ANKARA ÜNİVERSİTESİ TIP FAKÜLTESİ MECMUASI Cilt 56, Sayı 4, 2003 259-264

TANI TESTİ ÇALIŞMALARINDA METODOLOJİK STANDARTLARIN KULLANILMASI

Yasemin Genç*

ÖZET

Tanı testlerinin modern tıp alanında önemli bir yeri vardır. Teknolojik gelişmelere paralel olarak her geçen gün eskisinden daha iyi olduğu iddia edilen yeni testler önerilmektedir. Fakat birçok test genel kullanıma geçilmeden önce tüm detaylar göz önünde bulundurularak değerlendirilmemektedir. Testlerin ayırıcılık gücünü değerlendiren çalışmalarda yapılan metodolojik hatalar test performansının güvenilir bir şekilde elde edilememesine neden olmaktadır. Bu durum tanı testlerini değerlendiren çalışmalara metodolojik standartlar getirilmesi gereğini ortaya koymaktadır. Çalışmamızda bu standartlar tek tek ele alınmış ve standartların sağlanması için gerekenler sıralanmıştır.

Anahtar Kelimeler: *Altın Standart Testing Yokluğu, Hasta Spektrumunun Farklılığı, Seçilmiş Deneklerin Kullanılması, Tanıda Kararsızlık ya da Yetersiz Bulgu Sonuçları, Tekrarlanabilirlik, Yorumdan Kaynaklanan Yan*

SUMMARY

Use of Methodological Standards In Diagnostic Test Research

Diagnostic testing is an important component of modern medical care. Parallel to the technologic advances, new diagnostic tests that are supposed to be better are advised recently. Unfortunately, many diagnostic tests are not rigorously evaluated before general application. Studies examining diagnostic measures often have methodological flaws that impair their ability to provide reliable information on test performance. This situation implies that there is a need for fitting methodological standards to the researches evaluating diagnostic tests. In our study, we mentioned these standards one by one and explained needs for fitting them.

Key Words: *Imperfect Gold Standard, Spectrum and Subgroup effect, Verification Bias, Indeterminate and Uninterpretable results, Reproducibility, Interpretation-Review bias*

Tanı testleri, hasta ve sağlıklı bireylerin oluşturduğu heterojen bir kitlede bireylerin gerçek durumunu (gerçekten hasta olup olmadıklarını) ortaya çıkarmak amacıyla kullanılır. Doğruluğu kesin olarak kanıtlanmış referans testler (kesin test/altın standart test) ile bireylere "kesin hasta" ya da "kesin sağlıklı" tanısı koyulabilir. Fakat bu testlerin uygulanmalarının zor, maliyetlerinin yüksek ve bazı hastalıklarda girişimsel olmaları nedeniyle her şüpheli durumda kullanılmaları mümkün değildir. Bu nedenle bir çok bilim dalında referans testlere alternatif olacak tanı testleri geliştirilmeye çalışılır. Referans testlerle gerçek durumu (hasta-sağlıklı) belirlenmiş bireylere ilgilenilen tanı testi uygulanarak, testin ayırıcılık gücünü gösteren "doğruluk ölçütleri " elde edilir. İki sonuçlu (binary) tanı testlerinin ayırıcılık gücü "duyarlılık", "seçicilik", "pozitif tahmini değer" ve "negatif tahmini değer" gibi ölçütlerle, sıralı ya da sürekli sonuçlu tanı testlerinin ayırıcılık gücü ise "İşlem Karakteristiği Eğrisi(İKE) altında kalan alan" yardımıyla değerlendirilir.

Tanı testlerinin modern tıp alanındaki önemi gün geçtikçe artmaktadır. Teknolojinin gelişmesine paralel olarak her geçen gün hastalıkların ta-

* Ankara Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı

Geliş Tarihi: 19 Haziran 2003 Kabul Tarihi: 08 Eylül 2003

ranmasında ve tanı koymada daha iyi olduğu iddia edilen yeni yöntemler önerilmektedir. Bu durum tanı testlerinin tanı koyma gücü ve maliyetleri açısından karşılaştırılmalarını gerektirir. Fakat tanı testlerinin önemli işlevleri olmasına rağmen, testlerin ayırıcılık gücünü değerlendiren çalışmalara gereken önemin verilmediği ve birçok tanı testinin rutin kullanıma geçilmeden önce tüm detaylar göz önünde bulundurularak değerlendirilmediği yapılan çalışmalarla gösterilmiştir(1-3). Testlerin ayırıcılık gücünü belirlemek amacıyla planlanan çalışmaların deney düzeninde yapılan hatalar doğruluk ölçütlerinin yanlı kestirilmesine sebep olmaktadır. Bu durum tanı testlerini değerlendiren çalışmalara metodolojik standartlar getirilmesi gereğini ortaya koymuştur(4-6). Çalışmamızda bu standartlar tek tek ele alınmış ve standartların sağlanması için yapılması gerekenler sıralanmıştır.

1. Seçilmiş Denekler Üzerinde Çalışmadan Kaçınma (Verification Bias)

Doğruluk ölçütlerinin yanlı kestirimini etkileyen en önemli faktör bu ölçütlerin sadece belirli kriterlere göre seçilmiş denekler kullanılarak hesaplanmasıdır. Doğruluk ölçütlerinin yansız kestirilebilmesi için tanı testi ve referans test sonuçlarının birbirinden bağımsız olarak elde edilmesi gerekir. Fakat pratikte hastalara referans test uygulanıp uygulanmaması kararı tanı testi sonucuna ve bazı klinik parametrelere bağlıdır. Örneğin referans test girişimsel bir yöntem olduğunda, tanı testi sonucuna göre hastalık şüphesi olanlara referans test uygulama olasılığı, hastalık şüphesi olmayanlara göre daha yüksektir. Bu durumda doğruluk ölçütlerinin elde edildiği çalışma popülasyonu belli kriterlere göre seçilmiş bireylerden oluşur. Bu yaklaşım, klinik olarak ve maliyet-yarar hesabına göre doğru olsa da ölçütlerin yanlı kestirimine sebep olur. Bu tip bir yan "seçilmiş denekler üzerinde çalışmadan kaynaklanan yan" olarak adlandırılır.

Tanı testi çalışmalarında bu standardın sağlanması için tanı testi uygulanan bireylerin tümüne referans test uygulanması gerekir. Fakat referans testin girişimsel ya da pahalı olduğu durumlarda, doğruluk ölçütlerinin yansız kestirimlerini elde edebilmek için hastalık şüphesi düşük olan bireylere referans testi uygulamak etik ve pratik bir yol olmayabilir. Bu durumda ise yansız kestirimler elde etmek için geriye dönük düzeltme yöntemleri kullanılmalıdır. "Duyarlılık" ve "seçicilik" ölçütlerinin yansız kestirimi için Begg ve Greenes yöntemi(7), "İKE altında kalan alan" için ise Zhou (8) yöntemi en yaygın kullanılan düzeltme yöntemleridir.

2. Referans Testin Yokluğu (Imperfect Gold Standard)

Yeni geliştirilen bir tanı testinin performansı, bu test ile referans testin sonuçlarının karşılaştırılmasıyla belirlenir. Fakat pratikte tüm hastalıklar için referans test elde etmek mümkün değildir. Bu nedenle tanı testinin doğruluk ölçütleri, göreli olarak diğerlerine göre en doğru sonuçları veren fakat sonuçları kesin doğru olmayan bir teste göre hesaplanır. Bu tür testler "kesin olmayan referans test"ler (imperfect gold standard) olarak adlandırılır.

Tanı testlerinin değerlendirilmesinde kesin olmayan referans testlerin kullanılmasının birkaç nedeni vardır. Bunlardan ilki, anjina pektoris ve migren gibi bazı hastalıkların referans testinin olmamasıdır. Bir diğer neden ise, referans test olduğu halde uygulanmasının teknik olarak uygun olmaması ya da uygulandığında hastayı büyük risk altında bırakmasıdır. Örneğin, Alzheimer hastalığının kesin tanısı ancak hasta öldükten sonra beyin otopsisinin incelenmesiyle koyulabilir. Ya da, pulmoner wedge basıncının ölçülmesi konjestif kalp yetmezliğinin ayırıcı tanısında kesin tanı yöntemi olmakla beraber girişimsel olması nedeniyle hasta için çok risklidir. Kesin olmayan referans testlerin kullanılmasının bir diğer nedeni ise, bazı durumlarda tedaviye başlamak için referans testin sonucunun beklenmesinin hastanın hayatının riske atılmasına sebep olmasıdır. Örneğin hastayı şoka sokacak derecede arter yaralanmasının kesin tanısını koymak için anjiyografi yapmayı beklemek hastanın hayatını riske etmektir. Böyle bir durumda referans testin sonucu beklenmeden kesin olmayan referans testlerle nihai karar verilip tedaviye başlanır.

Tanı testi çalışmalarında bu standarda uymak için referans test olmadığı durumda, yeni tanı testinin "duyarlılık", "seçicilik", "pozitif tahmini değer" ve "negatif tahmini değer" ölçütlerini elde ederken Discrepant Resolution (DR)(9), Composi-

te Referans Standart (CRS)(10) ve Latent Class Model(11) gibi yöntemlerden yararlanılmalıdır. DR yönteminde altın standart olmayan referans test ile yeni tanı testinin uyuşmadığı durumlara bir çözücü test uygulanarak referans testin kaçırdığı gerçek pozitif durumlar yakalanmaya çalışılır. CRS yönteminde ise çeşitli referans testlerin sonuçları kombine olarak kullanılır.

3. Yorumdan Kaynaklanan Yandan Kaçınma (Interpretation-Review bias)

Subjektif kriterlere dayanan tanı testlerinin doğruluk ölçütleri daha önceki test sonuçları ve diğer klinik bulguların gözlemcilerin kararlarını etkilemesi sonucu yanlı kestirilebilir. Bu durum genellikle testin performansının olduğundan daha yüksek bulunmasına sebep olur.

Yorumdan kaynaklanan yan değişik biçimlerde ortaya çıkabilir.

Geriye dönük (retrospective) çalışmalarda, tanı testi genellikle referans test sonucuna göre hasta ya da sağlıklı olarak gruplandırılmış kişilere uygulanır. Gözlemcinin referans test sonuçlarını önceden bilmesi tanı testi sonuçlarını etkiler. Bu da doğruluk ölçütlerinin yanlı (test review bias) kestirimine sebep olur. Benzer bir yan kaynağı da referans testin yokluğunda (imperfect gold standard) ortaya çıkar. Bu durumda referans test subjektif kriterler yardımıyla yorumlandığından tanı testinin sonuçlarının bilinmesi referans testin sonuçlarını etkiler (diagnostic review bias) (6).

Tanı testi çalışmalarında bu standardın sağlanması için araştırmacıların tanı testini değerlendirirken referans test sonuçlarını bilmemeleri gerekir.

Hastanın demografik özelliklerinin (yaş, cinsiyet gibi) veya klinik bulgularının (hastalığın şiddeti, hastanın şikayetleri, gibi) referans testin değerlendirilmesinden daha önce biliniyor olması da tanı testinin performansının gerçekte olduğundan daha yüksek bulunmasına sebep olur(12,13). Fakat alan çalışmalarında bu bilgilerin elde edilmesini önlemek mümkün değildir. Bu bilgiler olmadan testin ham performansını elde etmek için özel deneysel çalışma planlamak gerekir. Araştırmacılar, alan çalışması yardımıyla elde edilen doğruluk ölçütlerinin sadece testin ayırıcılık gücünü değil aynı zamanda klinik bilgilerin de ayırıcılık gücünü yansıttığını bilmelidirler.

4. Hasta Spektrumunun Farklılığı (Spectrum and Subgroup effect)

Tanı testlerinin doğruluk ölçütleri çalışılan hasta popülasyonunun demografik özelliklerine (yaş ve cinsiyet gibi) ve hastalığın şiddeti, süresi gibi klinik parametrelere bağlı olarak değişebilir. İlk olarak Hlatky ve arkadaşları egzersiz elektrokardiyografi testinin sonuçlarının atipik ve tipik kroner kalp hastalarında farklı ayırıcılık gücüne sahip olduğunu ortaya koyarak bu yan kaynağına dikkat çekmişlerdir(14). Yine kanser taramalarında test duyarlılıklarının büyük tümörlerde küçük tümörlere göre daha yüksek bulunması bu yan kaynağına bir örnek olabilir.

Bir tanı testinin performansı tüm grupta düşük olduğu halde hastalığın bir alt grubunda çok yüksek bulunabilir. Bu nedenle hasta spektrumu göz önünde bulundurulmadan elde edilen doğruluk ölçütleri tüm hasta grubu için ancak "ortalama bir değer" verebilir.

Tanı testi çalışmalarında bu standardın sağlanması için doğruluk ölçütleri birlikte değişenlerin (covariates) tüm kombinasyonlarında ayrı ayrı hesaplanmalı yada alt grupların genişlikleri göz önünde bulundurularak ortalamaları hesaplanmalıdır (15). Birlikte değişenlerin sayısı fazla olduğunda her bir kombinasyona düşen denek sayısı az olabilir. Bu durumda duyarlılık, seçicilik ve tahmini değerlerin kestirimi lojistik regresyon yardımıyla yapılabilir(16,17). Test sonucu sıralı ölçekli olan tanı testleri için ise Tosteson ve Begg tarafından önerilen "sıralı regresyon modeli" kullanılarak birlikte değişenlerin etkisi giderilebilir ve tanı testinin ayırıcılık gücünü gösteren İKE altında kalan alanın yansız kestirimi elde edilebilir(18).

5. Tanıda Kararsızlık ya da Yetersiz Bulgu Sonuçları (Indeterminate and Uninterpretable results)

Klinik testler her zaman yorumlanabilir, net sonuçlar vermezler. Bazı test sonuçları tanıda kararsızlığa (equivocal) neden olurken bazıları da yetersiz bulgu (nondiagnostic) nedeniyle yorumlanamaz. Örneğin kontrol mamografilerinde mikrokalsifikasyonların saptanması şüpheli (tanıda kararsız kalınan) sonuçlara bir örnektir. İncelenen hastada barsak gazı olmasının abdominal ultrason görüntülerini engellemesi ya da ince-iğne aspirasyonu ile alınan hücre miktarının tanı koymak için yeter-

li olmaması durumu da yetersiz bulgu sonucunda tanı testinin yorumlanamamasına örnek olarak verilebilir.

Tanı testlerini değerlendirirken bu gibi sonuçların göz ardı edilmesi ya da hasta ya da sağlıklı gruplarından birine dahil edilmesi doğruluk ölçütlerinin yanlı kestirilmesine sebep olur. Tanı testi iki sonuçlu (binary) olduğunda, şüpheli ya da yetersiz bulgu nedeniyle yorumlanamayan sonuçların hasta grubuna dahil edilmesi duyarlılığın gerçekte olduğundan daha yüksek, seçiciliğin ise gerçekte olduğundan daha düşük kestirilmesine neden olur. Bu sonuçların sağlıklı grubuna dahil edilmesi ise tersi bir duruma yol açar.

Tanı testi sonucu sıralı ölçekli olduğunda şüpheli sonuçlar bir sıra numarası ile temsil edilerek yan kaynağı olmaktan çıkarılabilir. Fakat yetersiz bulgu nedeniyle yorumlanamayan sonuçları sıralı ölçekli testlere dahil etmek mümkün değildir.

Tanı testi çalışmalarında bu standardın sağlanması için bu gibi test sonuçlarının oranlarının ayrı ayrı verilmesi gerekir. Ayrıca doğruluk ölçütleri hesaplanırken bu verilerin kullanılıp kullanılmadığı da mutlaka belirtilmelidir. Bu bilgiler, testin klinik yararlılığının ve uygulama maliyetinin belirlenmesi yoluyla optimal tanı stratejisi geliştirilmesinde de kullanılabilir.

6. Testin Tekrarlanabilirlik Düzeyi (Reproducibility)

Subjektif kriterlere göre değerlendirilen testlerin performansı testi değerlendiren gözlemciye (örneğin radyolog), kullanılan ekipmana ve uygulanan laboratuar prosedürüne göre değişkenlik gösterir. Örneğin bir tanı testinin ayırıcılık gücü, uzman bir kişinin özel ekipman kullanmasıyla çok yüksek bulunurken uzman olmayan bir kişinin standart ekipman kullanmasıyla zayıf bulunabilir. Ayrıca aynı gözlemcinin, aynı ekipmanın ya da aynı laboratuarın ölçümleri tekrarlandığında da farklı sonuçlar elde edilebilir. Fakat bu etkiler, bir testin doğruluk ölçütlerinin değişik çalışmalarda farklı bulunmasına sebep olur. Özellikle radyoloji alanında görüntülerden elde edilen tanı sonuçlarında gözlemcilerden kaynaklanan değişimin varlığına dikkat çekilmektedir.

Tekrarlanabilirlik tanı testlerinin değerlendirilmesinde önemli bir özelliktir. Özellikle sonuçları subjektif olan tanı testlerini değerlendirme çalışmalarında tekrarlanabilirliğin düzeyi mutlaka belirtilmeli ya da doğuluk ölçütleri uygun istatistiksel yöntemlerle düzeltilerek verilmelidir.

Bu amaçla kullanılan istatistiksel yöntemler verilerin iki sonuçlu olup olmadığına ve gözlemci sayısına bağlı olarak değişmektedir.

İki sonuçlu tanı testleri için iki gözlemcinin uyumu Kappa Katsayısı, Ağırlıklandırılmış Kappa Katsayısı, Sınıf-içi Korelasyon Katsayısı ve Tetrachoric Korelasyon Katsayısı kullanılarak hesaplanabilir. İki sonuçlu testler için ikiden çok gözlemcinin uyumu Sınıf-içi Korelasyon Katsayısı, Genelleştirilmiş Kappa Katsayısı, Latent Trait Model ve Latent Class Modeller kullanılarak hesaplanabilir.

Sıralı ölçekli tanı testleri için iki gözlemcinin uyumu Polychoric Korelasyon Katsayısı ile incelenirken ikiden çok gözlemcinin uyumu Latent Trait Modeller kullanılarak araştırılır(19).

7. Doğruluk Ölçütleri İçin Güven Sınırları

Herhangi bir örneklemden elde edilen "duyarlılık", "seçicilik" ve "İKE altında kalan alan" gibi ölçütleri evrene genellemek için bunlara ait aralık tahminlerini yapmak gerekir. Tanı testi çalışmalarında doğruluk ölçütlerine ait güven sınırlarının genellikle verilmediği gözlenmektedir.

"Duyarlılık" ve "Seçicilik" için güven sınırları Wilson Score yöntemiyle aşağıdaki gibi hesaplanabilir(20,21).

Tanı Testi	Referans Test	
	Hasta	Sağlıklı
Pozitif	a	b
Negatif	c	d

$$Duyarlılık\ için\ Güven\ Sınırları = \frac{2(a+c)Duy + Z^2_{\alpha/2} \mp Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} + 4(a+c)Duy(1-Duy)}}{(2((a+c)+Z^2_{\alpha/2}))}$$

$$Seçicilik\ için\ Güven\ Sınırları = \frac{2(b+d)Seç + Z^2_{\alpha/2} \mp Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} + 4(b+d)Seç(1-Seç)}}{(2((b+d)+Z^2_{\alpha/2}))}$$

"İşlem Karakteristiği eğrisi altında kalan alan" için güven sınırı hesaplamaları alan hesaplaması için kullanılan yönteme göre değişmektedir. İstatistik paket programlarında (Örneğin SPSS) alan hesaplamasının yanında güven sınırlarını hesaplatmak da mümkündür.

KAYNAKLAR

1. Reid MC, Lachs MS, Feinstein AR: Use of methodological standards in diagnostic test research. JAMA 1995, 274:645-651
2. Sheps SB, Schecter MT: The assessment of diagnostic tests: A survey of current medical research. JAMA 1984,252:2418-2422
3. Jaeschke R, Guyatt GH, Sackett DL: User's Guides to the medical literature III.How to use an article about a diagnostic test B. What are the results and will they help me in caring for my patients? JAMA 1994, 271:703-707
4. Greenhalgh T: How to read a paper. Papers that report diagnostic or screening tests.BMJ 1997, 315: 540-543
5. Mower WR: Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999, 33:85-91
6. Begg CB, Greenes RA: Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983, 39:207-215
7. Zhou, X.H.: A nonparametric ML estimator for the ROC curve area in the presence of verification bias. Biometrics 1996, 52:299-305
8. Miller CW: Bias in Discrepant Analysis: Whwn two wrongs don't make a right. Journal of Clinical Epidemiology 1998, 51:219-231
9. Hadgu A: The discrepancy in discrepant analysis, Lancet, 1996, 348:592-593
10. Alonza TA, Pepe MS: Using a combination of referance tests to assess the accuracy of a new diagnostic test, Statistics in Medicine 1999,18:2987-3003
11. Doubilet DE, Herman PG: Interpretation of radiographs effect of clinical history. Am J Roentgenol 1981, 137:1055-1058
12. Barbaum KS, Franken EA, Dorfman DD: Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs. Invest Radiol 1986:21,532-539
13. Hlatky MA, Pryor DB, Harrell FE ve ark: Factors affecting sensitivity and specificity of exercise electrocardiograph.Multivariate analysis. Am J Med 1984, 77:64-71
14. Lachs MS, Nachamkin I, Edelstein PH ve ark: Spectrum bias in the evaluation of diagnostic tests: Lessons from the rapid dipstick test for ürinary tract infection. Annals of Internal Medicine 1992,117:135-140
15. Mulherin SA, Miller WC: Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002,137:598-602
16. Coughlin SS, Trock B, Criqui MH ve ark: The lojistic modeling of sensitivity, specificity and predictive value of a diagnostic tests. J Clin Epidemiol 1992, 45:1-7
17. Tosteson ANA, Begg CB: A general regression methodology for ROC curve estimation, Med Desic Making 1988,8:204-215
18. Uebersax JS: A review of modeling approches for the analysis of observer agreement.Invest Radiol 1992,27:738-743
19. Newcombe, Robert G: "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods, Statistics in Medicine 1998, 17: 857-872
20. Wilson, E. B: "Probable Inference, the Law of Succession, and Statistical Inference, Journal of the American Statistical Association 1927, 22: 209-212