대한진단검사의학회지 제26권 제4호 2006
Korean J Lab Med 2006;26:299-306



# 체외진단키트 성능평가를 위한 웹기반의 검체 수 산정 프로그램 개발

황상현[1] · 오흥범[2] · 채정민[3] · 서민관[3] · 정순영[3] · 최성은[4] · 이관제[4]

부산대학교 의학전문대학원 부산대학교병원 진단검사의학과[1], 울산대학교 의과대학 서울아산병원 진단검사의학과[2], 고려대학교 컴퓨터교육학과[3], 동국대학교 통계학과[4]

# Development of a Web-based Program to Calculate Sample Size for Evaluating the Performance of In Vitro Diagnostic Kits

Sang-Hyun Hwang, M.D.[1], Heung-Bum Oh, M.D.[2], Jeong-Min Chae[3], Min-Kwan Seo[3], Soon-Young Jung, Ph.D.[3], Sung-Eun Choi[4], Kwan Jeh Lee, Ph.D.[4]

Departments of Laboratory Medicine, Pusan National University Hospital, Pusan National University School of Medicine[1], Busan; Department of Laboratory Medicine, Asan Medical Center, University of Ulsan College of Medicine[2], Seoul; Department of Computer Science Education, Korea University[3], Seoul; Department of Statistics, Dongkuk University[4], Seoul, Korea

**Background** : Many studies evaluating the performance of in vitro diagnostic kits have been criticized for the lack of reliability. To attain reliability those evaluation studies should be preceded by sample size calculation ensuring statistical power. This study was intended to develop a web-based system to estimate the sample size, which was often neglected because it would require expert knowledge in statistics.

**Methods** : For sample size calculation, we extracted essential parameters from the performance studies on the 3rd generation anti-hepatitis C virus (HCV) kits reported in the literature. We developed a system with PHP web-script language and MySQL. The statistical models used in this system were as follows; one sample without power consideration (model 1), one sample with power consideration (model 2), and two samples with power consideration (model 3).

**Results** : Among the articles published between 1989 and 2005, 13 articles that evaluated the performance of anti-HCV kits were identified by searching with Medical Subject Headings (MeSH). The diagnostic sensitivity was 83-100% with a median of 145 samples (range; 12-1,091) and the specificity was 97-100% with a median of 1,025 samples (range; 33-4,381). The estimated sample size would be 280 in the model 1, 817 in the model 2, and 1,510 in the model 3, when we set 2% prevalence of HCV infection, 95% sensitivity of a conventional kit, 97% sensitivity of a new kit , 95% significance level (two-sided test), 2% allowable error, and 80% power.

**Conclusions** : Our study indicates that an insufficient sample size is still a problem in performance evaluation. Our system should be helpful in increasing the reliability of performance evaluation by providing an appropriate sample size. *(Korean J Lab Med 2006;26:299-306)*

Key Words : *Sample size calculation, Performance Evaluation, Sensitivity, Specificity In vitro diagnostics, HCV*

**접 수 :** 2006년 1월 25일 **접수번호 :** KJLM1920
**수정본접수 :** 2006년 4월 13일
**게재승인일 :** 2006년 5월 1일
**교 신 저 자 :** 오 흥 범
우 138-736 서울시 송파구 풍납동 388-1
서울아산병원 진단검사의학과
전화 : 02-3010-4505, Fax : 02-478-0884
E-mail : hboh@amc.seoul.kr

*본 연구는 식품의약품 안전청에서 2005년도에 시행한 용역연구개발사업의 연구결과입니다.

## 서 론

혈액을 이용한 체외진단키트는 그 동안 주로 질병 진단을 목적으로 사용되었으나 최근에는 치료 후 추적관찰, 예후예측, 질병예방의 목적으로까지 그 영역이 확대되고 있다. 따라서 검사키트 성

능에 대한 정확하고 신뢰성 높은 평가와 해석이 필요하게 되었다. 체외진단키트의 성능평가 항목에는 여러 가지가 있지만 가장 중요한 항목은 진단민감도, 진단특이도, 기존 검사와의 일치율이라 할 수 있다[1-4]. 지금까지 이런 주요 항목에 대한 성능평가 연구들이 발표되었으나 신뢰성이 떨어지는 보고가 많다는 지적이 있었다. Reid 등[5]의 연구에 의하면, 1978년부터 1993년 사이 4개의 주요 의학저널에 발표된 1,302건의 성능 평가 연구들에 대한 조사에서 50% 정도만이 적절한 방법을 적용한 것으로 나타났다. 2004년의 한 보고에서도 성능평가에 있어 여전히 방법적인 문제가 있음이 지적되었다[6].

검체 수 산정은 체외진단키트의 올바른 성능평가를 위해 반드시 필요한 과정이다. 필요한 환자수보다 적은 환자수를 사용하면 기존 진단키트와의 성능차이를 정확하게 추정할 수 없다. '대조군과 차이가 없다'는 결론을 내린 71개의 무작위 환자-대조군 임상시험 중 57개(80%) 연구는 불충분한 검체 수로 인한 것이라는 보고가 있었다[7]. 반면 필요한 환자수보다 더 많은 환자 수를 사용하는 것은 비용-효율성이 떨어지며 일부에서는 환자의 임의추출이 어려워 잘못된 결론에 도달할 수 있다는 보고가 있었다[8]. 검체 수 산정 과정은 통계학적 전문지식을 요구하기 때문에 종종 무시되어 온 것이 사실이다.

본 연구에서는 3세대 C형 간염바이러스(hepatitis C virus, HCV) 항체 검사를 중심으로 기존 성능평가의 문제점을 알아보고, 아울러 성능평가에 필요한 요소들을 찾아낸 후 이를 토대로 검체 수 산정 프로그램을 웹기반으로 개발하여 누구나 쉽게 이용할 수 있도록 하고자 하였다.

## 재료 및 방법

### 1. HCV 항체 검사의 성능평가 연구문헌 검색

본 연구에서는 1989년부터 2005년 동안 MEDLINE 데이터베이스에서 HCV와 관련된 주제를 검색하였다. Medical Subject Headings (MeSH) 용어로 "hepatitis C", "mass screening", 'hepatitis C antibodies", "sensitivity and specificity", "diagnostic accuracy"를 사용하여 검색하고, 3세대 HCV 항체검사 키트의 성능평가 문헌 중에서 검체 수, 진단민감도, 진단특이도, 대상 집단, 참고검사법 등을 추출하였다. 각 연구 결과의 진단민감도, 진단특이도의 95% 신뢰구간(confidence interval, CI)은 직접 계산하여 도식화하였다[9].

### 2. 검체 수 산정 프로그램 개발

검체 수 산정프로그램을 웹기반으로 www.koreanhla.com에 구현하였다. 개발도구로는 공개 소프트웨어인 PHP 웹 스크립트 언어와 MySQL 데이터베이스를 사용하였다. 검체 수 산정을 위한 통계 모델로 다음 세 가지 방법을 이용하였다. 1) 검정력을 고려하지 않은 단일 집단에서의 검체 수 산정(모델 1, one sample without power consideration, one sample with power consideration), 2) 검정력을 고려한 단일 집단에서의 검체 수 산정(모델 2, one population with power consideration), 3) 검정력을 고려한 두 집단에서의 검체 수 산정(모델 3, two populations with power consideration) 통계 모델이었다.

입력 화면은 성능을 평가하고자 하는 새로운 체외진단키트(new kit)와 비교 대상이 되는 기존 체외진단키트(reference kit)의 진단민감도, 진단특이도, 허용오차, 유의수준, 검정력, 검정방법, 검체 수 산정모델을 선택한 후 "검체 수 산정" 버튼을 누르면 적절한 검체 수가 계산되도록 구성하였다(Fig. 1, 2A). 또한, 진단민감도, 진단특이도, 유의수준, 검정력의 변화에 따른 검체 수의 변화를 한 눈에 알아볼 수 있도록 검체 수 "matrix 보기" 기능을 구현하였다(Fig. 2B).

### 3. 검체 수 산정에 사용된 통계모델

세 가지 통계 모델은 Arkin 등[10]의 논문을 참고로 하였다. 유병률을 고려하는 경우는 Nancy 등[11]의 논문을 참고로 하였는데, 세 가지 통계 모델을 통해 계산된 검체 수(n)를 유병률(p)로 나누는 것으로 다음의 식과 같다.

유병률을 반영한 검체 수 $N = n/p$

1) 검정력을 고려하지 않은 단일 집단에서의 검체 수 산정(one

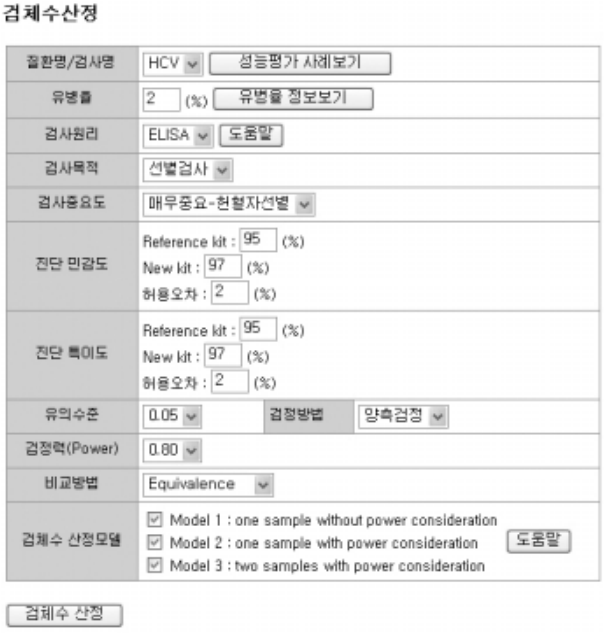


Fig. 1. The screenshot of the main input menu. The statistical parameters including sensitivity, specificity and allowable error limit were assigned. Significant level, power, and statistical models can be chosen via a select box or a check box, respectively.

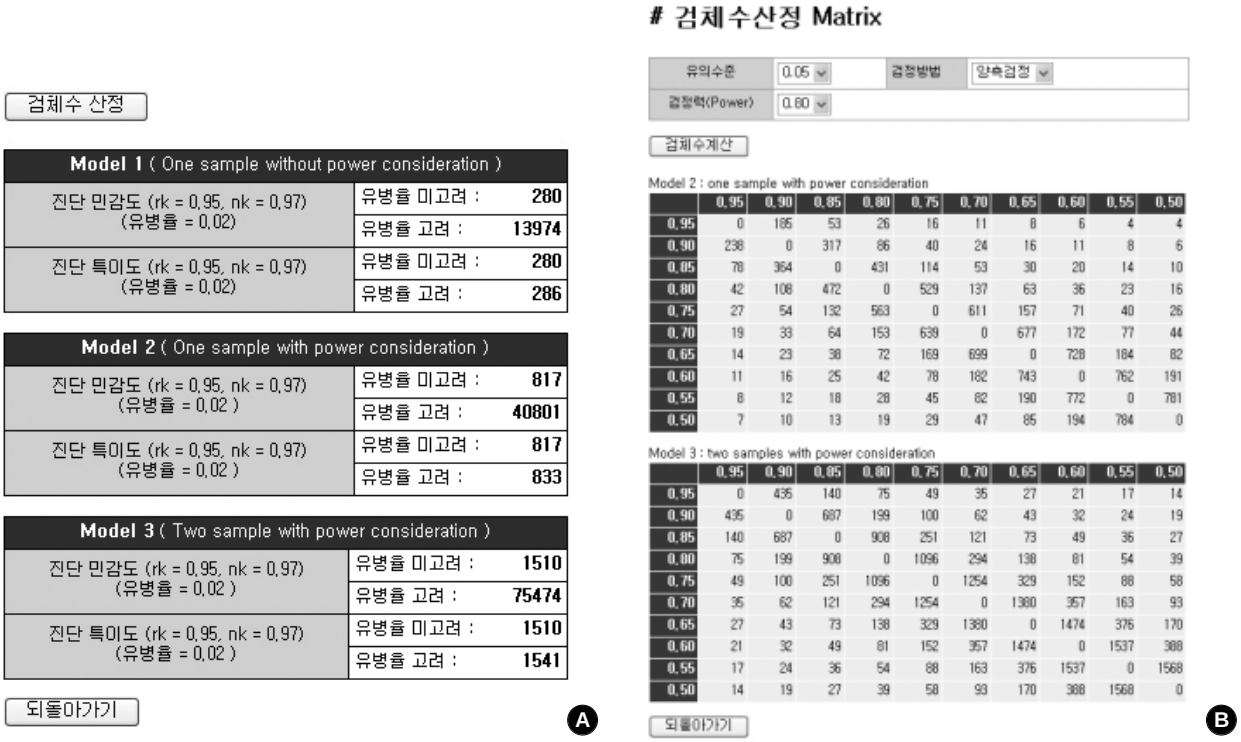


Fig. 2. Output screenshot of the program. (A) Sample size was estimated by giving prevalence of HCV=2%, sensitivity of the reference test=95%, sensitivity of a new test=97%, significant level (two-sided)=95%, allowable error=2%, and power=80%. (B) When the power and significant level were selected, the estimation of sample size for comparison of two in vitro diagnostic devices was tabulated by the statistical models.

sample without power consideration)

$$n = Z_{\alpha 2}^{2}\, p(1-p)/r^{2}$$

$n$: 필요 검체 수

$Z_{\alpha 2}$: 양측검정에서의 Z값(95% 신뢰구간에서는 $Z_{\alpha 2}=1.96$)

$r$: 허용 오차 혹은 신뢰구간의 절반

$p$: 진단민감도 또는 진단특이도

2) 검정력을 고려한 단일 집단에서의 검체 수 산정(one sample with power consideration)

$$n = \left\{ \frac{Z_{\alpha 2}\sqrt{\pi_a (1-\pi_a)} + Z_{\beta 1}\sqrt{\pi_b (1-\pi_b)}}{\pi_a - \pi_b} \right\}^2$$

$n$: 표본크기(필요 검체 수)

$\pi_a$: 새로운 체외검사키트 A에서의 진단민감도 또는 진단특이도

$\pi_b$: 예상되거나 문헌에 알려져 있는 체외검사키트의 진단민감도 또는 진단특이도

$Z_{\alpha}$: $1-\alpha = pr\{Z | -Z_{\alpha 2} \le Z \le Z_{\alpha 2}\}$

$Z_{\beta}$: $1-\beta = pr\{Z | Z \ge Z_{\beta 1}\}$

3) 검정력을 고려한 두 집단에서의 검체 수 산정(two sample with power consideration)

$$n = \left\{ \frac{Z_{\alpha 2}\sqrt{2\pi_0 (1-\pi_0)} + Z_{\beta 1}\sqrt{\pi_a (1-\pi_a) + \pi_b (1-\pi_b)}}{\pi_a - \pi_b} \right\}^2$$

n: 표본크기(필요 검체 수)

$\pi_a$: 비교하고자 하는 체외검사키트 A의 진단민감도 또는 진단특이도

$\pi_b$: 새로운 체외검사키트 B의 진단민감도 또는 진단특이도

$\pi_0$: 두 비율이 동일하다는 가정 하에서 결합비율

$Z_{\alpha}$: $1-\alpha = pr\{Z | -Z_{\alpha 2} \le Z \le Z_{\alpha 2}\}$

$Z_{\beta}$: $1-\beta = pr\{Z | Z \ge Z_{\beta 1}\}$

## 결 과

### 1. HCV 항체검사 성능평가 문헌검색 결과

1989년에서 2005년 사이에 보고된 문헌들 중에서, "hepatitis C virus"와 "mass screening" MeSH 용어의 조합으로 검색한 경우 233건의 문헌이 추출되었는데, 이 중 1개의 문헌이 HCV 항체검사의 성능평가와 관련된 것이었다. "hepatitis C antibodies" 및 "sensitivity and specificity" 검색의 경우는 351건의 문헌이

Table 1. Performance evaluation studies of hepatitis C virus serological assays

| Studies | Reference | Assay | Spectrum of patients | Reference standard |
|---|---|---|---|---|
| Stuyver | [21] | Ortho HCV 3.0 | Blood donors with persistently increased ALT and histological chronic hepatitis or steatosis | HCV RNA by RT-PCR |
| Lavanchy | [22] | Ortho HCV 3.0 | Hemodialysed patients | HCV RNA by RT-PCR |
| Courouce | [23] | Cobas Core anti-HCV EIA | 2500 sera from patients and blood donors | ELISA 2.0 and RIBA 3.0 |
| Hennig (a) | [24] | AxSYM HCV version 3.0 | 4383 blood donors | Abott Matrix HCV 2.0 and HCV RNA by RT-PCR |
| Hennig (b) | [24] | IMX HCV version 3.0 | | |
| Jonas (a) | [25] | ARCHITECT Anti-HCV | 3811 blood donors, 1984 randomly selected clinical specimens | RIBA |
| Jonas (b) | [25] | ARCHITECT Anti-HCV | | |
| Abdel-Hamid | [26] | Abott HCV EIA 3.0 | 1134 serum samples collected in a community-based study | RIBA and RT-PCR |
| Zachary (a) | [27] | Monolisa anti-HCV Plus version 2 | 2020 routine serum samples | |
| Zachary (b) | [27] | AxSYM HCV version 3.0 | 253 anti-HCV seropositive patients, | RIBA and RT-PCR |
| Zachary (c) | [27] | Ortho Vitros ECi anti-HCV | 394 anti-HCV negative blood donors | |
| Judd (a) | [28] | OraSure (Epitope, Beaverton, OR, USA)-Ortho HCV 3.0 SABe | | RIBA and RT-PCR |
| Judd (b) | [28] | Salivette (Sarstedt, Leicester, UK)-Ortho HCV 3.0 SABe | | |
| Ismail | [29] | Ortho Vitros ECi Anti-HCV | 177 anti-HCV-seropositive samples (Abbott EIA) | RIBA 3.0 |
| Huber | [30] | Cobas Core anti-HCV EIA | 1090 patients with acute liver disease or suspected chronic hepatitis | HCV RNA by RT-PCR |
| Prince | [31] | ELISA 3.0, unspecified | 301 blood donors with elevated ALT levels | HCV RNA by RT-PCR |
| Vrielink (a) | [32] | Abbott HCV EIA 3.0 | 403 blood donor samples, 212 non-A, non-B hepatitis patients, 253 multi-transfused patients, 1055 first-time blood donors | HCV RNA by RT-PCR and RIBA 2.0 |
| Vrielink (b) | [32] | Murex anti-HCV | | |
| Vrielink (c) | [32] | Ortho HCV 3.0 | | |
| Busch | [33] | ELISA 3.0, unspecified | 1091 blood donors | HCV RNA by PCR and RIBA 2.0 |

Table 2. Comparison of confidence intervals and *P* value for the assessment of a case-control difference (confidence interval of control; -4, 4)

| Category | Confidence interval | Assessment using confidence interval | Assessment using *P* value |
|---|---|---|---|
| A | (7,15) | Superior to control | Significant |
| B | (1,9) | Non-inferior to control | Significant |
| C | (-1,7) | Non-inferior to control | Non-significant |
| D | (1,4) | Equivalent to control | Significant |
| E | (-1,2) | Equivalent to control | Non-significant |
| F | (-9,-1) | Inferior to control | Significant |
| G | (-7,1) | Inferior to control | Non-significant |
| H | (-15,-7) | Inferior to control | Significant |
| I | (-7,9) | Inconclusive | Non-significant |

추출되었으며 이 중 48개 문헌이 성능평가와 관련이 있었다. 이들 48개 문헌 중 1개는 3세대 HCV 항체검사 성능평가에 대한 계통적 문헌고찰 논문이었다[12]. "hepatitis C antibodies"와 "diagnostic accuracy"로는 14개의 문헌이 추출되었고, 이 중 2건이 성능평가와 관련이 있었다. 이렇게 검색 추출한 51개의 문헌 중에서 1세대, 2세대 HCV 항체 검사와 면역 블롯검사 등 본 연구 주제와 관련이 없는 48개의 연구 문헌을 제외한 10개의 문헌이 3세대 항체검사의 성능평가 연구였다. Colin 등의 문헌고찰 논문을 통하여 추가적으로 3개의 관련문헌을 찾을 수 있어서 총 13개 문헌을 얻을 수 있었다(Table 1).

평가된 제품들은 Ortho HCV 3.0 (Ortho Diagnostic Systems, NJ, USA), Abbott HCV EIA 3.0 (Abbott Diagnostic, Abbott Park, IL, USA), Abbott AxSYM HCV version 3.0 (Abbott Diagnostic), Abbott IMX HCV version 3.0 (Abbott Diagnostic), Abbott ARCHITECT Anti-HCV (Abbott Diagnostic), Monolisa anti-HCV plus version 2 (Bio-Rad, Marnes-La-Coquette, France), Ortho HCV 3.0 SABe (Ortho Diagnostics, Amersham, UK), Ortho Vitros ECi Anti-HCV (Ortho Diagnostic Systems), Cobas Core anti-HCV EIA (Roche Diagnostics, Basel, Switzerland), Murex anti-HCV (Murex Diagnostics, Dartford, UK)이었다.

전체 13개의 성능평가 연구 중 5개는 모든 양성 검체에 대하여 HCV reverse transcriptase-polymerase chain reaction (RT-PCR)을 참고검사법으로 이용하였다. 체외진단키트의 성능평가 연구들의 진단민감도는 83-100%로 대상 환자군의 특성과 검체 수에 따라 차이가 있었고, 진단민감도 평가를 위한 HCV 감염 환

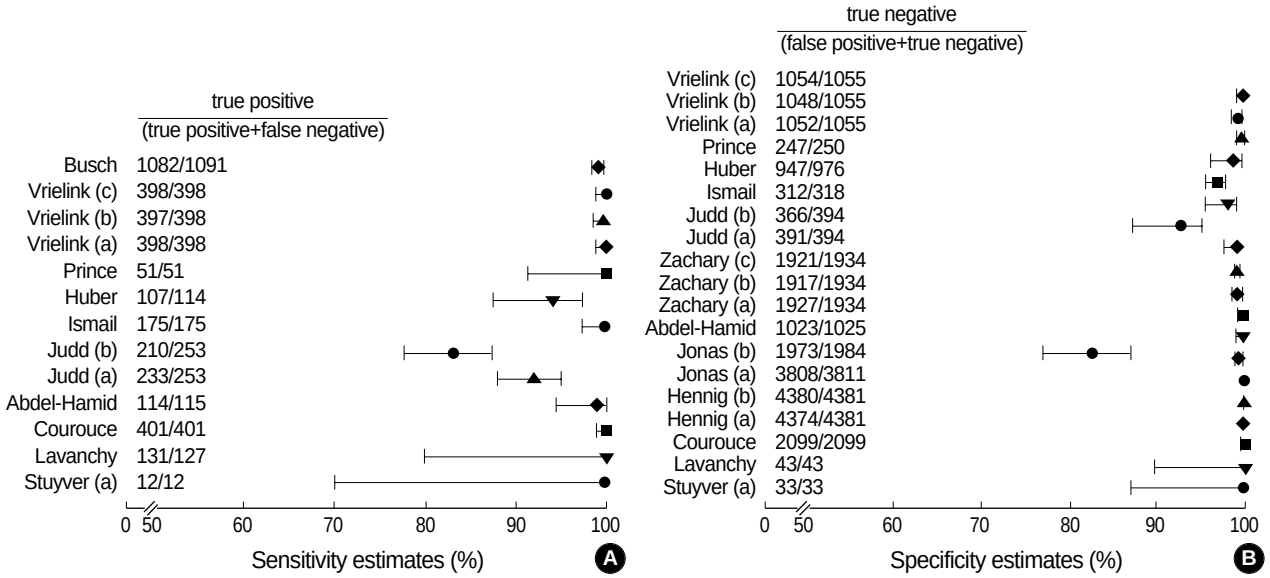


Fig. 3. Estimates from the studies of sensitivity and specificity of anti-HCV antibody tests. Points indicate estimates of sensitivity (A) and specificity (B). Horizontal lines are 95% confidence intervals for the estimates.

자 수는 중앙값이 145명(12-1,091)이었다. 진단특이도는 97-100%였으며 대부분의 연구에서 대조군 검체 수는 중앙값이 1,025명(33-4,381명)이었다(Fig. 3).

### 2. 개발된 프로그램을 이용한 검체 수 산정

검체 수 산정을 위하여 필요한 파라미터 값으로 1) 검사명, 2) 유병률, 3) 검사원리, 4) 검사목적, 5) 검사중요도, 6) 진단민감도, 7) 진단특이도, 8) 유의수준, 9) 검정력, 10) 비교방법, 11) 검체 수 산정모델을 선정하였다. 3), 4), 5) 항은 검체 수 산정에 필요한 파라미터는 아니나 기본 정보로 입력하도록 하였고 비교방법으로는 동등성(equivalence) 비교가 자동으로 선택이 되도록 하였다. 모델 2와 모델 3의 경우는 두 체외진단키트의 진단민감도가 50%에서 95%까지 5% 단위로 변화할 때 필요한 검체 수를 matrix 형태로 표현되도록 구현하였다.

HCV 유병율 2%, 기존 검사법의 진단민감도 95%, 새로 개발된 체외진단키트의 진단민감도 97%, 허용오차가 2%, 양측검정, 유의수준 0.05, 검정력 80%, 동등성 비교 등의 파라미터를 입력하고 검체 수를 산정하는 입력화면 구성은 Fig. 1과 같고 그 결과는 Fig. 2A와 같다. 모델 1에서는 280명, 모델 2에서는 817명, 모델 3에서는 1,510명의 검체가 산정되었다. 50%에서 95%까지 5%단위로 변화할 때 필요한 검체 수는 Fig. 2B와 같았다.

## 고 찰

본 연구에서는 3가지 통계모델을 프로그램에 구현하였다. 모델 1 (one sample without power consideration)은 검정력을 고려하지 않는 경우이며 모델 2 (one sample with power consideration)는 검정력을 고려한 경우의 검체 수 산정 모델이다. 두 모델은 기존 체외진단키트의 성능을 문헌고찰을 통해 알아보고 새로운 체외진단키트의 성능을 임상적 요구에 부합되는 범위에서 설정한 후 이를 검증하는 경우에 사용되는 검체 수 산정 방법이다. 모델 1과 모델 2에서는 기존 체외진단키트와 직접 비교를 시행하는 것은 아니며 임상상이 잘 정의된 검체 혹은 참고검사법에 의해 이미 검사가 이루어진 검체를 사용하는 경우이다.

통계적 가설검정에는 제1종 오류와 제2종 오류가 발생한다. 제1종 오류는 두 체외진단키트의 성능이 실제로는 동일한데, 유의한 차이가 있다고 판단하는 오류이다. 이것의 발생 확률을 $\alpha$라 하고 일반적으로는 0.05로 설정한다. 제2종 오류는 두 체외진단키트의 성능이 실제적으로는 차이가 있음에도 불구하고 차이가 없다고 판단하는 오류이다. 제2종 오류의 발생확률을 $\beta$라 하면 검정력은 제2종 오류의 여집합 즉 1-$\beta$로 결정된다. 따라서 검정력은 두 체외진단키트의 성능이 실제적으로 차이가 있을 때 차이가 있다고 결론을 내릴 수 있는 확률이다. 체외진단키트의 성능평가 연구에서 두 체외진단키트 성능 간 통계적으로 유의한 차이를 발견하지 못했을 경우 검정력이 낮아 발생한 문제가 아닌지 반드시 확인하여야 한다. 그 동안 대부분의 성능평가 연구에서는 이러한 제2종 오류를 간과해 온 경향이 있다. 따라서 본 연구에서는 모델 1과 모델 2를 구분하여 검체 수를 제시하였다.

모델 3 (two sample with power consideration)은 검정력을 고려한 상태에서 기존 체외진단키트와 새로운 체외진단키트를 동시에 직접 비교하는 경우의 검체 수 산정방법이다. 통상적으로 검정력은 80%를 사용하지만 연구자가 검사의 중요도에 따라 조정할 수 있다. 충분하지 못한 검체 수로 연구를 수행하면 검정력이 낮아 기존의 우수한 성능을 보이는 체외진단키트와 차이가 없다는

결론을 내리게 되어, 마치 기존의 체외진단키트와 동등한 성능을 보이는 것으로 오해를 일으킬 수 있다.

최근 유럽의 기준에 따르면[13], HCV 항체 검사의 진단민감도 평가를 위해서는 약 400개의 HCV 양성 환자의 검체를 검사하도록 규정하고 있다. 이는 본 프로그램의 모델 1에서 95%의 신뢰수준으로 약 2%의 오차 내에서 95% 정도로 예상되는 HCV 항체 검사의 진단민감도를 구할 때 필요한 검체 수에 해당한다(필요 검체 수=457). 또한 HCV 항체 검사의 진단특이도 평가를 위해 최소 5,300개 이상의 검체로 평가하도록 요구하고 있다[13]. 이는 본 프로그램의 모델 2에서 참고검사법의 진단특이도 95%, 평가대상 검사의 진단특이도가 96%이고, 80% 검정력으로 비교하고자 하는 경우에 5,299개 이상의 검체가 필요한 것과 유사한 결과라고 할 수 있다. 본 연구에서 기존의 HCV 항체 검사 성능평가 연구를 고찰한 결과, 진단민감도 평가를 위한 HCV 감염 환자 수는 중앙값이 145명(12-1,091), 진단특이도 평가를 위한 대조군의 중앙값은 1,025명(33-4,381명)으로 유럽의 기준에 훨씬 미치지 못함을 알 수 있었다. 이는 Fig. 3에서 보듯이 신뢰구간의 폭이 넓어 신뢰성이 낮음을 시사한다. 따라서 향후에는 성능평가 수행 전에 적절한 검정력 또는 신뢰구간 확보를 위한 검체 수 산정과정이 필요하다고 여겨지며 본 연구에서 개발된 프로그램이 큰 도움이 될 것으로 사료된다.

신뢰성 있는 체외진단키트의 성능평가를 위해서는 검체 수 산정에 앞서 적절한 연구 설계가 필수적이다. Rutjes 등[14]은 대조군을 정상인으로만 구성하는 연구 설계에서는 의사환자군을 대조군으로 사용한 경우보다 체외진단키트 성능이 2-3배 더 과장되어 보일 수 있음을 보고하였다. 이러한 오류를 피하기 위해서는 체외진단키트의 경우에도 신약 임상시험의 경우와 유사하게 제1상, 제2상, 제3상 및 제4상 등의 단계별 평가가 필요하다[15, 16]. 이는 처음부터 모든 가능한 경우를 고려한 평가를 시행하는 것보다 비용-효율적인 측면도 있고 해당 체외진단키트의 효과를 체계적으로 조사할 수 있는 방법이다.

제1상은 정상인을 대상으로 한 연구로서 정량검사의 경우는 참고치 설정, 정성검사의 경우는 cut-off 설정들을 시행하는 것이다. 제2상에서는 환자-대조군 연구를 시행하는 것으로 2상은 세부적으로 phase IIa, IIb, IIc로 나눌 수 있다. Phase IIa는 환자-대조군 성능평가, phase IIb는 질환의 중증도에 따라 환자군을 더 세분화하고 대조군과의 차이를 비교함으로써 spectrum bias에 의한 차이를 없애는 것이며, phase IIc는 환자군과 유사한 임상증상을 보이는 대상을 대조군에 포함하는 것이다. 제2상까지가 체외진단키트의 기본 성능평가라 할 수 있다. 제2상 연구결과의 신뢰를 높이기 위해 1999년부터 standards for reporting of diagnostic accuracy (STARD) initiative가 시작되었다[17]. 이는 신약 임상시험에 대한 연구결과 보고 발의안인 consolidated standards of reporting trials (CONSORT) 모델에 따라 체외진단키트 성능평가 연구보고의 질적 향상을 도모하고자 한 것이다. 이를 위해 STARD에서는 점검표를 제시하고 있는데 이에 대한 이해와 적용이 국내에서도 필요할 것으로 사료된다.

단일인구 집단을 대상으로 무작위 검사를 시행하거나 일정 기간 동안 전향적으로 연속검사를 시행하는 제3상 혹은 제4상 연구 설계의 경우에는 단일 모집단을 대상으로 성능을 평가하는 것이므로 유병율을 고려한 검체 수 산정이 필요하다. 이런 경우에는 매우 많은 검체가 필요하게 되는데, 예를 들어 유방암 유병률이 1%라고 했을 때 50명의 유방암 환자가 포함되기 위해서는 약 5천명의 일반인을 대상으로 연구를 시행해야 한다. 따라서 드문 질환의 경우는 현실적으로 적용하기 힘든 문제점을 안고 있다. 일반적으로 체외진단키트의 허가는 제2상 수준에서 이루어지지만 HCV 같은 헌혈자 선별검사 등은 시판 후 제3상, 제4상 단계의 평가를 고려해야 할 것이다. 이를 위하여 본 프로그램에서는 검사의 중요도라는 파라미터를 도입하였으나 현재의 프로그램에서는 검체 수 산정에 사용하지 아니하였다. 향후 다양한 통계모델을 도입하였을 경우 어떤 모델을 적용하는 것이 타당한지를 가늠하는 지표가 될 수 있을 것이다.

새로 개발된 체외진단키트의 허가를 위한 성능평가에는 흔히 기존 체외진단키트와의 동등성 여부를 알아보는 것이 일반적이다. 기존 체외진단키트와 비교하는 방법에는 동등성 비교(equivalence test) 외에도 비열등성 비교(non-inferiority test), 우월성 비교(superiority test) 등이 있다. 동등성 비교란 두 체외진단키트의 성능이 동등하다는 통계적 귀무가설에 대한 가설검정을 통하여 판정하는 것이라면 후자는 신뢰구간과 임상의의 요구를 종합하여 판정하는 것이다. 적은 수의 검체를 이용하면 새로운 체외진단키트의 성능이 기존의 체외진단키트와 동등하다는 잘못된 판정을 내릴 우려가 있는 반면, 많은 수의 검체를 이용하면 새로운 체외진단키트와 기존 체외진단키트 성능 간에 차이가 있다고 판정되어 새로운 체외진단키트가 우수하지 않으면 판매허가를 얻지 못하는 문제가 발생하게 된다. 이런 경우 새로운 체외진단키트가 기존키트에 비해 우수하지 않더라도 임상적 요구에 부합하면 적절한 체외진단키트로 판정할 수 있어야 한다.

Table 2는 신뢰구간을 토대로 임상적 판단을 수반하여 결정을 내리는 경우와 $P$ 값을 토대로 결정을 내리는 경우를 비교하고 있다. Table 2에서 임상의는 신뢰구간(-4, 4)이면 동등한 효과로 판정하고자 하는 것이다. 유사 성능의 체외진단키트와 비교하는 경우에는 일반적으로 동등성(equivalence) 평가를 위한 검체 수를 구한 후 성능을 평가할 수 있겠지만 참고검사법과 비교하는 경우에는 비열등성(non-inferiority) 비교를 위한 검체 수를 이용하여 비교평가를 시행한 후 신뢰구간을 토대로 한 판단이 필요할 것이다. 본 연구에서는 우월성 혹은 비열등성 비교를 위한 검체 수 산정 방법에 대해서는 구현하지 않았다. 향후 이에 대한 프로그램 보완이 필요하다.

검체 수 산정을 위해서는 우선 기존에 시판되고 있는 체외진단키트들의 성능을 고찰해야 하는데, 그 동안 사용되어 온 방법은 주요 저널의 검색과 review 저널의 참고문헌 검색 등이었다. 저널 검색은 MEDLINE과 EMBASE를 이용하고 있는데, 이를 이

용한 문헌 고찰의 문제점은 두 데이터 베이스간 일치율이 약 35% 정도로 낮고[18], MeSH 용어로 "sensitivity, specificity"로 찾는다면 1992년에서 1995년 사이의 전체 성능평가 연구의 51% 정도 밖에는 찾을 수 없다는 것이다[19]. 따라서 체외진단키트의 성능평가 문헌검색 및 문헌의 주요 정보를 데이터베이스화하는 체계적인 연구가 향후 필요할 것이다. 또한 인터넷 상의 많은 문헌을 토대로 한 meta-analysis가 중요하게 되었다. 이를 위해서는 체외진단키트의 진단민감도와 진단특이도를 보고할 때 신뢰구간을 함께 보고해야 한다[20].

결론적으로, 본 연구를 통해 저자들은 세 가지 통계모델을 이용한 검체 수 산정 프로그램을 개발할 수 있었다. 웹 기반으로 개발하여 누구나 접근이 용이하였기 때문에 진단검사의학 분야에서 널리 활용될 수 있을 것으로 기대된다. 향후 다양한 통계모델을 구현하고 다양한 실험디자인에서 적절한 통계모델은 무엇인지를 제시하는 프로그램 보완이 지속적으로 이루어져야 할 것이다.

## 요 약

**배경 :** 체외진단키트의 성능평가 연구보고에 신뢰성이 떨어진다는 지적이 있어왔다. 올바른 성능평가를 위해서는 적절한 검체 수 산정이 우선되어야 한다. 그러나 검체 수 산정 과정은 통계학적 전문지식을 요구하기 때문에 종종 무시되어 온 것이 사실이다. 본 연구를 통해 웹기반으로 검체 수 산정 프로그램을 개발하고자 하였다.

**방법 :** 3세대 hepatitis C virus (HCV) 항체의 성능평가 연구들에 대한 문헌 고찰을 통해 기존 연구의 신뢰도를 분석하였고 아울러 검체 수 산정에 필요한 파라미터와 그 값을 추출하였다. 검체 수 산정 프로그램은 PHP 웹 스크립트 언어와 MySQL을 이용하였다. 프로그램에 사용된 통계모델은 검정력을 고려하지 않은 단일 집단에서의 검체 수 산정(모델 1), 검정력을 고려한 단일 집단에서의 검체 수 산정(모델 2), 검정력을 고려한 두 집단에서의 검체 수 산정(모델 3) 등 세 가지이었다.

**결과 :** 1989년에서 2005년 사이에 보고된 문헌들 중에서 Medical Subject Headings (MeSH) 용어를 통해 총 13개의 항-HCV 성능평가 연구보고를 수집하였다. 문헌상의 진단민감도는 83-100%로 평가에 이용된 대상 환자 수는 중앙값이 145명(12-1,091)이었다. 진단특이도는 97-100%였으며 대상 정상인 수는 중앙값이 1,025명(33-4,381명)이었다. HCV 유병률 2%, 기존 검사법의 진단민감도 95%, 새로 개발된 체외진단키트의 진단민감도 97%, 허용오차가 2%, 양측검정, 유의수준 0.05, 검정력 80%, 동등성 비교 등의 파라미터 값에서의 적절한 검체 수는 모델 1에서 280명, 모델 2에서 817명, 모델 3에서 1,510명이었다.

**결론 :** 본 연구를 통해 부적절한 검체 수를 이용한 성능평가는 여전한 문제임을 확인할 수 있었다. 저자들이 개발한 웹기반 프로그램은 향후 성능평가 연구의 신뢰성을 높이는 데 활용될 수 있을 것이다.

## 참고문헌

1. How to read clinical journals: II. To learn about a diagnostic test. Can Med Assoc J 1981;124:703-10.
2. Deeks JJ and Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004;329:168-9.
3. Sim J and Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005;85: 257-68.
4. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001;323:157-62.
5. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645-51.
6. Lumbreras-Lacarra B, Ramos-Rincon JM, Hernandez-Aguado I. Methodology in diagnostic laboratory test research in clinical chemistry and clinical chemistry and laboratory medicine. Clin Chem 2004; 50:530-6.
7. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med 1978;299:690-4.
8. Moore AD and Joseph L. Sample size considerations for superiority trials in systemic lupus erythematosus (SLE). Lupus 1999;8:612-9.
9. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998;17:857-72.
10. Arkin CF and Wachtel MS. How many patients are necessary to assess test performance? JAMA 1990;263:275-8.
11. Wang JT, Wang TH, Sheu JC, Tsai SJ, Hsieh YS, Lin DT, et al. Hepatitis C virus infection in volunteer blood donors in Taiwan. Evaluation by hepatitis C antibody assays and the polymerase chain reaction. Arch Pathol Lab Med 1993;117:152-6.
12. Colin C, Lanoir D, Touzet S, Meyaud-Kraemer L, Bailly F, Trepo C; HEPATITIS Group. Sensitivity and specificity of third-generation hepatitis C virus antibody detection assays: an analysis of the literature. J Viral Hepat 2001;8:87-95.
13. Dati F, Denoyel G, van Helden J. European performance evaluations of the ADVIA Centaur infectious disease assays: requirements for performance evaluation according to the European directive on in vitro diagnostics. J Clin Virol 2004;30:S6-10.
14. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51:1335-41.
15. Sackett DL and Haynes RB. The architecture of diagnostic research. BMJ 2002;324:539-41.
16. Gluud C and Gluud LL. Evidence based diagnostics. BMJ 2005;330:

724-6.
17. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003; 138:40-4.
18. Smith BJ, Darzins PJ, Quinn M, Heller RF. Modern methods of searching the medical literature. Med J Aust 1992;157:603-11.
19. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000;53:65-9.
20. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:1-12.
21. Stuyver L, Claeys H, Wyseur A, Van Arnhem W, De Beenhouwer H, Uytendaele S, et al. Hepatitis C virus in a hemodialysis unit: molecular evidence for nosocomial transmission. Kidney Int 1996;49:889-95.
22. Lavanchy D, Steinmann J, Moritz A, Frei PC. Evaluation of a new automated third-generation anti-HCV enzyme immunoassay. J Clin Lab Anal 1996;10:269-76.
23. Courouce AM, Bouchardeau F, Girault A, Le Marrec N. Significance of NS3 and NS5 antigens in screening for HCV antibody. Lancet 1994; 343:853-4.
24. Hennig H, Schlenke P, Kirchner H, Bauer I, Schulte-Kellinghaus B, Bludau H. Evaluation of newly developed microparticle enzyme immunoassays for the detection of HCV antibodies. J Virol Methods 2000;84:181-90.
25. Jonas G, Pelzer C, Beckert C, Hausmann M, Kapprell HP. Performance characteristics of the ARCHITECT anti-HCV assay. J Clin Virol 2005;34:97-103.
26. Abdel-Hamid M, El-Daly M, El-Kafrawy S, Mikhail N, Strickland GT, Fix AD. Comparison of second- and third-generation enzyme immunoassays for detecting antibodies to hepatitis C virus. J Clin Microbiol 2002;40:1656-9.
27. Zachary P, Ullmann M, Djeddi S, Wendling MJ, Schvoerer E, Stoll-Keller F, et al. Evaluation of two commercial enzyme immunoassays for diagnosis of hepatitis C in the conditions of a virology laboratory. Pathol Biol 2004;52:511-6.
28. Judd A, Parry J, Hickman M, McDonald T, Jordan L, Lewis K, et al. Evaluation of a modified commercial assay in detecting antibody to hepatitis C virus in oral fluids and dried blood spots. J Med Virol 2003;71:49-55.
29. Ismail N, Fish GE, Smith MB. Laboratory evaluation of a fully automated chemiluminescence immunoassay for rapid detection of HBsAg, antibodies to HBsAg, and antibodies to hepatitis C virus. J Clin Microbiol 2004;42:610-7.
30. Huber KR, Sebesta C, Bauer K. Detection of common hepatitis C virus subtypes with a third-generation enzyme immunoassay. Hepatology 1996;24:471-3.
31. Prince AM, Scheffel JW, Moore B. A search for hepatitis C virus polymerase chain reaction-positive but seronegative subjects among blood donors with elevated alanine aminotransferase. Transfusion 1997; 37:211-4.
32. Vrielink H, Zaaijer HL, Reesink HW, van der Poel CL, Cuypers HT, Lelie PN. Sensitivity and specificity of three third-generation anti-hepatitis C virus ELISAs. Vox Sang 1995;69:14-7.
33. Busch MP, Watanabe KK, Smith JW, Hermansen SW, Thomson RA. False-negative testing errors in routine viral marker screening of blood donors. For the Retrovirus Epidemiology Donor Study. Transfusion 2000;40:585-9.