Tıbbi araştırmalarda güvenilirlik ve geçerlilik

Zekeriya Aktürk, H. Acemoglu

Angaben zur Veröffentlichung / Publication details:

Aktürk, Zekeriya, and H. Acemoglu. 2012. "Tıbbi araştırmalarda güvenilirlik ve geçerlilik." *Dicle Tip Dergisi* 39 (2): 316–19. https://doi.org/10.5798/diclemedj.0921.2012.02.0150.

Dicle Tıp Dergisi /
Dicle Medical Journal

2012; 39 (2): 316-319
doi: 10.5798/diclemedj.0921.2012.02.0150

DERLEME / REVIEW ARTICLE

Tıbbi araştırmalarda güvenilirlik ve geçerlilik

Reliability and validity in medical research

Zekeriya Aktürk[1], Hamit Acemoğlu[2]

[1] *Atatürk Üniversitesi Tıp Fakültesi Aile Hekimliği AD, Erzurum, Türkiye*

[2] *Atatürk Üniversitesi Tıp Fakültesi Tıp Eğitimi AD, Erzurum, Türkiye*

Geliş Tarihi / Received: 30.09.2011, Kabul Tarihi / Accepted: 05.03.2012

ÖZET

Tıbbi araştırmalarda sıklıkla araştırma ölçeklerine başvururuz. Sadece psikometrik ölçümler için değil, her çeşit ölçümde güvenilirlik ve geçerlilik kavramlarını gündeme getirebiliriz. Bu yazıda güvenilirlik ve geçerlilik kavramlarının örneklendirilerek açıklanması amaçlanmıştır. Bir ölçeğin/ölçümün güvenilirliğini ve geçerliliğini bilimsel yöntemlerle değerlendirmek mümkündür. Bir ölçeğin güvenilirliğinden bahsedince akla stabilliği (aynı örneklemde yapılan tekrarlayan ölçümlerden aynı sonucun alınması), eşdeğerliliği ve homojenliği gelir. Homojenlik açısından güvenilirlik ölçeğin iç özelliğiyle (iç tutarlılık "internal consistency") ilgilidir; aynı yapıyı ölçen maddelerin ne kadar benzer sonuçlar verdiğinin ölçülmesidir. Yani aynı özelliği ölçmede farklı maddelerin ne kadar tutarlı olduğunun incelenmesidir. Bu kavramlara göre aşağıdaki güvenilirlik çeşitleri ile ölçebiliriz: paralel formlar, test/tekrar test (test/retest), gözlemciler arası güvenilirlik, yarıya bölme (split half) ve Cronbach alfa. Geçerlilik açısından içerik geçerliliği (Content validity), tahmin ettirici geçerlilik (Predictive validity), yapı geçerliliği (Construct validity), eş zamanlı geçerlilik (Concurrent validity) ve Görünüm geçerliliğinden (Face validity) bahsedilebilir. Bilim ölçüm demektir. Yanlış ölçümler yapmak kanıtlarımızın ve tedavilerimizin güvenilmez olmasına yol açar. Bu nedenle ölçüm araçlarımızın geçerli ve güvenilir sonuçlar vermesine azami önem göstermeliyiz.

Anahtar kelimeler: ölçme, geçerlilik, güvenilirlik

ABSTRACT

Scientists commonly refer to study instruments during medical research. In fact, the reliability and validity issues go beyond psychometric studies and can be linked with any kind measurements. In this study we aimed to explain the reliability and validity concepts by giving examples. It is possible to evaluate the reliability and validity of an instrument by scientific methods. If we speak of reliability, we have to mention stability (having the same results in repeated measurements from the same sample), equivalence, and homogeneity. Homogeneity is related with internal consistency; it measures how close results are obtained from items intending to measure the same structure. In other words, how consistent are the different items in measuring the same feature? Accordingly, the following types of reliability can be measured: parallel forms, test/retest, inter-observer reliability, split half, and Cronbach alfa. From the point of validity, we will discuss content validity, predictive validity, construct validity, concurrent validity, and face validity. Science means measurement. Wrong measurements will make our evidence and thus the treatments unreliable. Therefore, medical researchers have to give utmost importance in receiving valid and reliable results from the instruments they use.

Key words: measurement, validity, reliability

GİRİŞ

Ölçme, "Bir niteliğin gözlenip, gözlem sonucunun sayı ve sembollerle gösterilmesidir"[1] İstatistikte güvenilirlik, ölçme aracının kendi içinde kararlılığı ve tutarlılığı olup, standart hatanın az olması demektir.[2] Geçerlilik ise bir ölçme aracının ölçmek istediği değişkeni ölçüp ölçmediği, ölçüyorsa onu başka değişkenlerden ne derece ayırarak ölçtüğüdür.

Bir testin söz konusu bir durumu ölçebilmesi için (a) bahse konu edilen durum var olmalıdır ve (b) ölçümü hedeflenen durumdaki değişimler ölçüm

Yazışma Adresi /Correspondence: Dr. Hamit Acemoğlu
Atatürk Üniversitesi Tıp Fakültesi Tıp Eğitimi AD, Erzurum Eposta: hamitacemoglu@yahoo.com

sonuçlarını da değiştirmelidir.[3] Tıbbi araştırmalarda sıklıkla araştırma ölçeklerine başvurulur. Yaşam kalitesi için SF-36,[4] ağrı şiddeti için WOMAC,[5] depresyon için Beck depresyon ölçeği,[6] benlik saygısı için Rosenberg [7] kullanıldığı gibi, daha birçok sağlık durumunu değerlendirmek için ölçekleri kullanırız.

Aslında sadece psikometrik ölçümler için değil, her çeşit ölçümde güvenilirlik ve geçerlilik kavramlarını gündeme getirebiliriz. Örneğin biyokimyasal analiz yapan cihazların da doğru ölçüp ölçmediklerinden ve her defasında aynı sonucu vermelerinden emin olmak gerekir.

Bu yazıda güvenilirlik ve geçerlilik kavramlarının örneklendirilerek açıklanması amaçlanmıştır.

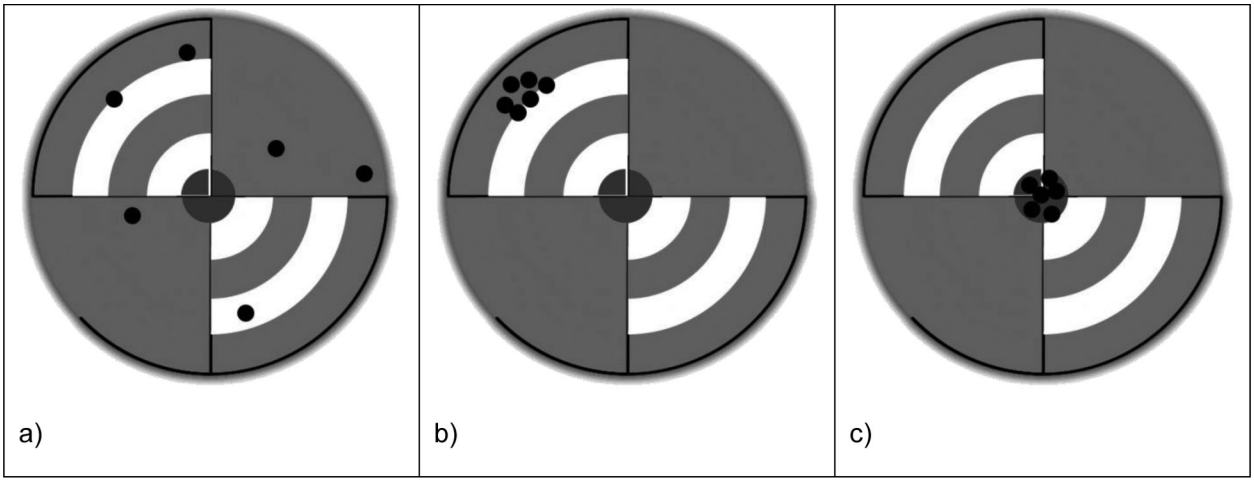
Güvenilirlik ve Geçerlilik Kavramları

Özellikle sözlü sınavlarda şikayet edilen bir durum vardır: eğitici iyi bir sınav yaptığını düşünür ama öğrenciler aynı görüşte olmayabilir. Sınavın eğitimin içeriğini ne kadar ölçtüğü sorgulanabileceği gibi, aynı sınava tekrar girilse geçme/kalma durumunun çok farklı olacağı da iddia edilebilir.

Bu durumda alt ekstremite kemiklerinin anlatıldığı anatomi dersinin sınavında birkaç çeşit eğiticiden bahsedebiliriz. Birinci eğitici tipi sınavda hem konuyla alakasız sorular sorar, hem de her öğrenciye farklı içerikte sorular sorar. Hedefi tutturmaktan uzak olan ve hem de her ölçümde farklı sonuçlar veren bu ölçüme "Hem güvenilmez hem de geçersiz" ölçüm denilebilir (Şekil 1a).

İkinci eğitici tipi ise yine anlatılan içerikle ilgili sorular sormaz ama tüm öğrencilere örneğin üst ekstremite kemikleriyle ilgili sorular sorar. Bu eğiticinin ölçümü için "Güvenilir ama geçersiz" bir ölçüm diyebiliriz (Şekil 1b).

Üçüncü eğitici tipi ise sorularını alt ektsremite kemikleri hakkında hazırlamıştır ve tüm öğrencilere de aynı içerikte sorular sormaktadır. Bu eğiticinin ölçümü için ise "Hem güvenilir, hem de ama geçerli" bir ölçüm diyebiliriz (Şekil 1c).



Şekil 1. a) Güvenilmez ve geçersiz ölçüm, b) Güvenilir ama geçersiz ölçüm, c) Güvenilir ve geçerli ölçüm.

Şüphesiz her ölçüm aracından hem güvenilir, hem de geçerli olması beklenir. Ancak, tamamen tutarsız bir eğiticidense, en azından yanlış yerden de sınav yapsa hep aynı "ters köşeye yatıran" eğiticinin tercih edileceği gibi, güvenilir olmayan bir ölçek için geçerliliğin bir değerinin olmayacağını söyleyebiliriz.

GÜVENİLİRLİK ÇEŞİTLERİ

Bir ölçeğin güvenilirliğinden bahsedince akla stabilliği (Aynı örneklemde yapılan tekrarlayan ölçümlerden aynı sonucun alınması halinde ölçeğin stabil olduğu söylenir.), eşdeğerliliği ve homojenliği gelir.[8] Homojenlik açısından güvenilirlik ölçeğin iç özelliğiyle ilgilidir (iç tutarlılık da "internal consistency" denir). Aynı yapıyı ölçen maddelerin ne kadar benzer sonuçlar verdiğinin ölçülmesidir. Yani aynı özelliği ölçmede farklı maddelerin ne kadar tutarlı olduğunun incelenmesidir. Bu kavramları açacak olursak aşağıdaki güvenilirlik çeşitlerinden bahsedebiliriz:

1. Paralel formlar

Paralel form güvenilirliğine bakmak için aynı kişilere iki farklı ölçek uygulanıp aralarındaki korelasyona bakılabilir. Korelasyon ne kadar yüksekse ölçeklerin o kadar eşdeğer olduğunu söyleriz. Ölçülen aynı şeydir. Sadece ifade tarzları ve/veya soruların tasarımı farklıdır.

Bunu yapmanın bir yolu, soruları hazırlamak ve rastgele ikiye ayırarak uygulamaktır. Paralel formlar uygulaması ölçeğin stabilliğini de ölçer.

2. Test/tekrar test (Test/retest)

Aynı araştırma aracı aynı deneklere benzer şartlar altında 2 veya daha fazla kez uygulanır. Burada iki ölçüm arasında fazla bir farklılık olmamalıdır. Tabii ki, ölçüm yapılan zaman aralığının ne kadar olduğu da önemlidir. Uygulama ne kadar erken tekrarlanırsa o kadar benzer sonuçlar elde edilir. Diğer taraftan çok geciktirilmesi ölçülen durumun değişmesine yol açabilir. 2-4 hafta gibi zaman aralıkları genel anlamda uygun kabul edilse de (9) hafıza faktörü, bireyin konuya duyarlaşması ve zaman içerisinde oluşabilecek değişiklikler de dikkate alınarak tekrar test yapılmalıdır.

3. Gözlemciler arası güvenilirlik

Uygulayıcılar arası (Interrater) güvenilirliği tek bir formun iki uygulayıcı tarafından uygulanması ve aralarındaki korelasyona bakılması ile ölçülür.

Ölçek kategorik bir ölçüm yapıyorsa (evet/hayır gibi) iki araştırmacının uygulamasında ne kadar uyum olduğuna bakılır. Aralarındaki uyum (örn. %82) rapor edilir. Ölçek nümerik bir ölçüm yapıyorsa iki araştırmacının uygulamasının ne kadar korelasyon (Intraclass Correlation Coefficient - ICC) gösterdiğine bakılır. Phi (basit korelasyon), Kappa (rastlantı açısından düzeltme yapılmış) ve Kendall's tau (sıralı veriler için) katsayıları hesaplanabilir.

4. Gözlemci içi güvenilirlik

Gözlemci içi (intra-rater) güvenilirlik aynı değerlendiricinin yaptığı birden fazla ölçümün arasındaki uyum derecesidir. Aynı gözlemcinin aynı ölçüm araç ve gereçlerini kullanarak yaptığı ölçümler birbirinden farklılık gösterebilir. Anlaşılacağı gibi bu kullanılan ölçüm araçlarına değil, araştırmacıya bağlı bir durumdur. Ölçümün numerik olduğu durumlarda sınıf içi korelasyona (intraclass correlation), kategorik olduğu durumda ise Cohen'in kappa katsayısına bakılarak değerlendirme yapılabilir.

5. Yarıya bölme (Split half)

Bir özelliği ölçmek için kullanılan tüm maddeler rastgele ikiye ayrılır. Ölçek bir grup bireye uygulanır ve her iki yarımın puanları hesaplanır. Bu iki yarımın karşılaştırılmasıyla (Guttman Split-Half katsayısı) güvenilirliğin derecesi belirlenir.

6. Cronbach alfa

Cronbach alfa yarıya bölmenin (matematiksel anlamda) eşdeğeridir. Güvenilirlik hesaplarında sıkça kullanılan bir katsayıdır. Maddeler arası korelasyon ortalamasını da dikkate alarak iç güvenilirliği hesaplar.

Cronbach alfa hesaplamasında ölçek maddeleri rastgele ikiye ayrılarak karşılaştırılır. Bu rastgele ikiye ayırma işlemi tüm ihtimaller için tekrarlanır. Benzer bir ölçüm de Kuder-Richardson'dur.

Özetleyecek olursak, güvenilirlik açısından paralel formlar ve uygulayıcılar arası güvenilirlik testin eşdeğerliliğini, yine uygulayıcılar arası güvenilirlik ve test/tekrar test testin stabilliğini, yarıya bölme, Kuder-Richardson ve Cronbach alfa gibi ölçümler ise homojenliğini belirler.

GEÇERLİLİK ÇEŞİTLERİ

Geçerlilik açısından içerik geçerliliği (Content validity), tahmin ettirici geçerlilik (Predictive validity), yapı geçerliliği (Construct validity), eş zamanlı geçerlilik (Concurrent validity) ve Görünüm geçerliliğinden (Face validity) bahsedilebilir.

1. İçerik geçerliliği

İçerik geçerliğinden anlaşılan ölçeğin içeriğinin gerçekten ölçülmesi hedeflenen durumla ilgili olup olmamasıdır. Depresyonu taramak için oluşturduğumuz bir ölçekte keyifsizlik, suçluluk hissi, intihar düşüncesi gibi maddeler bekleriz; gastrointestinal kanamayla ilgili soruların olması içerik açısından geçersiz olduğunu düşündürür.

İçeriğin boyutlarının belirlenmesini belki ölçek geliştirmedeki en zor kısımdır. Bu amaçla bir uzmanlar grubundan yararlanılabilir ve literatür desteği gerekir.

2. Tahmin ettirici (=criterion= predictive =ölçüt) geçerlilik

Araştırma aracının gerçek yaşamda durumları ne kadar tahmin ettirici olduğuyla ilgilidir. Depresyon ölçeğinde intihar riski saptananların ne kadarı intihar ediyor? Ya da trafik sınavında yüksek puan alanlar trafikte ne kadar iyi araç kullanıyor?

3. Yapı geçerliliği

Aracın ölçülmeye çalışılan teorik psiko sosyal yapı ile ne kadar korelasyon gösterdiği ile ilgilidir. "Bu ölçek ölçmeye çalıştığımız fenomeni ne kadar ölçüyor?" sorusuna cevap aranmasıdır. Altta yatan fenomenle ilgili farklı konseptleri ölçmeye çalışır. Bu amaçla madde analizi yapılabilir.

4. Eşzamanlı geçerlilik

Eş zamanlı geçerliliği test etmek için ölçek aynı veya ilişkili bir yapıyı inceleyen ve daha önce geçerliliği ispat edilmiş başka bir ölçekle eşzamanlı olarak uygulanır. Bu da tahmin ettirici geçerlilik gibi bir ölçütü tahmin etmeye ne kadar yaradığını gösterir. Yeni geliştirilen depresyon ölçeğinin Beck depresyon ölçeği ile birlikte uygulanmasını örnek olarak verebiliriz.

5. Görünüm geçerliliği

Bir arabanın hızının dış görünüşünden tahmin edilmesi gibidir. Maddelerin görünüşü, okunabilirliği, uygulama kolaylığı gibi konular açısından değerlendirme yapılır. Bu amaçla Tablo 1'deki soruların sorulması faydalı olabilir.

Tablo 1. Bir ölçeğin görünüm geçerliliğini değerlendirmek için katılımcılara sorulabilecek sorular

1	Anket hakkındaki genel görüşleriniz nelerdir?
2	Açıklamalarla ilgili görüşleriniz nelerdir?
3	Bu anketi doldurmanız ne kadar sürdü?
4	Soruların sayısı kabul edilebilir mi?
5	Soruların sırası mantıklı mı?
6	Cevap vermede zorlandığınız veya anlayamadığınız sorular oldu mu?
7	Genel olarak anketin anlaşılabilirliği ve sadeliği nasıldır?
8	Soruların derecelendirmesiyle ilgili problem yaşadınız mı, bu konuda başka bir öneriniz var mı?
9	Anketle ilgili herhangi başka bir öneriniz var mı?
10	Anketi doldurmak için yardıma ihtiyaç duydunuz mu? Kim yardım etti?

SONUÇ

"Bilim ölçüm demektir. Yanlış ölçümler yapmak kanıtlarımızın ve tedavilerimizin güvenilmez olmasına yol açar"[10] Bu nedenle ölçüm araçlarımızın geçerli ve güvenilir sonuçlar vermesine azami önem göstermeliyiz.

Nasıl bir tansiyon aletinin veya laboratuvarda kullandığımız bir cihazın kalibre edilmiş olması gerekiyorsa, psikometrik ölçümler için kullandığımız ölçeklerin de geçerli ve güvenilir olması önemlidir. Araştırmalarda kullanacağımız uluslararası ölçeklerin Türkçe geçerliliğinin olup olmamasının yanında kendi hazırladığımız sınav ve soruların da geçerli ve güvenilir olmasına dikkat etmeliyiz.

KAYNAKLAR

1. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. Fort Worth: Holt, Rinehart and Winston; 1986.
2. Meeker WQ, Escobar LA. Statistical Methods for Reliability Data. Hoboken, New Jersey: Wiley; 1998.
3. Borsboom D, Mellenbergh GJ, Heerden Jv. The concept of validity. Psychological Review 2004;111(4):1061-1.
4. Filiz TM, Topsever P, Uludağ C, Görpelioğlu S, Çınar N. Türk kadınlarında üriner inkontinans şiddeti ve yaşın jenerik yaşam kalitesi sf-36 üzerine etkileri. Turkiye Klinikleri J Med Sci 2007;27(2):189-4.
5. Paker N, Buğdaycı D, Sabırlı F, Özel S, Ersoy S. Diz incinme ve osteoartrit sonuç skoru: Türkçe sürümünün güvenilirlik ve geçerlilik çalışması. Turkiye Klinikleri J Med Sci 2007;27(3):350-6.
6. Aktürk Z, Dağdeviren N, Türe M, Tuğlu C. Birinci basamak için beck depresyon tarama ölçeği'nin türkçe çevriminin geçerlik ve güvenirliği. Türkiye Aile Hekimliği Dergisi 2005;9(3):117-2.
7. Sayar K, Bilen A, Arıkan M. Kronik ağrı hastalarında öfke, benlik saygısı ve aleksitimi. T Klin J Psychiatry 2001;2:36-42.
8. Bannigan K, Watson R. Reliability and validity in a nutshell. J Clin Nursing 2009;18(23):3237-3.
9. Carmines EG, Zeller RA. Reliability and validity assessment. California: Sage Publications; 1979.
10. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric ınstruments: Theory and application. Am J Med 2006;119(2):e7-6.