표준화 환자를 이용한 임상수행능력평가시험 (CPX)에서 교수와 표준화 환자의 평가 결과 비교

이화여자대학교 의과대학 의학교육실,
이화여자대학교 의과대학 의학교육실 의학교육실임상수행능력평가를 위한 서울 · 경기 컨소시엄 위원회[1]

권복규 · 김나진 · 이순남 · 어은경 · 박혜숙 · 이동현 · 박미혜 · 오지영 · 한재진[1] · 허정원 · 유경하[1]

= Abstract =

Comparison of the Evaluation Results of Faculty with Those of Standardized Patients in a Clinical Performance Examination Experience

Ivo Kwon, MD, **Najin Kim, Soon Nam Lee,** MD, **Eunkyung Eo,** MD, **Hyesook Park,** MD, **Dong Hyeon Lee,** MD, **Mi Hae Park,** MD, **Jee-Young Oh,** MD, **Jae Jin Han[1],** MD, **Jung-Won Huh,** MD, **Kyung Ha Ryu[1],** MD

Office of Medical Education, Ewha Womans University College of Medicine, Office of Medical Education, Ewha Womans University College of Medicine, Seoul · Gyeonggi CPX Consortium[1]

Purpose: To compare the evaluation results of faculties to those of Standardized Patients (SP) participating in a Clinical Performance Examination (CPX) administered at Ewha Womans University College of Medicine.

Methods: The CPX was taken by 77 fourth year medical students. Cases and checklist were developed by the medical school consortium in capital area. Six cases were used and 24 SPs participated and evaluated the students' performances. The whole session was recorded on videotapes so that 6 medical school faculties could analyze and evaluate the students' performances as well. The results were compared and analyzed by SPSS package.

Results: The agreement between the faculties and the SPs was relatively good (r=0.79), but not good enough. In every case, SPs gave higher marks than did the faculties. Clear disease entity cases like "hepatitis" and "anemia" showed better agreement than obscure clinical contexts such as "bad news delivery". Better agreement was seen in the items of physical exam category (r=0.91), but the agreement was very poor in the items of doctor-patient (Dr-Pt) relationship category (r=0.54). The construction of checklist and the character of each evaluation item should influence the differences.

Conclusion: More detailed guidelines and clear/specific evaluating items are necessary to improve the agreement rate. In certain categories like physical exam and brief history taking, the SP's evaluation can replace the faculties', but for complex contexts like Dr-Pt relationship.

Key Words: Evaluation, Clinical performance examination, Faculty, Standardized patient

교신저자: 유경하, 이화여자대학교 의과대학 소아과학교실
서울시 양천구 목6동 911-1번지
Tel: 02)2650-2678, Fax: 02)2653-3718, E-mail: ykh@ewha.ac.kr

서 론

표준화 환자를 사용한 임상수행능력평가는 우리나라에는 1994년에 도입되어 2002년까지 26개 의과대학이 이를 시행하였다. 임상수행능력평가는 객관구조화진료시험 (Objective Structured Clinical Exam, OSCE)과 진료수행시험 (Clinical Performance Exam, CPX)으로 구분할 수 있으며 OSCE가 단편적인 임상수기의 수행 여부 중심이라면 CPX는 보다 포괄적인 진료 현실에 가깝게 구성된 시험이라 할 수 있다. 우리나라에서는 도입 단계에서 두 방법이 맞물려 스테이션이 어느 형태에 가까운지 혼동되는 경우도 있지만 (박훈기 등, 2003) CPX가 보다 포괄적이며, 환자가 호소하는 문제 (problem)중심으로 되어 있고 1차적인 방법이 되어야 하며 학생의 임상능력을 평가하는 최선의 방법이라는 데는 이견이 없는 듯하다 (Abrahamson, 1998).

그러나 CPX를 제대로 수행하는 데 있어서 가장 중요한 요소 중 하나는 학생의 수행 (performance)에 대한 평가 (assessment)가 제대로 이루어지고 있는가이다. 가장 좋은 방법은 잘 훈련된 임상 교수가 참가하여 평가하는 것이지만 임상 교수의 바쁜 일정과 각 의과대학의 한정된 자원을 고려한다면 이는 쉽지 않은 일이다. 그러므로 시험 운영의 효율성을 고려하여 표준화 환자를 학생 평가에 활용하여 필요한 교수 수를 최소한으로 줄이는 방법을 사용하게 되고, 여러 연구들은 OSCE 수행 시 표준화 환자의 점검표 채점이 임상 교수의 채점과 비교할 때 80~100%의 높은 일치율을 보인다는 사실을 증명하였다 (박훈기 등, 2003).

그러나 CPX에서 표준화 환자의 평가와 임상 교수의 평가를 비교한 국내 연구는 많지 않으며 따라서 표준화 환자의 평가를 그대로 학생 평가로 인정하기 위해서는 검증이 필요하다. 또한 표준화 환자의 평가를 교수 평가와 비교하여 봄으로써 시험문제 (증례)의 개발이 제대로 이루어졌는지, 그리고 표준화 환자의 점검표 작성에 대한 훈련이 제대로 이루어졌는지를 파악할 수 있을 것이다. 따라서 본 연구는 일개 의과대학에서 시행한 진료수행시험에서 표준화 환자 평가와 교수 평가의 일치도를 분석하여 CPX에서 표준화 환자 평가를 어떻게 적용할 수 있을지를 모색해 보고 관련된 문제점을 논의하고자 하였다.

대상 및 방법

가. CPX 수행

2004년 7월 26일부터 28일까지 본과 4학년 학생 77명을 대상으로 CPX를 실시하였다. 수도권 8개 의대 CPX consortium에서 개발한 증례 6개를 사용하였으며 한 증례 당 2~3명의 표준화 환자가 참여하였다. 전체 동원된 표준화 환자의 숫자는 24명이었으며, 이들이 학생 평가를 담당하였다. 학생은 오전과 오후에 걸쳐 시험을 보았고, 시험의 전 과정은 비디오로 녹화되었다. 시험의 관리와 감독은 6명의 교수진이 돌아가면서 담당하였으며 표준화 환자 트레이너가 함께 관찰하였다. 미리 증례를 경험한 학생들에게 다른 학생들에게 문제를 누설하지 말도록 주의를 주었다. 녹화한 비디오는 나중에 교수들이 판독하였으며 이를 표준화 환자가 작성한 평가와 비교하였다. 6개의 증례는 복통, 나쁜 소식 전하기, 불면증, 기침, 간염보균자와 빈혈이었다.

나. 평가

2004년 11월 11일부터 12일 사이에 평가 워크숍을 겸한 교수 평가를 합숙을 하며 실시하였다. 4개의 증례에 대해서는 학생 24명에 대해 교수 1인이 단독으로 평가한 점수와, 교수 2인이 공동으로 평가한 점수를 서로 비교해 보았으며 2개 증례는 교수 1인의 평가로 점수화하였다. 단독평가와 공동평가를 함께 한 증례는 복통, 나쁜 소식 전하기, 불면증, 기침이었다. 6개 증례 중 4개만을 비교한 것과 77명의 학생 중 24명을 평가한 것은 1박2일의 기간 중 전체 학생의 비디오를 전부 볼 수가 없어서였으며 나머지 학생에 대해서는 교수 1인이 비디오 평가를 한 것으로 점수를 산출하였다.

표준화 환자 점검표는 각 증례에 따라 약간의 차이는 있지만, 학생의사에 대한 만족도를 최우수/아

Table I. Categorization and Number of the Evaluation Items of the 6 Standardized Patients

	Abdominal pain	Bad News delivery	Insomnia	Cough	Hepatitis carrier	Anemia
History	12	3	9	14	9	13
Physical examination	5		5	4	3	8
Information sharing	3	5	1	1	6	
Doctor-patient relation	7	7	7	7	7	7
Attitude	6	8	2	6	6	6
Sum	33	23	24	32	31	34

주 잘함/잘함/개선 요망/최저 수준/수준 미달의 6단계로 표시하게 하였으며 병력 (history)에 대한 지문을 3~14개로 예/아니오로 표시하게 하였으며 신체진찰에 대한 평가항목이 3~8문항으로 제대로 했음/제대로 못했음/하지 않았음의 3개 영역으로 표시하게 하였다. 임상의사의 예절에 대한 항목이 2~8문항으로 예/아니오/해당 없음의 3개 영역으로 표시하게 하였으며 환자-의사관계를 7개 문항으로 하여 최우수-수준미달의 6단계로 표시하게 하였다. 학생의사가 잘한 점과 고쳐야 할 점을 구체적으로 적어 주는 것으로 채점은 마치게 되어 있다. 기타 증례의 성격에 따라 필요한 부분, 예컨대 환자 교육이나 나쁜 소식 전하기와 같은 내용을 가지고 평가문항을 만들어 여기에 예/아니오 등으로 표시하기도 하였다.

이와 같은 평가지 문항을 영역별로 분류하여 해당 문항의 개수를 열거하면 Table I과 같다.

다. 분석

교수평가자와 표준화 환자의 평가를 Pearson 상관분석을 이용하여 사례별로 상관계수를 산출하였다. 또한 교수평가자와 표준화 환자의 평가의 차이를 전체 점수와 사례별 점수로 비교하였으며 짝지은 t- test를 이용하여 유의성 검정을 하였다. 병력청취, 신체진찰, 환자교육, 의사환자관계, 의사태도 영역으로 구분하여 각 영역에서의 사례별 상관계수와 유의성을 평가하였다. 또한 문항간에 교수평가자와 표준화 환자간의 채점의 일치율을 산출하였다. 모든 통계분석은 SPSS 10.0 통계패키지를 이용하여 분석하였으며 유의성은 P값 0.05 수준에서 시행하였다.

결 과

가. 교수 평가와 표준화 환자 평가의 평균 점수차

점검표의 각 항목별로 평가점수의 평균치를 비교하여 작성하였다. 각 증례의 만점은 '복통'은 76점, '기침'은 79점, '나쁜 소식 전하기'는 59점, '불면증'은 63점, '간염보균자'는 68점, '빈혈'은 80점이나 모두 100점 만점으로 점수를 환산하여 비교하였다. '복통' 증례에서는 교수평가가 43.2점 (SD=5.37)으로 표준화 환자 평가점수 45.2점 (SD=7.84)과 유의한 차이가 없었다. '기침' 증례에서는 교수평가가 40.6점 (SD=5.50)으로 표준화 환자 평가점수 44.3점 (SD=7.37)보다 다소 낮았다. '나쁜 소식 전하기' 증례에서는 교수평가가 32.6점 (SD=6.05)으로 표준화 환자 평가점수 39.6점 (SD=5.08)보다 유의하게 낮았다. '불면증' 증례에서는 교수평가가 32.0점 (SD=6.53)으로 표준화자 평가점수 40.0점 (SD=4.40)과 가장 유의한 차이를 보였다. '간염보균자' 증례에서는 교수평가가 33.3점 (SD=8.44)으로 표준화 환자 평가점수 36.4점 (SD=5.73)보다 다소 낮았다. '빈혈' 증례에서는 교수평가가 39.2점 (SD=4.86)으로 표준화 환자 평가점수 39.7점 (SD=7.66)과 거의 동일하였다.

통계적으로 유의한 차이를 보인 증례는 '기침', '나쁜 소식 전하기', '불면증', '간염보균자'이었고, 동일한 평균치를 보인 증례는 '빈혈'이었으며 전체적으로 각 증례에서 교수평가 보다 표준화 환자 평가점수의 평균치가 높았다 (Fig. 1).

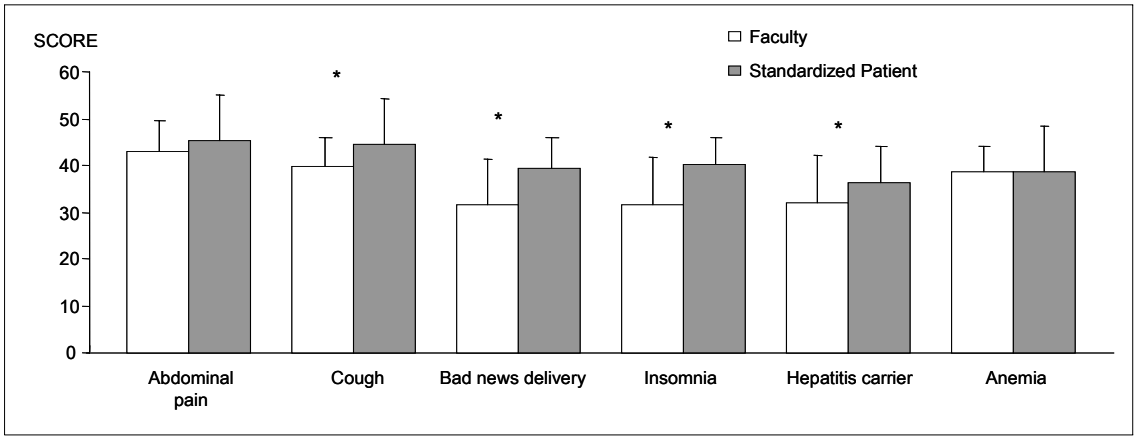

Fig. 1. Average scores of each items (* P < 0.05, by the paired t-test)

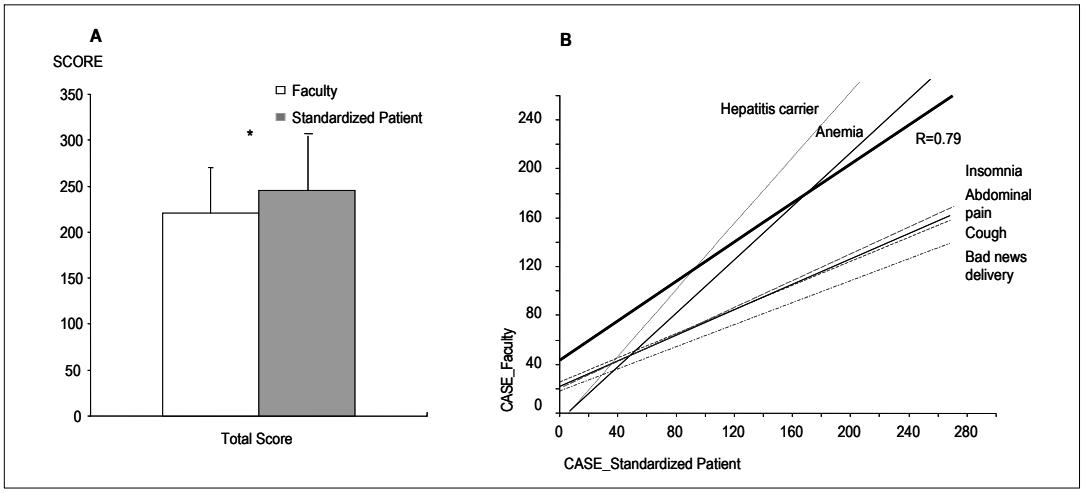

Fig. 2. Correlation between faculty and standardized patient evaluation according the cases (* P < 0.05, by the paired t-test)

나. 교수 평가와 표준화 환자 평가의 상관도

6개 증례의 총점은 교수평가가 220.5점으로 표준화 환자의 평가점수 245.1점보다 통계적으로 유의하게 낮았다 (Fig. 2A). 표준화 환자 평가와 교수 평가간의 피어슨 상관계수는 '나쁜 소식 전하기' 증례에서 0.33으로 가장 낮은 값을 보였다. '기침'은 상관계수 0.59, '복통'은 0.61, '불면증'은 0.62로 거의 비슷한 값을 나타내며 유의한 상관관계를 보였다. '빈혈'에서는 상관계수 0.74, '간염보균자'는 0.78로 비교적 높은 값을 보였다.

표준화 환자 평가와 교수 평가간의 상관계수는 각 증례별로 다른 값을 보였으나 전체적인 상관계수는 0.79로 비교적 높은 값을 보이며 유의한 상관관계를 보였다 (Fig. 2B).

다. 각 영역별 교수 평가와 표준화 환자 평가의 차이

1) 병력청취 영역

교수평가의 평균은 33.7점으로 표준화 환자의 평가점수 43.0점보다 통계적으로 유의하게 낮았다

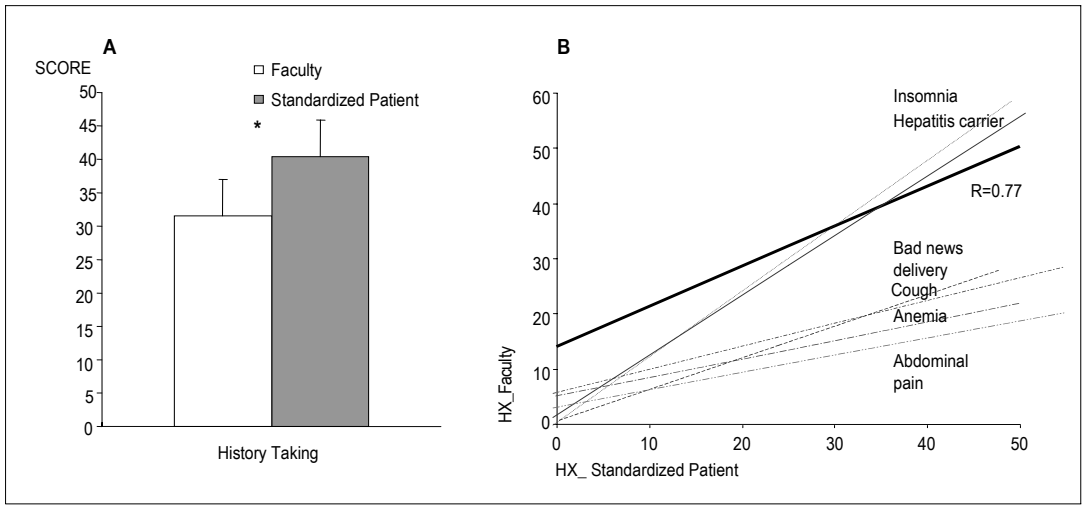

Fig. 3. Correlation between faculty and standardized patient evaluation in history taking (* P < 0.05, by the paired t-test)

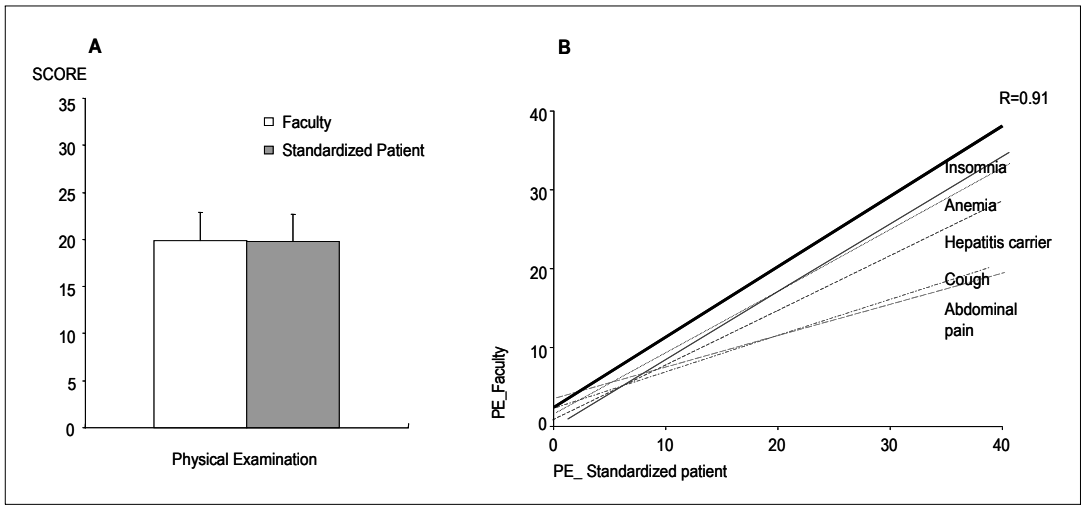

Fig. 4. Correlation between faculty and standardized patient evaluation in physical examination

(Fig. 3A). 표준화 환자 평가와 교수 평가간의 피어슨 상관계수는 '복통' 증례에서 0.30, '빈혈'에서 0.35로 낮았다. '기침'은 0.46, '나쁜 소식 전하기'는 0.55로 비슷한 값을 나타내며 유의한 상관관계를 보였다. '간염보균자'에서는 상관계수 0.70, '불면증'은 0.87로 비교적 높은 값을 보였다. 표준화 환자 평가와 교수 평가간의 상관계수는 각 증례에서 다른 값을 보였으나 전체적인 상관계수는 0.77로 비교적 높은 값으로 유의한 상관관계를 보였다 (Fig. 3B).

2) 신체진찰 영역

교수평가와 표준화 환자의 평가점수 평균치가 20.0점으로 같았다 (Fig. 4A). 표준화 환자 평가와 교수 평가간의 피어슨 상관계수는 '복통'과 '기침' 각 증례에서 0.46과 0.57로 유의한 상관관계를 보였다. '간염보균자'에서는 상관계수 0.68, '빈혈'은 0.89,

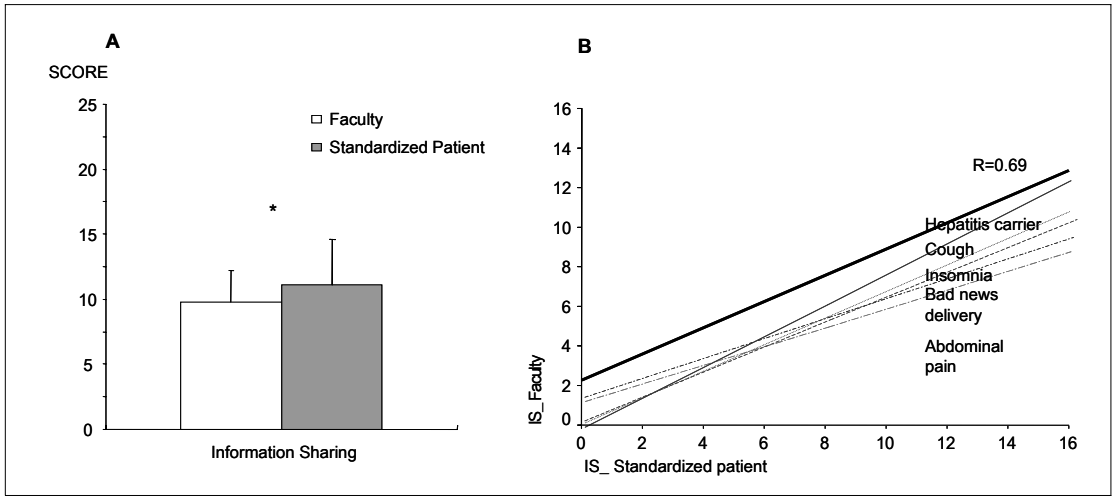

Fig. 5. Correlation between faculty and standardized patient evaluation in information sharing (* P < 0.05, by the paired t-test)

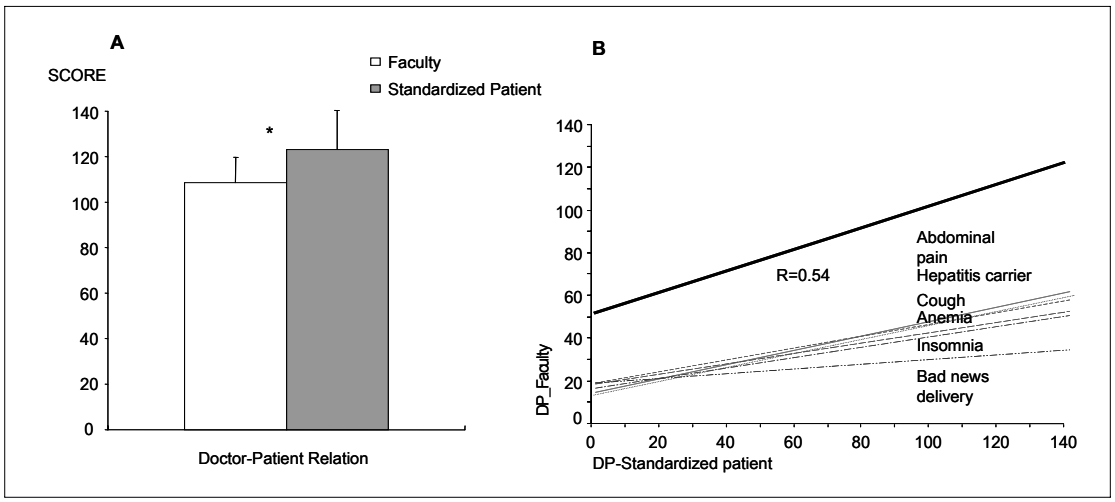

Fig. 6. Correlation between faculty and standardized patient evaluation in doctor-patient relation (* P < 0.05, by the paired t-test)

'불면증'은 0.93으로 높은 값을 보였다. 표준화 환자 평가와 교수 평가간의 상관계수는 각 증례에서 조금씩 다른 값을 보였으며 전체적인 상관계수는 0.91로 가장 높은 값을 보이며 유의한 상관관계를 보였다 (Fig. 4B).

3) 환자 교육 영역

교수평가가 9.6점으로 표준화 환자의 평가점수 11.0점보다 낮았으며 통계적으로 유의한 차이를 보였다 (Fig. 5A). 표준화 환자 평가와 교수 평가간의 피어슨 상관계수는 '복통' 증례에서 0.46, '나쁜 소식 전하기'에서 0.48, '불면증'에서 0.56, '기침'에서 0.61로 거의 비슷한 값을 나타내며 유의한 상관관계를 보였다. '간염보균자' 증례에서는 상관계수 0.69로 높은 값을 보였다. 표준화 환자 평가와 교수 평가간의 상관계수는 각 문항에서 조금씩 다른 값을

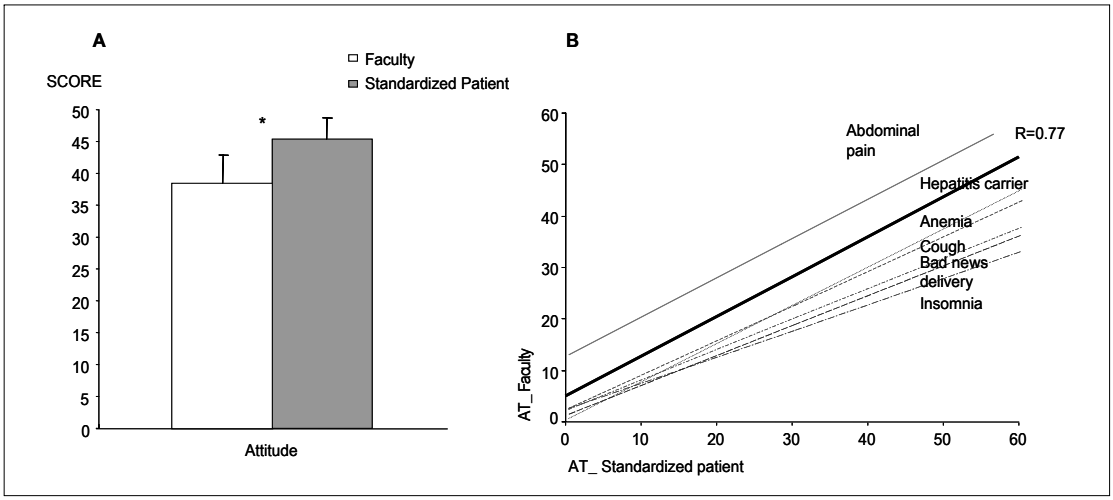

Fig. 7. Correlation between faculty and standardized patient evaluation in attitude (* P < 0.05, by the paired t-test)

보였으며 전체적인 상관계수는 0.69로 유의한 상관관계를 보였다 (Fig. 5B).

4) 의사-환자 관계 영역

교수평가가 116.3점으로 표준화 환자의 평가점수 123.6점과 통계적으로 유의한 차이를 보였다 (Fig. 6A). 표준화 환자 평가와 교수 평가간의 피어슨 상관계수는 '나쁜 소식 전하기' 증례에서 0.09로 가장 낮은 값을 보였다. '불면증' 증례에서는 상관계수 0.19, '빈혈'은 0.45, '기침'은 0.46, '간염보균자'는 0.50, '복통'은 0.51로 비슷한 값을 나타내며 유의한 상관관계를 보였다. 표준화 환자 평가와 교수 평가간의 상관계수는 각 문항에서 비슷한 값을 보였으며 전체적인 상관계수는 0.54로 가장 낮은 값을 보이며 유의한 상관관계를 보였다 (Fig. 6B).

5) 의사 태도 영역

교수평가가 39.2점으로 표준화 환자의 평가점수 44.4점과 통계적으로 유의한 차이를 보였다 (Fig. 7A). 표준화 환자 평가와 교수 평가간의 피어슨 상관계수는 '불면증' 증례에서 0.35로 낮은 값을 보였다. '나쁜 소식 전하기' 증례에서는 상관계수 0.51, '기침'은 0.56, '빈혈'은 0.67, '간염보균자'에서 0.68로 비슷한 값을 나타내며 유의한 상관관계를 보였다. '복통' 증례에서는 상관계수 0.77로 높은 값을 보였다. 표준화 환자 평가와 교수 평가간의 상관계수는 각 문항에서 조금씩 다른 값을 보였으며 전체적인 상관계수는 0.77로 비교적 높은 값을 보이며 유의한 상관관계를 보였다 (Fig. 7B).

라. 평가문항별 일치도

각 증례에서 교수와 표준화 환자 채점자간의 평가문항 별 일치도를 보기 위해 전반적 일치도 카파 (Kappa)를 구하였다. 복통 증례에서 두 채점자간의 점검표를 분석한 결과로 Q23이 카파 1.00으로 가장 높은 값을 나타내었고 Q4, Q12, Q9는 카파 0.70이상으로 강한 상관성을 보였다. 중등도의 상관성을 보인 것은 Q13, Q7, Q24로 각각 카파 0.49, 0.59였다. 0.01~0.35의 카파를 보인 것은 약한 상관성으로 가장 많았다 (Table II).

마. 신체진찰 영역과 의사-환자관계 영역에서 교수와 표준화 환자간의 채점표별 일치도

신체진찰 영역과 의사-환자관계 영역에서 교수와 표준화 환자 채점자간의 점검표별 일치도는 '교수와 표준화 환자가 동일하게 평가한 학생 수 / 표본화한 학생수'로 나타내었다. 0~9/24는 약한 상관성, 10~14/24는 중등도의 상관성, 15~24/24는 강한

Table II. Agreement between Faculty and Standardized Patient Evaluation in Each Items of Abdominal Pain Case

No	Item	Kappa
Q4	통증 양상	0.93
Q6	통증의 기간, 주기 및 지속성	0.25
Q7	복부 통증의 정도	0.59
Q8	통증과 연관된 증상	0.17
Q9	체중 변동	0.70
Q10	열이나 오한의 유무	0.24
Q12	대변과 함께 출혈의 유무	0.72
Q13	통증의 유발이나 완화 인자	0.49
Q14	대장암이나 용종의 가족력	0.35
Q15	과민성장증후군의 원인에 대한 설명	0.06
Q16	과민성장증후군의 치료에 대한 설명	0.31
Q23	손 씻기	1.00
Q24	환자의 몸 가려주기	0.59
Q27	전문가로서의 자질	0.26
Q29	적극적으로 경청	0.01
Q33	환자의 필요 충족	0.13

상관성을 보인 것으로 가정하였다. 신체진찰 영역에서는 대부분 강한 상관성을 보였으며 의사-환자관계 영역에서 약한 상관성을 보인 것은 48.8%, 중등도의 상관성을 보인 것은 41.5%, 강한 상관성을 보인 것은 9.7%로 대부분이 중등도 이하의 상관성을 보였다 (Table III).

고 찰

6개 증례 전체에서 표준화 환자와 교수의 평가 상관성은 r=0.79이었다. 본 연구는 전체적으로는 좋은 상관관계를 보여주지만 개별 증례 중 '나쁜 소식 전하기'의 상관계수가 0.33으로 나타났는데, 기존 연구에서는 표준화 환자와 교수간에 80-100%의 퍼센트 일치도를 보여주었고 '나쁜 소식 전하기'의 상관계수는 0.60로 보고하였다 (박훈기 등, 2003). 전반적으로 교수보다는 표준화 환자가 더 높은 점수를 주었으며 이는 기존의 연구에서도 확인된다 (김주자 등, 2004).

채점자간에 차이가 나는 원인으로는 점검표 자체의 애매한 채점 기준, 채점자간 일치도 향상 훈련 부족, 평가자의 피로도, 점검표의 항목 수와 척도 종류 등을 들 수 있다. 위의 차이에는 이 모든 요인이 함께 영향을 미친 것으로 생각된다. 그러나 신체진찰 영역에서 0.91의 높은 상관계수를 보인 것을 고려한다면 가장 큰 원인은 애매한 채점 기준에 있는 것으로 보인다. 그 다음으로 높은 상관계수를 보인 것은 '병력 청취'로 r=0.77이었다. 점검표를 검토해 보면 '신체진찰' 영역은 구체적인 행위를 수행하였는가를 3단계 척도 (제대로 했음/제대로 못했음/하지 않았음)로 나누어 표시하게 되어 있는데 이는 평가자의 훈련이 쉽고, 평가자 간에 의견의 일치를 보기도 역시 쉽기 때문에 높은 상관성을 보인 것이다. 또한 병력청취 역시 "...한 내용을 물어보았다"를 단지 '예/아니오'로만 표시하게 되어 있기 때문에 객관적인 검증이 가능하며 평가도 쉬웠던 것으로 생각된다.

반면 의사-환자관계 영역에서 상관계수는 0.54로 가장 낮았는데 "전문가로서 자질을 보여주었다.", "질병에 관한 정보를 효과적으로 수집했다", "나의 말을 적극적으로 들어주었다.", "나와 유대관계를 잘 형성하였다.", "내가 필요한 말을 할 수 있

Table III. Agreement between Faculty and Standardized patient Evaluation in Some Items of Physical Examination and Doctor-Patient Relationship

Cases	Physical Examination		Doctor-Patient Relation	
	Q	Number ratio	Q	Number ratio
Abdominal pain	Q18	18/24	Q28	8/24
	Q19	17/24	Q29	13/24
	Q20	14/24	Q30	12/24
	Q22	13/24	Q31	6/24
			Q32	12/24
			Q33	4/24
Cough	Q17	15/24	Q26	9/24
	Q18	20/24	Q27	8/24
	Q19	21/24	Q28	15/24
	Q20	13/24	Q29	15/24
			Q30	14/24
			Q31	14/24
			Q32	9/24
Insomnia	Q12	15/24	Q18	7/24
	Q13	17/24	Q19	4/24
	Q14	17/24	Q20	10/24
	Q15	11/24	Q21	9/24
	Q16	15/24	Q22	15/24
			Q23	9/24
			Q24	3/24
Bad news delivery			Q17	11/24
			Q18	4/24
			Q19	7/24
			Q20	5/24
			Q21	7/24
			Q22	10/24
			Q23	9/24
Hepatitis carrier	Q12	13/24	Q25	4/24
	Q13	22/24	Q26	5/24
	Q14	13/24	Q27	12/24
			Q28	8/24
			Q29	15/24
			Q30	14/24
			Q31	10/24
Anemia	Q16	22/24	Q28	10/24
	Q17	17/24	Q29	10/24
	Q18	18/24	Q30	14/24
	Q19	21/24	Q31	12/24
	Q20	20/24	Q32	11/24
	Q21	16/24	Q33	10/24
	Q22	21/24	Q34	7/24
	Q23	24/24		

도록 격려해 주었다.", "나의 느낌에 공감해 주었다.", "나의 필요를 충족시켜 주었다." 등의 다양한 해석이 가능한 물음에 최우수/아주 잘함/잘함/개선요망/최저 수준/수준 미달의 6단계로 평가하도록 한 점검표가 주된 요인이라고 생각된다. 이러한 양상은 기존 연구에서도 동일하게 나타나는데 순천향의과대학에서 실시한 CPX의 교수와 표준화 환자간 평가 차이에 관한 연구 (김주자 등, 2004)는 병력 청취 항목에서 Kappa값이 전반적으로 낮았으나 진찰소견, 진단 및 관리 항목에서는 Kappa값이 비교적 높았다고 보고하였다.

아울러 의사-환자관계에 대한 인식이 교수 평가자와 표준화 환자 평가자간에 동일하지 않을 수 있다는 점이 지적되어야 한다. 또한 학생과 직접 면담 중에 그의 태도를 평가하는 것과, 제3자가 촬영된 화면을 보면서 후에 평가하는 것은 방법론적으로 동일하다고 볼 수 없다. 따라서 CPX에서 의사-환자관계의 평가에 대해서는 보다 심도 있는 연구가 필요하다. 그리고 표준화 환자의 주관적인 느낌을, 제3자인 교수의 평가보다 우선시할 것인가에 대해서도 더 많은 논의가 필요할 것으로 본다.

6개 증례의 총점을 보면 교수 평가는 220.5점으로 표준화 환자 평가 245.1점보다 통계적으로 유의하게 낮았다. 특히 병력 청취 범주에서 교수 평가가 33.7점으로 표준화 환자의 43.0점보다 무려 10점이나 낮았다. 최근의 연구 (Heine et al., 2003)는 병력 청취에서의 표준화 환자 오류가 신체진찰에서보다 훨씬 많았으며 그 원인은 대부분 학생을 배려 (students' favor)하는 데서 기인한다. 그러므로 병력 청취 범주에 대해서는 표준화 환자의 평가 교육을 더욱 철저하게 하여야 할 것이다.

평가문항별 일치도를 보면 일치도가 강하게 나타난 문항, 중등도로 나타난 문항, 약하게 나타난 문항을 구별할 수 있다. 가장 강하게 나타난 문항은 "진찰을 하기 전에 손을 씻었는가"이다. 이처럼 문항의 의도가 분명하고 다양한 해석의 가능성이 없을수록 평가자 간의 일치도는 높게 나타날 수밖에 없다. 중등도로 나타난 문항들은 여러 질문을 그 의도로 해석할 여지가 있는 문항들이다. 예컨대 "복부 통증의 정도를 물어보았다."는 평가 문항은 다양한 질문에 해당될 수 있다. "얼마나 아프세요?", "많이 아프세요?", "심하게 아프세요?", "견딜만 하세요?"등의 질문이 "정도"로 해석될 수 있다. 여기서 표준화 환자의 학생에 대한 배려가 드러나는데 교수 평가자는 의학적으로 의미 있는 (medically meaningful) 질문을 한 학생에 대해서만 그러한 문항의 의도를 만족시켰다고 보는 반면, 표준화 환자는 비슷한 질문을 한 학생들도 모두 그렇게 여길 가능성이 있다. 그러므로 표준화 환자의 평가는 일관되게 교수 평가보다 높게 나타난다.

가장 일치도가 낮은 항목은 대단히 주관적인 평가항목들이다. "전문가로서의 자질을 보여주었다." (Kappa 0.26), "나의 말을 적극적으로 들어주었다."(Kappa 0.01) "나의 필요를 충족시켜 주었다" (Kappa 0.13)와 같은 문항은 표준화 환자가 과연 평가할 수 있는 항목인지 의심스럽다. 평가의 일치도, 재현성, 신뢰성 면에서 이러한 문항은 문제가 많다. 그러므로 표준화 환자를 사용한 CPX시험에서 의사-환자관계의 수행능력을 어떻게 평가하여야 하는지에 대해서는 더 많은 경험과 연구가 필요할 것이다.

결론적으로 평가의 일치도와 신뢰성을 높이기 위해서는 학생의 "행위 (performance)"에 초점을 맞춘 척도를 단순화한 점검표의 개발이 필요하다. 그러나 CPX의 원래 취지를 고려할 때 그렇게 하였을 경우 전체적인 임상수행능력을 본다는 평가의 의미가 줄어들 것이라는 우려가 있을 수 있다. 그러므로 이를 보완하기 위해서 보다 상세한 채점기준표의 개발, 그리고 예상되는 여러 상황에 대한 표준 평가 지침의 제작과 아울러 표준화 환자에 대한 철저한 교육이 필요할 것이다. 신체진찰과 병력 청취 영역에서 훈련된 표준화 환자의 평가는 교수 평가와 큰 차이를 보이지 않으므로 이 부분은 그대로 인정해도 좋을 것이다. 다만 학생의 전문가적 수행능력 (professional performance)을 평가할 때 고려할 사항으로, 환자의 입장에서 보는 전문가의 자질과 전문가인 교수의 입장에서 보는 자질 평가의 기준이 다를 수 있으므로 이를 동일하게 비교하기보다는 분리하여 비교하는 것이 합당할 것이며, 교수의 비디오 평

가는 이런 면에서 의미가 있을 것으로 생각된다.

마지막으로 이 연구는 표준화 환자의 평가 환경과 교수의 평가 환경이 근본적으로 다른 데서 기인하는 한계를 가진다. 이를 극복하기 위해서 표준화 환자와 동시에 교수가 함께 입실하여 평가를 하거나, 혹은 일방유리를 통해 직접 관찰을 하며 평가를 하는 등의 방법을 고려해 볼 수 있겠으나 역시 진료를 당하는 입장과, 제3자로서 관찰하는 입장이 다르다는 근원적인 문제를 해결하기는 어려울 것이다. 환자의 만족도나 의사-환자관계에서는 그러한 점이 특히 두드러진다. 평가 교수와 표준화 환자 평가자와의 지속적인 대화와 피드백, 그리고 상세한 표준평가 지침의 개발이 이 문제를 해결하는 데 도움이 될 것이다.

참 고 문 헌

김병수, 이영미, 안덕선, 박정률(2001). 임상의학입문 평가를 위한 객관적임상실기시험 경험. *한국의학교육*, 13(2), 289-298.

김주자, 이경재, 최규연, 이동환(2004). 일개 의과대학에서 실시한 표준화 환자(SP)를 이용한 임상수행능력평가시험(CPX)분석. *한국의학교육*, 16(1), 51-61.

박훈기, 이정권, 황환식, 이재웅, 최윤영, 김혁, 안동현(2003). 객관구조화진료시험(OSCE)에서 교수와 표준화 환자 사이의 점검표 채점의 일치도. *한국의학교육*, 15(2), 141-150.

서보양, 이두진, 권굉보, 강복수(1998). 객관적으로 구조화된 임상시험의 시행 경험. *한국의학교육*, 10(2), 363-381.

Abrahamson S(1998). 임상실기능력을 어떻게 평가할 것인가? *한국의학교육*, 10(1), 153-164.

Heine N, Garman K, Wallace P, Bartos R, Richards A(2003). An analysis of standardized patient checklist errors and effect on student scores. *Med Edu*, 37(2), 99-104.

Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G(1996). Evaluating communication skills in the objective structured clinical examination format: reliability and generalizability. *Med Edu*, 30, 38-43.

Yelland M(1998). Standardized patients in the assessment of general practice consulting skills. *Med Edu*, 32, 8-13.

Wind L(2004). Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). *Med Edu*, 38, 39-44.