Çok Testli Çok Değerlendiricili ROC Çalışmalarında Tanı Testleri Arasındaki İlişkinin Diagnostik Doğruluk Sonuçlarına Etkisi

The Effect of the Relation Between Diagnosis Tests to the Diagnostic Accuracy Results in Multi-Reader Multi Case ROC Studies

Gülhan OREKİCİ TEMEL,[a]
E. Arzu KANIK[a]

[a]Biyoistatistik AD,
Mersin Üniversitesi Tıp Fakültesi,
Mersin

Geliş Tarihi/*Received:* 17.01.2011
Kabul Tarihi/*Accepted:* 24.03.2011

Yazışma Adresi/*Correspondence:*
E. Arzu KANIK
Mersin Üniversitesi Tıp Fakültesi,
Biyoistatistik AD, Mersin,
TÜRKİYE/TURKEY
arzukanik@gmail.com

ÖZET Amaç: Bu çalışmanın amacı çok tanı testi ve çok değerlendirici var olduğu çalışmalarda tanı testleri arasındaki ilişkinin, örnek genişliği, değerlendirici sayısı, tanı testi sayısı ve kategorisi çerçevesinde diagnostik doğruluk üzerine etkisini ortaya koymaktır. **Gereç ve Yöntemler:** Etki değerlendirici sayısı, tanı testi sayısı, kategorisi ve örnek genişliği çerçevesinde değerlendirilmiştir. Vakaların örnek büyüklüğü (15,30 ve 100), tanı testi sayısı (t=2,5,7), tanı testinin ölçüm düzeyi (2,3,5,7,10) ve değerlendirici sayısı (2,5,7) alınarak kombinasyonlar hazırlanmıştır. Deneme planlarına ait verilerin üretimi her kombinasyon için Matlab 7.0 paket programında 1000 kez üretilmiş ve bu 1000'lik verilerin analizleri DBM Metodunun algoritmasını kullanan LABMRMC 1.0 paket programında gerçekleştirilmiştir. **Bulgular:** İkili tanı testleri için değerlendirici sayısı ne olursa olsun gerçekleşen Tip I Hatalar %5'in çok üzerindedir. Değerlendirici sayısı ve örnek genişliği artıkça bu hataların arttığı gözlenmektedir. **Sonuç:** Tanı testleri arasında çok güçlü ve zayıf korelasyonun var olduğu deneme planlarında, tanı testi sayısı, değerlendirici sayısı ve örnek genişliğine ait tüm durumlarda tanı testinin ikili yapıda olması değerlendiricilerin diagnostik doğrulukları üzerinde olumsuz etkiye neden olmaktadır. Tanı testinin kategorisi en az 3 olduğunda bu sorun gözlenmemektedir.

Anahtar Kelimeler: DBM; OR; Çok Testli Çok Değerlendiricili ROC

ABSTRACT Objective: The aim of this study is to reveal the effect of the relation between diagnosis tests to the diagnostic accuracy in multi-reader multi case studies with the limits of their sample size, the number of multi-readers, the number of diagnosis tests and categories. **Material and Methods:** Effect has been evaluated with the limits of the number of readers, the number of diagnosis tests and categories and the sample size. Combinations have been prepared by getting observations' sample size (15,30 and 100), the number of diagnosis test (t=2,5,7), the diagnosis test's measurement level (2,3,5,7,10) and the number of readers. The production of the experimental design data have been produced 1000 times on Matlab 7.0 program for each combination and the analysis of 1000 data have been performed on LABMRMC 1.0 program that uses algorithm DBM methods. **Results:** Whatever the number of the readers for the binary diagnosis tests are, Type 1 Errors are over %5; and it has been observed that as the number of the readers and sample size increase, these values also increase. **Conclusion:** Very strong and weak correlation between diagnosis tests in experimental design, the diagnosis test's being in binary in all number of diagnosis tests, the number of readers and sample size cases have been caused negative effect on readers accuracy. This problem hasn't been observed when the category of diagnosis test at least three.

Key Words: MRMC ROC; DBM; OR

Turkiye Klinikleri J Biostat 2011;3(2):63-73

ROC verilerinin analizinde genellikle geleneksel metodlar kullanılır. Bu durumda ya tek bir değerlendiricinin bir grup vaka üzerindeki sonuçları ya da birden fazla değerlendiricinin yine bir grup vaka üzerindeki sonuçları analiz edilir.[1] Bazı durumlarda ölçülen değişken tek bir değerlendirici için ya da bir değerlendirici grubu için oluşturulmuş olabilir. Bu durumda da değerlendirici grubu için ya da her bir tanı testi sonuçları için ortalama eğri altında kalan alanlar hesaplanarak performans değerlendirmesi yapılır.[2] Ancak klinikte karşılaşılan veri yapısı aslında bu kadar basit değildir. Tanı amaçlı veriler, korelasyonlu veri ya da kümelenmiş veri şeklinde olabilir. Bir vakadan farklı tanı testleri için ölçümler alındığında korelasyonlu veri, bir vakanın farklı bölgelerinden örneğin sağ ve sol gözünden ya da sağ ve sol el bileklerinden ölçümler alındığında kümelenmiş veriler elde edilir. Bu deneme düzenine farklı değerlendiricilerin verileri de eklendiğinde veri analizinde geleneksel ROC analizinden farklı olarak hem tanı testlerinin vaka gruplarını ayırmadaki performansını, hem de değerlendiricilerin performansını Çok Testli Çok Değerlendiricili ROC analizleriyle çözümlemek gerekir.[3]

Her değerlendiricinin her vakayı değerlendirdiği tam çapraz deneme planlı ROC verilerinin değerlendirilmesini 1992 yılında Dorfman ve ark. çalışmalarında ilk defa çok testli çok değerlendiricili (multi reader multi case, MRMC), ÇTÇD olarak isimlendirmişlerdir.[4] Çok değerlendiricili ROC çalışmaları analiz edilirken kullanılan farklı istatistik analiz yöntemleri vardır.[5] Metodlardan en sık kullanılanı Dorfman ve ark. tarafından önerilen Dorfman-Berbaum-Metz (DBM) yöntemi ve Obuchowski ve Rockette tarafından önerilen düzeltilmiş F metodudur kısaca OR olarak adlandırılır.[4,6,7] DBM sözde değerler (pseudovalues) için üç yönlü ANOVA yaparken, OR doğruluk tahminleri için ilişkilendirilmiş hataların iki yönlü ANOVA analizini yapar.

ÇTÇD deneme planlarında diagnostik doğruluk tartışılırken, tanı testleri arasındaki uyum/uyumsuzluk durumunun, tanı testi ve değerlendirici sayısının ve tanı testlerinin kategorisinin birlikte değerlendirilmesi gerekmektedir. Fakat ÇTÇD ROC analizlerinde bu faktörlerin ve kombinasyonlarının diagnostik doğruluk üzerine etkisinin araştırıldığı bir çalışma literatürde mevcut değildir. Bu çalışmada tanı testleri arasındaki güçlü ve zayıf ilişkinin tanı testlerinin ve değerlendiricilerin doğrulukları üzerine etkisine bakılmıştır.

GEREÇ VE YÖNTEMLER

ÇTÇD ROC VERİLERİNİN DENEME TASARIMI

ÇTÇD ROC çalışmalarında en genel deneme tasarımı R değerlendiricinin C vakayı K tanı testini birlikte değerlendirdiği faktöriyel deneme tasarımıdır.

Kabul edelim ki, R değerlendirici, C vaka ve K tanı testi var olsun. Bu durumda her değerlendiricinin C×K tane sonucu ve bütün çalışmanın ise R×K×C tane sonucu vardır. ÇTÇD analizlerinin deneme düzenleri bir tür faktöriyel deneme düzenidir. Her değerlendirici bir ya da birden fazla test sonucunu değerlendirir. Tablo 1'de ÇTÇD ROC analizi için deneme tasarımı görülmektedir.[3,8]

DBM METODU TEORİSİ

Çok değerlendiricili ROC çalışmalarında en sık kullanılan metod DBM metodudur. Dorfman ROC verilerinin çok değerlendiricili ile yapıldığı çalışmalarda karışık etkili (mixed effect) ANOVA modelini kullanmaktadır. Modelin testinde amaç "değerlendiricilerin doğrulukları arasında fark yoktur", "diagnostik testlerin doğruluğu arasında fark yoktur" ve "değerlendiricilerin ortalama doğruluğu bütün diagnostik testler için aynıdır" ifadeli yokluk hipotezlerini test etmektir. Yokluk hipotezinin testinde Quenouille-Tukey Jackknife sözde değerleri ile ANOVA hesaplaması yapılır. Karışık etkili ANOVA, sözde değerler ile yapılır. Sözde değerler orijinal verinin bir çeşit transformasyonu gibi düşünülür.

t tane tanı şekli (i=1,2,...,t), r tane değerlendirici (j=1,2,..,r) ve c tane de vakanın bulunduğu bir veri seti (k=1,2,...,c) olsun. ROC verilerinin analiz edilirken her değerlendirici için Maksimum Likelihood (ML) tahmin edicisi kullanılarak, her değerlendirici-tanı kombinasyonu için binormal ROC eğrisi elde edilir. Her değerlendirici-tanı kombinasyonu için ROC eğrisinden bir A_z indeksi elde

TABLO 1: ÇTÇD Doğruluk Analizleri için Faktöriyel Deneme Tasarımı.

Vaka (Hasta/Sağlam)	Tanı Testleri															
	1				2				....				K			
	Değerlendirici				Değerlendirici				...				Değerlendirici			
	R_1	R_2	.	R_r	R_1	R_2		R_r	.	.	.	.	R_1	R_2	...	R_r
1	Y_{111}	Y_{121}	.	Y_{1r1}	Y_{112}	Y_{122}		Y_{1r2}	.	.	.	.	Y_{11k}	Y_{12k}	...	Y_{1rk}
2	Y_{211}	Y_{221}	.	Y_{2r1}	Y_{212}	Y_{222}		Y_{2r2}	.	.	.	.	Y_{21k}	Y_{22k}	...	Y_{2rk}
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...	.
c	Y_{C11}	Y_{C21}		Y_{Cr1}	Y_{C12}	Y_{C22}		Y_{Cr2}	.	.	.	.	Y_{C1k}	Y_{C2k}		Y_{Crk}

edilir. ML tahmini j değerlendirici tarafından i tanı değişkeni için bütün vakalar üzerinden yapılır ve A_{ij} ile gösterilir. Sonra 1. vaka silinir ve ML ile A_{ij}'nin yeni bir tahmini hesaplanır. Sonra 1. vaka örnekleme yeniden dahil edilir ve 2. vaka silinir ve A_{ij}'nin yeni bir tahmini hesaplanır ve böylece bu işlemler bir döngü şeklinde devam eder. Bu durum örneklemdeki her vaka bir kez silininceye kadar devam eder. Bu işlemle, eğri altında kalan alan c kez hesaplanır. Çünkü örneklem üzerinde c tane vaka vardır.

Bu durumda $A_{ij(k)}$ k. vaka silindiğinde A_{ij}'nin ML tahminini gösterir. j değerlendirici ve i tanı için k tane sözde değer A_{ij*k} ile gösterilir ve 1 nolu eşitlik ile hesaplanır.

$$A_{ij*k} = cA_{ij} - (c-1)A_{ij(k)} \quad (1)$$

Burada c ve c-1 faktörleri k. vaka silinerek oluşturulan alt kümede k. vakanın ağırlığını gösterir. Bu ilişki, bütün vakalar üzerinde k. vakanın ortalama etkisi olarak düşünülebilir. Silinen vakanın hasta ya da sağlam olup olmamasına bağlı olarak, negatif ya da pozitif vaka sayısından bir vaka düşülür.[9-12]

DBM metodu karma etkili doğrusal ANOVA modeli olarak kabul edilir.[4] Bu metodun modeli eşitlik 2'de verilmiştir.

$$Y_{ijk} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \varepsilon_{ijk} \quad (2)$$

Burada

μ Populasyon ortalamasını,

τ_i i tanı testinin sabit etkisini

R_j j değerlendiricinin tesadüfi etkisini

C_k k vakanın tesadüfi etkisini

ε_{ijk} k vaka, j değerlendirici ve i. tanın hata miktarını gösterir.

Parantez ile verilen ifadeler interaksiyon etkisini gösterir.

İnteraksiyon terimlerinin (tanı_testi×değerlendirici, tanı_testi×vaka, değerlendirici×vaka ve tanı_testi×vaka×değerlendirici) hepsi rastgele etkilidir.

Bir tanı testinin performansı için F istatistiği eşitlik 3'de verilmiştir.

$$\text{Fistatistiği} = \frac{MS(T)_{pseudo}}{MS(T\times R)_{pseudo} - MS(T\times C)_{pseudo} - MS(T\times R\times C)_{pseudo}} \quad (3)$$

MS(T): Tanı etkisinin kareler toplamını

MS(T×R): Tanı-değerlendirici interaksiyon kareler toplamı

MS(T×C): Tanı-vaka interaksiyon kareler toplamı

MS(T×R×C): Tanı-değerlendirici-vaka interaksiyon kareler toplamını göstersin.

Karar eğer F değeri; $(F_{\alpha(1),\alpha(2)})(1-\alpha)$ F tablo değerinden büyükse tanı testlerinin diagnostik

doğrulukları arasındaki farklılık anlamlıdır. F istatistiğine karşılık gelen kritik tablo değerlerinin hesaplaması eşitlik 4 ve eşitlik 5'de verilmiştir.

Burada

$$df_1 = (t-1) \quad (4)$$

$$df_{satt} = \frac{[MS(T\times R)_{pseudo} + MS(T\times C)_{pseudo} - MS(T\times R\times C)_{pseudo}]^2}{\frac{MS(T\times R)^2_{pseudo}}{(t-1)(r-1)} + \frac{MS(T\times C)^2_{pseudo}}{(t-1)(c-1)} + \frac{MS(T\times R\times C)^2_{pseudo}}{(t-1)(r-1)(c-1)}} \quad (5)$$

ÇTÇD ROC analizlerinde performans değeri eğri altında kalan alandır.[6-11]

BENZETİM DENEMELERİ

Çok değerlendirici ve çok tanı testinin bulunduğu çalışmalarda tanı testleri arasındaki korelasyonun diagnostik doğruluk sonuçlarını (eğri altında kalan alan) etkilediği düşünülmektedir. Ayrıca tanı testinin sayısı ve tanı testinin ölçüm düzeyi, değerlendirici sayısı, vaka sayısı da doğruluk sonuçlarını etkilemektedir. Etki değerlendirici sayıları, tanı testi sayısı, kategorisi ve örnek genişliği çerçevesinde değerlendirilecektir. Vakaların örnek büyüklüğü (15, 30 ve 100), tanı testi sayısı (t=2,5,7), tanı testinin ölçüm düzeyi (2,3,5,7,10) ve değerlendirici sayısı (2,5,7) alınarak kombinasyonlar hazırlanmıştır. Tanı testleri arası çok güçlü korelasyon için iki, beş ve yedi tanı testi sonucunda her bir kombinasyonu için tanı testi sonuçları arasında 0.90'lık korelasyonu yakalamaları sağlanmıştır. Tanı testleri arası çok zayıf korelasyon için iki, beş ve yedi tanı testi sonucunda her bir kombinasyonu için tanı testi sonuçları arasında 0.10'lık korelasyonu yakalamaları sağlanmıştır. Her koşulda tanı testlerinin ayırma gücü %50 olarak alınmıştır. Her bir kombinasyon için üretilen verilerde gerçekte değerlendiricilerin ortalama doğrulukları bakımından fark yoktur ve bu farksızlık bütün diagnostik testler için aynıdır. Yani değerlendiriciler ve tanı testlerinin doğrulukları arasında farklılık yoktur. Bu nedenle değerlendiricilerin diagnostik doğrulukları için Hipotez Takımı I ve tanı testlerinin diagnostik doğrulukları için de Hipotez Takımı II aşağıdaki gibi kurulabilir.

Hipotez 1:

H_0= Değerlendiricilerin ortalama diagnostik doğrulukları bakımından fark yoktur.

H_1= Değerlendiricilerin ortalama diagnostik doğrulukları bakımından fark vardır.

Üretilmiş olan bütün kombinasyonlarda Hipotez Takımı I için Tip I Hata yapma olasılığını, değerlendiricilerin diagnostik doğrulukları arasında gerçekte farklılık yokken var olarak bulunma ihtimali olarak ifade edebiliriz.

Hipotez II:

H_0= Tanı testlerinin ortalama diagnostik doğrulukları bakımından fark yoktur.

H_1= Tanı testlerinin ortalama diagnostik doğrulukları bakımından fark vardır.

Bu durumda da üretilmiş olan bütün kombinasyonlarda Hipotez Takımı II için Tip I Hata yapma olasılığını, tanı testlerinin diagnostik doğrulukları arasında gerçekte farklılık yokken var olarak bulunma ihtimali olarak ifade edebiliriz.

Deneme planlarına ait verilerin üretimi her kombinasyon için Matlab 7.0 paket programında 1000 kez üretilmiş ve bu 1000 genişlikli set analizleri DBM Metodunun algoritmasını kullanan LABMRMC 1.0 paket programında gerçekleştirilmiştir. Sonuçların grafik gösterimi SPSS 17.0 paket programında yapılmıştır.

BULGULAR

Değerlendirici sayısı, tanı testi sayısı ve kategorisi, tanı testleri arasında korelasyonun çok yüksek ve çok zayıf olma ihtimalleri için üretilmiş olan kombinasyonlardaki tanı testlerinin diagnostik doğrulukları arasında gerçekte farklılık yokken var olarak bulunma ihtimalini gösteren Tip I Hata yapma olasılıkları iki tanı testi için Tablo 2, beş tanı testi için Tablo 3 ve yedi tanı testi için Tablo 4'de verilmiştir. Tanı testleri arasında korelasyonun çok zayıf ve çok güçlü olduğu durumlar için değerlendirici sayısı, örnek genişliği, tanı testinin kategorilerine göre tanı testlerinin diagnostik doğrulukları arasında gerçekte fark yokken var olarak bulunma ihtimalini gösteren, Tip I Hata yapma olasılıkları Şekil 1'de sunulmuştur.

Tablo 2, 3 ve 4 incelendiğine iki, beş ve yedi tanı testi arasında çok güçlü ve çok zayıf korelasyon

TABLO 2: İki tanı testi arasındaki korelasyon çok güçlü ve zayıf olduğunda tanı testlerinin diagnostik doğruluk farklılıklarına ilişkin, olabilecek Tip I Hata yapma olasılıkları (%)

Değerlendirici Sayısı	Örnek Genişliği	İki Tanı Testi Arasındaki Korelasyon Çok Güçlü					İki Tanı Test Arasındaki Korelasyon Çok Zayıf				
		Tanı Testinin Kategori Sayısı					Tanı Testinin Kategori Sayısı				
		K=2	K=3	K=5	K=7	K=10	K=2	K=3	K=5	K=7	K=10
İki değerlendirici	N=15	0.2	0.0	0.0	0.0	0.0	0.0	2.2	1.7	2.2	1.4
	N=30	0.0	0.9	0.5	1.5	0.6	0.0	2.1	2.7	2.3	2.6
	N=100	0.0	1.3	1.9	2.2	2.1	0.1	3.3	3.7	3.0	3.4
Beş değerlendirici	N=15	0.0	0.0	0.0	0.0	0.0	0.0	1.6	2.0	1.4	0.2
	N=30	0.0	0.8	0.5	0.6	0.6	0.0	2.2	1.4	1.5	1.6
	N=100	0.0	1.9	1.7	1.7	1.4	0.0	2.7	2.8	1.9	2.2
Yedi değerlendirici	N=15	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.8	1.4	1.6
	N=30	0.0	0.6	0.4	1.0	0.4	0.0	2.8	2.0	2.8	2.8
	N=100	0.0	1.3	1.6	1.7	0.9	0.0	3.2	2.8	3.3	2.1

TABLO 3: Beş tanı testi arasındaki korelasyon çok güçlü ve zayıf olduğunda tanı testlerinin diagnostik doğruluk farklılıklarına ilişkin, olabilecek Tip I Hata yapma olasılıkları (%).

Değerlendirici Sayısı	Örnek Genişliği	Beş Tanı Testi Arasında Korelasyon Çok Güçlü					Beş Tanı Test Arasında Korelasyon Çok Zayıf				
		Tanı Testinin Kategori Sayısı					Tanı Testinin Kategori Sayısı				
		K=2	K=3	K=5	K=7	K=10	K=2	K=3	K=5	K=7	K=10
İki değerlendirici	N=15	0.1	0.0	0.2	0.2	0.0	0.0	1.4	1.4	1.5	1.5
	N=30	0.0	1.3	1.1	1.4	0.9	0.0	2.8	2.3	2.3	2.1
	N=100	0.0	1.7	1.6	2.4	2.9	0.0	3.5	1.9	0.2	3.0
Beş değerlendirici	N=15	0.0	0.0	0.0	0.0	0.0	0.0	1.6	1.4	1.3	1.5
	N=30	0.0	1.1	0.7	0.4	0.8	0.0	1.4	1.9	2.4	1.6
	N=100	-	-	-	-	-	-	-	-	-	-
Yedi değerlendirici	N=15	0.0	1	0.0	0.0	0.0	0.0	1.4	1.7	1.9	1.6
	N=30	0.0	0.6	1.8	1.5	1.4	0.0	2.3	2.3	2	1.4
	N=100	-	-	-	-	-	-	-	-	-	-

-Paket programın kapasitesi yeterli gelmediği için hesaplanamamaktadır.

TABLO 4: Yedi tanı testi arasındaki korelasyon çok güçlü ve zayıf olduğunda tanı testlerinin diagnostik doğruluk farklılıklarına ilişkin, olabilecek Tip I Hata yapma olasılıkları (%).

Değerlendirici Sayısı	Örnek Genişliği	Yedi Tanı Testi Arasında Korelasyon Çok Güçlü					Yedi Tanı Test Arasında Korelasyon Çok Zayıf				
		Tanı Testinin Kategori Sayısı					Tanı Testinin Kategori Sayısı				
		K=2	K=3	K=5	K=7	K=10	K=2	K=3	K=5	K=7	K=10
İki değerlendirici	N=15	0.0	0.0	0.0	0.0	0.0	0.0	1.4	1.3	1.3	1.2
	N=30	3.5	3.5	2.6	1.9	1.4	0.0	2.6	2.2	2.1	1.2
	N=100	0.0	1.8	1.9	1.7	1.5	0.0	3.4	2.8	2.4	2.1
Beş değerlendirici	N=15	0.0	0.0	0.0	0.0	0.0	0.0	1.8	1.4	1.1	1.2
	N=30	0.0	0.9	1.3	1.1	1.0	0.0	3.1	2.5	2.6	1.4
	N=100	-	-	-	-	-	-	-	-	-	-
Yedi değerlendirici	N=15	0.0	0.0	0.0	0.0	0.0	0.0	1.8	1.8	1.7	1.6
	N=30	0.0	1.0	0.1	1.3	1.3	0.0	2.3	3.4	1.9	2.0
	N=100	-	-	-	-	-	-	-	-	-	-

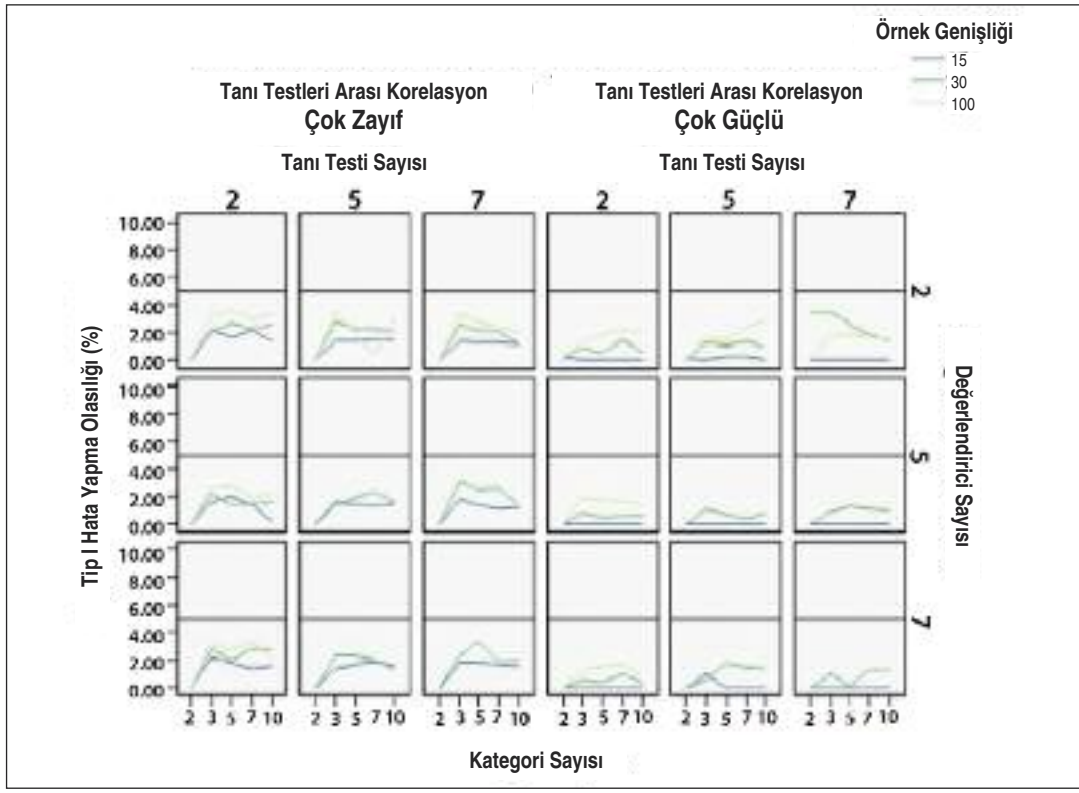
-Paket programın kapasitesi yeterli gelmediği için hesaplanamamaktadır.

varken değerlendirici sayısı ve örnek genişliği ve tanı testinin kategorisinden etkilenmeksizin tanı testlerinin diagnostik doğrulukları arasında gerçekte farklılık yokken var olarak bulunma ihtimalini gösteren Tip I Hata yapma olasılıkları daima %5'in altındadır.

Değerlendirici sayısı, tanı testi sayısı ve kategorisi, tanı testleri arasındaki korelasyonun çok yüksek ve çok zayıf olma ihtimalleri için üretilmiş olan kombinasyonlardaki değerlendiricilerin diagnostik doğrulukları arasında gerçekte farklılık yokken, farklılık var olarak bulunma ihtimalini gösteren Tip I Hata yapma olasılıkları iki tanı testi için Tablo 5, beş tanı testi için Tablo 6 ve yedi tanı testi için Tablo 7'de verilmiştir. Bu kombinasyonlardaki değerlendiricilerin diagnostik doğrulukları arasında gerçekte fark yokken fark var olarak bulunma olasılığını gösteren Tip I Hata yapma olasılıkları Şekil 2'de sunulmuştur.

Tablo 5 incelendiğinde iki tanı testi arasında korelasyon çok güçlü olduğunda değerlendirici sayısı ve örnek genişliğinden etkilenmeksizin tanı testinin kategorisi 3-5-7 ve 10 olduğu durumda Tip I Hata yapma olasılığı %6'nın altındadır. Fakat tanı testlerinin kategorisi ikili yapıda olduğu durumda 2 değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %40.9, 30 için %60.8 ve 100 için %65.4'dür. Yedi değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %78.7, 30 için %89.6 ve 100 için %97.3'dür. Yedi tanı testinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %86.0, 30 için %99.4 ve 100 için %99.5'dir.

Bununla birlikte iki tanı testi arasında korelasyon çok zayıf yine kategori sayısı 3-5-7 ve 10 olduğu durumda Tip I Hata yapma olasılığı örnek genişliği ve değerlendirici sayısından etkilenmeksizin %6'nın altındadır. Fakat tanı testi ikili yapıda



ŞEKİL 1: İki-beş ve yedi tanı testi arasında korelasyon çok güçlü ve çok zayıf iken, tanı testlerinin diagnostik doğruluk farklılıklarına ilişkin, olabilecek Tip I Hata yapma olasılıklarının grafik gösterimi.

olduğu durumda 2 değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %41.6, 30 için %60.4 ve 100 için %63.2'dir. Beş değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %70.1, 30 için %89.9 ve 100 için %96.9'dur. Yedi değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %86.6, 30 için %95.6 ve 100 için %99.2'dir.

Tablo 6 incelendiğinde beş tanı testi arasında korelasyon çok güçlü olduğunda, değerlendirici sayısı ve örnek genişliğinden etkilenmeksizin tanı testinin kategorisi 3,5,7 ve 10 olduğu durumda Tip I Hata yapma olasılığı %6'nın altındadır. Fakat tanı testi ikili yapıda ve 2 değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %39.1, 30 için %65.4 ve 100 için %60.5'dir. Beş değerlendiricinin var olduğu durumda olasılıklar örnek genişliği 15 için %75.4, 30 için %90.6'dir. Yedi değerlendiricinin var olduğu durumda da örnek genişliği 15 için %84.9 ve 30 için %94.2'dür.

Bununla birlikte beş tanı testi arasında korelasyon çok zayıf, tanı testi ikili yapıda ve 2 değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %41.8, 30 için %63.2 ve 100 için %62.7'dir. Beş değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %76.6, 30 için %94.0'dır. Yedi değerlendiricinin var olduğu durumda da bu olasılık örnek genişliği 15 için %89.9 ve 30 için %98.4'dür.

Tablo 7 incelendiğinde yedi tanı testi arasında çok güçlü bir korelasyon var olduğunda, değerlendirici sayısı ve örnek genişliğinden etkilenmeksizin tanı testinin kategorisi 3, 5, 7 ve 10 olduğu durumda Tip I Hata yapma olasılığı %5'in altındadır. Fakat tanı testi ikili yapıda ve 2 değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %40.4, 30 için %61.7 ve 100 için %62.3'dür. Beş değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %77.0, 30 için %90.0'dir. Yedi değerlendiricinin

TABLO 5: İki tanı testi arasında korelasyon çok güçlü ve çok zayıf olduğunda, değerlendiricilerin diagnostik doğruluk farklılıklarına ilişkin, olabilecek Tip I Hata yapma olasılıkları (%)

Değerlendirici Sayısı	Örnek Genişliği	İki Tanı Testi Arasında Korelasyon Çok Güçlü					İki Tanı Testi Arasında Korelasyon Çok Zayıf				
		Tanı Testinin Kategori Sayısı					Tanı Testinin Kategori Sayısı				
		K=2	K=3	K=5	K=7	K=10	K=2	K=3	K=5	K=7	K=10
İki değerlendirici	N=15	40.9	5.9	5.5	4.5	3.9	41.6	5.5	3.7	3.3	4.2
	N=30	60.3	4.8	3.3	4.6	3.5	60.4	4.8	3.5	3.1	3.3
	N=100	65.4	5.2	5.3	4.3	4.1	63.2	4.5	5.0	4.4	5.5
Beş değerlendirici	N=15	78.7	4.7	3.6	3.1	4.2	70.1	3.9	3.5	3.4	2.7
	N=30	89.6	4.5	3.7	5.4	3.1	89.9	5.2	3.3	3.2	3.1
	N=100	97.3	5.1	5.7	5.2	5.2	96.9	5.0	4.7	5.0	4.4
Yedi değerlendirici	N=15	86.0	4.6	3.7	2.9	3.5	86.6	4.7	3.0	2.6	3.1
	N=30	94.4	5.1	3.4	4.8	2.3	95.6	4.3	3.2	3.2	2.8
	N=100	99.5	5.3	4.3	4.9	4.9	99.2	4.8	4.3	4.7	5.2

-Paket programın kapasitesi yeterli gelmediği için hesaplanamamaktadır.

var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %83.0 ve 30 için %94.9'dur.

Bununla birlikte beş tanı testi arasında korelasyon çok zayıf, tanı testleri ikili yapıda ve 2 değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %46.6, 30 için %60.3 ve 100 için %61.1'dir. Beş değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %83.0, 30 için %89.4'dür. Yedi değerlendiricinin var olduğu durumda hata yapma olasılığı örnek genişliği 15 için %75.4 ve 30 için %96.6'dır.

TARTIŞMA

Dorfman ve ark.nın 1992 yılında ortaya attığı DBM metodu ile tanı testleri arasındaki korelasyonun diagnostik doğruluk sonuçlarını etkilemediğini iddia etmişlerdir. Klinikte diagnostik doğruluk sonuçları değerlendirilirken değerlendiricilerin ve tanı testlerinin diagnostik doğrulukları çerçevesinde değerlendirmek gerekir.

Tanı testleri arasında çok güçlü ve zayıf korelasyon olduğu durumlarda tanı testi sayısı, değerlendirici sayısı ve kategorisi ve hatta örnek genişliğinden etkilenmeksizin tanı testlerinin diagnostik doğrulukları arasında gerçekte farklılık yokken var olarak bulunma olaslığını gösteren Tip I Hatalar %5'in altında elde edilmiştir. Bu bulgular DBM metodunda söz edilen ve bu metodla tanı testleri arası korelasyonun tanı testlerinin performanslarının klasik yöntemlerle karşılaştırılmasında yaratacağı olumsuz etkileri giderilebileceği savını desteklemektedir.

İki, beş ve yedi tanı testi arasında çok güçlü ve zayıf korelasyon varken değerlendirici sayısı ve ör-

TABLO 6: Beş tanı testi arasında korelasyon çok güçlü ve zayıf olduğunda, değerlendiricilerin diagnostik doğruluk farklılıklarına ilişkin olabilecek Tip I Hata yapma olasılıkları (%).

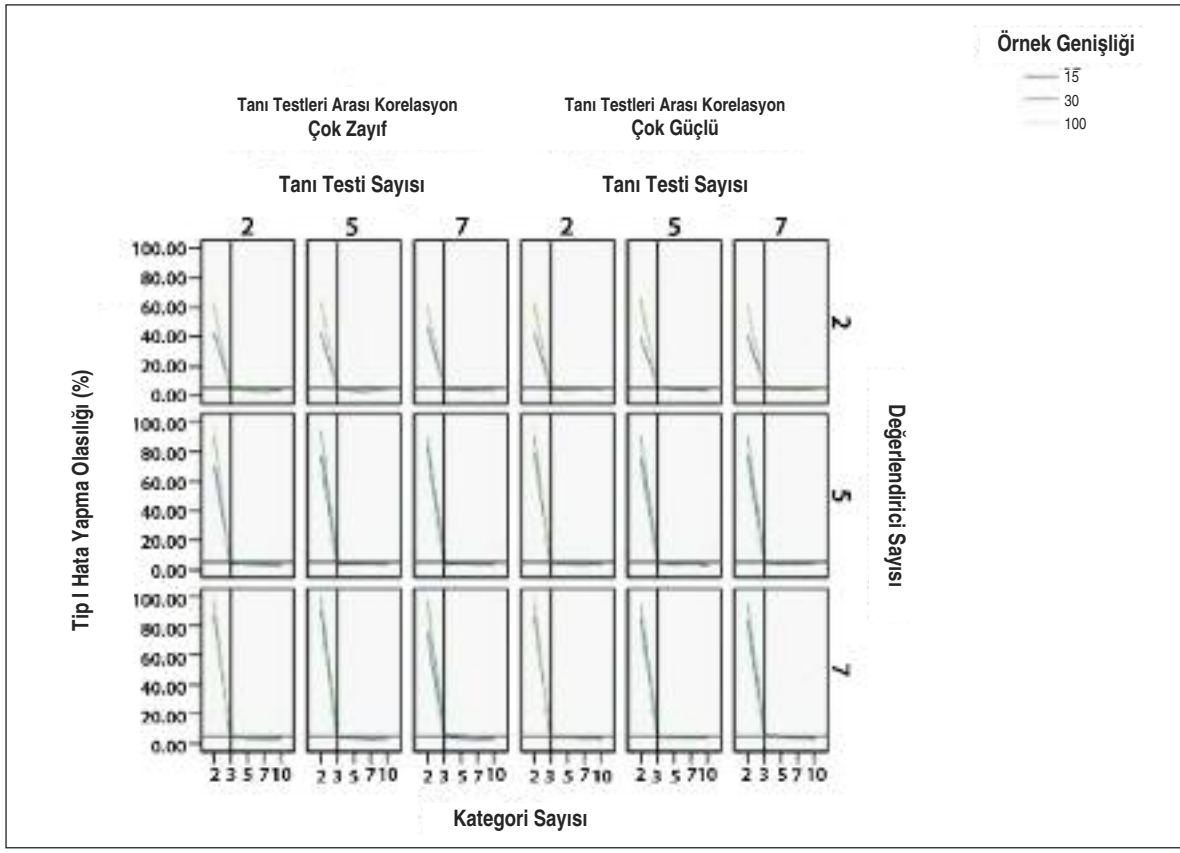
Değerlendirici Sayısı	Örnek Genişliği	Beş Tanı Testi Arasında Korelasyon Çok Güçlü					Beş Tanı Test Arasında Korelasyon Çok Zayıf				
		Tanı Testinin Kategori Sayısı					Tanı Testinin Kategori Sayısı				
		K=2	K=3	K=5	K=7	K=10	K=2	K=3	K=5	K=7	K=10
İki değerlendirici	N=15	**39.1**	5.6	4.5	3.9	3.2	**41.8**	5.4	2.7	3.2	3.9
	N=30	**65.4**	5.5	4.6	4.3	4.6	**63.2**	4.9	4.5	5.6	4.0
	N=100	**60.5**	5.8	5.2	5.5	5.8	**62.7**	5.6	4.8	4.6	4.2
Beş değerlendirici	N=15	**75.4**	4.9	3.1	4.3	2.3	**75.6**	3.4	3.9	3.6	3.3
	N=30	**90.6**	5.1	4.4	3.5	4.5	**94.0**	4.5	4.9	3.9	4.0
	N=100	-	-	-	-	-	-	-	-	-	-
Yedi değerlendirici	N=15	**84.9**	5.3	4.2	4.0	3.4	**89.9**	4.6	3.1	2.4	3.0
	N=30	**94.2**	4.5	3.8	3.2	4.0	**98.4**	4.4	4.2	4.4	4.5
	N=100	-	-	-	-	-	-	-	-	-	-

-Paket programın kapasitesi yeterli gelmediği için hesaplanamamaktadır.

TABLO 7: Yedi tanı testi arasında korelasyon çok güçlü ve zayıf olduğunda, değerlendiricilerin diagnostik doğruluk farklılıklarına ilişkin olabilecek Tip I Hata yapma olasılıkları (%).

Değerlendirici Sayısı	Örnek Genişliği	Yedi Tanı Testi Arasında Korelasyon Çok Güçlü					Yedi Tanı Testi Arasında Korelasyon Çok Zayıf				
		Tanı Testinin Kategori Sayısı					Tanı Testinin Kategori Sayısı				
		K=2	K=3	K=5	K=7	K=10	K=2	K=3	K=5	K=7	K=10
İki değerlendirici	N=15	**40.4**	5.9	3.9	4.3	3.6	**46.6**	3.7	3.7	3.6	3.6
	N=30	**61.7**	4.4	4.6	3.5	4.6	**60.3**	4.5	3.9	5.0	5.9
	N=100	**62.3**	4.5	4.8	4.9	4.2	**61.1**	4.3	5.4	4.9	4.2
Beş değerlendirici	N=15	**77.0**	4.6	3.7	3.7	4.4	**83.0**	3.5	4.0	3.3	3.2
	N=30	**90.0**	4.8	3.3	4.3	4.5	**89.4**	5.6	4.6	3.9	4.5
	N=100	-	-	-	-	-	-	-	-	-	-
Yedi değerlendirici	N=15	**83.0**	4.8	4.4	3.2	2.9	**75.4**	4.2	3.0	2.3	3.0
	N=30	**94.9**	6.1	5.2	3.9	3.6	**96.6**	6.3	5.5	4.6	2.6
	N=100	-	-	-	-	-	-	-	-	-	-

-Paket programın kapasitesi yeterli gelmediği için hesaplanamamaktadır.



ŞEKİL 2: İki, beş ve yedi tanı testi arasında çok güçlü ve çok zayıf korelasyon varken, değerlendiriciler arasındaki diagnostik doğruluk farklılığına ilişkin, Tip I Hata yapma olasılıklarının grafik gösterimi.

nek genişliğinden etkilenmeksizin tanı testinin kategorisi 3, 5, 7 ve 10 olduğu durumda değerlendiricilerin diagnostik doğrulukları arasında gerçekte farklılık olmadığında farklılık var olarak bulunma olasılıklarını gösteren Tip I Hata değerleri %6'nın altındadır. Fakat ikili tanı testleri için değerlendirici sayısı ne olursa olsun gerçekleşen Tip I Hatalar %5'in çok üzerindedir ve değerlendirici sayısı ve örnek genişliği artıkça bu değerlerin arttığı ve hatta %90'ların üzerine çıktığı gözlenmektedir. Tanı testinin kategorisinin ikili yapıda olduğu durumdaki bu bulgular DBM metodunu desteklememektedir.[4] Ayrıca Dorfman ve ark.nın çalışmasında tanı testlerinin kategori düzeyi ile ilgili herhangi bir kısıtlama bulunmamaktadır.[4] Literatürde tanı testinin kategorisi ikili yapı için çok testli çok değerlendiricili deneme planını kullanarak yapılan simülasyon denemeleri bulunmaktadır. DBM metodunu kullanmalarına karşın Tip I hatalar üzerine bir sonuç yayınlamamışlardır.[13]

SONUÇ

Jackknife yeniden örnekleme yöntemini kullanan Dorfman ve ark. tanı testleri arasındaki korelasyonun diagnostik doğruluk sonuçlarını etkilemediğini ortaya atmışlardır.[4] Bu çalışmada yapılan simülasyon denemeleri de bu sonucu kısmen doğrulamaktadır. Tanı testleri arasında çok güçlü ve zayıf korelasyonun var olduğu deneme planlarında değerlendirici sayısı, tanı testi sayısı, tanı testinin kategorisi ve örnek genişliği için üretilen tüm durumlarda bu korelasyonun tanı testlerinin diagnostik doğrulukları üzerine olumsuz bir etkisi gözlenmemiştir. Bu sonuç DBM metodunu destekler niteliktedir.[4] Bu nedenle amaç tanı testlerinin diagnostik doğrulukları olduğu çalışmalarda tanı testinin kategori düzeyi doğruluk sonuçlarını etkilememektedir.

Tanı testleri arasında çok güçlü ve zayıf korelasyonun var olduğu deneme planlarında, tanı tes-

ti sayısı, değerlendirici sayısı ve örnek genişliğine ait tüm durumlarda tanı testinin ikili yapıda olması değerlendiricilerin diagnostik doğrulukları üzerinde olumsuz etkiye neden olmaktadır. Bu sonuç ise Dorfman ve ark.nın çalışmasını desteklememektedir.[4] Bu nedenle çok tanı testli ve çok değerlendiricinin bulunduğu ve asıl amacın değerlendiricilerin diagnostik doğrulukları olduğu diagnostik çalışmalarda, tanı testinin kategorisinin en az 3'lü yapıda olması gerektiği önerilmektedir.

KAYNAKLAR

1. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143(1):29-36.
2. Beam CA Analysis of clustered data in receiver operating characteristic studies. Stat Methods Med Res 1998;7(4):324-36.
3. Zhou X, Obuchowski N, McClish D. The Design of Diagnostic Accuracy Studies, Analysis of Correlated ROC Data. Statistical Methods in Diagnostic Medicine. 1st ed. New York: John Wiley; 2002. p:77-307.
4. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992;27(9):723-31.
5. Obuchowski NA, Beiden SV, Berbaum KS, Hillis SL, Ishwaran H, Song HH, et al. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. Acad Radiol 2004;11(9):980-95.
6. Dofrman DD, Berbaum KS, Lenth RV. Multireader, multicase receiver operating characteristic methodology: a bootstrap analysis. Acad Radiol 1995;2(7):626-33.
7. Obuchowski NA, Rockette HE. Hypothesis Testing of the Diagnostic Accuracy for Multiple Diagnostic tests: An Approach with Dependent Observations. Commun Stat Simul Computation 1995;24(2):285-308.
8. Obuchowski NA. Multireader receiver operating characteristic studies: a comparison of study designs. Acad Radiol 1995;2(8):709-16.
9. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. Stat Med 2005;24(10):1579-607.
10. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy BA. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. Acad Radiol 1998;5(9): 591-602.
11. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. Acad Radiol 2004;11(11):1260-73.
12. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. Acad Radiol 2005;12(12):1534-41.
13. Gallas BD, Pennello GA, Myers KJ. Multireader multicase variance analysis for binary data. J Opt Soc Am A Opt Image Sci Vis 2007;24(12):B70-80.