 간접적 근거를 제시함으로서 종종 권고를 위한 낮은 근거수준을 제공합니다(Appendix 2) [12].

진단검사에 대한 타당한 연구설계는 타당한 진단적 불확실성이 존재하는 대표성 있고(representative) 연속적으로(consecutive) 모집된 환자군을 모집한 것으로 예를 들면 일반적 진료현장에서 임상 의사가 그 진단법을 적용하게되는 환자군을 의미한다. 만약 이러한 기준을 충족하지 못한다면, 예를 들어 중증 환자군과 건강한 대조군을 모집한다면, 진단정확도는 명백히 잘못 해석될 수 있을 것이다[19,20]. 타당한 연구는 우리가 연구하고자 하는 중재 검사와 적절한 참고표준검사(reference, 흔히 황금표준으로 불리는 [gold standard])를 비교하여야 한다. 모든 대상 환자군에서 그러한 비교를 시행하지 못한다면 비뚤림 위험은 높아진다. 만약 진단검사를 시행하는 또는 해석하는 연구자가 참고표준검사 또는 황금표준검사 결과를 먼저 인지하고 있다면 연구 비뚤림 위험은 역시 높아질 것이다. 진료지침 개발 패널들은 진단 정확도 연구의 비뚤림 평가를 위해 이미 개발된 도구를 사용하고 만약 심각한 비뚤림 위험(serious limitations)이 있다면 근거수준을 낮출 수 있다[21-25].

2) 직접성(Directness)

직접성을 평가하는 것은 진단검사에 대한 임상 권고를 개발하는 진료지침 개발 패널들에게 아마도 힘든 과정일 것입니다. 예를 들면, 새로운 검사가 낮은 위험도와 가격이 낮다면 하기 쉽겠지만, 위양성 및 위음성이 발생할 수 있습니다. 심혈관질환 진단을 위해 침습적인 혈관조영술을 전산화 단층촬영으로 대체하는 것을 생각해 봅시다(Tables 3, 4 and Appendices 3, 4) [5,12]. 진양성은 효과가 입증된 치료(약물요법, 혈관성형술[angioplasty] 및 스텐트, 혈관우회로 조성술[bypass surgery])를 시행할 수 있게 하며 진음성은 참고표준검사(여기서는 침습적 혈관조영술)의 부작용으로부터 환자를 보호할 수 있습니다. 반면에, 위양성은 명백한 불필요한 치료를 받는

Table 1. Examples and implications of different testing scenarios

Example of a new test and reference test or strategy	Putative benefit of new test	Diagnostic accuracy		Patients' outcomes and expected impact on management				Balance between presumed outcomes, test complications, and cost
		Sensitivity	Specificity	True positives	True negatives	False positives	False negatives	
Shorter version of the dementia test compared with the original mini-mental state exam for diagnosis of dementia	Simpler test, less time	Equal	Equal	Presumed influence on patient-important outcomes: Uncertain benefit from earlier diagnosis and treatment	Almost certain benefit from reassurance	Likely anxiety and possible morbidity from additional testing and treatment	Possible detriment from delayed diagnosis	Evidence of a shorter time and similar test accuracy (and thus patients' outcomes) would generally support the usefulness of new tests
				Directness of evidence (test results) for outcomes important to patients: Some uncertainty	No uncertainty	Some uncertainty	Major uncertainty	
Helical computed tomography for renal calculus compared to intravenous pyelogram (IVP)	Detection of more (but smaller) calculi	Greater	Equal	Presumed influence on patient-important outcomes: Certain benefit for larger stones; less clear benefit for smaller stones, and unnecessary treatment can result	Almost certain benefit from avoiding unnecessary tests	Likely detriment from unnecessary additional invasive tests	Likely detrimental for large stones; less certain for small stones, but possible detriment from unnecessary additional invasive tests for other potential causes of complaints	Fewer complications and downsides compared with IVP would support the usefulness of the new tests, but the balance between desirable and undesirable effects is unclear in view of uncertain consequences of identifying smaller stones
				Directness of evidence (test results) for outcomes important to patients: Some uncertainty	No uncertainty	No uncertainty	Major uncertainty	
Computed tomography for coronary artery disease compared to coronary angiography	Less invasive testing, but misses some cases	Slightly less	Less	Presumed influence on patient-important outcomes: Benefit from treatment and fewer complications	Benefit from reassurance and fewer complications	Harm from unnecessary treatment	Detriment from delayed diagnosis or myocardial insult	Undesirable consequences of more false positives and false negatives with computed tomography are unacceptable despite the higher rate of rare complications (infarction and death) and higher cost of angiography
				Directness of evidence (test results) for outcomes important to patients: No uncertainty	No uncertainty	No uncertainty	Some uncertainty	

Adapted from the article of Schünemann et al. BMJ 2008;336:1106-10 [12].
See Appendix 1 (complete translate in Korean).

부작용을 겪을 수 있으며 위음성은 심혈관계 질환의 발생 위험을 낮출 수 있는 치료를 받지 못하게 할 수 있습니다.

따라서 위양성과 위음성을 낮추는 것이 환자에게 도움이 될 것은 자명합니다. 진단에 불분명한 검사 결과의 영향은 다소 불분명하지만 명백히 바람직하지 않은 결과를 초래할 것입니다. 더군다나 비록 드물게 발생하기는 하나 침습적 혈관조영술의 부작용인 경색 또는 사망은 의심할 여지없이 중요할 것입니다. 임상진료지침 개발 패널들은 진단검사의 바람직한 결과와 바람직하지 않은 효과를 비교할 때 반드시 환자에게 이러한 결과들의 중요성을 고려하여야 합니다. 예를 들어 심혈관질환 위험도가 낮은 환자군에서 전산화 단층촬영은 많은 수의 위양성을 초래하여 불필요한 불안 및 추가적 검사를 야기할 수 있습니다(Table 4) [12]. 또한 전산화 단층촬영은 심혈관질환이 있는 환자의 1% (위음성)를 발견하지 못할 수 있습니다.

진단 관련 질문에 대하여 임상진료지침 개발 패널들은 다른 중재법에 대한 임상진료지침개발과정에서 마주치게 되는 비직접성과 관련된 일련의 문제들을 마주할 수 있습니다[2]. 진단 정확도는 환자집단에 따라 다르므로, 패널들은 연구에서 시행된 새로운 중재 검사, 참고표준검사 및 환자군이 권고가 사용되는 의료 환경(setting)과 환자집단에 비교할 만한 것인지 고려하여야 합니다.

마지막으로 두 가지 이상의 새로운 대체검사 및 진단과정을 평가할 때, 패널들은 이러한 진단과정이 참고표준검사와 직접

Table 2. Factors that decrease the quality of evidence for studies of diagnostic accuracy and how they differ from evidence for other interventions

Factors that determine and can decrease the quality of evidence	Explanations and differences from the quality of evidence for other interventions
Study design	Different criteria for accuracy studies—Cross-sectional or cohort studies in patients with diagnostic uncertainty and a direct comparison of the test results with an appropriate reference standard are considered high quality and can move to moderate, low, or very low depending on other factors
Limitations (risk of bias)	Different criteria for accuracy studies—Consecutive patients should be recruited as a single cohort and not classified by disease state, and selection, as well as the referral process, should be described clearly.tests should be done in all patients in the same patient population for the new test and well-described reference standard; the evaluators should be blind to the results of the alternative test and reference standard
Indirectness:	
Outcomes	Similar criteria—Panels assessing diagnostic tests often face an absence of direct evidence regarding the impact on patient-important outcomes. They must make deductions from studies of diagnostic tests about the balance between the presumed influences on the patient-important outcomes of any differences in true and false positives and true and false negatives in relation to complications and costs of the test. Therefore, accuracy studies typically provide low-quality evidence for making recommendations owing to the indirectness of the outcomes, similar to surrogate outcomes for treatments
Patient populations, diagnostic test, comparison test, and indirect comparisons	Similar criteria—Quality of evidence can be reduced if important differences exist between the populations studied and those for whom the recommendation is intended (in previous testing, spectrum of disease or comorbidity); if important differences exist in the tests studied and diagnostic expertise of people applying them in studies compared to settings for which recommendations are intended; or if the tests being compared are each compared with a reference (gold) standard in different studies and not directly compared in same studies
Important inconsistency in study results	Similar criteria—For accuracy studies, unexplained inconsistency in sensitivity, specificity, or likelihood ratios (rather than relative risk or mean differences) can reduce the quality of evidence
Imprecise evidence	Similar criteria—For accuracy studies, wide confidence intervals for the estimates of test accuracy or true and false positive and negative rates can reduce the quality of evidence
High probability of publication bias	Similar criteria—High risk of publication bias (for example, evidence from small studies for new intervention or test, or asymmetry in the funnel plot) can lower the quality of evidence

Adapted from the article of Schünemann et al. BMJ 2008;336:1106-10 [12].
See Appendix 2 (complete translate in Korean).

Table 3. Key findings of diagnostic accuracy studies—should multislice spiral computed tomography rather than conventional coronary angiography[a] be used to diagnose coronary artery disease in a population with a low (20%) pre-test probability?

Measure	Test finding (95% confidence Interval)
Pooled sensitivity	0.96 (0.94 to 0.98)
Pooled specificity	0.74 (0.065 to 0.84)
Positive likelihood ratio[b]	5.4 (3.4 to 8.3)
Negative likelihood ratio[b]	0.05 (0.03 to 0.09)

[a]Assuming that the reference standard, angiography, does not yield false positives or false negatives. [b]Average likelihood ratios from Hamon et al. [5]. Adapted from the article of Schünemann et al. BMJ 2008; 336:1106-10 [12].
See Appendix 3 (complete translate in Korean).

적으로 (동일 연구에서) 또는 비적접적으로 (서로 다른 연구에서) 비교가 가능한지를 고려하여야 한다[26-28].

연구 근거수준 최종결과

Table 5는 침습적 혈관조영술의 대체 진단요법으로 전산화 단층촬영의 근거수준 및 근거표를 보여줍니다. 환자에게 중요한 의료결과에 대한 검사 결과(진양성, 위양성, 진음성) 근거의 직접성의 불확실성은 거의 보이지 않습니다(Table 1) [12]. 그러나, 검사 정확도에서의 제한점으로 위음성이 환자에게 중요한 의료결과에 해로운 결과를 초래하는 정도에 대한 불확실성은 근거수준을 '높음'에서 '중등도'로 낮추는 요인이 되었습니다(Table 5 and Appendix 5) [12]. 각각의 연구 간에 설명할 수 없는 이질성(heterogeneity)은 모든 의료결과에 대한 근거수준을 추가적으로 낮추는 요인이 되었습니다. 환자에게 중요한 의료결과에 대한 위음성의 영향(추론을 통한)에 대한 불확실성은 근거수준을 '높음'에서 '낮음'으로 낮추는 요인이 되었습니다(Table 1) [12].

권고 만들기

진단검사 부작용과 관련한 진양성, 진음성, 위양성 및 위음성 결과가 초래하는 환자에게 중요한 의료결과의 차이는 임상진료지침 개발 패널이 진단검사를 적용하거나 적용하지 않는 권고를 결정하는 요인입니다[13]. 권고강도(strength of a recommendation)에 영향을 미치는 다른 요인으로는 근거수준, 진단검사 그리고 환자에게 중요한 의료결과와 관련된 가치와 선호도의 불확실성(uncertainty), 그리고 비용입니다.

심혈관 전산화 단층촬영은 침습적 혈관조영술이 초래할 수 있는 심근경색 및 사망을 피할 수 있게 합니다. 그러나 이러한 부작용의 발생은 매우 드뭅니다. 따라서, 혈관조영술

Table 4. Consequences of the key findings of diagnostic accuracy studies—should multislice spiral computed tomography rather than conventional coronary angiography[a] be used to diagnose coronary artery disease in a population with a low (20%) pre-test probability?

Consequences	No. per 1,000 patients	Importance[b]
True-positive results[c]	192	8
True-negative results[d]	592	8
False-positive results[e]	208	7
False-negative results[f]	8	9
Inconclusive results[g,i]	–	5
Complications[h,i]	–	5
Cost[i]	–	5

All results given per 1,000 patients tested for the prevalence of 20% and likelihood ratios shown in Table 3.
[a]Assuming that the reference standard, angiography, does not yield false positives or false negatives. [b]On a 9 point scale, GRADE recommends classifying these outcomes as not important (score 1-3), important (4-6), and critical (7-9) to a decision. [c]Important because it mandates drugs, angioplasty, stents, and bypass surgery. [d]Important because it spares patients unnecessary interventions associated with adverse effects. [e]Important because patients are exposed to unnecessary potential adverse effects from drugs and invasive procedures. [f]Important because of the increased risk of coronary events as a result of patients not receiving efficacious treatment. [g]Uninterpretable, indeterminate, or intermediate test results: important because they generate anxiety, uncertainty as to how to proceed, further testing, and possible negative consequences of either treating or not treating. [h]Not reliably reported, important because although rare, they can be serious. [i]Although the data for these consequences are not reported for simplicity, or because they are not known precisely based on the available data, they are important.
Adapted from the article of Schünemann et al. BMJ 2008;336:1106-10 [12].
See Appendix 4 (complete translate in Korean).

을 대체하는 진단방법으로 심혈관 전산화 단층촬영을 평가하는 임상진료지침 개발 패널들은, 적은 비용에도 불구하고, 침습적 혈관조영술을 대체하는 진단검사로서 전산화 단층촬영을 사용하지 않도록(agianst computed tomography) 약한 권고를 만들 수 있습니다. 이 권고는 많은 수의 위양성 및 위음성(효과적으로 치료될 수 있는 심혈관질환 환자를 놓칠 수 있는) 결과에 기반하고 있습니다. 또한 이 권고는 새로운 검사법에 대한 낮은 근거수준과 환자의 가치와 선호도에 기반하고 있습니다. 비침습적이고 부작용이 적은 검사에 대한 일반적 선호에도 불구하고 대다수의 환자들은 위양성 및 위음성이 초래하는 위험도를 고려할 때 좀 더 침습적인 검사(혈관조영술)를 선호할 것입니다.

결론

다른 중재(치료법)에 대한 권고처럼, 진단검사에 대한 근거수준 및 권고강도 평가를 위한 GRADE 방법론은 임상진료권고를 만드는 데 있어 포괄적이고 투명한 방법을 제공할 것입니다. 검사 결과는 환자에게 중요한 의료결과에 대한 대리지표임을 인지하는 것은 본 방법론의 핵심입니다. 이 방법론의 적용하기 위하여 임상 의사는 진단검사의 정확도가 어떠하든지 간에 그 결과가 환자의 의료결과의 개선을 가져오는지를 명확히 인식하는 것이 필요합니다.

Table 5. Quality assessment of diagnostic accuracy studies—example: should multislice spiral computed tomography be used instead of conventional coronary angiography for the diagnosis of coronary artery disease?[a]

No of studies	Design	Limitations	Indirectness	Inconsistency	Imprecise data	Publication bias	Quality
True positives (patients with coronary artery disease)							
21 studies (1,570 patients)	Cross-sectional studies[b]	No serious limitations	Little or no uncertainty	Serious inconsistency[d]	No serious imprecision	Unlikely[e]	⊕⊕⊕◯ Moderate
True negatives (patients without coronary artery disease)							
21 studies (1,570 patients)	Cross-sectional studies[b]	No serious limitations	Little or no uncertainty	Serious inconsistency[d]	No serious imprecision	Unlikely[e]	⊕⊕⊕◯ Moderate
False positives (patients incorrectly classified as having coronary artery disease)							
21 studies (1,570 patients)	Cross-sectional studies[b]	No serious limitations	Little or no uncertainty	Serious inconsistency[d]	No serious imprecision	Unlikely[e]	⊕⊕⊕◯ Moderate
False positives (patients incorrectly classified as having coronary artery disease)							
21 studies (1,570 patients)	Cross-sectional studies[b]	No serious limitations	Some uncertainty[c]	Serious inconsistency[d]	No serious imprecision	Unlikely[e]	⊕⊕◯◯ Low

[a]Full quality assessment would include a row for the outcomes important to patients associated with each possible test result (true positive, true negative, false positive, false negative, and inconclusive) as well as complications and costs of the test (see table 3); simplified summary of the quality of evidence for critical outcomes presented here. [b]All patients were selected to have conventional coronary angiography and generally presented with a high probability of coronary artery disease (median prevalence in included studies 63.5%, range 6.6-100%). [c]Some uncertainty about directness for false negatives related to the detrimental effects of delayed diagnosis or myocardial insult, reducing the quality of evidence for consequences of false-negative test results from high to moderate. [d]Statistically significant, unexplained heterogeneity of results for sensitivity (proportion of patients with positive coronary angiography with positive computed tomography scan), specificity (proportion of patients with negative coronary angiography with negative computed tomography scan), likelihood ratios, and diagnostic odds ratios, reducing the quality of evidence for consequences of true positive, true negative, and false-positive results from high to moderate and of false-negative results from moderate to low. [e]Possibility of publication bias not excluded but not considered sufficient to downgrade the quality of evidence.
Adapted from the article of Schünemann et al. BMJ 2008;336:1106-10 [12].
See Appendix 5 (complete translate in Korean).

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

AUTHOR CONTRIBUTIONS

J.H.J.: contacting BMJ editorial office to get the approval, translating the article, and drafting the manuscript, D.K.K, J.Y.K., H.J.J., and H.W.K.: helping to translate and draft the manuscript, E.C.H.: helping to translate and draft the manuscript, and final approval.

ORCID

Jae Hung Jung, https://orcid.org/0000-0002-4990-7098
Do Kyung Kim, https://orcid.org/0000-0002-3696-8756
Ho Won Kang, https://orcid.org/0000-0002-8164-4427
Ja Yoon Ku, https://orcid.org/0000-0003-3460-9386
Hyun Jin Jung, https://orcid.org/0000-0002-1895-7180
Hong Wook Kim, https://orcid.org/0000-0002-3847-1401
Eu Chang Hwang, https://orcid.org/0000-0002-2031-124X

REFERENCES

1. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001; 323:157-62.
2. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. CMAJ 1988;138:697-703.
3. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. J Gen Intern Med 1989;4:288-95.
4. Guyatt G, Montori V, Devereaux PJ, Schünemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. ACP J Club 2004;140:A11-2.
5. Hamon M, Biondi-Zoccai GG, Malagutti P, Agostoni P, Morello R, Valgimigli M, et al. Diagnostic performance of multislice spiral computed tomography of coronary arteries as compared with conventional invasive coronary angiography: a meta-analysis. J Am Coll Cardiol 2006;48:1896-910.
6. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089-92.
7. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844-7.
8. 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.
9. Moe GW, Howlett J, Januzzi JL, Zowall H; Canadian Multicenter Improved Management of Patients With Congestive Heart Failure (IMPROVE-CHF) Study Investigators. N-terminal pro-B-type natriuretic peptide testing improves the management of patients with suspected acute heart failure: primary results of the Canadian prospective randomized multicenter IMPROVE-CHF study. Circulation 2007;115:3103-10.
10. Worster A, Preyra I, Weaver B, Haines T. The accuracy of non-contrast helical computed tomography versus intravenous pyelography in the diagnosis of suspected acute urolithiasis: a meta-analysis. Ann Emerg Med 2002;40:280-6.
11. Worster A, Haines T. Does replacing intravenous pyelography with noncontrast helical computed tomography benefit patients with suspected acute urolithiasis? Can Assoc Radiol J 2002;53: 144-8.
12. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al.; GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106-10.
13. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al.; GRADE Working Group. Going from evidence to recommendations. BMJ 2008;336:1049-51.
14. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ; GRADE Working Group. What is "quality of evidence" and why is it important to clinicians? BMJ 2008;336: 995-8.
15. 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.
16. Maat-Kievit A, Vegter-van der Vlis M, Zoeteweij M, Losekoot M, van Haeringen A, Roos R. Paradox of a better test for Huntington's disease. J Neurol Neurosurg Psychiatry 2000;69:579-83.
17. Walker FO. Huntington's disease. Semin Neurol 2007;27:143-50.
18. Almqvist EW, Brinkman RR, Wiggins S, Hayden MR; Canadian Collaborative Study of Predictive Testing. Psychological consequences and predictors of adverse events in the first 5 years after predictive testing for Huntington's disease. Clin Genet 2003;64:300-9.
19. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469-76.
20. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.
21. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al.; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Ann Intern Med 2003; 138:40-4.

22. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.
23. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 2006;6:9.
24. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al.; GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.
25. Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al.; ATS Documents Development and Implementation Committee. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. Am J Respir Crit Care Med 2006; 174:605-14.
26. Fletcher RH. Carcinoembryonic antigen. Ann Intern Med 1986; 104:66-73.
27. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77:64-71.
28. Levy D, Labib SB, Anderson KM, Christiansen JC, Kannel WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. Circulation 1990;81:815-20.

Appendix 1. 다양한 진단검사 임상시나리오의 예 및 의미

예	예측되는 이득	진단 정확도		환자 의료결과 및 예측되는 치료 결과				예측되는 의료결과, 진단검사 부작용 및 비용
		민감도	특이도	진양성	진음성	위양성	위음성	
치매 진단을 위한 치매선별검사 원전 대비 치매선별검사 축약본	단순한 검사 검사시간 단축	동일	동일	환자에게 중요한 의료결과에 예측되는 영향: 조기 진단 및 치료의 불확실한 이득	확진에 따른 거의 확실한 이득	추가적인 검사 또는 치료로 인한 발생가능한 불안 증대 및 가능한 이환률	진단지연에 따른 가능한 위해	비슷한 진단 정확도 및 짧은 검사 시간은 새로운 검사법의 유용성을 지지함
				환자에게 중요한 의료결과에 대한 직접성: 약간의 불확실성	불확실성 없음	약간의 불확실성	중대한 불확실성	
신결석 진단을 위한 신우조영술 대비 나선식 전산화 단층촬영	더 많은 결석 진단(보다 작은 크기의 결석 진단)	높음	동일	환자에게 중요한 의료결과에 예측되는 영향: 크기가 큰 결석에 대한 확실한 이득 크기가 작은 결석에 대한 덜 확실한 이득과 이로인해 초래될 수 있는 불필요한 치료	불필요한 검사를 피할수 있는 확실한 이득	불필요한 추가적인 침습적 검사로 인한 발생가능한 위해	큰 결석으로 인한 발생가능한 위해 작은 결석에 대한 적은 확실성 그러나 다른 요인 평가를 위한 불필요한 침습적 검사로 인한 위해	신우조영술 대비 적은 부작용은 새로운 검사법의 융용성을 지지함, 그러나 크기가 작은 결석의 진단 증가에 따른 바람직한 및 바람직하지 않은 결과의 차이는 불확실함
				환자에게 중요한 의료결과에 대한 직접성: 약간의 불확실성	불확실성 없음	불확실성 없음	중대한 불확실성	
심혈관질환 진단을 위한 혈관조영술 대비 전산화 단층촬영	덜 침습적, 환자 진단 실패 가능성	약간 낮음	낮음	환자에게 중요한 의료결과에 예측되는 영향: 치료 이득 및 적은 합병증 발생률	확진에 따른 이득 및 적은 합병증 발생률	불필요한 치료로 인한 위해	진단지연에 따른 위해 또는 심근손상	전산화단층촬영의 더 많은 위음성 및 위양성이 초래하는 바람직하지 않은 결과는 비록 혈관조영술의 심근경색 또는 사망과 같은 드문 합병증의 높은 발생률과 높은 비용을 고려하더라도 적용하기 어려움
				환자에게 중요한 의료결과에 대한 직접성: 불확실성 없음	불확실성 없음	불확실성 없음	약간의 불확실성	

Appendix 2. 진단검사연구에서 근거수준을 낮추는 요인과 중재연구의 근거수준 평가와의 차이

근거수준을 낮출 수 있는 요인	풀이 및 중재연구에서의 근거수준과의 차이
연구설계	중재연구와 다른 기준—적절한 참고표준검사를 이용한 직접적 비교를 시행한 단면연구 및 코호트 연구는 근거수준 '높음'으로 간주되며 다른 요인에 따라 '중등도', '낮음', '매우 낮음'으로 근거수준은 낮아질 수 있음
연구 비뚤림	중재연구와 다른 기준—단일 코호트로서 연속적인 환자모집 과정, 질병 중등도에 따른 환자분류가 아니며 의뢰과정 및 선별과정이 명확히 기술되어야 함. 동일환 환자군내 모든 환자는 중재 검사뿐만 아니라 참고표준검사를 시행받아여 함. 연구 평가자는 중재 검사 및 참고표준검사에 모두 맹검되어야 함.
직접성:	
의료결과	중재연구와 비슷한 기준—진료지침 개발 패널들은 진단검사와 환자에게 중요한 의료결과에 대한 직접적 근거를 찾기 어려울 수 있음. 패널들은 진단검사 연구로부터 부작용 그리고 검사 비용과 연관된 진양성 및 위양성 그리고 진음성 및 위음성과 관련된 환자에게 중요한 의료결과에 예측되는 이득과 위해의 차이에 대한 추론이 필요함. 결국, 진단 정확도 연구는 의료결과와의 비직접성(중재의 대리결과와 비슷하게)으로 인하여 일반적으로 낮은 근거수준으로 평가됨.
환자군, 진단검사, 비교검사, 비직접 비교군	중재연구와 비슷한 기준—연구와 권고가 시행되는 환자군에 중대한 차이가 있다면 근거수준은 낮아질 수 있음 (질환의 중등도 또는 동반질환); 임상 권고가 시행되는 의료 환경과 비교하여 연구에서의 검사법과 검사를 시행하는 전문성의 차이가 있다면 또는 참고표준검사와 중재 검사가 각기 다는 연구에서 비교되었고 동일 연구내에서는 비교되지 않은 경우
연구결과의 중요한 비일관성	중재연구와 비슷한 기준—진단검사 연구에서, 민감도, 특이도, 우도의 설명할 수 없는 비일관성은 근거수준을 낮출 수 있음.
비정밀성	중재연구와 비슷한 기준—진단검사 연구에서, 효과 추정치(진양성, 진음성, 위양성, 위음성)에 대한 넓은 신뢰수준은 근거수준을 낮출 수 있음.
높은 출판비뚤림 가능성	중재연구와 비슷한 기준—높은 출판비뚤림 위험(예를 들면, 새로운 중재 또는 검사에 대한 적은수의 환자군을 대상으로 한 연구, 깔때기 그림의 비대칭성)은 근거수준을 낮출 수 있음.

Appendix 3. 진단 정확도 연구에서 주요사항—낮은 관상동맥질환 유병률(20%)을 가진 모집단에서 고식적 심혈관 조영술[a]과 비교하여 다중나선식 전산화 단층촬영은 관상동맥질환의 진단검사로 이용될 수 있는가?

구분	효과 추정치 (95% 신뢰구간)
통합 민감도	0.96 (0.94 to 0.98)
통합 특이도	0.74 (0.065 to 0.84)
통합 양성 우도비[b]	5.4 (3.4 to 8.3)
통합 음성 우도비[b]	0.05 (0.03 to 0.09)

[a]참고표준검사인 혈관조영술은 위양성 또는 위음성이 발생하지 않음을 가정. [b]평균우도비[5].

Appendix 4. 진단 정확도 연구에서 주요사항의 결과—낮은 관상동맥질환 유병률(20%)을 가진 모집단에서 고식적 심혈관 조영술[a]과 비교하여 다중나선식 전산화 단층촬영은 관상동맥질환의 진단검사로 이용될 수 있는가?

결과	1,000명당 환자수	중요도[b]
진양성[c]	192	8
진음성[d]	592	8
위양성[e]	208	7
위음성[f]	8	9
불분명한 검사 결과[g,i]	-	5
부작용[h,i]	-	5
비용[i]	-	5

유병률 20%, 1,000명당 환자수.
[a]참고표준검사인 혈관조영술은 위양성 또는 위음성이 발생하지 않음을 가정. [b]GRADE 건강결과 구분: 9점 척도. 중요하지 않은(1-3점), 중요한(4-6점), 핵심적인(7-9점). [c]약물요법, 혈관성형술, 스텐트, 혈관우회조성술이 필수적이므로 중요함. [d]부작용을 동반할 수 있는 불필요한 중재를 피할 수 있어 중요함. [e]환자가 약물요법이나 침습적 시술로 인한 부작용에 노출될 수 있어 중요함. [f]효과적인 치료를 받지 못해 심혈관질환의 위험도 증가될 수 있어 중요함. [g]해석하기 어렵고, 분명하지 않은 결과 불안을 가중시키고, 추가적인 검사를 어떻게 진행할지 불확실하며, 치료 또는 미치료와 연관된 부정적 결과 가능성으로 인해 중요함. [h]결과보고의 신뢰성이 낮음 드문 경우이나, 심각한 부작용 가능성이 있어 중요함. [i]비록 이러한 결과들은 간략히 보고되지 않았다 할지다도 또 이용가능한 정보를 기반으로 정확히 알려지지 않았더라도, 중요한 결과들임.

Appendix 5. 진단검사 연구에서 근거수준 평가—예시: 다중나선식 전산화 단층촬영은 관상동맥질환 진단을 위해 고식적 심혈관 조영술을 대체할 수 있는가?[a]

연구수	연구설계	비뚤림 위험	비직접성	비일관성	비정밀성	출판비뚤림	근거 수준
진양성(관상동맥질환 있음)							
연구: 21 (환자: 1,570명)	단면연구[b]	비뚤림 위험 낮음	낮은 불확실성	심각함[d]	비정밀성 없음	가능성 적음[e]	⊕⊕⊕◯ 중등도
진음성(관상동맥질환 없음)							
연구: 21 (환자: 1,570명)	단면연구[b]	비뚤림 위험 낮음	낮은 불확실성	심각함[d]	비정밀성 없음	가능성 적음[e]	⊕⊕⊕◯ 중등도
위양성(환자가 관상동맥이 있을것으로 잘못 분류됨)							
연구: 21 (환자: 1,570명)	단면연구[b]	비뚤림 위험 낮음	낮은 불확실성	심각함[d]	비정밀성 없음	가능성 적음[e]	⊕⊕⊕◯ 중등도
위음성(환자가 관상동맥이 없을것으로 잘못 분류됨)							
연구: 21 (환자: 1,570명)	단면연구[b]	비뚤림 위험 낮음	약간의 불확실성[c]	심각함[d]	비정밀성 없음	가능성 적음[e]	⊕⊕◯◯ 낮음

[a]근거수준 평가: 진단 정확도 결과(진양성, 진음성, 위양성, 위음성, 불분명한 결과)와 연관된 환자에게 중요한 의료결과뿐만아니라 검사의 부작용 및 비용을 포함함. [b]고식적 관상동맥 조영술을 시행받은 모든 환자는 일반적으로 관상동맥질환의 높은 유병률을 가지고 있음(포함된 모든 연구의 중간 유병률 63.5%, 범위 6.6-100%). [c]진단지연 또는 심근 손상과 같은 위해와 위음성과의 직접성에 약간의 불확실성은 위음성 결과에 대한 근거수준을 '높음'에서 '중등도'로 낮춤. [d]민감도(전산화 단층촬영 양성 환자수/혈관조영술상 양성 환자수), 특이도(전산화 단층촬영 음성 환자수/ 혈관조영술상 음성 환자수), 우도비, 질병 교차비의 통계적으로 유의한 설명할 수 없는 이질성은 진양성, 진음성, 위양성의 근거수준을 '중등도'에서 '낮음'으로 낮춤. [e]출판비뚤림의 가능성을 완전히 배제할수 없으나 근거수준을 낮추기에 충분하지 않음.