Bölüm 3

TRIPOD+AI ve TRIPOD-LLM Rehberlerine Uyum: Klinik Yapay Zeka Modelleri Nasıl Raporlanmalı?

Buğra Varol[1]

Özet

Bu bölüm, klinik yapay zeka (Artificial Intelligence, AI) ve makine öğrenimi (Machine Learning, ML) temelli tahmin modellerinin şeffaf raporlanmasına yönelik Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Artificial Intelligence (TRIPOD+AI, 2024) ile Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Models (TRIPOD-LLM, 2025) kılavuzlarını sistematik biçimde değerlendirmektedir. Öncelikle TRIPOD'ın klasik sürümünden bu yana metodolojik gereksinimlerin nasıl evrildiğini ortaya koymaktadır. İkinci kısımda, TRIPOD+AI'nin on dört ek maddesi ve TRIPOD-LLM'nin büyük dil modelleri özel ilkeleri üzerinden veri yönetimi, model geliştirme-doğrulama, yorumlanabilirlik, etik ve yanlılık analizleri için ayrıntılı kontrol listeleri sunmaktadır. Eksik veri yönetimi, veri sızıntısı, tekil metrik kullanımı ve haricî doğrulama eksikliği gibi yaygın hataları örnekleyerek bunlara yönelik önleme stratejilerini tartışmaktadır. Son bölüm, biyoistatistikçilerin kuramsal rehberlik rolünü vurgulamakta; federatif öğrenme, gizlilik-korumalı AI ve adillik-odaklı değerlendirmeler ışığında geleceğe yönelik araştırma gündemini çizmektedir.

1. Giriş

Klinik tıp, yapay zeka (artificial intelligence, AI) ve makine öğrenimi (machine learning, ML) teknolojilerinin entegrasyonuyla hızlı bir dönüşüm yaşamaktadır. AI uygulamaları, görüntü analizi, tanı süreçleri, kişiselleştirilmiş tedavi planları ve prognostik tahminler gibi birçok klinik alanda sağlık hizmetlerinin kalitesini yükseltme ve tanısal doğruluğu artırma

1 Dr., Adnan Menderes Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı, Aydın/Türkiye, bugravarol87@gmail.com, ORCID ID: 0000-0001-8052-7782

https://doi.org/10.58830/ozgur.pub780.c3256

potansiyeline sahiptir. Özellikle klinik tahmin modelleri, klinisyenlerin daha bilinçli kararlar almalarına ve kaynakların etkin kullanımına destek sağlamaktadır (Topol, 2019; Esteva ve ark., 2019).

Ancak AI tabanlı klinik tahmin modellerinin geliştirme ve doğrulama süreçlerinin şeffaf ve sağlam metodolojilere dayanması kritik öneme sahiptir. Bu modellerin raporlanmasında görülen eksiklikler ve tutarsızlıklar, modellerin bağımsız değerlendirilmesini, tekrarlanabilirliğini ve klinik uygulamaya geçişini zorlaştırmaktadır (Altman ve ark., 2012). Bu soruna yönelik olarak geliştirilen TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) rehberi, tahmin modellerinin şeffaf raporlanması için standartlar sunmaktadır (Collins ve ark., 2015; Moons ve ark., 2015). Ancak geleneksel istatistiksel yaklaşımların yanı sıra AI ve ML yöntemlerinin yaygınlaşması, orijinal TRIPOD rehberinin yetersiz kalmasına neden olmuştur. Bu nedenle, 2024 yılında yayımlanan TRIPOD+AI ve 2025 yılında geliştirilen TRIPOD-LLM rehberleri, genel AI/ML ve büyük dil modelleri (Large Language Models; LLM) için raporlama standartlarını güncelleyerek, AI modellerinin "kara kutu" yapısını aşmayı ve klinik entegrasyonu kolaylaştırmayı hedeflemektedir (Collins ve ark., 2024; Gallifant ve ark., 2025).

Bu bölümde, TRIPOD+AI ve TRIPOD-LLM rehberlerinin temel prensipleri, yapıları ve klinik ML çalışmalarına entegrasyonu incelenecek; ayrıca sıkça karşılaşılan raporlama hataları ve bu hatalardan kaçınma stratejileri tartışılarak klinik AI modellerinin güvenilirliğini ve şeffaflığını artırmaya yönelik pratik bilgiler sunulacaktır.

2. TRIPOD Kılavuzunun Evrimi: Temelden Yapay Zekaya

Klinik tahmin modelleri, hastaların sağlık durumları hakkında öngörülerde bulunarak klinik karar süreçlerine değerli katkılar sağlar. Bu modellerin geliştirme ve doğrulama süreçleri karmaşık istatistiksel yöntemler gerektirir ve sonuçların güvenilirliği kullanılan metodolojinin kalitesiyle yakından ilişkilidir. Ancak bu alandaki araştırmaların raporlama kalitesi sıklıkla eleştirilmekte; modellerin nasıl oluşturulduğu, test edildiği ve uygulanabilirliği konusunda yeterli bilgi verilmemesi nedeniyle modellerin bağımsız değerlendirilmesi ve klinik uygulamalara aktarılması güçleşmektedir (Altman ve ark., 2012).

Raporlamadaki bu eksiklikleri gidermek üzere, Collins ve ark. (2015) tarafından TRIPOD rehberi geliştirilmiştir. TRIPOD, çok değişkenli tahmin modelleri üzerine yapılan çalışmaların şeffaf raporlanmasını sağlamak için 22 maddelik bir kontrol listesi sunmaktadır. Bu liste, çalışma tasarımı, veri

kaynakları, model geliştirme stratejileri, doğrulama yöntemleri ve model sınırlılıkları gibi kritik detayların eksiksiz sunulmasını hedeflemektedir. TRIPOD rehberinin temel amacı, tahmin modellerinin değerlendirilmesini kolaylaştırmak, tekrarlanabilirliğini artırmak, sistematik derleme ve meta-analizlerin kalitesini yükseltmek ve klinik uygulamalara entegrasyonlarını hızlandırmaktır (Moons ve ark., 2015; Heus ve ark., 2020)

Ancak ML ve AI yöntemlerinin klinik uygulamalarda artan kullanımı, TRIPOD rehberinin bu yeni metodolojilerin özel ihtiyaçlarını tam olarak karşılayamadığını göstermiştir. ML modelleri genellikle daha fazla belirteç kullanmakta, doğrusal olmayan ilişkileri modellemekte ve veri ön işleme ile hiperparametre optimizasyonu gibi ek adımlar içermektedir. Bu nedenle, AI yöntemlerine özgü ihtiyaçları karşılamak üzere TRIPOD+AI ve TRIPOD-LLM gibi rehberlerin geliştirilmesi gerekli hale gelmiştir (Collins ve ark., 2024; Gallifant ve ark., 2025).

3. TRIPOD+AI Rehberi: Yapay Zeka Tabanlı Tahmin Modelleri İçin Raporlama Standardı

AI ve ML algoritmalarının klinik tahmin modellerinde kullanımı, geleneksel istatistiksel yöntemlere kıyasla bazı benzersiz zorluklar ve dikkat edilmesi gereken noktalar içerir. Bu modeller genellikle yüksek boyutlu verilerle çalışır, karmaşık örüntüleri öğrenebilir ve bazen yorumlanması zor "kara kutu" niteliğinde olabilirler (Stiglic ve ark., 2020; Lambert ve ark., 2024). Bu özellikler, model geliştirme ve doğrulama süreçlerinin şeffaf bir şekilde raporlanmasını daha da kritik hale getirmektedir. Standart TRIPOD bildirgesinin bu yeni nesil modellere özgü bazı detayları tam olarak kapsamadığı anlaşıldığında, Collins ve ark. (2024) tarafından TRIPOD+AI rehberi geliştirilmiştir. 2024 yılında BMJ'de yayımlanan bu rehber, orijinal TRIPOD çerçevesini temel alarak AI/ML modellerine özgü ek raporlama maddeleri sunar.

TRIPOD+AI, orijinal TRIPOD'daki 22 ana maddeye ek olarak AI/ML ile ilgili 14 yeni madde (ve alt madde) içermektedir. Bu eklemeler, çalışmanın her bölümünde – başlık, özet, giriş, yöntemler, sonuçlar, tartışma ve diğer bilgiler kısımlarında – AI/ML'ye özgü detayların nasıl raporlanacağına dair spesifik yönlendirmeler sağlar (Collins ve ark., 2024).

TRIPOD+AI rehberinin getirdiği temel yenilikler ve vurguladığı noktalar birkaç ana başlık altında toplanabilir.

3.1. Veri Yönetimi, Model Geliştirme ve Algoritmik Seçim

Öncelikle, kullanılan veri kümelerinin – eğitim, iç doğrulama (hiperparametre ayarı veya çapraz doğrulama), test ve varsa harici doğrulama – kökeni, nasıl toplandığı, model geliştirme için uygunluğu ve potansiyel yanlılık kaynakları ayrıntılı biçimde açıklanmalıdır. Verinin temsil ediciliği ve hedeflenen popülasyona uygunluğu tartışılmalıdır. Uygulanan veri ön işleme adımları (örneğin, normalizasyon veya standardizasyon gibi ölçeklendirme işlemleri, eksik veri atama yöntemleri ve seçilme gerekçeleri, veri artırma teknikleri) net olarak belirtilmelidir (Karrar, 2022). Bu adımların model performansına etkisi ve seçilen yöntemlerin gerekçeleri sunulmalıdır. Ayrıca, kullanılan değişkenlerin nasıl seçildiği veya türetildiği; bu aşamada yararlanılan algoritmalar – örneğin Mutlak Sapma Daraltma ve Seçim Operatörü (Least Absolute Shrinkage and Selection Operator, LASSO) ya da rastgele orman temelli değişken önem sıralaması – ile seçim kriterleri ayrıntılı biçimde tanımlanmalıdır. (Ding ve ark., 2025). Model geliştirme aşamasında kullanılan spesifik AI/ML algoritması (örneğin, destek vektör makineleri, rastgele ormanlar, derin öğrenme ağları) ve bu algoritmanın seçilme gerekçesi açıklanmalı; modelin mimarisi (örneğin, derin öğrenme modelinde katman sayısı, nöron sayısı, aktivasyon fonksiyonları) ayrıntılarıyla sunulmalıdır. Model geliştirme ve analiz için kullanılan yazılımlar, programlama dilleri (Python, R vb.) ve kütüphaneler (örneğin, Scikit-learn, TensorFlow, PyTorch gibi) sürüm numaralarıyla birlikte belirtilerek çalışmanın tekrarlanabilirliği artırılmalıdır. Son olarak, model performansını optimize etmek için ayarlanan hiperparametreler ve bu ayarlama sürecinde kullanılan yöntemler (örneğin, grid search, random search veya Bayesian optimizasyon) ile değerlendirme yaklaşımları (örneğin, çapraz geçerlilik sonuçları) açıklanmalıdır (Bischl ve ark., 2023; Ilemobayo ve ark., 2024).

3.2. Model Performansının Değerlendirilmesi, Yorumlanabilirlik ve Tekrarlanabilirlik

Model performansını değerlendirirken kullanılan metrikler ve bu metriklerin seçilme gerekçeleri açıklanmalıdır. Özellikle dengesiz veri kümelerinde hangi performans metriklerinin daha uygun olduğu tartışılmalı, model tahminlerinin gerçek sonuçlarla uyumu (kalibrasyon) da değerlendirilmelidir. Modelin nihai performansının, model geliştirme aşamasında hiç görmediği ayrı bir test veri kümesinde değerlendirildiği net bir şekilde belirtilmeli ve olası veri sızıntısını (data leakage) önlemek için alınan önlemler vurgulanmalıdır. Geliştirilen modelin performansı, mümkünse mevcut klinik standartlar veya daha basit modellerle kıyaslanmalıdır (Saito ve Rehmsmeier, 2015). Bazı AI modelleri tahminleriyle birlikte bir

belirsizlik ölçüsü de sunabilir; eğer model bu yeteneğe sahipse, bunun nasıl hesaplandığı ve raporlandığı açıklanmalıdır. Modelin yorumlanabilirliği açısından, eğer model doğası gereği bir "kara kutu" niteliğindeyse (örneğin, kompleks bir derin öğrenme modeli), modelin tahminlerini açıklamak için kullanılan yöntemler (örneğin, SHAP – SHapley Additive exPlanations – veya LIME – Local Interpretable Model-agnostic Explanations – gibi açıklayıcı algoritmalar) ve bu yöntemlerin bulguları sunulmalı, hangi özelliklerin model tahminlerini ne şekilde etkilediği gösterilmelidir (Salih ve ark., 2025). Model çıktılarının klinik karar verme sürecine entegrasyonu ve klinisyenler tarafından anlaşılabilmesi de tartışılmalıdır. Tekrarlanabilirliği sağlamak adına, model geliştirme kodunun, kullanılan veri kümelerinin (etik ve gizlilik kısıtları çerçevesinde) ve eğitilmiş modelin kendisinin (mümkünse) erişilebilir hale getirilmesi teşvik edilmelidir. Bu amaçla kullanılan platformlar (örneğin, GitHub gibi) belirtilerek araştırmanın tekrarlanabilirliği kolaylaştırılmalıdır (Edin ve Johansson, 2023).

3.3. Etik Hususlar ve Yanlılık Değerlendirmesi

Etik gereklilikler bağlamında, çalışma için alınan etik kurul onayları ve hasta verilerinin gizliliğini korumak için alınan önlemler belirtilmelidir. Yanlılık (bias) değerlendirmesi ise veri toplama, model geliştirme veya model değerlendirme aşamalarında ortaya çıkabilecek potansiyel yanlılıkların (örneğin, seçim yanlılığı, ölçüm yanlılığı, algoritmik yanlılık) tanımlanmasını ve bu yanlılıkları azaltmak için alınan önlemlerin tartışılmasını içerir. Modelin farklı alt gruplardaki (örneğin cinsiyet veya etnik köken gibi) performansı da değerlendirilerek olası adaletsizlikler raporlanmalıdır (Char ve ark., 2018; Obermeyer ve ark., 2019).

4. TRIPOD+AI'nin Makalenin Bölümlerine Entegrasyonu

TRIPOD+AI kontrol listesi, bir AI/ML tahmin modeli çalışmasının raporlanmasında makale bölümlerine göre spesifik öneriler sunarak aşağıdaki düzende bütüncül bir çerçeve oluşturur (Bathula ve ark., 2024; Collins ve ark., 2024):

Başlık ve özet, çalışmanın bir AI tahmin modeli içerdiğini, kullanılan yaklaşımı ve temel bulguları vurgulamalıdır. ***Giriş bölümü*** ise modelin hedeflediği klinik problemi, yöntemin seçilme gerekçesini ve gerekli görülürse etik uyumla ilgili bilgileri açıklamalıdır. ***Yöntemler bölümü***, model geliştirme ve doğrulama ile ilgili tüm kritik detayları (veri kaynakları, bireyler, değişken tanımları, örneklem büyüklüğü, eksik veri yönetimi, model geliştirme ve test süreçleri, performans değerlendirme, yorumlanabilirlik, yazılım kullanımı) içermelidir. ***Sonuçlar bölümünde*** çalışma kapsamındaki birey akışının, temel

demografik ve klinik özelliklerin, modelin test veri kümesindeki performans ölçütlerinin ilgili güven aralıklarıyla birlikte ve uygulanmışsa alt grup analizlerinin ayrıntılı biçimde raporlanmalıdır. ***Tartışma bölümü***, bulguların özetini; modelin sınırlılıklarını (yanlılıklar dahil), sonuçların yorumunu ve genellenebilirliğini; modelin olası klinik etkilerini ve gelecek araştırma yönelimlerini kapsamalıdır. Modelin klinik iş akışına entegrasyonu, etik çıkarımlar ve sorumlu kullanım tartışılmalıdır. ***Diğer Bilgiler*** kısmında ise çalışma fon kaynakları, çıkar çatışmaları, etik onay ayrıntıları ve AI/ML modelleri için kod ile veri paylaşım durumu belirtilmelidir.

TRIPOD+AI rehberinin benimsenmesi, AI tabanlı klinik tahmin modelleri alanındaki araştırmaların kalitesini, şeffaflığını ve tekrarlanabilirliğini artırarak bu güçlü araçların klinik pratiğe güvenli ve etkili bir şekilde entegrasyonunu hızlandıracaktır.

5. TRIPOD-LLM Rehberi: Büyük Dil Modelleriyle Geliştirilen Tahmin Modellerinin Raporlanması

LLM'ler, yapılandırılmamış metin verilerinden karmaşık örüntüleri öğrenme ve bu örüntülere dayanarak tahminler yapabilme yetenekleri sayesinde klinik tahmin modelleri geliştirmede giderek önem kazanmaktadır. Klinik uygulamalarda LLM'ler, elektronik sağlık kayıtlarındaki doktor notları gibi serbest metinleri analiz ederek tanı veya risk tahminlerinde kullanılmaya başlanmıştır (Thirunavukarasu ve ark., 2023). LLM'lerin bu esnekliği onları çok güçlü kılmakla birlikte, kendilerine özgü riskleri ve belirsizlikleri de beraberinde getirmektedir (örneğin, yanlış veya uydurma bilgi üretme eğilimleri olan "halüsinasyonlar", eğitim verilerine gömülü önyargılar ve benzeri riskler). Standart TRIPOD ve TRIPOD+AI rehberleri, LLM tabanlı modellerin bu kendine has yönlerini tam olarak karşılayamadığı için, Gallifant ve ark. (2025) TRIPOD-LLM rehberini geliştirmişlerdir. TRIPOD-LLM, TRIPOD+AI'nin üzerine inşa edilmiş olup LLM'lerle geliştirilen modellerin raporlanmasında ek standartlar getirmektedir. Özellikle aşağıdaki alanlara vurgu yapar.

5.1. LLM Modelinin Özellikleri ve Eğitim Süreci

Kullanılan temel LLM modelinin adı, boyutu (parametre sayısı) ve varsa versiyonu ile sağlayıcısı raporlanmalıdır. Modelin ön eğitim aşamasında hangi veri kümeleri üzerinde çalıştığı ve bu verilerin kapsamı açıklanmalıdır. Geliştirilen tahmin görevi için modelin nasıl ince ayar (fine-tuning) yapıldığı; kullanılan kurum verilerinin tipi, dönem aralığı ve etik onay durumu belirtilmelidir. Girdi metinlerinin formatı ve modelin ürettiği çıktının formatı tanımlanmalı; model tahminlerinin belirli "prompt"larla

yönlendirildiği durumlarda kullanılan prompt örnekleri ve tasarım tercihleri açıklanmalıdır (Vrdoljak ve ark., 2025; Wang ve ark., 2023).

5.2. Performans Değerlendirmesi ve Güvenlik Önlemleri

LLM tabanlı modelin başarısını değerlendirirken klasik tahmin performans metriklerinin yanı sıra metin üretim kalitesini ve doğruluğunu ölçen kriterler de kullanılabilir. Modelin yanlış veya uydurma bilgi üretme (halüsinasyon) eğilimi test edilmeli ve en aza indirmek için alınan önlemler raporlanmalıdır. Farklı ifade biçimleri veya farklı hasta alt grupları karşısında modelin ne kadar sağlam olduğu incelenmelidir. Eğer model gerçek hasta verileriyle eğitildiyse, kişisel sağlık bilgilerinin sızmaması için anonimleştirme ya da benzeri gizlilik koruma yöntemlerinin uygulandığı ve modelin eğitim verilerindeki hassas bilgileri ezberlemediği/sızdırmadığı değerlendirilmelidir. Ayrıca, modelin farklı demografik veya klinik gruplar arasında adil performans gösterip göstermediği analiz edilmelidir (Asgari ve ark., 2025; Shool ve ark., 2025).

5.3. Kullanım Ortamı ve Sınırlılıklar

LLM tabanlı modelin mevcut klinik iş akışlarına nasıl entegre edilebileceği ve kullanımı için ek altyapı veya uzman eğitimi gerekip gerekmediği tartışılmalıdır. Modelin zamanla güncelliğini yitirmemesi için bir güncelleme planı ve gerçek dünyadaki performansının izlenmesine dair stratejiler sunulmalıdır. Ayrıca, modelin eğitimi ve çalıştırılması için gereken hesaplama kaynakları (örneğin, güçlü GPU ihtiyacı) belirtilerek bu gereksinimlerin uygulamaya etkisi değerlendirilmelidir. Son olarak, LLM tabanlı yaklaşıma özgü olabilecek etik veya yasal kısıtlar şeffaflıkla ele alınmalıdır (Dennstädt ve ark., 2025).

6. TRIPOD+AI ve TRIPOD-LLM Rehberlerinin Araştırma Sürecine Entegrasyonu

TRIPOD+AI ve TRIPOD-LLM rehberlerinin klinik ML araştırmalarına entegrasyonu, çalışmaların kalitesini artırmak, bulguların güvenilirliğini pekiştirmek ve sonuçların klinik pratiğe daha hızlı ve güvenli bir şekilde aktarılmasını sağlamak için hayati bir adımdır. Bu entegrasyon, araştırma sürecinin en başından sonuna kadar bilinçli bir çaba gerektirir (Collins ve ark., 2024; Gallifant ve ark., 2025).

6.1. Planlama, veri toplama ve ön işleme

Araştırmacılar, çalışma sorusunu formüle ederken ve protokolü oluştururken ilgili TRIPOD+AI veya TRIPOD-LLM kontrol listesini dikkatlice incelemelidir. Bu sayede en baştan hangi bilgilerin toplanacağı

ve raporlanacağı konusunda net bir yol haritası belirlenir. Modelin hedef popülasyonu ve kullanılacak veri kaynakları (TRIPOD+AI Madde 5a-AI; TRIPOD-LLM rehberine göre temel LLM ve ince ayar verileri) açıkça tanımlanmalı; veri toplama süreçleri, olası yanlılıklar ve etik hususlar planlama aşamasında göz önünde bulundurulmalıdır. Tahmin edilecek sonuç (outcome) ve olası açıklayıcı değişkenler net bir şekilde tanımlanmalı, nasıl ölçülüp kodlandıkları belirtilmelidir (TRIPOD Madde 6, 7). Özellikle AI modelleri için yeterli veri büyüklüğü sağlamak kritik olduğundan, mümkünse örneklem büyüklüğü ya da güç analizi yapılarak gerekçelendirilmelidir (TRIPOD Madde 8). Kullanılacak AI/ML algoritması veya temel LLM (TRIPOD-LLM) ile model geliştirme stratejisi (eğitim, model optimizasyonu ve test aşamalarının planı) en başta belirlenmelidir (TRIPOD+AI Madde 10e-AI; TRIPOD-LLM). Veri sızıntısını önlemek için uygun doğrulama şemaları (örneğin, zaman temelli veya coğrafi ayırma) düşünülmelidir. Veri toplama ve ön işleme aşamasında da, verilerin toplanması, temizlenmesi ve ön işlenmesi adımları titizlikle belgelenmelidir. Eksik verilerin nasıl ele alındığı, uygulanan özellik mühendisliği adımları ve bunların gerekçeleri kayıt altına alınmalıdır. Veri işleme ve analiz için kullanılan kodlar, bir sürüm kontrol sistemi (örneğin, Git) ile yönetilmeli ve iyi dokümante edilerek tekrarlanabilirlik artırılmalıdır (Collins ve ark., 2024; Gallifant ve ark., 2025).

6.2. Model geliştirme, doğrulama ve raporlama

Model geliştirme ve doğrulama aşamalarında, modelin eğitim süreci, hiperparametre ayarları (TRIPOD+AI Madde 10f-AI) ve kullanılan yazılım/kütüphaneler (TRIPOD+AI Madde 10e-AI) ayrıntılı olarak kaydedilmelidir (Collins ve ark., 2024). LLM'ler için *prompt* tasarımı ve ince ayar prosedürleri (TRIPOD-LLM rehberi) benzer şekilde belgelendirilmelidir (Gallifant ve ark., 2025). Model performansı, uygun metrikler kullanılarak hem iç (örneğin çapraz doğrulama) hem de dış (bağımsız bir test veri kümesi) doğrulamalarla değerlendirilmelidir (TRIPOD+AI Madde 10g-AI). Kalibrasyon, ayrımcılık gücü ve mümkünse klinik yarar analizleri yapılmalıdır (Van Calster ve ark., 2019). Özellikle karmaşık modellerde tahminlerin nasıl oluşturulduğunu anlamak için yorumlanabilirlik yöntemleri uygulanmalı ve sonuçları not edilmelidir (TRIPOD+AI Madde 10h-AI; TRIPOD-LLM rehberi). Modelin farklı alt gruplardaki performansı incelenmeli, potansiyel yanlılıklar değerlendirilmelidir (TRIPOD+AI Madde 21a-AI; TRIPOD-LLM). Araştırma sonuçları raporlanırken TRIPOD+AI veya TRIPOD-LLM kontrol listesi bir rehber olarak kullanılmalı ve her bir maddenin makalenin ilgili bölümünde ele alındığından emin olunmalıdır; nitekim birçok dergi makale ile birlikte bu kontrol listesinin sunulmasını talep etmektedir.

Yöntemler ve sonuçlar bölümleri, başka bir araştırmacının çalışmayı tekrarlayabileceği veya bulguları bağımsız olarak değerlendirebileceği kadar açık ve detaylı yazılmalıdır. Mümkün olduğunca ve etik kurallar izin verdiği ölçüde, analiz kodlarının, modelin ve anonimleştirilmiş verilerin paylaşılması teşvik edilmelidir (TRIPOD+AI Madde 22a-AI; TRIPOD-LLM). Bu şeffaflık, bilimin ilerlemesini hızlandıracaktır. Son olarak, çalışmanın ve modelin sınırlılıkları dürüstçe tartışılmalı; potansiyel yanlılık kaynakları ve bunların sonuçlar üzerindeki etkileri ele alınmalıdır (TRIPOD Madde 18) (Collins ve ark., 2024; Gallifant ve ark., 2025).

6.3. Entegrasyonda Karşılaşılan Zorluklar ve Çözüm Önerileri

TRIPOD+AI ve TRIPOD-LLM rehberlerinin entegrasyonu sürecinde bazı zorluklarla karşılaşılabilir. Bu zorlukların farkında olunması ve proaktif çözüm stratejileri geliştirilmesi önemlidir (Collins ve ark., 2024; Gallifant ve ark., 2025):

- ***Farkındalık Eksikliği:*** Araştırmacıların bu rehberlerin varlığından veya öneminden haberdar olmaması, rehberlerin kullanılmamasına yol açabilir. Bu sorunun aşılması için eğitim programları, çalıştaylar ve dergi politikaları aracılığıyla rehberlerin tanıtılması ve farkındalığın artırılması sağlanmalıdır.
- ***Ek Yük Algısı:*** Detaylı raporlama, araştırmacılar tarafından ek zaman ve çaba gerektiren bir yük olarak algılanabilir. Oysa iyi raporlama, araştırmanın kalitesini ve etkisini artıran bir yatırımdır (Altman ve ark., 2012). Bu nedenle, protokol aşamasından itibaren planlama yapılarak gereksinim duyulacak bilgilerin önceden belirlenmesi, raporlama yükünü azaltmaya yardımcı olacaktır.
- ***Hızla Gelişen AI Alanı:*** AI/ML tekniklerinin sürekli gelişmesi ve yeni algoritmaların ortaya çıkmasıyla rehberlerin güncel kalması zorlaşabilir. Bu durum, rehberlerin periyodik olarak gözden geçirilip güncellenmesini gerektirir. Ayrıca araştırmacıların temel raporlama prensiplerine odaklanarak ortaya çıkan yeni yöntemleri de şeffaf bir biçimde raporlaması önemlidir.
- ***"Kara Kutu" Modellerin Karmaşıklığı:*** Bazı ileri AI modellerinin iç işleyişini tam olarak açıklamak güç olabilir (Marey ve ark., 2024). TRIPOD+AI ve TRIPOD-LLM rehberleri, yorumlanabilirlik yöntemlerinin kullanımını ve raporlanmasını teşvik ederek bu zorluğa çözüm sunmaktadır. Araştırmacıların model davranışını anlamak için çaba göstermesi ve modellerinin sınırlarını açıkça belirtmesi gerektiği vurgulanmaktadır.

- ***Veri ve Kod Paylaşımındaki Engeller:*** Gizlilik endişeleri, kurumsal politikalar veya fikri mülkiyet kaygıları nedeniyle veri ve kod paylaşımı kısıtlanabilir. Bu engelleri aşmak için veri anonimleştirme teknikleri, güvenli veri havuzları ve açık bilim ilkeleri teşvik edilmelidir. Veri paylaşımının bilimsel iş birliği ve ilerleme için faydaları vurgulanmalıdır.

Rehberlere uyum sağlamak başlangıçta ek bir çaba gibi görünse de uzun vadede klinik ML alanının bilimsel titizliğini ve güvenilirliğini artıracak, dolayısıyla hasta bakımına olan olumlu etkileri güçlendirecektir. Bu süreçte biyoistatistikçiler metodolojik danışmanlık vererek ve raporlama standartlarının uygulanmasına öncülük ederek kilit bir rol oynayabilir.

7. Raporlamada Sık Karşılaşılan Hatalar ve Çözümler

TRIPOD+AI ve TRIPOD-LLM gibi rehberlere rağmen, klinik AI ve ML modeli çalışmalarının raporlanmasında hala sıkça görülen hatalar vardır. Bu hatalar, çalışmaların sağlıklı değerlendirilmesini, tekrarlanmasını ve güvenle klinik uygulamaya aktarılmasını zorlaştırmaktadır. Aşağıda yaygın raporlama hataları ve rehberler ışığında bunlardan kaçınma yolları sunulmuştur (Collins ve ark., 2024; Gallifant ve ark., 2025):

7.1. Veri Kaynakları ve İşleme ile İlgili Hatalar

- ***Yetersiz veri kümesi tanımı:*** Eğitim, doğrulama ve test için kullanılan veri kümelerinin kaynağı, toplanma yöntemleri, dahil etme/dışlama kriterleri, zaman aralığı ile temel demografik ve klinik özellikleri yeterince tanımlanmamış olabilir. Çözüm olarak, her bir veri kümesi için ayrıntılı "veri kartı" benzeri tanımlamalar yapılmalı; veri setinin temsil ettiği popülasyon ve olası yanlılıkları tartışılmalıdır. TRIPOD ve TRIPOD+AI rehberlerindeki ilgili maddeler titizlikle uygulanmalı, LLM çalışmalarında TRIPOD-LLM rehberinin önerileri de dikkate alınmalıdır.
- ***Eksik veri yönetiminin belirsizliği:*** Hangi değişkenlerde ne kadar eksik veri olduğu ve eksik verilerin nasıl ele alındığı (örneğin, veri setinden çıkarma veya tekli/çoklu değer atama) hakkında bilgi verilmemesi yaygın bir eksikliktir. Çözüm için eksik verilerin hangi oranda ve ne şekilde dağıldığı şeffaflıkla raporlanmalı; kullanılan eksik veri değer atama yöntemleri, bu yöntemlerin varsayımları ve seçilme gerekçeleri ayrıntılı açıklanmalıdır. Farklı eksik veri yöntemlerinin sonuçlara etkisini değerlendirmek üzere duyarlılık analizleri yapılması da önerilir.

- ***Veri ön işleme adımlarının yetersiz açıklanması:*** Normalizasyon, kategorik değişken kodlaması, veri artırma (augmentation) gibi ön işleme adımlarının atlanması veya yüzeysel geçilmesi sık rastlanan bir hatadır. Çözüm olarak, uygulanan tüm ön işleme adımları kullanılan spesifik yöntem ve parametreleriyle (örneğin, normalizasyonda min-max ölçekleme mi yoksa z-skoru mu kullanıldığı) birlikte açıkça belirtilmeli ve bu adımların model performansına etkisi tartışılmalıdır.
- ***Değişken seçimi veya mühendisliğinin şeffaf olmaması*:** Modele dahil edilen son değişkenlerin nasıl seçildiği veya türetildiği belirsiz bırakılabilir; hangi değişken seçme algoritmalarının, eşik değerlerin veya uzman görüşüne dayalı kriterlerin kullanıldığı açıklanmayabilir. Çözüm olarak, değişken seçimi ya da değişken türetme süreçlerinde kullanılan tüm yöntemler (filtre, sarmalayıcı veya gömülü yöntemler), uygulanan istatistiksel kriterler ya da uzman görüşü yaklaşımları detaylı biçimde açıklanmalıdır. Seçilmeyen değişken bilinçli olarak çıkarıldıysa bunlar ve nedenleri de belirtilmelidir.

7.2. Model Geliştirme ve Algoritma ile İlgili Hatalar

- ***Model veya algoritmanın yetersiz tanımlanması:*** Kullanılan AI/ML algoritmasının veya LLM'in tam olarak belirtilmemesi; model mimarisi, hiperparametreler ve model doğrulama süreçleri hakkında detay verilmemesi sık görülen bir eksikliktir. Örneğin yalnızca "derin öğrenme modeli kullanıldı" demek, hangi mimarinin kaç katmanlı olduğu veya nasıl eğitildiği konusunda bilgi vermez. Çözüm olarak algoritmanın tam adı, versiyonu ve kullanılan yazılım kütüphaneleri (sürüm numaralarıyla) belirtilmelidir. Derin öğrenme modelleri için katman sayısı, her katmandaki nöron sayısı, aktivasyon fonksiyonları gibi mimari ayrıntılar sunulmalıdır. Hiperparametre ayarlama süreci (kullanılan arama stratejisi, değerlendirme metrikleri, çapraz doğrulama şeması) şeffafça raporlanmalıdır. LLM kullanıldıysa *prompt* tasarımı ve ince ayar prosedürleri de TRIPOD-LLM rehberine uygun biçimde açıklanmalıdır (Collins ve ark., 2024; Gallifant ve ark., 2025).
- ***Veri sızıntısı (data leakage):*** Test verisinden veya geleceğe ait bilgilerden gelen verilerin modelin eğitim/optimizasyon sürecine sızması, model performansının gerçekte olduğundan daha yüksek görünmesine yol açar. Örneğin tüm veri üzerinde normalizasyon yapılması veya değişken seçiminin tüm veri kullanılarak yapılması bu tür bir sızıntıya neden olabilir (Xiao ve ark., 2025). Çözüm, veri bölünmesinin (eğitim, doğrulama, test) en başta yapılması ve test verisinin model geliştirme ile doğrulama süreçlerinden tamamen izole

tutulmasıdır. Çapraz doğrulama kullanılacaksa, her katlamada veri ön işleme ve değişken seçiminin sadece o katlamanın eğitim verisi üzerinde uygulandığına dikkat edilmelidir. Bu önlemlerin alındığı açıkça raporlanmalıdır (Gallifant ve ark., 2025).

7.3. Model Performansı ve Değerlendirmesi ile İlgili Hatalar

- ***Uygun olmayan veya sınırlı metrik kullanımı:*** Sadece AUC-ROC (Area Under the Receiver Operating Characteristic Curve, AUC-ROC) gibi tek bir metrik raporlamak, özellikle dengesiz veri setlerinde yanıltıcı olabilir. Kalibrasyon analizinin yapılmaması da önemli bir eksikliktir (Van Calster ve ark., 2019). Bu nedenle çalışmanın amacına ve veri yapısına uygun olarak AUC-ROC, Area Under the Precision-Recall Curve (AUC-PR), duyarlılık, özgüllük, pozitif/negatif tahmin değeri, F1 skoru, Brier skoru gibi birden fazla performans ölçütü güven aralıkları ile birlikte raporlanmalıdır. Ayrıca model tahminlerinin gerçek sonuçlarla uyumunu gösteren kalibrasyon eğrileri ve ilgili istatistikler sunulmalı; metrik seçiminin gerekçesi açıklanmalıdır (Collins ve ark., 2015; Collins ve ark., 2024).
- ***Yalnızca iç doğrulama yapılması (dış doğrulama eksikliği):*** Model performansının sadece eğitim verisi veya aynı kaynaktan gelen bir test alt kümesi ile değerlendirilmesi, modelin farklı popülasyon veya ortamlardaki genel performansı hakkında bilgi vermez (Steyerberg ve ark., 2013). Çözüm olarak, mümkünse model tamamen bağımsız bir dış doğrulama kohortunda test edilmelidir. Bu mümkün değilse bu eksiklik sınırlılıklar bölümünde açıkça belirtilmeli ve gelecekte harici doğrulama yapılması gereği vurgulanmalıdır. Alternatif olarak, modelin güvenirliğini değerlendirmek için *bootstrap* veya tekrarlı çapraz doğrulama gibi yöntemler uygulanabilir (Collins ve ark., 2015).
- ***Modelin paylaşılmaması veya yeniden kullanım için detay sunulmaması:*** Geliştirilen modelin (örneğin, bir regresyon modelinin katsayıları, bir karar ağacının kuralları veya eğitilmiş bir AI modelinin ağırlıkları) başkaları tarafından kullanılması veya incelenmesi için erişilebilir olmaması ya da modelin nasıl kullanılacağına dair yönergelerin sunulmaması önemli bir hatadır. Çözüm, modelin tam spesifikasyonunun (mümkünse matematiksel denklemi veya kural seti) veya eğitilmiş model dosyasının kendisinin (örneğin, herkese açık bir depo aracılığıyla) paylaşılması ve modelin nasıl kullanılacağına dair talimatların sağlanmasıdır. Bu yaklaşım, çalışmanın tekrarlanabilirliğini ve modelin pratik uygulamasını kolaylaştırır (Collins ve ark., 2015; Collins ve ark., 2024).

7.4. Yorumlama, Sınırlılıklar ve Etik Hususlarla İlgili Hatalar

- ***Modelin yorumlanabilirliğinin ihmal edilmesi:*** "Kara kutu" modellerde (özellikle derin öğrenme veya bazı LLM uygulamalarında) modelin nasıl tahmin yaptığına dair hiçbir açıklama sunulmaması, klinisyenlerin modele güvenmesini ve onu anlamasını zorlaştırır (Lambert ve ark., 2024). Çözüm olarak SHAP, LIME gibi modelden bağımsız yorumlama yöntemleri veya modele özgü dikkat mekanizmaları kullanılarak modelin hangi özellikleri ne ölçüde kullandığı gösterilmeli ve elde edilen açıklamalar klinik bağlamda anlaşılır şekilde yorumlanmalıdır (Collins ve ark., 2024). TRIPOD-LLM rehberi de LLM modelleri için benzer yorumlanabilirlik çabalarını teşvik etmektedir (Gallifant ve ark., 2025).
- ***Sınırlılıkların ve yanlılıkların yüzeysel tartışılması:*** Çalışmanın metodolojik sınırlılıkları, veri kaynaklarındaki potansiyel yanlılıklar (örneğin, seçilim yanlılığı veya spektrum yanlılığı) ve algoritmanın neden olabileceği yanlılıklar derinlemesine ele alınmadığında modelin güvenilirliği sorgulanır (Ferrara, 2024). Modelin farklı alt gruplardaki performansı değerlendirilmezse adalet (fairness) konusu belirsiz kalır. Çözüm, kapsamlı bir sınırlılıklar bölümü yazarak potansiyel yanlılık kaynaklarını ve bunların sonuçlara etkilerini dürüstçe değerlendirmektir. Modelin farklı demografik gruplardaki adilliği analiz edilip raporlanmalıdır (Collins ve ark., 2015; Collins ve ark., 2024). TRIPOD-LLM rehberi de LLM'lere özgü yanlılık ve sınırlılık tartışmalarını vurgular (Gallifant ve ark., 2025).
- ***Etik ve gizlilik konularının ihmal edilmesi:*** Etik kurul onayı, bilgilendirilmiş onam ve hasta verilerinin gizliliğini korumak için alınan önlemler hakkında bilgi verilmemesi; özellikle LLM modellerinde, modelin eğitim verilerini "ezberleyerek" hassas bilgileri sızdırma riskinin göz ardı edilmesi kritik bir hatadır (Omiye ve ark., 2024). Çözüm, çalışma için alınan etik kurul onayı ve hasta onamının belirtilmesi; hasta verilerinin gizliliğini korumak için uygulanan yöntemlerin (örneğin, verilerin anonimleştirilmesi) açıklanmasıdır (Collins ve ark., 2024). Modelin sorumlu geliştirilmesi ve kullanılması için ortaya çıkan etik çıkarımlar tartışılmalı; LLM'lerde olası veri ezberleme ve hassas bilgi sızdırma riskleri TRIPOD-LLM rehberi ışığında ele alınmalıdır (Gallifant ve ark., 2025).

Bu yaygın hatalardan kaçınmak, araştırmacıların TRIPOD+AI ve TRIPOD-LLM rehberlerindeki maddeleri titizlikle uygulamaları, çalışmalarını eleştirel bir gözle değerlendirmeleri ve raporlamada şeffaflığı

önceliklendirmeleri ile mümkündür. Ayrıca dergi editörleri ve hakemler de bu standartların uygulanmasını talep ederek iyi raporlama uygulamalarının alanda yerleşmesine önemli katkı sağlayabilir.

8. Gelecekteki Yönelimler ve Sonuç

Klinik AI ve ML alanı baş döndürücü bir hızla gelişmeye devam etmektedir. Yeni algoritmalar, daha büyük ve karmaşık veri kümeleri ve yenilikçi uygulamalar sürekli ortaya çıkmaktadır. Bu dinamik ortamda, geliştirilen tahmin modellerinin şeffaf, tekrarlanabilir ve güvenilir bir şekilde raporlanması, bilimin ilerlemesi ve bu teknolojilerin hasta yararına güvenli bir şekilde kullanılabilmesi için her zamankinden daha kritik bir öneme sahiptir (Moher ve ark., 2014; Van Calster ve ark., 2019).

TRIPOD+AI ve TRIPOD-LLM gibi rehberler bu hedefe ulaşmada önemli araçlardır. Ancak bu rehberler statik belgeler değildir; AI/ML alanındaki gelişmeler ve ortaya çıkan yeni zorluklar karşısında periyodik olarak güncellenmeleri ve genişletilmeleri gerekecektir. Önümüzdeki dönemde raporlama standartlarıyla ilgili aşağıdaki gelişmeler öngörülebilir:

- ***Daha Dinamik ve İnteraktif Rehberler:*** Raporlama rehberlerinin, araştırmacıların spesifik çalışmalarına yönelik öneriler sunabilen çevrimiçi araçlar veya yazılımlar haline gelmesi beklenebilir.
- ***Federatif Öğrenme ve Gizlilik-Koruma Yöntemleri İçin Uzantılar:*** Birden fazla kurumun verisini merkezi bir havuz oluşturmadan birleştiren ve veri gizliliğini koruyarak model geliştirmeyi mümkün kılan *federatif öğrenme* gibi teknikler yaygınlaştıkça, bu modeller için özel raporlama standartlarına ihtiyaç duyulacaktır. Benzer şekilde, homomorfik şifreleme ve diferansiyel gizlilik gibi gizliliği koruyan AI uygulamaları için de rehberlerde yeni maddeler geliştirilmesi gerekebilir.
- ***Klinik Uygulama ve İzleme Odaklı Raporlama:*** Modelin geliştirilip doğrulanmasının ötesinde, gerçek dünyada klinik uygulamaya nasıl entegre edildiği, kullanım sonrasında performansının nasıl izlendiği ve modelin zamanla güncelliğini yitirmemesi (model drift) için nasıl yönetildiği konularında daha ayrıntılı raporlama standartları talep edilebilir. Örneğin "TRIPOD-IMPLEMENT" gibi, uygulama ve izlemeye odaklı bir uzantı rehberi gündeme gelebilir.
- ***Hasta ve Kamu Katılımının Raporlanması:*** AI modellerinin geliştirme ve değerlendirme süreçlerine hasta ve toplum temsilcilerinin katılımı giderek daha fazla önem kazanmaktadır. Bu katılımın nasıl

sağlandığı ve sonuçlara etkisinin nasıl olduğu konusunda raporlama standartları oluşturulabilir.

- ***AI Etiği ve Adillik İçin Derinlemesine Kılavuzlar:*** Algoritmik yanlılık, adillik ve hesap verebilirlik konuları ön plana çıktıkça, bu hususların değerlendirilmesi ve raporlanmasına dair daha detaylı ve spesifik rehberlere ihtiyaç duyulacaktır.

Sonuç olarak, TRIPOD+AI ve TRIPOD-LLM rehberleri klinik AI ve ML tahmin modellerinin raporlanmasında güncel en iyi uygulamaları temsil etmektedir. Bu rehberlere uyum, araştırma kalitesini yükseltir, bilimsel iletişimi kolaylaştırır, yanlılık riskini azaltır ve nihayetinde hastalar için daha güvenli ve etkili AI çözümlerinin geliştirilmesine katkı sunar. Biyoistatistikçiler, veri bilimciler, klinisyenler ve araştırmacılar olarak hepimizin sorumluluğu bu standartları benimsemek, uygulamak ve sürekli iyileştirilmelerine katkıda bulunmaktır. Şeffaflık ve titizlik, AI'nin tıptaki vaadini gerçeğe dönüştürmenin temel taşlarıdır. Bu rehberlerin sağladığı çerçeve sayesinde sadece daha iyi makaleler yazmakla kalmayacak, aynı zamanda daha sağlam bilim yaparak toplum sağlığına daha büyük katkılar sunabileceğiz.

Referanslar

Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. *PLoS Med.* 2012;9(5):e1001216. doi:10.1371/journal.pmed.1001216

Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, Pimenta D. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit Med.* 2025;8(1):274. doi:10.1038/s41746-025-01670-7

Bathula A, Gupta SK, Merugu S, Saba L, Khanna NN, Laird JR, ve ark. Blockchain, artificial intelligence, and healthcare: the tripod of future—a narrative review. *Artif Intell Rev.* 2024;57(9):238. doi:10.1007/s10462-024-10873-5

Bischl B, Binder M, Lang M, Pielok T, Richter J, Coors S, Lindauer M. Hyperparameter optimization: foundations, algorithms, best practices, and open challenges. *WIREs Data Min Knowl Discov.* 2023;13(2):e1484. doi:10.1002/widm.1484

Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med.* 2018;378(11):981-983. doi:10.1056/NEJMp1714229

Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, ve ark. TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. doi:10.1136/bmj-2023-078378

Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* 2015;350:g7594. doi:10.1136/bmj.g7594

Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digit Med.* 2025;8:143. doi:10.1038/s41746-025-01476-7

Ding J, Du J, Wang H, Xiao S. A novel two-stage feature selection method based on random forest and improved genetic algorithm for enhancing classification in machine learning. *Sci Rep.* 2025;15:16828. doi:10.1038/s41598-025-01761-1

Edin J, Johansson F. Automated medical coding on MIMIC-III and MIMIC-IV: a critical review and replicability study. *Proc ACM SIGIR Conf Res Dev Inf Retr.* 2023:2572-2582. doi:10.1145/3539618.3591918

Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, ve ark. A guide to deep learning in healthcare. *Nat Med.* 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z

Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. *Sci.* 2024;6(1):3. doi:10.3390/sci6010003

Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, ve ark. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med.* 2025;31(1):60-69. doi:10.1038/s41591-024-03425-5

Heus P, Reitsma JB, Collins GS, Damen JA, Scholten RJ, Altman DG, ve ark. Transparent reporting of multivariable prediction models in journal and conference abstracts: TRIPOD for abstracts. *Ann Intern Med.* 2020;173(1):42-47. doi:10.7326/M20-0193

Ilemobayo A, Durodola O, Alade O, Awotunde OJ, Adewumi TO, Falana O, Ogungbire A, Osinuga A, Ogunbiyi D, Ifeanyi A, Odezuligbo IE, Edu OE. Hyperparameter tuning in machine learning: a comprehensive review. J Eng Res Rep. 2024;26(6):388-395. doi:10.9734/jerr/2024/v26i61188

Karrar AE. The effect of using data pre-processing by imputations in handling missing values. *Indones J Electr Eng Inform.* 2022;10(2):375-384. doi:10.52549/ijeei.v10i2.3730

Lambert B, Forbes F, Doyle S, Dehaene H, Dojat M. Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artif Intell Med.* 2024;150:102830. doi:10.1016/j.artmed.2024.102830

Marey A, Arjmand P, Alerab ADS, Eslami MJ, Saad AM, Sanchez N, ve ark. Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. *Egypt J Radiol Nucl Med.* 2024;55(1):183. doi:10.1186/s43055-024-01356-2

Moher D, Altman DG, Schulz K, Simera I, Wager E. Importance of transparent reporting of health research. In: *Guidelines for Reporting Health Research: A User's Manual.* Chichester: Wiley; 2014:4-6. doi:10.1002/9781118715598.ch1

Moons KGM, Altman DG, Reitsma JB, Vergouwe Y, Mallett S, Collins GS, ve ark. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-W73. doi:10.7326/M14-0698

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453. doi:10.1126/science.aax2342

Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med*. 2024;177(2):210-220. doi:10.7326/M23-2772

Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432

Salih AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, ve ark. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv Intell Syst*. 2025;7(1):2400304. doi:10.1002/aisy.202400304

Shool S, Adimi S, Amleshi RS, Bitaraf E, Golpira R. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak*. 2025;25:117. doi:10.1186/s12911-025-02954-4

Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, ve ark. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381. doi:10.1371/journal.pmed.1001381

Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Min Knowl Discov*. 2020;10(5):e1379. doi:10.1002/widm.1379

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8

Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7

Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. doi:10.1186/s12916-019-1466-7

Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare (Basel)*. 2025;13(6):603. doi:10.3390/healthcare13060603

Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv e-prints*. 2023;arXiv:2306.09968. doi:10.48550/arXiv.2306.09968

Xiao X, Zhang Y, Xu J, Ren W, Zhang J. Assessment methods and protection strategies for data leakage risks in large language models. *J Ind Eng Appl Sci*. 2025;3(2):6-15. doi:10.70393/6a69656173.323736