

근거 및 권고수준 등급화 체계 현황 분석

서 경 화¹ · 이 선 희^{2*} · 신 인 순² · 임 선 미¹ · 장 지 은² · 정 유 민² · 박 윤 형^{1,3} | ¹대한의사협회 의료정책연구소, ²이화여자대학교 의과대학 예방의학교실, ³순천향대학교 의과대학 예방의학교실

Trend analysis of grading systems for level of evidence and strength of recommendation

Kyung Hwa Seo, MBA¹ · Sun Hee Lee, MD^{2*} · Ein Soon Shin, PhD² · Sun Mi Lim, PhD¹ · Ji Eun Jang, BS² · Yu Min Jung, RN² · Yoon Hyung Park, MD^{1,3}

¹Research Institute for Healthcare Policy, Korean Medical Association, Seoul, ²Department of Preventive Medicine, Ewha Womans University School of Medicine, Seoul, ³Department of Preventive Medicine, Soonchunhyang University College of Medicine, Cheonan, Korea

*Corresponding author: Sun Hee Lee, E-mail: lsh0270@ewha.ac.kr

Received May 17, 2011 · Accepted May 24, 2011

When clinicians or healthcare professionals are to make decisions, they can judge the quality of evidence and reliability of recommendations by 'Level of evidence' and 'Grade of recommendation'. Because of this, the step of grading evidence and recommendations is very important in developing clinical practice guidelines. The objective of this study was to identify the various grading systems and criteria of the clinical practice guidelines. We reviewed 101 guidelines from the National Guideline Clearinghouse Database and chose 66 guidelines to analyze in terms of the grading systems for level of evidence and strength of recommendation. The grading systems for 'Level of evidence' were classified into 4 types by criteria such as study design, study quality, consistency, limitations, strength of evidence, and validity. Type II was the most common evidence grading system applied by 12 organizations (37.5%) and 30 guidelines (45.5%). The grading systems for 'Grade of recommendation' were classified into 4 types by criteria such as level/quality of evidence, strength of recommendations, study quality, consistency, applicability, balance between benefit and harm, and effectiveness/usefulness. Type I was the most common recommendation grading system applied by 9 organizations (33.4%) and 23 guidelines (40.4%). A formal grading system based on consistent and clear approaches is needed because the process of grading work can be subjective when clinical practice guideline users are making decisions. It is necessary for clinical practice guideline developers to have a common criterion so that they can judge the grade of evidence and recommendations objectively in the development of clinical practice guidelines.

Keywords: Grading system; Level of evidence; Quality of evidence; Strength of recommendation; Clinical practice guidelines

서론

임상의뿐만 아니라 관련 분야의 전문 보건의료 인력들에게까지, 체계적인 방법론에 의하여 개발된 근거기반 진료지침을 활용하는 추세가 자리를 잡아가고 있다. 임상현장에서 임상진료지침(이하 진료지침)을 실제 활용하는 과정에서 진료지침에 사용된 근거의 수준, 권고안의 등급화 또는 권고 강도는 환자의 진단 및 치료에 대한 결정적인 판단을 내리는데 핵심적 의사결정자료가 되므로 매우 중요하며 특히 이들 지표들의 측정 기준과 관련하여 신뢰도 및 도출과정에 대한 정확한 이해가 필수적이다[1]. 근거 수준이란 근거의 질적 수준을 평가하여 등급화한 지표로 제시되는데, 근거의 수집을 통해 얻은 중재방법(intervention)의 효과를 측정할 값이 얼마나 신뢰할 만한지를 나타내는 척도라고 할 수 있다[1,2]. 권고수준은 근거수준을 평가한 후 그 수준을 등급화한 결과외에도 이익과 손실의 균형, 비용효과 혹은 비용편익 측면 등을 총체적으로 감안하여 권고 정도를 등급화한 지표이다. 임상에서의 활용도 등을 감안하여 최종적으로는 ‘권고의 강도’로 제안되고 있는데, 대체로 근거수준이 높으면 권고강도도 높아지지만 진료지침 사례에 따라 근거수준이 낮은 경우에도 권고강도의 수준이 높은 경우도 발생할 수 있다[1,2].

근거수준과 권고수준은 다음과 같은 등급화 단계를 거쳐 결정된다. 첫째, 체계적 문헌검색을 통하여 진료지침 개발을 위한 근거 자료를 확보하고 둘째, 연구에 포함시킬 기준과 부합하는 관련 문헌 혹은 근거를 수집한 후 셋째, 근거의 유형별 즉, 무작위배정비교임상시험 연구와 비무작위 연구 문헌을 구분하고 개발된 평가 척도를 사용하여 질적 수준을 평가하며, 넷째, 문헌 혹은 근거들의 질 평가 결과를 토대로 전체적인 근거들의 중재방법에 대한 효과를 파악하는 단계를 거쳐 최종적으로 권고문안의 등급을 결정하게 된다[3].

질적 수준이 높은 근거에 기반하여 체계적으로 개발된 진료지침에 포함되어 있는 권고 문안들은 임상의를 포함한 전문가들이 보다 정확한 의사결정을 할 수 있도록 도움을 줄 수 있기 때문에 각 권고 문안별로 합리적인 근거수준과 권고수준을 등급화하여 표기하는 것은 근거기반 진료지침개발에 있어 매우 중요한 단계라고 할 수 있다. 그러나 임상학회

별로, 진료지침 개발 기관별로 진료지침 개발과정에서 다양한 근거수준 및 권고수준의 등급체계를 사용하고 있을 뿐만 아니라, 각 등급체계별로 서로 다른 기준을 적용하여 등급을 분류하고 있어 사용자의 의사결정 과정에 혼란을 줄 수 있다[4,5]. 따라서 임상학회 및 진료지침 개발 기관에서 진료지침 개발 시 적용한 다양한 근거수준과 권고수준 등급체계의 현황을 파악하고, 각 등급을 분류하는 기준에 따라 등급체계 유형을 분석하는 등 등급체계 활용실태를 분석한 후, 보편적으로 활용되면서도 국내 실정에 부합된 등급체계를 제안하는 것은 진료지침의 질을 높이고 궁극적으로 의료의 질향상과 효과성을 높이는 데 중요한 연구과제라고 할 수 있겠다.

연구방법

이 연구에서는 국제적으로 개발·활용되고 있는 주요 진료지침들을 대상으로, 진료지침의 근거수준과 권고수준을 비교분석하여 등급화 유형을 도출하고자 하였다. 이 연구의 주요 분석대상이 된 진료지침은 미국의 Agency of Healthcare Research and Quality (AHRQ)에서 운영하고 있는 진료지침 데이터베이스 National Guideline Clearinghouse (NGC) [6]에 질환별로 분류·탑재되어 있는 임상진료지침들을 검색하였고, 진료지침별 등급체계와 관련된 자료는 NGC의 하위메뉴 중 ‘Guideline Synthesis’의 내용을 참고하였다[7]. 2011년 2월 21일 현재 질환별로 등록되어 있는 진료지침은 총 101개였고, 질환별로 최소 2개 최대 5개의 서로 다른 기관에서 개발된 진료지침들이 탑재되어 있었다. 이들 진료지침 중에서 근거수준의 등급화 체계에 관한 내용을 포함하고 있는 진료지침은 66개였으며 이들 진료지침을 분석한 후 근거수준 등급화 유형을 분류하였다. 권고수준의 등급화 체계에 대해서는 해당 내용을 포함하고 있는 57개 진료지침을 분석하여 유형을 분류하였다.

기관별, 임상학회별로 사용하고 있는 근거수준 및 권고수준의 등급화 체계에 대한 현황분석은 다음과 같은 절차를 거쳐 수행되었다. 첫째, 기관별 등급체계 및 기준을 파악하기 위해 NGC 데이터베이스 하위메뉴 ‘Guideline Synthesis’에 탑재되어 있는 진료지침들에 대해서 ‘Strength of evidence

Table 1. Classification by criteria of evidence grading system

Items	Evidence grading system				Total
	Type I	Type II	Type III	Type IV	
Criteria	<ul style="list-style-type: none"> • Study design • Study quality (consistency) 	<ul style="list-style-type: none"> • Study design 	<ul style="list-style-type: none"> • Study design • limitation 	<ul style="list-style-type: none"> • Strength of study design 	
Guidelines (%)	23 (34.8)	30 (45.5)	10 (15.2)	3 (4.5)	66 (100.0)
Organizations (%)	11 (34.4)	12 (37.5)	7 (21.9)	2 (6.2)	32 (100.0)
	<ul style="list-style-type: none"> • SIGN • NCCC/NICE • NCCPC/NICE • NCCWCH/NICE • SMOH • ACN/NZGG • AAOS • MDPH • HFSA • PHS • ACP 	<ul style="list-style-type: none"> • BASHH • RCOG • RNAO • HIGN • NCCMH/NICE • VA/DoD • SOGC • UMHS • ACOG • EFNS • ESC • WOCN 	<ul style="list-style-type: none"> • AAP • GOLD • ACCP • ASGA • Am Diabetes Assoc • ACCP/AACVPR • AACE 	<ul style="list-style-type: none"> • ADA • ACR 	

See the Appendix for abbreviations of organizations.

and recommendations grading scheme'에 서술된 근거 수준 및 권고수준의 등급체계 및 등급 분류 기준들을 수집하였다. 둘째, 수집된 기관별 등급체계에 기여한 요인 혹은 기준을 구성하는 요소들을 파악하고 나열하였다. 셋째, 두 번째 단계에서 등급체계 형태 및 등급별 기준내용들을 비교분석한 것을 바탕으로 유사한 기준들에 의해 분류되는 등급체계를 같은 유형으로 범주화하였고, 마지막으로 범주화된 등급체계 유형별로 진료지침의 종류 및 해당 등급체계를 활용하고 있는 기관 현황을 분석하였다.

연구결과

1. 근거수준 등급체계 유형

분석 결과 근거수준에 대한 등급체계는 총 4가지 유형으로 분류되었다. 각 유형별 등급체계의 세부기준, 적용된 진료지침의 수, 그리고 활용 기관수 및 명칭을 요약정리하면 Table 1과 같다. 체계적 문헌고찰을 통해 각 연구들로부터 정리해 낸 근거들의 등급을 분류하는 기준요소로는 연구 설계방법(study design), 연구의 질(study quality), 일관성(consistency), 한계점(limitation) 및 연구설계 방법의 강도(strength of study design)를 꼽을 수 있었고, 특히 연구 설

계방법은 모든 유형에서 공통적으로 근거수준 등급의 분류 기준으로 고려하고 있음을 알 수 있었다.

유형 I로 분류된 근거수준 등급체계의 주된 기준으로는 연구설계 방법 뿐 아니라, 전반적인 연구의 질적인 측면을 동시에 감안하였고, 일부 기관들에서는 일관성에 관한 기준도 적용하고 있었다. 다른 유형들에 비해 등급별 세부기준 내용을 매우 구체적으로 기술하고 있음이 특징적이었다. 유형 I에 속하는 등급체계는 연구대상 34.4%인 11개 범주로서 Scottish Intercollegiate Guidelines Network (SIGN), National Institute for Health and Clinical Excellence (NICE), Singapore Ministry of Health (SMOH)의 등급체계가 대표적이었으며, 이 등급체계를 활용하여 개발된 진료지침은 연구대상 진료지침의 34.8%인 23개였다. 유형 I 등급체계를 사용하고 있는 기관들은 대부분 SIGN의 등급체계를 그대로 사용하고 있거나 일부 수정하여 활용하고 있음을 알 수 있었다(Table 2).

유형 II로 분류된 근거수준 등급체계는 연구의 설계방법만을 등급 분류 기준으로 고려하고 있었으며, 유형 I의 세부기준 내용에서 언급되지 않은 비무작위화당 통제실험(controlled trial without randomization) 및 준-실험연구(quasi-experimental study)로부터 얻은 근거의 등급도 포

Table 2. Comparison of the level of evidence grading system by organization: Type I^{a)}

Criteria Organizations	High quality RCT	Well conducted RCT	RCT with high risk of bias	High quality case-control/cohort study	Low quality RCT or cohort study	Outcomes research (ecological studies)	Systematic review of case-control study	Well conducted case-control/cohort study	Case-control/cohort study with high risk of bias	Non-analytic studies (case series/ case report)	Expert opinion
SIGN	1++	1+	1-	2++				2+	2-	3	4
NCCC/NICE	1++	1+	1-	2++				2+	2-	3	4
NCCPC/NICE	1++	1+	1-	2++				2+	2-	3	4
NCCWCH/NICE	1++	1+	1-	2++				2+	2-	3	4
SMOH	1++	1+	1-	2++				2+	2-	3	4
ACN/NZGG	I	II	III-1	III-2				III-3		IV	
AAOS	Level I	Level II		Level III						Level IV	Level V
MDPH	Category A			Category B						Category C	Category D
HFSA	A			B							C
PHS	A		B								C
ACP	1a	1b	1c	2a	2b	2c	3a	3b		4	5

RCT, randomized controlled trial. See the Appendix for abbreviations of organizations.

^{a)}The main grading criteria was study design and study quality (consistency).**Table 3.** Comparison of the level of evidence grading system by organization: Type II^{a)}

Criteria Organizations	Meta-analysis of RCT	At least one RCT	At least one controlled study without randomisation	Quasi-experimental study	Non-experimental descriptive study (case-control study)	Expert committee reports/opinion
BASHH	Ia	Ib	IIa	IIb	III	IV
RCOG	Ia	Ib	IIa	IIb	III	IV
RNAO	Level Ia	Level Ib	Level IIa	Level IIb	Level III	Level IV
HIGN	Level I	Level II	Level III	Level IV	Level V	Level VI
NCCMH/NICE	I		IIa	IIb	III	IV
VA/DoD	I		II-1	II-2	II-3	III
SOGC	I		II-1	II-2	II-3	III
UMHS	A		B	C		D
ACOG	I	II-1			II-2	III
EFNS	Class I	Class II				Class III
ESC	A	B				C
WOCN	Level A		Level B			Level C

RCT, randomized controlled trial. See the Appendix for abbreviations of organizations.

^{a)}The main grading criteria was study design.

함하고 있었다. British Association of Sexual Health and HIV (BASHH), Royal College of Obstetricians and Gynaecologists (RCOG), Registered Nurses' Association of Ontario (RNAO) 등 연구대상의 37.5%인 12개 기관에서 유형 II 분류 체계를 사용하고 있었으며, 이 등급체계를 활용하여 개발된 진료지침은 연구대상 진료지침의 45.5%인

tion of Ontario (RNAO) 등 연구대상의 37.5%인 12개 기관에서 유형 II 분류 체계를 사용하고 있었으며, 이 등급체계를 활용하여 개발된 진료지침은 연구대상 진료지침의 45.5%인

Table 4. Comparison of the level of evidence grading system by organization: Type III^{a)}

Criteria Organizations	Well-designed RCT	RCT with minor limitation (inconsistent results/ methodological flaws)	Overwhelming evidence from observational study/ RCT with serious flaws	Observational study (case-control/ cohort study)	Expert opinion/ case reports	Exceptional situations (validating studies cannot be performed)
AAP	A	B		C	D	X
GOLD	A	B		C	D	
ACCP	A	B	C			
ASGE	A	B	C+	C	D	
Am Diabetes Assoc	A	B	C		E	
ACCP/AACVPR	High (A)	Moderate (B)		Low (C)		
AACE	1	2		3	4	

RCT, randomized controlled trial. See the Appendix for abbreviations of organizations.

^{a)} The main grading criteria was study design and limitation.

Table 5. Comparison of the level of evidence grading system by organization: Type IV^{a)}

Criteria Organizations	Studies of strong design	Studies of strong design with minor methodological concerns	Studies of weak design	No reliable evidence/ expert opinion	No evidence
ADA	Grade I	Grade II	Grade III	Grade IV	Grade V
ACR	Category 1	Category 2	Category 3	Category 4	

See the Appendix for abbreviations of organizations.

^{a)} The main grading criteria was strength of study design.

30개였다(Table 3). 결과를 놓고 볼 때, 유형 II로 분류된 근거수준 등급체계가 가장 많이 사용되고 있는 등급체계였다.

유형 III으로 분류된 근거수준 등급체계는 연구 설계방법과 연구의 한계점을 분류기준으로 고려하고 있었는데, 연구 설계 방법의 세부 기준 내용으로는 3가지 즉, 무작위할당통제연구, 관찰연구, 사례보고 및 전문가 합의로 구분하고 있었다. 그리고 무작위할당통제연구로부터 얻은 근거에 대해서는, 잘 설계된 연구로부터 얻은 근거와 한계점(방법론적인 결함이나 일관성이 없는 결과 등)이 있는 연구로부터 얻은 근거로 구분하여 등급을 분류하고 있었다. 유형 III을 활용하여 개발된 진료지침은 연구대상 진료지침의 15.2%인 10개였으며, American Academy of Pediatrics (AAP), Global Initiative for Chronic Obstructive Lung Disease (GOLD)를 포함하여 연구대상기관들 중의 21.9%인 7개 기관에서 유형 III 등급체계를 적용하여 진료지침을 개발하였다(Table 4).

마지막으로 유형 IV로 분류된 근거수준 등급체계는 연구 설계방법의 강도를 분류 기준으로 고려하고 있었다. American Dietetic Association (ADA) 그리고 American College of Radiology (ACR) 2개 기관(6.2%)에서 진료지침 개발시 활용하고 있었으며, 연구대상 진료지침의 4.5%인 3개의 지침개발에 유형 IV 등급체계가 적용되었다. 연구 설계방법의 강도를 기준으로 근거의 등급을 분류하는 점이 다른 유형과의 차이점이며, 신뢰할 만한 근거가 없는 경우 전문가의 견해에 근거하여 Grade IV 등급을 적용하고 있었다. 그러나 다른 유형에 비해 근거의 등급 분류 기준이 구체적이지 않았다(Table 5).

2. 권고수준 등급체계 유형

권고수준의 등급체계를 분류해 본 결과 4가지 유형으로 구분할 수 있었으며, 각 유형별 등급분류 세부기준 내용과

Table 6. Classification by criteria of recommendation grading system

Items	Recommendation grading system				Total
	Type I	Type II	Type III	Type IV	
Criteria	<ul style="list-style-type: none"> • Study quality • Level of evidence (consistency/ applicability) 	<ul style="list-style-type: none"> • Level of evidence • Strength of recommendations 	<ul style="list-style-type: none"> • Level of evidence 	<ul style="list-style-type: none"> • Level of evidence • Balance between benefit and harm 	
Guidelines (%)	23 (40.4)	15 (26.3)	11 (19.3)	8 (14.0)	57 (100.0)
Organizations (%)	9 (33.4)	8 (29.6)	6 (22.2)	4 (14.8)	27 (100.0)
	<ul style="list-style-type: none"> • SIGN • SMOH • ACN/NZGG • NCCWH/NICE • NCCMH/NICE • EFNS • BASHH • MDPH • AACE 	<ul style="list-style-type: none"> • ADA • KPCMI • AAP • VA/DoD • ASGE • HFSA • UMHS • APA 	<ul style="list-style-type: none"> • AAOS • RAO • SOGC • ACOG • NHFA/CCANZ • TES 	<ul style="list-style-type: none"> • USPSTF • AGA • ESC • CAPO/BSC 	

See the Appendix for abbreviations of organizations.

Table 7. Comparison grading system of recommendation by organization: Type I^{a)}

Organizations	Criteria	Evidence from level 1+/ 1+, directly applicable, consistency results	Evidence from level 2+/ 2+, directly applicable, consistency results	Evidence from level of 2+/ 2+, directly applicable, consistency results	Evidence level 3 or 4	Good practice points based on clinical experience	No conclusive level of evidence category
SIGN		Grade A	Grade B	Grade C	Grade D	GPP	
SMOH		A	B	C	D	GPP	
NCCWCH/NICE		A	B	C	D	GPP	
ACN/NZGG		A	B	C	D	GPP	
NCCMH/NICE		A	B		C	GPP	
EFNS		Level A rating	Level B rating	Level C rating		GPP	
MDPH		Category A	Category B	Category C	Category D		
BASHH		A		B	C		
AACE		A	B		C		D

See the Appendix for abbreviations of organizations.

^{a)}The main grading criteria was study quality and level of evidence (consistency/ applicability).

적용된 진료지침의 수 그리고 활용 기관 수 및 기관명을 요약제시하면 Table 6과 같다. 권고수준의 등급체계에서는 근거수준에 비해 다양한 요소들을 등급 분류 기준으로 고려하고 있음을 알 수 있었다. 기준 항목에는 연구의 질(study quality), 일관성(consistency), 적용가능성(applicability), 근거의 질(level/quality of evidence), 권고수준의 강도

(strength of recommendations), 이익과 손실의 균형(balance between benefit and harm) 등과 같은 세부 기준들이 사용되었다.

유형 I로 분류된 권고수준 등급체계는 SIGN, SMOH, NICE, Australian Cancer Network/New Zealand Guidelines Group (ACN/NZGG) 등을 포함하여 연구대상의

Table 8. Comparison grading system of recommendation by organization: Type II^{a)}

Organizations	Criteria	Strong recommended, good evidence	Strong recommended, can apply to most clinical settings	Strong recommended, fair evidence	Intermediate -strength, may change in stronger evidence	Intermediate -strength, may differ depending on circumstances	Weak recommended, alternative approaches in better circumstances
ADA		Strong		Fair			Weak
KPCMI		A (evidence based)		B (evidence based)			
AAP		Strong recommendation		Recommendation			
VA/DoD		A		B		C	
ASGE		1A	1B	1C+	1C	2A	2B
HFSA		Recommended		Should be considered		May be considered	
UMHS		I		II			
APA		I		II		III	
Organizations	Criteria	Very weak recommended, likely alternative approaches in better circumstances	Weak recommended, likely to change as data become available	No recommended	Recommend against routine/ prevention setting	Insufficient evidence	Based on Expert opinion
ADA						Insufficient evidence	Consensus
KPCMI				C (evidence based)		D (evidence based)	Consensus-based
AAP			Option	No recommendation		I	
VA/DoD				C	D	I	
ASGE		2C	3	Not recommended			
HFSA							
UMHS				III			
APA							

See the Appendix for abbreviations of organizations.

^{a)}The main grading criteria was level of evidence and strength of recommendations.

33.4%인 9개 기관에서 진료지침을 개발하는 데 활용하고 있었으며, 이를 적용하여 개발된 진료지침의 수는 23개 (40.4%)였다. 유형 I 등급체계는 주로 연구의 질 및 근거수준을 기준으로 등급을 구분하고 있었고, 일부 기관에서는 일관성 및 적용가능성을 등급 분류 기준으로 고려하고 있었다. 근거수준 등급체계 유형 I과 유사하게 권고수준 등급체계 유형 I에서도 대부분 SIGN의 등급체계를 그대로 또는 수정하여 적용하고 있음을 알 수 있었다(Table 7). 또한 유형 I

로 분류된 권고수준 등급체계가 가장 많이 사용되고 있는 등급체계였다.

유형 II로 분류된 권고수준 등급체계는 근거의 수준과 더불어 권고에 대한 강도의 크기를 감안하여 각 등급을 구분하고 있었으며, 주로 'strong', 'fair', 'weak' 등으로 강도의 크기를 나타내고 있었다. American Society for Gastrointestinal Endoscopy (ASGE)의 경우 권고사항에 대한 강도를 매우 구체적으로 구분한 반면, ADA나 AAP, Heart Failure

Table 9. Comparison grading system of recommendation by organization: Type III^{a)}

Criteria Organizations	Good evidence to recommend the clinical action	Fair evidence to recommend the clinical action	Existing evidence is conflicting	Fair evidence to recommend against clinical action	Good evidence to recommend against clinical action	Inefficient evidence
AAOS	A	B		C		I
RNAO	A	B	C	D	E	I
SOGC	A	B	C	D	E	I
ACOG	Level A		Level B	Level C		
NHFA/CSANZ	A	B		C		D
TES	++++	+++O	++OO	+OOO		

See the Appendix for abbreviations of organizations.

^{a)}The main grading criteria was level of evidence.

Society of America (HFSA), American Psychiatric Association (APA)와 같은 기관에서는 권고의 강도 및 권고여부를 3등급 수준으로 간단하게 구분하고 있었다. 또한 KPCMI, AAP 등 5개 기관에서는 ‘권고하지 않음(no recommended)’ 등급을 포함하기도 하였으며, 근거가 부족한 경우와 적절한 근거가 없을 때 전문가 합의 기반에 대한 등급을 정하고 있는 기관도 있었다. 연구대상 진료지침의 26.3%인 15개 진료지침이 유형 II에 속하는 등급체계를 적용하여 개발되었으며, 연구대상의 29.6%인 8개의 기관에서 사용하고 있었다(Table 8).

유형 III으로 분류된 권고수준 등급체계는 대표적으로 American Academy of Orthopaedic Surgeons (AAOS), RNAO, Society of Obstetricians and Gynaecologists of Canada (SOGC) 등의 기관에서 적용하고 있었으며, 이 유형에 해당하는 등급체계는 권고사항의 등급 결정 시 주로 근거 수준에 대한 내용을 기준으로 고려하고 있었다. 근거 수준은 크게 ‘good’, ‘fair’로 구분하고 진료행위를 권고할 만한 근거들인지 그렇지 않은 근거들인지를 구분하여 등급을 분류한 것이 특징이며, 기존 근거에 대한 의견에 차이가 있는 경우도 별도의 등급으로 분류하고 있었다. 연구대상의 22.2%인 6개 기관에서 유형 III에 해당하는 등급체계를 사용하고 있었으며, 연구대상 진료지침의 19.3%인 11개 진료지침이 유형 III에 속하는 등급체계를 적용하여 개발되었다(Table 9).

마지막으로 유형 IV로 분류된 권고수준 등급체계는 U.S. Preventive Service Task Force (USPSTF), American

Gastroenterological Association Institute (AGA), European Society of Cardiology (ESC), Committee on Perioperative Evaluation/ Brazilian Society of Cardiology (CAPO/BSC) 4개 기관(14.8%)에서 진료지침 개발 시 적용하고 있는 등급체계였다. 이 유형에 해당하는 등급체계들은 근거의 수준과 함께 이익과 손실에 대한 크기의 비교를 기준으로 권고수준의 등급을 구분하고 있었다. 지침 적용 시 순이익이 지속적인 경우 가장 높은 등급을 적용하였고, 이익보다 손해가 더 큰 경우 그리고 근거가 부족한 경우 낮은 등급을 적용하였다. 유형 IV 등급체계를 적용하여 개발된 진료지침의 수는 연구대상 진료지침의 14.0%인 8개였다(Table 10).

결론

본 연구에서는 임상진료지침을 개발할 때 진료지침 개발 기관 및 임상학회별로, 상이한 기준에 의거한 근거수준 및 권고수준 등급체계를 사용하고 있는 국제적인 실태를 파악함으로써 국내 적용방안을 모색하고자 시도되었다. 국제적인 현황파악을 위해 NGC에 탑재된 총 101개의 진료지침 중에서 근거수준 등급체계가 명시된 진료지침 66개 그리고 권고수준 등급체계가 명시된 진료지침 57개를 대상으로 등급체계들을 분석하였다. 또한 등급화에 적용된 세부 기준들의 특성에 따라 그 유형을 분류해 봄으로써 국제적으로 가장 많이 사용되고 있는 등급체계를 구분해 내고, 이를 기반으로 진료지침 사용자 측면에서 표준화된 근거수준 및 권고수준 등급

Table 10. Comparison grading system of recommendation by organization: Type IV^{a)}

Criteria Organizations	Net benefit is substantial	Net benefit is moderate	Net benefit is small	No net benefit or harms outweigh the benefit	Insufficient evidence
USPSTF	A	B	C	D	I statement
AGA	Grade A	Grade B	Grade C	Grade D	Grade Insuff
ESC	Class I	Class II	Class IIa	Class IIb	Class III
CAPO/BSC	Class I	Class II	Class IIa	Class IIb	

See the Appendix for abbreviations of organizations.

^{a)}The main grading criteria was level of evidence and balance between benefit and harm.

체계를 개발하기 위한 기초 자료로 활용함으로써 상이한 등급 체계의 난립으로 인한 혼선을 최소화하고자 하였다.

분석결과 근거수준에 대한 등급체계는 총 4가지 유형으로 분류되었으며, 연구 설계방법, 연구의 질, 일관성, 한계점 및 연구설계 방법의 강도가 주요 분류 기준으로 고려되고 있음을 확인하였다. 특히 연구 설계방법은 모든 유형에서 공통적인 근거수준 등급의 분류 기준이었다. 유형별 분류에서는 유형 II로 분류된 근거수준 등급체계가 가장 많이 사용되고 있는 바, 이 등급체계의 특징은, 등급 분류 기준으로 연구의 설계방법만을 고려하고 있으면서 비무작위할당 통제실험 및 준-실험연구로부터 얻은 근거의 등급도 포함하는 것으로 요약될 수 있다. 권고수준의 등급체계도 4가지 유형으로 분류되었고, 권고수준 등급체계에서는 근거수준 등급체계에 비해 연구의 질, 일관성, 적용가능성, 근거의 질, 권고수준의 강도, 그리고 이익과 손실의 균형 등 다양한 요소들을 등급 분류 기준으로 고려하고 있었으며, 연구의 질 및 근거 수준을 기준으로 분류한 유형 I의 권고수준 등급체계가 가장 많이 사용되고 있는 등급체계였다.

따라서 실증적인 자료 분석을 토대로 기관별, 임상학회별로 진료지침 개발 시 적용하고 있는 등급체계가 매우 다양하고, 등급체계에서 고려하고 있는 등급 분류 기준 역시 상이하다는 것을 확인 할 수 있었으며, 51개 기관에서 활용하고 있는 총 6개 등급체계를 평가한 Atkins 등[8]의 연구 결과와 유사한 결과를 알 수 있었다. 본 연구에서는 Atkins 등[8]의 연구에서 적용한 등급체계 평가기준을 적용하여 분석을 한 것은 아니지만, 등급체계 기준으로 고려된 요소들만으로도 기관별 등급체계가 다양함을 확인할 수 있었다.

본 연구 결과에서 알 수 있듯이 기관별, 임상학회별로 개발된 진료지침들에서 비슷한 수준의 근거에 대해 다른 등급을 부여하고, 상이한 등급체계를 적용하게 될 경우 결국 권고사항의 등급화에도 영향을 미치게 되어, 임상 및 보건의료분야 전문가들이 임상에서 진료지침을 사용하게 되었을 때 오히려 중요한 의사결정에 혼란을 주거나 시간을 지연시키는 심각한 문제가 발생할 수 있으므로, 근거수준 및 권고수준의 등급체계를 활용함에 있어서 매우 신중함을 기할 필요가 있다고 판단된다[3,9,10]. 특히 각 진료지침의 특성 및 진료지침이 활용되는 의료문화적 상황에 맞게 다양한 등급체계를 유연하게 사용하는 것은 현실적 불가피성이 있으나 가능한 보편적인 등급체계를 활용하는 것이 진료지침의 공신력을 높일 수 있으며 이 연구에서 제시된 바와 같이 근거수준 등급체계는 유형 II, 권고등급 분류체계는 유형 I 범주가 가장 많이 활용되고 있음을 국내 표준 설정시 참고할 필요가 있겠다.

국내의 경우에는 진료지침 개발 노력이 시작단계에 있고 국내 실정에 부합된 방법론에 대한 논의도 충분하지 않아 기존에 개발된 진료지침들에는 등급체계조차 포함되지 않는 경우도 있고, 등급체계가 제시된 경우라 하더라도 체계적인 과정을 통해 선정되지 않아[5] 향후 많은 논의가 필요한 분야라고 할 수 있다[11]. 등급체계의 타당성을 높이기 위해서는 국내 실정에 기반한 다양한 근거연구가 활성화되어야 하겠지만, 국내 실정에 적합한 등급체계 표준화에 대한 과학적 토의가 지속될 필요가 있으며, 이 연구는 이러한 논의의 실증적인 기초자료로 활용될 수 있을 것으로 기대된다.

핵심용어: 등급화체계; 근거수준; 근거의 질; 권고수준의 강도; 임상진료지침

REFERENCES

1. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schunemann HJ; GRADE Working Group. Going from evidence to recommendations. *BMJ* 2008;336:1049-1051.
2. The GRADE working group. GRADE handbook. The Cochrane Collaboration; 2009.
3. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Varonen H, Vist GE, Williams JW Jr, Zaza S; GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.
4. Timmermans S, Mauck A. The promises and pitfalls of evidence-based medicine. *Health Aff (Millwood)* 2005;24:18-28.
5. Owens DK, Lohr KN, Atkins D, Treadwell JR, Reston JT, Bass EB, Chang S, Helfand M. Grading the strength of a body of evidence when comparing medical interventions. Rockville, MD: Agency for Healthcare Research and Quality; 2009.
6. Agency for Healthcare Research and Quality. National Guideline Clearinghouse [Internet]. Rockville, MD: Agency for Healthcare Research and Quality; [cited 2011 Feb 21]. Available from: <http://guideline.gov/index.aspx>.
7. Agency for Healthcare Research and Quality. National Guideline Clearinghouse: Guideline Synthesis [Internet]. Rockville, MD: Agency for Healthcare Research and Quality; [cited 2011 Feb 21]. Available from: <http://guideline.gov/syntheses/index.aspx>.
8. Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, Liberati A, O'Connell D, Oxman AD, Phillips B, Schunemann H, Edejer TT, Vist GE, Williams JW Jr; GRADE Working Group. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. *BMC Health Serv Res* 2004; 4:38.
9. Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schunemann H. An emerging consensus on grading recommendations? *Evid Based Med* 2006;11:2-4.
10. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64:401-406.
11. Ansari MT, Tsertsivadze A, Moher D. Grading quality of evidence and strength of recommendations: a perspective. *PLoS Med* 2009;6:e1000151.

Appendix. Abbreviations of organizations

AACE: American Association of Clinical Endocrinologists
 AAOS: American Academy of Orthopaedic Surgeons
 AAP: American Academy of Pediatrics
 ACCP/AACVPR: American College of Chest Physicians/ American Association of Cardiovascular and Pulmonary Rehabilitation
 ACN/NZGG: Australian Cancer Network/ New Zealand Guidelines Group
 ACOG: American College of Obstetricians and Gynecologists
 ACP: American College of Physicians
 ACR: American College of Radiology
 ADA: American Dietetic Association
 AGA: American Gastroenterological Association Institute
 Am Diabetes Assoc: American Diabetes Association
 APA: American Psychiatric Association
 ASGE: American Society for Gastrointestinal Endoscopy
 BASHH: British Association of Sexual Health and HIV
 CAPO/BSC: Committee on Perioperative Evaluation, Brazilian Society of Cardiology
 EFNS: European Federation of Neurologic Studies
 ESC: European Society of Cardiology
 GOLD: Global Initiative for Chronic Obstructive Lung Disease
 HFSA: Heart Failure Society of America
 HIGN: Hartford Institute for Geriatric Nursing
 KPCMI: Kaiser Permanente Care Management Institute
 MDPH: Massachusetts Department of Public Health
 NCCC: National Collaborating Center for Chronic Conditions
 NCCMH: National Collaborating Centre for Mental Health
 NCCPC: National Collaborating Centre for Primary Care
 NCCWCH: National Collaborating Centre for Women's and Children's Health
 NHFA/CSANZ: National Heart Foundation of Australia/ Cardiac Society of Australia and New Zealand
 NICE: National Institute for Health and Clinical Excellence
 PHS: Public Health Service
 RCOG: Royal College of Obstetricians and Gynaecologists
 RNAO: Registered Nurses' Association of Ontario
 SIGN: Scottish Intercollegiate Guidelines Network
 SMOH: Singapore Ministry of Health
 SOGC: Society of Obstetricians and Gynaecologists of Canada
 TES: The Endocrine Society
 UMHS: University of Michigan Health System
 USPSTF: U.S. Preventive Services Task Force
 VA/DoD: Department of Veterans Affairs, Department of Defense
 WOCN: Wound, Ostomy, and Continence Nurses Society



Peer Reviewers' Commentary

환자치료에 필요한 의학적 지식은 기초의학지식을 근거로 많은 임상시험을 거쳐서 만들어진다. 의학 분야는 전문학술잡지 수, 학술대회, 산출되는 논문의 수 등에서 다른 분야를 압도하고 있으며, 현재도 세계 각국의 대학과 연구소, 병원 등에서 막대한 논문이 쏟아지고 있다. 그중에서 많은 수의 표본으로 시행된 연구, 연구 설계가 잘된 연구의 결과를 찾아 환자진료에 사용하는 것이 의사들의 바람이다. 이 논문은 의학연구결과를 질적으로 평가하여 그 결과를 어느 정도 받아들여야 하는지와 받아들인 결과를 사용할 때 어느 정도로 사용해야 하는지에 대해 정리한 논문이다. 앞으로 각과별로 임상진료지침을 작성할 때 지표로 삼을 내용이라고 본다.

[정리:편집위원회]