Ankara Üniversitesi Tıp Fakültesi Mecmuası 2012, 65 (3)
DOI: 10.1501/Tıpfak_000000825

TEMEL TIP BİLİMLERİ/ *BASIC SCIENCES*
Araştırma Makalesi / Research Article

Tanı Doğruluğu Çalışmalarının Kalitelerinin Değerlendirilmesi: STARD Kriterlerinin Türkçe Uyarlaması

Evaluation Of Quality Of Diagnostic Accuracy Studies: Turkish Adaptation Of STARD Criteria

Yasemin Genç[1], Rabia Albayrak[2], Can Ateş[1], Mustafa Agah Tekindal[3], Pınar Selvi[4], Sibel Perçinel[5], Koray Ceyhan[6]

[1] *Ankara Üniversitesi Tıp Fakültesi Biyoistatistik AD*
[2] *Ankara Üniversitesi Ziraat Fakültesi Zootekni Bölümü Biyometri ve Genetik AD*
[3] *Başkent Üniversitesi Tıp Fakültesi Biyoistatistik AD*
[4] *Çukurova Üniversitesi Tıp Fakültesi Biyoistatistik AD.*
[5] *Ankara Üniversitesi Tıp Fakültesi Tıbbi Patoloji AD.*
[6] *Ankara Üniversitesi Tıp Fakültesi Tıbbi Patoloji AD. Sitopatoloji BD.*

Amaç: Bu çalışmada, tanı doğruluğu çalışmalarının planlanması ve sonuçlarının raporlanması konusunda bir standardın oluşturulması amacıyla çoğunluğunu epidemiyolog ve biyoistatistikçilerin oluşturduğu bir grup tarafından 2003 yılında yayınlanan STARD (The **STA**ndards for **R**eporting of **D**iagnostic Accuracy) kriterlerinin Türkçeye çevrilmesi ve ülkemizde yapılan çalışmalarda kullanımının yaygınlaştırılması amaçlanmıştır.

Gereç ve Yöntem: 2003 yılında yayınlanan orijinal STARD kriterleri, ilk 4 yazar (RA, MAT, CA, PS) tarafından birbirlerinden bağımsız olarak Türkçeye çevrilmiş ve YG'nin liderliğinde görüş birliğine varılarak son haline getirilmiştir. Dil eşdeğerliliğinin sağlanması amacıyla, Türkçe uyarlaması yapılan STARD kriterleri profesyonel bir tercüman tarafından kontrol edilmiş ve gerekli düzeltmeler yapılmıştır. Son hali verilen Türkçe STARD kriterleri, geri çevirme yöntemi kullanılarak bir başka profesyonel tercüman tarafından Türkçeden İngilizceye çevrilmiş ve orijinal metin ile üst düzeyde bir benzerlik olduğu saptanmıştır. Daha sonra Türkçeye uyarlanmış kontrol listesinin araştırmacılar tarafından doğru bir şekilde anlaşılıp anlaşılmadığını ve tekrar edilebilirlik düzeyini değerlendirmek amacı ile aynı uzmanlık düzeyine sahip iki pataloğa (SP, KC) 20 tanı doğruluğu çalışması verilmiş ve Türkçe uyarlaması yapılmış olan STARD kriterlerini kullanarak, yayınların kalitelerini değerlendirmeleri istenmiştir. Gözlemciler arasındaki uyum her bir kriter için gözlemciler arası uyum yüzdesi ve Cohen'in Kappa istatistiği kullanılarak incelenmiştir. Uyum ayrıca toplam raporlanan kriter sayısı kullanılarak Bland-Altman yöntemi ve Sınıf İçi Korelasyon Katsayısı (SKK) kullanılarak da değerlendirilmiştir.

Bulgular: İncelenen tanı doğruluğu çalışmalarında kriterler rapor edilme sıklıkları bakımından büyük bir varyasyon göstermektedir. Kriterlerin bir bölümü bütün makalelerde rapor edilirken bir bölümü ise neredeyse hiç rapor edilmemiştir. Makale başına toplam raporlanan kriter sayısı 15.9±3.2 (Gözlemci I'in sonucu) olup 25 kriterin en az 10'unun, en fazla 21'inin raporlanmış olduğu gözlenmiştir. Kriterler için ayrı ayrı hesaplanan gözlemciler arası uyum yüzdelerinin bir kriter için %60 diğerleri için %75 ve üzerinde olduğu gözlenmiştir. İki kriter için Kappa istatistiği elde edilememiş, 3 kriter için önemsiz düzeyde bulunmuş, 20 kriter için ise 0.32'nin üzerinde bulunmuştur. %95 Uyum Limitleri -2.2-3.5 olarak bulunmuş olup bu sonuç iki gözlemci arasındaki uyumun kabul edilebilir düzeyde olduğunu göstermiş, bunun yanında SKK da 0.945 gibi oldukça yüksek bir değer olarak bulunmuştur.

Sonuç: Türkçe uyarlaması yapılmış STARD kriterlerinin ülkemizde tanı doğruluğu çalışması yapacak olan araştırmacıların yanında dergi hakem ve editörlerin için de bir kılavuz olması umut edilmektedir. Çalışmalarda metodolojik standartlara uygun ortak bir ölçüt kullanılması, bireysel çalışmaların yanında meta analizi çalışmalarının da kalitelerinin artmasını sağlayacaktır.

Anahtar Sözcükler: ***STARD kriterleri, Tanı doğruluğu çalışmaları***

Aim: The aim of this study is translate the STARD (The STAndards for Reporting of Diagnostic Accuracy) criteria which were developed in 2003 by a group of epidemiologists and biostatisticians in order to establish a standard for planning and reporting the results of the diagnostic accuracy studies.

Materials and Design: The STARD criteria that were published in 2003 have firstly been translated into Turkish language by the first four authors (RA, MAT, CA, and PS) independently and afterwards four translated versions have been combined with consensus of the whole group headed by YG. For assuring the language equivalency between the original and the translated STARD statements, the translated criteria was also checked by a professional interpreter and necessary corrections were made. The final version of the translated criteria is translated into English by another professional interpreter using reverse-translation method and found to be quite similar with the original. For checking the understandability and repeatability of the translated version of the STARD criteria, 2 pathologists were asked to use the translated STARD criteria and evaluate the quality of 20 diagnostic accuracy studies. The inter-rater agreement for each criterion was investigated by using consistency percentages and Cohen's kappa statistic. Moreover, the agreement is also assessed by using Bland-Altman method and Within Group Correlation by using the total number of reported criteria.

Results: The investigated 20 diagnostic accuracy studies were found to be variable with respect to reported criteria percentages. Some of the criteria were found to be reported in all studies whereas some other criteria were reported in none of the studies. Total number of reported criteria per article is found to be 15.9±3.2 (Rater 1) with a minimum of 10 and a maximum of 21. Agreement percentages per criteria were found to be 60% for one of the criterion, and 75% or above for the others. Kappa statistic could not be estimated for two items, was found to be not significant for 3 items and was above 0.32 for 20 items. The limits for agreement were found as -2.2-3.5, and it shows that the inter-rater agreement has an acceptable level along with SKK is calculated to be 0.945.

Conclusion: It is expected that the translated STARD criteria will be a guideline for the researchers in the diagnostics accuracy studies as well as for the journal reviewers and editors. The use of Turkish translated version of STARD as a standard measure will increase the quality of diagnostic accuracy studies as well as the meta analysis studies.

Key Words: ***STARD statements, Diagnostic accuracy studies***

İndeks testler (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 standartlar (altın standart test) ile bireylere "kesin hasta" ya da "kesin sağlıklı" tanısı koyulabilir. Fakat bu testlerin uygulanmalarının zor,

Geliş Tarihi: 11.04.2012 • Kabul Tarihi: 10.02.2014

İletişim

Dr. Can Ateş
Tel : 0 312 595 81 39
E-posta : *can.ates@gmail.com*
Ankara Üniversitesi Tıp Fakültesi Biyoistatistik AD.Morfoloji Kampüsü Sıhhiye, Ankara

maliyetlerinin yüksek ve bazı hastalıklarda girişimsel olmaları nedeniyle her şüpheli durumda kullanılmaları mümkün olmayabilir. Bu nedenle birçok bilim dalında referans standartlara alternatif olacak indeks testler geliştirilmeye çalışılır. Belirli kriterlere göre seçilmiş bir grup kişiye referans standart ve indeks test uygulanarak indeks testin ayırıcılık gücünü gösteren "doğruluk ölçütleri " elde edilir. Yeni geliştirilen indeks testlerin ayırıcılık gücünü belirlemek ya da testlerin ayırıcılık gücünü karşılaştırmak amacıyla yapılan çalışmalara "tanı doğruluğu" çalışmaları adı verilir (1).

Tanı doğruluğu çalışmalarının modern tıp alanındaki önemi gün geçtikçe artmaktadır. Teknolojinin gelişmesine paralel olarak her geçen gün hastalıkların taranmasında ve tanı koymada daha iyi olduğu iddia edilen yeni yöntemler önerilmektedir. Bu durum yeni geliştirilen testlerinin tanı koyma güçlerinin tahmin edilmesi ve aynı amaçla kullanılan diğer testlerle karşılaştırılmasını gerektirir. Fakat yapılan çalışmalar, indeks testlerin ayırıcılık gücünü değerlendiren çalışmaların planlanması, yürütülmesi ve sonuçlarının raporlanmasına gereken önemin verilmediğini ortaya koymaktadır (1, 2). Reid ve arkadaşları (3) 1995 yılında yaptıkları bir çalışmada dört önemli tıbbi dergide (New England Journal of Medicine, Journal of the American Medical Association, British Medical Journal, ve Lancet) yayınlanan 112 tanı testi çalışmasını yedi temel standarda uygunlukları bakımından değerlendirmiş ve çalışmaların hiçbirinin standartların tümünü sağlamadığını gözlemişlerdir (3). Lijmer ve ark. (4) 1999 da yaptıkları bir çalışmada tanı doğruluğu çalışmalarının içermesi gereken en kritik bilgileri dahi içermediğini belirlemişlerdir (4).

Bunun yanında sistematik derleme ve meta analizi çalışmalarının artması, tanı doğruluğu çalışmalarının yetersiz bir şekilde raporlandığını daha çarpıcı bir şekilde ortaya koymuştur. 1996 yılında tanı doğruluğu çalışmaları alanında yapılan sistematik derlemeleri konu alan bir toplantıda, yayınlanmış tanı doğruluğu çalışmalarının birçoğunda değerlendirilen teste ilişkin duyarlılık, seçicilik ve işlem karakteristiği eğrisi (İKE) altında kalan alan gibi temel ölçütlerin bile yer almadığı belirtilmiş, bu durumun sistematik derleme çalışmalarını imkansız hale getirdiği ortaya konmuştur. Ayrıca Nelemans ve ark. (5) De Vries ve ark. (6) tarafından yapılan meta analizi çalışmalarında da tanı doğruluğu çalışmalarının nasıl planlandığı ve gerçekleştirildiği konusunda çok kritik bilgilerin dahi yer almadığı bildirilmiştir (5, 6).

1999 yılında Roma'da yapılan Cochrane colloquium toplantısında özellikle araştırma sonuçlarının raporlanması konusunda bir standardın sağlanması amacıyla çalışmalarda sık rastlanan hataları içeren bir kontrol listesi oluşturmaya karar verilmiştir. Bu toplantıda temelleri atılan çalışma sonucunda, QUADAS (**QU**ality **A**ssesment of Diagnostic **A**ccuracy **S**tudies) ve STARD (The **STA**ndards for **R**eporting of **D**iagnostic Accuracy) isimleri ile iki kontrol listesi geliştirilmiş ve 2003 yılında yayınlanmıştır (7, 8).

QUADAS; tanı doğruluğu çalışmalarının kalitelerini artırmak amacıyla geliştirilmiştir. Kontrol listesinin oluşturulması sürecinde, aralarında epidemiyolog ve biyoistatistikçilerin de bulunduğu çalışma grubu üyeleri dört kez bir araya gelmiş (Delphi prosedürü) ve 28 maddeden oluşan kontrol listesi, ortak karar ile 14 maddeye indirilmiştir. Listedeki her bir madde için "Evet", "Hayır" ve "Belirsiz" olmak üzere üç ayrı yanıt seçeneği bulunmaktadır.

STARD adıyla bilinen kontrol listesi ise tanı doğruluğu çalışmalarının sonuçlarının raporlanmalarına yönelik kalitenin artırılması amacıyla geliştirilmiştir. Kontrol listesinin geliştirilmesi sürecinde, tanı doğruluğu çalışmalarının değerlendirilmesi amacıyla geliştirilmiş 33 ayrı kontrol listesi olduğu saptanmıştır. Bu listelerin incelenmesi sonucunda 75 maddeden oluşan yeni bir liste hazırlanmış sonrasında yapılan geniş katılımlı bir toplantı ile liste 25 maddeye indirgenerek son haline getirilmiştir. Oluşturulan yeni STARD kriterleri 2003 yılında aynı anda 8 tıbbi dergide (Radiology, American Journal of Clinical Pathology, Annals of Internal Medicine, British Medical Journal, Clinical Biochemistry, Clinical Chemistry, Clinical Chemistry of Laboratory Medicine, and Lancet) yayınlanarak ilan edilmiştir. STARD kriterlerinin yayınlanmasından bu yana 200'den fazla süreli yayında Yazarlara Bilgi/Yazım Kuralları kısmında kullanılması önerilmiştir.

Tanı doğruluğu çalışmalarının tam ve doğru olarak raporlanması, okuyuculara sonuçlarda var olabilecek yanlılığı anlama fırsatı sağlayabileceği gibi elde edilen sonuçların uygulanabilirliği ve genellenebilirliği hakkında da bilgi vermektedir. Çalışmamızın amacı, bilim çevreleri tarafından kabul görmüş ve birçok dile çevrilmiş STARD kriterlerini Türkçeye çevirerek tanı doğruluğu çalışması yapan araştırmacılarımıza çalışmalarını tam ve doğru olarak raporlamaları konusunda katkıda bulunmaktır.

GEREÇ VE YÖNTEM

Dil Eşdeğerliliği Çalışması

STARD kriterlerinin Türkçeye uyarlanması sürecinde ilk olarak, orijinal STARD bildirimi, ilk dört yazar (RA, MAT, CA, PS) tarafından bağımsız olarak Türkçeye çevrilmiştir. Çeviri aşamasında 2003 yılında STARD grubu tarafından yayınlanan ve maddelerin örneklerle açıklamalarının verildiği makaleden yararlanılmıştır (9, 10). Dört farklı çeviri metni tanı doğruluğu çalışmalarının planlanması konusunda uzman YG'nin liderliğinde görüş birliğine varılarak son haline getirilmiştir. Dil eşdeğerliliğinin sağlanması amacıyla, Türkçe uyarlaması yapılan STARD kriterleri, profesyonel bir tercüman tarafından kontrol edilmiş ve gerekli düzenlemeler yapılmıştır. Son hali verilen kontrol listesi, Türkçe ve İngilizce dillerine eş düzeyde hakim bir başka profesyonel tercüman tarafından Türkçeden İngilizceye çevrilmiş ve orijinal STARD kriterleri ile İngilizceye çevrilmiş STARD kriterlerini karşılaştırıldığında aralarında üst düzeyde bir benzerlik olduğunu saptamıştır. Dilimize çevrilen kontrol listesinin araştırmacılar tarafından doğru bir şekilde anlaşılıp anlaşılmadığını ve tekrar edilebilirlik düzeyini belirlemek amacı ile 20 tanı doğruluğu çalışması aynı uzmanlık düzeyine sahip iki patologa (Yazar V, VI) verilmiş ve Türkçe uyarlaması yapılmış olan STARD kontrol listesini kullanılarak, yayınların kalitelerini değerlendir meleri istenmiştir. STARD kriterlerinin Türkçe uyarlaması Tablo 1'de, bu süreçte izlenen yola ilişkin akış diyagramı ise Şekil 1'de yer almaktadır.

Yayın seçimi

Değerlendirmede kullanılan yayınlar, dört kriter ile Türk Tıp Dizininde yapılan tarama sonucunda belirlenmiştir. Bu kriterler, (1) 2009-2010 yıllarında yayınlanmış olması, (2) yayın dilinin Türkçe olması, (3) çalışmanın insanlar üzerinde yapılmış olması ve (4) başlık, özet ya da anahtar kelimelerde "duyarlılık" ve/veya "seçicilik/özgüllük" ve/veya "ROC/İKE/İşlem Karakteristiği Eğrisi" ve/veya "tanı doğruluğu/performansı" kelimelerinin yer almasıdır. Elektronik tarama sonucunda 97 makalenin kriterleri sağladığı belirlenmiş fakat çalışmaların tam metinleri incelendiğinde sadece 20 makalenin tanı doğruluğu çalışması olduğu gözlenmiştir.

İstatistik Yöntemler

Her bir kriterin makalelerde rapor edilme yüzdeleri ve gözlemciler arası uyum yüzdeleri hesaplanmış tır. Ayrıca, gözlemciler arası uyum düzeyi, Cohen'in Kappa istatistiği ve Bland Altman yöntemi kullanılarak, güvenirlik ise Sınıf içi Korelasyon Katsayısı (SKK) kullanılarak değerlendirilmiş olup, tekrar edilebilirlik üç farklı istatistiksel yöntem ile ölçülmüş tür. Uyum istatistikleri, gözlemci ler arası farklılıkların dağılımı hakkında bilgi sağlarken güvenirlik, iki gözlemcinin yüksek ve düşük kalitede raporlanan tanı doğruluğu çalışmalarını ayırt etme yetenekleri hakkında bilgi verir.

İki gözlemci arasındaki uyum, her bir kriter için ayrı ayrı Cohen'in Kappa istatistiği hesaplanarak incelenmiştir. Kappa değerlerinin yorumlanmasında Landis ve Koch tarafından önerilen sınıflandırma dikkate alınmıştır. Buna göre Kappa değerinin $\kappa < 0{,}01$ olması hiç uyumun olmadığını, 0,01-0,20 olması önemsiz uyumun varlığını, 0,21-0,40 zayıf, 0,41-0,60 orta düzeyde, 0,61-0,80 yeterli ve 0,81-1,00 ise mükemmel uyumun varlığını göstermektedir (9).

İki gözlemci arasındaki uyum, toplam raporlanan STARD kriteri sayıları kullanılarak Bland Altman yöntemi ve SKK ile de değerlendirilmiştir. Yirmi beş STARD kriteri bulunduğundan, toplam raporlanan kriter sayısı, her bir makale için 0 ile 25 arasında değişmektedir. Gözlemciler arası uyum, Bland Altman yöntemiyle iki gözlemcinin makalelere verdiği toplam puanlar arasındaki farkların ortalaması (d) ve standart sapması (SS_d) kullanılarak "uyum limitleri" hesaplanarak ölçülmüştür. $d \pm 1{,}96$ SS_d formülü yardımıyla hesaplanan "%95 Uyum Limiti", yan (d) ve rastgele hata (SS_d)'nın toplamıdır. İki değerlendirme arasındaki yan miktarı, d için %95 güven aralığı kullanılarak tahmin edilmiştir. "d" için %95 güven aralığı

$$d \mp 1.96\ x \left[\frac{SS_d}{\sqrt{n}}\right]$$ formülü

kullanılarak hesaplanmış olup, "n" toplam makale sayısını göstermektedir. Aralığın 0'ı içermesi gözlemciler arasında sistematik farklılığın olmadığını gösterir. Ayrıca iki gözlemcinin değerlendirmeleri sonucunda her bir makale için elde ettikleri toplam kriter sayısı ortalamaları X eksenine ve farkları Y eksenine konularak Bland-Altman grafiği çizilmiştir. Bu grafik farkların büyüklüğünü, yönünü ve dağılım aralığını göstermenin yanında ortalamalar arttıkça farkların artıp artmadığını da göstermektedir.

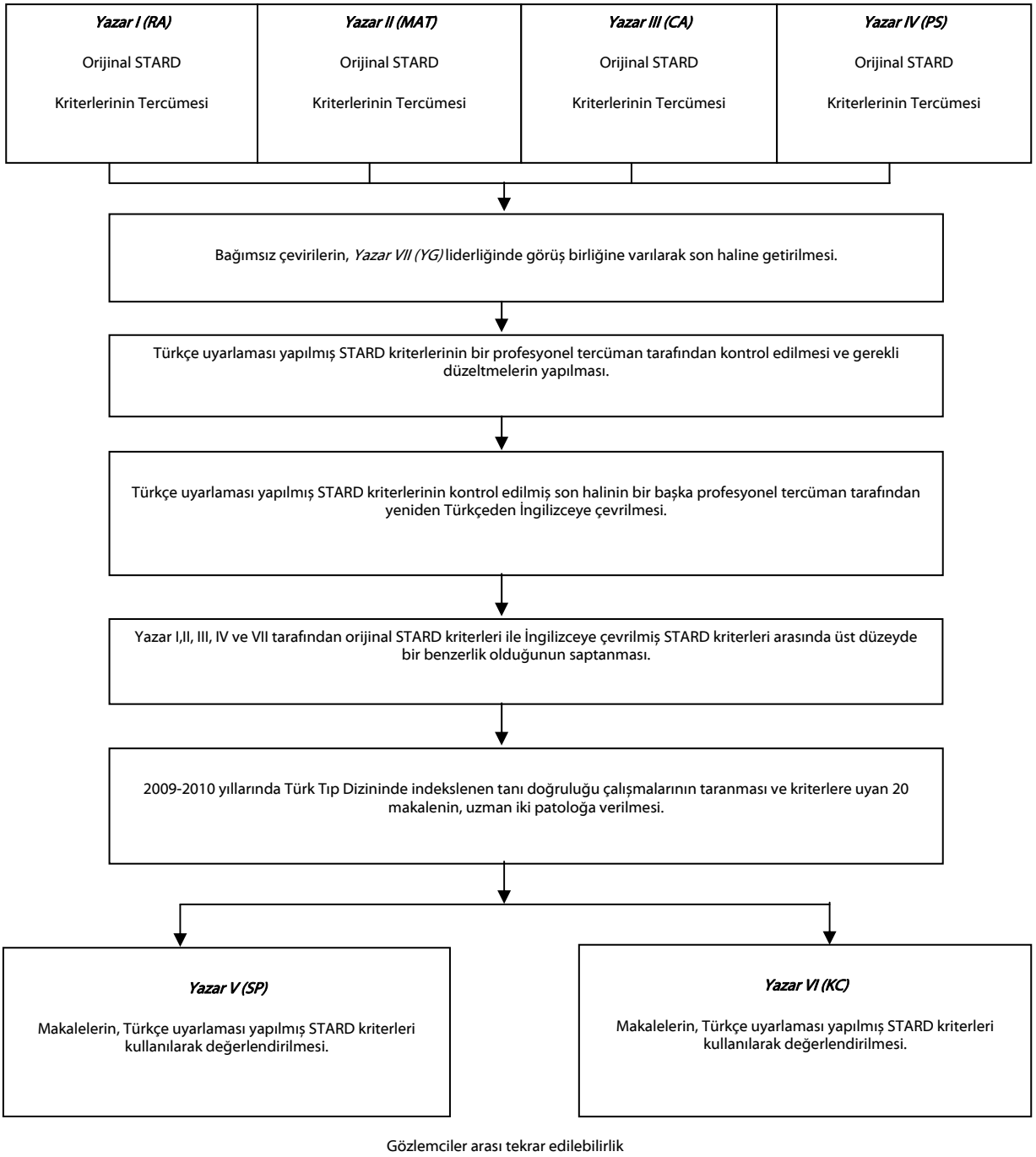
İki yönlü rastgele etkili model (Two way random effect model) kullanılarak hesaplanan SKK, makalelerin raporlanma kalitelerinin varyansının, toplam varyansa (makalelerin toplam puanlarının değişkenliği, değerlendiricilerden kaynaklanan değişkenlik, rastgele hata toplamı) oranlaması olarak tanımlanır (10, 11). SKK, 0 ile +1 aralığında değişir ve 0,75'in üzerinde olması istenir (12, 13).

BULGULAR

STARD kriterlerinin rapor edilme yüzdeleri, her bir kriter için gözlemciler arası uyum yüzdeleri ve Cohen'in Kappa istatistikleri Tablo 2'de verilmiştir.

Tablo 1. STARD kriterlerinin Türkçe uyarlaması

Bölüm ve Konu	Madde		Sayfa Numarası
BAŞLIK/ÖZET/ ANAHTAR KELİMELER	1	Makaleyi bir tanı doğruluğu çalışması olarak tanımlayın (MeSH başlığının duyarlılık ve seçicilik olması önerilir).	
GİRİŞ	2	Tanı doğruluğunun tahmin edilmesi, tanı testlerinin karşılaştırılması ya da tanı doğruluğunun alt gruplarda karşılaştırılması gibi ifadeler kullanarak çalışmanın amacını ve/veya hipotezlerini belirtin.	
METOD			
Denekler	3	Çalışma populasyonunu tanımlayın: Dahil etme ve çıkarma kriterleri, verilerin toplandığı ortam ve mekanlar.	
	4	Katılımcıların çalışmaya alınma prosedürünü tanımlayın: Katılımcı alımı, var olan semptomlara mı, önceki testin sonuçlarına mı, yoksa referans standart (altın standart test) ve/veya indeks test (tanı testi) sonuçlarına mı dayanmaktadır?	
	5	Katılımcı örneklemini tanımlayın: Madde 3 ve 4 deki seçilme kriterlerini sağlayan katılımcılar, çalışmaya ardışık olarak mı alındı? Eğer değilse katılımcıların nasıl seçildiğini belirtin.	
	6	Verilerin toplanma biçimini tanımlayın: Veri toplama biçimi indeks test (tanı testi) ve referans standart (altın standart test) uygulanmadan önce mi (ileriye dönük planlanmış çalışma) uygulandıktan sonra mı (geriye dönük planlanmış çalışma) belirlendi?	
Test Yöntemleri	7	Referans standardı (altın standart testi) ve bilimsel dayanağını tanımlayın.	
	8	Materyal ve metodun teknik özelliklerini, ölçümlerin ne zaman ve nasıl yapıldığını da belirtilerek detaylı olarak açıklayın ve/veya indeks test (tanı testi) ve referans standart (altın standart test) için referanslar gösterin.	
	9	İndeks test (tanı testi) ve referans standart (altın standart test) için kullanılan birimleri, kesim noktalarını ve belirlenen kategorilerin geçerliliğini gerekçeli olarak tanımlayın.	
	10	İndeks testi (tanı testini) ve referans standardı (altın standart testi) uygulayan kişilerin sayısını, eğitimini ve uzmanlık düzeyini belirtin.	
	11	İndeks testi (tanı testini) ve referans standardı (altın standart testi) uygulayan gözlemcilerin, katılımcıların klinik bilgilerine ve diğer test sonuçlarına kör olup olmadıklarını belirtin.	
İstatistiksel Yöntemler	12	Tanı doğruluğu ölçütlerinin hesaplanmasında veya karşılaştırılmasında kullanılan yöntemleri ve belirsizlik tahmini (örneğin, %95 güven aralığı) için kullanılan istatistiksel yöntemleri belirtin.	
	13	Eğer yapılmış ise testin tekrar edilebilirliğini hesaplamak için kullanılan yöntemleri belirtin.	
SONUÇLAR			
Katılımcılar	14	Çalışmanın yapıldığı zamanı, katılımcı alımına başlama ve katılımcı alımını sonlandırma tarihleri ile birlikte belirtin.	
	15	Çalışma populasyonunun klinik ve demografik özelliklerini (örneğin, yaş, cinsiyet, var olan semptomların spektrumu, eşlik eden hastalıklar, uygulanan tedaviler, tedavi merkezleri) raporlayın.	
	16	İndeks test (tanı testi) ve/veya referans standart (altın standart test) uygulanıp/uygulanmayıp, dahil etme kriterlerini sağlayan katılımcıların sayısını bildirin. Katılımcıların neden her iki testi de almadıklarını tanımlayın (akış diyagramı önerilir).	
Test sonuçları	17	İndeks testler (tanı testleri) ile referans standart (altın standart test) arasında geçen zamanı ve bu süreçte herhangi bir tedavi uygulanıp uygulanmadığını belirtin.	
	18	Hedef koşullara sahip olanlarda hastalık şiddetinin (kriter tanımla) dağılımı ve hedef şartlara uymayan katılımcılara ait diğer teşhisleri raporlayın.	
	19	İndeks testlerin (tanı testlerinin) sonuçlarını (belirlenemeyen ve kayıp sonuçlar da dahil olmak üzere) referans standardın (altın standart testin) sonuçları ile çaprazlayarak tablo halinde verin; sürekli sonuçlar için ise test sonuçlarının referans standart (altın standart test) sonuçlarına göre dağılımını verin.	
	20	İndeks test (tanı testi) veya referans standart (altın standart test) uygulanırken karşılaşılan olumsuz durumları raporlayın.	
Tahminler	21	Tanı doğruluğu tahminlerini ve istatistiksel belirsizlik ölçütlerini (örneğin, %95 güven aralığı) raporlayın.	
	22	İndeks testlerde (tanı testlerinde) şüpheli sonuçların, eksik (kayıp) yanıtların ve sapan değerlerin nasıl ele alındığını raporlayın.	
	23	Eğer yapıldıysa, katılımcıların alt grupları, gözlemciler veya merkezler arasında tanı doğruluğunun değişkenlik tahminlerini raporlayın.	
	24	Eğer yapıldıysa, testlerin tekrarlanabilirlik tahminlerini raporlayın.	
TARTIŞMA	25	Çalışma bulgularının klinik uygulanabilirliğini tartışın.	



Şekil 1. STARD kriterlerinin Türkçeye uyarlanması sürecini gösteren akış diyagramı

Tablo 2. STARD kriterlerinin rapor edilme yüzdeleri, her bir kriter için gözlemciler arası uyum yüzdeleri ve Cohen'in Kappa istatistikleri (n=20)

	Kriterlerin Rapor Edilme Yüzdeleri		Gözlemciler Arası Uyum Yüzdesi	
	Gözlemci I	Gözlemci II		
Maddeler	n(%)	n(%)	%	Cohen Kappa
Başlık / Özet / Anahtar Kelimeler				
1	20(100)	20(100)	100	1,00
Giriş				
2	19(95)	19(95)	100	1,00
Metod				
3	20(100)	20(100)	100	1,00
4	20(100)	18(90)	90	-
5	2(10)	7(35)	75	0,34
6	10(50)	14(70)	80	0,60
7	18(90)	15(75)	85	0,50
8	19(95)	18(90)	95	0,64
9	15(75)	15(75)	80	0,47
10	6(30)	4(20)	90	0,74
11	6(30)	5(25)	95	0,88
12	13(65)	12(60)	85	0,68
13	3(15)	2(10)	95	0,77
Sonuçlar				
14	16(80)	15(75)	95	0,86
15	18(90)	17(85)	85	0,32
16	19(95)	17(85)	80	-0,08
17	14(70)	14(70)	100	1,00
18	14(70)	12(60)	90	0,78
19	17(85)	16(80)	85	0,48
20	9(45)	7(35)	60	0,18
21	10(50)	9(45)	95	0,90
22	5(25)	6(30)	85	0,63
23	1(5)	2(10)	85	-0,07
24	3(15)	0(0)	85	-
Tartışma				
25	20(100)	20(100)	100	1,00

STARD Kriterlerinin Raporlanması

Değerlendirilen 20 tanı doğruluğu çalışması sonucunda rapor edilmeleri bakımdan kriterler arasında büyük bir varyasyonun olduğu gözlenmiştir. Kriterlerden bazıları (madde 1, madde 3, madde 25) bütün makalelerde de rapor edilmişken bazıları ise (madde 23, madde 24) neredeyse hiç rapor edilmemiştir. İki gözlemcinin sonuçları birlikte değerlendirildiğinde, makalelerin en az %75'inde raporlanan 12 kriter bulunmaktadır. Dört kriter makalelerin %74-%50'sinde rapor edilirken, 9 kriter makalelerin %50-%0'ında raporlanmıştır (Tablo 2). Gözlemci I'in değerlendirmesi sonucunda toplam raporlanan STARD kriteri sayısı ortalaması makale başına 15,9±3,2 iken Gözlemci II için de bu değer oldukça benzer bulunmuştur. Makalelerde 25 kriterin en az 10'u en fazla 21'i raporlanmış olup STARD kriterlerinin tümünü raporlayan herhangi bir makaleye rastlanmamıştır (Şekil 2).

Gözlemciler Arası Uyum

Gözlemciler arası uyum yüzdelerine bakıldığında en düşük uyum %60 ile Madde 20 için bulunmuş olup iki gözlemci 20 makalenin 12'sinde aynı cevabı (kriter sağlanıyor/kriter sağlanmıyor) vermiştir. Diğer tüm maddeler için uyum yüzdesi %75'in üzerindedir (Tablo 2).

Her bir kriter için gözlemciler arası uyum Cohen'in Kappa istatistiği kullanılarak incelendiğinde bu değerlerin 8 maddede 0,81-1,00, 6 maddede 0,61-0,80, 4 maddede

0,41-0,60, 2 maddede 0,21-0,40 aralığında, 3 madde de ise 0,20 den daha düşük düzeyde olduğu gözlenmiştir. İki madde için ise kappa istatistikleri hesaplanamamıştır.

Bland-Altman grafiği iki gözlemcinin değerlendirmeleri sonucunda elde ettikleri toplam raporlanan kriter sayısı ortalamaları arttıkça fark değerlerinin artmadığını göstermektedir. Ayrıca farklar ortalama etrafında simetrik bir şekilde dağılmakta olup fazla yaygın değildir (Şekil 2).

Raporlanan toplam STARD kriteri sayıları kullanılarak gözlemciler arasındaki uyum Bland-Altman yöntemiyle değerlendirildiğinde %95 Uyum Limitleri -2,2 – 3,5 olarak bulunmuştur. Buna göre %95 güvenirlikle Gözlemci II bir makalede 2,2 kriterin daha fazla, 3,5 kriterin daha az raporlandığını belirtebilir. Bu sonuçlar iki gözlemci arasındaki uyumun kabul edilebilir düzeyde olduğunu göstermektedir (Şekil 2). Ayrıca farklara ilişkin %95 Güven Aralığının 0,0-1,3 olarak bulunduğu ve 0'ı içerdiği gözlenmiştir. Bu sonuç iki gözlemcinin değerlendirmelerinde sistematik bir yanlılığın olmadığını ve Uyum Limitlerine ilişkin sonuçların güvenle kullanılabileceğini göstermektedir.

Güvenirlik, ölçümlerin tekrarlanabilirliği ya da tekrarlı ölçümlerin tutarlılığı olarak tanımlanır. Sağlık alanında en yaygın kullanılan güvenirlik çalışmaları, gözlemci-içi ve gözlemciler arası uyumdur (10). Buradan yola çıkarak Türkçeye uyarlanmış STARD kontrol listesinin her iki gözlemci tarafından da aynı şekilde anlaşılıp yorumlandığı elde edilen SKK ile açıklanabilir. Çalışmamızda iki gözlemcinin her bir makale için elde ettikleri toplam raporlanan STARD kriteri sayıları kullanılarak hesaplanan SKK, 0,945 (%95 GA; 0,852-0,979) gibi oldukça yüksek bir değer bulunmuştur. Hesaplanan bu istatistiğe göre iki patolog arasındaki uyumun istatistiksel olarak anlamlı düzeyde olduğu söylenebilir ($p < 0,001$).

TARTIŞMA

Çalışmamızda en küçük gözlemciler arası uyum % 60 ile "indeks test veya referans standart uygulanırken karşılaşılan olumsuz durumların raporlanması" kriteri için elde edilmiş olup ikinci en düşük değer %75 ile "katılımcı örnekleminin tanımlanması" kriteri için hesaplanmıştır. Diğer tüm maddeler için gözlemciler arası uyum yüzdesi %80 ile %100 arasında bulunmuştur.

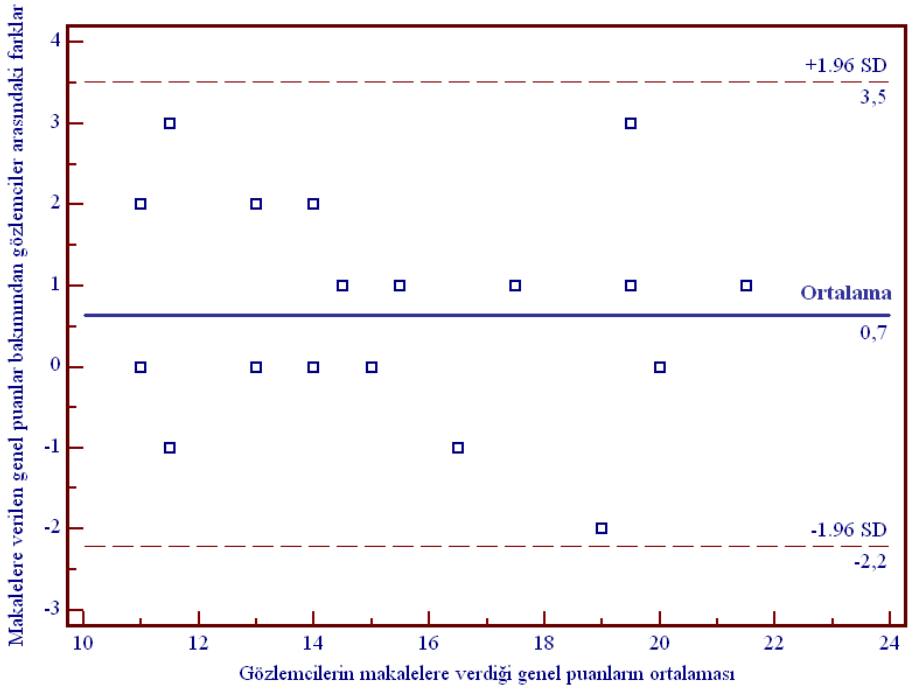
Ancak katılımcıların örnekleme nasıl alındığının (madde 5), indeks testi ve referans testi uygulayan kişilerin sayısının, eğitiminin, uzmanlık düzeyinin (madde 10) ve bunların katılımcıların klinik bilgilerine ve diğer test sonuçlarına kör olup olmadıklarının (madde 11) açıklandığı maddelerle birlikte, testin tekrar edilebilirliğini hesaplamak için kullanılan yöntemlerin belirtildiği 13. maddenin de çok düşük düzeyde raporlandıkları görülmüştür. Benzer şekilde şüpheli sonuçların, eksik (kayıp) yanıtların ve sapan değerlerin nasıl ele alındığı (madde 22), katılımcı alt grupları, gözlemciler veya merkezler arasında tanı doğruluğunun değişkenlik tahminleri (madde 23) ve tekrarlanabilirlik tahminleri (madde 24) araştırıcılar tarafından önem verilmeyen diğer maddeler arasında yer almaktadır.

Bunların yanında, kontrol listesinde bulunmaları gereken makale bölümleri de belirtilmesine rağmen, araştırmacıların bazı maddelere ilişkin açıklamaları farklı bölümlerde yaptıkları gözlenmiş bu konuda konuda da ortak bir yaklaşımın olmadığı sonucuna varılmıştır. Örneğin, çalışmanın amacı ve/veya hipotezinin belirtilmesinin istendiği 2. maddeye ilişkin açıklamanın giriş yerine tartışma bölümünde açıklandığı pek çok makaleye rastlanılmıştır. Yine değerlendirilen makalelerde dikkat çeken bir nokta da özellikle çalışma popülasyonunun tanımlanması ve dâhil etme/ hariç bırakma kriterlerinin tanımlandığı 3. ve 4. maddelerin açıklanmasına temel tıp, mikrobiyoloji ve psikiyatri alanında yapılan çalışmalarda yeterli önemin verilmediği saptanırken, kardiyoloji alanında yapılan çalışmalarda bu maddelere ilişkin açıklamaların özenle ele alındığı gözlenmiştir.

Stengel ve arkadaşlarının 62 tanı doğruluğu makalesini içeren benzer ça çalışmasında gözlemciler arası uyum yüzdeleri %58 ile %98 arasında değişirken Smidt ve arkadaşlarının 32 makaleyi içeren çalışmalarında uyum yüzdeleri %63 ile %100 arasında değişmektedir (14,15).

Türkçe STARD kriterleri kullanılarak yapılan değerlendirmeler sonucu elde edilen uyum yüzdelerinin orijinal STARD kriterleri kullanılarak yapılan bu iki çalışmayla benzer sonuçlar vermesi, gözlemcilerin kriterleri benzer şekilde yorumladığını göstermektedir. Yüksek gözlemciler arası uyum yüzdelerine rağmen iki maddede Cohen'in Kappa istatistikleri; -0,08 (madde 16) ve -0,07 (madde 23) gibi uyum olmadığını gösterir değerler bulunmuştur. Dikkatlice incelendiğinde bunun prevelansın çok yüksek (ya da çok düşük) olmasından kaynaklandığı anlaşılmıştır (16,17).

Makalelerde kriterlerin raporlanma oranlarına bakıldığında hemen hemen tüm çalışmalarda özet ya da anahtar kelimeler bölümlerinden en az birinde çalışmanın tanı doğruluğu çalışması olarak belirtildiği, çalışmanın amacının,



Gözlemci I		Gözlemci II		Gözlemci I- Gözlemci II (d)		% 95 Uyum Limitleri
Ort (SS)	Min-Maks	Ort (SS)	Min-Maks	Ort_d (SS_d)	%95 GA	
15.9 (3.2)	11 - 22	15.2 (3.6)	10 - 21	0.65(1.5)	0.0 - 1.3	-2.2 - 3.5

Ort=Aritmetik Ortalama, SS=Standart Sapma, Min= Minimum, Maks= Maksimum, d=Fark, Ort_d =Farkların ortalaması, SS_d= Farkların Standart Sapması, GA= Güven Aralığı

Şekil 2. 20 makalede toplam raporlanan STARD kriteri sayılarına ilişkin Bland-Altman grafiği

çalışma popülasyonunun özelliklerinin ve katılımcıların çalışmaya alınma prosedürlerinin tanımlandığı gözlenmiştir. Bunun yanında çalışmalarda materyal ve metodun teknik özellikleri, çalışma popülasyonunun klinik ve demografik özellikleri büyük oranda verilmiş ve çalışma bulgularının klinik uygulanabilirliği tartışılmıştır.

Toplam raporlanan kriter sayısına bakıldığında çalışmamızda gözlemci 1, makale başına ortalama 15,9±3,2 kriterin raporlandığını belirtmişken gözlemci 2 bu değeri, 15,2 ± 3,6 olarak bulmuştur. Smidt ve arkadaşları yaptıkları çalışmada etki faktörü en az 4 olan dergilerde 2000 yılında yayınlanan 32 makalede ortalama 12,08±3,9 STARD kriterinin raporlandığını belirtmişlerdir. Türk Tıp Dizininde yayınlanan Türkçe tanı doğruluğu çalışmalarında raporlanan toplam kriter sayısının daha fazla olması, çalışmamızın 2009-2010 yılında yayınlanan çalışmaları kapsamasından kaynaklandığı düşünülmektedir. Yıllar içinde daha fazla sayıda STARD kriterinin raporlandığını görmek sevindirici olsa da bu düzeyin yeterli olmadığı düşünülmektedir. Değerlendirilen 20 çalışmanın hiç birinde akış diyagramına yer verilmemiştir. Detaylı bir biçimde hazırlanan akış diyagramı özellikle Metod bölümünde yer alan kriterlerden birçoğunu içerecektir.

SONUÇ

Türkçe uyarlaması yapılmış STARD kriterlerinin ülkemizde tanı doğruluğu çalışması yapacak olan araştırmacıların yanında dergi hakem ve editörlerin için de bir kılavuz olması umut edilmektedir. Çalışmalarda metodolojik standartlara uygun ortak bir ölçüt kullanılması, bireysel çalışmaların yanında meta analizi çalışmalarının da kalitelerinin artmasını sağlayacaktır

KAYNAKLAR

1. Genç Y. Tanı Testi Çalışmalarında Metodolojik Standartların Kullanılması. Ankara Üniversitesi Tıp Fakültesi Mecmuası. 2003;56:4:259-264.

2. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: A new meta-analytic method. Med Decis Making 1993;13:313-321.

3. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: Getting better but still not good. JAMA 1995;274:645-651.

4. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of deign-related bias in studies of diagnostic tests. JAMA 1999;282:1061-1066.

5. Nelemans PJ, Leiner T, Henrica CW, Joseph M.A, Engelshoven V. Peripheral arterial disease: Meta-analysis of the diagnostic performance of MR angiography. Radiol 2000;217:105-114.

6. De Vries S.O., Hunnink MGM, Polak JF. Summary receiver operating characteristic curves as a technique for metaanalysis of the diagnostic performance of duplex ultrasonography in peripheral ar-terial disease. Acad Radiol 1996;3:361-369.

7. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. Clin Chem 2003;49:1 1-6.

8. 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.

9. Whiting PF, Weswood ME, Rutjes AW, et al. Eva-luation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 2006;6:9.

10. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. Clin Chem 2003;49:1 7-18.

11. Ateş C, Öztuna D, Genç Y. Sağlık araştırmalarında sınıf içi korelasyon katsayısının kullanımı. Türkiye Klinikleri J Bios-tat. 2009;1:59-64.

12. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol. Med. 1989;19: 61-70.

13. Fleiss JF, Chapter 1: Reliability of Measurement. In the Design and Analysis of Clinical Experiments. John Wiley & Sons, London. 1986;1-33.

14. Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. Association between compliance with methodological standarts of diagnostic research and reported test accuracy: metaanalysis of focused assesment of US for trauma. Radiol 2005;236:102-111.

15. Smidt N, Rutjes AWS, Van der Windt AWM et al. Quality of reporting of diagnostic accuracy studies. Radiol 2005;235:347-353.

16. Smidt N, Rutjes AWS, Van der Windt DAWM at al. Reproducibility of the STARD checklist: An instrument to assess the quality of reporting of diagnostic accuracy studies. BMC Medical Res Methodol 2006;6:12.

17. Vach W. The dependence of Cohen's Kappa on the prevalence does not matter. J Clin Epidemiol 2005;58:655-661.