İSTATİSTİKTE

GÜNCEL KONULAR-IV

EDİTÖR

PROF. DR. ÖZLEM ALPU

EĞİTİM
yayınevi

İSTATİSTİKTE GÜNCEL KONULAR-IV

Editör: Prof. Dr. Özlem Alpu

Yayınevi Grubu Genel Başkanı: Yusuf Ziya Aydoğan (yza@egitimyayinevi.com)
Genel Yayın Yönetmeni: Yusuf Yavuz (yusufyavuz@egitimyayinevi.com)
Sayfa Tasarımı: Bilge Zümra Aydoğan
Kapak Tasarımı: Eğitim Yayınevi Tasarım Birimi

T.C. Kültür ve Turizm Bakanlığı
Yayıncı Sertifika No: 76780

E-ISBN: 978-625-5014-24-5
1. Baskı, Eylül 2026

Kütüphane Kimlik Kartı
İSTATİSTİKTE GÜNCEL KONULAR-IV
Editör: Prof. Dr. Özlem Alpu
E-ISBN: 978-625-5014-24-5
V+145 s., 160x240 mm
Kaynakça var, dizin yok.

EĞİTİM yayınevi

Yayınevi Türkiye Ofis:
Konya: Eğitim Yayınevi Tic. Ltd. Şti., Fevzi Çakmak Mah. 10721 Sok. B Blok, No: 16/B, Safakent, Karatay, Konya, Türkiye
İstanbul: Salon Yayınları, Atakent mah., Yasemen sok., No: 4/B, Ümraniye, İstanbul, Türkiye
Santral: +90 332 351 92 85
Editör hatları: +90 533 151 50 42, +90 507 151 50 43
bilgi@egitimyayinevi.com

Yayınevi Amerika Ofis: New York: Egitim Publishing Group, Inc.
P.O. Box 768/Armonk, New York, 10504-0768, United States of America
americaoffice@egitimyayinevi.com

Lojistik ve Sevkiyat Merkezi: Kitapmatik Lojistik ve Sevkiyat Merkezi, Fevzi Çakmak Mah. 10721 Sok. B Blok, No: 16/B, Safakent, Karatay, Konya, Türkiye
İnternet Satış: www.kitapmatik.com.tr
Whatsapp hattı: +90 553 950 50 37
bilgi@kitapmatik.com.tr

Kitabevi Şubesi: Eğitim Kitabevi, Şükran mah. Rampalı 121, Meram, Konya, Türkiye
Whatsapp hattı: +90 501 651 92 85
bilgi@egitimkitabevi.com

EĞİTİM YAYINEVİ GRUBU

İÇİNDEKİLER

ÖNSÖZ

Bu kitap, yüksek boyutlu verilerin ve veri odaklı teknolojilerin hızla yaygınlaşması doğrultusunda, istatistiksel yöntemlerin yalnızca doğru uygulanmasını değil, aynı zamanda yöntemlerin varsayımlarının, sınırlılıklarının ve gerçek yaşam problemlerindeki karşılıklarının anlaşılması için istatistiğin farklı disiplinlerle kesiştiği güncel ve uygulamaya dönük konuları ele alan dördüncü kitabımız olup, altı bölümden oluşmaktadır.

Kitabın ilk bölümünde, Açıklanabilir Yapay Zekâ (Explainable Artificial Intelligence, XAI) kavramsal, yöntemsel ve uygulamalı boyutlarıyla incelenmektedir. Yapay zekâ ve makine öğrenmesi modellerinin tahmin gücündeki gelişmeler, beraberinde bu modellerin kararlarının nasıl ve neden üretildiğinin anlaşılması gerekliliğini de ortaya çıkarmıştır. İstatistiksel modelleme ile yapay zekâ arasındaki giderek güçlenen ilişki dikkate alındığında, güvenilir ve sorumlu veri analitiğinin temel unsurlarından biri olarak değerlendirilen açıklanabilirlik yaklaşımları ve yaygın kullanılan teknikler ayrıntılı biçimde ele alınmış, bu yaklaşımların güçlü ve zayıf yönleri üzerinde durulmuştur.

İkinci bölümde, özellikle klinik araştırmalar ve hasta odaklı karar verme süreçleri açısından önem taşıyan Minimum Klinik Önemli Fark (MKÖF) kavramı üzerinde durulmaktadır. Bölüm, araştırmacıları MKÖF puanının belirlenmesinde kullanılan metodolojiler hakkında bilgilendirmek ve herhangi bir sonuç aracı için bu puanı etkileyebilecek çeşitli faktörleri tartışmayı amaçlamaktadır.

Üçüncü bölümde yer verilen Genellenebilirlik Teorisi, birden fazla değerlendirici, ölçüm tekniği veya uygulama koşulunun bulunduğu araştırmalarda ölçümlerin güvenirliğinin daha doğru ve kapsamlı biçimde değerlendirilmesine olanak tanımaktadır. Bölümde bu teorinin kavramsal yapısı ele alınarak, ölçümlerde farklı hata kaynaklarına değinilmiş, ölçme kuramı türleri açıklanmış, genellenebilirlik kuramının dayandığı istatistiksel modeller tek ve iki yüzeyli evrenler ve desenler bakımından incelenmiştir.

Dördüncü bölümde, $k \times k$ boyutlu eşleştirilmiş kategorik verilerin analizinde kullanılan istatistiksel yöntemler ele alınmaktadır.

Kategorik verilerin yapısına uygun yöntemlerin seçilmesi ve özellikle eşleştirilmiş gözlemler arasındaki uyum ve değişimin doğru biçimde değerlendirilmesi, uygulamalı araştırmalar açısından önemli bir metodolojik gerekliliktir. Bu gereklilik neticesinde bölümün, farklı alanlarda karşılaşılan karmaşık kategorik veri yapılarının analizine yönelik güncel bir bakış açısı sunması amaçlanmıştır.

Beşinci bölümde, tanı testi performans ölçütleri ele alınmaktadır. Duyarlılık, özgüllük ve benzeri ölçütlerin yanı sıra tanısal performansın değerlendirilmesine ilişkin temel istatistiksel yaklaşımlar, sağlık araştırmalarında elde edilen sonuçların doğru yorumlanması açısından büyük önem taşımaktadır. Tanı testlerinin yalnızca doğru sınıflandırma kapasitesinin değil, klinik karar süreçlerine sağladığı katkının da değerlendirilmesi, bu bölümün temel amaçlarından biridir.

Kitabın son bölümünde ise modern biyoinformatiğin önemli çalışma alanlarından biri olan RNA dizileme verilerinde diferansiyel gen ifadesi analizi ele alınmakta ve düzensiz gen ifade profillerinin taranmasına yönelik istatistiksel yaklaşımlar incelenmektedir. Yüksek boyutlu ve karmaşık yapıya sahip genomik verilerin analizinde istatistiksel düşüncenin rolünü ortaya koyan bu bölüm, güncel biyolojik araştırmalar ile istatistiksel yöntemler arasındaki güçlü etkileşime dikkat çekmektedir.

Bu kitabın; istatistik, sağlık bilimleri, biyoloji, psikometri, veri bilimi ve yapay zekâ gibi farklı alanlarda çalışan araştırmacılara ve konuya ilgi duyan uygulamacılara güncel bir başvuru kaynağı sunması amaçlanmaktadır.

Bu eserin ortaya çıkmasında emeği geçen tüm bölüm yazarlarına, bilimsel katkıları ve titiz çalışmaları için teşekkür eder; kitabın güncel istatistiksel yöntemlerin anlaşılmasına, farklı disiplinler arasında bilimsel iletişimin güçlenmesine ve araştırmacıların yöntemsel bakış açılarının gelişmesine katkı sağlamasını dilerim.

Prof. Dr. Özlem ALPU

AÇIKLANABİLİR YAPAY ZEKÂ: KAVRAMSAL ÇERÇEVE, YAKLAŞIMLAR VE TEKNİKLERİ

Sevgi Abdalla[1]

GİRİŞ

Yapay zekânın alt dallarından biri olan makine öğrenimi (ML), endüstri ve kamu yönetimi başta olmak üzere günlük yaşamın her alanında kullanılan bir araç haline gelmiştir. Yapay zekâ tabanlı bu sistemler, görüntü işleme, doğal dil işleme, konuşma tanıma, biyomedikal analiz, finansal tahminleme, otonom araçlar, siber güvenlik ve karar destek sistemleri gibi birçok alanda insan performansına yaklaşan, hatta bazı görevlerde insan performansını aşan sonuçlar üretmektedir. Özellikle derin öğrenmenin temeli olan sinir ağlarının karmaşık veri yapılarından örüntü çıkarabilme yeteneği, yapay zekânın gerçek yaşam problemlerinin çözümünde yaygın olarak kullanılmasını mümkün kılmıştır. Bütün bu olumlu gelişmelerle birlikte yapay zekâ modellerinin artan başarısı, yeni bir problemi doğurmuştur.

Yüksek doğruluk oranları sağlayan makine öğrenimi algoritmaları çoğunlukla milyonlarca parametre içeren karmaşık yapılardan oluşmaktadır. Özellikle derin öğrenme mimarilerinde kararın hangi değişkenlerden etkilenerek üretildiğini anlamak çoğu zaman mümkün değildir. Bu nedenle bu tür modeller literatürde kara kutu (black box) modeller olarak tanımlanmaktadır (Arrieta vd., 2020; Ali vd., 2023). Kara kutu yaklaşımında modelin girdileri ve çıktıları gözlemlenebilmekte; ancak karar verme sürecinde hangi içsel mekanizmaların rol oynadığı açıklanamamaktadır. Bu durum, yüksek doğruluk sağlayan sistemlerin güvenilirliği, şeffaflığı ve hesap verebilirliği konusunda önemli tartışmalara neden olmaktadır.

Yapay zekâ sistemlerine duyulan güvenin artırılması amacıyla geliştirilen yaklaşımların başında Açıklanabilir Yapay Zekâ (Explainable Artificial Intelligence, XAI) gelmektedir. XAI, yapay zekâ modellerinin karar mekanizmalarının insanlar tarafından

[1] Dr. Öğr. Üyesi, Eskişehir Osmangazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Eskişehir, sayhan@ogu.edu.tr, ORCID: orcid.org/0000-0003-4177-5868

anlaşılabilecek şekilde açıklanmasını amaçlayan yöntemlerin genel adıdır (Arrieta vd., 2020). Başka bir ifadeyle XAI, yapay zekâ modellerinin yalnızca "ne" karar verdiğini değil, aynı zamanda "neden" ve "nasıl" bu kararı verdiğini ortaya koymayı hedeflemektedir. Bu yaklaşım sayesinde geliştiriciler, model davranışlarını analiz edebilmekte, alanın uzmanları kararların güvenilirliğini değerlendirebilmekte ve son kullanıcılar yapay zekâ sistemlerine daha fazla güven duyabilmektedir.

XAI'nın temel amacı yüksek doğruluktan vazgeçmeden yorumlanabilirlik ile performans arasında dengeli bir yapı oluşturmaktır. Son yıllarda Avrupa Birliği Yapay Zekâ Yasası (EU AI Act), Genel Veri Koruma Tüzüğü (GDPR), OECD Yapay Zekâ İlkeleri ve NIST AI Risk Management Framework gibi uluslararası düzenlemeler; yapay zekâ sistemlerinin şeffaf (Adadi & Berrada, 2018; Guidotti vd., 2018), adil ve hesap verebilir olmasını teşvik etmektedir. Özellikle yüksek riskli yapay zekâ uygulamalarında kararların açıklanabilir olması, bireylerin otomatik kararları sorgulayabilmesi ve geliştirilen sistemlerin denetlenebilir olması temel ilkeler arasında yer almaktadır. Dolayısıyla, açıklanabilirlik kavramı yalnızca teknik bir gereklilik olarak değerlendirilmemektedir.

XAI alanındaki çalışmaların hızla artmasının temel nedenlerinden biri de kullanıcıların ihtiyaçlarının birbirinden farklı olmasıdır. Dolayısıyla, tek tip açıklamaların tüm kullanıcı grupları için yeterli olmadığı, hedef kullanıcı kitlesine göre tasarlanması gerektiği aşikardır.

Bu bölümde XAI kavramı kapsamlı bir bakış açısıyla ele alınmaktadır. İzleyen kısımlarda XAI'nın kavramsal temelleri ve literatürde kullanılan temel terminoloji açıklanacak; daha sonra açıklanabilirlik yaklaşımları ve yaygın kullanılan XAI yöntemleri ayrıntılı biçimde incelenecektir. Ek olarak, mevcut yöntemlerin güçlü ve zayıf yönleri tartışılarak açıklanabilirlik alanında karşılaşılan temel zorluklar değerlendirilecek ve gelecekteki araştırma eğilimleri üzerinde durulacaktır. Böylece bölümün hem yapay zekâ araştırmacıları hem de XAI yöntemlerini kendi çalışma alanlarında uygulamak isteyen araştırmacılar için bütüncül bir kaynak niteliği taşıması amaçlanmaktadır.

AÇIKLANABİLİR YAPAY ZEKÂNIN KAVRAMSAL ÇERÇEVESİ

Açıklanabilir Yapay Zekâ (XAI), en genel anlamıyla yapay zekâ sistemlerinin karar verme süreçlerini insanların anlayabileceği biçimde açıklayabilmesini sağlayan yöntem ve tekniklerin bütününü ifade etmektedir (Arrieta vd., 2020). Bununla birlikte literatürde XAI için tek ve evrensel kabul görmüş bir tanım bulunmamaktadır (Adadi & Berrada, 2018; Guidotti vd., 2018). Literatürde yaygın kabul gören tanımlardan birine göre XAI, belirli bir hedef kullanıcı için yapay zekâ sisteminin çalışma prensibini anlaşılır hâle getiren açıklanabilir çıktı üretebilme yeteneğidir (Ali vd., 2023). Bu tanımın dikkat çekici yönü, açıklamanın yalnızca teknik olarak doğru olmasını değil, aynı zamanda uzmanın bilgi düzeyine uygun olmasını da gerekli görmesidir (Miller, 2019). Dolayısıyla açıklanabilirlik, yalnızca algoritmanın özelliklerinden değil, açıklamayı değerlendiren ve yorumlayan kişinin bilgi düzeyi, beklentileri ve kullanım amacı gibi faktörlerden de etkilenmektedir (Rong vd., 2024).

Açıklanabilirlik ve Yorumlanabilirlik Kavramları

XAI literatüründe en sık karşılaşılan kavramlardan ikisi açıklanabilirlik (explainability) ve yorumlanabilirlik (interpretability) kavramlarıdır. "**Yorumlanabilirlik",** modelin iç yapısının insanlar tarafından doğrudan anlaşılabilmesini ifade etmektedir. Başka bir ifadeyle yorumlanabilir modeller, herhangi bir ek açıklama algoritmasına ihtiyaç duymadan incelenebilen modellerdir. Karar ağaçları, doğrusal regresyon ve lojistik regresyon gibi klasik makine öğrenmesi algoritmaları bu gruba örnek olarak verilebilir. Bu modellerin karar mekanizması model parametreleri incelenerek doğrudan anlaşılabilmektedir. Buna karşılık "**açıklanabilirlik**" özellikle kara kutu modeller tarafından üretilen kararların sonradan geliştirilen yöntemlerle açıklanmasını ifade etmektedir. Burada amaç, modelin tüm matematiksel yapısını açıklamak yerine, belirli bir kararın hangi değişkenlerden etkilendiği veya modelin hangi örüntülere dayanarak tahminde bulunduğunu kullanıcıya anlaşılır biçimde sunmaktır.

Tablo 1. Açıklanabilirlik ve Yorumlanabilirlik Kavramlarının Karşılaştırılması

Özellik	Açıklanabilirlik	Yorumlanabilirlik
Temel Amaç	Modelin Kararını açıklamak	Modeli yorumlamak
Model Tipi	Kara-kutu	Beyaz-kutu
Açıklama Zamanı	Model eğitildikten sonra	Model tasarımında
Ek Algoritma Gereksinimi	Var	Yok
Örnek Modeller	SHAP, LIME, Grad-CAM	Karar Ağacı Doğrusal Regresyon

SHAP, LIME, Integrated Gradients ve Grad-CAM gibi yöntemler bu yaklaşımın en yaygın örneklerindendir. Bu iki kavram arasındaki temel fark Tablo 1'de özetlenmiştir. Yorumlanabilirlik, modelin yapısal özellikleriyle ilişkiliyken, açıklanabilirlik daha çok model davranışlarının kullanıcıya aktarılmasına odaklanır. Günümüzde yüksek doğruluk gerektiren problemlerde çoğunlukla kara kutu modeller kullanıldığından XAI araştırmalarının büyük bölümü yorumlanabilirlik kavramı üzerinde yoğunlaşmaktadır (Adadi & Berrada, 2018; Guidotti vd., 2018).

XAI ile İlişkili Temel Kavramlar

Açıklanabilir yapay zekâ yalnızca model kararlarının açıklanmasını hedefleyen bağımsız bir araştırma alanı değildir. Günümüzde XAI; güvenilir yapay zekâ (Trustworthy AI), sorumlu yapay zekâ (Responsible AI), şeffaflık (Transparency), adalet (Fairness), sağlamlık (Robustness), hesap verebilirlik (Accountability) ve etik yapay zekâ (Ethical AI) gibi birçok kavramla birlikte değerlendirilmektedir. Bu kavramların tamamı, yapay zekâ sistemlerinin toplum tarafından güvenle kullanılabilmesini sağlayan temel ilkeleri oluşturmaktadır. (Arrieta vd., 2020; Ali vd., 2023)

Şeffaflık (Transparency), modelin karar mekanizmasının mümkün olduğunca görünür ve anlaşılabilir olmasıdır. Şeffaf bir modelde kullanıcı yalnızca son tahmini değil, bu tahmine ulaşılırken izlenen süreci de inceleyebilmektedir. Özellikle doğası gereği yorumlanabilir

modeller yüksek şeffaflık sağlamaktadır. Bununla birlikte karmaşık derin öğrenme modellerinde tam şeffaflık çoğu zaman mümkün olmadığından, açıklama algoritmaları bu eksikliği gidermeye çalışmaktadır. (Lipton, 2018; Molnar, 2022)

Adalet (Fairness): Adalet ilkesi, modelin herhangi bir demografik grubu sistematik olarak dezavantajlı duruma düşürmemesini ifade etmektedir. Günümüzde kredi değerlendirme, işe alım ve suç tahmini gibi uygulamalarda açıklanabilirlik ile adalet birlikte değerlendirilmektedir. Açıklanabilirlik yöntemleri, modelin hangi değişkenleri kullandığını göstererek potansiyel önyargıların belirlenmesine katkı sağlamaktadır. (Doshi-Velez & Kim, 2017)

Sağlamlık (Robustness), küçük veri değişiklikleri karşısında modelin benzer kararlar verebilme yeteneğini ifade etmektedir. Bir model yalnızca doğru tahmin üretmekle kalmamalı; aynı zamanda küçük gürültülerden veya kasıtlı değişikliklerden minimum düzeyde etkilenmelidir. XAI yöntemleri modelin hangi özelliklere duyarlı olduğunu göstererek sağlamlığın değerlendirilmesine de katkı sunmaktadır.

Güvenilir Yapay Zekâ (Trustworthy AI), Literatürde doğruluk, açıklanabilirlik, tarafsızlık, güvenlik, sağlamlık ve hesap verebilirlik gibi birçok ilkenin birlikte sağlandığı yapay zekâ sistemlerini ifade etmektedir. Açıklanabilirlik bu yapının merkezinde yer almakta ve diğer ilkelerle doğrudan ilişkilendirilmektedir. Nitekim açıklanamayan bir modelin güvenilir olduğunun kabul edilmesi oldukça güçtür (Adadi & Berrada, 2018; Guidotti vd., 2018).

AÇIKLANABILIR YAPAY ZEKÂ YAKLAŞIMLARI

XAI alanında geliştirilen yöntemler, model türlerine ve açıklanma gereksinimlerine göre çeşitlilik göstermektedir. Örneğin, doğrusal regresyon modeli doğası gereği yorumlanabilirken, milyonlarca parametre içeren bir derin öğrenme sinir ağı için aynı durum söz konusu değildir. Benzer şekilde, bir veri bilimcisi modelin genel davranışını anlamak isterken, bir hekim yalnızca bir hastaya özel teşhise ilişkin kararın nedenini öğrenmek isteyebilir. XAI yöntemleri; (i) doğası gereği açıklanabilirlik (ante-hoc/intrinsic) ve sonradan açıklanabilirlik (post-hoc) yöntemleri, (ii) model-bağımlı ve model-

bağımsız yöntemler ve (iii) yerel (local) ve global açıklanabilirlik yöntemleri olmak üzere üç kritere göre sınıflandırılmaktadır.

Doğası Gereği (Ante-hoc / Intrinsic) ve Sonradan (Post-hoc) Açıklanabilir Yöntemler

Doğası gereği (Ante-hoc) açıklanabilirlik, herhangi bir ek açıklama yöntemine ihtiyaç duymadan probleme ilişkin kararın, doğrudan modelden elde edildiği veya incelenebildiği modeller için kullanılan bir tanımdır. Bu modellerde kararın hangi değişkenlerden etkilendiği, model parametreleri analiz edilerek kolaylıkla anlaşılabilmektedir. Bu nedenle literatürde bu modeller, white-box (beyaz kutu) veya bazı durumlarda glass-box (şeffaf kutu) modeller olarak da adlandırılmaktadır. Bu grupta yer alan yöntemlere doğrusal Regresyon (Linear Regression), lojistik Regresyon (Logistic Regression), karar ağaçları (Decision Trees), kural tabanlı sistemler (RuleBased Systems) ve bazı Bayes ağları örnek olarak verilebilir. Bu modellerde her değişkenin çıktıya etkisi doğrudan gözlemlenebildiği için karar sürecinin açıklanması oldukça kolaydır. Özellikle sağlık, finans ve kamu yönetimi gibi alanlarda yorumlanabilir modeller tercih edilmektedir. Ancak bu yöntemlerin performansları yüksek boyutlu ve doğrusal olmayan problemlerde sınırlıdır. Günümüzde görüntü işleme, doğal dil işleme ve büyük veri analitiği gibi karmaşık problemlerde derin öğrenme modellerinin ulaştığı doğruluk düzeyine çoğu zaman erişememektedir. Bu nedenle, araştırmacılar (Lipton, 2018; Molnar, 2022) yüksek doğruluk ile açıklanabilirlik arasında denge kurabilecek yeni yöntemler geliştirmeye yönelmiştir. Bu yöntemlere de sonradan açıklanabilir yöntemler yani post-hoc yöntemler denilmektedir.

Sonradan (Post-hoc) açıklanabilirlik, model eğitildikten sonra geliştirilen ek yöntemlerle modele bağlı kararların açıklanmasını ifade etmektedir. Bu yaklaşım günümüzde XAI araştırmalarının en yoğun çalışılan alanını oluşturmaktadır. Çünkü modern yapay zekâ uygulamalarının büyük bölümü derin öğrenme, topluluk öğrenmesi (ensemble learning) veya gradyan artırmalı karar ağaçları gibi karmaşık kara kutu modeller üzerine kuruludur. Post-hoc yöntemlerde modelin iç yapısı değiştirilmez. Bunun yerine modelin sağladığı çıktılar analiz edilerek kullanıcıya anlaşılır açıklamalar sunulur. Böylece modelin yüksek doğruluk performansı korunurken karara

ilişkin ek bilgiler elde edilir. Post-hoc yöntemlerin amaçları en etkili değişkenleri belirlemek, modelin hangi örüntüleri öğrendiğini göstermek, hatalı öğrenmeleri ortaya çıkarmak, kullanıcı güvenini artırmak, model geliştirme sürecine geri bildirim sağlamak olarak sıralanabilir (Adadi & Berrada, 2018; Guidotti vd., 2018).

Günümüzde en yaygın kullanılan post-hoc yöntemler; LIME, SHAP, GradCAM, Integrated Gradients, DeepLIFT, Layer-wise Relevance Propagation (LRP), karşı olgusal açıklamalar ve Feature Importance gibi tekniklerdir. Bu yöntemlerden en önemli dört tanesi bir sonraki bölümde ayrıntılı olarak ele alınmaktadır (Ribeiro vd., 2016).

Model Bağımlı ve Model Bağımsız Yöntemler

Model bağımlı yöntemler, yalnızca belirli model aileleri için geliştirilen, modelin mimarisine ve iç parametrelerine erişim gerektiren XAI teknikleridir. Örneğin; Grad-CAM yönteminin yalnızca CNN modellerinde kullanılması durumu model bağımlı yönteme örnektir (Selvaraju vd., 2017). Model-bağımsız (model agnostic) yöntemler ise modelin iç yapısına erişim gerektirmez. Model yalnızca bir fonksiyon olarak değerlendirilir; girdiler değiştirilerek çıktılardaki değişim analiz edilir. Bu gruba, LIME, SHAP, PDP, özellik önemi teknikleri örnek olarak verilebilir (Ribeiro vd., 2016). Model-bağımsız yöntemlerin en önemli avantajı, aynı açıklama yönteminin farklı algoritmalarda kullanılabilmesidir. Bu nedenle günümüzde birçok ticari XAI platformu model-bağımsız açıklama tekniklerini tercih etmektedir.

Local ve Global Açıklanabilirlik Yöntemleri

XAI literatüründe önemli olarak kabul edilen bu ayrım, açıklamanın kapsamına ilişkindir. Açıklamalar belirli bir tahmine odaklanabileceği gibi modelin genel davranışını da açıklayabilir. Bu doğrultuda yöntemler yerel ve global açıklamalar olarak iki gruba ayrılmaktadır.

Yerel açıklamalar yalnızca tek bir tahminin neden üretildiğini ortaya koymayı amaçlamaktadır. Örneğin, bir hastaya kanser tanısı konulmasının hangi değişkenlerden kaynaklandığı veya bir kredi başvurusunun neden reddedildiği gibi sorular yerel açıklamalarla yanıtlanmaktadır. Yerel açıklamalar özellikle karar destek sistemlerinde kullanıcı güveninin artırılması açısından önemlidir. Son kullanıcılar çoğu zaman modelin tamamının nasıl çalıştığını değil,

yalnızca kendi durumlarıyla ilgili kararın gerekçesini öğrenmek istemektedir. Yerel açıklamalarda modelin tamamı açıklanmaz; yalnızca ilgili gözlem üzerinde etkili olan değişkenler analiz edilir. LIME ve SHAP'in yerel açıklama modları bu yaklaşımın en bilinen örnekleridir. Yerel açıklamalar özellikle klinik karar destek sistemleri, kredi değerlendirme sistemleri ve adli bilişim uygulamalarında yaygın olarak kullanılmaktadır. (Ribeiro vd., 2016)

Global açıklanabilirlik ise modelin genel davranışını anlamaya yöneliktir. Amaç, modelin tüm veri kümesi üzerinde nasıl karar verdiğini ve hangi değişkenlerin sistematik olarak daha etkili olduğunu belirlemektir. Daha çok veri bilimciler ve model geliştiriciler tarafından kullanılmaktadır. Özellik önem sıralamaları (Feature Importance), SHAP, Summary Plot ve Partial Dependence Plot gibi yöntemler bu grupta değerlendirilmektedir. Yerel ve global açıklamalar birbirinin alternatifi değil, tamamlayıcısıdır.

Yerel açıklamalar bireysel kararların gerekçelerini ortaya koyarken, global açıklamalar model davranışının genel karakteristiklerini göstermektedir (Lundberg & Lee, 2017).

YAYGIN KULLANILAN XAI TEKNİKLERİ

Yüksek performanslı makine öğrenimi ve derin öğrenme modellerinin yaygınlaşmasıyla birlikte, bu modellerin karar süreçlerinin anlaşılmasını sağlayacak çok sayıda açıklanabilirlik yöntemi geliştirilmiştir. Özellikle son on yılda önerilen yöntemlerin büyük bölümü, model doğruluğunu korurken kullanıcıların model davranışlarını anlamalarını kolaylaştırmaktadır. Bu yöntemler, açıklamanın kapsamına, model türüne, matematiksel altyapısına ve kullanım amacına göre farklılık göstermektedir. Bununla birlikte ortak amaçları, yapay zekâ modellerinin karar süreçlerini daha şeffaf, güvenilir ve yorumlanabilir hâle getirmektir.

Literatürde çok sayıda XAI yöntemi önerilmiş olmasına rağmen, uygulamada en yaygın kullanılan teknikler LIME, SHAP, Grad-CAM ve Integrated Gradients (bütünleşik gradyanlar)dır. Bu yöntemler farklı problem türleri için geliştirilmiş olmakla birlikte birbirlerini tamamlayıcı özelliklere de sahiptir (Ribeiro vd., 2016).

LIME (Local Interpretable Model-agnostic)

LIME, Ribeiro, Singh ve Guestrin (2016) tarafından geliştirilen ve günümüzde en yaygın kullanılan model-bağımsız açıklama yöntemlerinden biridir. Temel amacı, karmaşık bir makine öğrenimi modelinin yalnızca belirli bir tahmine ilişkin kararını açıklamaktır. LIME'ın temel varsayımı, global olarak karmaşık davranan bir modelin belirli bir gözlemin çevresinde daha basit bir yapıyla temsil edilebileceğidir. Bu nedenle yöntem, açıklanmak istenen örneğin çevresinde yeni sentetik örnekler üretmekte, bu örnekleri orijinal modele göndererek tahminler elde etmekte ve daha sonra bu yerel veri kümesi üzerinde doğrusal bir vekil model (surrogate model) oluşturmaktadır. Böylece karmaşık model yerine, yerel olarak yorumlanabilir bir model kullanılarak hangi değişkenlerin kararı etkilediği belirlenmektedir. Bu yaklaşım özellikle bireysel kararların açıklanmasında önemli avantajlar sunmaktadır. LIME'ın en önemli avantajları model bağımsız çalışabilmesi, hemen hemen tüm makine öğrenimi algoritmalarına uygulanabilmesi, hesaplama maliyetinin görece düşük olması, kullanıcı dostu görselleştirmeler sunabilmesidir (Ribeiro vd., 2016). Bununla birlikte yöntemin en önemli sınırlılığı, farklı örnekleme süreçlerinde farklı açıklamalar üretebilmesi nedeniyle kararlılık (stability) sorunudur. Ayrıca yalnızca yerel açıklamalar üretmesi, modelin genel davranışı hakkında bilgi vermemektedir. Ek olarak özellikler arasındaki ilişkileri her zaman doğru temsil edememektedir.

Matematiksel Temel

LIME'ın temel varsayımı, global olarak oldukça karmaşık davranan bir modelin belirli bir gözlem çevresinde daha basit ve yorumlanabilir

Bu yaklaşım optimizasyon problemi

$$\xi(x) = arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (1)$$

Eşitlik 1'deki fonksiyonla temsil edilmektedir.

Tablo 2. LIME Yönteminin Bileşenleri

Sembol	Açıklama
$f(x)$	Açıklanmak istenen kara kutu model
$g(x)$	Yerel vekil (surrogate) model
G	Yorumlanabilir modeller kümesi
L	Gerçek model ile vekil model arasındaki hata
(π_x)	Gözleme olan uzaklığa göre ağırlık fonksiyonu
$\Omega(g)$	Model karmaşıklığını gösteren ceza terimi

Açıklanmak istenen x noktasına yakın olan türetilmiş z örneklerine daha fazla ağırlık vermek için genellikle bir üstel çekirdek fonksiyonu (Exponential Kernel) kullanılır (Bknz. Eşitlik 2):

$$\pi_x(z) = exp\left(-\frac{D(x,z)^2}{\sigma^2}\right) \quad (2)$$

Burada:

D(x, z), orijinal x noktası ile türetilen z noktası arasındaki uzaklık fonksiyonudur (Örn: Sürekli veriler için Öklid mesafesi veya metinler için Cosine mesafesi).

σ, çekirdeğin genişliğini (bant genişliğini) kontrol eden parametredir.

LIME, yerel modeli eğitirken genellikle ağırlıklı en küçük kareler yöntemiyle eğitilen doğrusal bir vekil model kullanır.

z' türetilmiş örneklerin ikili gösterimi olmak üzere kayıp fonksiyonu açık haliyle Eşitlik 3'teki gibi yazılır:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in Z} \pi_x(z)(f(z) - g(z'))^2 \quad (3)$$

Burada, $\mathcal{Z}$ açıklanacak x noktasının etrafına küçük rastgele gürültüler (perturbations) eklenerek oluşturulmuş yeni veri setidir.

Eğer yorumlanabilir model sınıfı (G) olarak doğrusal regresyon seçilirse, g(z') fonksiyonu Eşitlik 4'te ifade edilmiştir:

$$g(z') = w_0 + w_1 z'_1 + w_2 z'_2 + \cdots + w_d z'_d = w_0 + \sum_{i=1}^{d} w_i z'_i \quad (4)$$

Buradaki w_i katsayıları (ağırlıkları), ilgili özniteliğin yerel olarak tahmine olan etkisinin yönünü ve büyüklüğünü (LIME skorunu) verir.

Bu denklemde amaç, gerçek modele mümkün olduğunca yakın sonuçlar üreten ancak aynı zamanda yorumlanabilirliği yüksek olan bir vekil model oluşturmaktır.

Algoritmanın Çalışma Prensibi

LIME algoritması temel olarak beş adımdan oluşmaktadır.

Adım 1: Açıklanmak istenen gözlem seçilir.

Adım 2: Seçilen gözlemin çevresinde çok sayıda sentetik örnek oluşturulur.

Adım 3: Bu örnekler kara kutu modele gönderilerek yeni tahminler elde edilir.

Adım 4: Gözleme yakın örneklere daha yüksek ağırlık değeri atanır.
Adım 5: Ağırlıklı doğrusal regresyon modeli oluşturularak özellik önemleri hesaplanır.

Bu süreç sonunda elde edilen regresyon katsayıları, modelin ilgili tahminini oluştururken hangi değişkenlere daha fazla önem verdiğini göstermektedir.

LIME için Örnek Uygulama

Bir kredi değerlendirme modelinin müşterinin başvurusunu reddettiği varsayılsın.

LIME analizi sonucunda Tablo 3'teki gibi bir çıktı elde edilmiş olsun.

Tablo 3. LIME Yöntemi Örnek Uygulama Çıktıları

Özellik	Katkı
Düşük gelir	−0.42
Yüksek kredi borcu	−0.31
Düzenli ödeme geçmişi	+0.18
Uzun çalışma süresi	+0.09

Bu çıktı değerlerine göre, modelin kredi kararını en fazla düşük gelir düzeyi ve mevcut borç miktarına göre verdiğini göstermektedir.

SHAP (SHapley Additive Explanations)

SHAP, Lundberg ve Lee (2017) tarafından geliştirilen ve günümüzde en güvenilir açıklama yöntemlerinden biri olarak kabul edilen model-bağımsız açıklama yaklaşımıdır. Yöntemin matematiksel temeli oyun teorisinde kullanılan Shapley değerlerine dayanmaktadır. Shapley değeri, her oyuncunun ortak başarıya yaptığı katkıyı adil biçimde hesaplamak amacıyla geliştirilmiştir. SHAP bu yaklaşımı makine öğrenimine uyarlayarak her özelliğin model tahminine yaptığı katkıyı hesaplamaktadır (Lundberg & Lee, 2017). Örneğin bir kredi değerlendirme modelinde gelir düzeyi, kredi geçmişi, yaş, mevcut borç miktarı gibi değişkenlerin nihai kredi kararına ne kadar katkı yaptığı ayrı ayrı hesaplanabilmektedir. Güçlü matematiksel temele sahip olması, yerel ve global açıklamalar sunabilmesi, özellik katkılarının tutarlı biçimde hesaplanması, farklı modeller arasında karşılaştırma yapılabilmesi. SHAP'ın önemli üstünlükleri arasındadır. Ancak özellikle çok büyük veri kümelerinde ve derin öğrenme modellerinde hesaplama maliyeti önemli ölçüde artmaktadır. Bu nedenle TreeSHAP, KernelSHAP ve DeepSHAP gibi optimize edilmiş sürümler geliştirilmiştir (Lundberg & Lee, 2017).

Matematiksel Temel

Bir özelliğin katkısı aşağıdaki Eşitlik 5 ile hesaplanmaktadır.

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right] \quad (5)$$

Burada N toplam özellik sayısını, S ise i özelliğini içermeyen alt kümeleri ifade eder.

Tablo 4. SHAP Yönteminin Bileşenleri

Sembol	Açıklama
N	Özellik kümesi
S	i özelliği dışındaki özellik alt kümesi
$f_{(S)}(x)$	Alt küme ile elde edilen model çıktısı
$f_{(S \cup i)}(x)$	$İ$ özelliği eklendiğinde oluşan çıktı
ϕ_i	Özelliğin Shapley katkısı

Toplam model çıktısı ise;

$$f(x) = \phi_0 + \sum_{i=1}^{N} \phi_i \quad (6)$$

Fonksiyonu ile elde edilmektedir. Burada; ϕ_0 modelin temel tahminini ve ϕ_i her özelliğin katkısını göstermektedir. SHAP yönteminin çalışma prensibi izleyen kısımda açıklanmıştır.

Adım 1: Model tahmini oluşturulur.

Adım 2: Özelliklerin tüm olası kombinasyonları değerlendirilir.

Adım 3: Her özelliğin modele yaptığı marjinal katkı hesaplanır.

Adım 4: Katkılar ağırlıklandırılır.

Adım 5: Her özellik için nihai Shapley değeri hesaplanır.

Bu süreç sonunda her değişken için pozitif veya negatif katkı değeri elde edilmektedir.

SHAP için Örnek Uygulama

Bir diyabet risk tahmin modelinde SHAP analizi sonucunda aşağıdaki katkılar elde edilmiş olsun.

Tablo 5. SHAP Yöntemi için Örnek Uygulama Çıktıları

Özellik	SHAP Değeri
Glukoz düzeyi	+0.71
VKİ	+0.43
Yaş	+0.18
Fiziksel aktivite	−0.12
HDL kolesterol	−0.21

Bu sonuçlara göre glukoz düzeyi ve vücut kitle indeksinin model tahminini artırdığı, fiziksel aktivite ve HDL kolesterol düzeyinin ise riski azalttığı yorumlanabilir.

Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM (Gradient-weighted Class Activation Mapping), Selvaraju ve arkadaşları (2017) tarafından geliştirilen, özellikle Evrişimsel Sinir Ağları (Convolutional Neural Networks-CNN) için tasarlanmış gradyan tabanlı bir açıklanabilir yapay zekâ yöntemidir. Yöntemin temel amacı, modelin belirli bir sınıfa ilişkin tahmin üretirken giriş görüntüsünün hangi bölgelerine odaklandığını görselleştirmektir.

Grad-CAM, son evrişim katmanındaki özellik haritalarını (feature maps) ve bunlara ait gradyan bilgilerini kullanarak bir sınıf aktivasyon ısı haritası oluşturur. Elde edilen ısı haritası (heatmap), modelin kararında en etkili olan görüntü bölgelerini göstermektedir. Bu özelliği nedeniyle Grad-CAM, özellikle tıbbi görüntü analizi, nesne tanıma ve otonom sürüş sistemlerinde yaygın olarak kullanılmaktadır (Selvaraju vd., 2017).

Matematiksel Temel

Grad-CAM'in temel varsayımı, belirli bir sınıfa ait çıktı skorunun, son evrişim katmanındaki özellik haritalarıyla güçlü bir ilişki içinde olduğudur. Bu nedenle yöntem, ilgili sınıfa ait skorun son evrişim katmanındaki aktivasyonlara göre türevini hesaplayarak her özellik haritasının sınıflandırmaya katkısını belirlemektedir.

CNN modelinde son evrişim katmanında Ak ile temsil edilen k. özellik haritası bulunsun.

Modelin c sınıfına ait skoru yc ile temsil edilmektedir. İlk olarak sınıf skorunun özellik haritalarına göre gradyanı hesaplanmaktadır.

$$\frac{\partial y^c}{\partial A_{ij}^k}$$

Özellik Haritalarının Ağırlıklarının Hesaplanması

Grad-CAM'de her özellik haritası için tek bir önem katsayısı hesaplanmaktadır. Bu katsayı $\alpha_k{}^c$ ile gösterilmektedir ve eşitlik 7'de verilmiştir.

$$\alpha_k^c = \frac{1}{z}\sum\sum\frac{\partial y^c}{\partial A_{ij}^k} \qquad (7)$$

Global Ortalama Havuzlama kullanılarak hesaplanmaktadır. Burada Z, özellik haritasındaki toplam piksel sayısını göstermektedir. Bu denklem aslında ilgili özellik haritasının sınıflandırmaya yaptığı ortalama katkıyı hesaplamaktadır. Eğer $\alpha_k{}^c$ büyükse, ilgili özellik haritası model kararında önemli rol oynamaktadır

Aktivasyon Haritasının Oluşturulması

Özellik haritalarının ağırlıkları hesaplandıktan sonra nihai sınıf aktivasyon haritası oluşturulmaktadır. Eşitlik 8'de verilen fonksiyon yöntemin temel fonksiyonudur.

$$L_{Grad-Ca}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \qquad (8)$$

ReLU(·) ise yalnızca pozitif katkı yapan bölgelerin seçilmesini sağlamaktadır.

$$Relu(x) = \max(0,\ x) \qquad (9)$$

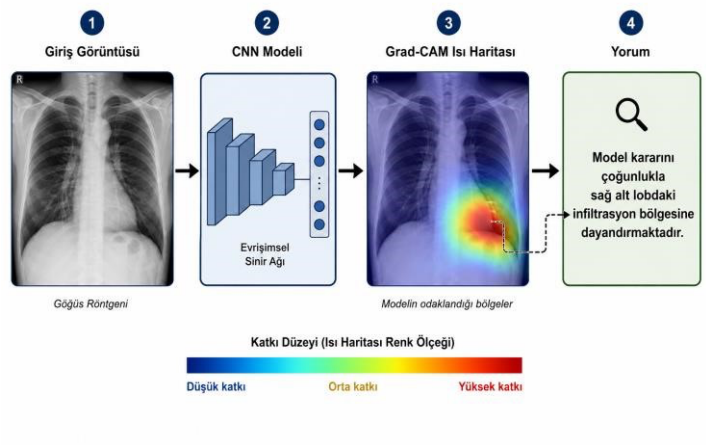
şeklinde tanımlanmaktadır. Bu adım sonucunda negatif katkılar sıfırlanmakta yalnızca hedef sınıfı destekleyen bölgeler görüntülenmektedir. Grad-CAM'in temel fikri, bir sınıfa ilişkin kararın hangi özellik haritaları tarafından üretildiğini gradyanlar yardımıyla belirlemektir. Gradyan büyükse, ilgili özellik haritasındaki değişim çıktıyı önemli ölçüde değiştirmektedir. Dolayısıyla yüksek gradyana

sahip özellik haritaları için daha önemlidir. Global Average Pooling sayesinde her özellik haritası tek bir katsayı ile temsil edilmektedir. Bu katsayı, ilgili özellik haritasının sınıfa yaptığı katkıyı göstermektedir.

Grad-Cam için Örnek Uygulama

Şekil 1'de Grad-CAM yönteminin bir göğüs röntgeni üzerindeki temel işleyişi örneklenmektedir. İlk aşamada giriş görüntüsü evrişimsel sinir ağına (CNN) aktarılmakta, ardından modelin hedef sınıfa ilişkin kararını destekleyen uzamsal bölgeler gradyan bilgisi kullanılarak belirlenmektedir. Oluşturulan ısı haritasında mavi tonlar düşük, sarı tonlar orta ve kırmızı tonlar yüksek katkıyı temsil etmektedir.

Bu örnekte, bir akciğer grafisinin CNN modeli tarafından **Pnömoni** olarak sınıflandırıldığı varsayılsın. Grad-CAM analizi sonucunda oluşturulan ısı haritasında sağ alt akciğer lobunda yoğun kırmızı bölgeler gözlenmiştir. Bu durum, modelin kararını verirken görüntünün özellikle bu anatomik bölgesine odaklandığını göstermektedir. Eğer ısı haritası akciğer dışındaki alanlarda yoğunlaşmış olsaydı, bu durum modelin klinik açıdan anlamsız örüntüler öğrenmiş olabileceğini ifade edecekti. Bu nedenle Grad-CAM yalnızca açıklama üretmek için değil, aynı zamanda model doğrulama ve hata analizi amacıyla da kullanılmaktadır.



Şekil 1. Grad-Cam için Örnek Uygulama Modeli

Integrated Gradients (Bütünleşik Gradyanlar)

Integrated Gradients (IG), Sundararajan, Taly ve Yan (2017) tarafından geliştirilen, derin öğrenme modellerinin karar mekanizmasını açıklamak amacıyla kullanılan gradyan tabanlı bir açıklanabilir yapay zekâ yöntemidir. Yöntemin temel amacı, bir modelin tahminine her giriş özelliğinin ne ölçüde katkı sağladığını nicel olarak belirlemektir. Klasik gradyan tabanlı yöntemlerden farklı olarak Integrated Gradients, yalnızca gerçek girdide hesaplanan gradyanları kullanmak yerine, referans (baseline) bir girdiden gerçek girdiye kadar uzanan doğrusal yol boyunca hesaplanan gradyanların integralini alır. Böylece özellikle derin sinir ağlarında karşılaşılan gradyan doygunluğu (gradient saturation) problemi büyük ölçüde azaltılmakta ve daha güvenilir özellik katkıları elde edilmektedir (Sundararajan vd., 2017).

Matematiksel Temel

Integrated Gradients yöntemi, modeli $F:R^n \rightarrow R$ fonksiyonu ile ifade etmektedir. Burada $x=(x_1,x_2,\ldots,x_n)$ gerçek girdi, $x'=(x_1',x_2',\ldots,x_n')$ ise referans girdiyi göstermektedir. Bir girdi özelliğinin modele yaptığı katkı aşağıdaki Eşitlik 10'daki fonksiyon yardımıyla hesaplanmaktadır.

$$IG_i(x) = (x_i - x_i') \int_0^1 [\partial F(x' + \alpha(x - x'))/\partial x_i] d\alpha \quad (10)$$

Burada;

F(x): Açıklanmak istenen derin öğrenme modeli

x : Gerçek girdi verisi

x′: Referans (baseline) girdi

α: Referans ile gerçek girdi arasında oluşturulan ara noktaları belirleyen katsayı

$IG_i(x)$: i. özelliğin model çıktısına yaptığı katkıdır.

Bu yaklaşım, referans girdiden gerçek girdiye kadar olan doğrusal yol boyunca modelin gradyan bilgisini bütünleştirerek her özelliğin karar üzerindeki toplam etkisini hesaplamaktadır. Uygulamada integral işlemi doğrudan çözülemediği için belirli sayıda ara nokta oluşturularak Riemann toplamı ile yaklaşık olarak hesaplanmaktadır.

IG yönteminin önemli özelliklerinden biri tamlık (Completeness) ilkesini sağlamasıdır. Bu ilkeye göre tüm özellik katkılarının toplamı, modelin gerçek girdi ile referans girdi arasındaki çıktı farkına eşittir.

Bu özellik, hesaplanan katkı değerlerinin model çıktısını eksiksiz biçimde temsil ettiğini göstermektedir.

Algoritmanın Çalışma Prensibi

Adım 1. Açıklanacak veri örneği belirlenir.

Adım 2. Problemi temsil eden uygun bir referans (baseline) girdi seçilir.

Adım 3. Referans ile gerçek girdi arasında doğrusal ara örnekler oluşturulur.

Adım 4. Her ara örnekte modelin gradyan bilgisi hesaplanır.

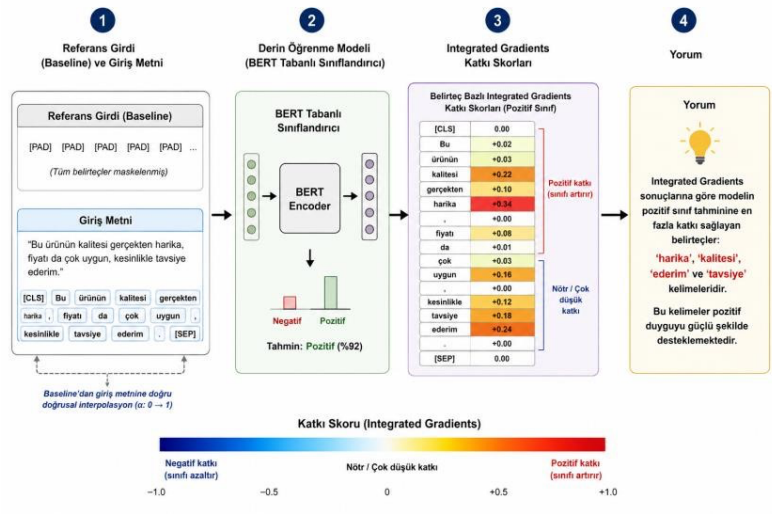
Adım 5. Gradyanların integrali alınarak her girdi özelliğinin katkısı belirlenir.

Bu süreç sonunda her girdi değişkeni için pozitif veya negatif katkı puanı elde edilmektedir. Pozitif değerler ilgili özelliğin model tahminini artırdığını, negatif değerler ise tahmini azalttığını göstermektedir.

IG için Örnek Uygulama

Şekil 2'de IG yönteminin duygu analizi problemindeki uygulaması gösterilmektedir. İlk aşamada, tüm belirteçlerin maskelendiği referans girdi (baseline) ile gerçek metin tanımlanmakta ve bu iki girdi arasında doğrusal enterpolasyon oluşturulmaktadır. Ardından girdi verisi BERT tabanlı sınıflandırıcı tarafından analiz edilerek metnin pozitif sınıfa ait olduğu tahmin edilmektedir. Üçüncü aşamada IG yöntemi kullanılarak her kelimenin sınıflandırma sonucuna yaptığı katkı hesaplanmaktadır. Katkı skorları incelendiğinde "harika", "kalitesi", "ederim" ve "tavsiye" kelimelerinin pozitif sınıf tahminine en yüksek katkıyı sağladığı görülmektedir. Buna karşılık katkı değeri düşük olan kelimelerin model kararına etkisinin olmadığı anlaşılmaktadır. Bu örnek, özellikle doğal dil işleme uygulamalarında bu tür açıklamalar, modelin anlamsal olarak doğru ipuçlarına odaklanıp odaklanmadığının

değerlendirilmesine ve karar süreçlerinin daha şeffaf hâle getirilmesine önemli katkı sağlamaktadır.



Şekil 2. IG Yöntemi için Örnek Uygulama Modeli

SONUÇ VE ÖNERİLER

Bu bölümde XAI kavramı ile birlikte LIME, SHAP, Integrated Gradients ve Grad-CAM yöntemleri incelenmiş; bu yöntemlerin çalışma prensipleri, matematiksel temelleri ve kullanım alanları değerlendirilmiştir. İncelenen yöntemlerin her birinin farklı model türleri ve uygulama alanları için çeşitli avantajlar sunduğu, ancak tek başına tüm açıklanabilirlik gereksinimlerini karşılamadığı görülmüştür. Bu nedenle açıklama yönteminin seçiminde model yapısı, uygulama alanı ve kullanıcı gereksinimlerinin birlikte değerlendirilmesi önem taşımaktadır.

Gelecekte yapılacak çalışmalarda, farklı XAI yöntemlerinin birlikte kullanıldığı hibrit yaklaşımların geliştirilmesi, büyük dil modelleri ve üretken yapay zekâ sistemleri için daha etkili açıklama algoritmalarının oluşturulması ve açıklanabilirlik yaklaşımlarının etik ve yasal düzenlemelerle uyumlu şekilde standartlaştırılması önerilmektedir. Bu doğrultuda açıklanabilir yapay zekâ, güvenilir ve sorumlu yapay zekâ sistemlerinin geliştirilmesinde temel bileşenlerden biri olmaya devam edecektir.

KAYNAKÇA

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S.,

Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://arxiv.org/abs/1702.08608

European Commission High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI.

European Union. (2024). Artificial Intelligence Act (AI Act).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), Article 93. https://doi.org/10.1145/3236009

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. https://doi.org/10.1145/3233231

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv. https://arxiv.org/abs/1902.01876

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework* (AI RMF 1.0).

Organisation for Economic Co-operation and Development. (2019). OECD principles on artificial intelligence.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer. https://doi.org/10.1007/9783-030-28954-6

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Artificial Intelligence Review*, 56, 3299–3339.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017).

GradCAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

https://doi.org/10.1109/ICCV.2017.74

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks.

Proceedings of the 34th International Conference on Machine Learning, 3319–3328.

MİNİMUM KLİNİK ÖNEMLİ FARK DEĞERİ

İsmet Doğan[**], Nurhan Doğan[***]

GİRİŞ

Hastaların sağlık durumu, tedavi, içsel patolojik faktörler, hastaların davranışları veya diğer faktörler nedeniyle değişime tabi olduğundan hastalar, sağlık durumlarındaki değişim derecesi bakımından farklılık gösterir. Hasta içi değişimleri veya hastalar arası değişim farklılıklarını değerlendirirken iki temel soru ortaya çıkar:

Müdahale öncesi ile müdahale sonrası gözlemlenen değişim/fark, gerçek bir değişimi/farkı yansıtıyor mu ve değişim/fark tahminine ilişkin belirsizlik nedir?

Müdahale öncesi ile müdahale sonrası gözlemlenen değişim/fark önemli midir?

Birinci soru, istatistiksel değerlendirme ile ilgilidir. İkinci soru ise, müdahale öncesi ile müdahale sonrası sonuçlardaki minimum önemli değişim ve farkı tanımlama ve değerlendirme ile ilgilidir. İstatistiksel değerlendirme için en uygun yaklaşım tartışmalarında, kavramlar ve yöntemler konusunda önemli ölçüde fikir birliği mevcuttur. Bu durum, sağlık sonuçlarındaki minimum önemli değişiklik ve farkın tanımlanması ve değerlendirilmesiyle ilgili alandaki kavramlar ve yöntemler konusundaki önemli kafa karışıklığıyla tam bir tezat oluşturmaktadır (Dekker ve ark., 2024). İstatistiksel değerlendirmede anlamlı fark, yalnızca şans eseri oluşması muhtemel olmayan bir fark olarak tanımlanır. Tıp alanında ise anlamlı fark istatistiksel tanımdan daha inceliklidir ve hastanın sağlık algısından etkilenebilir. İstatistiksel olarak anlamlı bir fark, klinik ortamda hiç anlamlı olmayabilir (Tripathi ve ark., 2024). Klinik önemli fark kavramı, istatistiksel olarak anlamlı

** Prof. Dr., Afyonkarahisar Sağlık Bilimleri Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Afyonkarahisar, ismet.dogan@afsu.edu.tr, ORCID: 0000-0001-9251-3564

*** Prof. Dr., Afyonkarahisar Sağlık Bilimleri Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Afyonkarahisar, nurhandogan@hotmail.com, ORCID: 0000-0001-7224-6091

fark kavramının eksikliklerini gidermenin bir yolu olarak geliştirilmiştir. Klinik önemli fark, hastanın anlamlı ve değerli bulacağı, tekrar yapma seçeneği olsa müdahalenin tekrarını düşüneceği bir değişikliği temsil eder. Minimum klinik önemli fark (MKÖF), bu tür bir değişiklik için eşik değerdir ve bu eşik değerden daha büyük herhangi bir değişiklik miktarı anlamlı veya önemli olarak kabul edilir (Copay ve ark., 2007). MKÖF terimi ilk olarak Jaeschke ve ark. (1989) tarafından "hastanın faydalı olarak algıladığı ve rahatsız edici yan etkiler ve aşırı maliyet olmaması durumunda hastanın yönetiminde bir değişikliği gerektirecek olan, ilgi alanındaki en küçük puan farkı" olarak tanımlanmıştır (Jaeschke ve ark., 1989). Kavram, bir sonuç ölçüm aracındaki değişim eşiklerini belirlemek ve bu değişimin hasta için gerçekten anlamlı olmasını sağlamak amacıyla ortaya atılmıştır. Jaeschke ve ark. (1989)'nın iddiası, müdahale sonrası değişimi ölçen araçların kullanımı sırasında istatistiksel olarak anlamlı değişiklikler sıklıkla meydana gelse de bazı durumlarda anlamlı değişimin klinik açıdan çok az öneme sahip olduğu şeklinde ifade edilmektedir (Cook, 2008). Klinik uygulama ve tıbbi araştırmalar, ağrı, işlevsellik, tedavilerden memnuniyet, yaşam kalitesi gibi çeşitli sağlık durumlarındaki veya farklı sonuçlardaki değişikliklerin değerlendirilmesini içerir. Bu değerlendirmelerden kaynaklanan zorluklardan biri, farklılıkların istatistiksel olarak anlamlı bir değişikliği temsil edip etmediğini ve eğer öyleyse, bunun hastalar için gerçekten önemli bir klinik fayda mı yoksa zarar mı oluşturduğunu belirlemektir. Hasta tarafından bildirilen sonuçlar, hem hastaların bakış açısını hem de hastalığın ve tedavilerin yarattığı etkiyi içerecek şekilde giderek daha yaygın hale gelmektedir (Salas Apaza ve ark., 2021).

MKÖF kavramının önemi ile ilgili olarak;

- Her türlü sonuç ölçümlerine uygulanabilse de, en yaygın olarak hasta tarafından bildirilen sonuç ölçümlerinin yorumlanmasında uygulanmaktadır. Hasta tarafından bildirilen bu ölçümler gerek hastalık yükünü değerlendirmede gerekse yeni ve gelişmekte olan tedavilerin etkinliğini değerlendirmede ne kadar önemli ise, bu ölçümlerdeki değişikliklerin klinik bağlamda

yorumlanması da aynı derecede önemlidir (Tripathi ve ark., 2024),

- Değerlerinin hesaplanması ve bunların tedavi başarısı için eşik olarak kullanılması, hastaların ihtiyaç ve tercihlerinin dikkate alınmasını sağlar ve farklı gruplar arasında klinik olarak anlamlı bir iyileşme yaşayan hasta oranının karşılaştırılmasına olanak tanır. Üstelik klinisyenler bir müdahalenin hastalarının yaşamları üzerindeki etkisini daha iyi anlayabilir, örneklem büyüklüğü hesaplamaları daha sağlam hale gelebilir ve sağlık politikası yapıcıları hangi tedavilerin geri ödemeyi hak ettiğine karar verebilir. MKÖF değerlerinin tedaviye özgü değil, genel olarak hastalık kategorisine özgü olduğunu unutmamak gerekir. Bu değerler, hastanın klinik fayda algısına dayanır ve bu algı, yalnızca tedavi yönteminden değil, teşhisinden ve sonrasında ortaya çıkan semptomlarından da etkilenir (Klukowska ve ark., 2024),
- Test sonuçlarının klinik önemini değerlendirmek ve klinik karar verme süreçlerine temel oluşturabilir. İlk olarak Yaşam Kalitesi Ölçeği gibi hasta tarafından bildirilen sonuçların puanlarındaki değişikliklerin klinik önemini daha iyi açıklamak için önerilmiş, daha sonra farklı alanlara kademeli olarak genişletilerek istatistik ve klinik uygulama arasındaki boşluğu doldurmuştur (Zhang ve ark., 2023),
- Hastaların önemli olarak algıladığı, ister yararlı ister zararlı olsun, hasta tarafından bildirilen ilgi konusu sonuçtaki en küçük değişikliğin ölçüsünü sağlar ve bu değişiklik hasta veya hekimin tedavi yönetiminde bir değişiklik düşünmesine yol açar. Minimum önemli fark bilgisi, karar vericilerin tedavi etkisinin büyüklüğünü daha iyi yorumlamasına ve yararlı ve zararlı sonuçlar arasındaki dengeyi değerlendirmesine olanak tanır (Johnston ve ark., 2015),

ifadeleri yer almaktadır.

MKÖF'ün zayıf yönleri,

- Evrensel ve sabit bir özellik değildir ve hasta popülasyonları veya hastalığa özgü durumlar arasında aktarılamaz,
- Evrensel kabul görmüş bir metodolojinin bulunmaması, aynı sonuç ölçütü için bildirilen değerlerin seçilen yönteme bağlı olarak geniş bir aralıkta değişmesine yol açmaktadır,
- Belirli bir sonuç ölçütü için evrensel olarak kabul edilmiş bir değer mevcut değildir; bu değerler, incelenen popülasyona ve seçilen metodolojiye bağlı olarak değişiklik göstermektedir,
- Grubun ortalama puanına dayalı tek bir nokta tahmini olarak bildirilen değerler, gerçek değişim puanı dağılımını yansıtan güven aralıklarından yoksundur. Tek bir nokta tahmininin kullanılması, gerçekte iyileşmiş olan bireylerin ortalamanın altında kalarak iyileşmemiş olarak yanlış sınıflandırılmasına yol açabilir,

olarak ifade edilmektedir (Wright ve ark., 2012). Bir hastanın iyileşip iyileşmediği veya kötüleşip kötüleşmediği sorusu klinik uygulama için temeldir ve elde edilen bilgiler prognoz, tedavi ve yönetimle ilgili kararlar alınmasında kullanılır. Sonucun ölçülmesi, klinik uygulamanın yürütülmesinde kritik bir adımdır ve küresel değişim derecelendirme ölçekleri (KDÖ) klinik araştırmalarda çok yaygın olarak kullanılmaktadır (Kamper ve ark., 2009). Küresel değişim derecelendirmesi, Jaeschke ve ark. (1989) tarafından, klinik olarak önemli (orta ve büyük) değişiklikler ile daha küçük, klinik olarak önemsiz değişiklikler arasında bir kesme noktası oluşturmak üzere tasarlanmış ve klinik sonuç ölçümü araştırma çalışmalarında bir çapa görevi görmesi için önerilmiştir (Schmitt ve Abbott, 2015). Jaeschke ve ark. (1989) tarafından önerilen ölçek;

Çok çok daha iyi (+7 puan)
Çok daha iyi (+6 puan)
Oldukça daha iyi (+5 puan)
Orta derecede daha iyi (+4 puan)
Bir miktar daha iyi (+3 puan)
Biraz daha iyi (+2 puan)
Neredeyse aynı, hemen hemen hiç daha iyi değil (+1 puan)

Değişiklik yok (0 puan)
Neredeyse aynı, hemen hemen hiç daha kötü değil (-1 puan)
Biraz daha kötü (-2 puan)
Bir miktar daha kötü (-3 puan)
Orta derecede daha kötü (-4 puan)
Oldukça daha kötü (-5 puan)
Çok daha kötü (-6 puan)
Çok çok daha kötü (-7 puan)
biçimindedir. Hastaların her takip ziyaretinde yaşadıkları değişiklik miktarı -3 ile -1 veya +1 ile +3 arasında ise küçük değişiklik, -5 ve -4 veya +4 ve +5 arasında ise orta düzeyde değişiklik ve -7 ve -6 veya +6 ve +7 arasında ise büyük değişiklik olarak ifade edilmiştir. Tanımı gereği MKÖF, -3 ile -1 ve +1 ile +3 arasındaki değişikliklerle temsil edilmektedir (Jaeschke ve ark., 1989). Klinik araştırmalarda istatistiksel olarak anlamlı sonuçları klinik önemle ilişkilendirmenin, çalışma bulgularının yanlış yorumlanmasını önlemek ve hastaların gereksiz tedavilere maruz kalmasını engellemek için giderek daha fazla önem kazandığı bir gerçektir. İstatistiksel olarak anlamlı fark kavramının dezavantajlarının üstesinden gelmek için geliştirilen ve hastanın hissettiği bir değişikliği temsil eden MKÖF kavramı dikkate değerdir (Mishra ve ark., 2024). Bu çalışmanın amacı, klinisyenleri MKÖF puanının belirlenmesinde kullanılan metodolojiler hakkında bilgilendirmek ve herhangi bir sonuç aracı için bu puanı etkileyebilecek çeşitli faktörleri tartışmaktır.

YÖNTEMLER

Sağlıkla ilgili yaşam kalitesi için önemli sorunlardan biri, hastaların sağlıkla ilgili yaşam kalitesi puanlarının zaman içinde değişmesi veya hastalar arasında farklılık göstermesi durumunda nasıl yorumlanması gerektiğiyle ilgilidir. Altın standart bir metodolojinin bulunmaması nedeniyle, sağlıkla ilişkili yaşam kalitesi puanlarının yorumlanması güçtür ve bu puanların yorumlanabilirliğini değerlendirmek amacıyla çeşitli yaklaşımlar kullanılmaktadır (Schünemann ve Guyatt, 2005). Bir müdahalenin MKÖF değeri farklı yöntemler kullanılarak hesaplanabilmekte olup, bu hesaplamalar için en uygun metodoloji konusunda literatürde ortak bir görüş bulunmamaktadır. Bu değerin belirlenmesinde kullanılan yöntemler:

- Çapa tabanlı yöntem,
- Dağılım tabanlı yöntem,
- Uzman görüşü tabanlı/panel tabanlı yöntem,
- Literatür taraması.

başlıkları altında toplanmaktadır. Çapa tabanlı yöntemler, hastaların bildirdiği sonuç ölçüm puanlarındaki değişimi bir referans ölçüt (çapa) ile ilişkilendirerek değerlendirir. Bu yöntemler klinik anlamlılığı ortaya koyabilmekle birlikte, ölçüm hatasını dikkate almamaları önemli bir sınırlılık oluşturmaktadır. Ayrıca, uygun bir çapanın belirlenmesindeki güçlükler ve elde edilen değerin kullanılan çapa türüne bağlı olarak değişkenlik göstermesi, bu yaklaşımların uygulanabilirliğini kısıtlamaktadır. Dağılım tabanlı yöntemler, bir hasta grubunun başlangıçta bildirdiği sonuç ölçüm puanlarının istatistiksel dağılım özelliklerine dayanarak klinik açıdan anlamlı olabilecek değişim düzeyini tanımlar. Ölçüm hatasını dikkate almaları ve uygulanmalarının görece kolay olması bu yöntemlerin başlıca avantajlarıdır. Bununla birlikte, klinik yorum sağlamamaları ve farklı örneklemlerden elde edilen sonuçların değişkenlik gösterebilmesi önemli dezavantajlar arasında yer almaktadır. Ayrıca, dağılım tabanlı yöntemler için henüz kabul edilmiş bir altın standart kriter bulunmamaktadır. Görüş tabanlı yöntemler ve literatür taramaları sırasıyla uzman görüşlerine ve yayımlanmış bilimsel çalışmalara dayanmaktadır. Bu yaklaşımlar çoğunlukla çapa ve dağılım tabanlı yöntemlerin tamamlayıcısı olarak değerlendirilmektedir. Klinik açıdan anlamlı değişimin belirlenmesinde çapa tabanlı, dağılım tabanlı ve görüş tabanlı yöntemlerin birlikte kullanılması önerilmektedir. Her bir yöntem grubuna ait iyi tanımlanmış metodolojik yaklaşımlar aşağıda ayrıntılı olarak sunulmaktadır (Hu ve ark., 2009, Sedaghat, 2019; Klukowska ve ark., 2024).

Çapa Tabanlı Yöntemler

Çapa tabanlı yöntemler, bir sonuç ölçümü ile hastanın kendi algısı arasında karşılaştırma yapılmasına olanak tanır. Müdahale öncesi ile müdahale sonrası değişiklikleri bir çapa sorusuyla karşılaştırır. Hastanın kendi deneyimine dayanarak, tedaviden sonra başlangıç durumuna göre iyileşip iyileşmediğini belirlemek için "X

müdahalesinden sonra kendinizi daha iyi hissediyor musunuz?" sorusunun referans olarak kullanılması önerilmektedir. Bu durumda, hastanın değişim izlenimini anlamak için küresel bir ağrı derecelendirme ölçeği ("çok daha kötü", "biraz daha kötü", "neredeyse aynı", "biraz daha iyi" ve "çok daha iyi") kullanılabilir. Çapa sorusunun hastalar için kolay anlaşılabilir olması gerekir. Tipik çapalar, sağlık durumundaki değişiklik, semptomların varlığı, hastalık şiddeti, tedaviye yanıt, ölüm veya iş kaybı gibi gelecekteki olayların prognozu ile ilgili derecelendirmeler olabilir (Salas Apaza ve ark., 2021). MKÖF'ü belirlemek için genellikle tercih edilen yöntemlerdir. Çapa seçimi için, mevcut çapalar öznel çapalar ve nesnel çapalar olarak ikiye ayrılabilir. Öznel çapa, geçmiş dönemdeki hastalık değişimleri hakkındaki bir değerlendirmedir ve birden fazla önyargıya açıktır. Nesnel çapa ise laboratuvar inceleme göstergeleri, fizyolojik inceleme göstergeleri ve klinik sonuçları seçebilir. Bununla birlikte, çapa seçiminin uygunluğu, çapa ile araştırma göstergelerindeki veya ölçekteki puanlardaki değişim arasındaki korelasyon testine bağlıdır; yani, seçilen dış çapa ile hedef ölçüm aracı arasında orta düzeyde bir korelasyon $(r > 0.3 \; veya \; r > 0.35)$ olmalıdır. Ayrıca, son çalışmalar seçilen çapanın güvenilirliğinin ve mevcut durum önyargısının derecesinin değerlendirilebileceğini göstermiş olsa da bu hâlâ sonraki bir doğrulama yöntemi olarak kullanılmakta ve yalnızca geçiş derecelendirmeleriyle sınırlıdır. Çapa seçim stratejileri üzerine yapılan araştırmalarda hâlâ eksiklik bulunmaktadır. Aynı zamanda, uygun çapa seçildikten sonra, hastaların çapa kesme değerine göre gruplandırılması gerekir. Bu değer, hastaları çapaya göre hafif değişim grupları ve değişmeyen gruplar olarak ayırmak için kullanılan eşik değerdir. Bununla birlikte, çapa kesme değerinin belirlenmesinde henüz bir fikir birliğine varılmamıştır (Zhang ve ark., 2023). Çapa tabanlı yaklaşımın avantajı, sonuç ölçüm puanındaki değişimin, hastanın bakış açısını hesaba katan anlamlı bir dış çapa ile bağlantılı olmasıdır. En yaygın olarak, bir hastadaki değişimi belirlemeye çalışırken MKÖF'ü tanımlamak için küresel değişim derecelendirme puanı kullanılır. Bu, bir bakım dönemi boyunca müdahalenin etkili olup olmadığını anlamak isteyen klinisyen için klinik olarak önemlidir. Çapa tabanlı yöntemler, uzun vadeli yanıt verme üzerindeki hatırlama yanlılığı etkisi

nedeniyle eleştirilmiştir. Hatırlama yanlılığı, bir hastanın en yakın zamanda olanları en iyi hatırlaması ve daha uzak geçmişi daha az net hatırlaması durumunda ortaya çıkar. Küresel derecelendirmeleri kullanan çapa tabanlı yöntemler, küresel enstrümanın ölçüm hassasiyetini hesaba katamamaları nedeniyle de eleştirilmiştir. Bu, bir durumun kötüleşmesi ile bir durumun iyileşmesi karşılaştırıldığında önem kazanır. Genellikle, durumun kötüleşmesi için gereken minimum değişim miktarı, durumun olumlu yönde değişmesi için gereken miktardan daha azdır. Eğer referans noktası hassasiyetten yoksunsa, hastanın gerçek tepkisi gizlenecektir. Çapa tabanlı yöntemlerin sınırlılıkları;

- Çapa sorusu, birden fazla sonuç türünü yansıtabilen, hasta tarafından bildirilen bir sonuç ölçümündeki değişiklikleri tam olarak yakalayamayabilir.
- Çapa sorusu uygulandığında, tedavi sonrası zaman noktasında tedavi öncesi durum için hatırlama yanlılığı söz konusudur.

şeklinde ifade edilmektedir (Sedaghat, 2019).

Hasta içi skor değişimi: Yöntem her hasta için, sonuç ölçme aracının müdahale öncesi ve müdahale sonrası puanları arasındaki değişimi hesaplamaya odaklanır. MKÖF, iyileşme gösteren hastaların $(i = 1, 2, 3, \dots, k)$ ortalama Δ skorudur.

$$MK\ddot{O}F_{hasta\,i\c{c}i} = \left(\frac{\sum_{i=1}^{k} \Delta_i}{k}\right)_{iyile\c{s}me\ g\ddot{o}steren} \quad (1)$$

Δ: Müdahale öncesi ile müdahale sonrası skorlar arasındaki değişim miktarı,

Hastalar arası skor değişimi: MKÖF, iyileşme gösteren hastaların $(i = 1, 2, 3, \dots, k)$ ortalama Δ skorundan, iyileşme göstermeyenlerin $(j = 1, 2, 3, \dots, r)$ ortalama Δ skorunun çıkarılmasıyla elde edilir.

$$MK\ddot{O}F_{hastalar\ arası} = \left(\frac{\sum_{i=1}^{k} \Delta_i}{k}\right)_{iyile\c{s}me\ g\ddot{o}steren} - \left(\frac{\sum_{j=1}^{r} \Delta_j}{r}\right)_{iyile\c{s}me\ g\ddot{o}stermeyen} \quad (2)$$

Δ: Müdahale öncesi ile müdahale sonrası skorlar arasındaki değişim miktarı,

ROC eğrisi yaklaşımı: MKÖF, eğrinin en sol üst köşesindeki noktaya karşılık gelen Delta (Δ) skorudur. Bu nokta, en az yanlış sınıflandırmayla ilişkili skor değişimine karşılık gelir. Hastaları iyileşenler ve iyileşmeyenler olarak ikiye ayırmak için çapa üzerinde farklı kesme noktaları kullanılabilir. Duyarlılık, dışsal bir kriterde iyileşme bildiren ve hasta tarafından bildirilen sonuç puanları MKÖF değerinin üzerinde olan hastaların oranını ifade eder. Özgüllük ise iyileşme bildirmeyen ve hasta tarafından bildirilen sonuç puanları MKÖF değerinin altında olan hastaların oranını ifade eder. MKÖF için üzerinde anlaşılmış ideal bir duyarlılık veya özgüllük düzeyi yoktur, ancak araştırmacılar genellikle dengeyi hedefler. Alıcı işletim karakteristik (ROC) eğrileri, "iyileşmiş" ve "değişmemiş" hastaları en iyi şekilde ayırt eden hasta tarafından bildirilen sonuç puanını belirlemek için sıklıkla kullanılır. ROC eğrisinin altındaki alan, puanların bu gruplar arasında ne kadar iyi ayrım yaptığını yansıtır. 0.7-0.8 arası bir alan iyi, 0.8-0.9 arası ise doğruluk açısından mükemmel kabul edilir, ancak ROC analizi için hasta gruplarının seçimi yine de biraz keyfi olabilir (Mishra ve ark., 2024).

Sosyal karşılaştırma yaklaşımı: Daha az kullanılan bir yaklaşımda ise hastalar birbirleriyle sağlık durumları hakkında konuşurlar ve ardından kendilerini konuştukları hastaya kıyasla "biraz daha iyi", "biraz daha kötü" veya "aynı" olarak değerlendirirler. MKÖF, kendilerini "biraz daha iyi" veya "biraz daha kötü" olarak değerlendirenler ile diğer hastayla "aynı" olduğunu söyleyenler arasındaki puan farkıdır. Gruplar, hastalık düzeyi, durumun ciddiyeti, hastalıkla ilgili olmayan dış kriterler ve varsayımsal sağlık durumları gibi kriterlere göre karşılaştırılabilir. Bu yaklaşımın zorluğu, hastaları birbirleriyle uygun şekilde eşleştirmektir (Mishra ve ark., 2024).

Dağılım Tabanlı Yöntemler

Uygun bir çapa bulmak zor olduğunda, dağılım tabanlı yöntem benimsenir. MKÖF değerlerini belirlemek için kullanılan dağılım tabanlı yöntemler, örneklemin istatistiksel özelliklerine ve dolayısıyla değişimin şans eseri meydana gelme olasılığına göre istatistiksel olarak

anlamlı değişikliklere dayanmaktadır. Verilerin dağılımına dayanarak MKÖF'ü belirlemeyi amaçlar. Ancak istatistiksel bir bakış açısı ile elde edilen tahmini sonuçlar tek başına MKÖF'ü bilimsel olarak açıklayamaz (Zhang ve ark., 2023). Dağılım tabanlı yöntemler, bir farkın şansın ötesinde anlamlı olma olasılığının, verilerin yayılımından nasıl tahmin edilebileceğini belirlemeye çalışır. Bir çalışmadan elde edilen sonuçların istatistiksel özelliklerine dayanırlar. Dağılım tabanlı yöntemler bireysel hasta değerlendirmelerinden türetilmediği için (hastalar için önemli ve klinik olarak anlamlı sonuçları belirleyemediği anlamına gelir) MKÖF'ü belirlemek için muhtemelen kullanılmamalıdırlar. Mantığı, yalnızca minimum tespit edilebilir etkiyi, yani rastgele ölçüm hatasına atfedilmesi muhtemel olmayan bir etkiyi belirleyebilen istatistiksel akıl yürütmeye dayanmaktadır. MKÖF'ü hesaplamak için standart sapma, etki büyüklüğü, ortalama standart hata vb. içermektedir (Salas Apaza ve ark., 2021). Dağılım tabanlı yöntemlerin avantajı, rastgele varyasyonun belirli bir seviyesinin ötesindeki değişimi hesaba katabilme yeteneği ve harici bir kriter gerektirmemesi nedeniyle basit olmasıdır. Dağılım tabanlı yöntemlerin sınırlılıkları;

- Hem kötüleşme hem de iyileşme için benzer sonuçlar üretir (bir müdahalenin iyileşmesi ve bozulması arasında net bir ayrım yapılmamaktadır). Bu da yorumlamayı daha kolay ancak daha şüpheli hale getirir, çünkü kötüleşme için iyileşmeden daha yüksek bir MKÖF sıklıkla gözlemlenir.
- Hasta tarafından bildirilen sonuçtaki değişime dayanmaz ve bu nedenle sonuçtaki değişimin derecesini veya önemini yansıtacak bir ölçü içermez.
- Etki büyüklüğü veya hasta tarafından bildirilen sonuç ölçümündeki değişkenlik derecesi ile ilgili varsayımlara bağlı olarak birkaç farklı değer olarak hesaplanabilir.
- Standart sapmaya dayalı yöntemler örneklem/kohort özelinde olabilir (genelleştirilebilirlik eksikliği).
- Standart hata en iyi test-tekrar test güvenilirliği ile hesaplanır, ancak test-tekrar test güvenilirliğinin belirlenmesindeki zorluk nedeniyle kullanılmayabilir,

- Klinik olarak anlamlı iyileşmeyi belirlemek için üzerinde anlaşılmış az sayıda ölçüt olması,
- İstatistiksel anlamlılıktan belirgin şekilde farklı olan, hastanın klinik olarak önemli değişime ilişkin bakış açısını ele almaz,
- Örnekleme özgüdür.

olarak ifade edilmektedir (Sedaghat, 2019).

Ölçüm standart hatası (SEM) yöntemi

$$MKÖF = X * \sqrt{1 - ICC} * SD_{başlangıç puanlar} \quad (3)$$

SD: Standart sapma,

ICC: Sınıf içi korelasyon katsayısı,

X değeri olarak küçük etki için 1.00, orta etki için 1.96, büyük etki için ise 2.77 değerleri kullanılır. Ölçüm standart hatası (SEM), bir sonucun tekrarlı ölçümlerinin gerçek puandan ne kadar farklı olabileceğini belirleyerek, sonuç ölçümünün güvenirliğini kavramsallaştırır. MKÖF ile ilgili çalışmalarda, güvenirlik ölçütü olarak ICC veya Cronbach alfa kullanılmasına rağmen, Cronbach alfa değerinin MKÖF hesaplamasında yetersiz bir güvenirlik ölçütü olduğu belirtilmektedir (Sedaghat, 2019).

Etki büyüklüğü (ES) yöntemi

$$MKÖF_{ES\ tabanlı} = 0{,}2 * SD_{başlangıç puanlar} \quad (4)$$

SD: Standart sapma,

Standartlaştırılmış yanıt ortalaması (SRM) yöntemi

$$MKÖF_{SRM\ tabanlı} = 0{,}2 * SD_{değişim puanları} \quad (5)$$

SD: Standart sapma,

Standart sapma (SD) yöntemi

$$MKÖF_{SD\ tabanlı} = 0{,}5 * SD_{başlangıç puanlar} \quad (6)$$

SD: Standart sapma,

Minimum tespit edilebilir değişim (%95 MDC) yöntemi

$$MKÖF = 1{,}96 * SEM * \sqrt{2} \quad (7)$$

$$SEM = \sqrt{1 - ICC} * SD_{başlangıç\ puanlar}$$

SEM: Standart Error of Measurement,

SD: Standart sapma,

ICC: Sınıf içi korelasyon katsayısı,

MKÖF ile ilgili çalışmalarda, güvenirlik ölçütü olarak sıklıkla sınıf içi korelasyon katsayısı (ICC) veya Cronbach alfa katsayısı kullanılmakla birlikte, Cronbach alfa değerinin bu hesaplamalar için yeterli bir güvenirlik ölçütü olmadığı bildirilmektedir (Sedaghat, 2019).

Güvenilir değişim indeksi yöntemi

İndeks işlevsel olarak, ölçüm hatası nedeniyle ölçümdeki beklenen değişkenliği, her biri ölçüm hatasından etkilenen iki ölçüm zaman noktasının kullanılmasından kaynaklanan değişimi tahmin ederken beklenen değişkenliği ve yalnızca normal değişkenlikten kaynaklanan değişimi dışlamak için gereken değişim eşiğini belirlemek için sıfır hipotezi anlamlılık testi $\alpha = 0.05\ (çift\ yönlü)$ mantığını kullanır. Güvenilir değişim indeksi (GDİ), bir örneklemdeki tüm bireyleri değerlendirmek için kullanılan eşiktir. Eşikten daha büyük değişime sahip olanlar güvenilir bir şekilde iyileşmiş/kötüleşmiş olarak, eşikten daha az değişime sahip olanlar ise iyileşmemiş/kötüleşmemiş olarak kabul edilir (Crenshaw ve Monson, 2025).

$$SEM = \sqrt{1 - ICC} * SD_{başlangıç\ puanlar} \quad (8)$$

SD: Standart sapma,

ICC: Sınıf içi korelasyon katsayısı,

$$s_{diff} = \sqrt{2 * SEM^2} \quad (9)$$

$$GDİ = 1.96 * s_{diff} \quad (10)$$

GDİ, bir örneklemdeki tüm bireyleri değerlendirmek için kullanılabildiği gibi her bir hasta için de hesaplanabilmektedir. Bu durumda;

$$GD\dot{I} = \frac{Bireysel\ Hasta\ Değişim\ Puanı}{S_{diff}} \quad (11)$$

kullanılmalıdır. Teorik olarak, bireysel $GD\dot{I} > 1.96$ ise, hastanın istatistiksel olarak güvenilir bir şekilde tanımlanabilir bir iyileşme $(p < 0.05)$ yaşadığı kabul edilebilir. Yine de bu, değişimin hasta için klinik olarak anlamlı olup olmadığını değil, değişimin yalnızca ölçüm hatasına atfedilmemesi gerektiğini ve muhtemelen gerçek puan değişiminin bir bileşenini içerdiğini yansıtır. Bu nedenle, bu yöntem, tüm dağılım tabanlı yöntemlerde olduğu gibi, hasta tercihlerine değil, örneklemin istatistiksel özelliklerine dayandığı için MKÖF hesaplamalarında önerilmez (Klukowska ve ark., 2024).

Uzman Görüşü Tabanlı/Panel Tabanlı Yöntemler

Uzman görüş birliği yöntemi, MKÖF'ü belirlemek için grup kararı ve görüş birliği yöntemine dayanır, ancak sonuçlar yalnızca uzmanların öznel yargısına dayanır ve yalnızca yardımcı bir yöntem olarak kullanılır. Literatür analizi yöntemi ve uzman görüş birliği yöntemi nispeten belirli bir ilgi alanına, ya da gruba odaklanan yaklaşımlardır (Klukowska ve ark., 2024).

Delphi yöntemi

Delphi Yöntemi, tıbbi bir konu hakkında fikir birliği oluşturmak için uzmanların kolektif görüşünü kullanan sistematik bir yaklaşımdır. Çoğunlukla en iyi uygulama kılavuzlarını geliştirmek için kullanılmıştır. Bununla birlikte, MKÖF belirlemesine yardımcı olmak için de kullanılabilir. Yöntem, üyelerden oluşan bir panele anket veya soru formları dağıtmaya odaklanır. Anonimleştirilmiş cevaplar gruplandırılır ve sonraki turlarda uzman panelle tekrar paylaşılır. Bu, uzmanların görüşlerini yansıtmalarına ve diğerlerinin yanıtlarının güçlü ve zayıf yönlerini değerlendirmelerine olanak tanır. Fikir birliğine varılana kadar süreç tekrarlanır. Anonimliğin sağlanması, belirli bir katılımcının kendi görüşünün diğer kişisel faktörlerden etkilenmesi veya görülmesiyle ilgili olası herhangi bir önyargıyı önler

(Klukowska ve ark., 2024). Ancak, MKÖF konusunda bir fikir birliğine varmak için birkaç tur geri bildirim gerekebileceğinden, zaman alıcı bir süreç olabilir. Ayrıca, uzman panelinin önyargılarına da tabidir. Eğer panel uzman popülasyonunu temsil etmiyorsa, MKÖF daha geniş popülasyona genelleştirilemeyebilir (Mishra ve ark., 2024).

Literatür taraması

Literatür analizi, yayınlanmış literatürü sistematik olarak inceler ve sonuçları MKÖF için referans temeli olarak sentezler. Açıkçası, bu yöntem ikincil literatüre dayanır ve MKÖF'ün belirlenmesinde yalnızca yardımcı bir rol oynamalıdır (Klukowska ve ark., 2024).

Diğer Yöntemler

MKÖF'ün temel değerlerden %30 azaltma olarak hesaplanması

Son yıllarda, MKÖF hesaplamalarına alternatif olarak temel değerlerden %30'luk basit bir azaltma yöntemi ortaya çıkmıştır.

Eşler arası-t istatistiği (Wright ve ark., 2012)

$$MK\ddot{O}F = \frac{x_1 - x_0}{\sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n(n-1)}}} \quad (12)$$

d_i: *i.* hastanın ön test ve son test puanları arasındaki fark,

$\bar{d}$: ön test ve son test puanları arasındaki farkların ortalaması,

Standartlaştırılmış yanıt ortalaması (Standardized response mean) (Wright ve ark., 2012)

$$MK\ddot{O}F = \frac{x_1 - x_0}{\sqrt{\frac{\Sigma(d_i - \bar{d})^2}{(n-1)}}} \quad (13)$$

d_i: *i.* hastanın ön test ve son test puanları arasındaki fark,

$\bar{d}$: ön test ve son test puanları arasındaki farkların ortalaması,

Tepki hızı istatistiği (Responsiveness statistic**)** (Wright ve ark., 2012)

$$MKÖF = \frac{x_1 - x_0}{\sqrt{\frac{\sum\left(d_{i-stabil} - \bar{d}_{stabil}\right)^2}{(n-1)}}} \quad (14)$$

d_i: *i.* stabil hastanın ön test ve son test puanları arasındaki fark,

$\bar{d}$: ön test ve son test puanları arasındaki farkların ortalaması,

Çapa tabanlı minimal önemli değişim dağılım modeli

Teorik olarak, çapa ve dağılım tabanlı yaklaşımların birleştirilmesi daha sağlam sonuçlar sağlayabilir. Önerilen stratejiler arasında, farklı yöntemlerle elde edilen değerlerin ortalamasının alınarak iki yöntemin basit bir şekilde birleştirilmesi bulunmaktadır. Örneğin, bir hastanın klinik olarak anlamlı değişim gösterdiğinin kabul edilebilmesi için hem hasta memnuniyetine dayalı ROC tabanlı bir çapa kriterinin hem de %95 güven düzeyine dayanan minimum tespit edilebilir değişim (MDC) kriterinin birlikte sağlanması gerekebilir. Bu yöntemde, iyileşme için üst %95 ve kötüleşme için alt %95 sınırları, yalnızca müdahale öncesi ve müdahale sonrası skorların önemli değişim göstermediği belirlenen hasta grubuna dayanarak bulunur. Sınırların elde edilmesinde Eşitlik 15 ile verilen formül kullanılır (kötüleşme veya iyileşme için sırasıyla 1,645 × SD eklemek yerine çıkartılır) (Klukowska ve ark., 2024).

$$95\%\ Üst\ Sınır = Ortalama\ Değişim_{Değişim\ Görülmeyen\ Hasta\ Grubu} + 1.645 * SD_{Değişim\ Puanları} \quad (15)$$

SD: Standart sapma

UYGULAMA

Bir araştırmacı, enfeksiyon polikliniğine başvuran ve 14 gün süreyle standart tedavi uygulanan 34 hastanın klinik yanıtını değerlendirmeyi amaçlamaktadır. Tedavinin etkinliğini objektif olarak

ortaya koyabilmek için inflamasyon göstergesi olan C-reaktif protein (CRP) düzeylerini kullanmayı planlamıştır. Bu kapsamda her hastadan tedavi öncesinde (0. gün) ve tedavi sürecinde 3., 7., 10. ve 14. günlerde olmak üzere toplam beş farklı zaman noktasında CRP ölçümü alınmıştır. Araştırmacı, tedavi sürecinde CRP değerlerindeki değişimin yalnızca istatistiksel olarak anlamlı olup olmadığını değil, aynı zamanda klinik açıdan anlamlı bir iyileşmeyi yansıtıp yansıtmadığını da belirlemek istemektedir. Bu nedenle minimum klinik önemli farkın (MKÖF) hesaplanmasını hedeflemektedir. MKÖF'ün belirlenmesinde, tedavinin nihai etkisini yansıtması nedeniyle başlangıç ve 14. gün CRP değerleri arasındaki değişim esas alınmıştır. Veriler Tablo 1'de verildiği gibidir.

Tablo 1. Veri Seti (n=34)

Hasta	Tedavi Öncesi CRP	Tedavi Sonrası CRP	Değişim (Δ) (Sonra-Önce)	Algılanan değişim (Anchor)
1	233	170	-63	Biraz iyi
2	228	68	-160	İyileşti
3	303	82	-221	İyileşti
4	247	40	-207	İyileşti
5	44	7	-37	İyileşti
6	213	96	-117	İyileşti
7	62	20	-42	İyileşti
8	180	67	-113	Biraz iyi
9	337	47	-290	İyileşti
10	312	181	-131	Biraz iyi
11	301	103	-198	İyileşti
12	153	149	-4	Aynı
13	169	10	-159	İyileşti
14	128	84	-44	Biraz iyi
15	324	186	-138	Biraz iyi
16	149	25	-124	İyileşti
17	262	114	-148	İyileşti
18	299	25	-274	İyileşti
19	73	17	-56	İyileşti
20	343	64	-279	İyileşti

21	492	85	-407	İyileşti
22	141	85	-56	Biraz iyi
23	343	55	-288	İyileşti
24	255	57	-198	İyileşti
25	261	96	-165	İyileşti
26	17	46	29	Kötüleşti
27	318	7	-311	İyileşti
28	279	174	-105	Biraz iyi
29	141	133	-8	Aynı
30	247	69	-178	İyileşti
31	272	60	-212	İyileşti
32	155	105	-50	Biraz iyi
33	31	75	44	Kötüleşti
34	173	28	-145	İyileşti
Ortalama	220.15	77.35	-142.79	
Standart Sapma	106.556	51.225	105.394	
Standart Hata	18.274	8.785	18.075	

Shapiro-Wilk testi kullanılarak verilerin dağılımının normal dağılıma uymadığı belirlenmiştir. 14 günlük tedavinin etkili olup olmadığı Wilcoxon testi ile değerlendirilmiş olup, tedavinin etkili olduğu sonucuna ulaşılmıştır ($p < 0.001$). MKÖF değerinin belirlenmesinde dağılım tabanlı SEM yöntemi kullanılmıştır. Buna göre;

$$MKÖF = X * \sqrt{1 - ICC} * SD_{başlangıçpuanlar}$$

$$MKÖF = 1.00 * \sqrt{1 - 0{,}90} * 106{,}556 = 33{,}7$$

Yorum: Hesaplamada ölçümlerin güvenirliğini yansıtan sınıf içi korelasyon katsayısı (ICC) 0,90 olarak alınmış olup, bu değer CRP ölçümlerinin yüksek düzeyde güvenilir olduğunu göstermektedir. SEM yöntemine dayalı olarak yapılan hesaplama sonucunda MKÖF değeri 33,7 olarak bulunmuştur. Bu sonuç, CRP düzeylerinde en az 33,7 birimlik bir değişimin, ölçüm hatasının ötesinde, gerçek ve klinik açıdan anlamlı bir değişimi temsil ettiğini göstermektedir. Tedavi

sürecinde CRP değerinde 33,7 birimin üzerinde gözlenen azalmalar, hastalarda klinik olarak anlamlı bir biyokimyasal iyileşme göstergesi olarak değerlendirilebilir.

Aynı veri seti için çapa tabanlı hasta içi skor değişimi yöntemi kullanılarak MKÖF değeri hesaplanmıştır. Tablo 1'de verilen sonuçlara göre, 2-7, 9, 11, 13, 16-21, 23-25, 27, 30, 31 ve 34'ncü hastalar iyileşme gösteren hastalardır. Dolayısıyla,

$$MKÖF_{hasta\,içi} = (\Delta/n)_{iyileşme\ gösteren}$$

$$MKÖF_{hasta\,içi} = \frac{(-160) + (-221) + (-207) + \cdots\ldots\ldots\ldots + (-178) + (-212) + (-145)}{22}$$

$$= -\frac{4216}{22} = -191{,}64$$

$$MKÖF_{hasta\,içi} = |-191{,}64| = 191{,}64$$

Yorum: MKÖF değeri 191,64 olarak bulunmuştur. Hastaların klinik olarak anlamlı bir iyileşme gösterdiği kabul edilen grupta, tedavi öncesine kıyasla CRP düzeylerinde ortalama 191,64 birimlik bir düşüş gözlenmiştir.

Bu iki yöntem birlikte değerlendirildiğinde, 33,7 birimlik düşüşün ölçüm hatasının ötesinde gerçek değişimi, 191,64 birimlik düşüşün ise klinik olarak anlamlı iyileşmeyi temsil ettiği söylenebilir. 191,64 değeri, klinik olarak anlamlı iyileşme gösteren hastalarda CRP'de gözlenen ortalama düşüşün büyüklüğünü verir ve bu nedenle klinik eşik olarak kullanılması daha uygundur. 33,7 değeri ise ölçüm hatasının ötesinde gerçek değişimi gösteren istatistiksel alt sınırdır.

TARTIŞMA

MKÖF, tıp araştırmalarında ve uygulamalarında giderek daha fazla ilgi ve önem kazanmaktadır. Geleneksel olarak çalışmalar, tedavi ve kontrol grupları arasında ilgi duyulan bir sonuç ölçümündeki ortalama iyileşmeyi/değişimi istatistiksel testler kullanarak tahmin etmekte ve karşılaştırmaktadır. Bu testler, tedavi grubundaki iyileşmenin şans eseri beklenenden daha fazla istatistiksel olarak anlamlı olup olmadığını (yani $p < 0.05$) belirlemeyi amaçlamaktadır. İstatistiksel anlamlılık klinik anlamlılığa eşit değildir. Grup içi iyileşmenin büyüklüğü veya

gruplar arası iyileşme farklılıkları (başlangıçtaki değişkenliğe göre), genellikle grup içi ve gruplar arası etki büyüklüğünün ölçüleri olarak rapor edilmektedir. İstatistiksel testler "Tedavi grubu, kontrol grubuna göre şans eseri beklenilenden daha fazla iyileşme gösteriyor mu?" sorusunu yanıtlamak için kullanılabilir. Etki büyüklükleri ise, "Ortalama olarak, grup içi iyileşme (veya gruplar arası iyileşme karşılaştırması) küçük mü yoksa büyük mü (denekler arasındaki değişkenlik derecesine göre)?" sorusunu yanıtlayarak istatistiksel anlamlılığın ötesine, klinik anlamlılığa doğru yorumlamayı genişletmeye çalışmaktadır. MKÖF, istatistiksel ve klinik anlamlılık arasındaki uyumun artırılması ve çalışmaların daha etkin biçimde tasarlanması amacıyla kullanılabilir. İstatistiksel testler, etki büyüklüğü ve bu ölçüt birlikte ele alındığında, araştırmacılar ve klinisyenler için tedavi etkinliğine ilişkin grup düzeyinde tamamlayıcı ve özgün bilgiler sunmaktadır. Bununla birlikte, söz konusu yaklaşım bireysel düzeydeki değişimlerin değerlendirilmesinde de kullanılabilmektedir. Ancak sahip olduğu potansiyele rağmen, hesaplanmasına ilişkin metodolojik yaklaşımlar literatürde tartışmalı olmaya devam etmektedir (Malec ve Ketchum, 2020). Sağlık hizmeti maliyetlerindeki artış, bilgilendirilmiş ve etkili sağlık hizmeti müdahalelerinin sunulmasında hesap verebilirliği zorunlu kılmaktadır. Klinik uygulamada bu puan, istatistiksel güç ve örneklem büyüklüğünün belirlenmesinde ve bir müdahalenin etkinliğini yansıtan terapötik eşik değerinin saptanmasında kullanılabilir. Her ne kadar cazip bir kavram olsa da tek bir sonuç ölçeği için literatürde çok sayıda farklı değer bulunması nedeniyle, klinisyenlerin bu puanları sorgulamadan kabul etmeleri önerilmemektedir. Bu değerlerin nasıl belirlendiğinin anlaşılması, daha nüanslı bir yorumlamayı ve klinik uygulamada daha bilinçli bir kullanımını mümkün kılacaktır (Wright ve ark., 2012). Hasta tarafından bildirilen sonuç ölçütleri kullanılarak değerlendirilen klinik iyileşmelerin yorumlanmasında, MKÖF önemli ve gerekli bir araç olarak kabul edilmektedir. Ancak bu ölçütlere dayalı bir değer belirlenmesi, seçilen metodolojiye bağlı olarak değişen ve nihai sonucu etkileyebilen birçok varsayım ve ayrıntı içermektedir. Klinik sonuç verilerine uygulanması ve yorumlanması, yalnızca kullanılan hesaplama yöntemlerinin bilinmesini değil, aynı zamanda bu yöntemlerin güçlü ve zayıf yönlerine ilişkin bilgi sahibi olmayı da

gerektirmektedir. MKÖF'ün çalışmalara dahil edilmesi, araştırmanın kalitesini artırabilir ve hastaların bakımını iyileştirebilir (Mishra ve ark., 2024). Literatürde MKÖF'ün belirlenmesine yönelik çok sayıda yaklaşım bulunmasına karşın, hangi yöntemin en doğru sonucu verdiği konusunda tam bir görüş birliği bulunmamaktadır. Bunun temel nedeni, klinik olarak önemli değişim kavramının hem istatistiksel hem de hasta algısına dayalı öznel bileşenler içermesidir. Çapa tabanlı yöntemler hasta perspektifini doğrudan yansıtması nedeniyle klinik anlamlılığın değerlendirilmesinde önemli avantajlar sağlarken, dağılım tabanlı yöntemler ölçüm hatasını ve örneklem özelliklerini dikkate alarak değişimin istatistiksel güvenilirliğini ortaya koymaktadır. Bu nedenle güncel literatürde tek bir yönteme dayalı MKÖF hesaplamaları yerine, farklı yöntemlerden elde edilen sonuçların birlikte değerlendirilmesi önerilmektedir. Özellikle hasta tarafından bildirilen sonuç ölçütlerinde, çapa tabanlı ve dağılım tabanlı yaklaşımların birlikte kullanılması, klinik ve istatistiksel anlamlılığın daha dengeli biçimde yorumlanmasına katkı sağlayabilmektedir. Bu yaklaşım, hem klinik araştırmalarda hem de rutin sağlık hizmetlerinde daha güvenilir kararların alınmasına yardımcı olabilir.

SONUÇ VE ÖNERİLER

Klinik çalışmalarda MKÖF hesaplamak;

- Klinik çalışmalarda örneklem büyüklüğü hesaplamalarında bu değerin belirlenmesi öncelikli bir gerekliliktir. Doğru tahmin edilen bir eşik, çalışmanın klinik olarak anlamlı bir tedavi etkisini saptayabilmesi için yeterli güce ulaşmasına katkı sağlayabilir. Eşik değeri küçük olduğunda, tedavi grupları arasındaki anlamlı farklılıkları ortaya koymak için daha düşük örneklem büyüklükleri yeterli olabilmekte; bu durum çalışmanın maliyet ve süresini azaltıcı bir etki doğurabilmektedir.
- İkinci olarak, bu eşik değerinin anlaşılması, tedavi etkilerinin klinik öneminin yorumlanmasına katkı sağlayabilir. Söz konusu değer, bir tedavi etkinin hastalar açısından klinik olarak anlamlı sayılacak düzeyde olup olmadığını belirlemeye yardımcı olur.

- Üçüncü olarak, bu eşik değer, klinik olarak anlamlı değişikliklere daha duyarlı yeni sonuç ölçütlerinin geliştirilmesine rehberlik edebilir. Eşik değeri daha küçük olan ölçütler, hasta sonuçlarındaki daha küçük ancak klinik açıdan anlamlı değişiklikleri saptama olasılığını artırabilir; bu da çalışmalarda duyarlılığı yükseltebilir ve tedavi etkilerinin tespit edilmesine katkı sağlayabilir.
- Son olarak, bu eşik değerinin anlaşılması, klinik çalışmalarda uygun uç noktaların seçilmesine katkı sağlayabilir. Bazı durumlarda kullanılan uç noktalar, hasta merkezli sonuçlarla örtüşmeyebilir veya klinik açıdan anlamlı bir eşik değeri içermeyebilir. Bu nedenle, eşik değerinin değerlendirilmesi, hastalar için daha ilgili ve anlamlı olan uç noktaların belirlenmesine yardımcı olabilir.

gibi nedenlerden dolayı çok önemlidir. Bununla birlikte MKÖF değerlerinin evrensel ve değişmez eşikler olarak değerlendirilmemesi gerekmektedir. Aynı sonuç ölçütü için elde edilen MKÖF değerleri; hasta popülasyonu, hastalığın özellikleri, kullanılan ölçüm aracı, izlem süresi ve hesaplama yöntemi gibi birçok faktörden etkilenebilmektedir. Bu nedenle bir çalışmada elde edilen MKÖF değerinin farklı hasta gruplarına veya farklı klinik koşullara doğrudan aktarılması her zaman uygun olmayabilir. Araştırmacıların ve klinisyenlerin MKÖF değerlerini yorumlarken kullanılan metodolojiyi dikkate almaları ve sonuçları klinik bağlam içerisinde değerlendirmeleri önem taşımaktadır. Sonuç olarak MKÖF, istatistiksel anlamlılık ile klinik anlamlılık arasında köprü kuran önemli bir ölçüt olup, hasta merkezli araştırmaların planlanması, yürütülmesi ve yorumlanmasında değerli bilgiler sunmaktadır. Ancak MKÖF değerlerinin mutlak doğrular olarak değil, klinik bağlam, hasta özellikleri ve kullanılan hesaplama yöntemiyle birlikte değerlendirilmesi gerekmektedir. Bu yaklaşım, araştırma sonuçlarının daha doğru yorumlanmasına ve klinik kararların hasta yararını en üst düzeye çıkaracak biçimde verilmesine katkı sağlayacaktır.

KAYNAKÇA

Cook, C.E. (2008). Clinimetrics Corner: The Minimal Clinically Important Change Score (MCID): A Necessary Pretense. *Journal of Manual & Manipulative Therapy*, 16(4), E82-3.

Copay, A.G., Subach, B.R., Glassman, S.D., Polly, D.W. Jr. ve Schuler, T.C. (2007). Understanding The Minimum Clinically Important Difference: A Review of Concepts and Methods. *Spine*, 7(5), 541-546.

Dekker, J., de Boer, M. ve Ostelo, R. (2024). Minimal Important Change and Difference in Health Outcome: An Overview of Approaches, Concepts, and Methods. *Osteoarthritis and Cartilage*, 32(1), 8-17.

Hu, G., Huang, Q., Huang, Z. ve Sun, Z. (2009). Methods to Determine Minimal Clinically Important Difference. *Zhong Nan Da Xue Xue Bao Yi Xue Ban*, 34(11), 1058-1062.

Jaeschke, R., Singer, J. ve Guyatt, G.H. (1989). Measurement of Health Status. Ascertaining The Minimal Clinically Important Difference. *Controlled Clinical Trials*, 10(4), 407-415.

Johnston, B.C., Ebrahim, S., Carrasco-Labra, A., Furukawa, T.A., Patrick, D.L., Crawford, M.W., Hemmelgarn, B.R., Schunemann, H.J., Guyatt, G.H. ve Nesrallah, G. (2015). Minimally Important Difference Estimates and Methods: A Protocol. *BMJ Open*, 5(10), e007953.

Kamper, S.J., Maher, C.G. ve Mackay, G. (2009). Global Rating of Change Scales: A Review of Strengths and Weaknesses and Considerations for Design. *Journal of Manual & Manipulative Therapy*, 17(3), 163-70.

Klukowska, A.M., Vandertop, W.P., Schröder, M.L. ve Staartjes, V.E. (2024). Calculation of The Minimum Clinically Important Difference (MCID) Using Different Methodologies: Case Study and Practical Guide. *European Spine Journal*, 33(9), 3388-3400.

Malec, J.F. ve Ketchum, J.M. (2020). A Standard Method for Determining The Minimal Clinically Important Difference for Rehabilitation Measures. *Archives of Physical Medicine and Rehabilitation*, 101(6), 1090-1094.

Mishra, B., Sudheer, P., Agarwal, A., Nilima, N., Srivastava, M.V.P. ve Vishnu, V.Y. (2024). Minimal Clinically Important Difference of Scales Reported in Stroke Trials: A Review. *Brain Sciences*, 14, 80.

Salas Apaza, J.A., Franco, J.V.A., Meza, N., Madrid, E., Loézar, C. ve Garegnani, L. (2021). Minimal Clinically Important Difference: The basics. *Medwave*, 21(3), e8149.

Schmitt, J. ve Abbott, J.H. (2015). Global Ratings of Change Do Not Accurately Reflect Functional Change Over Time in Clinical Practice. *Journal of Orthopaedic & Sports Physical Therapy*, 45(2), 106-111.

Schünemann, H.J. ve Guyatt, G.H. (2005). Commentary-Goodbye M(C)ID! Hello MID, Where Do You Come From? *Health Services Research*, 40(2), 593-597.

Sedaghat, A.R. (2019). Understanding the minimal clinically important difference (MCID) of patient-reported outcome measures. *Otolaryngology–Head and Neck Surgery*, 161(4), 551-560.

Tripathi, S.H., Min, S., Cody, A.S., Shukla, G., Houssein, F.A., Howard, J.S., Hu, A., Previtera, M.J., Phillips, K.M. ve Sedaghat, A.R. (2024). Variability in Minimal Clinically Important Difference Calculation and Reporting in The Otolaryngology Literature. *Laryngoscope*, 134(5), 2059-2069.

Wright, A., Hannon, J., Hegedus, E.J. ve Kavchak, A.E. (2012). Clinimetrics Corner: A Closer Look at The Minimal Clinically Important Difference (MCID). *Journal of Manual & Manipulative Therapy*, 20(3), 160-166.

Zhang, Y., Xi, X. ve Huang, Y. (2023). The Anchor Design of Anchor-Based Method to Determine The Minimal Clinically Important Difference: A Systematic Review. *Health and Quality of Life Outcomes*, 21(1), 74.

GENELLENEBİLİRLİK TEORİSİ

Mine Özşen [**], İlker Ercan [***]

GİRİŞ

Bilimsel araştırmaların temel amacı deneysel olayın açıklanmasını, sayılmasını ve öngörülebilmesini sağlayan genel prensipleri ortaya koymaktır. Bu amacın gerçekleştirilebilmesi verilerin toplanması ve karşılaştırılması ile mümkündür. Bir niteliğin gözlenip, gözlem sonucunun sayı ve sembollerle ifade edilmesine ölçme adı verilir (Aktürk ve Acemoğlu, 2012).

Gerçekleştirilen tüm ölçme işlemlerinde hata yapılma ihtimalinin varlığı bilinen ve göz ardı edilmemesi gereken bir durumdur. Soyut özellikleri ölçen alanlarda gerek ölçülen özellik gerek elde edilen sonuçlar açısından her zaman bir miktar hata payının olduğu unutulmamalıdır. Ağırlıklı olarak somut özelliklerin ölçüldüğü sağlık araştırmalarında da bir miktar hata payı mevcuttur ancak yapılan hatanın tanı ve/veya tedavi etkinliğini belirlemesini, maliyet hesabının yapılmasını, neden-sonuç ilişkisinin kurulmasını, risk faktörlerinin belirlenmesini ve tıbbi kılavuzların oluşturulmasını etkilemesi sebebiyle oldukça dikkatli olunmalıdır. Dikkat edilerek hata oranının en aza indirilmesi gerekliliğinin en önemli nedenlerinden biri de oluşacak hatanın dünya genelinde sağlık hizmeti alan her bireyi etkileme durumudur (Yıldız ve Okyay, 2019).

Ölçme işleminde ölçme aracı ve ölçüm işlemi olmak üzere iki boyut vardır. Ölçme hatalarını en aza indirebilmek kullanılacak aracın güvenilir ve geçerli olması ile mümkündür. Yani kullanılacak araç tesadüfi hatalardan arınmış, farklı zamanlarda

[**] Doç. Dr., Bursa Uludağ Üniversitesi, Tıp Fakültesi, Tıbbi Patoloji Anabilim Dalı, Bursa, mineozsen@uludag.edu.tr, ORCID: https://orcid.org/0000-0002-5771-7649
[***] Prof. Dr., Bursa Uludağ Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Bursa, iercan@msn.com, ORCID: https://orcid.org/0000-0002-2382-290X

ancak aynı koşullarda uygulandığında aynı sonuçları vermeli ve istenilen özelliği tutarlı bir şekilde ölçmelidir. Sağlık bilimlerinde yapılan ölçümlerde güvenirlik, insanların sağlık ve hayatlarına yönelik kararların alınmasında etkili olması nedeniyle önemli bir konudur. Hekimler ve sağlık çalışanları verecekleri kararlarda ölçümlerin ve değerlendirmelerin ne kadar güvenilir olduğuna dair bilgi sahibi olmak isterler ki, bu da kanıta dayalı tıbbın temel unsurlarından biridir (Kartal ve Bardakçı, 2018).

Sağlık alanında ölçme ve değerlendirmenin en önemli olduğu alanlardan biri patolojik değerlendirmedir. Özellikle dakikalar içerisinde karar verilmesi gereken, rutin pratikte yapılandan farklı bir şekilde doku tespiti yapılması nedeniyle artefaktların gözlenebildiği, operasyonun seyri ve operasyon sonrası tedavi protokolleri üzerinde temel yol göstericilerden biri olan intraoperatif konsültasyon (frozen) ölçme ve değerlendirme konusunda yüksek hassasiyet isteyen bir alandır.

Günümüzde tüm dünyada yaygın olarak kullanılan intraoperatif konsültasyon ilk kez 1891 yılında Johns Hopkins Hastanesinde çalışan bir patolog tarafından dondurulmuş kesitin tanıtılması ile kullanıma girmiştir. Santral sinir sistemi lezyonlarında klinik bilginin yanı sıra gelişen teknolojiyle birlikte giderek artan nörogörüntüleme tekniklerine rağmen intraoperatif konsültasyon hala ihtiyaç duyulan bir uygulamadır. Cerrahi esnasında cerraha yönlendirici bilgi sağlayan, cerrahi tedavi planını yönlendiren, örneklenen alanın operasyon öncesi klinikoradyolojik değerlendirmede öngörülen lezyonu temsil edip etmediği konusunda bilgi veren, alınan örneğin yeterliliğini değerlendiren bir yöntem olması nedeniyle santral sinir sistemi lezyonlarının değerlendirmesinde rutin pratikte sıkça kullanılmaktadır (Amraei vd, 2017; Prayson, 2018).

Santral sinir sistemi lezyonlarında intraoperatif konsültasyon, histopatolojik tanının intraoperatif cerrahiyi etkileyeceği,

ameliyat esnasından klinikoradyolojik olarak şüphelenilenden farklı bir lezyonla karşılaşıldığı, biyopsi tanısı ya da tümör derecesi bilinmek istendiği veya cerrahi sınır değerlendirilmesi gereken durumlarda uygulanmaktadır (Kurdi vd., 2022). Hasta yönetiminde sağladığı yararların yanı sıra intraoperatif konsültasyon tekniğinin çeşitli kısıtlılıkları olduğu da bilinmektedir. Bu kısıtlılıklar şu başlıklar altında toplanabilir; lezyonun heterojeniteye sahip olması, cerrahtan kaynaklanan hatalar, patolog tarafından yapılan yorumlama hataları, patoloji teknisyeninden kaynaklanan yetersizlikler, koter, ezilme, kuruma ya da donma artefaktları. Özellikle yüksek dereceli gliomaların lenfoma ve metastatik tümörlerden ayrımı, başta primeri bilinmeyen ya da primer tümör tanısı olmayan olgularda metastatik tümörlerin tanınması, iğsi hücreli ve vaskülaritesi yüksek neoplazilerin ayrımı ve non-neoplastik lezyonlar santral sinir sistemi lezyonlarının intraoperatif değerlendirmesinde problem yaratan konulardır. Gerek intraoperatif konsültasyon tekniğinin sahip olduğu kısıtlılıklar gerekse santral sinir sistemi lezyonlarının içerisinde özellikle deneyim ve bilgi gerektiren, tanıda zorluk yaratabilen lezyonların olması nöropatoloji materyallerinin intraoperatif tanısı ile permanent tanısı arasında uyumsuzluklar olmasına sebep olmaktadır. Literatürde yer alan farklı çalışmalarda tanısal doğruluk oranlarının %78,4 ile %95 arasında değiştiği gözlenmektedir (Zin ve Zulkarnain, 2019; Obeidat vd., 2019). Tanısal doğruluğun arttırmanın yolu tanı hatalarına neden olan durumların iyi anlaşılmasından geçmektedir.

İntraoperatif konsültasyon gibi çoklu hata kaynaklarının bulunabildiği değerlendirme ya da ölçme işlemlerinde çoklu hata kaynaklarının değerlendirilmesine olanak tanıyan istatistiksel yaklaşımlardan biri de genellenebilirlik teorisidir (G teori). Varyans analizi ve klasik test kuramının bir uzantısı olarak düşünülen G teori bir değerlendirme/ölçme işleminden elde

edilen sonuçlardaki tutarsızlığı değerlendirmek amacıyla, modele alınan farklı hata kaynaklarını aynı anda hesaba katarak güvenirlilik bilgisi sağlayan bir kuramdır. G teori tek bir çalışmada potansiyel tüm hata kaynaklarının tek tek veya birbirleriyle etkileşiminden ortaya çıkan hataları değerlendirerek bir kat sayısının elde edilmesine olanak tanır. Elde edilen sonuçlar bilimsel araştırmalar için oldukça önemli bir sorun teşkil eden bu hataların azaltılması için çalışmalar yapılmasına adına bir kaynak oluşturmaktadır (Brennan, 2001; Shavelson, 1991; Brennan, 2011; Güler ve Taşdelen Teker, 2015).

Santral sinir sistemi lezyonlarında intraoperatif konsültasyonun tanısal doğruluk oranları ile ilgili çeşitli çalışmalar mevcuttur ancak literatürde daha önce tek bir çalışmada tüm değişim kaynaklarını yalın etki ve etkileşimden ortaya çıkan değişimi değerlendirmeye olanak sağlayan G teorisinden yararlanılarak yapılan bir güvenirlik değerlendirmesi bulunmamaktadır. Neoplastik primer santral sinir sistemi lezyonlarında intraoperatif ve permanent değerlendirme arasında ne kadar güvenilirlik olduğu, gözlenen uyumsuzlukta hangi değişim/hata kaynağının ne kadar öneme sahip olduğunu araştıran bu çalışmada G teorisinden yararlanılmıştır.

GENEL BİLGİLER

Ölçüm, yalnızca incelenen değişkenin değil, aynı zamanda ölçüm sürecine etki eden çeşitli faktörlerin de etkisi altındadır. İdeal koşullarda, gözlenen farklılıkların yalnızca ölçülen değişkenden kaynaklanması beklenir; ancak ölçüm sürecine etki eden çeşitli faktörlerden dolayı her ölçüm ideal koşullarda gerçekleşmemektedir. Ölçülmek istenen özelliğin gerçek değeri (t_i) ile ölçümden elde edilen değeri (x_i) arasındaki fark olarak ifade edilen ölçme hatası (e_i), ölçüm işleminden, ölçme aracından, ölçme işlemini yapan kişiden, deneklerden ya da ortamdan kaynaklanıyor olabilir. Hata eşitlik (1)'de gösterildiği gibi

gözlenen ölçme sonucu ile gerçek ölçme sonucu arasındaki farktır (Sümbüloğlu ve Sümbüloğlu, 2004).

$$e_i = x_i - t_i \quad (1)$$

Ölçülmek istenen özelliğin gerçek değerinin bilinmesi mümkün olmadığından hatanın büyüklüğü ve yönü de kesin olarak saptanamaz (Öncü, 1994). Bu durum nedeniyle, "gerçek değer" kavramı yerine çoğu zaman "ortalama değer" kavramı kullanılmaktadır. Ortalama değer gerçek değerin tam karşılığı olmasa da, onu en iyi temsil eden ölçüt olarak kabul edilmektedir. Ortalama değerin ölçme hatasının belirlenmesinde kullanılabilmesi için hataların rastgele nitelikte olması gereklidir (Armağan, 1983).

Ölçümde hatanın tamamen ortadan kaldırılması olası değildir. Bu nedenle, hata kaynaklarının ve türlerinin bilinmesi, ölçüm sonuçlarının doğru yorumlanabilmesi açısından önemlidir (Ercan ve Kan, 2006).

Ölçme sonuçlarını etkileyen hata türleri çeşitli faktörlerden kaynaklanabilir. Bu hataların tamamen ortadan kaldırılması mümkün olmasa da en aza indirilmeleri ölçümün güvenilirliğini artırmaktadır. Başlıca hata kaynakları şu şekilde gruplandırılabilir (Ercan ve Kan, 2006).

i. Kişisel Sürekli Faktörler

Bireyin değişmeyen ya da çok az değişen yaş, cinsiyet, zekâ, eğitim geçmişi, bilgi birikimi veya sosyoekonomik düzey gibi özellikleri ölçümde farklılıklara yol açabilir.

ii. Kişisel Geçici Faktörler

Sağlık durumu gibi kısa sürede değişebilen koşullar ölçüm sonuçlarını etkileyebilir.

iii. Çevresel Etkiler

Ölçümün yapıldığı ortamın gürültülü, yetersiz aydınlatmalı veya uygun olmayan koşullara sahip olması sonuçları değiştirebilir. Bu nedenle ölçüm öncesinde çevrenin standart hâle getirilmesi gerekmektedir.

iv. Standart Olmayan Uygulama

Ölçüm sürecinde yönergelerden sapılması (örneğin, görüşmecinin ek sorular sorması, açıklamaları değiştirmesi ya da laboratuvar deneyini farklı koşullarda yürütmesi) sonuçların tutarlılığını bozabilir.

v. Ölçme Aracına İlişkin Etkiler

Kullanılan aracın geçerlik ve güvenirlik açısından yetersizliği, maddelerin belirsiz ifadeler içermesi, farklı kişilerde farklı anlamlar uyandırması ya da soruların sayıca yetersiz kalması ölçüm sonuçlarını olumsuz etkileyebilir.

vi. Mekanik Hatalar

Yanlış işaretleme, okunaksız yazma, bazı soruların atlanması ya da kullanılan cihazdaki arızalar ölçüm hatalarına neden olabilir.

vii. Maddelerin Yetersiz Örneklenmesi

Ölçme aracı, ölçülmek istenen özelliğin tüm boyutlarını kapsayamayabilir. Seçilen maddeler evreni yeterince temsil etmiyorsa ölçüm sonuçlarında sapmalar görülebilir.

viii. Ölçülen Özellikten Kaynaklanan Etkiler

Ölçülmek istenen özellik zaman içinde değişkenlik gösteriyorsa (örneğin ruhsal durum, tutumlar), ölçüm sonuçlarının istikrarlı olması zorlaşır.

İdeal bir ölçümde istenen tüm farklılıkların ölçülen özelliğin birimler arasındaki farklılığından kaynaklanması, diğer faktörlerin etkisinin olmamasıdır. İstatistiki özelliklerine göre hata iki gruba ayrılmaktadır (Ercan ve Kan, 2006);

a. Sistemik Hata: Ölçüm sürecinde elde edilen değerlerin gerçek değerlerden sürekli olarak aynı yönde sapma eğilimi göstermesidir. Sistematik hatalar, ölçümlerin gerçek değerden sürekli olarak daha yüksek veya daha düşük sonuç vermesiyle ortaya çıkar ve sabit ya da değişken karakterde olabilir.

b. Rastgele Hata: Ölçüm sürecinde bilinçsiz rastlantılara bağlı olarak ortaya çıkan sapmalardır. Bu tür hatalar ölçüm sonuçlarının gerçek değer çevresinde dengeli bir dağılım göstermesine sebebiyle, çoklu ölçümlerin ortalaması genellikle gerçek değere yakındır. Çoğu durumda ölçüm sonuçları üzerindeki etkileri ihmal edilebilir düzeydedir.

Ölçme Kuramları

Ölçme kuramlarının temelleri 1950'li yıllarda oluşturulmaya başlanmıştır ancak farklı bilim disiplinlerinin daha yeni yaklaşımlara ihtiyaç duyması ve kuramlarda gözlenen çeşitli eksiklikler araştırma ve gelişmelerin günümüzde de devam etmesini gerektirmiştir. Ölçme kuramları Klasik test kuramı, Madde Tepki Kuramı ve Genellenebilirlik Kuramı şeklinde üç başlık altında özetlenebilir. Tarihsel olarak klasik test kuramının geliştirilmesi ile başlayan bu süreç kuramın eksikliklerinin görülmesi ile birlikte bir sonraki kuramın geliştirilmesine neden olmuştur (Erkuş vd., 2017).

Klasik Test Kuramı

Ölçmede en eski ve sık kullanılan kuramlardan biri olan klasik test kuramı "$x_i = t_i + e_i$" eşitliği ile ifade edilen bir kuramdır. x_i i. deneğin gözlenen puanını, t_i i. deneğin gerçek puanı ve e_i ise i.

deneğin tesadüfi hata puanını temsil etmektedir (Crocker ve Algina, 2008).

Klasik test kuramı, gruba bağımlı oluşu nedeniyle aşağıdaki sınırlılıkları göstermektedir (Akyıldız, & Şahin, 2017):

a. Örneklemdeki tüm bireyler için ölçmenin standart hatası aynıdır.

b. Az sayıdaki madde aracılığıyla yüksek bir güvenirlik değeri elde etmek zordur.

c. Testler eşdeğer olmadığı sürece bir testten elde edilen puanları başka bir testten elde edilen puanlarla karşılaştırmak mümkün değildir.

d. Madde parametrelerinin doğru tahmin edilmesi örneklemin çok iyi seçilmiş olmasıyla mümkündür.

e. Testten elde edilen puanların ne anlam ifade ettiğini sadece puandan çıkarsamak mümkün değildir.

f. Test puanları sıralama ölçeği düzeyindedir.

g. Bir test içinde farklı madde formatlarını kullanmak testten elde edilecek toplam puanda dengesizliklere yol açmaktadır.

h. Ortalama, standart sapma, madde güçlüğü, madde ayırt ediciliği gibi tüm parametreler sadece o grup için geçerlidir.

ı. Her maddenin puan değeri aynıdır.

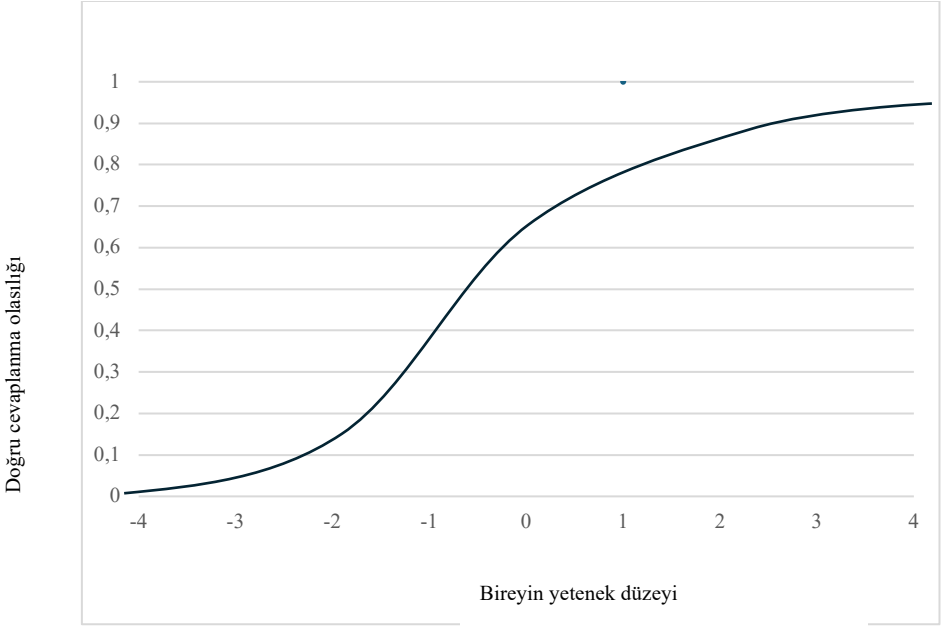
Klasik test kuramının sahip olduğu kısıtlılıklar madde tepki kuramının geliştirilmesine neden olmuştur.

Madde Tepki Kuramı

Klasik test kuramından farklı olarak grup ya da örneklemden bağımsız olarak madde parametrelerinin tahmin edilebilmesi, beceri ölçülerinin uygulanan madde örnekleminden bağımsız olması, her bir maddenin testle ölçülmek istenen beceriler ile eşleştirilerek birey bazında değerlendirilmesi ve testlerin bilgisayar ortamına aktarılarak her bireye kendi yetenek seviyesine uygun madde sunulması gibi yeniliklere sahip olan madde tepki kuramı, bireyin cevabının yetenek düzeyi ile

ilişkisini matematiksel olarak ortaya koyan kuramdır (Akyıldız ve Şahin, 2017).

Bireyin cevabının yetenek düzeyi ile ilişkisi "madde karakteristik eğrisi" adı verilen matematiksel bir açıklama ile ifade edilmektedir. Madde karakteristik eğrisi, madde ile ölçülmeye çalışılan becerinin o maddeye doğru yanıt verme olasılığı üzerindeki regresyonunu ifade eder. Eğri üzerinde y ekseni bir maddenin doğru cevaplanma olasılığını, x ekseni ise bireyin yetenek düzeyini (theta- Θ) ifade eder. Yetenek düzeyi arttıkça doğru cevaplanma olasılığı da artış göstermektedir (Akyıldız ve Şahin, 2017; Baker, 2016)



Şekil 1. Madde Karakteristik Eğrisi

Testi oluşturan maddelerin tek bir özelliği ölçmesi (tek boyutluluk), belirli bir yetenek düzeyine göre maddelere verilen cevapların bağımsız olması (yerel bağımsızlık) ve model ile verinin uyuşması (uyum) madde tepki kuramında sağlanması

gereken varsayımlardır. Doğru kestirim yapılması ancak bu varsayımların sağlanması ile mümkün olmaktadır (DeMars, 2016).

Genellenebilirlik Kuramı

1963 yılında Cronbach L, Rajaratnam N. ve Gleser G tarafından Cronbach'ın iç tutarlılık alfa güvenirliliği üzerine yayımladığı makalesinin bir uzantısı olarak ortaya atılan ve sonrasında çeşitli araştırmacılar tarafından geliştirilen genellenebilirlik kuramı, ortaya çıkan tek bir hata kaynağından ziyade ölçmeye etki eden farklı hata kaynaklarının bir arada incelenmesine olanak sağlayan bir kuramdır (Cronbach, Rajaratnam ve Gleser, 1963; Cronbach, 1951).

Varsayımları sağlandığında güçlü istatistiksel testlerden biri olan varyans analizini (ANOVA) kullanarak kapsamlı bir kavramsal temel oluşturan G kuramı bu özellikleri ile klasik test kuramının bir uzantısı olarak görülmektedir. Klasik test kuramı heterojen yapılar söz konusu olduğunda veya değerlerlendiricilerin varyans ve eğilimleri farklılaştığında sınırlı kalmaktadır. G kuramının klasik test kuramına temel üstünlüğü çeşitli hata kaynaklarını ayırmayı, maddelere, formlara ya da durumlardan kaynaklanan hatalar gibi olası tüm kaynaklarını aynı anda saptayabilmesidir (Atılgan, 2019).

Klasik test kuramının dayandığı model eşitlik (2)'de gösterilmektedir.

$$x_i = t_i + e_i \quad (2)$$

x_i: i. deneğin gözlenen puanını

t_i: i. deneğin gerçek puanı

e_i: i. deneğin tesadüfi hata puanını ifade etmektedir.

G kuramının dayandığı model ise eşitlik (3)'te gösterilmektedir.

$$x_i = \mu_p + e_1 + e_2 + ... + e_K \quad (3)$$

x_i: i. deneğin gözlenen puanını

μ_p: Evren puanı

$e_1, e_2, ... , e_k$: Hata kaynaklarını ifade etmektedir.

Genellenebilirlik kuramının klasik test kuramına üstünlüklerini şu şekilde sıralanabilir;

a. Genellenebilirlik kuramında çoklu hata kaynaklarını tek bir analiz ile değerlendirir.
b. Genellenebilirlik kuramı her bir hata kaynağının ölçme üzerinde ne kadar öneme sahip olduğunu belirler.
c. Klasik test kuramı ile sadece göreli karar verilebilecek bilgi sağlanırken genellenebilirlik kuramı ile G katsayısı ve phi katsayısı olmak üzere iki farklı güvenirlik katsayısı hesaplanabilir. G katsayısı göreli kararlar, phi katsayısı ise mutlak kararlar için hesaplanmaktadır.
d. Genellenebilirlik kuramı en uygun güvenirlik katsayısının elde edilebileceği karar çalışmaları yapılabilmesine olanak sağlar.
e. Genellenebilirlik kuramı sayesinde, yapılması planlanan uygulamalarda etkin çözüm prosedürleri geliştirilebilir (Güler, 2009).

Genellenebilirlik kuramının anlaşılabilmesi için kuramın çerçevesini oluşturan temel kavramların bilinmesi gerekmektedir. Bu kavramlar izleyen başlıklar altında toplanabilir:

a. Yüzey, Koşul ve Evren
b. Ölçme Objesi
c. Genellenebilirlik ve Karar çalışmaları
d. Çaprazlanmış ve İç içe Desen
e. Bağıl ve Mutlak Değerlendirme

f. Tesadüfi ve Sabit Yüzeyler.

a. Yüzey, Koşul ve Evren**:** Genellenebilirlik kuramından ele alınan hata/değişkenlik kaynaklarına "yüzey (facet)" adı verilir. Yüzeyin her bir düzeyi ise koşul olarak adlandırılır. Alınabilecek tüm koşullardan elde edilen ölçme sonuçları evreni "kabul edilebilir gözlemler evreni" olarak, araştırmacının genellemek istediği koşulların seti ise "genelleme evreni" olarak tanımlanır. Kavramları örnekler ile açıklamak gerekirse, değerlendiricilerin maddeleri puanladıkları bir ölçme işleminde "değerlendiriciler ve maddeler" yüzeyi temsil ederken, her bir değerlendirici ise koşulu temsil etmektedir. Olası herhangi bir madde ile olası herhangi bir değerlendirici içerebilen evren kabul edilebilir gözlemler evreni iken, olası tüm durumlar genelleme evrenini oluşturmaktadır (Atılgan, 2019 ; Güler vd., 2012).

b. Ölçme objesi: Genellenebilirlik kuramında popülasyon terimi ile ifade edilen ölçme objesi istenilen kararların alınacağı ölçmenin hedefi durumundadır. Çoğu araştırmada ölçme objesi bireylerdir ancak bu bir zorunluluk değildir. Farklı değişkenler de ölçme objesi olabilir. Örnekle açıklamak gerekirse, çoktan seçmeli yirmi sorunun yer aldığı bir sınavda sınava girecek olan her bir birey ölçme objesidir. Aynı sınavda sınavın zorluk derecesini belirlemek adına sorunlara 1 ile 5 arasında bir puan verildiğinde ise ölçme objesi konumuna sorular geçmektedir.

Bireylerin doğası gereği var olan bir durum olan bireyler arası farklılıklar klasik test kuramında hata kaynağı olarak ele alınmamaktadır. Benzer durum genellenebilirlik kuramı içinde geçerlidir. Bireyler arası farklılıklar genellenebilirlik kuramında da yüzey (değişkenlik kaynağı) olarak kabul edilmez (Crocker ve Algina, 2008; Güler vd., 2012).

c. Genellenebilirlik ve Karar çalışmaları: Ölçme yapılan örneklemi ölçmenin evrenine genellemek amacıyla, olası tüm değişkenlik kaynaklarını inceleyerek hata kaynaklarını belirleyen

ve etkilerini ortaya koyarak kabul edilebilir gözlemler evrenini tanımlayan çalışmalar genellenebilirlik çalışmaları olarak tanımlanır. Genellenebilirlik çalışmaları kapsamında; testin kararlılığı (farklı zamanlarda testten elde edilen cevapların aynı kalması), ölçme aracının farklı formlarındaki puanların tutarlılığı, alt ölçek ya da maddeler arasındaki ilişkiler araştırılabilir. Ölçme konusunda karar verebilmek adına ölçmedeki varyans kaynakları hakkında bilgi sağlaması ve değişkenlik kaynaklarından gelen hataları belirleyerek ve azaltılması adına çalışmalar yapılması da Genellenebilirlik çalışmalarının konusudur.

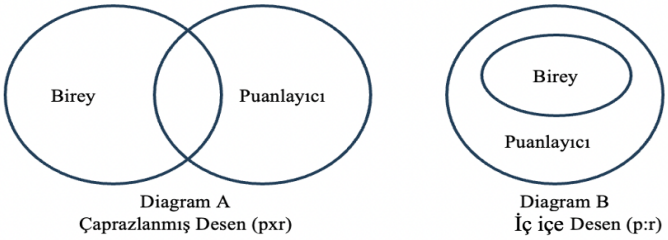
Genellenebilirlik çalışmalarında varyans bileşenlerinin kestirimi temel amaçken, karar çalışmalarında varyans bileşenlerinin kestirimi, kullanımı ve yorumlanması söz konusudur. Karar çalışmalarında belirli bir amaç için karar vermek adına yeni veriler toplanabilir ancak Genellenebilirlik çalışmasından elde edilen verilerden yararlanmak da pratik uygulamada tercih edilebilen bir durumdur. Bireyleri tanımlayarak doğru bir seçme ya da yerleştirme yapılması, farklı grupların karşılaştırılması ya da değişkenler arasındaki ilişkinin belirlenmesi amacıyla Karar çalışmaları yapılabilmektedir. Karar çalışmalarında yapılan ölçmenin en uygun desenine ulaşılması istenmektedir; ancak bir çalışmayı sadece desenine bakarak Karar ya da Genellenebilirlik çalışması olarak tanımlamak mümkün değildir. Çalışmanın tipini belirleyen temel unsur araştırmanın amacıdır. Örneğin; bir ilde devlet okulları ile özel okullarda eğitim gören eşit sayıdaki öğrenciye uygulanan standart bir başarı testinde amaç testin her iki grup içinde eşit güvenirlilikte olup olmadığını belirlemek ise bu çalışma Genellenebilirlik çalışmasıdır. Amaç gruplar arasındaki başarı seviyelerinin ortalamasını karşılaştırmak şeklinde belirlenirse de çalışma Karar çalışması olacaktır (Brennan, 1992).

d. Çaprazlanmış ve İç içe Desen: Çalışma desenleri çaprazlanmış (crossed) ve iç içe (nested) olmak üzere iki başlık

altında toplanır. Kullanılan desene göre analizlerde farklılıklar ortaya çıkmaktadır.

Bir yüzeyin tüm koşullarının diğer bir yüzeyin tüm koşulları ile örtüştüğü durumlarda desen çaprazlanmış desen iken, bir yüzeyin tüm koşullarının diğer bir yüzeyin bazı koşulları ile örtüştüğü durumlar iç içe desendir. Çaprazlanmış desende değişkenlik kaynaklarının arasına "x" işareti konulur ve desen "b x m x p" şeklinde gösterilir. İç içe desende ise değişkenlik kaynaklarının arasına ":" işareti konularak desen "p : m : b" ifadesi ile gösterilmektedir. p, m ve b ifadeleri madde, durum, zaman, puanlayıcı gibi değişkenlik kaynaklarını temsil eder. Örnek üzerinden açıklanacak olursa birden fazla sayıda puanlayıcının bireylerin maddelere verdikleri yanıtlar üzerinden değerlendirme yaptıkları bir durumda "b" bireyleri, "m" maddeleri ve "p" puanlayıcıları temsil etmektedir. Çaprazlanmış desenlerden elde edilen veriler eşleştirilmiş veri olarak adlandırılırken iç içe desenden elde edilen veriler bağımsız verilerdir.

Mümkün tüm yüzeyler ve bu yüzeyler arasındaki etkileşimlerden kaynaklanan hataların kestirimlerine imkân sağlamaları nedeniyle çalışmalarda genellikle tümüyle çaprazlanmış desenler tercih edilmektedir. Ancak bir desen içerisinde bazı yüzeylerin çaprazlanmış bazılarınınsa iç içe olması da mümkündür (Shavelson ve Webb, 1991; Güler vd., 2012).



Şekil 2. Tek Yüzeyli Çaprazlanmış ve İç İçe Desen Venn Şeması Örneği

(Shavelson, & Webb, 1991).

e. Bağıl ve Mutlak Değerlendirme: G teorisi araştırmacıların hem bağıl değerlendirme hem de mutlak değerlendirme yapabilmesine imkân sağlamaktadır. Bağıl değerlendirmede bireyin ölçümden elde ettiği sonuçlara bağlı olarak ölçmenin uygulandığı grup içerisinde birey sıralanırken, mutlak değerlendirmede bireyin değerlendirme sonucu daha önce belirlenmiş bir ölçütle kıyaslanmaktadır. Bir başka ifade ile mutlak değerlendirme gruptan bağımsızken, bağıl değerlendirme gruptan yola çıkarak belirlenen bir ölçüte bağlıdır. Örneğin sınav geçme notu 70 olarak belirlenmiş bir fakültede yapılan bu değerlendirme mutlak değerlendirme iken sınava giren tüm öğrencileri aldıkları puanlar doğrultusunda sıralamaya koyan bir değerlendirmede kişinin başarısı diğer sınava giren kişilerin başarılarına bağlı olarak değişmekte ve bu durum bağıl değerlendirme olarak ifade edilmektedir.

Bağıl ve mutlak değerlendirme genellenebilirlik katsayısı (G veya Ep^2) ve güvenirlik katsayısı (Φ veya "phi") olmak üzere iki farklı şekilde hesaplanmaktadır. Genellenebilirlik katsayısı yapılan ölçümlerden elde edilen puanların evren puanına genellenmesindeki doğruluk oranını vermektedir. Değerleri 0 ile 1 arasında değişen G katsayısında kabul edilebilir değer 0,80 ve üzeridir (Shavelson ve Webb, 1991; Cardinet vd., 2010).

Genellenebilirlik katsayısının hesaplandığı formül eşitlik (4-14)'te gösterilmektedir (Dimitrov, 2006; Ercan vd., 2008).

$$\sigma^2_{crt,e} = MS_{crt,e} \quad (4)$$

c: Olgu

r: Değerlendirici

t: Teknik

e: Diğer hata kaynaklarını ifade etmektedir.

$MS_{crt,e}$: Üçlü etkileşimden ve modele dahil edilmeyen hata terimlerine karşılık gelen kareler ortalaması

$$\sigma^2_{cr} = (MS_{cr} - MS_{crt,\,e}) / n_t \quad \ldots\ldots(5)$$

σ^2_{cr}: Olgu ve değerlendirici etkileşiminden ortaya çıkan varyans,

MS_{cr}: Olgu ve değerlendirici etkileşiminden elde edilen kareler ortalaması,

n_t: Teknik sayısını ifade etmektedir.

$$\sigma^2_{ct} = (MS_{ct} - MS_{crt,\,e}) / n_r \quad (6)$$

σ^2_{ct}: Olgu ve teknik etkileşiminden ortaya çıkan varyans

MS_{ct} : Değerlendirici ve teknik etkileşiminden elde edilen kareler ortalaması

n_r: Değerlendirici sayısını ifade etmektedir.

$$\sigma^2_{rt} = (MS_{rt} - MS_{crt,\,e}) / n_c \quad (7)$$

σ^2_{rt} : Değerlendirici ve teknik etkileşiminden ortaya çıkan varyans.

MS_{rt} : Değerlendirici ve teknik etkileşiminden elde edilen kareler ortalaması

n_c : Olgu sayısını ifade etmektedir.

Yukarıda verilen varyans bileşenleri G katsayısının hesaplanması için yeterlidir. Bununla birlikte, iki yüzlü çapraz tasarım modelinde diğer varyans bileşenleri de hesaplanmalıdır. Bu varyans bileşenleri aşağıdaki gibi tahmin edilir:

$$\sigma^2_{c} = (MS_c - MS_{cr} - MS_{ct} + MS_{crt,\,e}) / n_r\, n_t \quad (8)$$

$$\sigma^2_r = (MS_r - MS_{cr} - MS_{rt} + MS_{crt,\,e}) / n_c\, n_t \quad (9)$$

$$\sigma^2_t = (MS_t - MS_{ct} - MS_{rt} + MS_{crt,\,e}) / n_c\, n_r \quad (10)$$

σ^2_c : Olgu varyansı

σ^2_r : Değerlendirici varyansı

σ^2_t: Teknik varyansını ifade etmektedir.

Göreli ve mutlak hata varyansları aşağıdaki gibi tahmin edilir:

$$\sigma^2_{Rel} = (\sigma^2_{cr}/n_r) + (\sigma^2_{ct}/n_o) + (\sigma^2_{crt,e}/n_r\, n_t) \quad (11)$$

$$\sigma^2_{Abs} = (\sigma^2_r/n_r) + (\sigma^2_t/n_t) + (\sigma^2_{rt}/n_r\, n_t) + \sigma^2_{Rel} \quad (12)$$

σ^2_{Rel} : Göreli hata varyansı

σ^2_{Abs} : Mutlak hata varyansını ifade etmektedir.

Genellenebilirlik katsayısı (G) ve güvenirlik katsayısı (phi) katsayıları aşağıdaki gibi tahmin edilir:

$$G = \sigma^2_c / (\sigma^2_c + \sigma^2_{Rel}) \quad (13)$$

Değerleri 0-1 aralığında değişen "phi" katsayısında G katsayısından farklı olarak bağıl hata varyansı yerine mutlak hata varyansı kullanılır. Phi katsayısı için eşitlik (14)'te yer alan formül ile hesaplanmaktadır.

$$\Phi = \sigma^2_c / (\sigma^2_c + \sigma^2_{Abs}) \quad (14)$$

f. Tesadüfi ve sabit yüzeyler: Araştırmacının kararı doğrultusunda seçilen örneklemin büyüklüğüne bağlı olarak tesadüfi ve sabit yüzeyden bahsedilebilir. Örneklem evrenden rastgele seçilmişse, evrenden alınan örneklem aynı büyüklükteki başka örneklemler ile değiştirilebiliyorsa ya da örneklem büyüklüğü evren büyüklüğünden çok küçük ise bu örneklemler "rastgele yüzeyler" olarak adlandırılır.

Sabit yüzeylerde ise evrenin küçük olmasına bağlı olarak evren üzerinde çalışılması, seçilen koşullar haricinde genelleme yapılmaması yani araştırmacının genelleme yapmak amacı içerisinde olduğu durumlarda kullanılır *(*Shavelson vd., 1991 ; Güler vd., 2012).

Genellenebilirlik Kuramının Dayandığı İstatistiksel Model

Genellenebilirlik kuramı birey puanı, evren puanı ve etkisi, değişkenlerin/hataların etkisi ve değişken/hataların etkileşiminin etkilerini varyans analizi kullanarak ayrıştıran ve bu sayede değişkenlik kaynaklarının miktarlarını belirlemeye çalışan bir yöntemdir.

Varyanslara bağlı çalışan bir metot olan varyans analizinde örneğin kişilerin başarı puanlarındaki değişkenlik, birey etkisi, madde etkisi, birey-madde etkileşimi olan artık, sistematik ve tesadüfi hatalar olarak ele alınabilir. Bir bireyin tek yüzeyli çaprazlanmış desende puanını etkileyen dört bileşen söz konusudur. Bunlar; tüm bireyler için sabit olan genel ortalama, bireyin evren puanı ile genel ortalama arasındaki farkı ifade eden birey etkisi, maddenin güçlük düzeyini gösteren madde etkisi ve birey ile maddenin etkileşiminin etkisini olan artık etkisidir.

Çok yüzeyli desenlerde analize ek değişkenlik kaynakları da dahil olmaktadır. Örneğin yukarıdaki örnek tekrar ele alınırsa, kişilerin başarı puanlarını farklı değerlendiricilerin değerlendirdiği varsayılırsa desen çok yüzeyli desen haline gelir. Çok yüzeyli desenlerde tek yüzeyli ölçmelerde yer alan genel ortalama, birey etkisi, madde etkisi ve artık etkisine ilaveten değerlendirici etkisi, birey-madde etkisi, birey-değerlendirici etkisi, madde-değerlendirici etkisi de gözlenen puanın bileşenleri arasında yer alır (Shavelson vd., 1991).

Tek Yüzeyli Evrenler ve Desenler

Tek yüzeyli evrende genellenebilirlik ve karar çalışması için çaprazlanmış ve iç içe olmak üzere yapılabilecek iki desen mevcuttur. "b" sembolü bireyleri ve "m" sembolünün maddeleri ifade ettiği kabul edilirse genellenebilirlik çalışmaları için çaprazlanmış ve iç içe desen sırasıyla "bxm" ve "b:m" şeklinde ifade edilirken, karar çalışmaları için bu ifadeler sırasıyla "bxM" ve "b:M" şeklinde gösterilir.

Tek yüzeyli çaprazlanmış desende bir bireyin bir madde üzerinde gözlenen puanına ait bileşenler eşitlik (15)'da gösterilmektedir (Atılgan, 2019 ; Güler vd., 2012).

$X_{bm} = \mu$ Genel ortalama (15)

$X_{bm} = \mu + (\mu_b - \mu)$ Birey etkisi

$X_{bm} = \mu + (\mu_m - \mu)$ Madde etkisi

$X_{bm} = \mu + (X_{bm} - \mu_b - \mu_m)$ Artık etkisi

Tek yüzeyli iç içe desende bir bireyin bir madde üzerinde gözlenen puanına ait bileşenler ise eşitlik (16)'de ifade edilmiştir.

$X_{mb} = \mu$ Genel ortalama (16)

$X_{mb} = \mu + (\mu_b - \mu)$ Birey etkisi

$X_{mb} = \mu + (X_{mb} - \mu_b)$ Artık etkisi

Madde etkisinin ayrı bir bileşen olarak yer almaması ve iç içe desende artık etkisinin çaprazlanmış desenden farklı olması tek yüzeyli evrenlerde iç içe ve çaprazlanmış desenlerin temel farkıdır (Güler vd., 2012).

Tek yüzeyli tasarımlarda, Genellenebilirlik Teorisi ile elde edilen G katsayısı, Cronbach alfa katsayısı ile özdeştir ve aynı veriler üzerinde yapılan hesaplamalarda iki katsayı da aynı değeri vermektedir. Bu eşdeğerlik yalnızca tek yüzeyli evrenler için geçerlidir; çok yüzeyli tasarımlarda ise genellenebilirlik teorisi farklı değişim kaynaklarını da değerlendirmeye aldığından G katsayısı Cronbach alfa ile farklı değerler almaktadır (Dimitrov, 2006).

İki Yüzeyli Evrenler ve Desenler

Genellenebilirlik analizleri, varyans bileşenlerinin kestirildiği genellenebilirlik çalışmalarının yer aldığı birinci aşama ve kestirilen varyans bileşenlerini kullanarak elde edilen evren puan varyansı, bağıl ve mutlak hata, genellenebilirlik katsayısı ve güvenirlik katsayılarının hesaplandığı karar çalışmalarının yer aldığı ikinci aşama olmak üzere iki aşamadan oluşmaktadır.

İki yüzeyli desenlerde birey, madde ve puanlayıcının ortalamaları desende yer alan ana etkileri oluştururken ana etkiler dışında kalan tüm etkiler etkileşim etkisini oluşturmaktadır. İki yüzeyli desenlerde çalışmanın içeriğine göre değişkenlik kaynakları farklılık gösteren başlıca altı tip desen bulanmaktadır (Shavelson ve Webb, 1991). Örneğin değişkenlik kaynakları birey (b), madde (m) ve puanlayıcı (p) olan iki yüzeyin yer aldığı bir çalışmada desenler şu şekildedir;

i. b x m x p; tüm değişkenlik kaynaklarının birbiriyle çaprazlandığı desendir. Tümüyle çaprazlanmış desen olarak da isimlendirilir. Tümüyle çaprazlanmış desenlerin gösteriminde yüzeylerin sıralanışı önem arz etmemektedir. Yani "b x m x p" gösterimi ile "b x p x m" gösterimi aynı deseni ifade etmektedir.

ii. b x (m : p); maddelerin puanlayıcılar içinde yuvalandığı ve bireylerin madde ve puanlayıcı yüzeyleri ile çaprazlandığı desendir.

iii. (m : b) x p; maddelerin bireyler içinde iç içe yer aldığı ve maddeler ile bireylerin puanlayıcı yüzeyi ile çaprazlandığı desendir.

iv. m : (b x p); bireyler ile puanlayıcıların çaprazlandığı ve maddelerin birey-puanlayıcı kombinasyonu ile yuvalandığı desendir.

v. (m x p) : b; maddeler ile puanlayıcıların çaprazlandığı ve her iki yüzeyin de bireyler ile yuvalandığı desendir.

vi. m : p : b; maddeler, puanlayıcılar ve puanlayıcılar da bireyler içinde iç içedir. Tümüyle iç içe desen olarak da isimlendirilir. Tümüyle iç içe desenler değişkenlik kaynakları hakkında en az bilginin elde edildiği desenlerdir.

SONUÇ VE ÖNERİLER

Genellenebilirlik teorisi, ölçme süreçlerinde ortaya çıkabilecek farklı hata kaynaklarını aynı anda değerlendirebilmesi nedeniyle klasik güvenirlik yaklaşımlarına göre daha kapsamlı bir yöntem sunmaktadır. Yalnızca genel bir güvenirlik katsayısı vermekle kalmayıp, farklı hata kaynaklarının ölçüm sonuçları üzerindeki etkilerini ve bu kaynakların toplam değişkenliğe katkılarını ortaya koyması önemli bir avantajdır. Bu özelliği sayesinde araştırmacılara ölçme sürecinin hangi aşamalarında iyileştirmeye ihtiyaç duyulduğu konusunda ayrıntılı

bilgi sağlamaktadır. Özellikle birden fazla değerlendirici, ölçüm tekniği veya uygulama koşulunun bulunduğu araştırmalarda ölçümlerin güvenirliğinin daha doğru ve kapsamlı biçimde değerlendirilmesine olanak tanımaktadır. Bu nedenle genellenebilirlik teorisi, ölçme süreçlerinin değerlendirilmesi ve hata kaynaklarının azaltılmasına yönelik stratejilerin geliştirilmesinde önemli bir istatistiksel yöntem olarak öne çıkmaktadır.

KAYNAKÇA

Aktürk, Z. ve Acemoğlu, H. (2012). Tıbbi Araştırmalarda Güvenilirlik ve Geçerlilik. *Dicle Tıp Dergisi*, 39(2), 316-319.

Akyıldız, M. ve Şahin, M. D. (2017). Açıköğretimde Kullanılan Sınavlardan Klasik Test Kuramına ve Madde Tepki Kuramına Göre Elde Edilen Yetenek Ölçülerinin Karşılaştırılması. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, 3(4), 141-159.

Aldape, K., Simmons, M. L., Davis, R. L., Miike, R., Wiencke, J. K., Barger, G. ve Wrensch, M. (2000). Discrepancies in Diagnoses of Neuroepithelial Neoplasms: The San Francisco Bay Area Adult Glioma Study. *Cancer*, 88(9), 2342-2349.

Amraei, R., Moradi, A., Zham, H., Ahadi, M., Baikpour, M. ve Rakhshan, A. (2017). A Comparison Between the Diagnostic Accuracy of Frozen Section and Permanent Section Analyses in Central Nervous System. *Asian Pacific Journal of Cancer Prevention*, 18(3), 659-666.

Armağan, İ. (1983). *Yöntembilim-2: Bilimsel Araştırma Yöntemleri*. İzmir: Dokuz Eylül Üniversitesi Güzel Sanatlar Fakültesi Yayınları.

Atılgan, H. (2019). *Genellenebilirlik Kuramı ve Uygulaması*. Ankara: Anı Yayıncılık.

Baker, F. B. (2016). *Madde Tepki Kuramının Temelleri* (çev. N. Güler ve M. İlhan). Ankara: Pegem Akademi.

Brennan, R. L. (1992). *Elements of Generalizability Theory*. Iowa City: American College Testing.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, 24(1), 1-21.

Cardinet, J., Johnson, S. ve Pini, G. (2010). *Applying Generalizability Theory Using EduG*. New York: Routledge.

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16(3), 297-334.

Cronbach, L. J., Rajaratnam, N. ve Gleser, G. C. (1963). Theory of Generalizability: A Liberalization of Reliability Theory. *British Journal of Statistical Psychology*, 16(2), 137-163.

Crocker, L. ve Algina, J. (2008). *Introduction to Classical and Modern Test Theory*. Mason, OH: Cengage Learning.

DeMars, C. (2016). *Madde Tepki Kuramı*. İçinde E. H. Özberk ve H. Kelecioğlu (Editörler), Madde Tepki Kuramı (s. 1-30). Ankara: Nobel Akademik Yayıncılık.

Dimitrov, D. M. (2006). *Reliability*. Boston: Houghton Mifflin.

Doğan, İ. ve Doğan, N. (2019). Ölçek Geliştirmede Kullanılan İçerik Geçerliği Üzerine Genel Bir Bakış. *Türkiye Klinikleri Journal of Biostatistics*, 11(2), 143-151.

Ercan, İ. ve Kan, İ. (2006). Ölçme ve Ölçmede Hata. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, 7(1), 51-56.

Ercan, İ., Ocakoğlu, G., Güney, İ. ve Yazıcı, B. (2008). Adaptation of Generalizability Theory for Inter-Rater Reliability for Landmark Localization. *International Journal of Tomography and Statistics*, 9(S08), 51-58.

Erkuş, A., Sünbül, Ö., Sünbül, S. Ö., Yormaz, S. ve Aşiret, S. (2017). *Psikolojide Ölçme ve Ölçek Geliştirme-II*. Ankara: Pegem Akademi.

Güler, N. (2009). Genellenebilirlik Kuramı ve SPSS ile GENOVA Programlarıyla Hesaplanan G ve K Çalışmalarına İlişkin Sonuçların Karşılaştırılması. *Eğitim ve Bilim*, 34(154), 93-103.

Güler, N., Kaya Uyanık, G. ve Taşdelen Teker, G. (2012). *Genellenebilirlik Kuramı*. Ankara: Pegem Akademi.

Güler, N. ve Taşdelen Teker, G. (2015). Açık Uçlu Maddelerde Farklı Yaklaşımlarla Elde Edilen Puanlayıcılar Arası Güvenirliğin Değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 6(1), 12-24.

Kartal, M. ve Bardakçı, S. (2018). *SPSS ve AMOS Uygulamalı Örneklerle Güvenirlik ve Geçerlik Analizleri*. Ankara: Akademisyen Kitabevi.

Kurdi, M., Baeesa, S., Maghrabi, Y., Bardeesi, A., Saeedi, R., Al-Sinani, T. ve Hakamy, S. (2022). Diagnostic Discrepancies Between Intraoperative Frozen Section and Permanent Histopathological Diagnosis of Brain Tumors. *Turkish Journal of Pathology*, 38(1), 34-39.

Obeidat, F. N., Awad, H. A., Mansour, A. T., Hajeer, M. H., Al-Jalabi, M. A. ve Abudalu, L. E. (2019). Accuracy of Frozen-Section Diagnosis of Brain Tumors: An 11-Year Experience From a Tertiary Care Center. *Turkish Neurosurgery*, 29(2), 242-246.

Öncü, H. (1994). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Matser Basım.

Prayson, R. A. (2018). Accuracy of Frozen Section in Determining Meningioma Subtype and Grade. *Annals of Diagnostic Pathology*, 35, 7-10.

Rathore, S., Chaddad, A., Iftikhar, M. A., Bilello, M. ve Abdulkadir, A. (2021). Combining MRI and Histologic Imaging Features for Predicting Overall Survival in Patients With Glioma. *Radiology: Imaging Cancer*, 3(4), e200108.

Scott, C. B., Nelson, J. S., Farnan, N. C., Curran, W. J., Murray, K. J., Fischbach, A. J. ve Nelson, D. F. (1995). Central Pathology Review in Clinical Trials for Patients With Malignant Glioma: A Report of Radiation Therapy Oncology Group 83-02. *Cancer*, 76(2), 307-313.

Shavelson, R. J. ve Webb, N. M. (1991). *Generalizability Theory: A Primer*. Thousand Oaks: Sage Publications.

Shavelson, R. J., Webb, N. M. ve Rowley, G. L. (1991). *Generalizability Theory*. in R. L. Linn (Editör), *Educational Measurement* (s. 233-256). New York: Macmillan.

Sümbüloğlu, V. ve Sümbüloğlu, K. (2004). *Sağlık Bilimlerinde Araştırma Yöntemleri*. Ankara: Hatiboğlu Yayınları.

Wang, X., Wang, R., Yang, S., Zhang, J., Wang, M., Zhang, D. ve Han, X. (2022). Combining Radiology and Pathology for Automatic Glioma Classification. *Frontiers in Bioengineering and Biotechnology*, 10, 841958.

Yıldız, F. ve Okyay, P. (2019). Sağlık Araştırmalarında Yan Tutma (Bias) ve Yan Tutmanın Değerlendirilmesi. *ESTÜDAM Halk Sağlığı Dergisi*, 4(2), 219-231.

Zin, A. A. M. ve Zulkarnain, S. (2019). Diagnostic Accuracy of Cytology Smear and Frozen Section in Glioma. *Asian Pacific Journal of Cancer Prevention*, 20(2), 321-325.

k × k BOYUTLU EŞLEŞTİRİLMİŞ KATEGORİK VERİLERİN ANALİZİNDE KULLANILAN İSTATİSTİKSEL YÖNTEMLER

Hatice Ortaç[**], Gökhan Ocakoğlu[***]

GİRİŞ

Eşleştirilmiş kategorik veri yapıları; aynı gözlem biriminin farklı zaman noktalarında, ardışık ölçümlerde veya bağımsız iki farklı yöntem/değerlendirici tarafından sınıflandırıldığı araştırma tasarımlarında temel bir rol oynamaktadır. Tıbbi ve biyolojik araştırmalarda tedavi öncesi ve sonrası evre değişimlerinin incelenmesi, tanısal testlerin uyumunun değerlendirilmesi veya boylamsal çalışmalarda kategorik durum geçişlerinin modellenmesi gibi pek çok araştırma probleminde asıl amaç, iki ölçüm arasındaki marjinal dağılımların istatistiksel olarak anlamlı bir farklılık gösterip göstermediğini test etmektir. İkili (binary) yanıt değişkenine sahip 2×2 boyutlu temel tasarımlarda, bu amaçla McNemar (1947) testi altın standart olarak yaygın bir şekilde kullanılmaktadır. Ne var ki, klinik ve epidemiyolojik uygulamaların birçoğunda yanıt değişkeni ikiden fazla sıralı (ordinal) veya sırasız (nominal) kategori barındırmaktadır (örn. hastalık evreleri, memnuniyet düzeyleri veya çoklu sınıflandırma kriterleri). Bu tür $k \times k$ boyutlu kare kontenjans tablolarında klasik McNemar testi yetersiz kalmakta; dolayısıyla marjinal homojenlik, simetri ve yarı-simetri (quasi-symmetry) hipotezlerini eşzamanlı ve esnek bir şekilde sınayabilen daha genel istatistiksel yaklaşımlara ihtiyaç duyulmaktadır.

[**] Doktora öğrencisi, Bursa Uludağ Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalı, Bursa, Türkiye, haticeortac21@gmail.com, ORCID: orcid.org/0000-0002-1199-1706

[***] Prof. Dr., Bursa Uludağ Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı, Bursa, Türkiye, ocakoglu@uludag.edu.tr, ORCID: orcid.org/0000-0002-1114-6051

Bu kısıtlılıkları aşmak amacıyla literatürde geliştirilen temel istatistiksel yaklaşımlar, ağırlıklı olarak marjinal homojenlik (marginal homogeneity) ve simetri (symmetry) hipotezleri etrafında şekillenmektedir. Marjinal homojenlik hipotezi, karesel tablonun satır ve sütun marjinal olasılık dağılımlarının birbirine denk olup olmadığını sınarken; simetri hipotezi, köşegen dışı (off-diagonal) hücre frekanslarının (n_{ij} ve n_{ji}, $i \neq j$ olmak üzere) karşılıklı olarak birbirine eşitliğini zorunlu kılan çok daha katı bir yapısal koşulu test etmektedir. Bishop vd. (1975) tarafından literatüre kazandırılan temel teorik çerçeveye göre tam simetri (complete symmetry) modeli; yarı-simetri (quasi-symmetry) ve marjinal homojenlik koşullarının eşzamanlı olarak sağlanmasına eşdeğerdir. Bu teorik ayrıştırma (decomposition), $k \times k$ boyutlu eşleştirilmiş verilerin analizinde tercih edilecek spesifik test istatistiklerinin (örneğin Bowker, Stuart-Maxwell veya Bhapkar testleri) hiyerarşik ilişkisini ve model seçimi dinamiklerini anlamak açısından kritik bir zemin oluşturmaktadır.

McNemar (1947), yalnızca ikili (binary) yanıt kategorisine sahip eşleştirilmiş verilerde marjinal değişimi test eden yalın ve etkin bir istatistiksel yöntem önermiştir. Bowker (1948), bu temel yaklaşımı $k \times k$ boyutlu çok kategorili tablolara genişleterek, tam simetri (complete symmetry) hipotezini sınayan ilk genelleştirilmiş çözümü literatüre kazandırmıştır. İlerleyen süreçte Stuart (1955), marjinal homojenlik hipotezine odaklanarak McNemar testinin doğrudan bir genellemesi niteliğinde olan skor tipi (score-type) bir test istatistiği önermiş; Maxwell (1970) ise bu yaklaşımı bağımsız olarak ele alıp geliştirerek uygulamalı istatistik literatüründe yaygınlaşmasını sağlamıştır (Stuart-Maxwell testi). Metodolojik ilerleme sürecinin kritik bir dönüm noktası olarak Caussinus (1965), simetri ile marjinal homojenlik arasındaki matematiksel ilişkiyi log-lineer modeller çerçevesinde yarı-simetri (quasi-symmetry) kavramı üzerinden temellendirirken; Bhapkar (1966), marjinal

homojenlik hipotezini Wald tipi bir test formülasyonuyla sınayan güçlü bir alternatif yapı sunmuştur. Bishop vd. (1975) ise Caussinus'un (1965) ortaya koyduğu "tam simetri = yarı-simetri + marjinal homojenlik" ayrıştırmasını (decomposition) çok yönlü tablolara entegre ederek, bu farklı yaklaşımları tutarlı bir teorik çerçevede birleştirmiştir. Bu kronolojik metodolojik seyir, ele alınan istatistiksel testlerin birbirinden yalıtılmış bağımsız çözümler olmaktan ziyade, eşleştirilmiş kategorik veri analizinin temelini oluşturan kuramsal çerçevenin birbirini izleyen ve tamamlayan yapıtaşları olduğunu açıkça ortaya koymaktadır.

Bu bölümde; $k \times k$ boyutlu eşleştirilmiş kategorik verilerin analizinde temel dayanak noktası oluşturan başlıca istatistiksel yöntemler sistematik olarak ele alınmaktadır. Bu kapsamda sırasıyla; marjinal homojenlik hipotezini sınayan Stuart-Maxwell ve Bhapkar testleri, tam simetri koşulunu değerlendiren Bowker testi ve son olarak yarı-simetri (quasi-symmetry) ile tam simetri log-lineer modellerinin karşılaştırılmasına dayanan Olabilirlik Oranı Testi (Likelihood Ratio Test - LRT) tabanlı Marjinal Homojenlik (LRT-MH) yaklaşımı teorik ve pratik boyutlarıyla detaylandırılmaktadır.

Stuart-Maxwell Testi

Stuart-Maxwell testi, k × k boyutlu eşleştirilmiş kontenjans tablolarında marjinal dağılımların homojenliğine ilişkin varsayımı değerlendirmek amacıyla önerilmiş, asimptotik teoriye dayalı bir istatistiksel yöntemdir (Stuart, 1955). McNemar (1947) testinin ikiden fazla kategoriye genelleştirilmiş biçimi olarak kabul edilen bu test, Maxwell'in (1970) de aynı yöntemi ele alıp yaygınlaştırmasıyla literatürde "Stuart-Maxwell testi" olarak adlandırılmıştır. Testin temel amacı, her bir kategorideki satır ve sütun marjinallerinin birbirine eşit olup olmadığını değerlendirmektir. Tabloda yer alan gözlem frekansları n_{ij} ile

gösterilmek üzere, her bir kategoriye ilişkin fark değeri Eşitlik (1)'de tanımlanmaktadır.

$$d_i = n_{i.} - n_{.i} \quad (i = 1, 2, \ldots, k) \tag{1}$$

d_i : i. kategoriye ilişkin satır ve sütun marjinal frekansları arasındaki farkı,

n_i : i. satırın toplam frekansını (ilk ölçümde i. kategoride yer alan birey sayısını),

$n_{.i}$: i. sütunun toplam frekansını (ikinci ölçümde i. kategoride yer alan birey sayısını),

k : kategori sayısıdır.

Marjinal homojenlik sıfır hipotezi altında bu farklara ait varyans ve kovaryans değerleri sırasıyla Eşitlik (2a) ve Eşitlik (2b)'de ifade edilmiştir.

$$V(d_i) = n(p_{i.} + p_{.i} - 2p_{ii}) \tag{2a}$$

$$Cov(d_i, d_j) = -n(p_{ij} + p_{ji}) \tag{2b}$$

Formüllerde; n toplam örneklem büyüklüğünü; $p_{i.}$ ve $p_{.i}$ sırasıyla i. satır ve i. sütuna ait marjinal olasılıkları; p_{ii} köşegen hücre olasılığını (uyum durumunu); p_{ij} ve p_{ji} ise köşegen dışı (off-diagonal) hücre olasılıklarını temsil etmektedir. Toplam marjinal farkların sıfıra eşit olması ($\sum d_i = 0$) nedeniyle tam boyutlu kovaryans matrisi tekil (singular) bir yapı sergilemektedir. Bu nedenle, bilinmeyen olasılıkların gözlenen hücre frekansları üzerinden tahmin edildiği modele, serbestlik derecesini düzeltmek amacıyla $k - 1$ boyutlu indirgenmiş bir fark vektörü (d) dahil edilmekte olup fark vektörü Eşitlik (3)'te verilmiştir.

$$d = (d_1, d_2, \ldots, d_{k-1})' \tag{3}$$

d : k-1 boyutlu fark vektörüdür.

Bu fark vektörü kullanılarak oluşturulan Stuart-Maxwell test istatistiği (Q) Eşitlik (4)'te ifade edilmiştir.

$$Q = d' \hat{\Sigma}^{-1} d \quad (4)$$

Burada d', fark vektörünün transpozunu; $\hat{\Sigma}^{-1}$ ise d vektörüne ait tahmin edilen varyans-kovaryans matrisinin tersini (inverse) göstermektedir. Elde edilen Q istatistiği, asimptotik olarak $k-1$ serbestlik dereceli Ki-kare (χ^2) dağılımına uymaktadır.

Stuart (1955), bu istatistiğin sıfır hipotezi altında serbestlik derecesi *(k − 1)* olan χ^2 dağılımına yaklaştığını göstermiştir. Buna göre, hesaplanan p değerinin belirlenen anlamlılık düzeyinden (genellikle $\alpha = 0.05$) küçük olması durumunda marjinal homojenlik hipotezi reddedilmekte ve satır ile sütun marjinal dağılımları arasında istatistiksel açıdan anlamlı bir asimetri (farklılık) bulunduğu sonucuna varılmaktadır. Bu genel yapının $k = 2$ özel durumuna indirgenmesi, Stuart-Maxwell yaklaşımının McNemar testiyle olan matematiksel bağını son derece net bir biçimde ortaya koymaktadır. Boyut $k = 2$ olarak sınırlandığında, matriste tek bir bağımsız marjinal fark değeri ($d_1 = n_{12} - n_{21}$) kalmakta ve ilgili varyans ifadesi $n(p_{12} + p_{21}) \approx n_{12} + n_{21}$ formuna sadeleşmektedir. Bu cebirsel indirgeme, hesaplanan Q istatistiğini tam olarak süreklilik düzeltmesi içermeyen (uncorrected) klasik McNemar test istatistiğine, diğer bir deyişle Eşitlik (5)'teki forma eşitlemektedir:

$$Q = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \quad (5)$$

Stuart-Maxwell testinin dayandığı istatistiksel hipotezler şu şekilde formüle edilmektedir:

H_0: $p_{i.} = p_{.i}$, (tüm $i = 1, \ldots, k$ için marjinal homojenlik geçerlidir)

H_1: $p_{i.} \neq p_{.i}$ (en az bir i için)

Testin asimptotik geçerliliği, çok değişkenli normallik (multivariate normality) varsayımına dayanmaktadır; diğer bir deyişle, örneklem büyüklüğü (n) arttıkça, d fark vektörünün

dağılımının çok değişkenli normal dağılıma yakınsadığı kabul edilir. Bu yakınsama, Merkezi Limit Teoremi'nin (Central Limit Theorem) çok değişkenli kategorik veri bağlamındaki bir uzantısıdır ve pratikte her bir hücredeki beklenen frekansın belirli bir eşik değerinin (genellikle ≥ 5) üzerinde olmasını gerektirir. Beklenen frekansların düşük olduğu seyrek (sparse) veri yapılarında, asimptotik χ^2 yaklaşımı gerçek test istatistiği dağılımını yeterince iyi temsil edemeyebilir. Bu ihlal durumunda testin nominal Tip-I hata oranı (α) sistematik olarak sapma (şişme veya muhafazakarlaşma) eğilimi gösterebilmektedir.

Yöntemin pratik uygulamalarındaki temel bir sınırlılık, tahmin edilen varyans-kovaryans matrisinin ($\hat{\Sigma}$) tersinin alınabilmesi için matrisin tekil olmayan (non-singular) bir yapıda olması gerekliliğidir. Belirli bir kategoride tam uyum söz konusu olduğunda (yani ilgili satır ve sütun marjinal frekansları köşegen hücreler dışında tam olarak sıfır değerini aldığında), marjinal fark vektöründe ardışık bağımlılıklar oluşur ve matris tekilleşerek (singularity) matematiksel olarak çözümsüz hale gelir. Geleneksel yaklaşımlar ve yazılım algoritmaları genellikle bu sorunu aşmak için ilgili marjinal kategoriyi analizden çıkarmakta ve testi kalan $k-1$ adet kategori üzerinden, indirgenmiş bir serbestlik derecesi ile yürütmektedir. Bu husus, özellikle örneklem hacminin kısıtlı olduğu veya bazı kategorilerin nadir gözlendiği klinik uygulamalarda analiz stratejisi belirlenirken dikkatle ele alınmalıdır.

Tarihsel bağlamda Maxwell (1970), Stuart'ın (1955) çalışmasından bağımsız olarak, aynı örneklemin iki farklı değerlendirici (rater) tarafından bağımsız sınıflandırıldığı tasarımlarda değerlendiriciler arası uyumu (agreement) ve marjinal değişimi incelemek amacıyla benzer bir test istatistiği önermiştir. Farklı çıkış noktalarından elde edilen bu iki istatistiğin matematiksel ve asimptotik olarak birbirine tam özdeş olduğu ise daha sonra Keefe (1982) tarafından cebirsel olarak kanıtlanmıştır.

Bhapkar Testi

Bhapkar testi, k × k boyutlu eşleştirilmiş kontenjans tablolarında marjinal homojenlik hipotezini değerlendirmek amacıyla Bhapkar (1966) tarafından önerilen alternatif bir asimptotik yöntemdir. Bhapkar (1966), doğrusal hipotezlerin sınanmasında kullanılan Wald (1943) istatistiği ile Neyman'ın (1949) minimum χ^2 yaklaşımına dayanan χ_1^2 istatistiğinin, hücre frekansları pozitif olduğu sürece cebirsel olarak özdeş olduğunu göstermiştir; doğrusal olmayan hipotezler için de bu özdeşlik, Neyman'ın doğrusallaştırma tekniğiyle elde edilen χ_1^2 istatistiği için geçerlidir. Bu testte gözlenen hücre frekansları n_{ij}, ilgili olasılıklar p_{ij} ile ilişkilendirilir ve marjinal homojenlik hipotezi H_0: $p_{i.} = p_{.i}$ biçiminde tanımlanır. Hipotezler aşağıdaki şekilde kurulur:

H_0: $p_{i.} = p_{.i}$, tüm $i = 1, \ldots, k$ için (marjinal homojenlik geçerlidir)

H_1: en az bir i için $p_{i.} \neq p_{.i}$

Bu durumda fark vektörü Eşitlik (6)'te gösterildiği şekilde ifade edilir.

$$c = (c_1, c_2, \ldots, c_{k-1})' \tag{6}$$

Burada her bir c_i satır-sütun farklarını temsil eder. Test istatistiği Eşitlik (7)'da gösterildiği şekilde hesaplanmaktadır.

$$X^2 = c'\, G^{-1} c \tag{7}$$

Burada G, fark vektörüne ait örnek kovaryans matrisini temsil etmekte olup, bu matrisin bileşenleri Eşitlik (8)'de verilmiştir.

$$g_{kk'} = n[\delta_{kk'}(p_{k.} + p_{.k'}) - p_{kk'} - p_{k'k}] - n\, c_k c_{k'} \tag{8}$$

$\delta_{kk'}$: Kronecker delta'yı temsil etmekte olup,

$k = k'$ iken 1, $k \neq k'$ iken 0 değerini almaktadır,

c_k ve $c_{k'}$ ise Eşitlik (6)'da tanımlanan fark vektörünün ilgili bileşenleridir.

Bu formülasyon, Stuart (1955) tarafından önerilen ve dikdörtgen parantezdeki $-c_k c_{k'}$ terimini atlayan tanımdan farklıdır; Bhapkar (1966), bu ek düzeltme teriminin, hipotez yanlış olsa dahi tutarlı bir kovaryans tahmini sağlaması bakımından tercih edilmesi gerektiğini göstermiştir.

Hipotezin geçerli olduğu durumda, test istatistiği serbestlik derecesi $(k-1)$ olan χ^2 dağılımına uymaktadır.

Bhapkar ve Stuart-Maxwell testleri arasındaki temel fark, varyans-kovaryans matrisinin tahmin ediliş biçiminde yatmaktadır. Stuart-Maxwell testi, sıfır hipotezi altında (yani marjinal homojenlik geçerliyken) hesaplanan bir varyans tahminine dayanan score tipi bir yaklaşım izlerken, Bhapkar testi, hipotezden bağımsız olarak doğrudan gözlenen verilerden hesaplanan bir varyans tahminine (Wald tipi bir yaklaşıma) dayanmaktadır.

Bhapkar (1966), çalışmasında yalnızca yeni bir test önermekle kalmamış, aynı zamanda kategorik veriler için o döneme kadar birbirinden bağımsız geliştirilen çeşitli ki-kare tipi test kriterlerinin (Wald, 1943; Neyman, 1949 gibi) asimptotik olarak nasıl birbirine indirgendiğini de göstermiştir.

Bowker Simetri Testi

Bowker (1948) tarafından önerilen simetri testi, $k \times k$ boyutlu eşleştirilmiş kontenjans tablolarında marjinal homojenliğe kıyasla daha güçlü ve katı bir hipotezi, yani tam simetriyi (complete symmetry) sınamaktadır. McNemar (1947) testinin doğrudan bir genellemesi niteliğinde olan bu istatistiksel yaklaşım, orijinal çalışmada dikence balığı (*Gasterosteus*) türünün sağ ve sol yanlarındaki yanal plaka desenlerinin karşılaştırılması problemi bağlamında literatüre kazandırılmıştır.

Tam simetri hipotezi, karesel tablonun köşegen dışındaki (off-diagonal) her bir simetrik hücre çiftinin (n_{ij} ve n_{ji}, $i \neq j$ olmak üzere) karşılıklı olarak birbirine eşit olmasını zorunlu kılar ve istatistiksel olarak aşağıdaki şekilde formüle edilir:

H_0: $p_{ij} = p_{ji}$, (tüm $i \neq j$ için tam simetri geçerlidir)

H_1: $p_{ij} \neq p_{ji}$ (en az bir (i, j) çifti için)

Bu hipotez, marjinal homojenlikten çok daha kısıtlayıcı bir koşuldur; zira bir matriste tam simetri sağlandığında, marjinal homojenlik koşulu da matematiksel bir zorunluluk olarak otomatik biçimde sağlanır. Ancak tersi geçerli değildir; marjinal homojenliğin varlığı, tablonun tam simetrik yapıda olduğunu garanti etmez. Söz konusu hiyerarşik yapı nedeniyle Bowker testi; Stuart-Maxwell ve Bhapkar testlerinden yapısal olarak farklılaşan, fakat onlarla kavramsal ve log-lineer teorik düzlemde ayrılmaz biçimde ilişkili olan temel bir hipotezi değerlendirmektedir.

Bowker'ın (1948) orijinal türetimi, $k \times k$ boyutlu karesel tablodaki hücre frekanslarının çok terimli (multinomial) dağılıma sahip olduğu varsayımına dayanmaktadır. Sıfır hipotezi (H_0) altında, köşegen dışındaki her bir (i, j) ve (j, i) hücre çiftindeki toplam gözlem frekansı ($n_{ij} + n_{ji}$) sabit tutulduğunda, bu toplam içindeki dağılımın ilgili iki hücre arasında eşit olasılıkla ($p = 1/2$) paylaşılması beklenmektedir. Bu durum, her bir asimetrik hücre çiftinin koşullu (conditional) olarak bir binom dağılıma uyduğu ve dolayısıyla söz konusu hücre çiftlerine klasik bir McNemar testinin uygulanabileceği anlamına gelmektedir. Bowker'ın literatüre temel metodolojik katkısı; sayısı $k(k-1)/2$ olan bu koşullu binom testlerinin asimptotik olarak birbirinden bağımsız olduğunu kanıtlaması ve bu bağımsız bileşenleri tek bir genel Ki-kare (χ^2) test istatistiği çatısı altında birleştirerek çok kategorili veri yapılarına uygun hale getirmiş olmasıdır.

Test istatistiği, Eşitlik (9)'da gösterildiği üzere, karesel matrisin köşegen dışı (off-diagonal) her bir simetrik hücre çiftine ait gözlenen frekans farklarının karesinin, ilgili iki hücrenin frekans toplamına oranlanması ve bu değerlerin tüm simetrik çiftler üzerinden toplanmasıyla hesaplanmaktadır:

$$\chi_B^2 = \sum_{i<j} \frac{(n_{ij}-n_{ji})^2}{n_{ij}+n_{ji}} \quad (9)$$

Burada; χ_B^2 Bowker simetri test istatistiğini, $\sum_{i<j}$ köşegen üstündeki ($i < j$) her bir hücre çifti üzerinden alınan toplamı, n_{ij} ve n_{ji} ise köşegen dışı simetrik (i, j) ve (j, i) hücrelerindeki gözlenen frekansları temsil etmektedir. Hesaplanan bu istatistik, sıfır hipotezi altında asimptotik olarak $k(k-1)/2$ serbestlik dereceli Ki-kare (χ^2) dağılımına yakınsamaktadır. Boyutun $k = 2$ olduğu özel durumda, matriste yalnızca tek bir köşegen dışı bağımsız hücre çifti kaldığından, Bowker istatistiği cebirsel olarak doğrudan klasik McNemar testine indirgenmektedir. Bowker testinin marjinal homojenlik testleriyle olan yapısal ilişkisi, Bishop vd. (1975) tarafından kuramsallaştırılan "tam simetri = yarı-simetri (quasi-symmetry) + marjinal homojenlik" ayrıştırması üzerinden çok daha net biçimde kavranabilir. Bu teorik ayrıştırma, çalışmanın bir sonraki bölümünde ele alınan Olabilirlik Oranı Testi (Likelihood Ratio Test) tabanlı Marjinal Homojenlik (LRT-MH) yaklaşımının da matematiksel temelini oluşturmaktadır. Uygulamalı istatistik literatüründe bilinen temel bir sınırlılık olarak; küçük örneklem hacimlerinde veya beklenen frekansların düşük olduğu seyrek tablolarda Bowker testinin asimptotik χ^2 yaklaşımı yetersiz kalabilmekte ve nominal Tip-I hata oranı hedeflenen düzeyden sapabilmektedir. Bu sorunu hafifletmek amacıyla literatürde Edwards (1948) süreklilik düzeltmesi (continuity correction) içeren bir modifikasyon önermiş; May ve Johnson (2001) ise Wald-tipi bir alternatif geliştirerek metodolojik çerçeveyi genişletmişlerdir.

Olabilirlik Oranı Testi ile Marjinal Homojenlik (LRT-MH)

Log-lineer modelleme çerçevesinde geliştirilen ve literatüre Caussinus (1965) tarafından kazandırılan yarı-simetri (quasi-symmetry) modeli, karesel kontenjans tablolarının analizinde tam simetri ile marjinal homojenlik kavramları arasındaki yapısal ilişkiyi kesin bir biçimde ortaya koymaktadır. Caussinus tarafından ispatlanan bu temel metodolojik teoreme göre; bir kontenjans tablosunda tam simetri koşulunun sağlanması, söz konusu tablonun yarı-simetri modelini ve marjinal homojenlik hipotezini eşzamanlı (simultaneous) olarak sağlamasına matematiksel açıdan eşdeğerdir. Karesel matrisler için kurulan bu temel teorem, izleyen yıllarda üç yönlü tablolar için Bishop vd. (1975) tarafından, genel çok yönlü (multi-way) veri yapıları için ise Bhapkar ve Darroch (1990) tarafından genişletilerek eşleştirilmiş verilerin analizinde log-lineer modellerin kuramsal bütünlüğü tamamlanmıştır. Kuazi-simetri modeli, Eşitlik (10)'da gösterildiği şekilde log-lineer biçimde ifade edilir.

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad \gamma_{ij} = \gamma_{ji} \tag{10}$$

m_{ij} : (i, j) hücresindeki beklenen frekansı,

α_i ve β_j : sırasıyla satır ve sütun ana etkilerini,

γ_{ij} : simetrik bir etkileşim terimini ($\gamma_{ij} = \gamma_{ji}$) temsil etmektedir.

Bu model, satır ve sütun ana etkilerinin birbirinden farklı olmasına izin vererek, tam simetriden daha esnek bir yapı sunmaktadır.

Yarı-simetri (quasi-symmetry) modelinin matematiksel altyapısı log-lineer formda Eşitlik (11)'deki gibi ifade edilmektedir:

$$\log(m_{ij}) = \mu + \alpha_i + \alpha_j + \gamma_{ij}, \quad (\gamma_{ij} = \gamma_{ji}) \tag{11}$$

Daha önce belirtildiği üzere tam simetri modeli; yarı-simetri (quasi-symmetry) modelinin satır ve sütun ana etkilerine eşitlik kısıtı ($\alpha_i = \beta_i$) getirilmesiyle elde edilen iç içe geçmiş (nested)

bir alt modeldir. Bu hiyerarşik yapı sayesinde, söz konusu iki modelin olabilirlik oranı (deviance - G^2) istatistikleri arasındaki fark, marjinal homojenlik hipotezini yarı-simetri koşulu altında doğrudan test etmek amacıyla kullanılabilmekte ve eşitlik (12)'deki şekilde hesaplanmaktadır. Yarı-simetri modelinin veri setine uygun olduğu varsayımı çerçevesinde, marjinal homojenliğe ilişkin hipotezler aşağıdaki gibi formüle edilmektedir;

H_0: $p_{i.} = p_{.i}$ (tüm i için; marjinal homojenlik ve eşdeğer olarak tam simetri geçerlidir)

H_1: $p_{i.} \neq p_{.i}$ (en az bir i için yarı-simetri geçerliyken tam simetri kısıtları ihlal edilmiştir)

$$\Delta G^2 = G^2_{\text{Simetri}} - G^2_{\text{Yarı-Simetri}} = -2\ln\left(\frac{L_{\text{Simetri}}}{L_{\text{Yarı-Simetri}}}\right) \qquad (12)$$

Eşitlik (12)'de tanımlanan bu test istatistiğinin asimptotik özelliklerini karakterize etmesi, Wilks'in (1938) klasik genel olabilirlik oranı (likelihood ratio) teoremine dayanmaktadır. Sıfır hipotezinin doğruluğu altında hesaplanan ΔG^2 istatistiği, iç içe geçmiş iki modelin tahmin edilen parametre sayıları arasındaki farka karşılık gelen $k-1$ serbestlik dereceli Ki-kare (χ^2) dağılımına yakınsamaktadır.

LRT-MH yaklaşımının Stuart-Maxwell ve Bhapkar testlerinden temel metodolojik farkı; marjinal homojenlik hipotezini doğrudan bir Wald veya Skor (score) istatistiği üzerinden değil, iç içe geçmiş (nested) iki log-lineer modelin deviance (G^2) istatistikleri üzerinden dolaylı olarak sınamasıdır. Bu karakteristik özelliği sayesinde LRT-MH yaklaşımı, $k \times k$ boyutlu kategorik veri analizlerinde log-lineer modelleme çerçevesiyle (framework) tam bütünleşik ve esnek bir alternatif sunmaktadır.

Serbestlik derecesi (df) hesabı bağlamında; yarı-simetri (quasi-symmetry) modelinin serbest parametre sayısı, tam simetri

modelinin parametre sayısından tam olarak $k-1$ adet daha fazladır. Bu cebirsel fark, satır ve sütun marjinal ana etkilerinin (α_i ve β_j) tam simetri modelinde birbirine eşitlenmeye zorlanmasından kaynaklanmaktadır. Dolayısıyla, bu iki model arasındaki olabilirlik oranı farkının (ΔG^2) serbestlik derecesi de doğal bir sonuç olarak $k-1$ değerini almaktadır. Elde edilen bu asimptotik değerin Stuart-Maxwell ve Bhapkar testlerinin serbestlik dereceleriyle birebir örtüşmesi, söz konusu üç testin aynı temel hipotezi (marjinal homojenlik) farklı istatistiksel paradigmalar üzerinden sınadığını bir kez daha doğrulamaktadır.

Kısıtlanmamış doymuş model (saturated model), yarı-simetri ve tam simetri modelleri arasındaki bu hiyerarşik yapı, genel log-lineer model kuramı çerçevesinde cebirsel olarak daha da netleştirilebilmektedir. Herhangi bir kısıt içermeyen doymuş bir $k \times k$ tablo, tüm hücre olasılıklarının toplamının bire eşit olması ($\sum p_{ij} = 1$) koşulu altında $k^2 - 1$ adet serbest parametreye sahiptir. Buna karşın tam simetri modelinde, yalnızca köşegen (diagonal) ve köşegen üstü (upper-diagonal) hücrelerin birbirinden bağımsız olduğu varsayıldığından, hesaplanabilir serbest parametre sayısı $k(k+1)/2 - 1$ değerine düşmektedir. Doymuş model ile tam simetri modeli arasındaki serbestlik derecesi farkı, Eşitlik 13'te ifade edildiği üzere doğrudan $k(k-1)/2$ sonucunu vermektedir:

$$\Delta df = (k^2 - 1) - \left[\frac{k(k+1)}{2} - 1\right] = \frac{k(k-1)}{2} \qquad (13)$$

Ortaya çıkan bu matematiksel eşitlik; tam simetri modelinin doymuş modele kıyasla uyum iyiliğini test eden Olabilirlik Oranı Testi'nin, Bowker simetri testi ile aynı hipotezi, aynı serbestlik derecesi altında sınadığını ve büyük örneklemlerde asimptotik olarak eşdeğer (asymptotically equivalent) sonuçlar ürettiğini teorik olarak kanıtlamaktadır. Yarı-simetri modeli ise kısıtlılık bağlamında bu iki uç nokta (doymuş model ve tam simetri) arasında yer alan hiyerarşik bir ara model olup; serbest parametre

sayısı bakımından tam simetri modelinden $k-1$ adet daha zengindir. Söz konusu kademeli parametrizasyon yapısı, "Tam Simetri = Yarı-simetri + Marjinal Homojenlik" ayrıştırmasının yalnızca teorik düzeyde değil, serbestlik dereceleri düzeyinde de mutlak bir tutarlılıkla yansıdığını açıkça göstermektedir.

Tomizawa ve Tahata (2007), simetri, yarı-simetri ve marjinal homojenlik (marjinal simetri) modelleri arasındaki bu yapısal ayrıştırma ilişkisini yeniden ele alarak, çok yönlü (multi-way) tablolar için temel bir asimptotik özellik ortaya koymuşlardır. Yazarlar, söz konusu veri yapılarında tam simetri modeline ait olabilirlik oranı istatistiğinin (G^2), yarı-simetri ve marjinal homojenlik modellerine ait olabilirlik oranı istatistiklerinin toplamına asimptotik olarak eşdeğer olduğunu kanıtlayarak bu teorik log-lineer ayrıştırmanın esnekliğini ve genişletilebilirliğini ispatlamışlardır.

Yanıt kategorilerinin sıralı (ordinal) bir yapıya sahip olduğu spesifik araştırma tasarımlarında ise Agresti (1983), marjinal homojenlik hipotezini sınamak amacıyla köşegen-parametre simetri (diagonal-parameter symmetry) modeline dayanan alternatif bir olabilirlik oranı testi önermiştir. Bu özelleştirilmiş yaklaşım, kategoriler arasındaki doğal sıra bilgisini doğrudan modele katarak, sırasız (nominal) kategoriler için tasarlanmış genel LRT-MH yöntemine kıyasla istatistiksel gücü (statistical power) belirgin şekilde artırmaktadır. Dolayısıyla, karesel tablodaki sınıfların sıralı olduğu durumlarda, veri kaybını önlemek ve marjinal asimetriyi daha hassas tespit edebilmek adına Agresti'nin (1983) önerdiği bu ordinal modelin kullanılması metodolojik açıdan çok daha uygundur.

Log-lineer yarı-simetri modelindeki marjinal parametrelerin genellikle kapalı formda cebirsel bir çözümü bulunmadığından, maksimum olabilirlik tahmin süreci zorunlu olarak iteratif yöntemlere dayanmaktadır. Bu amaçla literatürde ve istatistiksel

paket programlarında en yaygın kullanılan hesaplama yöntemi, teorik kökeni Deming ve Stephan'ın (1940) çalışmasına dayanan İteratif Oransal Uyum (Iterative Proportional Fitting – IPF) algoritmasıdır.

LRT-MH yaklaşımında test istatistiği olarak kullanılan ΔG^2 değeri, özünde simetri ve yarı-simetri modellerinin sırasıyla kısıtlanmamış doymuş (saturated) modele göre hesaplanan sapma (deviance) değerleri arasındaki farkı ifade etmektedir. Doymuş modelin referans olabilirlik değeri her iki modelin deviance formülasyonunda da ortak bir terim olarak yer aldığından, iki istatistik birbirinden çıkarıldığında bu ortak terimler cebirsel olarak birbirini yok etmekte ve fark denklemi bütünüyle sadeleşerek Eşitlik 12'deki yalın forma ($G^2_{\text{Simetri}} - G^2_{\text{Yarı-Simetri}}$) indirgenmektedir.

SONUÇ

Bu çalışmada, $k \times k$ boyutlu eşleştirilmiş kategorik verilerin analizinde temel alınan dört farklı istatistiksel yaklaşım —Stuart-Maxwell testi, Bhapkar testi, Bowker simetri testi ve Olabilirlik Oranı Testi tabanlı Marjinal Homojenlik (LRT-MH) yaklaşımı— kuramsal ve pratik boyutlarıyla incelenmiştir. Söz konusu testler, marjinal homojenlik ve tam simetri (complete symmetry) kavramları etrafında şekillenmekte olup, veriyi modellemede farklı metodolojik dayanaklara sahiptir. Stuart-Maxwell ve Bhapkar testleri, marjinal homojenliği doğrudan sırasıyla skor-tipi ve Wald-tipi istatistikler üzerinden sınarken; LRT-MH yaklaşımı, yarı-simetri (quasi-symmetry) modelinin geçerli olduğu varsayımı altında, marjinal homojenliğe karşılık gelen kısıtları olabilirlik oranı (deviance - G^2) yöntemiyle dolaylı olarak değerlendirmektedir. Bowker testi ise söz konusu yaklaşımlardan ayrılarak, marjinal eşitliğe kıyasla yapısal olarak çok daha kısıtlayıcı bir koşul olan tam simetri hipotezini test etmektedir.

Stuart-Maxwell ve Bhapkar testleri, temel marjinal homojenlik hipotezini sınarken varyans-kovaryans matrisinin tahmin edilme sürecinde farklı analitik stratejiler benimserler. Stuart-Maxwell testi bu matrisi sıfır hipotezi (H_0) kısıtları altında elde ederken, Bhapkar testi doğrudan kısıtlanmamış gözlenen veriler üzerinden tahminleme yapar. Büyük örneklemlerde asimptotik olarak birbirlerine eşdeğer (Keefe, 1982) olmalarına karşın, bu temel algoritmik ayrışma, testleri özellikle hesaplama felsefesi açısından birbirinden farklılaştıran en önemli niteliktir.

Öte yandan, Bowker testi ve LRT-MH yaklaşımı veri analizine farklı ve tamamlayıcı bir kuramsal zemin sunmaktadır. Bowker testi, yalnızca marjinal toplamların denkliğini sınırlandırmayıp köşegen dışı (off-diagonal) her bir simetrik hücre çiftinin uyumunu ayrı ayrı şart koştuğundan; araştırmacının marjinal değişimden ziyade hücresel düzeyde katı bir geçiş asimetrisi aradığı araştırma tasarımlarında tercih edilmelidir. LRT-MH yaklaşımı ise genel log-lineer modelleme mimarisiyle (framework) tam bütünleşik çalışması sayesinde; yarı-simetri gibi hiyerarşik ara modellerin de aynı kuramsal çatı altında değerlendirilmesine olanak tanımakta ve kategorik veri analizinde çok daha kapsamlı bir uyum iyiliği (goodness-of-fit) karşılaştırmasını mümkün kılmaktadır.

Özetle, $k \times k$ boyutlu eşleştirilmiş kontenjans tablolarının analizinde, her tür veri yapısı için geçerli evrensel ve tek bir "en iyi" testten söz etmek mümkün değildir. Doğru istatistiksel testin seçimi; spesifik olarak sınanmak istenen araştırma hipotezinin kuramsal doğasına (marjinal homojenlik veya daha kısıtlayıcı bir yapı olan tam simetri), tablonun satır/sütun boyutuna (k), örneklem hacmine ve hücrelerdeki veri seyrekliğine (sparseness) bağlı olarak, istatistiksel bir gerekçelendirme (justification) ile yapılmalıdır.

KAYNAKÇA

Agresti, A. (1983). Testing marginal homogeneity for ordinal categorical variables. *Biometrics*, 39(2), 505–510.

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association*, 61(313), 228–235.

Bhapkar, V. P., ve Darroch, J. N. (1990). Marginal symmetry and quasi symmetry of general order. *Journal of Multivariate Analysis*, 34(2), 173–184.

Bishop, Y. M. M., Fienberg, S. E., ve Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43(244), 572–574.

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des Sciences de l'Université de Toulouse*, 29, 77–183.

Deming, W. E., ve Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4), 427–444.

Edwards, A. L. (1948). Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika*, 13(3), 185–187.

Keefe, T. J. (1982). On the relationship between two tests for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 69(3), 683–684.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*, 116(535), 651–655.

May, W. L., ve Johnson, W. D. (2001). Symmetry in square contingency tables: Tests of hypotheses and confidence interval construction. *Journal of Biopharmaceutical Statistics*, 11(1–2), 23–33.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.

Neyman, J. (1949). Contribution to the theory of the χ^2 test. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, 239–273.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42(3–4), 412–416.

Tomizawa, S., ve Tahata, K. (2007). The analysis of symmetry and asymmetry: Orthogonality of decomposition of symmetry into quasi-symmetry and marginal symmetry for multi-way tables. *Journal de la Société française de Statistique*, 148(3), 3–36.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3), 426–482.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60–62.

TANI TESTİ PERFORMANS ÖLÇÜTLERİ

Nurhan Doğan[2], İsmet Doğan[3]

GİRİŞ

Sağlık hizmetlerinde kullanılan bir tıbbi müdahalenin, cihazın ya da tanı testinin temel amacı, bireyin sağlık durumuna yönelik anlamlı bir katkı sağlamaktır. Yalnızca tanısal bir işlem gerçekleştirmek amacıyla test uygulanması; çelişkili veya yanlış yorumlanabilecek sonuçların ortaya çıkmasına, gereksiz tanıların konulmasına, hastalar açısından ek yüklerin oluşmasına ve sağlık hizmetlerinin maliyetinin artmasına neden olabilir. Bir tanı testinden öncelikle belirli bir hastalığa ilişkin riski doğru biçimde ortaya koyması beklenmekle birlikte, nihai olarak test sonucunun bireyin sağlık durumunun iyileştirilmesine katkıda bulunması gerekir (Leeflang ve Allerberger, 2019). Tanı testleri, klinik uygulamada bireyin belirli bir hastalığa sahip olup olmadığının mümkün olduğunca doğru biçimde belirlenmesini ve elde edilen sonuca uygun olarak erken ve etkili bir tedavinin planlanmasını sağlayan temel araçlardır. Yanlış tanı konulması, mevcut hastalığın gözden kaçırılması veya tanının gecikmesi, hastanın tedavi sürecini olumsuz yönde etkileyebileceğinden, doğru zamanda gerçekleştirilen güvenilir bir tanısal değerlendirme hastalık yönetiminin temel unsurlarından biridir. Bununla birlikte tanısal süreç durağan bir yapı göstermemektedir. Hastalıkların görülme sıklığı ve özellikleri, hastalıkların klinik seyri ve tanı amacıyla kullanılan yöntemler zaman içerisinde değiştiğinden, tanısal değerlendirme yaklaşımları da sürekli olarak gelişmektedir. Klinik uygulamada temel gereksinimlerden biri, belirli bir klinik durumdaki hasta için en uygun ve yararlı tanı testinin belirlenmesidir. Gereğinden fazla veya yetersiz tanı konulması, sırasıyla gereksiz tedavi uygulamalarına ya da gerekli tedavinin

2 Prof. Dr., Afyonkarahisar Sağlık Bilimleri Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Afyonkarahisar, nurhandogan@hotmail.com, ORCID: 0000-0001-7224-6091

3 Prof. Dr., Afyonkarahisar Sağlık Bilimleri Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Afyonkarahisar, ismet.dogan@afsu.edu.tr, ORCID: 0000-0001-9251-3564

verilmemesine neden olabilir. Her iki durum da hem bireylerin sağlığı hem de sağlık sistemlerinin kaynak kullanımı açısından önemli sonuçlar doğurmaktadır. Tıpta belirti ve semptomların değerlendirilmesinde kullanılan istatistiksel yöntemler ise araştırmanın hangi aşamasında bulunulduğuna göre farklılık gösterebilmekte ve doğrudan araştırma sorusu ile çalışma tasarımına bağlı olarak şekillenmektedir (Bolboaca, 2019). Tanı testlerinin doğruluğunu değerlendirmek amacıyla çeşitli istatistiksel yaklaşımlar geliştirilmiştir. Kullanılacak değerlendirme yönteminin belirlenmesinde, test sonucunun niteliği önemli bir rol oynamaktadır. Test sonuçları ikili veya sıralı ölçeklerde ifade edilen nitel veriler şeklinde olabileceği gibi, sürekli ölçümlerden oluşan nicel veriler biçiminde de elde edilebilir. Bununla birlikte, tıbbi araştırmaların pek çok alanında hastalık durumunu kesin olarak ortaya koyan gerçek bir altın standardın bulunmaması, tanı testlerinin doğruluğunun değerlendirilmesini güçleştiren önemli bir sorundur (Lim, 2021). Tanısal testlerin performansının değerlendirilmesi; kesin tanının konulması, hastalıkların taranması ve prognozun öngörülmesi gibi farklı klinik amaçlar açısından önem taşımaktadır. Bununla birlikte hiçbir tanı testi, hastalığı bulunan bireyler ile hastalığı bulunmayan bireyleri kusursuz biçimde birbirinden ayıramaz. Bu nedenle herhangi bir tanı testinin klinik kullanıma alınmasından önce tanısal doğruluğunun ortaya konulması gerekir. İkili sonuç veren tanı testlerinin değerlendirilmesinde duyarlılık, özgüllük, pozitif öngörü değeri ve negatif öngörü değeri en sık kullanılan temel performans ölçütleri arasında yer almaktadır (Dhamnetiya ve ark., 2021; Tamura ve ark., 2025). Bununla birlikte, bu ölçütlerin tamamı aynı özelliği değerlendirmemektedir. Bazı ölçütler testin hastalıklı ve sağlıklı bireyleri birbirinden ayırma kapasitesine odaklanırken, bazıları test sonucuna dayanarak hastalık durumunun öngörülmesine ilişkin bilgi sağlamaktadır (Shaikh, 2011). Yeni bir tıbbi testin klinik uygulamaya girmesinden önce farklı açılardan değerlendirilmesi gerekmektedir. Bu değerlendirme kapsamında analitik geçerlilik, testin laboratuvar koşullarında beklenen performansı gösterip göstermediğini; klinik geçerlilik, hedef hasta popülasyonunda hastalığı doğru biçimde ayırt edip edemediğini; klinik yarar ise testin kullanımının bireylerin sağlık sonuçlarında gerçek bir iyileşme sağlayıp sağlamadığını ortaya

koymaktadır. Klinik geçerliliğin değerlendirilmesine yönelik çalışmalar, aynı zamanda tanısal doğruluk çalışmaları olarak da adlandırılmakta ve testin hasta bireyler ile sağlıklı bireyleri birbirinden ne ölçüde doğru ayırabildiğini belirlemeyi amaçlamaktadır (Umemneku Chikere ve ark., 2019). Tanı, prognoz ve tedavi, klinik tıpta karar verme sürecinin üç temel bileşenini oluşturmaktadır. Bu süreç içerisinde doğru bir tanısal değerlendirme, hastanın prognozunun iyileştirilmesi açısından ilk ve en önemli aşamalardan biridir. Dolayısıyla tanısal araştırmalarda kullanılan istatistiksel yöntemlerin doğru biçimde anlaşılması ve yorumlanması, elde edilen bulguların klinik açıdan anlamlı biçimde değerlendirilmesi bakımından büyük önem taşımaktadır. İdeal bir tanı testinin hastalığı bulunan ve bulunmayan bireyleri kusursuz biçimde ayırması ve dolayısıyla %100 duyarlılık ile %100 özgüllüğe sahip olması beklenir. Ancak gerçeğe en yakın sonucu verdiği kabul edilen referans standartların çoğu zaman yüksek maliyetli olması, uzun zaman gerektirmesi ve özel uzmanlık ihtiyacı doğurması, bu yöntemlerin geniş popülasyonlarda rutin olarak kullanılmasını sınırlandırmaktadır. Bu nedenle günümüzde belirli biyobelirteçlerin tanısal değerinin araştırılması giderek önem kazanan çalışma alanlarından biri hâline gelmiştir. Bu araştırmaların önemli bir bölümü, mevcut altın standardın yerine kullanılabilecek, daha düşük maliyetli ve görece kolay ölçülebilen biyobelirteçlerin belirlenmesine odaklanmaktadır (Leonardis ve ark., 2023). Tanı testleri yalnızca klinik tanı sürecinde değil, aynı zamanda tarama programlarında ve bilimsel araştırmalarda da önemli bir yere sahiptir. Bu nedenle hastalığı bulunan bireylerle sağlıklı bireylerin güvenilir biçimde ayrılması ve hastanın sağlık durumunun kapsamlı biçimde değerlendirilmesi, etkili bir tanı testinin temel işlevleri arasında bulunmaktadır (Chen ve ark., 2025). Tanısal doğruluk, temel olarak bir testin hedeflenen hastalık durumunu sağlıklı durumdan ayırabilme kapasitesini ifade etmektedir. Bu kapasitenin değerlendirilmesinde duyarlılık, özgüllük, öngörü değerleri, olabilirlik oranları, ROC eğrisi altında kalan alan, Youden indeksi ve tanısal odds oranı gibi çok sayıda ölçütten yararlanılabilmektedir. Her bir ölçüt test performansının farklı bir boyutunu ortaya koymaktadır. Bazı ölçütler testin ayırt etme kapasitesini değerlendirirken, bazıları elde edilen test sonucunun hastalık durumunu öngörmedeki gücünü ortaya koymaktadır. Bunun

yanında, tanısal doğruluğu ifade eden ölçütlerin tümü aynı özelliklere sahip değildir. Bazıları hastalığın prevalansından etkilenirken, bazıları hastalığın spektrumuna, hastalığın nasıl tanımlandığına ve çalışmanın tasarımına duyarlıdır (Simundic, 2009). Tanı testlerinin hastalığın kesin olarak varlığını doğrulaması veya yokluğunu dışlaması her zaman mümkün değildir. Klinik uygulamada test sonuçları çoğunlukla hekimin hastanın belirli bir hastalığa sahip olma olasılığına ilişkin mevcut değerlendirmesini güçlendirmek veya zayıflatmak amacıyla kullanılmaktadır. Test sonuçlarının pozitif ya da negatif olarak sınıflandırılması, klinik karar verme sürecine katkı sağlamakla birlikte, tek başına test sonucunun klinik açıdan ne ölçüde güvenilir olduğunu ortaya koymaz. Klinisyenler duyarlılık ve özgüllük kavramlarına tanı testlerinin güvenilirliğini değerlendirmede sıklıkla başvursalar da bu iki ölçütün önemli sınırlılıkları bulunmaktadır. Çünkü duyarlılık ve özgüllük, hastalık durumu önceden bilinen bireylerde belirli bir test sonucunun görülme olasılığını ifade eder. Başka bir ifadeyle, belirli bir test sonucuyla karşılaşıldığında hastanın gerçekten hasta olma olasılığı konusunda doğrudan bilgi sağlamazlar. Ayrıca bir tanı testinin ikiden fazla olası sonuç ürettiği durumlarda sonuçların yalnızca pozitif veya negatif şeklinde sınıflandırılması yeterli olmayabilir. Böyle durumlarda duyarlılık ve özgüllük kavramlarının klasik biçimde kullanılması da uygun değildir (Hayden ve Brown, 1999). Bu nedenle tanı testlerinin değerlendirilmesinde yalnızca tek bir performans ölçütüne bağlı kalınmaması ve test sonucunun klinik bağlam içerisinde farklı göstergeler aracılığıyla ele alınması gerekmektedir. Tanı testleri, hekimlere hastanın sağlık durumuna ilişkin bilgi sağlayarak uygun tedavi yaklaşımının belirlenmesine katkıda bulunan önemli klinik araçlardır. Hasta ve sağlıklı bireylerin birbirinden doğru biçimde ayrılabilmesi, iyi bir tanı testinin temel özelliklerinden biridir (Samawi ve ark., 2024). Sağlık profesyonelleri günlük klinik uygulamada belirsizlik içeren durumlarla sıklıkla karşılaşmakta ve karar verme sürecinde hastaya ilişkin farklı bilgi kaynaklarından yararlanmaktadır. Anamnez, fizik muayene ve yardımcı tanı testlerinden elde edilen bilgiler birlikte değerlendirilerek hastanın mevcut klinik durumu belirlenmeye çalışılmaktadır. Bu bağlamda tanı testleri, özellikle şüphelenilen hastalığın mevcut olup olmadığının değerlendirilmesinde önemli bir destek sağlamaktadır. Bununla birlikte, bir tanı testinin

performansı bütün klinik koşullarda aynı olmayabilir. Testin uygulandığı hasta grubu, klinik durum ve hastalığın özellikleri gibi faktörler testin sonuçlarını etkileyebilir. Bu nedenle tanısal değerlendirmelerde yalnızca doğruluğun, yani duyarlılık ve özgüllüğün, değil; test sonucunun klinik kararlar ve sağlık sonuçları üzerindeki etkisinin de dikkate alınması gerekmektedir (Perez ve ark., 2021). Tanı testlerinin performansının değerlendirilmesinde kullanılan hesaplamalar genellikle bireylerin referans test sonucuna göre hasta ve sağlıklı olmak üzere iki gruba ayrılmasıyla başlatılmaktadır. Daha sonra test sonucu ile referans test sonucu birlikte değerlendirilerek 2x2'lik bir tablo oluşturulur. Bu tablodan elde edilen değerler kullanılarak farklı tanısal performans ölçütleri hesaplanabilmektedir. Bireylerin ilgili gruplara ayrılmasında altın standart olarak kabul edilen bir referans testten yararlanılabilir. Tanı testinin sonuçlarının niteliğinden bağımsız olarak uygun bir sınıflandırma yapıldığı sürece 2x2 tablosu oluşturulabilir ve çeşitli doğruluk ölçütleri hesaplanabilir. 2x2 tablosunun genel yapısı Tablo 1'de gösterilmiştir. Tanı testi ile referans test arasında güçlü bir uyum bulunması, testin yüksek tanısal doğruluğa sahip olduğuna ilişkin önemli bir gösterge olarak değerlendirilebilir.

Tablo 1. Test performans ölçütlerinin hesaplanmasında kullanılan 2x2 tablosu

		Referans Test Sonucu	
Medikal Test Sonucu		Hastalık Var	Hastalık Yok
	Pozitif	Gerçek Pozitif (GP)	Yalancı Pozitif (YP)
	Negatif	Yalancı Negatif (YN)	Gerçek Negatif (GN)

2x2 tablosunda ortaya çıkan dört farklı sonucun her birini ayrı ayrı değerlendirmek, farklı tanı testlerinin performansını karşılaştırırken pratik olmayabilir. Bu nedenle GP, YP, YN ve GN değerlerini çeşitli matematiksel ilişkiler aracılığıyla tek bir gösterge veya özet ölçüt hâline getiren çok sayıda performans ölçütü geliştirilmiştir (Chicco ve Jurman, 2022). Bu ölçütler, tanı testinin doğruluğu, faydalılığı ve klinik karar verme sürecine sağlayabileceği katkının değerlendirilmesine yardımcı olmaktadır (Bolboaca, 2019). Bununla birlikte, tanı testlerinin performansını tek bir genel istatistiksel ölçütle değerlendirme konusunda literatürde tam bir uzlaşı bulunmamaktadır. Farklı performans ölçütlerinin güçlü ve zayıf yönlerine ilişkin çok sayıda

yaklaşım ve eleştiri bulunmaktadır. En temel ölçütlerden biri olan doğruluk (accuracy), doğru sınıflandırılan bireylerin toplam birey sayısına oranını ifade etmektedir. Tanım gereği ikiden fazla kategorinin bulunduğu sınıflandırma problemlerinde de kullanılabilen bu ölçüt, veri setindeki sınıfların dağılımı dengeli olduğunda yararlı bilgiler sağlayabilir. Buna karşılık, sınıflar arasında belirgin bir dengesizlik bulunduğunda doğruluk yanıltıcı hâle gelebilir. Özellikle çoğunluk sınıfının çok daha fazla sayıda birey içerdiği veri setlerinde, sınıflandırıcının gerçek performansından daha iyimser bir sonuç ortaya çıkabilir (Chicco ve Jurman, 2020). Bu nedenle tanı testlerinin performansını değerlendirmek amacıyla yalnızca doğruluk değerine odaklanmak yerine, farklı ölçütlerin birlikte değerlendirilmesi daha kapsamlı bir yaklaşım sağlamaktadır. İzleyen bölümde tanı testlerinin performansını değerlendirmede kullanılan başlıca ölçütler ele alınarak bu ölçütlerin özellikleri, hesaplanma biçimleri ve yorumlanmaları açıklanmaktadır.

PERFORMANS ÖLÇÜTLERİ

Duyarlılık (Se) ve Özgüllük (Sp)

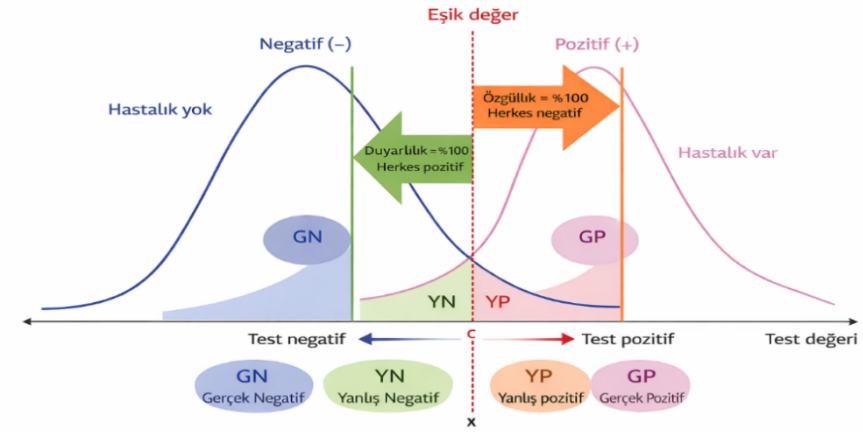
Tanısal testlerin ve risk ya da tahmin modellerinin performansını değerlendirmek amacıyla kullanılan temel ölçütlerin başında duyarlılık ve özgüllük gelmektedir. İngilizcede sensitivity, true positive rate veya recall olarak ifade edilen duyarlılık ile specificity veya true negative rate olarak adlandırılan özgüllük, 1947 yılından itibaren tanısal değerlendirmelerde yaygın biçimde kullanılmaktadır. Bununla birlikte, iki farklı tanı testinin karşılaştırılması söz konusu olduğunda, performansın tek bir genel göstergeyle özetlenmesi çoğu zaman daha işlevsel bir yaklaşım sunmaktadır. Bu amaçla geliştirilen ölçütlerin önemli bir bölümü duyarlılık ve özgüllük değerlerinin birlikte değerlendirilmesine dayanmaktadır. Tanısal doğruluk/etkinlik oranı, tanısal olabilirlik oranları, tanısal odds oranı, Youden indeksi, Kappa katsayısı ile pozitif ve negatif öngörü değerleri bu ölçütler arasında yer almaktadır. Duyarlılık ve özgüllük birbirinden bağımsız iki gösterge olarak ele alınabilse de aralarında göz ardı edilmemesi gereken bir ilişki bulunmaktadır. İlk bakışta bu durum şaşırtıcı olabilir; çünkü duyarlılık hastalığı bulunan bireylerden, özgüllük ise hastalığı

bulunmayan bireylerden elde edilen sonuçlar üzerinden hesaplanmaktadır. Dolayısıyla bu iki grubun birbirinden bağımsız olduğu düşünülebilir. Ancak tanı testlerinin çoğu kusursuz olmadığından, yanlış sınıflandırma olasılığı ortaya çıkmaktadır. Bu durum, duyarlılık ve özgüllük tahminleri arasında korelasyona yol açabilmektedir. Testin kusursuz olduğu durumda ise söz konusu korelasyon ortadan kalkmaktadır (Zhong, Liu ve Chen, 2020). Duyarlılık ve özgüllük, 2x2 sınıflandırma tablosunda yer alan gerçek pozitif (GP), yanlış negatif (YN), gerçek negatif (GN) ve yanlış pozitif (YP) değerlerinden yararlanılarak aşağıdaki eşitliklerle hesaplanır:

$$Duyarlılık = GP/(GP + YN) \quad (1)$$

$$Özgüllük = GN/(GN + YP) \quad (2)$$

Her iki ölçütün değeri 0 ile 1 arasında değişmektedir. Sıfıra yaklaşan değerler düşük performansı, 1'e yaklaşan değerler ise yüksek performansı ifade eder. Duyarlılık ve Özgüllük arasındaki denge Şekil 1'de gösterilmiştir.



Şekil 1. Duyarlılık ve özgüllük arasındaki denge. YN: yanlış negatif; YP: yanlış pozitif; GN: gerçek negatif; GP: gerçek pozitif.

Şekil 1'den de görüldüğü üzere olumlu tanı koyarken, eşik değerine (cut-off) bağlı olarak duyarlılık ve özgüllük arasında bir ödünleşim söz konusudur. Bir tanı testinde duyarlılık ve özgüllüğün aynı anda yükseltilmesi çoğu zaman mümkün değildir. Özellikle sürekli ölçekte elde edilen test sonuçlarında sınıflandırma için seçilen eşik değerinin (cut-off) değiştirilmesi, yanlış pozitif ve yanlış negatif sonuçların

sayısını etkileyerek bu iki ölçüt arasında bir denge kurulmasına neden olur. Eşik değerinin yükseltilmesi testin özgüllüğünü artırma eğilimindeyken duyarlılıkta azalmaya yol açabilir. Buna karşılık daha düşük bir eşik değerinin kullanılması, daha fazla bireyin hastalığı olan gruba atanmasına ve dolayısıyla duyarlılığın yükselmesine neden olur; ancak bu durumda yanlış pozitif sonuçların sayısı da artabilir (Lim, 2021). Bu nedenle klinik amaca uygun bir eşik değerinin belirlenmesinde duyarlılık ve özgüllük arasındaki dengenin dikkate alınması gerekir.

Pozitif ve Negatif Öngörü Değerleri

Pozitif öngörü değeri (PÖD; positive predictive value/precision), test sonucu pozitif bulunan bir bireyin gerçekten hastalığa sahip olma olasılığını ifade eder. Buna karşılık negatif öngörü değeri (NÖD; negative predictive value), test sonucu negatif olan bir bireyin gerçekte hastalığa sahip olmama olasılığını göstermektedir. PÖD ve NÖD'nin yorumlanmasında hastalığın toplumdaki görülme sıklığı önemli bir belirleyicidir. Özellikle araştırma örnekleminde hasta ve sağlıklı bireylerin sayısının araştırmacı tarafından önceden ve eşit oranlarda belirlenmesi durumunda, örneklemden elde edilen öngörü değerlerinin gerçek popülasyondaki klinik durumu doğrudan yansıtması beklenmemelidir. Duyarlılık ve özgüllükten farklı olarak PÖD ve NÖD, araştırılan popülasyondaki hastalık prevalansından belirgin biçimde etkilenmektedir. Bu nedenle bir çalışmada elde edilen PÖD ve NÖD değerlerinin, hastalık prevalansının farklı olduğu başka bir popülasyona doğrudan aktarılması uygun değildir (Shaikh, 2011). Prevalanstaki değişimin PÖD ve NÖD üzerindeki etkisi aynı yönde değildir. Hastalığın toplumdaki sıklığı arttıkça PÖD yükselirken NÖD düşme eğilimi gösterir. Bu değişimin PÖD üzerindeki etkisi genellikle daha belirgindir; NÖD ise prevalanstaki değişimden görece daha az etkilenmektedir (Hayden ve Brown, 1999). Klinik uygulama açısından bakıldığında, duyarlılık ve özgüllük testin teknik performansını ortaya koyarken PÖD ve NÖD, test sonucunun belirli bir hastadaki anlamının değerlendirilmesini kolaylaştırmaktadır. Bununla birlikte, yalnızca PÖD veya yalnızca NÖD üzerinden karar verilmesi yerine her iki göstergenin birlikte değerlendirilmesi gerekir (Habibzadeh, 2025). PÖD ve NÖD aşağıdaki eşitliklerle hesaplanmaktadır:

$$PÖD = GP/(GP + YP) \quad (3)$$

$$NÖD = GN/(GN + YN) \quad (4)$$

Her iki ölçüt için teorik aralık 0 ile 1 arasındadır. Değerlerinin 1 olması, ilgili test sonucunun gerçek durumla tamamen uyumlu olduğunu gösterirken, 0 değeri ilgili öngörü performansının bulunmadığı durumu ifade eder.

Pozitif ve Negatif Olabilirlik Oranı

Olabilirlik oranları, bir tanı testinden elde edilen sonucun hastalık olasılığını ne ölçüde değiştirdiğini değerlendirmeye yarayan önemli göstergelerdir. Pozitif olabilirlik oranı (LR+) test sonucu pozitif olduğunda, negatif olabilirlik oranı (LR-) ise test sonucu negatif olduğunda hastalık olasılığındaki değişimi değerlendirmek için kullanılır. LR+ için teorik olarak en düşük değer 1, üst sınır ise +∞ iken LR- için değerler 0 ile 1 arasında değişmektedir. Olabilirlik oranlarının temel üstünlüğü, testin hem duyarlılığını hem de özgüllüğünü tek bir değerlendirme çerçevesinde birleştirmesidir. Ayrıca test sonucundan önceki hastalık olasılığı ile test sonucundan sonraki olasılık arasında doğrudan bağlantı kurulmasına olanak tanır. Hastaya ilişkin klinik bilgiler kullanılarak belirlenen test öncesi olasılık, test sonucunun ardından olabilirlik oranı yardımıyla test sonrası olasılığa dönüştürülebilir. Test öncesi olasılığın belirlenmesinde hastalığın prevalansı önemli bir bilgi kaynağıdır. Bu yaklaşımda testin hastaya uyarlanmasından ziyade, mevcut klinik bilgiler ışığında hastanın test sonucunun nasıl yorumlanacağı ön plana çıkmaktadır (Shaikh, 2011). LR+ ve LR-, hastalık ve hastalık olmayan gruplardaki test sonuçlarının göreli olasılıklarını karşılaştırır. Bu nedenle prevalanstan bağımsız performans göstergeleri olarak kabul edilirler. Hastalık tanımı ve kullanılan referans standart değişmediği sürece, bir araştırmada hesaplanan olabilirlik oranlarının farklı klinik ortamlarda kullanılabilmesi önemli bir avantajdır (Chen ve ark., 2015). Olabilirlik oranlarının başlıca avantajları şu şekilde özetlenebilir (Grimes ve Schulz, 2005):

- Test sonuçlarının farklı düzeylerini klinik açıdan yorumlamaya olanak sağlaması,

- Farklı klinik ortamlara taşınabilir olması,
- Duyarlılık ve özgüllüğün tek başına kullanılmasından kaynaklanabilecek yanlış yorumlamaları azaltması,
- Klinik karar verme sürecine katkı sağlaması.

Bir olabilirlik oranının 1'e yakın olması, test sonucunun hastalık olasılığı üzerinde sınırlı bir değişiklik meydana getirdiğini gösterir. Buna karşılık LR+ değerinin yüksek, LR- değerinin düşük olması testin hastalığı ayırt etme gücünün daha yüksek olduğunu düşündürür. Olabilirlik oranları semptomlardan fizik muayene bulgularına, laboratuvar analizlerinden görüntüleme yöntemlerine ve skorlama sistemlerine kadar farklı klinik uygulamalarda kullanılabilmektedir. Uygun bir test öncesi olasılık ile birlikte değerlendirildiğinde, bu oranlar klinik kararın daha nicel bir temele oturtulmasını sağlayabilir (Grimes ve Schulz, 2005). LR+ ve LR- aşağıdaki eşitliklerden hesaplanmaktadır:

$$LR+= Se/(1-Sp) \quad (5)$$

$$LR-= (1-Se)/Sp \quad (6)$$

Hastalığın tanımlanmasında veya referans standardında değişiklik yapılması durumunda, hesaplanan performans göstergelerinin başka klinik ortamlara doğrudan aktarılması mümkün olmayabilir (Simundic, 2009).

Doğruluk

Doğruluk (*accuracy*), sınıflandırma sonucunun gerçek durumla ne oranda örtüştüğünü gösteren temel performans ölçütlerinden biridir. Başka bir ifadeyle, veri setindeki bireylerin ne kadarının doğru sınıfa atandığını ortaya koyar. Bununla birlikte doğruluk değeri tek başına değerlendirilmemeli; özellikle PÖD, NÖD ve diğer tanısal performans göstergeleriyle birlikte ele alınmalıdır (Shaikh, 2011). Doğruluk, sınıflardaki bireylerin dağılımından etkilenmektedir. Duyarlılık ve özgüllük sabit tutulduğunda, hastalık prevalansının azalması doğruluk değerinin yükselmesine neden olabilir. Ancak bu artış testin düşük prevalanslı popülasyonda daha başarılı hâle geldiği şeklinde yorumlanmamalıdır. Buradaki değişim, esas olarak popülasyondaki negatif bireylerin sayısının artması sonucunda doğru sınıflandırılan toplam birey sayısının değişmesinden kaynaklanmaktadır (Simundic,

2009; Chen ve ark., 2015). Doğruluk aşağıdaki eşitlikle hesaplanmaktadır:

$$Doğruluk = (GP + GN)/(GP + GN + YP + YN) \quad (7)$$

Doğruluk 0 ile 1 arasında değer alır. 1 değeri tüm bireylerin doğru sınıflandırıldığını, 0 değeri ise hiçbir bireyin doğru sınıflandırılmadığını gösterir. Ancak özellikle sınıfların dengesiz olduğu veri setlerinde yüksek doğruluk yanıltıcı olabileceğinden, bu ölçütün tek başına kullanılması uygun değildir.

Youden İndeksi

Youden indeksi, tanı testinin genel ayırt etme kapasitesini tek bir değerle ifade etmek amacıyla kullanılan ölçütlerden biridir. Literatürde bookmaker informedness veya doğru sınıflandırma gücü olarak da adlandırılmaktadır. Duyarlılık ile özgüllüğü birlikte dikkate alması nedeniyle farklı tanı testlerinin karşılaştırılmasında kullanılabilir. Youden indeksinin teorik aralığı 0 ile 1 arasındadır. Değerin 0 olması testin ayırt edici gücünün bulunmadığını, 1 olması ise mükemmel bir tanısal ayırım yapıldığını ifade eder. Bununla birlikte indeks, duyarlılık ve özgüllük arasındaki dağılıma duyarlı değildir. Örneğin $Se = 0.70$ ve $Sp = 0.80$ olan bir test ile $Se = 0.90$ ve $Sp = 0.60$ olan bir testin Youden indeksleri aynıdır ve her iki test için değer 0.50'dir. Dolayısıyla aynı Youden indeksine sahip iki testin klinik kullanım amaçları veya yanlış sınıflandırma profilleri birbirinden farklı olabilir (Shaikh, 2011). Youden indeksi prevalanstan bağımsız olmakla birlikte hastalık spektrumundan ve buna bağlı olarak duyarlılık, özgüllük ve olabilirlik oranlarından etkilenebilir. Değerin sıfırın altında olması, testin pratik açıdan anlamlı bir ayırt edicilik sağlamadığını düşündürür (Simundic, 2009; Chen ve ark., 2015). Youden indeksinin değeri eşitlik (8) kullanılarak hesaplanmaktadır.

$$Youden = Se + Sp - 1 \quad (8)$$

Tanısal Odds Oranı

Tanısal odds oranı (DOR; *diagnostic odds ratio*), bir tanı testinin ayırt edici gücünü tek bir gösterge üzerinden değerlendirmek amacıyla kullanılan ölçütlerden biridir. Tanı testlerinin performansını belirlemek için duyarlılık, özgüllük, öngörü değerleri ve ROC eğrisi altında kalan

alan (AUC) gibi çok sayıda gösterge kullanılmaktadır. DOR ise bu göstergeler arasında, testin pozitif ve negatif sonuçlarının hastalık durumuyla olan ilişkisini tek bir sayısal değerle özetleme özelliğine sahiptir (Huang, Yin ve Samawi, 2018). DOR, hastalığı bulunan bireylerde pozitif test sonucu elde edilme olasılığının, hastalığı bulunmayan bireylerde pozitif test sonucu elde edilme olasılığına oranını ifade eder. Değer yükseldikçe testin hastalığı olan ve olmayan bireyleri birbirinden ayırma gücü artar. DOR=1 olduğunda testin iki grup arasında ayırt edici bilgi sağlamadığı kabul edilir. Birden büyük değerler genel olarak daha iyi ayırt ediciliğe işaret ederken, 1'den küçük değerler testin ters yönde sınıflandırma eğiliminde olduğunu gösterebilir (Deeks, 2001). DOR, duyarlılık ve özgüllük ile birlikte PÖD, NÖD, olabilirlik oranları ve AUC gibi diğer tanısal göstergelerle beraber değerlendirilebilir (Casanova ve ark., 2022). Örneğin DOR=36 olması, hastalığı bulunan bireylerde pozitif test sonucunun ortaya çıkma odds'unun hastalığı bulunmayan bireylere göre 36 kat daha yüksek olduğunu ifade eder. 2x2 tablosunda hücrelerden herhangi birinin sıfır olması durumunda DOR tanımsız hâle gelebilir. Bu gibi durumlarda hücrelere 0,5 eklenmesi, yaklaşık bir DOR tahmini elde etmek için yaygın olarak kullanılan düzeltmelerden biridir (Glas ve ark., 2003; Raslich, Markert ve Stutes, 2007). DOR'un başlıca avantajları; duyarlılık ve özgüllüğü tek bir performans göstergesinde birleştirmesi, hesaplanmasının görece kolay olması, 2x2 tablosundan doğrudan elde edilebilmesi ve farklı tanı testlerinin karşılaştırılmasına olanak vermesidir. Bunun yanında bazı sınırlılıkları da bulunmaktadır. DOR mutlak hastalık riski hakkında bilgi vermez ve klinik karar verme sürecinde doğrudan yorumlanması güç olabilir. Ayrıca çok yüksek duyarlılık ancak düşük özgüllük veya bunun tersi gibi dengesiz performans profillerine sahip testlerin değerlendirilmesinde tek başına yeterli olmayabilir. Vaka-kontrol tasarımlarında DOR'un olduğundan yüksek tahmin edilebilmesi ve farklı referans standartların kullanılmasının sonuçları etkileyebilmesi de dikkate alınmalıdır (Sousa ve Ribeiro, 2009; Zhou, Obuchowski ve McClish, 2011). Bir tanı testinin klinik kullanıma alınmasından önce performansının uygun ölçütlerle değerlendirilmesi gerekir. Kullanılacak ölçütün seçimi, test sonucunun ikili, sıralı veya sürekli yapıda olmasına bağlı olarak değişebilir (Lim, 2021). DOR aşağıdaki eşitlikle hesaplanabilir:

$$DOR = \frac{GP}{YN} \Big/ \frac{YP}{GN} = \frac{duyarlılık}{(1 - duyarlılık)} \Big/ \frac{(1 - özgüllük)}{özgüllük} \quad (9)$$

DOR aynı zamanda test pozitifleri arasındaki hastalık odds'unun test negatifleri arasındaki hastalık odds'una oranı şeklinde de ifade edilebilir:

$$DOR = \frac{GP}{YN} \Big/ \frac{YN}{GN} = \frac{PPV}{(1 - PPV)} \Big/ \frac{(1 - NPV)}{NPV} \quad (10)$$

Olabilirlik oranlarıyla DOR arasında doğrudan bir ilişki bulunmaktadır:

$$DOR = \frac{GP}{YN} \Big/ \frac{YN}{GN} = \frac{LR(+)}{LR(-)} \quad (11)$$

DOR'un teorik aralığı 0 ile +∞ arasındadır. Eşitliklerden görülebileceği üzere prevalans DOR'un matematiksel hesabında doğrudan yer almamaktadır. Bununla birlikte, test performansının hastalık spektrumundan etkilenebilmesi nedeniyle klinik uygulamalarda DOR'un dolaylı olarak popülasyon özelliklerinden etkilenmesi mümkündür (Glas ve ark., 2003). DOR'un hesaplanması kolay olmasına rağmen klinik açıdan yorumlanması her zaman basit değildir. Genel olarak DOR'un 1'den büyük olması ayırt edici bir test performansını, daha yüksek değerler ise daha güçlü ayırt ediciliği ifade eder (Deeks, 2001). DOR için güven aralığı, logaritmik dönüşüm kullanılarak hesaplanabilir. Öncelikle log DOR'un standart hatası:

$$SE(\ln DOR) = \sqrt{\frac{1}{GP} + \frac{1}{GN} + \frac{1}{YP} + \frac{1}{YN}} \quad (12)$$

olarak hesaplanır. %95 güven aralığı ise:

$$\ln DOR \mp 1{,}96 * SE(\ln DOR) \quad (13)$$

eşitliği ile elde edilir. Son aşamada alt ve üst sınırların üstel dönüşümü alınarak DOR için güven aralığı hesaplanır (Glas ve ark., 2003).

Klinik Fayda İndeksi

Klinik fayda indeksi (CUI; *clinical utility index*), tanı testinin yalnızca teknik doğruluğunu değil, elde edilen test sonucunun klinik açıdan ne ölçüde yararlı olduğunu değerlendirmek amacıyla geliştirilmiştir. Mitchell (2008), klinik faydayı test sonucunun oluşum sıklığı ve ayırt edicilik özellikleriyle ilişkilendirmiştir. PÖD veya NÖD yüksek olsa bile ilgili test sonucunun toplumda ortaya çıkma sıklığı düşük olabilir. Bu nedenle testin gerçek klinik kullanım değerinin belirlenmesinde prevalansın da dikkate alınması gerekmektedir. Bu yaklaşım kapsamında pozitif sonuçlar için CUI+, negatif sonuçlar için ise CUI- kullanılmaktadır. CUI değerlerinin yorumlanması şu şekilde yapılmaktadır (Mitchell, 2008):

- CUI ≥ 0,81: Mükemmel fayda
- 0,64 ≤ CUI < 0,81: İyi fayda
- 0,49 ≤ CUI < 0,64: Orta düzeyde fayda
- 0,36 ≤ CUI < 0,49: Zayıf fayda
- CUI < 0,36: Çok zayıf fayda

Pozitif ve negatif klinik fayda indeksleri sırasıyla:

$$CUI+ = Se * PPV \quad (14)$$

$$CUI- = Sp * NPV \quad (15)$$

şeklinde hesaplanmaktadır. Bir testin klinik yararı; hastalığı olan ve olmayan bireylerdeki test sonuçlarının dağılımı, testin tanısal doğruluğu ve test öncesi hastalık olasılığı gibi faktörlerin ortak etkisiyle belirlenir. Test öncesi olasılığın tedavi eşiğine karşılık geldiği durumda CUI maksimum değerine ulaşmaktadır. Bu durumda CUI'nin sayısal değeri Youden indeksi ile aynı olabilir; ancak iki ölçütün kavramsal temelleri birbirinden farklıdır. Bu yaklaşım, testin klinik gereksinimin en yüksek olduğu durumda daha fazla fayda sağlayabileceğini ortaya koymaktadır. Bununla birlikte, test öncesi hastalık olasılığının güvenilir biçimde belirlenmesi uygulamada önemli bir güçlük oluşturmaktadır (Asberg, Mikkelsen ve Odsaeter, 2019).

Yaklaşık Korelasyon

Yaklaşık korelasyon (*approximate correlation*), gerçek sınıflandırma ile model veya test tarafından üretilen sınıflandırma arasındaki ilişkinin gücünü değerlendirmek amacıyla kullanılmaktadır. Baldi ve ark. (2000) tarafından önerilen yaklaşım, 2x2 sınıflandırma tablosundaki farklı hücreleri birlikte değerlendirerek sınıflandırma performansı hakkında genel bir gösterge sunmaktadır. İlk olarak ACP değeri:

$$ACP = \frac{1}{4}\left[\frac{GP}{GP+YN} + \frac{GP}{GP+YP} + \frac{GN}{GN+YP} + \frac{GN}{GN+YN}\right]$$

olarak hesaplanır ve yaklaşık korelasyon:

$$AC = 2 * (ACP - 0.5) \quad (16)$$

eşitliği ile elde edilir (Baldi ve ark., 2000). AC değeri -1 ile +1 arasında değişmektedir. +1'e yaklaşan değerler gerçek ve tahmin edilen sınıflandırmalar arasında güçlü bir uyuma, 0 civarındaki değerler rastlantısal sınıflandırmaya, -1'e yaklaşan değerler ise ters yönlü sınıflandırmaya işaret eder. Böylece ölçüt, yalnızca pozitif sınıfın değil, her iki sınıfın doğru sınıflandırılma durumunu dikkate almaktadır.

Dengeli Doğruluk

Dengeli doğruluk (*balanced accuracy*), özellikle sınıf dağılımının eşit olmadığı veri setlerinde doğruluğun tamamlayıcısı olarak değerlendirilebilecek önemli bir performans ölçütüdür. İkili sınıflandırmada her sınıfın doğru sınıflandırılma oranı ayrı ayrı hesaplanır ve bu oranların aritmetik ortalaması alınır. Çok sınıflı problemlerde ise her kategoriye eşit ağırlık verilmesi temel ilkedir. Veri setindeki sınıflar yaklaşık olarak eşit büyüklükte olduğunda klasik doğruluk ile dengeli doğruluk birbirine yakın sonuçlar verebilir. Buna karşılık sınıfların büyüklükleri arasında belirgin fark olduğunda doğruluk, çoğunluk sınıfının etkisiyle yüksek çıkabilir. Dengeli doğrulukta her sınıf nihai hesaplamaya eşit katkıda bulunduğundan, azınlık sınıfının performansı göz ardı edilmemiş olur. Bu nedenle dengesiz veri setlerinde özellikle yararlı bir ölçüt olarak kabul

edilmektedir (Grandini, Bagli ve Visani, 2008). İkili sınıflandırma için dengeli doğruluk:

$$BA = (Se + Sp)/2 \quad (17)$$

olarak hesaplanır. Değer 0 ile 1 arasında değişir ve yüksek değerler daha başarılı sınıflandırmayı gösterir.

F_1 Skoru

F_1 skoru, duyarlılık ile kesinlik (precision) arasındaki dengeyi tek bir değerle ifade etmek amacıyla kullanılan bir ölçüttür. Dice benzerlik katsayısıyla da ilişkilendirilen bu gösterge, iki temel performans bileşeninin harmonik ortalamasına dayanır. Özellikle pozitif sınıfın doğru belirlenmesine odaklanan sınıflandırma problemlerinde yaygın biçimde kullanılmaktadır. Duyarlılık veya precision değerlerinden herhangi biri sıfır olduğunda F_1 skoru da sıfır olur. Her iki değer sıfırdan farklı olduğunda ölçüt 0 ile 1 arasında değişir. F_1 skorunun önemli sınırlılıklarından biri, 2x2 tablosundaki tüm bilgiyi kullanmamasıdır. Ayrıca sınıf dağılımındaki değişimlerden etkilenmesi nedeniyle dengesiz veri setlerinde tek başına değerlendirilmesi yanıltıcı sonuçlar doğurabilir (Foody, 2023). F_1 skoru:

$$F_1 = 2 * \frac{GP}{2 * GP + YP + YN} \quad (18)$$

eşitliğiyle hesaplanmaktadır. Teorik aralığı 0 ile 1 arasındadır. 1 değeri mükemmel performansı, 0 değeri ise başarısız sınıflandırmayı ifade eder.

Sınıflandırma Belirginliği

Belirginlik (*Markedness*, MD), test sonucunun gerçek sınıfla ne ölçüde ilişkili olduğunu değerlendirmek amacıyla PÖD ve NÖD üzerinden tanımlanan bir ölçüttür. Bu ölçütün temel özelliği, testin tahmin edilen sınıflar açısından bilgi taşıma düzeyini ortaya koymasıdır. MD aşağıdaki şekilde hesaplanmaktadır:

$$MD = PÖD + NÖD - 1 \quad (19)$$

MD -1 ile +1 arasında değişir. +1 değeri pozitif ve negatif sınıfların kusursuz biçimde tahmin edildiğini, 0 değeri sınıflandırmanın şans düzeyinde bilgi sağladığını, -1 değeri ise sınıflandırmanın tamamen ters yönde olduğunu ifade eder. Bununla birlikte MD'nin PÖD ve NÖD'ye bağlı olması, dolayısıyla veri dağılımındaki değişikliklerden etkilenmesi, özellikle dengesiz veri setlerinde önemli bir sınırlılık oluşturmaktadır (Tharwat, 2021).

Yaygınlık Eşiği

Yaygınlık eşiği (*Prevalence Threshold*, PT), sınıflandırıcının belirli bir prevalans düzeyinde klinik veya pratik açıdan anlamlı bir performans sergileyebilmesi için gerekli olan eşik değeri ifade etmektedir. Chicco ve Jurman (2023) tarafından önerilen eşitlik kullanılarak:

$$PT = \frac{1}{\left\{1 + \sqrt{(GP/(GP+YN)) * \left((GN+YP)/YP\right)}\right\}} \quad (20)$$

şeklinde hesaplanmaktadır. PT değeri 0 ile 1 arasında değişir. Daha düşük PT değerleri, sınıflandırıcının düşük prevalans koşullarında dahi etkili olabildiğini gösterirken, yüksek PT değerleri anlamlı bir performans için daha yüksek prevalans gerektiğine işaret eder. Dolayısıyla diğer koşullar benzer olduğunda düşük PT değerleri tercih edilmektedir (Chicco ve Jurman, 2023).

Fowlkes-Mallows İndeksi

Fowlkes-Mallows indeksi (FMI), özellikle doğru pozitif sınıflandırmaların gücünü duyarlılık ve precision bileşenleri üzerinden değerlendiren bir ölçüttür. Aşağıdaki eşitlik kullanılarak hesaplanır (Chicco ve Jurman, 2023):

$$FMI = GP/\sqrt{(GP+YP) * (GP+YN)} \quad (21)$$

FMI 'nın değeri 0 ile 1 arasında değişmektedir. 1 değeri yanlış pozitif ve yanlış negatif sınıflandırmaların bulunmadığı ideal durumu

gösterirken, düşük değerler sınıflandırma performansının zayıf olduğunu ifade eder. Bu nedenle daha yüksek FMI değerleri daha başarılı sınıflandırma performansına karşılık gelir.

Yaygınlık

Yaygınlık veya prevalans (P), incelenen referans veri setinde pozitif sınıfa ait bireylerin toplam bireyler içerisindeki oranını göstermektedir. Bu nedenle prevalans yalnızca bir hastalığın toplumdaki görülme sıklığını ifade etmekle kalmaz, aynı zamanda veri setindeki sınıf dağılımının dengeli olup olmadığı hakkında da bilgi verir. Pozitif ve negatif sınıfların sayıları eşit olduğunda prevalans 0.5'tir ve sınıf dağılımı 1:1 oranındadır. Prevalansın 0.5'ten uzaklaşması, sınıflar arasındaki dengesizliğin artması anlamına gelir (Tsokani ve ark., 2022; Foody, 2023). Prevalans:

$$P = (GP + YN)/(GP + GN + YP + YN). \quad (22)$$

eşitliğiyle hesaplanmaktadır. Değer 0 ile 1 arasında değişir. $P = 0$ veri setinde pozitif sınıf bulunmadığını, $P = 1$ ise tüm gözlemlerin pozitif sınıfa ait olduğunu gösterir. $P = 0.5$ olması iki sınıfın eşit büyüklükte olduğunu ifade eder. Bu nedenle 0.5'ten uzaklaşma, sınıf dengesizliğinin derecesi hakkında bilgi vermektedir.

Jaccard İndeksi

Jaccard indeksi, gerçek ve tahmin edilen pozitif sınıflar arasındaki örtüşme düzeyini belirlemek amacıyla kullanılan benzerlik ölçütlerinden biridir. Özellikle pozitif sınıfa ilişkin doğru ve hatalı sınıflandırmaları birlikte dikkate alarak tahmin edilen pozitif kümenin gerçek pozitif kümeyle ne ölçüde örtüştüğünü gösterir (Abhishek ve Hamarneh, 2010).

$$J = GP/(GP + YP + YN) \quad (23)$$

Jaccard indeksi 0 ile 1 arasında değişmektedir. 1 değeri gerçek ve tahmin edilen pozitif sınıfların tam olarak örtüştüğünü, 0 değeri ise

doğru pozitif sınıflandırmanın bulunmadığını ifade eder. Yüksek Jaccard değerleri daha iyi sınıflandırma performansına işaret eder.

Cohen Kappa Katsayısı

Cohen Kappa katsayısı, gerçek sınıflandırma ile tahmin edilen sınıflandırma arasındaki uyumu ölçerken, yalnızca gözlenen uyumu değil, tesadüfen ortaya çıkabilecek uyumu da hesaba katan bir istatistiktir. Bu özelliği, farklı veri setlerinde veya farklı sınıf dağılımlarına sahip örneklerde yapılan sınıflandırmaların karşılaştırılmasına katkı sağlamaktadır (Grandini, Bagli ve Visani, 2008). İkili sınıflandırma için Kappa katsayısı:

$$Kappa = \frac{2 * (GP * GN - YP * YN)}{(GP + YP) * (GP + YN) + (GN + YP) * (GN + YN)} \quad (24)$$

eşitliği ile hesaplanmaktadır. Kappa katsayısı -1 ile +1 arasında değer alır. +1, gerçek ve tahmin edilen sınıflandırmalar arasında tam uyuma; 0, gözlenen uyumun tesadüfi uyum düzeyinde olmasına; negatif değerler ise sınıflandırmalar arasında ters yönlü bir ilişki bulunmasına karşılık gelir. Dolayısıyla yüksek Kappa değerleri, testin gerçek durum ile tahmin edilen durum arasında güçlü bir uyum sağladığını göstermektedir.

Brier Skoru

Brier skoru (BS), özellikle ikili sonuçlar için olasılıksal tahminlerin performansını değerlendirmek amacıyla kullanılan bir ölçüttür. Klinik ve epidemiyolojik çalışmalarda hastalık gelişimi, tedaviye yanıt veya tarama sonucu gibi ikili sonuçların tahmin edilmesinde kullanılabilmektedir. Bununla birlikte Brier skorunun yorumlanması, sonucun prevalansından etkilenmesi nedeniyle her zaman doğrudan değildir (Yang ve ark., 2022). Brier skorunun sıfır olması her durumda modelin kusursuz olduğu anlamına gelmez. Gerçek yaşam koşullarında, doğru biçimde tanımlanmış modellerde dahi sıfırdan farklı Brier skorları elde edilebilir. Bunun nedeni skorun yalnızca tahmin hatasını değil, aynı zamanda sonuç değişkeninin dağılımını ve rastlantısal varyasyonu da yansıtmasıdır. Bu nedenle farklı popülasyonlarda elde edilen Brier skorlarının doğrudan

karşılaştırılması dikkat gerektirir. Ayrıca Brier skoru doğrudan kalibrasyon ölçütü olarak değerlendirilmemelidir (Hoessly, 2026). Brier skoru ile ortalama karesel hata (MSE) arasında matematiksel bir bağlantı bulunmaktadır. Bununla birlikte iki ölçütün kullanım bağlamı farklıdır. Brier skoru ikili bir olayın gerçekleşmesine ilişkin tahmin edilen olasılık ile gerçekleşen sonucu karşılaştırırken, MSE daha genel olarak tahmin edilen ve gerçek sürekli değerler arasındaki farkın karesel ortalamasını ifade eder. Verilen sınıflandırma yaklaşımında Brier skoru:

$$BS = (YP + YN)/(GP + GN + YP + YN) \tag{25}$$

eşitliğiyle hesaplanmaktadır. Bu yaklaşımda Brier skoru tanısal doğruluğun tamamlayıcısı olarak ele alınmakta ve:

$$BS = 1 - tanısal\ doğruluk$$

olarak ifade edilmektedir (Shaikh, 2011). Brier skorunun değeri 0 ile 1 arasında değişir. Düşük değerler daha başarılı, yüksek değerler ise daha hatalı sınıflandırmayı gösterir. Bu nedenle en iyi değer 0, en kötü değer ise 1 olarak kabul edilmektedir.

Kullback–Leibler Sapması/Uzaklığı

Kullback–Leibler sapması (D_{KL}), iki olasılık dağılımının birbirinden ne ölçüde farklı olduğunu belirlemek için kullanılan bilgi kuramsal bir ölçüttür. Makine öğrenmesi, istatistik ve bilgi teorisi başta olmak üzere çeşitli alanlarda dağılımlar arasındaki farklılığın nicel olarak ifade edilmesinde kullanılmaktadır. Ayrıca aykırı gözlemlerin belirlenmesi ve örnekler arasındaki benzerliğin değerlendirilmesi gibi uygulamalarda da teorik bir temel sağlayabilir (Zhong, Liu ve Chen, 2020; Zhang ve ark., 2023). P ve Q dağılımları arasındaki Kullback–Leibler sapması:

$$D_{KL} = \sum_{k} ln\left(\frac{P(k)}{Q(k)}\right) P(k) \tag{26}$$

şeklinde tanımlanır. D_{KL} klasik anlamda bir uzaklık metriği değildir. Bunun temel nedeni, iki dağılımın yer değiştirilmesiyle aynı değerin elde edilmesinin gerekmemesidir; yani genel

olarak $D_{KL}(P, Q) \neq D_{KL}(Q, P)$. P ve Q aynı dağılım olduğunda $D_{KL} = 0$ olur; dağılımlar farklılaştıkça değer pozitif hâle gelir (Zhong, Liu ve Chen, 2020). D_{KL}'nin çeşitli avantajları bulunmaktadır. Dağılımlar arasındaki farklılığı tek bir sayısal değerle ifade edebilmesi, olasılık dağılımlarının karşılaştırılmasında kullanılabilmesi ve makine öğrenmesi uygulamalarına görece kolay biçimde uyarlanabilmesi bunlardan bazılarıdır. Bunun yanında yüksek boyutlu dağılımlarda hesaplama maliyetinin artması, üçgen eşitsizliğini sağlamaması ve bazı dağılım biçimlerinde özellikle kuyruk davranışları nedeniyle sorunlar oluşturabilmesi önemli sınırlılıklar arasındadır (Hu ve Hong, 2013). Tıbbi tanı açısından Kullback-Leibler sapmasının kullanımı da araştırılmış ve tanısal belirteçlerin performansını değerlendirmek amacıyla bir gösterge olarak ele alınmıştır (Samawi ve ark., 2020). Lee (1999), bu yaklaşımı tanı testlerinin performansını karakterize etmek için kullanmış ve hastalığı belirleme ile hastalığı dışlama potansiyelinin değerlendirilmesine yönelik genel bir ölçüt olarak önermiştir. Gerçekte hasta ve sağlıklı bireylerden elde edilen test sonuçlarını temsil eden dağılımlar f ve g olarak tanımlandığında, hastalığı belirleme potansiyeline ilişkin Kullback–Leibler sapması:

$$D(f \,||\, g) = (1 - Se) * log_e \frac{1 - Se}{Sp} + Se * log_e \frac{Se}{1 - Sp} \qquad (27)$$

şeklinde hesaplanabilir. Hastalığı dışlama potansiyeli ise:

$$D(g \,||\, f) = Sp * log_e \frac{Sp}{1 - Se} + (1 - Sp) * log_e \frac{1 - Sp}{Se} \qquad (28)$$

eşitliği ile ifade edilmektedir. Burada Se duyarlılığı, Sp ise özgüllüğü göstermektedir. $D(f \,||\, g)$ ve $D(g \,||\, f)$ değerleri sırasıyla tanı testinin hastalığı belirleme ve hastalığı dışlama potansiyeli bağlamında yorumlanmaktadır (Lee, 1999). Bu iki ölçütün simetrik olmaması nedeniyle aynı tanı testi için elde edilen değerler birbirinden farklı olabilir. Teorik olarak değerler 0 ile +∞ arasında değişmektedir (Bonnici, 2024). Bu yaklaşım, duyarlılık ve özgüllük toplamları aynı olan testlerin farklı tanısal özelliklere sahip olabileceğini göstermesi

açısından önemlidir. Örneğin $Se = 0.80$ ve $Sp = 0.90$ olan bir test ile $Se = 0.90$ ve $Sp = 0.80$ olan başka bir testin Youden indeksleri aynıdır. Buna karşın Kullback–Leibler sapmalarının yönü ve büyüklüğü farklı olabilir. İlk test için $D(f \parallel g)$ =1.36 ve ikinci test için 1.15 bulunurken, $D(g \parallel f)$ değerleri sırasıyla 1,15 ve 1,36 olarak hesaplanmaktadır. Bu sonuçlar ilk testin hastalığı belirleme, ikinci testin ise hastalığı dışlama yönündeki performansının göreli olarak daha yüksek olduğunu göstermektedir (Lee, 1999). $D(f \parallel g)$ ve $D(g \parallel f)$ kullanılırken bazı matematiksel koşulların dikkate alınması gerekir:

- Duyarlılık veya özgüllükten herhangi birinin hesaplanamadığı durumda bu ölçütler de hesaplanamaz.
- Duyarlılık veya özgüllük değerlerinden herhangi birinin 0 ya da 1 olması durumunda ilgili logaritmik ifadeler nedeniyle ölçütün hesaplanmasında sorun ortaya çıkabilir.
- Yanlış pozitif veya yanlış negatif sayıların gerçek pozitif ve gerçek negatif sayılara göre aşırı yüksek olması, sapma değerlerinin büyümesine neden olabilir.
- $YP > YN$ olduğunda $D(g \parallel f) > D(f \parallel g)$ ilişkisi ortaya çıkabilir.
- $GP > GN$ olduğunda $D(f \parallel g) > D(g \parallel f)$ olması beklenebilir.
- Aynı 2x2 tablo için duyarlılık ve özgüllük toplamı 1 olduğunda her iki sapma değeri de 0'a eşit olur.

Bu özellikler, Kullback–Leibler yaklaşımının duyarlılık ve özgüllüğün basit toplamından farklı olarak testin hangi yöndeki ayırt ediciliğinin daha baskın olduğunu ortaya koyabilmesine olanak sağlamaktadır.

Matthews Korelasyon Katsayısı

Matthews korelasyon katsayısı (MCC), ilk olarak Matthews (1975) tarafından biyokimyasal bir çalışma kapsamında geliştirilmiş ve daha sonra ikili sınıflandırma problemlerinin değerlendirilmesinde kullanılmaya başlanmıştır. Özellikle sınıf dağılımlarının dengesiz olduğu veri setlerinde performansı daha dengeli biçimde yansıtabilmesi nedeniyle son yıllarda daha fazla ilgi görmektedir (Matthews, 1975; Chicco, Warrens ve Jurman, 2021). MCC, gerçek ve tahmin edilen sınıflandırmalar arasındaki ilişkiyi korelasyon yaklaşımıyla değerlendirmektedir. İkili sınıflandırma için:

$$MCC = \frac{GP * GN - YP * YN}{\sqrt{(GP + YP) * (GP + YN) * (GN + YP) * (GN + YN)}} \quad (29)$$

eşitliği kullanılmaktadır. MCC, -1 ile +1 arasında değer alır. +1, tahmin edilen ve gerçek sınıflandırmaların tamamen uyumlu olduğu ideal durumu; 0, iki sınıflandırma arasındaki ilişkinin şans düzeyinde olduğunu; -1 ise gerçek ve tahmin edilen sınıflandırmaların tamamen ters yönde olduğunu ifade eder (Chicco, Starovoitov ve Jurman, 2021; Tamura ve ark., 2025). MCC'nin önemli özelliklerinden biri, sınıf dengesizliğinin bulunduğu durumlarda yalnızca çoğunluk sınıfının performansına dayalı yüksek doğruluk değerlerinden kaynaklanabilecek yanıltıcılığı azaltabilmesidir. Bu nedenle dengesiz sınıflandırma problemlerinde doğruluk ve F_1 skoru gibi ölçütlere kıyasla daha bilgilendirici bir sonuç sağlayabildiği bildirilmektedir (Chicco ve Jurman, 2023). MCC aynı zamanda Pearson korelasyon katsayısının ikili sınıflandırma verilerine uyarlanmış özel bir biçimi olarak değerlendirilebilir. Bununla birlikte MCC'nin her sınıflandırma problemi için otomatik olarak en uygun ölçüt olduğu söylenemez. Ölçütün seçimi araştırma sorusuna, sınıf dağılımına ve yanlış pozitif ile yanlış negatif sonuçların göreli önemine bağlı olarak yapılmalıdır. Buna rağmen, özellikle dengesiz sınıfların bulunduğu problemlerde sınıflandırmanın genel performansını değerlendirmek için güçlü ve dengeli bir gösterge olduğu belirtilmektedir (Stoica ve Babu, 2023;

Tamura ve ark., 2025). MCC'nin biyostatistik açısından önemi, sınıflandırma sonuçlarının hasta prognozu veya kritik klinik kararlar gibi sonuçları doğrudan etkileyebilmesinden kaynaklanmaktadır. Sınıf dengesizliğinin bulunduğu durumlarda bile dengeli bir performans değerlendirmesi sağlaması nedeniyle güvenilir ölçütlerden biri olarak kabul edilmektedir (Chicco ve Jurman, 2020; Itaya ve ark., 2025). MCC'nin yüksek bir değer üretebilmesi için duyarlılık, özgüllük, precision ve negatif öngörü değeri gibi temel bileşenlerin birlikte güçlü olması gerekir. Bu özellik, MCC'nin yalnızca doğruluğa veya F_1 skoruna dayalı değerlendirmelere kıyasla sınıflandırma performansı hakkında daha kapsamlı bilgi sunmasını sağlayabilir (Chicco ve Jurman, 2022; Tamura ve ark., 2025).

Öklid İndeksi

Öklid İndeksi bir hata ölçüsüdür. Diğer birçok performans ölçütünün aksine küçük değerler daha iyi performansı gösterir. Öklid indeksinin değeri aşağıda verilen eşitlik kullanılarak hesaplanmaktadır (Wu ve ark., 2022).

$$d = \sqrt{(1 - Se)^2 + (1 - Sp)^2} \tag{30}$$

Öklid İndeksi'nin değeri 0 ile $\sqrt{2} \approx 1.414$ arasında değişmektedir. İndeksin 0 değerini alması duyarlılık ve özgüllüğün 1'e eşit olduğu ideal sınıflandırma durumunu göstermektedir. İndeks değerinin artması, sınıflandırma sonucunun ideal noktadan uzaklaştığını ve performansın azaldığını ifade etmektedir. Bu nedenle düşük Öklid İndeksi değerleri daha iyi sınıflandırma performansını göstermektedir.

Çarpım İndeksi

Çarpım indeksinin değeri aşağıda verilen eşitlik kullanılarak hesaplanmaktadır (Wu ve ark., 2022).

$$CZ = Se * Sp \tag{31}$$

Çarpım İndeksi'nin değeri 0 ile 1 arasında değişmektedir. İndeksin 1 değerini alması duyarlılık ve özgüllüğün 1'e eşit olduğu mükemmel sınıflandırma durumunu göstermektedir. İndeksin 0 değerini alması ise duyarlılık veya özgüllük değerlerinden en az birinin sıfır olduğunu

ifade etmektedir. Yüksek CZ değerleri daha başarılı sınıflandırma performansını göstermektedir. En kötü değeri 0, en iyi değeri +1'dir.

Taguchi İndeksi

Taguchi indeksinin değeri aşağıda verilen eşitlik kullanılarak hesaplanmaktadır (Wu ve ark., 2022).

$$TI = K_1 * \left(\frac{1}{Sp} - 1\right) + K_2 * \left(\frac{1}{Se} - 1\right) \quad (32)$$

K1 ve K2 kayıp katsayılarıdır. Taguchi İndeksi'nin değeri 0 ile $+\infty$ arasında değişmektedir. İndeksin 0 değerini alması duyarlılık ve özgüllüğün 1'e eşit olduğu, yani mükemmel sınıflandırma durumunu göstermektedir. İndeks değerinin artması sınıflandırma hatalarının ve buna bağlı kayıp maliyetinin arttığını ifade etmektedir. Bu nedenle düşük Taguchi İndeksi değerleri daha iyi sınıflandırma performansını göstermektedir. En iyi değeri 0, en kötü değeri $+\infty$'dur.

Ayırt Edici Güç

Duyarlılık ve özgüllüğü özetleyen bir diğer ölçüt ise ayırt edici güçtür (discriminant power = DP). Ayırt edici güç, bir testin pozitif ve negatif örnekler arasında ne kadar iyi ayrım yaptığını değerlendirir. Duyarlılık ve özgüllük ölçütlerine bağlı olduğundan, dengesiz verilerle kullanılabilir. Ayırt edici güç değeri için testin ayırt edici gücü; $DP < 1$ ise zayıf, $1 \leq DP < 2$ ise sınırlı, $2 \leq DP < 3$ ise orta, diğer durumlarda ise iyi şeklinde değerlendirilir (Nellore, 2015). Ayırt edici güç değeri aşağıda verilen eşitlik kullanılarak hesaplanmaktadır.

$$DP = \frac{\sqrt{3}}{\pi}(logx + logy) \quad (33)$$

$$logx = \frac{Duyarlılık}{1 - Duyarlılık}$$

$$logy = \frac{Özgüllük}{1 - Özgüllük}$$

Ayırt Edici Güç'ün değeri 0 ile +∞ arasında değişmektedir. DP değerinin artması, testin pozitif ve negatif sınıfları birbirinden daha iyi ayırabildiğini göstermektedir. DP'nin 1'in altında olması zayıf ayırt edici gücü ifade ederken, yüksek DP değerleri güçlü sınıflandırma performansını göstermektedir.

Yanlış Bulgu Oranı

Yanlış bulgu oranı ($FDR; false\ discovery\ rate$), yanlış pozitiflerin sayısının, pozitif olarak tahmin edilen vaka sayısına bölünmesiyle elde edilir (Berrar, 2019).

$$FDR = \frac{YP}{YP + GP} \tag{34}$$

Yanlış bulgu oranının değeri 0 ile 1 arasında değişmektedir. FDR değerinin 0 olması yanlış pozitif bulunmadığını, 1 olması ise pozitif olarak sınıflandırılan tüm örneklerin yanlış olduğunu göstermektedir. Düşük FDR değerleri model performansının daha iyi olduğuna işaret etmektedir.

Birleşik Performans Ölçütü

Birleşik performans ölçütü ($UPM; unified\ performance\ measure$), F_1^+ ve F_1^-'nin harmonik ortalaması olarak tanımlanır. Dolayısıyla UPM hem pozitif hem de negatif sınıftaki performansı değerlendirir. Bu performans ölçütü, yalnızca dört temel ölçüt olan PÖD, Duyarlılık, Özgüllük ve NÖD'ün de yüksek değerlere sahip olması durumunda yüksek değerlere sahiptir. Ayrıca UPM, dengesiz verilerle de uyumludur (De Diego ve ark., 2022). Birleşik performans ölçütünün değeri aşağıda verilen eşitlik kullanılarak hesaplanmaktadır.

$$UPM = 2 * \frac{F_1^+ * F_1^-}{F_1^+ + F_1^-} \tag{35}$$

$$F_1^+ = 2 * \frac{PÖD * Duyarlılık}{PÖD + Duyarlılık}$$

$$F_1^- = 2 * \frac{\text{NÖD} * \text{Özgüllük}}{\text{NÖD} + \text{Özgüllük}}$$

Birleşik Performans Ölçütü'nün değeri 0 ile 1 arasında değişmektedir. UPM değerinin 1 olması, pozitif ve negatif sınıflar için hesaplanan tüm temel performans ölçütlerinin en yüksek değerlerini aldığı mükemmel sınıflandırma durumunu göstermektedir. UPM değerinin 0'a yaklaşması ise sınıflardan en az birindeki performansın oldukça düşük olduğunu ifade etmektedir. Bu nedenle yüksek UPM değerleri daha başarılı sınıflandırma performansını göstermektedir.

Geometrik Ortalama

Tüm sınıflandırıcıların temel amacı, özgüllükten ödün vermeden duyarlılığı artırmaktır. Bununla birlikte, duyarlılık ve özgüllük hedefleri genellikle çelişkilidir ve bu durum, özellikle veri kümesi dengesiz olduğunda iyi sonuç vermeyebilir. Bu nedenle, Geometrik Ortalama (GM) ölçütü hem duyarlılık hem de özgüllük ölçütlerini bir araya getirir (Tharwat, 2021). Geometrik ortalama değeri aşağıda verilen eşitlik kullanılarak hesaplanmaktadır.

$$GM = \sqrt{\text{Duyarlılık} * \text{Özgüllük}} \tag{36}$$

Geometrik Ortalama'nın değeri 0 ile 1 arasında değişmektedir. GM değerinin 1 olması duyarlılık ve özgüllüğün maksimum düzeyde olduğunu ve mükemmel sınıflandırma performansını gösterirken, 0 değeri en az bir sınıfın tamamen hatalı sınıflandırıldığını ifade etmektedir. GM, özellikle dengesiz veri setlerinde duyarlılık ve özgüllük arasındaki dengeyi değerlendirmek için kullanılan önemli bir ölçüttür.

UYGULAMA

Bir araştırmacı, yeni geliştirilen bir biyobelirteç testinin belirli bir hastalığın tanısındaki performansını değerlendirmek istemektedir. Hastalığın varlığı veya yokluğu, mevcut klinik uygulamalarda kabul gören referans yöntem (altın standart) kullanılarak belirlenmiştir. Çalışmaya başvuran toplam 200 birey değerlendirilmiş ve her birey

hem referans test hem de yeni tanı testi ile incelenmiştir. Referans yönteme göre 80 bireyde hastalık mevcut, 120 bireyde ise hastalık bulunmamaktadır. Yeni geliştirilen test sonuçları referans yöntem ile karşılaştırıldığında; 68 bireyin gerçek pozitif, 12 bireyin yalancı negatif, 96 bireyin gerçek negatif ve 24 bireyin yalancı pozitif olarak sınıflandığı belirlenmiştir. Elde edilen sonuçlara ilişkin 2x2 tablosu Tablo 2'de verilmiştir.

Tablo 2. Tanı Testi Sonuçları

		Referans Test Sonucu		
		Hastalık Var	Hastalık Yok	Toplam
Medikal Test Sonucu	Pozitif	68 (GP)	24 (YP)	92
	Negatif	12 (YN)	96 (GN)	108
	Toplam	80	120	200

Bu veri seti kullanılarak yukarıda anlatılan ölçütlerin değerleri hesaplanmış ve sonuçlar Tablo 3'te toplu olarak verilmiştir.

Tablo 3. Tanı Testi Performans Ölçütleri ve Değerleri

Ölçüt	Değer	Ölçüt	Değer	Ölçüt	Değer
Duyarlılık	0.85	Youden/BMI	0.65	Markedness	0.63
Özgüllük	0.80	DOR	22.67	Yaygınlık Eşiği	0.33
PÖD/Prec.	0.74	CUI+	0.63	Fowlkes-Mallows	0.79
NÖD	0.89	CUI-	0.71	Yaygınlık	0.40
LR+	4.25	MCC	0.64	Jaccard	0.65
LR-	0.19	Dengeli Doğruluk	0.83	Cohen Kappa	0,000131
Doğruluk	0.82	F_1 Skoru	0.79	Brier Skoru	0.18
$D(f \mid\mid g)$	0.98	$D(g \mid\mid f)$	1.05	Yaklaşık Korelasyon	0.64
Öklid İndeksi	0.25	Çarpım İndeksi	0.68	Taguchi İndeksi	0.43
Ayırt Edici Güç	5.33	Yanlış Bulgu Oranı	0.67	Birleşik Performans	0.82
Geometrik Ortalama	0.82				

Tablo 3'te verilen 31 farklı performans ölçütü birlikte değerlendirilmiştir. Buna göre; yeni testin duyarlılığı 0.85 olarak bulunmuştur. Bu değer, hastalığı olan bireylerin %85'inin doğru şekilde saptandığını ve testin özellikle hastalığı kaçırmama (YN düşüklüğü) açısından güçlü olduğunu göstermektedir. Özgüllük değeri

0.80 olup, hastalığı olmayan bireylerin büyük çoğunluğunun doğru şekilde dışlandığını göstermektedir.

Genel doğruluk 0.82 olarak hesaplanmış ve testin toplam sınıflama başarısının iyi düzeyde olduğunu ortaya koymuştur. Dengeli doğruluk (0.83), sınıf dağılımından bağımsız olarak testin performansının dengeli olduğunu göstermektedir. Youden İndeksi (0.65), testin duyarlılık ve özgüllüğü birlikte dikkate alındığında rastgele sınıflamaya göre anlamlı bir ayırt edicilik sağladığını göstermektedir. Bu değer, testin dengeli bir tanısal performansa sahip olduğunu desteklemektedir.

Pozitif öngörü değeri (0.74), test pozitif olduğunda hastalığın bulunma olasılığının orta-iyi düzeyde olduğunu göstermektedir. Negatif öngörü değeri (0.89) ise test negatif çıktığında bireyin büyük olasılıkla sağlıklı olduğunu ve testin hastalığı eleme amacıyla güçlü olduğunu göstermektedir.

F1 skoru (0.79), pozitif sınıf performansında hassasiyet ve kesinlik arasında dengeli bir yapı olduğunu göstermektedir. Markedness (0.63), pozitif ve negatif tahminlerin güvenilirliğinin orta-iyi düzeyde olduğunu desteklemektedir.

Pozitif olabilirlik oranı (LR+ = 4.25), testin hastalığı olan bireyleri olmayanlardan ayırmada anlamlı katkı sağladığını, ancak tek başına güçlü doğrulayıcı test düzeyine ulaşmadığını göstermektedir. Negatif olabilirlik oranı (LR− = 0.19), testin hastalığı dışlama kapasitesinin iyi olduğunu göstermektedir.

Tanısal odds oranı (DOR = 22.67), testin genel ayırt edici gücünün yüksek olduğunu ve hastalık varlığı ile test sonucu arasında güçlü bir ilişki bulunduğunu göstermektedir.

Hastalık yaygınlığı 0.40 olarak belirlenmiş, bu da orta düzey prevalansa işaret etmektedir. Yaygınlık eşiği (0.33), modelin karar sınırının görece dengeli olduğunu göstermektedir. Bu durum, test performansının aşırı sınıf dengesizliğinden etkilenmediğini desteklemektedir.

Brier skoru 0.18 olup, olasılık tahminlerinin genel olarak iyi kalibre edildiğini ve hata düzeyinin düşük olduğunu göstermektedir. Yanlış

bulgu oranı 0.67 olarak hesaplanmış olup, toplam hataların (FP+FN) klinik karar süreçlerinde dikkate alınması gerektiğini göstermektedir.

Matthews korelasyon katsayısı (0.64), testin gerçek ve tahmin edilen sınıflar arasında orta-iyi düzeyde korelasyon sağladığını göstermektedir.

Cohen Kappa değeri (≈0.00013) oldukça düşüktür ve şansa bağlı uyuma çok yakın bir sonuç vermektedir. Bu durum, prevalans etkisi ve sınıf dağılımının Kappa üzerinde yarattığı baskı ile açıklanabilir ve tek başına testin zayıf olduğu anlamına gelmemelidir.

Jaccard indeksi (0.65), doğru pozitiflerin toplam pozitif tahminlere oranının kabul edilebilir düzeyde olduğunu göstermektedir. Fowlkes–Mallows indeksi (0.79), hassasiyet ve kesinlik arasında güçlü bir denge olduğunu desteklemektedir.

CUI+ (0.63) ve CUI− (0.71) değerleri, sırasıyla pozitif ve negatif sınıflamada orta-iyi düzeyde güvenilirlik olduğunu göstermektedir.

$D(f \| g)$ = 0.98 ve $D(g \| f)$ = 1.05 değerleri, hasta ve sağlam gruplar arasında anlamlı bilgi ayrışması bulunduğunu göstermektedir. Bu durum testin sınıfları ayırabildiğini ancak tamamen ayrık bir yapı oluşturmadığını ifade eder.

Yaklaşık korelasyon (0.64), test sonucu ile gerçek hastalık durumu arasında orta düzeyde ilişki olduğunu göstermektedir.

Öklid indeksi (0.25), sınıflar arası ayrışmanın makul düzeyde olduğunu göstermektedir. Çarpım indeksi (0.68), testin genel performansının dengeli olduğunu desteklemektedir. Taguchi indeksi (0.43), sistemde orta düzey varyasyon ve hata bileşeni bulunduğunu göstermektedir. Geometrik ortalama (0.82), duyarlılık ve özgüllüğün dengeli ve birlikte yüksek düzeyde olduğunu göstermekte olup, testin her iki sınıfta da (hasta ve sağlam) tutarlı performans sergilediğini ortaya koymaktadır.

Ayırt edici güç (5.33), testin hasta ve sağlam bireyleri anlamlı düzeyde ayırabildiğini göstermektedir. Birleşik performans indeksi (0.82), tüm metrikler birlikte değerlendirildiğinde testin genel performansının iyi düzeyde olduğunu ortaya koymaktadır.

Tüm ölçütler birlikte değerlendirildiğinde elde edilen sonuçlar yeni geliştirilen tanı testinin özellikle tarama ve ön değerlendirme amaçlı kullanım için uygun olduğunu, ancak tek başına kesin tanı koydurucu bir yöntem olarak değerlendirilmesinin sınırlı olabileceğini göstermektedir.

TARTIŞMA

Teknolojinin ilerlemesiyle birlikte tanı testleri sürekli gelişmekte ve klinik uygulamalarda mevcut standartlara yeni alternatifler eklenmektedir. Tanı seçeneklerinin çeşitlenmesi, bu testlerin klinik tanı araçları içerisindeki yerinin doğru şekilde değerlendirilmesini gerekli kılmaktadır. Bu değerlendirme süreci, öncelikle duyarlılık, özgüllük, pozitif ve negatif öngörü değerleri, olabilirlik oranları ve alıcı işletim karakteristik (ROC) eğrisi gibi ölçütler aracılığıyla test doğruluğunun ortaya konulmasını içermektedir (Fardy ve Barrett, 2015). Tanı testlerinin performansını değerlendirmek amacıyla farklı yöntemler geliştirilmiştir. Bununla birlikte, bu yöntemlerin her biri test performansının farklı bir yönünü yansıttığından, tek bir ölçütün tüm klinik gereksinimleri karşılaması mümkün değildir. Özellikle duyarlılık ve özgüllük gibi temel ölçütler önemli olmakla birlikte, bunların birlikte ve klinik bağlam içerisinde değerlendirilmesi gerekmektedir. Tıbbi araştırmaların birçok alanında gerçek anlamda bir altın standart testin bulunmaması, tanısal doğruluk çalışmalarının değerlendirilmesini zorlaştırmaktadır. Bu durum, elde edilen sonuçların yorumlanmasında dikkatli olunmasını ve test performansının yalnızca sayısal değerler üzerinden değerlendirilmemesini gerektirmektedir. Tanı testlerinin değerlendirilmesi yalnızca hastalığın doğru şekilde saptanması açısından değil; aynı zamanda tarama, prognozun belirlenmesi ve klinik karar süreçlerine katkı sağlaması açısından da büyük önem taşımaktadır. Bu nedenle tanı testleri, farklı klinik amaçlara hizmet eden çok yönlü araçlar olarak değerlendirilmelidir. Bununla birlikte, kullanılan istatistiksel ölçütler testin ayırt edici gücü ve tahmin performansı gibi farklı özellikleri ortaya koymaktadır. Ancak hiçbir tekil ölçüt, tanı testinin klinik performansını tam olarak yansıtamaz. Bu nedenle tanısal değerlendirme sürecinde birden fazla ölçütün birlikte ele alınması gerekmektedir (Gogtay, 2017). Tanı testleri sağlık hizmetlerinde önemli bir role sahiptir ve hasta yönetiminde kritik

kararların verilmesine katkı sağlar. Sağlık profesyonelleri, belirsizlik içeren durumlarda klinik karar verirken bu testlerden yararlanmakta ve elde edilen sonuçları hastanın genel klinik durumu ile birlikte değerlendirmektedir. Sonuç olarak, tanı testlerinin değerlendirilmesi yalnızca istatistiksel doğruluk ölçütlerine dayandırılmamalı; klinik bağlam, hasta özellikleri ve sağlık hizmeti sonuçları birlikte ele alınmalıdır. Tanısal doğruluk önemli bir gösterge olmakla birlikte, tek başına klinik yararlılığı açıklamak için yeterli değildir.

SONUÇ VE ÖNERİLER

Tanısal çalışmalar, sağlık hizmetlerinde kullanılan testlerin klinik karar verme süreçlerine ne ölçüde katkı sağladığını ortaya koyan temel araştırma alanlarından biridir. Bu çalışmalar, yeni geliştirilen veya mevcut tanı testlerinin yalnızca teknik doğruluğunu değil, aynı zamanda klinik pratikte hastalığı olan ve olmayan bireyleri ayırt etme gücünü de değerlendirmeyi amaçlamaktadır. Bu nedenle tanısal doğruluk, bir testin klinik değerinin anlaşılmasında kritik bir bileşen olmakla birlikte tek başına yeterli bir gösterge değildir (Hoyer ve Zapf, 2021). Hem araştırmacıların hem de klinik uygulayıcıların tanı testlerinin doğruluğuna ilişkin sınırlamaların farkında olmaları ve doğruluğun, bir testin klinik değeri veya kullanışlılığını açıklayan tek boyut olmadığını bilmeleri gerekmektedir. Tanısal doğruluk, testin bireylerin sağlık durumunu ne ölçüde doğru yansıttığının dolaylı bir göstergesidir. Bir doğruluk çalışması planlanırken veya değerlendirilirken, duyarlılık ve özgüllük başta olmak üzere test performansının farklı koşullar altında değişkenlik gösterebileceği göz önünde bulundurulmalıdır (Leeflang ve Allerberger, 2019). Tanı testlerinin değerlendirilmesinde duyarlılık, özgüllük, öngörü değerleri, olabilirlik oranları ve ROC eğrisi gibi çok sayıda ölçüt kullanılmakta; her biri test performansının farklı bir yönünü ortaya koymaktadır. Ancak bu ölçütlerin hiçbirinin tek başına tüm klinik gereksinimleri karşılamadığı bilinmektedir. Özellikle sınıf dengesizliği, hastalık prevalansı ve testin uygulandığı popülasyon gibi faktörler, elde edilen sonuçların yorumlanmasını önemli ölçüde etkileyebilmektedir. Bu durum, tanısal doğruluğun çok boyutlu bir çerçevede ele alınmasını zorunlu kılmaktadır. Klinik uygulamada tanı testlerinin değeri yalnızca istatistiksel performanslarıyla değil, aynı zamanda hasta yönetimine olan katkılarıyla belirlenmektedir. Günlük pratikte anamnez, fizik

muayene ve laboratuvar testleri birlikte değerlendirilerek klinik kararlar verilmektedir. Bu süreçte tanı testleri çoğunlukla mevcut klinik şüpheyi doğrulamak veya alternatif tanıları dışlamak amacıyla kullanılmakta, böylece belirsizliğin azaltılmasına katkı sağlamaktadır. Bununla birlikte testlerin maliyet, erişilebilirlik, risk ve yorumlanabilirlik gibi yönleri de klinik faydanın önemli bileşenlerini oluşturmaktadır. Öte yandan tanı testlerinin performansı her zaman sabit değildir; testin uygulandığı popülasyona, hastalığın yaygınlığına ve klinik bağlama göre değişkenlik gösterebilir. Bu nedenle bir testin farklı ortamlarda aynı başarıyı göstermesi beklenmemelidir. Ayrıca bazı performans ölçütlerinin prevalanstan etkilenmesi, test sonuçlarının genellenebilirliğini sınırlayabilmektedir. Bu durum, özellikle klinik karar süreçlerinde birden fazla ölçütün birlikte değerlendirilmesini gerekli kılmaktadır. Sonuç olarak tanısal doğruluk, bir testin klinik yararlılığını açıklayan tek boyutlu bir kavram değildir. Tanı testlerinin güvenilir biçimde değerlendirilebilmesi için istatistiksel performans ölçütlerinin yanı sıra klinik bağlam, hasta popülasyonu özellikleri ve sağlık sonuçlarına olan etkiler birlikte ele alınmalıdır. Bu bütüncül yaklaşım, tanı testlerinin yalnızca teknik olarak değil, klinik açıdan da anlamlı ve doğru şekilde kullanılmasını sağlayacaktır.

KAYNAKÇA

Abhishek, K. ve Hamarneh, G. (2010). Matthews Correlation Coefficient Loss for Deep Convolutional Networks: Application to Skin Lesion Segmentation. *arXiv*, 13454v2 [eess.IV].

Asberg, A., Mikkelsen, G. ve Odsaeter, I.H. (2019). A New Index of Clinical Utility for Diagnostic Tests. *Scandinavian Journal of Clinical and Laboratory Investigation*, 79 (8), 560-565.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F. ve Nielsen, H. (2000). Assessing The Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinformatics*, 16(5), 412-424.

Berrar, D. (2019). Performance Measures for Binary Classification. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, Cannataro M. (eds.) Encyclopedia of Bioinformatics and Computational Biology. Reference Module in Life Sciences, 1. Elsevier, pp. 546–560.

Bolboaca, S.D. (2019). Medical Diagnostic Tests: a Review of Test Anatomy, Phases, and Statistical Treatment of Data. *Computational and Mathematical Methods in Medicine*, 1891569.

Bonnici, V. (2024). A Maximum Value for The Kullback–Leibler Divergence Between Quantized Distributions. *Information*, 15, 547.

Casanova, I.J., Campos, M., Juarez, J.M., Gomariz, A., Lorente-Ros, M. ve Lorente, J.A. (2022). Using The Diagnostic Odds Ratio to Select Patterns to Build an Interpretable Pattern-Based Classifier in a Clinical Domain: Multivariate Sequential Pattern Mining Study. *JMIR Medical Informatics*, 10(8), e32319.

Chen, F., Xue, Y., Tan, M.T. ve Chen, P. (2015). Efficient Statistical Tests to Compare Youden Index: Accounting for Contingency Correlation. *Statistics in Medicine*, 34(9), 1560-1576.

Chen, M., Samawi, H., Yin, J. ve Rochani, H. (2025). On Diagnostic Accuracy Measures and Optimal Cut-Point Selection Measures for Multi-Stage Diseases via Generalized Total Kullback–Leibler Divergence. *Communications in Statistics - Theory and Methods*, 1-39.

Chicco, D. ve Jurman, G. (2023). A Statistical Comparison Between Matthews Correlation Coefficient (MCC), Prevalence Threshold, and Fowlkes-Mallows Index. *Journal of Biomedical Informatics*, 144, 104426.

Chicco, D. ve Jurman, G. (2022). An Invitation to Greater Use of Matthews Correlation Coefficient in Robotics and Artificial Intelligence. *Frontiers in Robotics and AI*, 9, 876814.

Chicco, D. ve Jurman, G. (2020). The Advantages of The Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, 21(1), 6.

Chicco, D., Starovoitov, V. ve Jurman, G. (2021). The Benefits of The Matthews Correlation Coefficient (MCC) Over The Diagnostic Odds Ratio (DOR) in Binary Classification Assessment. *IEEE Access*, 9, 47112-47124.

Chicco, D., Warrens, M.J. ve Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE Access,* 9, 78368-78381.

De Diego, I.M., Redondo, A.R., Fernández, R.R., Navarro, J. ve Moguerza, JM. (2022). General Performance Score for Classification Problems. *Applied Intelligence*, 52, 12049-12063.

Deeks, J.J. (2001). Systematic Reviews of Evaluations of Diagnostic and Screening Tests. *BMJ*, 323, 157-162.

Dhamnetiya, D., Jha, R.P., Shalini, S. ve Bhattacharyya, K. (2021). How to Analyze The Diagnostic Performance of a New Test? Explained with Illustrations. *Journal of Laboratory Physicians*, 14(1), 90-98.

Fardy, J.M. ve Barrett, B.J. (2015). Evaluation of Diagnostic Tests. *Methods of Molecular Biology*, 1281, 289-300.

Foody, G.M. (2023). Challenges in The Real World Use of Classification Accuracy Metrics: From Recall and Precision to The Matthews Correlation Coefficient. *PLoS One*, 18(10), e0291908.

Glas, A.S., Lijmer, J.G., Prins, M.H., Bonsel, G.J. ve Bossuyt, P.M. (2003). The Diagnostic Odds Ratio: A Single Indicator of Test Performance. *Journal of Clinical Epidemiology*, 56(11), 1129-1135.

Gogtay, N.J. (2017). Statistical Evaluation of Diagnostic Tests - Part 2 [Pre-Test and Post-Test Probability and Odds, Likelihood Ratios, Receiver Operating Characteristic Curve, Youden's Index and Diagnostic Test Biases]. *Journal of the Association of Physicians of India*, 65(7), 86-91.

Grandini, M., Bagli, E. ve Visani, G. (2008). Metrics for Multi-Class Classification: An Overview. *arXiv*, 2008.05756v1 [stat.ML].

Grimes, D.A. ve Schulz, K.F. (2005). Refining Clinical Diagnosis with Likelihood Ratios. *Lancet*, 365(9469), 1500-1505.

Habibzadeh, F. (2025). Diagnostic Tests Performance Indices: An Overview. *Biochemia Medica (Zagreb)*, 35(1), 010101.

Hayden, S.R. ve Brown, M.D. (1999). Likelihood Ratio: A Powerful Tool for Incorporating The Results of a Diagnostic Test Into Clinical Decisionmaking. *Annals of Emergency Medicine*, 33(5), 575-580.

Hoessly, L. (2026). On Misconceptions About The Brier Score in Binary Prediction Models. *Global Epidemiology*, 11, 100242.

Hoyer, A. ve Zapf, A. (2021). Studies for The Evaluation of Diagnostic Tests-part 28 of a Series on Evaluation of Scientific Publications. *Deutsches Ärzteblatt International*, 118: 555–60.

Hu, Z. ve Hong, L.J. Kullback-Leibler Divergence Constrained Distributionally Robust Optimization. https://optimization-online.org/?p=12225 (accessed on 14 May 2025).

Huang, Y., Yin, J. ve Samawi, H. (2018). Methods Improving The Estimate of Diagnostic Odds Ratio, *Communications in Statistics-Simulation and Computation*, 47(2), 353-366.

Itaya, Y., Tamura, J., Hayashi, K. ve Yamamoto, K. (2025). Asymptotic Properties of Matthews Correlation Coefficient. *Statistics in Medicine*, 44, e10303.

Lee, W.C. (1999). Selecting Diagnostic Tests for Ruling Out or Ruling In Disease: The Use of The Kullback-Leibler Distance. *International Journal of Epidemiology*, 28(3), 521-525.

Leeflang, M.M.G. ve Allerberger, F. (2019). How to: Evaluate a Diagnostic Test. *Clinical Microbiology and Infection*, 25(1), 54-59.

Leonardis, D., D'Arrigo, G., Abd ElHafeez, S., Fusaro, M., Roumeliotis, S. ve Tripepi, G. (2023). Simple Statistical Measures for Assessing The Accuracy of a Diagnostic Test in Clinical Medicine. *Reports on Global Health Research,* 6(1), 148.

Lim, C.Y. (2021). Methods for Evaluating The Accuracy of Diagnostic Tests. *Cardiovascular Prevention and Pharmacotherapy*, 3(1), 15-20.

Matthews, B.W. (1975). Comparison of The Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2), 442-451.

Mitchell, A.J. (2008). The Clinical Significance of Subjective Memory Complaints in The Diagnosis of Mild Cognitive Impairment and Dementia: A Meta-Analysis. *International Journal of Geriatric Psychiatry*, 23(11), 1191-1202.

Nellore, S.B. (2015). Various Performance Measures in Binary Classification-An Overview of ROC Study. *International Journal of Innovative Science Engineering and Technology*, 2(9), 596-605.

Perez, I., Taito-Vicenti, I.Y., Gonzalez-Xuriguera, C.G., Carvajal, C., Franco, J.V.A. ve Loezar, C. (2021). How to Interpret Diagnostic Tests. *Medwave*, 21(7), e8432.

Raslich, M.A., Markert, R.J. ve Stutes SA. (2007). Selecting and Interpreting Diagnostic Tests. *Biochemia Medica (Zagreb)*,17, 151-161.

Samawi, H., Chen, D.G., Yin, J. ve Alsharman, M. (2024). Performance of Diagnostic Tests Based on Continuous Bivariate Markers. *Journal of Applied Statistics*, 51(3), 497-514.

Samawi, H.M., Yin, J., Zhang, X., Yu, L., Rochani, H., Vogel, R. ve Mo, C. (2020). Kullback-Leibler Divergence for Medical Diagnostics Accuracy and Cut-Point Selection Criterion: How It is Related to The Youden Index. *Journal of Applied Bioinformatics & Computational Biology*. 9 (2), 168.

Shaikh, S.A. (2011). Measures Derived From a 2x2 Table for An Accuracy of a Diagnostic Test. *Journal of Biometrics and Biostatistics*, 2(5), 1000128.

Simundic, A.M. (2009). Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC*, 19(4), 203-211.

Sousa, M.R. ve Ribeiro, A.L. (2009). Systematic Review and Meta-Analysis of Diagnostic and Prognostic Studies: A Tutorial. *Arquivos Brasileiros de Cardiologia*, 92(3), 229-238.

Stoica, P. ve Babu, P. (2023). Pearson-Matthews Correlation Coefficients for Binary and Multinary Classification and Hypothesis Testing. *arXiv*, 2305.05974v1 [eess.SP].

Tamura, J., Itaya, Y., Hayashi, K. ve Yamamoto, K. (2025). Statistical Inference of The Matthews Correlation Coefficient for Multiclass Classification. *arXiv*, 2503.06450v1 [stat.ME].

Tharwat, A. (2021). Classification Assessment Methods. *Applied Computing and Informatics*, 17(1), 168-192.

Tsokani, S., Veroniki, A.A., Pandis, N. ve Mavridis D. (2022). Assessing The Performance of Diagnostic Test Accuracy Measures. *American Journal of Orthodontics and Dentofacial Orthopedics,* 161(5), 748-751.

Umemneku Chikere, C.M., Wilson, K,, Graziadio, S., Vale, L. ve Allen, A.J. (2019). Diagnostic Test Evaluation Methodology: A Systematic Review of Methods Employed to Evaluate Diagnostic Tests in The Absence of Gold Standard. *PLoS ONE*, 14(10), e0223832.

Wu, F.C., Chuang, T.C., Chou, F.D. ve Lee, Y.C. (2022). Evaluating The Reliability of Diagnostic Performance Indices by Using Taguchi Quality Loss Function. *Annals of Operations Research*, 311(1), 437-449.

Yang, W., Jiang, J., Schnellinger, E.M., Kimmel, S.E. ve Guo, W. (2022). Modified Brier Score for Evaluating Prediction Accuracy for Binary Outcomes. *Statistical Methods in Medical Research*, 31(12), 2287-2296.

Zhang, Y., Liu, W., Chen, Z., Wang, J. ve Li, K. (2023). On The Properties of Kullback-Leibler Divergence Between Multivariate Gaussian Distributions. *arXiv*, 2102.05485v5 [cs.IT].

Zhong, J., Liu, R. ve Chen, P. (2020). Identifying Critical State of Complex Diseases by Single-Sample Kullback-Leibler Divergence. *BMC Genomics*, 21(1), 87.

Zhou, X., Obuchowski, N.A. ve McClish, D.K. (2011). Statistical Methods in Diagnostic Medicine. 2nd Edition. Hoboken, New Jersey: John Wiley & Sons, Inc.; p. 238-239.

RNA-DİZİLEME VERİLERİNDE DİFERANSİYEL GEN İFADESİ ANALİZİ VE DÜZENSİZ GEN İFADESİ PROFİLLERİNİN TARANMASI

Merve Başol Göksülük[2], Hamdi Furkan Kepenek[3], Dinçer Göksülük[4,*]

GİRİŞ

Gen ifadesi verileri: Mikrodizilimden RNA dizilemeye

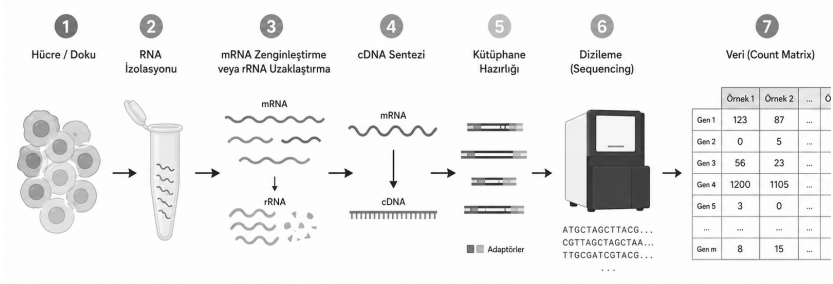
Gen ifadesi (gene expression), DNA'da saklanan bilginin RNA moleküllerine aktarılması ve bazı genler için protein üretimine kadar ilerlemesiyle ilişkili temel biyolojik süreçtir (Kukurba & Montgomery, 2015). Bir hücre veya dokuda belirli bir anda bulunan RNA moleküllerinin bütünü, biyolojik durumun dinamik bir görünümünü sunar. Bu görünüm; hastalık mekanizmalarının, hücresel farklılaşmanın ve tedaviye verilen moleküler yanıtın incelenmesinde kullanılabilir (Kukurba ve Montgomery, 2015; Wang vd., 2009b).

Yüksek çıktılı gen ifadesi çalışmalarının ilk yaygın teknolojik zemini mikrodizilim (microarray) platformlarıdır. Bu platformlarda gen ifadesi düzeyi, prob hibridizasyonu (probe hybridization) sonucunda elde edilen sürekli yoğunluk ölçümleriyle temsil edilir. Mikrodizilim çalışmaları, gen gruplarının biyolojik koşullar arasında karşılaştırılması ve hastalık alt tiplerinin araştırılması gibi birçok genomik problemin gelişmesine katkı sağlamıştır (Dudoit vd., 2002; Golub vd.,

[*] Sorumlu yazar: dincergoksuluk@sakarya.edu.tr
[2] Dr. Öğr. Üyesi, Sakarya Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Sakarya, Türkiye, E-posta: mbasolgoksuluk@sakarya.edu.tr, ORCID: orcid.org/0000-0002-2223-7856
[3] Hacettepe Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Ankara, Türkiye, E-posta: h.furkankepenek@gmail.com, ORCID: orcid.org/0009-0003-6339-4203
[4] Doç. Dr., Sakarya Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Sakarya, Türkiye. ORCID: orcid.org/0000-0002-2752-7668

1999). Bununla birlikte ölçüm önceden tasarlanmış problarla sınırlıdır; çapraz hibridizasyon, arka plan sinyali ve dar dinamik aralık özellikle düşük düzeyli ya da anotasyon (annotation) bilgisi sınırlı transkriptlerin (transcript) değerlendirilmesini güçleştirebilir.

RNA dizileme (RNA sequencing, RNA-Seq), yeni nesil dizileme teknolojilerini RNA moleküllerinin ölçümüne uygular. Şekil 1'de özetlenen süreç, toplam RNA (total RNA) moleküllerinin biyolojik örnekten izole edilmesiyle başlar. Çalışmanın amacına göre mesajcı RNA (messenger RNA, mRNA) zenginleştirilir veya ribozomal RNA (ribosomal RNA, rRNA) uzaklaştırılır; RNA molekülleri tamamlayıcı DNA (complementary DNA, cDNA) biçimine dönüştürülerek dizileme kütüphanesi hazırlanır. Yüksek verimli dizilemeyle elde edilen kısa okumalar referans genoma ya da transkriptoma hizalanır veya doğrudan transkriptlere atanır. Her gen ya da transkriptle eşleşen okumaların özetlenmesi sonucunda satırları genomik özellikleri, sütunları örnekleri gösteren sayım (count) matrisi elde edilir (Conesa vd., 2016; Kukurba ve Montgomery, 2015).



Şekil 1. RNA-Seq Deneyinde Biyolojik Örnekten Sayım Matrisine Uzanan Sürecin Şematik Gösterimi

Şekildeki akış tipik mRNA veya toplam RNA temelli bir deneyin genel çerçevesini gösterir. Küçük RNA (small RNA, sRNA) ve mikroRNA (microRNA, miRNA) dizilemesinde boyut

seçimi, adaptör bağlama ve kütüphane hazırlama adımları protokole göre değişebilir. Benzer biçimde eşleştirilmiş uçlu (paired-end) ya da tek uçlu (single-end) dizileme, okuma uzunluğu ve hizalama yaklaşımı elde edilen sayımların özelliklerini etkiler. Bu teknik kararlar biyolojik soruya göre verilmeli ve karşılaştırılan örneklerde mümkün olduğunca tutarlı uygulanmalıdır.

Mikrodizilim çoğunlukla sürekli yoğunluk değerleri üretirken RNA-Seq ayrık okuma sayımları üretir. RNA-Seq'in geniş dinamik aralığı düşük ve yüksek gen ifadesi düzeylerinin aynı deney içinde değerlendirilmesini kolaylaştırabilir; ancak sıfır değerlerin çokluğu ve ortalamaya bağlı değişkenlik standart sürekli veri yöntemlerinin doğrudan kullanılmasını sınırlar (Conesa vd., 2016; Love vd., 2014). Teknolojik üstünlük bu nedenle yalnız ölçülebilen özellik sayısıyla değil, üretilen veri ölçeğine uygun kalite kontrolü ve istatistiksel modellemeyle birlikte değerlendirilmelidir.

RNA-Seq, prob bölgeleriyle sınırlı olmadığından yeni transkriptlerin, alternatif izoformların, farklı kesip birleştirme (splice) olaylarının ve kodlamayan (non-coding) RNA türlerinin incelenmesine olanak verir. Buna karşılık ürettiği ayrık sayımlar doğrudan biyolojik gen ifadesi düzeyi olarak yorumlanamaz. Gözlenen değerler biyolojik sinyalin yanında kütüphane büyüklüğü (library size), dizileme derinliği (sequencing depth), örnek hazırlama ve hizalama gibi teknik etkenlerden de etkilenir. Biyolojik tekrarlar arasındaki değişkenlik çoğu zaman Poisson modelinin öngördüğünden büyüktür; aşırı saçılım (overdispersion) olarak adlandırılan bu özellik RNA-Seq için uygun bir sayım modeli gerektirir (Love vd., 2014; Marioni vd., 2008).

Gen ifadesi verileri bu çerçevede yalnız tek bir analitik amaca hizmet etmez; örneklerin bilinmeyen alt gruplarını keşfetmek,

bilinen sınıf etiketlerinden yeni örneklerin sınıfını tahmin etmek ve önceden tanımlanmış biyolojik koşullar arasında gen ifadesi düzeylerini karşılaştırmak için kullanılabilir (Dudoit vd., 2002; Golub vd., 1999). Bu kitap bölümünün ana odağı, sınıf karşılaştırmasının RNA-Seq çalışmalarındaki başlıca karşılığı olan diferansiyel gen ifadesi analizidir. Bu analiz, iki veya daha fazla biyolojik koşul arasında sistematik gen ifadesi değişimi bulunan genleri belirlemeyi amaçlar. Hastalıkla ilişkili süreçleri araştırmak, tedavi yanıtı veya direnç mekanizmalarını incelemek ve deneysel doğrulama için adayları önceliklendirmek amacıyla kullanılabilir.

RNA-Seq diferansiyel gen ifadesi analizi için çok sayıda yöntem geliştirilmiştir. DESeq2 ve edgeR, sayımları negatif binom modellerle ele alan yaklaşımlardır (Love vd., 2014; Robinson vd., 2010). Limma-voom, logaritmik ölçekte ortalama ile varyans arasındaki ilişkiyi her gözlem için kesinlik ağırlığı (precision weight) kullanarak doğrusal modele taşır; baySeq, PoissonSeq ve SAMseq ise farklı Bayesci, dağılımsal veya parametrik olmayan seçeneklere örnektir (Hardcastle ve Kelly, 2010; Law vd., 2014; Li ve Tibshirani, 2013; Li vd., 2012). Bu liste alanın tamamını temsil etmez ve yöntemlerin göreli performansı örneklem büyüklüğü, saçılım, aykırı değerler ve kütüphane dengesizliği gibi veri özelliklerine göre değişebilir (Rosati vd., 2024). Bu kitap bölümünde yöntemsel anlatı, normalizasyon, saçılım tahmini, deney tasarımı ve çoklu test kontrolünü bütünlüklü bir akış içinde ele alan DESeq2 çerçevesinde sürdürülmektedir. Biyolojik yorum ise ham sayımların doğrudan karşılaştırılmasına değil; deney tasarımı, teknik ölçek farkları ve örnekler arası değişkenliği birlikte ele alan istatistiksel çıkarıma dayanmalıdır. Bu nedenle sonraki bölümde önce sayım matrisi ve örnek bilgisi, ardından filtreleme, normalizasyon ve dönüşüm adımları ele alınacaktır.

YÖNTEM

RNA-Seq Veri Yapısı ve Analiz Öncesi Hazırlık

RNA-Seq iş akışının biyoinformatik çıktısı çoğunlukla bir sayım matrisidir (Şekil 1). Bu matrisin istatistiksel analize hazırlanması tek bir işlemden oluşmaz. Ölçülen özelliklerin ve örneklerin tanımlanması, analiz hedefine göre sınırlı bilgi taşıyan genlerin filtrelenmesi, örnekler arasındaki teknik ölçek farklarının normalizasyonla modellenmesi ve gerektiğinde varyansı dengeleyen bir dönüşüm uygulanması farklı amaçlara hizmet eder. Özellikle normalizasyon ile dönüşüm birbirinin yerine kullanılmamalıdır: ilki örneklerin sayım ölçeğini, ikincisi ise verinin sonraki kullanım için sayısal yapısını düzenler.

Bu bölüm, ham okuma dosyalarının kalite değerlendirmesi, adaptör temizliği ve hizalama gibi daha erken biyoinformatik adımların başarıyla tamamlandığını varsayar. Bununla birlikte bu adımlardaki düşük hizalanma oranı, örnek karışması veya belirgin kalite kaybı gibi sorunlar sayım matrisine taşınabilir. Dolayısıyla gen düzeyindeki hazırlıktan önce örnek düzeyindeki kalite ölçütleri de incelenmeli; sistematik olarak sorunlu bir örnek yalnız normalizasyonla düzeltilebilecekmiş gibi değerlendirilmemelidir (Conesa vd., 2016).

Sayım matrisi ve örnek bilgisi

Ham sayım matrisi

$$\mathbf{K} = \left(K_{ij}\right)_{p\times n}$$

ile gösterilsin. Burada K_{ij}, j. örnekte i. genomik özellikle eşleşen gözlenmiş okuma sayısıdır; p gen, transkript veya başka bir genomik özellik sayısını, n ise örnek sayısını gösterir. Sayımın yüksek olması genellikle daha fazla RNA molekülüyle ilişkilidir, ancak gen uzunluğu, örnek kompozisyonu, kütüphane hazırlama ve dizileme ölçeği de gözlenen değeri etkileyebilir. Bu nedenle K_{ij}, doğrudan ve hatasız bir biyolojik gen ifadesi düzeyi olarak

kabul edilemez (Conesa vd., 2016; Kukurba ve Montgomery, 2015).

Satır tanımlayıcılarının kullanılan genom anotasyonuyla uyumlu, sütun adlarının da örnek bilgi tablosuyla bire bir eşleşmesi gerekir. Aynı genin yinelenen satırlarda bulunması, gen ve transkript düzeylerinin karıştırılması veya örnek sırasının tasarım tablosuyla uyuşmaması model doğru kurulsa bile yanlış karşılaştırmalara yol açabilir. Bu nedenle boyut, tanımlayıcı, eksik değer ve örnek sırası kontrolleri istatistiksel modelden önce tamamlanmalıdır.

Matris, her sütunun hangi biyolojik koşula ve deneysel birime ait olduğunu gösteren örnek bilgisiyle birlikte değerlendirilmelidir. Bu bilgi grup, parti (batch), yaş, cinsiyet, eşleşmiş birey, zaman ve diğer teknik ya da biyolojik değişkenleri içerebilir. Değişkenlerin hangi sırayla modele girdiğinden önce, koşulla tam olarak karışıp karışmadıkları incelenmelidir. Örneğin bütün hasta örneklerinin bir partide, bütün kontrollerin başka bir partide işlenmesi durumunda hastalık ve parti etkileri veriden ayrı ayrı tahmin edilemez.

Biyolojik tekrar, aynı koşulu temsil eden bağımsız biyolojik birimlerden elde edilen örnekleri ifade eder; farklı hastalar, hayvanlar veya bağımsız hücre kültürleri buna örnektir. Teknik tekrar ise aynı biyolojik materyalin hazırlama veya ölçüm sürecinin yinelenmesidir. Teknik tekrarlar ölçüm değişkenliğini değerlendirebilir, ancak koşul içindeki doğal biyolojik değişkenliği temsil etmedikleri için biyolojik tekrarların yerini tutmaz. Saçılım ve standart hata tahminlerinin güvenilirliği esas olarak bağımsız biyolojik tekrarlara dayanır (Love vd., 2014).

Sınırlı bilgi taşıyan genlerin filtrelenmesi

Filtrelemenin amacı yalnızca sayımı düşük olan bütün genleri çıkarmak değildir; hedeflenen analiz için güvenilir tahmin veya

yorum üretmeye sınırlı katkı sağlayan özellikleri belirlemektir. Çok az örnekte gözlenen ya da örneklerin büyük bölümünde sıfır sayım alan bir gen, saçılım ve kat değişimi (fold change) tahmini için yetersiz bilgi taşıyabilir. Benzer biçimde belirsiz hizalanan okumaların baskın olduğu veya kullanılan anotasyonla güvenilir biçimde eşleştirilemeyen bir özellik teknik açıdan sorunlu olabilir. Bu genlerin tamamının test kümesinde tutulması çoklu test yükünü artırabilir ve düşük bilgiye dayalı kararsız tahminlerin öne çıkmasına yol açabilir (Love vd., 2014, 2026).

Uygun ölçüt analiz amacına bağlıdır. Diferansiyel gen ifadesi analizinde belirli sayıda örnekte asgari sayım veya gözlenme sıklığı kullanılabilir. Dönüştürülmüş veriyle yürütülen örnek ilişkisi, sınıflandırma ya da tamamlayıcı profil taramalarında ise çok düşük değişkenlik veya teknik kalite ölçütleri ayrıca değerlendirilebilir. Dolayısıyla "düşük sinyal" ya da "düşük kaliteli gen" gibi genel bir etiket yerine, hangi bilgi ölçütünün hangi analiz için neden yetersiz kabul edildiği açıkça tanımlanmalıdır. Filtreleme ölçütü grup etiketlerinden yararlanıyorsa kuralın sonuçları yapay biçimde güçlendirmediği de değerlendirilmelidir (Love vd., 2026).

Filtreleme kuralı sonuçlara bakılarak geriye dönük seçilmemeli; örneklem büyüklüğü, grup yapısı ve biyolojik hedef dikkate alınarak analiz öncesinde belirlenmelidir. Aşırı katı filtreleme zayıf fakat gerçek biyolojik değişimleri silebilir, aşırı gevşek filtreleme ise test gücünü azaltabilir. Buradaki ön filtreleme, DESeq2'nin hipotez testinden sonra uygulayabildiği bağımsız filtrelemeden farklıdır; bağımsız filtreleme çoklu test bağlamında "Hipotez testleri ve hata kontrolü" bölümünde ele alınacaktır.

Normalizasyon ve ölçekleme faktörleri

Aynı biyolojik gen ifadesi düzeyine sahip iki örnek, farklı kütüphane büyüklükleri veya kompozisyonları nedeniyle farklı

ham sayımlar üretebilir. Normalizasyonun amacı bu teknik ölçek farkını hesaba katarken biyolojik koşul farkını korumaktır. Yalnız toplam okuma sayısına göre ölçekleme, az sayıdaki çok yüksek gen ifadeli özelliğin kütüphaneyi baskıladığı durumlarda diğer genlerde yapay farklar oluşturabilir. Bu kompozisyon etkisi, sağlam bir göreli ölçek tahmininin neden gerekli olduğunu gösterir.

Literatürde toplam sayım (total count), üst çeyrek (upper quartile) ve kırpılmış M-değer ortalaması (trimmed mean of M-values, TMM) gibi yaklaşımlar bulunmaktadır (Bullard vd., 2010; Robinson ve Oshlack, 2010). Bu yaklaşımlar teknik ölçek farklarını farklı varsayımlar altında ele alır. DESeq2 ise varsayılan olarak medyan-oranı (median-ratio) yaklaşımını kullanır (Anders ve Huber, 2010; Love vd., 2014).

Önce her gen için örnekler arası geometrik ortalama kullanılarak sözde-referans (pseudo-reference) değer oluşturulur:

$$G_i = \left(\prod_{j=1}^{n} K_{ij} \right)^{1/n} .$$

Geometrik ortalaması sıfır olmayan genlerde her örneğin sayımı G_i'ye bölünür ve bu oranların medyanı, j. örneğin ölçekleme faktörü (size factor) olarak tahmin edilir:

$$s_j = \text{median}_{i:G_i>0} \left(\frac{K_{ij}}{G_i} \right)$$

Standart geometrik ortalama tanımında herhangi bir örnekte sıfır sayım bulunan genler oran hesabının dışında kalabilir; bu durum çok sıfırlı (zero-inflated) veri setlerinde kullanılan gen kümesini daraltabilir. Medyan kullanımı, az sayıdaki çok yüksek sayımlı genin örnek ölçeğini belirlemesini sınırlar. Yaklaşım çoğu genin diferansiyel olmadığı veya değişen genlerin büyük

çoğunluğunun tek yönde hareket etmediği varsayımına dayanır. Gen ifadesinin küresel olarak tek yönde kaydığı deneylerde haricî kontrol özellikleri ya da farklı bir normalizasyon stratejisi gerekebilir.

K_{ij}/s_j oranı betimsel olarak normalize edilmiş sayım şeklinde incelenebilir. Bununla birlikte DESeq2 çıkarımı bu oranları yeni gözlemler olarak modele vermez. Ham K_{ij} sayımları yanıt değişkeni olarak korunur; s_j ise beklenen sayımın teknik ölçek bileşeni olarak negatif binom modele eklenir. Böylece sayım doğası korunurken örnekler arası karşılaştırılabilirlik model içinde sağlanır. Ölçekleme faktörleri gen uzunluğunu genler arası karşılaştırma amacıyla düzeltmez; DESeq2'nin hedefi aynı genin örnekler arasındaki koşula bağlı değişimini modellemektir.

Sayım verisinin dönüştürülmesi ve varyansın dengelenmesi

RNA-Seq sayımlarında varyans genellikle ortalamayla birlikte büyür. Bu değişken varyanslı ortalama-varyans (mean-variance) ilişkisi, ham sayımlara sabit varyans ve normal dağılım varsayımlı yöntemlerin doğrudan uygulanmasını güçleştirir. Basit logaritmik dönüşüm yüksek sayımları sıkıştırabilir, ancak düşük sayımlardaki güçlü değişkenliği ve sıfır değerleri tek başına yeterince dengelemeyebilir. Dönüşümün hedefi, gen ifadesi düzeyi boyunca varyansı yaklaşık olarak sabit tutarak veriyi sabit varyanslı (homoscedastic) bir yapıya yaklaştırmaktır (Love vd., 2014, 2026).

DESeq2, varyans dengeleyici dönüşüm (variance stabilizing transformation, VST) ile düzenlileştirilmiş logaritmik dönüşüm (regularized logarithmic transformation, rlog) seçeneklerini sunar. Her ikisi de normalize edilmiş sayımlardaki ortalama-varyans bağımlılığını azaltır. VST büyük veri setlerinde daha hızlıdır; rlog, ölçekleme faktörleri arasındaki büyük farklara VST'den daha az duyarlı olabilse de örnek sayısı arttıkça hesaplama maliyeti yükselir. VST ve rlog dışında, dönüştürülmüş

sayımları farklı bir modelleme çerçevesinde kullanan yaklaşımlar da vardır. Örneğin limma-voom, log-CPM değerleriyle birlikte ortalama-varyans ilişkisine dayalı gözlem düzeyi kesinlik ağırlıkları üreten ayrı bir modelleme yaklaşımıdır (Law vd., 2014).

Dönüştürülmüş değerler örnekler arası uzaklıkların, temel bileşenlerin ve kümelerin incelenmesinde; ayrıca ısı haritası (heatmap) gibi görselleştirmelerde kullanılabilir. Bu kontroller, beklenmeyen örnek ayrışmalarını ve olası parti yapılarını model kurulmadan önce görünür hale getirebilir. Dönüşüm ayrıca "Düzensiz gen ifadesi profilleri ve ROC-türevli tamamlayıcı tarama" bölümünde ele alınan ROC-türevli tamamlayıcı profil taraması gibi sürekli ölçekli analizler için uygun bir çalışma ölçeği sağlayabilir. Dolayısıyla kullanım yalnız keşifsel görselleştirmeyle sınırlı değildir; seçilen yöntemin analiz akışında veri ölçeği gereksinimine göre belirlenir.

Buna karşılık DESeq2'nin çıkarımsal (inferential) hipotez testleri VST veya rlog değerleri üzerinden yürütülmez; ham sayımlar, ölçekleme faktörleri ve negatif binom model birlikte kullanılır. Dönüştürülmüş değerlerle gözlenen örüntüler kalite değerlendirmesi ve yorum için destekleyici olabilir, fakat p-değeri ve çoklu test hata kontrolünün yerini alamaz. Bu ayrım, bir sonraki bölümde diferansiyel gen ifadesi hedefi ve bu bölümde izlenen DESeq2 modelleme çerçevesi ele alınırken korunacaktır.

Diferansiyel gen ifadesi analizi: Modelleme ve çıkarım

Diferansiyel gen ifadesi analizi, biyolojik koşullar arasında gözlenen farkı aynı koşul içindeki örnekler arası değişkenlikle karşılaştırır. Amaç yalnız grup ortalamalarını sıralamak değil, deney tasarımında tanımlanan koşul etkisinin belirsizliğini hesaplayarak hangi genlerde sistematik bir değişim için yeterli kanıt bulunduğunu belirlemektir. RNA-Seq verisinde bu çıkarım; ayrık sayımlar, teknik ölçek farkları ve biyolojik tekrarlar

arasındaki aşırı değişkenlik nedeniyle sayım verisine uygun bir olasılıksal model gerektirir (Conesa vd., 2016; Love vd., 2014).

Önceki bölümde açıklanan filtreleme ve normalizasyon adımları modelin güvenilir girdilerini ve teknik ölçeğini hazırlar; dönüşüm ise farklı aşağı akış amaçları için ayrı bir veri temsili sağlar. DESeq2'nin hipotez testi bunların içinde ham sayım matrisini yanıt olarak kullanan çıkarımsal adımdır. Girişte belirlenen yöntemsel odak doğrultusunda bu bölümde negatif binom model, tasarım matrisi, saçılım tahmini ve hata kontrolü tek bir DESeq2 akışı içinde ele alınacaktır.

İstatistiksel hedef ve yöntemsel çerçeve

İki gruplu basit bir çalışmada q_{iA} ve q_{iB}, i. gen için teknik ölçek etkilerinden arındırılmış grup düzeyi beklenen gen ifadelerini göstersin. Temel sıfır hipotezi $H_{0i}: q_{iA} = q_{iB}$, eşdeğer olarak $\log_2(q_{iB}/q_{iA}) = 0$ biçimindedir. Sıfır hipotezinin reddedilmesi, gözlenen farkın model altında yalnız örnekleme değişkenliğiyle açıklanmasının güç olduğuna işaret eder. Reddedilmemesi ise iki grubun kesinlikle aynı olduğunu göstermez; örneklem büyüklüğü ve değişkenlik düzeyi nedeniyle mevcut verinin farkı desteklemeye yetmediği anlamına da gelebilir.

İstatistiksel anlamlılık biyolojik önemle aynı değildir. Bir genin düşük p-değeri üretmesi hastalıkla nedensel ilişkisini kanıtlamaz; örneklem büyüklüğü yüksek olduğunda küçük etkiler de güçlü istatistiksel kanıt oluşturabilir. Tersine, büyük görünen bir fark az sayıda tekrar veya yüksek saçılım nedeniyle belirsiz olabilir. Sonuçlar çalışma tasarımı, etki büyüklüğü (effect size), standart hata, düzeltilmiş p-değeri ve örnek düzeyindeki örüntüyle birlikte değerlendirilmelidir. Gen ontolojisi ve yolak analizleri gen kümelerini biyolojik bağlama yerleştirebilir; ancak bu analizler belirli bir aday için nedensellik veya klinik kullanım kanıtı sağlamaz.

Deney tasarımı, hangi etkinin tahmin edilebileceğini belirler. Biyolojik tekrarlar koşul içindeki doğal değişkenliğin tahminini sağlarken yaş, cinsiyet, parti, eşleşmiş birey ve zaman gibi değişkenler uygun olduğunda modele eklenebilir. Aynı bireyden önce ve sonra alınan örnekler bağımsızmış gibi modellenirse bireyler arası değişkenlik koşul etkisine karışabilir. Benzer biçimde bir değişken koşulla tam olarak karışmışsa etkiler veriden ayrıştırılamaz. Gelişmiş bir model, yetersiz veya karışmış deney tasarımını sonradan düzeltemez.

Negatif binom model ve beklenen sayım

K_{ij}, $i = 1, \dots, p$ geni için $j = 1, \dots, n$ örneğinde gözlenen ham sayım olsun. Poisson dağılımı sayım verisi için doğal bir başlangıçtır, ancak varyansın ortalamaya eşit olduğunu varsayar. Biyolojik tekrarlı RNA-Seq verilerinde varyans çoğunlukla bu düzeyden büyüktür. Yalnız Poisson varsayımına dayanan bir model bu ek değişkenliği göz ardı ederek standart hataları küçültebilir. DESeq2, gözlenen sayımı

$$K_{ij} \sim \mathrm{NB}\left(\mu_{ij}, \alpha_i\right) \tag{1}$$

ile modeller. Burada μ_{ij} gözlenen sayım değil, negatif binom dağılımının beklenen sayımıdır; $\alpha_i \geq 0$ ise gene özgü aşırı saçılım parametresidir. Kullanılan parametreleştirmede

$$\mathrm{E}\left(K_{ij}\right) = \mu_{ij}, \qquad \mathrm{Var}\left(K_{ij}\right) = \mu_{ij} + \alpha_i \mu_{ij}^2 \tag{2}$$

olur (Anders ve Huber, 2010; Love vd., 2014). İlk varyans bileşeni Poisson sayım değişkenliğini, ikinci bileşen ise ortalamayla büyüyebilen ek örnekler arası değişkenliği temsil eder. $\alpha_i = 0$ sınırında model Poisson dağılımına yaklaşır; α_i arttıkça aynı beklenen sayım çevresindeki dağılım genişler.

Örnek başına tek ölçekleme faktörünün kullanıldığı varsayılan durumda beklenen sayım, “Normalizasyon ve ölçekleme

faktörleri" bölümünde tahmin edilen teknik ölçek ile koşula bağlı gen ifadesi bileşenine ayrılır:

$$\mu_{ij} = s_j q_{ij}. \tag{3}$$

s_j, j. örneğin ölçekleme faktörü; q_{ij} ise bu teknik ölçekten arındırılmış koşula bağlı beklenen gen ifadesi bileşenidir. Dolayısıyla s_j ve q_{ij} önce μ_{ij}'yi belirler; gözlenen K_{ij} de ortalaması μ_{ij}, varyansı yukarıdaki bağıntıyla verilen dağılımdan oluşur. Örneğin $s_j = 1.5$ ve $q_{ij} = 100$ ise beklenen sayım $\mu_{ij} = 150$'dir; gözlenen sayımın tam olarak 150 olması gerekmez ve olası sapmanın büyüklüğü α_i'ye de bağlıdır.

K_{ij}/s_j oranı betimsel bir normalize değer sağlayabilse de DESeq2 genelleştirilmiş doğrusal modelinin yanıtı bu oran değil ham K_{ij} sayımıdır. Bu ayrım, normalizasyonun veriyi model öncesinde sürekli bir ölçüme dönüştürmediğini gösterir. Teknik ölçek beklenen değer üzerinden modele girerken negatif binom olabilirliği gözlenen tam sayı sayımlar için hesaplanır. Dönüştürülmüş VST veya rlog değerleri de bu olabilirlik hesabının girdisi değildir.

Tasarım matrisi, katsayılar ve kontrastlar

DESeq2, negatif binom dağılımını genelleştirilmiş doğrusal model (generalized linear model, GLM) ile deney tasarımına bağlar. $\mathbf{X}$, $n \times R$ boyutlu tasarım matrisi; $\mathbf{x}_j$, j. örneğin tasarım satırı ve $\boldsymbol{\beta}_i$, i. gene ait log2 ölçekli katsayı vektörü olsun. Koşula bağlı beklenen gen ifadesi bileşeni önce

$$\log_2(q_{ij}) = \mathbf{x}_j^\top \boldsymbol{\beta}_i, \qquad q_{ij} = 2^{\mathbf{x}_j^\top \boldsymbol{\beta}_i}$$

biçiminde modellenir. Bu gösterim $\mu_{ij} = s_j q_{ij}$ ayrıştırmasına yerleştirilip doğal logaritma alındığında model

$$\log(\mu_{ij}) = \log(s_j) + \log(2)\ \mathbf{x}_j^\top \boldsymbol{\beta}_i \tag{4}$$

biçiminde yazılır. Buradaki $\log(2)$ çarpanı, $\log_2$ ölçeğinde tanımlanan katsayıları doğal logaritmalı bağlantı ölçeğine dönüştürür. Bilinen $\log(s_j)$ terimi ofset (offset) olarak modele girer; böylece teknik ölçek düzeltilirken katsayılar biyolojik tasarımı temsil eder (Love vd., 2014, 2026). İki gruplu bir tasarımda sabit terim (intercept) referans grubun temel düzeyini, grup katsayısı ise karşılaştırılan grubun referansa göre log2 kat değişimini verir. Örneğin grup katsayısının 1 olması beklenen gen ifadesinin iki katına, -1 olması yarıya inmesine karşılık gelir.

Tasarım matrisi yalnız grup göstergesinden oluşmak zorunda değildir. Parti, eşleşmiş birey, zaman, sürekli kovaryatlar ve etkileşimler aynı model içinde temsil edilebilir. Ancak her ek terim veri tarafından desteklenmelidir; çok az örneğe karşı aşırı karmaşık bir tasarım katsayı belirsizliğini artırabilir veya model matrisini tam sıralı olmaktan çıkarabilir. Referans düzeylerin ve katsayıların analizden önce tanımlanması, yazılım çıktısındaki işaret ve karşılaştırmaların doğru yorumlanması için gereklidir.

Birden fazla katsayının doğrusal birleşimi kontrastla tanımlanabilir. Böylece çok düzeyli bir faktörde belirli iki grubun farkı, bir etkileşim içindeki koşul etkisi veya iki etkinin ortalaması aynı uyarlanmış modelden değerlendirilebilir. Kontrastın veriye bakılmadan biyolojik soruya göre tanımlanması, yalnız düşük p-değeri üreten karşılaştırmaların seçilmesini önler. Modelin temel avantajı, koşul etkisini diğer tanımlı değişkenlere göre uyarlanmış biçimde tahmin edebilmesidir; fakat ölçülmemiş karıştırıcılar ve tam karışma bu uyarlamanın dışında kalır.

Aşırı saçılım ve etki tahmini

Aşırı saçılım parametresi α_i, gen içindeki biyolojik tekrarların ne ölçüde değiştiğini belirlediği için çıkarımın merkezindedir. Her gen için yalnız o genin sayımlarına dayanan bir saçılım tahmini elde edilebilir; ancak az sayıda biyolojik tekrar veya

düşük sayım bulunduğunda bu tahmin yüksek örnekleme değişkenliği taşır. Saçılımın olduğundan küçük tahmin edilmesi varyansın ve standart hataların küçülmesine, olduğundan büyük tahmin edilmesi ise test gücünün azalmasına neden olabilir. Bu nedenle DESeq2, gen-bazlı olabilirlik tahminini doğrudan nihai değer olarak kullanmak yerine genler arasında paylaşılan bilgiyle düzenlileştirir.

DESeq2 bu sorunu genler arasında bilgi paylaşan ampirik Bayes yaklaşımıyla ele alır. Süreç üç kavramsal adımdan oluşur: önce her gen için gen düzeyi saçılım tahmini elde edilir; sonra ortalama normalize sayım ile saçılım arasındaki ortak eğilim kestirilir; son olarak gen düzeyi olabilirlik ile eğilim çevresindeki önsel bilgi birleştirilir (Love vd., 2014). İlk aşamada negatif binom GLM her gen için ayrı ayrı uyarlanır ve Cox–Reid düzeltilmiş profil olabilirliği kullanılarak gen-bazlı saçılım tahmini elde edilir. Örnek sayısı arttıkça veya genin sayımları saçılım hakkında daha belirgin bilgi taşıdıkça olabilirlik belirli bir değer çevresinde yoğunlaşır; düşük sayımlı ya da az tekrarlı genlerde ise daha yatay kalarak saçılım tahmininin belirsizliğini artırır.

İkinci aşamada gen-bazlı tahminler, genlerin ortalama normalize sayımlarıyla birlikte değerlendirilerek veri setine özgü ortalama–saçılım eğilimi kestirilir. Varsayılan parametrik biçim

$$\alpha_{\mathrm{tr}}(\bar{\mu}) = \frac{a_1}{\bar{\mu}} + \alpha_0 \qquad (5)$$

şeklindedir. $\frac{a_1}{\bar{\mu}}$ terimi düşük ortalama sayımlarda belirginleşen ortalamaya bağlı bileşeni, α_0 ise yüksek ortalama sayımlarda yaklaşılan saçılım düzeyini temsil eder. Bu eğilim bütün genlere atanacak ortak bir saçılım değildir; benzer ifade düzeyindeki genler için beklenen saçılımı tanımlar ve sonraki ampirik Bayes adımında önsel dağılımın merkezi olarak kullanılır.

Eğilim çevresindeki gene özgü saçılım değerleri logaritmik ölçekte

$$\log \alpha_i \sim \mathrm{N}\left(\log \alpha_{\mathrm{tr}}\left(\bar{\mu}_i\right), \sigma_d^2\right) \qquad (6)$$

önseliyle modellenir. Burada önselin merkezi, i. genin ortalama normalize sayımına karşılık gelen $\alpha_{\mathrm{tr}}(\bar{\mu}_i)$ eğilim değeridir; σ_d^2 ise genlerin bu eğilim çevresindeki yayılımını gösterir ve veriden tahmin edilir. Böylece eğri çevresindeki toplam değişkenliğin tamamı gerçek genler arası farklılık olarak değerlendirilmez; az sayıdaki örnekten kaynaklanan tahmin belirsizliği de hesaba katılır.

Son aşamada gen-bazlı düzeltilmiş olabilirlik ile bu önsel bilgi birleştirilerek her gen için maksimum sonsal (maximum a posteriori, MAP) saçılım tahmini elde edilir. Saçılım büzülmesi (shrinkage), gen-bazlı tahminleri sabit bir ağırlıkla değil, taşıdıkları bilgi düzeyine göre genel ortalama–saçılım eğilimine doğru düzenlileştirir: olabilirliği yatay ve bilgisi sınırlı genler eğilime daha fazla yaklaştırılırken saçılımı veriler tarafından güçlü biçimde desteklenen genlerde büzülme daha azdır. Genel eğilimin belirgin biçimde üzerinde bulunan saçılım aykırılarında ise önselin uygun olmadığı kabul edilerek gen-bazlı tahmin korunur; böylece gerçekten yüksek değişkenlik gösteren genlerin varyansının yapay olarak küçültülmesi önlenir (Love vd., 2014).

Elde edilen nihai saçılım, GLM katsayılarının kovaryansını ve standart hatalarını belirleyerek Wald ya da olabilirlik oranı testine taşınır. Saçılım büzülmesi gürültü düzeyini temsil eden varyans parametresini düzenlerken kat değişimi büzülmesi etki büyüklüğü tahminlerini kararlı hâle getiren ayrı bir işlemdir. Güncel DESeq2 iş akışında Wald testi büzülmemiş katsayılarla yürütülür; büzülmüş kat değişimleri etki büyüklüğünün raporlanması, sıralanması ve görselleştirilmesi amacıyla daha sonra hesaplanır (Love vd., 2026).

Hipotez testleri ve hata kontrolü

Tek bir katsayı veya önceden tanımlanmış kontrast çoğunlukla Wald testiyle değerlendirilir. $\mathbf{c}$ kontrast vektörü ve $\widehat{\boldsymbol{\Sigma}}_i$, *i*. gene ait katsayı tahminlerinin kovaryans matrisi olsun. Test edilen etki $\theta_i = \mathbf{c}^\top \boldsymbol{\beta}_i$ için

$$\hat{\theta}_i = \mathbf{c}^\top \widehat{\boldsymbol{\beta}}_i, \qquad H_{0i}: \mathbf{c}^\top \boldsymbol{\beta}_i = 0, \qquad Z_i = \frac{\mathbf{c}^\top \widehat{\boldsymbol{\beta}}_i}{\sqrt{\mathbf{c}^\top \widehat{\boldsymbol{\Sigma}}_i \mathbf{c}}} \tag{7}$$

istatistiği kullanılır. Tek bir katsayı sınanacaksa $\mathbf{c}$, ilgili katsayı için 1, diğerleri için 0 içeren birim vektör olarak seçilir; iki katsayının farkı veya birden fazla katsayının doğrusal birleşimi de uygun $\mathbf{c}$ tanımıyla aynı çerçevede sınanabilir. Düzenli koşullar altında Z_i, sıfır hipotezi altında yaklaşık standart normal dağılıma sahiptir. Testin yönü, referans düzeyi ve kontrast katsayıları birlikte raporlanmalıdır.

Olabilirlik oranı testi (likelihood ratio test, LRT), iç içe geçmiş tam ve indirgenmiş modelleri karşılaştırır:

$$D_i = 2\left(\ell_{\text{full},i} - \ell_{\text{reduced},i}\right). \tag{8}$$

D_i, sıfır hipotezi altında iki model arasındaki parametre sayısı farkına karşılık gelen serbestlik derecesiyle yaklaşık ki-kare dağılımına göre değerlendirilir. Wald testi tek bir etki için; LRT ise çok düzeyli faktör, zaman örüntüsü veya bir katsayı kümesinin modele toplu katkısı için daha uygundur. LRT'nin anlamlı olması hangi katsayının değiştiğini tek başına göstermez; ilgili katsayılar ve planlanmış karşılaştırmalar ayrıca incelenmelidir.

Binlerce genin eş zamanlı test edilmesi ham p-değerlerinde yanlış pozitif birikimine yol açar. DESeq2, Benjamini–Hochberg düzeltmesiyle yanlış keşif oranını (false discovery rate, FDR) kontrol etmeyi hedefler. Düzeltilmiş p-değeri, tek bir genin yanlış olma olasılığı değildir; seçilen gen kümesindeki beklenen yanlış keşif oranı bağlamında yorumlanır. Anlamlılık eşiği ile biyolojik

etki eşiği birbirinden ayrılmalı ve mümkünse analiz öncesinde tanımlanmalıdır (Love vd., 2014, 2026).

Çok düşük ortalama sayıma sahip genler çoğunlukla sınırlı test gücüne sahiptir. DESeq2, ortalama normalize sayımı kullanarak çoklu test düzeltmesinden önce bağımsız filtreleme yapabilir; sıfır hipotezi altında filtreleme ölçütü test istatistiğinden uygun biçimde bağımsızsa bu işlem FDR kontrolünü bozmadan keşif gücünü artırabilir. Bağımsız hipotez ağırlıklandırması (independent hypothesis weighting, IHW), uygun bir bağımsız kovaryata göre hipotezleri ağırlıklandırarak çoklu test gücünü artırmayı amaçlayan ileri bir seçenektir (Ignatiadis vd., 2016).

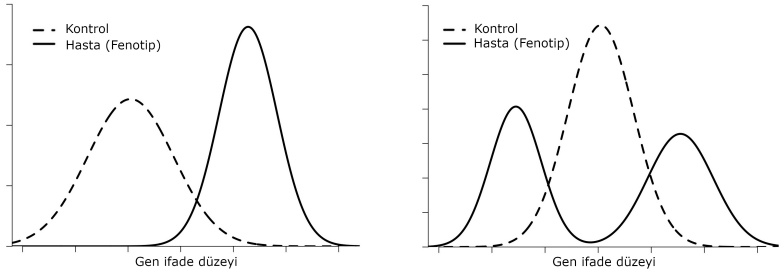
Sonuçlar yalnız "anlamlı" ve "anlamsız" ayrımına indirgenmemeli; log2 kat değişimi, standart hata ve düzeltilmiş p-değeri birlikte değerlendirilmelidir. Aykırı değerlerin katsayı ve saçılım tahminlerini yönlendirebileceği göz önünde bulundurulmalı; veri dışlama kararları yalnız istatistiksel sonuca göre verilmemeli ve uygulanan ölçütler açıkça raporlanmalıdır. Tasarım formülü, referans düzeyi, kontrast, filtreleme kuralı ve anlamlılık hedefi de analizin yeniden üretilebilirliği için belgelenmelidir.

DESeq2 sonuçları, tasarım matrisinde tanımlanan katsayı ve kontrastlarla sınırlıdır. Koşul ilişkisi tutarlı bir ortalama artışı veya azalışı olarak ortaya çıktığında bu çerçeve güçlü ve yorumlanabilirdir. Aynı grupta hem olağan dışı düşük hem olağan dışı yüksek gen ifadesi değerleri bilgi taşıyorsa tek bir ortalama farkı örüntünün tamamını özetleyemeyebilir. Sonraki bölümde bu sınır, düzensiz (improper) gen ifadesi profilleri ve tamamlayıcı tarama gereksinimi üzerinden ele alınacaktır (Başol Göksülük vd., 2026).

Düzensiz gen ifadesi profilleri ve ROC-türevli tamamlayıcı tarama

DESeq2, tasarım matrisinde tanımlanan katsayı ve kontrastlar aracılığıyla koşula bağlı beklenen gen ifadesi değişimini sınar. Basit iki gruplu bir karşılaştırmada, örneklerin çoğunda aynı yönde ortaya çıkan artış veya azalış bu hedefle iyi temsil edilir. Bununla birlikte aynı fenotip grubundaki örneklerin bir bölümü çok düşük, başka bir bölümü çok yüksek gen ifadesi gösterebilir. İki uç birbirini dengelediğinde grup ortalaması referans grubuna yakın kalabilir ve tek bir koşul katsayısı dağılım örüntüsünün tamamını özetleyemez. Buradaki sorun negatif binom modelin geçersizliği değil, modelde test edilen hedef ile araştırılan örüntünün farklı olmasıdır.

Bu kitapta düzenli (proper) ve düzensiz profil terimleri, Başol Göksülük vd. (2026) tarafından benimsenen operasyonel kullanım doğrultusunda kullanılmaktadır. Düzenli profilde koşulla ilişkili değişim çoğunlukla tek yönlüdür. Düzensiz profilde ise düşük ve yüksek uçların ikisi de fenotiple ilişkili bilgi taşıyabilir. Bu sınıflama alanın tamamında yerleşik, birbirini dışlayan bir taksonomi değildir. Tek tepeli (unimodal) veya iki tepeli (bimodal) dağılım, monoton olmayan ilişki, alt grup özgüllüğü ve varyans farklılığı birbiriyle örtüşebilse de eş anlamlı değildir; her düzensiz profil iki tepeli, her iki tepeli profil de incelenen koşulla ilişkili olmak zorunda değildir.



(a) Düzenli profil (b) Düzensiz profil

Şekil 2. Düzenli ve düzensiz gen ifadesi profillerinin temsili dağılımları. Kesikli eğriler referans grubunu, düz eğriler fenotiple ilişkili grubu göstermektedir.

Şekil 2'de düzenli profilde fenotiple ilişkili dağılım tek yönde yer değiştirirken düzensiz profilde düşük ve yüksek değerli iki alt yapı görülmektedir. Düzensizlik yalnız iki tepeli dağılımla sınırlı değildir: grup ortalamaları belirgin değişmeden fenotip grubunun varyansı artabilir veya hastalık ilişkisi bir denge aralığının iki ucundan sapma biçiminde ortaya çıkabilir. Kanser gibi heterojen hastalıklarda moleküler alt tipler, doku bileşimi ve farklı düzenleyici mekanizmalar aynı klinik sınıf içinde bu tür örüntüler oluşturabilir. Hem yetersiz hem aşırı moleküler etkinliğin olumsuz sonuçlarla ilişkilendirildiği *Goldilocks etkisi* ve iki tepeli miRNA profillerinin klinik alt gruplarla ilişkilendirildiği çalışmalar bu biyolojik olasılığa örnek sağlar (Ma ve Tangye, 2023; Moody vd., 2022). Yine de düzensiz bir profil tek başına nedensel mekanizma veya doğrulanmış biyobelirteç değildir; bağımsız veri ve deneysel kanıt gerektiren bir aday sinyaldir.

Düşük ve yüksek değerlerin aynı klinik grupta bulunması farklı biyolojik açıklamalarla uyumlu olabilir. Bir gen farklı moleküler alt tiplerde karşıt yönlerde düzenlenebilir, hücre bileşimi değişen örneklerde farklı gen ifadesi düzeyleri gösterebilir veya yalnız belirli bir mekanizmanın etkin olduğu hasta alt kümesinde uç değerlere ulaşabilir. Fenotip grubundaki varyans artışı da benzer bir dağılımsal görünüm oluşturabilir. Bu açıklamalar gözlenen profilden tek başına ayırt edilemez. Profil taramasının işlevi mekanizmayı kanıtlamak değil, tek yönlü koşul etkisiyle yeterince özetlenmeyen ve ayrıntılı incelemeyi hak eden genleri görünür kılmaktır.

Tespit zorlukları ve literatürdeki yaklaşımlar

Düzensiz profillerin sistematik olarak belirlenmesi üç nedenle zordur. İlk olarak düşük ve yüksek uçlar birlikte bulunduğunda kat değişimi küçük kalabilir; buna karşılık dağılım biçimi veya alt grup yapısı biyolojik bilgi taşıyabilir. İkinci olarak RNA-Seq sayımları teknik ölçek, sıfır yoğunluğu ve aşırı saçılımdan

etkilenir. Görülen uçların biyolojik heterojenlikten mi, parti etkisinden mi, örnek kalitesinden mi kaynaklandığı kalite kontrolü olmadan ayırt edilemez. Üçüncü olarak binlerce genin görsel olarak incelenmesi tekrarlanabilir değildir. Yüksek boyutlu taramada kullanılan ölçütün hangi örüntüyü özetlediği ve çıkarımsal hata kontrolü sağlayıp sağlamadığı açıkça ayrılmalıdır.

Literatürdeki yöntemler genel olarak birkaç ailede toplanabilir. Kanser aykırı değer profil analizi (cancer outlier profile analysis, COPA) ve aykırı değer toplamı (outlier sum), yalnız bazı tümör örneklerinde biriken uç sinyalleri ortalama farkından bağımsız olarak önceliklendirmiştir (MacDonald ve Ghosh, 2006; Tibshirani ve Hastie, 2007; Tomlins vd., 2005). Dağılım biçimini hedefleyen dip testi tek tepelilikten sapmayı sınar (Hartigan ve Hartigan, 1985); iki tepelilik indeksleri iki tepeli örüntüleri sayısal olarak özetler (Wang vd., 2009a); Gauss karışım modelleri (Gaussian mixture models, GMM) ve GMMchi ise gizli alt grup veya karışım yapısını araştırır (Liu vd., 2022). RNA-Seq'e yönelik SIBER (systematic identification of bimodally expressed genes using RNAseq data) iki tepeli gen ifadesi örüntülerini RNA-Seq bağlamında ele alır (Tong vd., 2013). Bu yöntem aileleri farklı dağılım, dönüşüm ve kümeleme varsayımlarına dayanır; bu nedenle ürettikleri adaylar aynı istatistiksel anlamı taşımaz.

Uç değer istatistikleri az sayıdaki belirgin örneğe duyarlıyken GMM alt grupların dağılımsal biçimini açıkça tanımlamayı gerektirir. Tek tepelilik testleri dağılım biçimini değerlendirir, ancak saptanan iki tepeliliğin fenotiple ilişkili olup olmadığını tek başına göstermez. Bu nedenle aday genler yorumlanırken yalnız skor büyüklüğüne değil, yöntemin kullandığı veri ölçeğine, dönüşüm seçimine, karışım modeli varsayımlarına ve örneklem yapısına da bakılmalıdır. Dolayısıyla yöntem seçimi yalnız aday

sayısına göre değil, hedeflenen örüntü ve yöntemin bu örüntüyü hangi varsayımla tanımladığına göre yapılmalıdır.

Alıcı işletim karakteristiği (receiver operating characteristic, ROC) tabanlı yaklaşımlar, belirteç dağılımının iki sınıfı ne ölçüde ayırabildiğine odaklanan başka bir çizgidir. Mikrodizilim bağlamında önerilen *not proper* ROC ve ok grafiği (arrow plot) yaklaşımları alt grup özgül örüntüleri görünür kılmayı amaçlamıştır (Parodi vd., 2008; Silva-Fortes vd., 2012). RNA-Seq verisinde de ROC-türevli skorlar, kalite kontrolü ve filtreleme sonrasında, örneğin varyans dengeleyici dönüşümle elde edilen karşılaştırılabilir değerler üzerinde incelenebilir (Başol Göksülük vd., 2026). Bu yaklaşımların DESeq2 ile görev ayrımı ve tamamlayıcı kullanımı sonraki alt bölümde açıklanmaktadır.

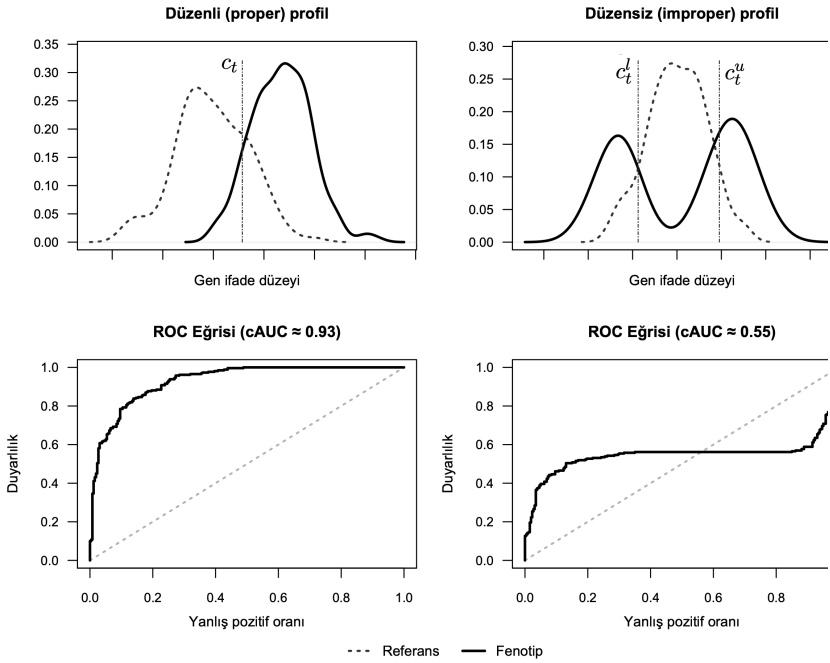
Tamamlayıcı ROC-türevli tarama

DESeq2 ile ROC-türevli tarama aynı soruya rakip yanıtlar üretmez. DESeq2, önceden tanımlanmış tasarım altında beklenen sayım etkisi için katsayı, standart hata ve hata kontrollü hipotez testi sunar. ROC skorları ise bir genin iki sınıf arasında tek yönlü veya iki uçlu ayrım örüntüsü gösterip göstermediğini betimler. Alt grup bilgisi önceden biliniyorsa tasarım matrisi ve etkileşimler bu yapıyı doğrudan modelleyebilir; alt grup bilinmiyorsa ROC-türevli ölçütler yalnız aday keşfi sağlar. Bu nedenle düşük DESeq2 kanıtı ile güçlü bir tamamlayıcı ROC skoru çelişki değil, farklı hedeflerin sonucudur.

Klasik ROC yaklaşımı tek yönlü bir ayrımı özetler. Belirteç yükseldikçe fenotip olasılığının arttığı varsayılırsa eşik üzerindeki değerler fenotiple ilişkili kabul edilir; ilişki ters yöndeyse yön buna göre çevrilir. Bu mantık, gen ifadesinin bir grupta tutarlı biçimde yüksek veya düşük olduğu düzenli profillere uygundur. Düzensiz profilde ise aynı grubun hem alt

hem üst uçları bilgi taşıdığından tek eşik iki sinyali birlikte temsil edemez.

Şekil 3'te düzenli senaryonun klasik ROC eğrisi köşegenin üzerinde seyrederek güçlü tek yönlü ayrım gösterir. Düzensiz senaryoda ise düşük ve yüksek değerler aynı fenotip grubunda toplandığından ROC eğrisi S biçimi alır; eğrinin köşegen üstündeki ve altındaki parçaları klasik alan hesabında kısmen dengelenebilir. Bu nedenle klasik alanın 0.5'e yakın olması her zaman sınıflar arasında dağılımsal ayrım bulunmadığını göstermez. Böyle bir durumda iki uçlu karar bölgesi veya eğrinin geometrik biçimi ek bilgi sağlayabilir.



Şekil 3. Düzenli ve düzensiz gen ifadesi senaryolarında temsili dağılımlar ve ROC eğrileri.

Klasik ROC eğrisi

ξ fenotiple ilişkili, η ise referans gruptaki gen ifadesi değerini göstersin. $x \in \mathbb{R}$ biyobelirteç ölçeğindeki genel bir değer olmak üzere iki grubun sürekli dağılım fonksiyonları

$$F_{\xi}(x) = \Pr(\xi \leq x), \qquad F_{\eta}(x) = \Pr(\eta \leq x)$$

olarak tanımlansın. Yüksek değerlerin fenotiple ilişkili olduğu yön altında, yanlış pozitif oranı (YPO; false positive rate, FPR) t'ye karşılık gelen kesim noktası c_t ile gösterilsin. Bu eşikten hareketle klasik ROC eğrisi $R(t)$ ve klasik ROC eğrisi altında kalan alan (classical area under the ROC curve, cAUC), ψ_{cAUC},

$$c_t = F_{\eta}^{-1}(1-t), \qquad t = 1 - F_{\eta}(c_t),$$
$$R(t) = 1 - F_{\xi}(c_t),\ \psi_{\mathrm{cAUC}} = \int_0^1 R\,(t)\,dt, \qquad 0 \leq t \leq 1 \qquad (9)$$

olarak tanımlanır. cAUC'nin 0.5'e yakın olması seçilen yöndeki ayrımın zayıf, 1'e yakın olması güçlü olduğunu gösterir. Bu olasılıksal sıralama özelliği düzenli profiller için yararlıdır; ancak S-biçimli eğride farklı yönlerdeki parçaları tek alanda topladığı için düzensizliği eksik özetleyebilir.

Sürekli dağılım varsayımı altında cAUC, fenotiple ilişkili gruptan rastgele seçilen bir değerin referans grubundan seçilen değerden daha yüksek olma olasılığı olarak yorumlanabilir. Bu yorum tek yönlü sıralamaya dayanır. Düşük ve yüksek uçların ikisi de fenotiple ilişkili olduğunda tek bir büyüklük sırası yeterli değildir; bir uç doğru yönde sıralanırken diğer uç ters yönde kalır. Düzensiz profilde klasik alanın sınırlılığı bu olasılıksal hedeften kaynaklanır.

Genelleştirilmiş ROC eğrisi

Genelleştirilmiş ROC eğrisi (generalized ROC curve, gROC), tek eşikli karar yerine Şekil 3'te görüldüğü gibi biyobelirteç dağılımının iki ucunu birlikte değerlendiren bir karar bölgesi

tanımlar (Martínez-Camblor vd., 2017). Klasik ROC eğrisinden farklı olarak toplam YPO t, dağılımın alt ve üst kuyruklarına paylaştırılır. Bu paylaşıma karşılık gelen genel alt ve üst biyobelirteç kesim noktaları sırasıyla c^l ve c^u ile gösterilsin; burada $c^l \leq c^u$'dur. $(-\infty, c^l) \cup (c^u, \infty)$ bölgesindeki değerler fenotiple ilişkili, $[c^l, c^u]$ aralığındaki değerler referans profiline ait olarak sınıflandırılır.

Süreklilik varsayımı altında sınır noktalarının açık veya kapalı seçilmesi olasılıkları değiştirmez. İki kesim noktalı karar kuralının duyarlılığı (Se), YPO'su ve seçiciliği (Sp)

$$\begin{aligned} \mathrm{Se}(c^l, c^u) &= \Pr\left(\{\xi < c^l\} \cup \{\xi > c^u\}\right) \\ &= F_\xi(c^l) + 1 - F_\xi(c^u), \qquad (10) \\ \mathrm{Sp}(c^l, c^u) &= \Pr(c^l \leq \eta \leq c^u) \\ &= F_\eta(c^u) - F_\eta(c^l) \\ &= 1 - \mathrm{YPO}(c^l, c^u), \qquad (11) \\ \mathrm{YPO}(c^l, c^u) &= \Pr\left(\{\eta < c^l\} \cup \{\eta > c^u\}\right) \\ &= F_\eta(c^l) + 1 - F_\eta(c^u) \qquad (12) \end{aligned}$$

ile gösterilir. Böylece referans grubundaki toplam YPO, alt kuyruktaki $F_\eta(c^l)$ ile üst kuyruktaki $1 - F_\eta(c^u)$ olasılıklarının toplamıdır.

Önceden belirlenen $t \in [0,1]$toplam YPO düzeyi için bu düzeyi sağlayan kesim noktası çiftleri

$$\mathcal{I}_t = \left\{(c^l, c^u) \in \mathbb{R}^2 : c^l \leq c^u,\ F_\eta(c^l) + 1 - F_\eta(c^u) = t\right\} \qquad (13)$$

kümesini oluşturur. Burada (c^l, c^u), $\mathcal{I}_t$ kümesini oluşturan uygun çiftlerden herhangi birini, t ise toplam YPO düzeyini gösterir. gROC eğrisi, her t düzeyinde $\mathcal{I}_t$ içindeki en yüksek duyarlılığı seçerek

$$R_g(t) = \sup_{(c^l, c^u) \in \mathcal{I}_t} \mathrm{Se}(c^l, c^u)$$
$$= \sup_{(c^l, c^u) \in \mathcal{I}_t} \left\{ F_\xi(c^l) + 1 - F_\xi(c^u) \right\} \quad (14)$$

biçiminde tanımlanır. En iyi kesim noktası çifti t'ye göre değişebileceğinden $R_g(t)$, tek bir sabit kesim noktası çiftinin performans eğrisi değildir. Supremuma ulaşan uygun çift varsa bu optimum çift $\left(c_t^{l,*}, c_t^{u,*}\right)$ ile gösterilebilir. Şekil 3'teki $\left(c_t^l, c_t^u\right)$ eşikleri belirli bir toplam YPO düzeyi altında seçilen bir uygun alt–üst kesim noktası çiftine karşılık gelir.

Genelleştirilmiş eğri altında kalan alan (generalized area under the curve, gAUC), ψ_{gAUC},

$$\psi_{\mathrm{gAUC}} = \int_0^1 R_g\,(t)\,dt \quad (15)$$

ile tanımlanır. cAUC'den farklı olarak gAUC, fenotip ve referans gruplarından seçilen iki değerin sıralanma olasılığı biçiminde doğrudan bir olasılıksal yoruma sahip değildir; seçicilik $1 - t$ değiştikçe iki eşikli karar kuralıyla ulaşılabilen en yüksek duyarlılıkların ortalamasını özetler (Martínez-Camblor vd., 2017). Kuramsal gAUC, popülasyon dağılımlarına ait bu fonksiyoneldir; veri analizinde dağılım fonksiyonları ampirik karşılıklarıyla değiştirildiğinde $\hat{R}_g(t)$ ve $\hat{\psi}_{\mathrm{gAUC}}$ elde edilir. Örnekleme, eşik kararlılığı ve aşırı uyum uyarıları bu ampirik tahmin bağlamında değerlendirilmelidir.

Yüksek gAUC tek başına düzensiz veya iki tepeli profil kanıtı değildir; genelleştirilmiş eğri klasik tek kuyruklu çözümleri de sınır durum olarak içerebildiğinden güçlü tek yönlü bir belirteç de yüksek gAUC üretebilir. Düzensiz profil yorumu daha çok düşük veya orta cAUC'ye karşı yüksek ampirik gAUC, optimum çözümde YPO'nun iki kuyruğa gerçekten paylaştırılması, S-biçimli ROC görünümü ve biyobelirteç dağılımının birlikte

değerlendirilmesine dayanmalıdır. Buna karşın S-biçimli ROC eğrisi de tek başına iki tepeli düzensizliği kanıtlamaz; fenotip grubunun referans gruba göre çok daha yüksek varyans göstermesi gibi farklı dağılım mekanizmaları benzer geometriler oluşturabilir.

Ampirik $\hat{\psi}_{\mathrm{gAUC}}$ hesabında $\left(c_t^{l,*}, c_t^{u,*}\right)$'nun aynı veri üzerinde en iyi ayrımı verecek biçimde seçilmesi özellikle küçük örneklemlerde aşırı uyuma yol açabilir. Bu kesim noktaları biyolojik olarak sabit sınırlar gibi yorumlanmamalı; aday profilin iki uçlu yapısını tanımlayan veri bağımlı değerler olarak raporlanmalıdır. Eşik kararlılığı yeniden örnekleme veya bağımsız veriyle değerlendirilebilir. Bu skor tek başına p-değeri veya yanlış keşif oranı kontrolü sağlamaz.

ROC eğrisi uzunluğu

ROC eğrisinin uzunluğu (length of the ROC curve, LROC), eğrinin birim kare içindeki geometrik uzunluğunu özetler. Franco-Pereira vd. (2020), bu ölçütü parametrik binormal ROC modeli altında ele almıştır. Bu modelden elde edilen yumuşatılmış ve türevlenebilir $R(t)$ eğrisi için

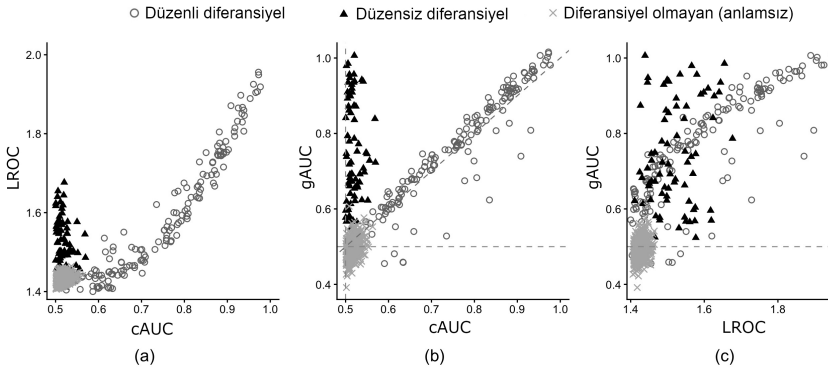
$$\mathrm{LROC} = \int_0^1 \sqrt{1 + \left(\frac{dR(t)}{dt}\right)^2}\, dt \qquad (16)$$

şeklinde tanımlanır. Köşegen üzerindeki rastgele ayrım çizgisinin uzunluğu $\sqrt{2}$, ideal ayrım sınırına yaklaşan eğrinin uzunluğu ise yaklaşık 2'dir. LROC alan yerine eğrinin biçimine duyarlı olduğundan cAUC'nin orta düzeyde kaldığı bazı S-biçimli örüntülerde tamamlayıcı bilgi verebilir. Denklem ampirik basamaklı eğriye doğrudan uygulanmadığından binormal model uyumunun yeterliliği ayrıca değerlendirilmelidir.

Tarama akışı ve sınırlılıklar

Tamamlayıcı taramada önce ham sayımların kalite kontrolü ve sınırlı bilgi taşıyan genlerin filtrelenmesi tamamlanır. DESeq2 çıkarımı ham sayımlar ve ölçekleme faktörleriyle yürütülür. ROC-türevli skorlar ise amaçlanan sürekli ölçekli karşılaştırmaya uygun olarak VST gibi dönüştürülmüş değerlerden hesaplanabilir. Genler cAUC, ampirik gAUC ve LROC ile sıralanır; düşük veya orta cAUC'ye karşı yüksek ampirik gAUC ya da LROC gösteren profiller dağılım grafikleri, örnek bilgisi ve teknik kalite ölçütleriyle birlikte incelenir.

Bu sıralama otomatik bir aday kararı değildir. Öncelikli genlerde grup dağılımları, tekil örneklerin etkisi ve olası parti yapısı kontrol edilmeli; farklı dönüşüm veya makul filtreleme seçimlerinde skorların korunup korunmadığı incelenmelidir. DESeq2 etki tahminleri ile ROC skorlarının yan yana raporlanması, koşul kontrastına ilişkin çıkarımsal kanıt ile dağılım örüntüsüne ilişkin keşifsel kanıtın birbirine karıştırılmasını önler.



Şekil 4. Benzetim çalışmasında gerçek sınıfı bilinen genlerin ROC-türevli skor uzayındaki yerleşimi: (a) cAUC–LROC, (b) cAUC–gAUC ve (c) LROC–gAUC ilişkileri. Düzenli diferansiyel, düzensiz diferansiyel ve diferansiyel olmayan etiketleri bir yöntem tarafından tahmin edilmemiş, benzetim tasarımında önceden tanımlanmıştır.

Şekil 4, gerçek durumları benzetim tasarımında bilinen genlerin ortak skor uzayında nasıl yerleştiğini göstermektedir.[5] Düzenli diferansiyel genler çoğunlukla cAUC ile gAUC ve LROC eksenlerinde birlikte yükselirken düzensiz genler düşük veya orta cAUC bölgesinde kalıp ampirik gAUC ya da LROC yönünde ayrışabilir. Bu görünüm, etiketlerin DESeq2 veya ROC skorlarından tahmin edildiği bir sınıflandırma başarısı değildir; bilinen sınıfların skorlarla ilişkisini açıklayan bir benzetim sonucudur (Başol Göksülük vd., 2026).

Skor eşikleri veri setine, dönüşüme ve beklenen heterojenliğe bağlıdır. Bu nedenle adaylar *istatistiksel olarak anlamlı diferansiyel gen* olarak değil, *izlemeye değer düzensiz profil adayları* olarak raporlanmalıdır. Skor tablosu gen düzeyi dağılımları, parti ve örnek kalitesi ile biyolojik gerekçeyle birlikte incelenmeli; bağımsız veri veya deneysel bulgular ayrı doğrulama kanıtı olarak sunulmalıdır. Bu görev ayrımı korunduğunda ROC-türevli ölçütler, DESeq2'nin çıkarımsal sonucunu değiştirmeden farklı bir örüntü hedefini görünür kılan tamamlayıcı araçlar sunar.

TARTIŞMA VE SONUÇ

RNA-Seq analizinde yöntem seçimi; biyolojik sorunun, veri ölçeğinin, deney tasarımının ve hedeflenen çıkarımın birbiriyle uyumlu olmasını gerektirir. Ayrık ve aşırı saçılımlı ham sayımların negatif binom dağılımıyla doğrudan modellenmesi, DESeq2 ve edgeR için istatistiksel olarak tutarlı bir temel sağlar (Love vd., 2014; Robinson vd., 2010). Bununla birlikte bu özellik, negatif binom yöntemlerinin her veri setinde koşulsuz olarak üstün olduğu anlamına gelmez. Limma-voom, logaritmik ölçekteki ortalama–varyans ilişkisini tahmin ederek gözlem düzeyinde kesinlik ağırlıkları kullanan geçerli bir alternatif sunar

[5] Şekil, Başol Göksülük vd. (2026) çalışmasındaki benzetim kurgusu ve yazarların paylaştığı kodlar temel alınarak yeniden üretilmiştir; görsel birebir kopya değildir.

(Law vd., 2014). Karşılaştırma çalışmalarında yöntemlerin göreli performansı örneklem büyüklüğü, saçılım, kütüphane dengesizliği, etkili gözlemler ve deney tasarımına göre değişebilmektedir (Baik vd., 2020; Schurch vd., 2016; Soneson ve Delorenzi, 2013). DESeq2'nin burada merkeze alınması bu nedenle evrensel bir üstünlük iddiasına değil, ham sayımdan hata kontrollü çıkarıma uzanan sürecin tek ve bütünlüklü bir model üzerinden açıklanmasına dayanır.

DESeq2 iş akışının güvenilirliği yalnız kullanılan dağılıma bağlı değildir. Biyolojik tekrarların yeterliliği, karıştırıcı değişkenlerin tasarım matrisinde uygun temsil edilmesi, kontrastların biyolojik soruya göre önceden tanımlanması ve çoklu test sonuçlarının etki büyüklükleriyle birlikte yorumlanması gerekir. Aykırı veya etkili gözlemler katsayı ve saçılım tahminlerini yönlendirebileceğinden, örnek ve gen düzeyindeki tanısal kontroller ile makul analiz kararlarına yönelik duyarlılık değerlendirmeleri önemlidir (Love vd., 2026, 2014). Bununla birlikte veri dışlama veya model değiştirme kararları yalnız daha düşük p-değerleri elde etme amacıyla verilmemeli; kullanılan ölçütler ve sonuçlara etkileri açıkça raporlanmalıdır.

Diferansiyel gen ifadesi analizi ile düzensiz profil taraması farklı kanıt hedeflerine sahiptir. DESeq2, tasarım matrisindeki etkiler için belirsizlik ve çoklu test düzeltmesiyle birlikte çıkarımsal sonuçlar üretir; cAUC, gAUC ve LROC ise ortalama farkıyla yeterince özetlenemeyen dağılım örüntülerini keşifsel olarak önceliklendirir. ROC-türevli skorlar p-değeri veya kontrollü FDR sağlamadığından iki sonuç türü birbirinin yerine geçirilmemelidir (Başol Göksülük vd., 2026).

ROC-türevli taramanın esnekliği aynı zamanda temel sınırlılığıdır. Veriden seçilen eşikler, özellikle küçük örneklemlerde ve etkili gözlemlerin varlığında aday ayrımını iyimser gösterebilir. Eşik ve sıralama kararlılığı bu nedenle ön

tanımlı kararlar, yeniden örnekleme, duyarlılık incelemeleri ve bağımsız veriyle değerlendirilmelidir. gAUC ve LROC doğrulanmış sınıflandırma performansı olarak değil, ayrıntılı gen düzeyi incelemeye yön veren aday sıralama araçları olarak yorumlanmalıdır (Başol Göksülük vd., 2026; Martínez-Camblor vd., 2017).

Sonuç olarak güvenilir bir RNA-Seq analizi, sayım verisine uygun modellemeyi güçlü deney tasarımı ve temkinli biyolojik yorumla birleştirmelidir. DESeq2, düzenli koşul etkileri için çıkarımsal bir çerçeve sunarken ROC-türevli tarama bu hedefin dışında kalabilecek dağılım örüntülerini görünür kılabilir. Her iki yaklaşımın varsayımları, hata kontrolü ve kanıt düzeyi ayrı tutulmalı; önceliklendirilen genler bağımsız veri ve uygun moleküler çalışmalarla doğrulanmadan biyobelirteç veya mekanizma olarak kabul edilmemelidir. Bu görev ayrımı, istatistiksel anlamlılık ile biyolojik önemi birbirine karıştırmadan daha kapsamlı bir değerlendirme yapılmasını sağlar.

KAYNAKÇA

Anders, S. ve Huber, W. (2010). Differential Expression Analysis for Sequence Count Data. *Genome Biology*, 11(10):R106.

Baik, B., Yoon, S., ve Nam, D. (2020). Benchmarking RNA-Seq Differential Expression Analysis Methods Using Spike-In And Simulation Data. *PLoS One*, 15(4):e0232271.

Başol Göksülük, M., Öztürk, E., Erkorkmaz, Ü., Deveci Özkan, A., Kepenek, H. F., ve Göksülük, D. (2026). Complementary ROC-Derived Indices for Screening Improper Expression Profiles in RNA-Seq Differential Expression Analysis. *Balkan Medical Journal*, 43(6):330–342.

Bullard, J. H., Purdom, E., Hansen, K. D., ve Dudoit, S. (2010). Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *BMC Bioinformatics*, 11(1):94.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., ve Mortazavi, A. (2016). A Survey of Best Practices for RNA-Seq Data Analysis. *Genome Biology*, 17(1):13.

Dudoit, S., Fridlyand, J., ve Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–87.

Franco-Pereira, A. M., Nakas, C. T., ve Pardo, M. C. (2020). Biomarker Assessment in ROC Curve Analysis Using the Length of the Curve As An Index of Diagnostic Accuracy: The Binormal Model Framework. *AStA Advances in Statistical Analysis*, 104(4):625–647.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., ve Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537.

Hardcastle, T. J. ve Kelly, K. A. (2010). baySeq: Empirical Bayesian Methods for Identifying Differential Expression in Sequence Count Data. *BMC Bioinformatics*, 11(1):422.

Hartigan, J. A. ve Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics*, 13(1):70–84.

Ignatiadis, N., Klaus, B., Zaugg, J. B., ve Huber, W. (2016). Data-Driven Hypothesis Weighting Increases Detection Power in Genome-Scale Multiple Testing. *Nature Methods*, 13(7):577–580.

Kukurba, K. R. ve Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb.top084970.

Law, C. W., Chen, Y., Shi, W., ve Smyth, G. K. (2014). voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts. *Genome Biology*, 15(2):R29.

Li, J. ve Tibshirani, R. (2013). Finding Consistent Patterns: A Nonparametric Approach for Identifying Differential Expression in RNA-Seq Data. *Statistical Methods in Medical Research*, 22(5):519–536.

Li, J., Witten, D. M., Johnstone, I. M., ve Tibshirani, R. (2012). Normalization, Testing, and False Discovery Rate Estimation for RNA-Sequencing Data. *Biostatistics*, 13(3):523–538.

Liu, T.-C., Kalugin, P. N., Wilding, J. L., ve Bodmer, W. F. (2022). GMMchi: Gene Expression Clustering Using Gaussian Mixture Modeling. *BMC Bioinformatics*, 23(1):457.

Love, M. I., Anders, S., ve Huber, W. (2026). Analyzing RNA-Seq Data With DESeq2. Bioconductor 3.23 Vignette, DESeq2 Version 1.52.0. (E.T. 2026-07-21).

Love, M. I., Huber, W., ve Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data With DESeq2. *Genome Biology*, 15(12):550. https://doi.org/10.1186/s13059-014-0550-8

Ma, C. S. ve Tangye, S. G. (2023). Inborn Errors of Immunity: The Goldilocks Effect—Susceptibility To Disease Due To A Little Too Much or A Little Too Little. *Clinical and Experimental Immunology*, 212(2):93–95.

MacDonald, J. W. ve Ghosh, D. (2006). COPA—Cancer Outlier Profile Analysis. *Bioinformatics*, 22(23):2950–2951.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., ve Gilad, Y. (2008). RNA-Seq: An Assessment of Technical Reproducibility And Comparison With Gene Expression Arrays. *Genome Research*, 18(9):1509–1517.

Martínez-Camblor, P., Corral, N., Rey, C., Pascual, J., ve Cernuda-Morollón, E. (2017). Receiver Operating Characteristic Curve Generalization for Non-Monotone Relationships. *Statistical Methods in Medical Research*, 26(1):113–123.

Moody, L., Xu, G. B., Pan, Y.-X., ve Chen, H. (2022). Genome-Wide Cross-Cancer Analysis Illustrates The Critical Role of Bimodal miRNA in Patient Survival and Drug Responses To PI3K Inhibitors. *PLOS Computational Biology*, 18(5):e1010109.

Parodi, S., Pistoia, V., ve Muselli, M. (2008). Not Proper ROC Curves As New Tool for The Analysis of Differentially Expressed Genes in Microarray Experiments. BMC *Bioinformatics*, 9(1):410.

Robinson, M. D., McCarthy, D. J., ve Smyth, G. K. (2010). edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics*, 26(1):139–140.

Robinson, M. D. ve Oshlack, A. (2010). A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data. *Genome Biology*, 11(3):R25.

Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E., ve Giordano, A. (2024). Differential Gene Expression Analysis Pipelines And Bioinformatic Tools for The Identification of Specific Biomarkers: A Review. Computational and Structural *Biotechnology Journal*, 23:1154–1168.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., ve Barton, G. J. (2016). How Many Biological Replicates Are Needed in An RNA-Seq Experiment and Which Differential Expression Tool Should You Use? *RNA*, 22(6):839–851.

Silva-Fortes, C., Amaral Turkman, M. A., ve Sousa, L. (2012). Arrow Plot: A New Graphical Tool For Selecting Up And Down Regulated Genes And Genes Differentially Expressed On Sample Subgroups. *BMC Bioinformatics*, 13(1):147.

Soneson, C. ve Delorenzi, M. (2013). A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data. *BMC Bioinformatics*, 14(1):91.

Tibshirani, R. ve Hastie, T. (2007). Outlier Sums for Differential Gene Expression Analysis. *Biostatistics*, 8(1):2–8.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A., ve Chinnaiyan, A. M. (2005). Recurrent Fusion Of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. *Science*, 310(5748):644–648.

Tong, P., Chen, Y., Su, X., ve Coombes, K. R. (2013). SIBER: Systematic Identification of Bimodally Expressed Genes Using RNAseq Data. *Bioinformatics*, 29(5):605–613.

Wang, J., Wen, S., Symmans, W. F., Pusztai, L., ve Coombes, K. R. (2009a). The Bimodality Index: A Criterion For Discovering and Ranking Bimodal Signatures From Cancer Gene Expression Profiling Data. *Cancer Informatics*, 7:199–219.

Wang, Z., Gerstein, M., ve Snyder, M. (2009b). RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.